BENCHMARKING UNCERTAINTY ESTIMATION IN LARGE LANGUAGE MODEL REPLIES FOR NATURAL SCIENCE QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are commonly used in question answering (QA) settings, including natural science and related research domains. Reliable uncertainty quantification (UQ) is critical for the trustworthy uptake of generated answers, yet existing approaches remain insufficiently validated in scientific QA. We introduce the first large-scale benchmark for evaluating UQ metrics in this setting, providing an extensible open-source framework to assess calibration across diverse models and datasets. Our study spans eleven LLM models in base, instruction-tuned and reasoning variants and covers eight scientific QA datasets, including both multiple-choice and arithmetic question answering tasks. At the token level, we find that instruction tuning induces strong probability mass polarization, reducing the reliability of token-level confidences as estimates of uncertainty. At the sequence level, we show that verbalized uncertainty estimates are systematically biased and poorly correlated with correctness, while answer frequency (consistency across samples) yields the most reliable calibration, albeit at high computational cost. These findings expose critical limitations of current UQ methods for LLMs and highlight concrete opportunities for developing scalable, well-calibrated confidence measures for scientific QA.

1 INTRODUCTION

Large language models (LLMs) have rapidly emerged as powerful tools for natural language processing, understanding and generation. Among their diverse applications, they are increasingly deployed in chat-based assistants for question answering (QA), automated agent-based systems, and serving as surrogates for traditional search engines (Jin et al., 2025; Xiong et al., 2024a; Kelly et al., 2023). Within this landscape, scientific QA constitutes a particularly critical and challenging task. It spans a wide range of use cases, from public science communication (Schäfer, 2023) and education across different levels of expertise (Welbl et al., 2017), to supporting research and knowledge discovery (Gu & Krenn, 2025). In these contexts, accuracy and trustworthiness are essential, as errors may misinform the public, impair learning, or distort scientific practice. A central obstacle to reliability is the phenomenon of hallucinations, where LLMs generate fluent and seemingly confident answers that are factually incorrect or misleading (Ji et al., 2022). While hallucinations are now widely recognized, effective methods to detect and mitigate them remain underdeveloped, particularly in high-stakes domains such as science.

Research into UQ aims to develop reliable, automated methods for quantifying how certain a model is in its own predictions (Guo et al., 2017). In the context of LLMs, UQ serves as a critical tool to mitigate hallucinations: by identifying outputs with high uncertainty, systems can flag potentially erroneous responses, abstain from answering, or route queries to alternative mechanisms – i.e. larger models, retrieval-augmented systems, or human reviewers. Beyond error mitigation, reliable UQ enhances transparency and user trustworthiness by providing interpretable indicators of answer reliability and soundness (Dhuliawala et al., 2023; Devic et al., 2025; Reyes et al., 2025).

2 KEY CONTRIBUTIONS

To address open questions regarding the applicability of UQ in scientific QA, this work benchmarks a range of uncertainty methods with a qualitative and quantitative focus on calibration. With this, we strive to fill gaps in available studies which we highlight in Section 3. We target structured tasks such as multiple-choice and arithmetic questions to enable rigorous ground truth verification. The evaluation centers on physics QA datasets as a representative natural science domain and controllable environment with strong potential for generalisation (Zhang et al., 2024b). We benchmark a selected subset of uncertainty methods across open-weight models. By evaluating how well these UQ methods capture the reliability of model predictions, we aim to characterize model behavior and limitations in a scientific context. In addition, we also investigate the most common information source for LLM uncertainty estimates: token probabilities.

We approach this task by identifying key conceptual approaches to uncertainty estimation in LLMs. We conduct a qualitative assessment of representative UQ methods. These are subsequently empirically evaluated on scientific QA datasets. We then study this data to answer two research questions:

- **(RQ1)** To what extent are the token probabilities a calibrated measure of confidence and what effect does the instruction-tuning or reasoning process have on the calibration?
- **(RQ2)** How do different UQ methods compare in reliable estimation of uncertainty in scientific QA?

