Uncertainty Estimation of Large Language Model Replies in Natural Sciences

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

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

Motivation

- **LLMs in science workflows:** Potential efficiency gains in literature review, knowledge comprehension, and data analysis.
- **Hallucination risk:** LLMs produce confidently phrased but factually incorrect outputs that are hard to identify.
- **Role of Uncertainty Quantification (UQ):** Automated UQ methods can flag low-confidence or potentially erroneous generations, supporting safer downstream use.
- **Benchmarking objective:** Develop a framework for reproducible benchmarking of uncertainty metrics; asses token-level calibration and reliability of sequence-level UQ metrics in scientific question answering using the framework.

Summary of Results

- **First large-scale UQ benchmark for LLMs in scientific QA**, covering token-level calibration and multiple sequence-level metrics.
- **Instruction tuning polarizes token probabilities:** Models become overconfident regardless of correctness, undermining token-level UQ.
- **Sequence-level findings**
  - **Verbalization-based metrics** are unreliable
  - **Frequency of Answer** better aligns with correctness but demands heavy computation and semantic clustering.
  - **Claim-Conditioned Probability** fails at token level due to vanishing confidence over long generations and poor semantic equivalence detection.
- **LM-Polygraph framework [2] re-engineered** into a modular, extensible framework for scalable, reproducible UQ benchmarking.

Implementation of Benchmarking Framework

- **Foundation:** Extends **LM-Polygraph** [2], which encapsulates each uncertainty metric computation steps in classes, defining required and provided dependencies.
- **Configurable execution nodes:** Generalizes modularity by representing every processing step as a runtime-configurable node.
- **Dynamic DAG construction:** Builds an acyclic directed graph at runtime from user specifications, enabling optimized execution order.
- **Performance optimizations:** Supports asynchronous processing, result reuse via caching, and batching to enable efficient computations.
- **Layered execution wrapper:** Manages resource allocation and persistent, disk-backed caching for scalable, reproducible benchmarking of large datasets.

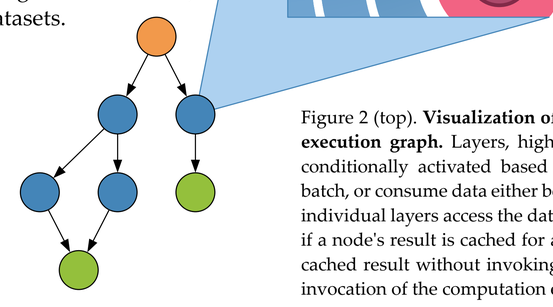


Figure 1 (right). **Abstract Visualization of the Order of Computation Steps as Acyclic Directed Graph in Reworked LM-Polygraph.** This features the dataset (orange), intermediate computation steps (blue) and calculation of the final uncertainty metric scores (green).

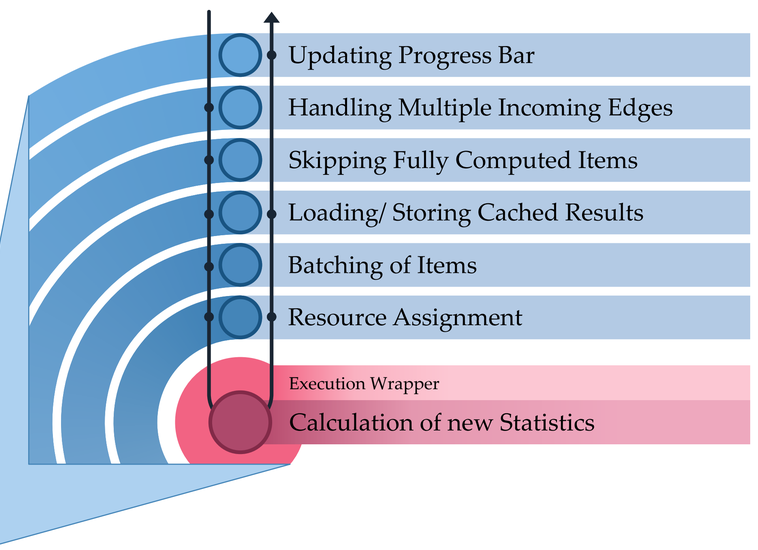


Figure 2 (top). **Visualization of data items passing through layers of nodes in the execution graph.** Layers, highlighted in blue, are applied as decorators and are conditionally activated based on the configuration. They may transform, filter, batch, or consume data either before or after computation. 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.

Token Level Calibration

Previous Work

- **Calibration comparison in GPT-4 Technical Report [1]:** Evaluated base vs. instruction-tuned GPT-4 using MMLU, rephrasing questions as classification task.
- **Confidence measurement:** Used label probabilities (A/B/C/D) as confidence scores.
- **Key findings:** Base GPT-4 showed strong calibration, instruction-tuned GPT-4 exhibited degraded calibration (see Figure 3).
- **Underlying hypothesis:** Instruction tuning shifts token-probability distributions away from the original training data "ground truth", degrading calibration significantly.