Finally, to facilitate ongoing research and rapid iteration as LLMs and UQ methods continue to evolve, we introduce a flexible and extensible framework for LLM benchmarking. We release this framework together with the implementation of benchmarks for calibration assessment presented in this paper, detailed reproduction instructions, leaf-node results from our benchmark runs, and additional visualizations in an open-access repository accompanying this paper[1].

Our paper is structured as follows: After the introduction and key contributions of Sections 1 to 2, we highlight the epistemic background of our analysis in Section 3. We present related work in Section 4, which in turn motivated our experimental setup of Section 5. We then describe our findings on token probability calibration in Section 6 (*RQ1*) and benchmark sequence-level uncertainty calibration in Section 7 (*RQ2*). Our report concludes with limitations in Section 8, conclusions in Section 9 and a brief outlook in Section 10.

3 BACKGROUND

3.1 CHALLENGES IN SCIENTIFIC QUESTION ANSWERING

Scientific QA involves unique challenges stemming from the structured, coherent and complex nature of scientific knowledge at large. Questions often require interpreting numerical values expressed in various lexical forms (“eighteen” instead of “18”), units (“pounds” instead of “lbs”), or domain-specific notations like chemical formulae and mathematical representations. In addition to parsing these elements, models must recall relevant formulae, relate their parameters, and grasp the relationships among variables. Many problems involve non-linear functions or require rearranging equations before applying them. Since LLMs do not possess intrinsic arithmetic capabilities, they depend on pattern recognition from training data or external tools such as code execution to perform calculations accurately. Multi-step reasoning as often required in scientific QA tasks are particularly difficult as errors in earlier steps propagate downstream and compromise the final result.

3.2 UNCERTAINTY QUANTIFICATION IN PREDICTIVE MODELS

Neural networks, including LLMs, face predictive uncertainty as their training data provide only a discrete and incomplete mapping of real-world artifacts (Hüllermeier & Waegeman, 2021).The paradigm of negative log likelihood training for next token prediction (Radford et al., 2018) in LLM pre-training enforces uncertainty of generated tokens. Besides obtaining this predictive uncertainty

[1] https://anonymous.4open.science/r/llm-uncertainty-bench-9B2B/

with dedicated methods (Gal et al., 2016), the act of validating the quantitative soundness of uncertainties comes as a challenge to many practitioners (Guo et al., 2017; Chung et al., 2021).

The field of UQ methods in LLMs is still nascent compared to methods for classification or regression tasks (Kendall & Gal, 2017; Kuleshov et al., 2018; Papamarkou et al., 2024) in general. The standard evaluation technique for UQ methods are calibration plots (calibrations for short), which visualise how well predicted confidence scores correlate to the true likelihood of correctness (Guo et al., 2017; Hendrycks & Gimpel, 2018). Many quantitative approaches rely on information-theoretic metrics, such as entropy or perplexity, but these often fail to capture language-specific nuances and require further validation in the LLM context. Recently, LLM-specific uncertainty estimation methods have emerged, such as Verbalized Uncertainty, P(True), and Claim Conditioned Probability (CCP) (see details in Section 3.3 and Section 7.1). However, they are studied in narrow domains, such as factual QA on biographies or encyclopedic content (Fadeeva et al., 2024). Consequently, it remains unclear how well these approaches generalize to scientific QA. Compounding the issue, current benchmarks are often tightly coupled to specific models, datasets, and UQ methods, limiting adaptability to rapidly evolving LLM architectures and use cases.

3.3 UQ METHODS AND CALIBRATION

A straight forward way of estimating uncertainty in LLMs is to reformulate multiple-choice question-answering (MCQA) items as classification tasks, prompting the model to output a single label token (e.g., A/B/C/D) representing the chosen answer.The confidence scores are derived from the probabilities assigned to each label. These probabilities can be used as confidence scores directly, i.e. excluding probability mass from non-label tokens. Alternatively, a normalization with respect to the total probability mass assigned to all possible answer labels is applied. By normalizing over the label set, the resulting confidence scores represent only relative preferences among the labels – not (un-)certainty in the options. This may obscure uncertainty that would otherwise be expressed through low absolute probabilities, making the normalized scores less reliable as measures of uncertainty (Wang et al., 2024a). The authors of (OpenAI, 2023) compared the calibration of responses from base model and instruction-tuned GPT-4 using this approach. Their results suggested good calibration in the base model, but significantly worse calibration in the instruction-tuned model. This sparked a controversy about the influence of instruction-tuning on the calibration of models.

We have conducted an extensive analysis of existing UQ methods, see Section A.4. Only methods producing normalized sequence-level uncertainty scores are included in our analysis to enable reliability UQ validation by calibration. Subsequently, we focus our work on the following UQ methods: *Verbalized Uncertainty* (Tian et al., 2023), *P(True)* (Kadavath et al., 2022), *Frequency of Answer* (Wang et al., 2023), *Claim-Conditioned Probability (CCP)* (Fadeeva et al., 2024). More details of these methods are discussed in Section 7.1.

4 RELATED WORK

UQ is particularly critical for LLMs because their token-level probabilistic generation can produce fluent yet confidently incorrect or misleading outputs known as hallucinations (Maynez et al., 2020). These hallucinations, classified as intrinsic contradictions or extrinsic fabrications, complicate trustworthiness and highlight the need for robust uncertainty detection (Sui et al., 2024; Zhang et al., 2023b; Banerjee et al., 2024). Many methods have been established to obtain predictive uncertainties for LLM generated text (Geng et al., 2024).

Token-level uncertainty estimation faces challenges due to varying semantic importance of tokens, while many methods assume equal token importance, leading to misrepresentations (Ullrich et al., 2025; Kuhn et al., 2023a). Linguistic calibration through epistemic markers (e.g., “might,” “potentially”) provides interpretable uncertainty cues, but models tend to be overconfident in these expressions, risking user over-reliance (Band et al., 2024; Zhou et al., 2024). Empirical work shows that base LLMs are generally better calibrated than instruction-tuned models, which often become overconfident and produce polarized token probability distributions that impair uncertainty representation (OpenAI, 2023; Tian et al., 2023; Cruz et al., 2024; Wang et al., 2025). Prompt design significantly affects uncertainty and calibration: small prompt variations, the use of epistemic phrasing, and simulating knowledge profiles influence both accuracy and confidence. Strategies have been

proposed to mitigate overconfidence in self-evaluations (Cao et al., 2024; He et al., 2024; Sclar et al., 2024; Zhou et al., 2023; Lu & Wang, 2024; Xiong et al., 2024b). This body of work underscores the complexity of uncertainty estimation in LLMs and the need for context-aware UQ methods.

First comprehensive studies of UQ effectiveness as well as their calibration were undertaken only recently by Huang et al. (2025) and Fadeeva et al. (2023a). For the latter, the authors share open-source UQ benchmark tooling (LM-Polygraph) publicly, underlining the importance of UQ analyses as LLM architectures progress rapidly. Both studies rate the effectiveness of UQ methods by virtue of two and one summary statistic respectively. This stands in contrast to existing best practices (calibration plots) to evaluate the quality of predictive uncertainties (Guo et al., 2017) both quantitatively and qualitatively.
We also observe early studies of calibration as a measure of quality control for UQ in LLMs: Tao et al. (2025) presented a vast analysis with respect to LLM models explored, but restrict themselves to one single dataset only. Liu et al. (2025) surveyed a variety of published UQ methods and presented available datasets. But they did not perform experiment for an empirical comparison. Multiple reports around the LM-Polygraph project (Fadeeva et al., 2023a) offer a structured comparison of UQ methods along axes like logit access, computational cost, and granularity of uncertainty of up to 28 uncertainty methods, but they focus on one single summary statistic to survey methods. Our work aspires to improve on this and present a comprehensive suite of results for multiple UQ methods, multiple datasets and LLMs, which is reproducible, fast and provides robust evidence with respect to qualitative and quantitative assessments.

5 General Experimental Setup

5.1 Dataset Selection

To ensure reliable and verifiable evaluation of uncertainty methods, we use multiple-choice question answering (MCQA) and arithmetic question answering (arithmetic QA) as our primary task formats. These formats provide either structured answer options with a single correct choice (MCQA) or verifiable numeric solutions (arithmetic QA), allowing for objective, automated correctness checks, unlike open-ended QA which often requires subjective human evaluation.