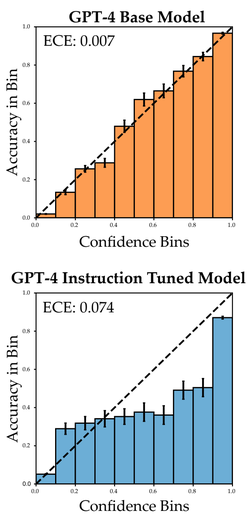


Figure 3 (right). **Calibration plots of the base model (top) and the instruction tuned (bottom) GPT-4 model on a subset of the MMLU dataset.** On the x-axis are bins according to the model's confidence (logprob) in each of the A/B/C/D choices for each question; on the y-axis is the accuracy within each bin. The dotted 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

- **Datasets:** Four multiple-choice datasets - MMLU, ArcReasoning, GSM8K, GPQA
- **Models:** three model-size pairs (7 B, 24 B, 70 B), each with a base and instruction-tuned version.
- **Confidence scoring:** Treated label probabilities as model confidence scores, both raw and normalized (excluding mass on non-label tokens).
- **Calibration analysis:** Compared calibration performance (e.g., calibration plots, ECE) between base vs. instruction-tuned models and between raw vs. normalized confidence.

Results

- **Normalization is essential:** Raw label probabilities are undermined by overall task comprehension and yield poor calibration; normalized scores enable meaningful confidence estimates.
- **Reasoning complexity raises ECE:** Calibration error grows with reasoning demands — GSM8K and GPQA exhibit higher ECE — highlighting token probabilities' limitations for multi-step and symbolic reasoning.
- **Fact Retrieval vs. Complex Tasks:** Token-level probabilities reliably capture aleatoric uncertainty in fact retrieval but become overconfident and inadequate for tasks that demand reasoning.
- **Mixed impact of instruction tuning on ECE:** Mistral 7B and Llama 70B show degraded ECE post-tuning, whereas ECE for Mistral 24B remains largely unchanged.
- **Universal polarization of token probabilities:** All instruction-tuned models concentrate probability mass on a single label, reducing nuance and reliability of token probabilities as confidence scores (see Figure 4).

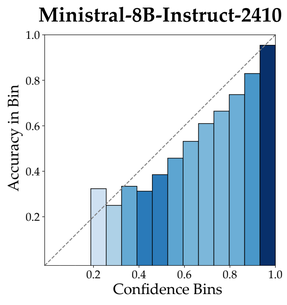
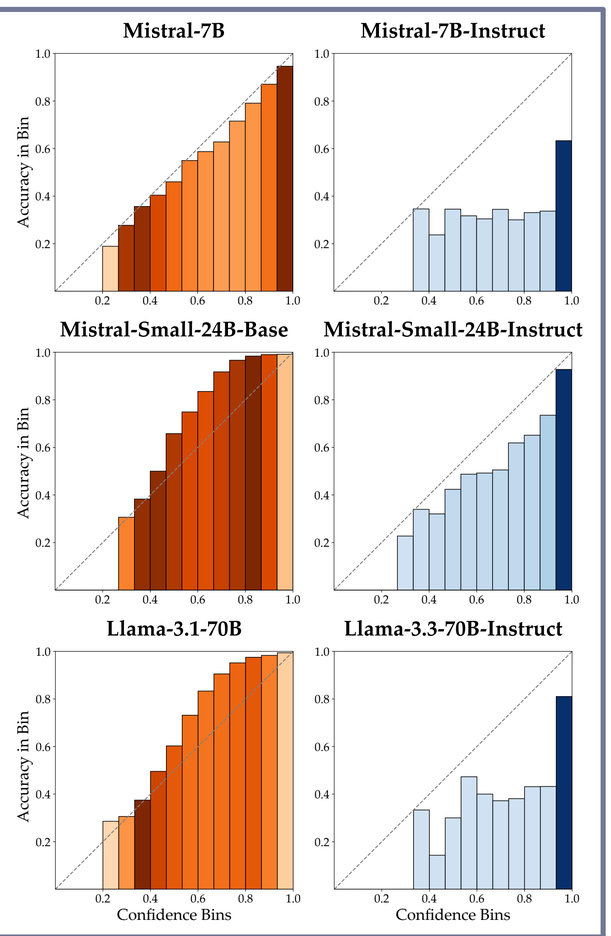


Figure 4 (right). **Impact of Instruction Tuning on Calibration (MMLU)**. Calibration plots show normalized entropy of label probabilities, focusing on the most likely label. Orange curves represent base models; blue curves show instruction-tuned versions. Darker shades indicate higher item counts per bin. Instruction tuning shifts counts toward the highest-confidence bin, implying reduced probability for other labels (not shown).

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.