Physics is chosen as the core domain due to its foundational role in the natural sciences, with supporting datasets from mathematics to capture formula-based reasoning and calculation. To study the effect of task complexity, the selected datasets span a range of difficulty levels and cognitive demands, from fact retrieval to multi-step reasoning.

The following provides an overview over seleted datasets (more details provided in Section A.3).

MMLU (Hendrycks et al., 2021) is a multiple-choice benchmark with $15,908$ questions across 57 academic subjects, including physics at varying levels. It is primarily testing comprehension, factual knowledge and single step reasoning.

ARC (Clark et al., 2018) tests scientific reasoning using grade-school science exam questions split into ARC-Easy and ARC-Challenge subsets. The latter emphasizes multi-step reasoning, making it a strong benchmark for evaluating UQ methods under complex conditions.

SciQ (Welbl et al., 2017) contains $13,679$ multiple-choice questions in physics, chemistry, and biology. It emphasizes conceptual understanding.

GPQA (Rein et al., 2023) is a graduate-level science benchmark with 448 expert-written questions designed to resist simple lookup. Its high difficulty and reasoning demands make it particularly useful for stress-testing UQ methods.

GSM8K (Cobbe et al., 2021) is a standard math word problem dataset with 8,500 arithmetic questions requiring step-by-step symbolic reasoning.

GSM-MC (Zhang et al., 2024b) is a multiple-choice variant of GSM8K, using model-generated distractors to reduce evaluation ambiguity.

SVAMP (Patel et al., 2021) is an arithmetic reasoning dataset containing 1,000 questions that introduce distracting information in the problem text. While featuring low to moderate computational complexity, these distractors are specifically designed to induce ambiguity and decision uncertainty,

challenging models can recognize and quantify uncertainty in the presence of misleading or irrelevant information.

SciBench (Wang et al., 2024b) comprises 692 college-level questions from math, chemistry, and physics textbooks. It targets advanced symbolic reasoning involving formulas and physical units.

5.2 Model Selection

Models were chosen to cover a broad spectrum of LLMs while ensuring reproducibility through the use of open-weight, publicly available models. To capture diversity in design, the selection spans five major providers (OpenAI, Mistral, Meta, Qwen, and Google) and includes models of varying sizes (from 7B to 70B parameters), types (base, instruction-tuned and reasoning models) and architectural designs, such as Mixture-of-Experts. To research the effect of instruction tuning on the calibration of label probabilities, instruction-tuned and reasoning models are complemented by their base model counterparts to enable controlled comparison. A detailed list of selected models can be found in Section A.2.

6 Benchmarking Label Probability Calibration

6.1 Methodology / Experimental Setup

While confidence scores represent a model's self-assessed probability of correctness, effective UQ requires calibration methods that align these scores with empirical accuracy (Guo et al., 2017). Expected Calibration Error (ECE) is a widely used metric for this purpose, comparing binned confidence estimates with actual correctness.

To ensure comparability while extending prior work, the experimental setup was designed to remain close to the GPT-4 Technical Report (OpenAI, 2023), with expansions in model coverage, dataset diversity, and calibration analysis. We evaluated all selected base, instruction-tuned and reasoning models across four MCQA datasets, chosen to represent increasing reasoning complexity: factual and single-step reasoning (MMLU), symbolic arithmetic reasoning (GSM8K), and multi-step reasoning (ARC-Reasoning, GPQA). Initial tests revealed substantial differences in task comprehension between base and instruction-tuned models. To address this, we designed four alternative prompts (see Section A.5.1), each using three-shot prompting to ensure task comprehension. We then selected the prompt that maximized the probability mass assigned to label tokens, thereby ensuring consistent task adherence across models.

Calibration performance was then assessed by comparing the model variants and raw and normalized label probabilities as confidence scores, using calibration plots and summary metrics such as ECE.

6.2 Key Findings

Figure 1 shows a subset of results that illustrate our key findings, which we discuss below. All results shown use Prompt Design 1 (see Section A.5.1), which was found to optimize task comprehension across model types (see Section A.5.2). Comprehensive plots showing results for all models, datasets and prompt designs are included in Section A.5.4.

While we had previously hypothesised that using raw label probabilities may enable the model to express general uncertainty, our results show that the task comprehension poses a bigger impact on the total probability mass assigned to label tokens. As a result, we find that raw probabilities are confounded by overall task comprehension and yield misleading calibration, while normalization (excluding non-label tokens) enables meaningful confidence estimates (see Figure A.3).

We further find that ECE increases with reasoning complexity (see Figure 1). Tasks requiring symbolic or multi-step reasoning, such as GSM8K and GPQA, exhibit substantially higher ECE compared to factual retrieval tasks, such as MMLU. This distinction highlights that token-level probabilities can reliably capture aleatoric uncertainty in fact retrieval, but become overconfident and unreliable when reasoning is required as a result of epistemic uncertainty.

Comparing base and instruction-tuned models, we reproduce the degradation of ECE observed by (OpenAI, 2023), although not uniformly across models to the same degree. However, visual calibra-

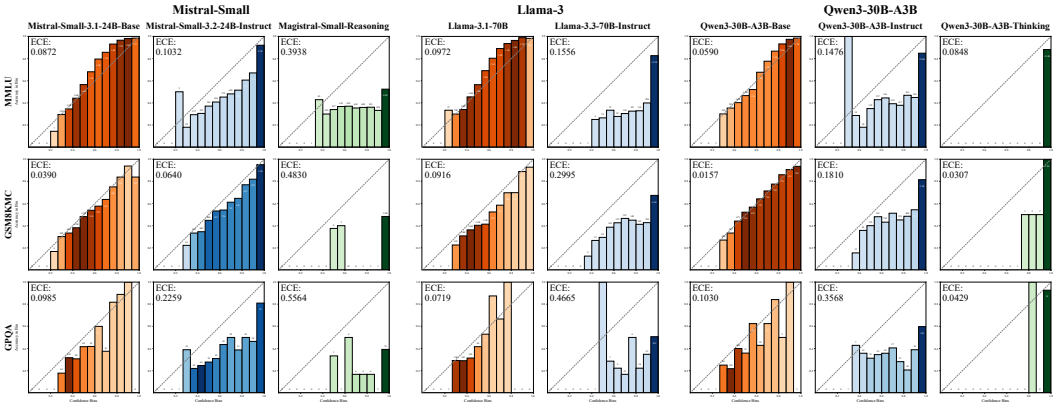


Figure 1: **Calibration Plots Using Normalized Label Probabilities and Only the Most Probable Label Per Prompt.** Columns correspond to three selected model families: base variants are shown in orange, instruction-tuned variants in blue, and reasoning variants in green. Rows refer to three QA datasets on MMLU, GSM8KMC and GPQA. Darker shading indicates a higher number of items within each confidence bin. Each bin is labelled with the sample count contained. The plot shows the results using Prompt Design 1.

tion analysis reveals systematic polarization of token probabilities. Instruction-tuned models tend to concentrate nearly all probability mass on a single label compared to their base model counterparts, thereby degrading the expressiveness of token-level confidence scores. This effect is evident in Figure 1, which shows a pronounced accumulation of items in the highest-confidence bucket. This effect is even more pronounced in reasoning models, where the reasoning process prior to responding with a label appears to commit to a single option, leaving minimal probability mass on the other options. The degree of polarization is consistent across most model families, with the notable exception of certain Mistral models, such as *Ministral-8B-Instruct* (see section A.5.4) and *Mistral-Small-3.2-24B-Instruct* (see Figure 1). Whether this difference arises from architectural design choices or from specific fine-tuning datasets and paradigms remains a subject of future research.

6.3 Issues with ECE

The observed contrast between mixed levels of ECE degradation and the uniform polarization that diminishes the practical usefulness of calibration scores reveals a fundamental limitation of ECE as a calibration metric. Although ECE measures the correlation between predicted confidence and the likelihood of correctness, it is not independent of the model's overall accuracy. As a result, the ECE of highly overconfident models that consistently produce high confidence scores is largely determined by the model's overall accuracy rather than the reliability of confidence estimates for individual predictions. Consequently, as model accuracy increases, ECE may appear deceptively low even when the model remains poorly calibrated at the instance level. This dependence undermines the utility of ECE for UQ, where the primary goal is to flag potentially incorrect answers through high uncertainty. This limitation is particularly relevant for LLMs, which suffer from overconfident in their generations (Zhou et al., 2024). As a result, we advocate for the use of ECE only in combination with visual assessment of calibration plots and complementary summary metrics.