Sequence Level Calibration of Uncertainty Metrics

Experiment Design

- **Models & datasets:** Benchmarked four instruction-tuned models (7B, 8B, 24B, 70B) on eight scientific QA datasets—five multiple-choice and three arithmetic—each with known ground truth.
- **Prompting strategies:** APriCoT [3] for multiple-choice and CoT for arithmetic tasks.
- **Sampling & scale:** Subsampled each dataset to 1,000 items and sampled 10 outputs per prompt, totaling 181,360 QA prompts per model.
- **Metrics evaluated:** Compared four uncertainty metrics, as highlighted on the right.

Verbalized Uncertainty [4]

Metric Explanation: Verbalized Uncertainty prompts the model, immediately after answering, to output a numeric confidence score

- **Biased score distribution:** Models consistently produced a narrow range of high verbalized confidence values (e.g., 0.0, 0.5, 0.9, 1.0, see Figure 6), likely influenced by training and instruction tuning.
- **Unreliable calibration:** Verbalized scores showed no meaningful correlation with answer correctness, indicating they are unreliable proxies for true model confidence.

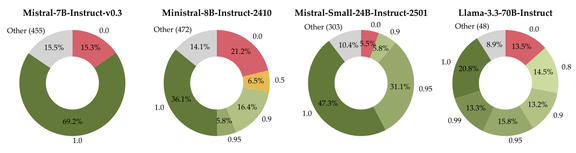
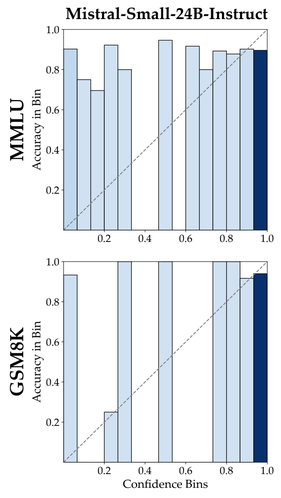


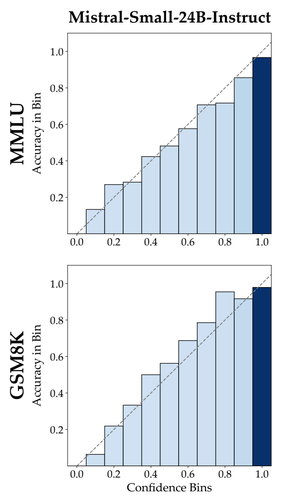
Figure 6. **Confidence Score Distribution for Verbalized Uncertainty per Model.**



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.

- **Reliable Calibration:** Higher answer frequency consistently aligns with correctness across task types, with increased diversity indicating uncertainty in harder datasets, yielding well-calibrated confidence estimates.
- **Limited applicability:** High sampling cost and reliance on semantic clustering make the method impractical for open-ended QA.



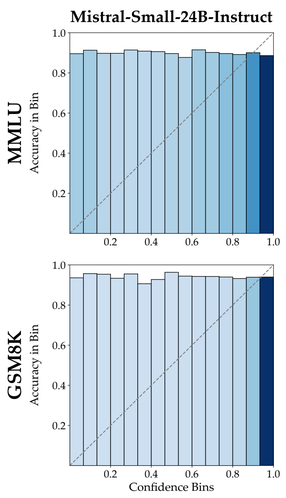
Summary of Results

- **Frequency of Answer** was the only metric providing well-calibrated confidence closely aligned with answer correctness.
- **Verbalized-based metrics** showed strong biases and weak correlation with accuracy.
- **Claim-Conditioned Probability** fails on on long form answers
- **Semantic consistency validated:** Frequency of Answer supports it as a useful signal but is limited by high sampling cost and complex semantic clustering, making it unsuitable for open-ended QA.
- **Conclusion:** Results highlight the need for more efficient and robust uncertainty metrics and general benchmarking thereof.

P(True) [5]

Metric Explanation: 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.

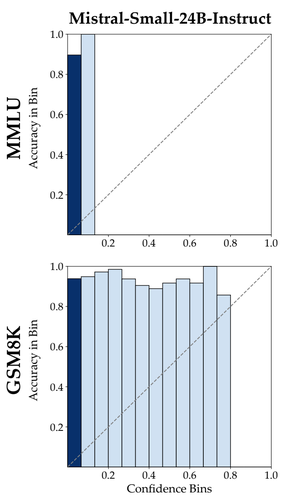
- **Polarized scores:** P(True) yields highly skewed confidence distributions, with most scores near 1.0 and minimal use of intermediate values, reflecting model response bias.
- **Unreliable calibration:** Confidence scores show little correlation with correctness, making P(True) an unreliable measure of uncertainty.



Claim-Conditioned Probability [6]

Metric Explanation: 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.

- **Unstable scores:** CCP confidence vanishes on longer outputs due to multiplicative token aggregation, and is further destabilized by NLI errors during semantic token equivalence checks and the inclusion of stop words.
- **Unreliable calibration:** Confidence scores show no alignment with correctness, making CCP unreliable for sequence-level uncertainty estimation.



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

TECHNISCHE UNIVERSITÄT DRESDEN

with Resources generously provided by

JÜLICH Forschungszentrum