7 Benchmarking Sequence-level calibration

7.1 Selection of Uncertainty Measures

The selection of the following four methods emphasizes relevance to scientific QA, theoretical soundness, computational feasibility, and empirical promise.

Verbalized Uncertainty (Tian et al., 2023) prompts the model subsequent to the answer generation to provide a self-assessed probability estimate of correctness in token space. While straightforward, it ignores token probability distributions and may be susceptible to training data bias.

P(True) (Kadavath et al., 2022) prompts the model subsequent to the answer generation to classify the answer as "(A) True" or "(B) False". The underlying token probabilities assigned to the corresponding labels "(A)" and "(B)" are then used as confidence scores.

Frequency of Answer estimates certainty by the proportion of semantically equivalent answers among multiple sampled generations for the same prompt. This is computationally expensive and semantic equivalence detection of answers is non-trivial in open ended questions answering. However, it captures semantic consistency and proxies other approaches like self-consistency prompting (Wang et al., 2023).

Claim-Conditioned Probability (CCP) (Fadeeva et al., 2024) evaluates uncertainty at the token level by determining the semantic consistency among the top probable token alternatives. This is done by clustering the token alternatives in tokens that entail and contradict the original meaning by comparing the chosen token with its alternatives using an NLI model. The token-level confidence score is the ratio of probability mass assigned to entailing tokens to the sum of entailing and contradicting tokens. Sequence-level confidence is calculated from product of token confidences.

7.2 Methodology / Experimental Setup

To evaluate the calibration of different uncertainty methods in long-form QA, we focus exclusively on instruction-tuned and reasoning models, as base models exhibit poor task comprehension and are unsuitable for QA tasks. For this experiment, all previously selected MCQA and arithmetic QA datasets were used.

MCQA introduces specific challenges: (1) **Selection bias.** Models may favor certain answer choices due to token frequency or formatting learned during training, regardless of semantic content (Myrzakhan et al., 2024). (2) **Positional bias.** Models can exhibit systematic preference for certain label positions (e.g., always selecting "A") (Zheng et al., 2024).
We address these biases using the APriCoT prompting strategy (Moore et al., 2025), which combines Chain-of-Thought reasoning with counterfactual prompting. Each answer choice is evaluated independently, and the model classifies it as correct or incorrect, producing a verifiable judgment. This isolates answer evaluation from ordering and formatting effects, reducing both selection and positional bias and improving calibration. Rephrasing MC questions as open-ended queries could avoid format-related biases, but many items depend on predefined choices (e.g., fill-in-the-blank or elimination logic). APriCoT approximates open-ended QA by eliciting reasoning over individual answers while preserving a verifiable format.

For arithmetic QA datasets, Chain-of-Thought (CoT) prompting is used to facilitate multi-step reasoning. To balance coverage and computational cost, each dataset is subsampled to 250 items, and 10 generations are sampled per prompt, resulting in a total of $57,500$ QA prompts per model.

7.3 Key Findings

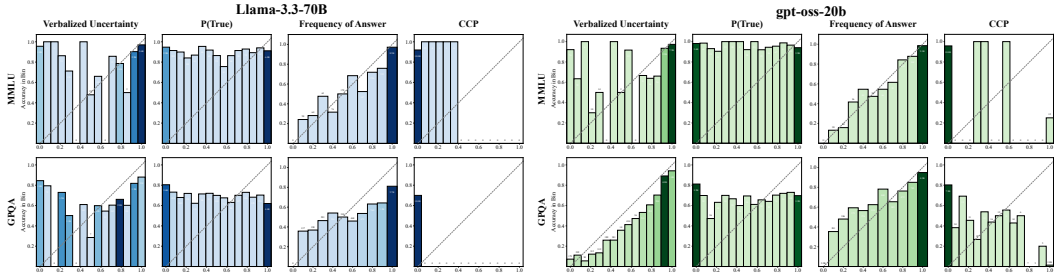
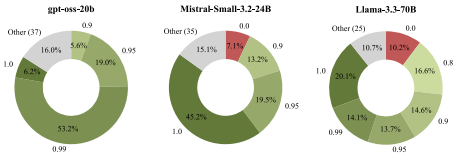


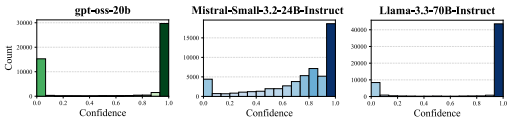
Figure 2: **Selected Calibration Plots For The Four Selected Methods.** Results for *Llama-3.3-70B* (left) and *gpt-oss-20b* (right) are shown, each for MMLU and GPQA for the four computed sequence level uncertainty methods. The full plots for all models and datasets per UQ method can be found in Section A.6.3.

Verbalized Uncertainty Our results show that, although all models reliably produced valid numeric outputs, they overwhelmingly defaulted to a small set of values (see Figure 3a), with higher confidences dominating the responses. This leads to a biased distribution of confidence scores, potentially driven by training data and instruction tuning. Calibration plots reveal no meaningful correlation between these verbalized scores and answer accuracy in nearly all models, indicating that Verbalized Uncertainty is not a reliable proxy for true model confidence. The *gpt-oss* models represent a notable exception: they provide well-calibrated scores, achieving low ECE values even on challenging datasets such as GPQA and SciBench (see Figure A.8). The reason for the reliability of these verbalized scores remains unclear, as no architectural or scale-related factors clearly distinguish these models from others.

P(True) We find that P(True) suffers from pronounced response bias. In several models P(True) overwhelmingly assigns near 1.0 confidence with little use of intermediate confidence scores, resulting in a polarized distribution of certainty scores (see Figure 3b). 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. This reflects again the polarisation in reasoning models observed in Figure 1.



(a) **Verbalized Uncertainty** Confidence scores seen in less than 5% of the total responses have been grouped into "Other", with the number of distinct confidence scores shown in brackets.



(b) **P(True)** Confidence scores are derived by probability mass assigned to labels `(A)` and `(B)`, representing the model's confidence that the answer is true or false respectively.

Figure 3: **Distribution of Confidence Scores Across Selected Representative Models.** Confidence scores have been aggregated across all datasets ($57,500$ prompts in total). See Section A.6.4 and Section A.6.5 for extensive plots for all models.

Frequency of Answer Our evaluation (Figure 2) shows that higher answer frequencies strongly correlate with correctness across both multiple-choice and arithmetic tasks. More challenging datasets (e.g., GPQA, SciBench) exhibiting greater answer diversity, which reflects 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, due to multiple generations per prompt, and its dependence on semantic clustering of outputs, it remains challenging or unapplicable to open-ended QA.

Claim Conditioned Probability In practice, we find that CCP suffers from vanishing sequence-level scores as generation length grows, i.e. multiplying a large number of token confidences drives overall confidence near 0. As a result, calibration plots (Figure 2) show no meaningful alignment with correctness. Furthermore, high impact of single NLI misclassifications and the inclusion of stop words in the aggregation further destabilize scores. These issues make CCP unreliable for sequence-level uncertainty estimation. But, its token-level insights could still inform targeted analyses once aggregation and domain-specific entailment are improved.

8 Limitations

This study focuses on scientific QA in structured formats (multiple-choice and arithmetic), which constrains the generalizability of the findings to other domains and task types such as open-ended generation or summarization. All datasets are in English, and the potential impact of cross-linguistic variation, input formatting, and prompt phrasing on calibration was not examined. Despite mitigation efforts using prompting strategies like APriCoT, the multiple-choice setting itself may introduce systematic biases, such as steering effects. Furthermore, the benchmark employs controlled

inference conditions (e.g. fixed temperature, fixed decoding parameters) that may not capture the variability of real-world deployments. Finally, the evaluation of sequence-level uncertainty was limited to a subset of UQ methods with normalized outputs, selected to enable calibration analysis via calibration plots and summary metrics. This leaves unnormalized or claim-level metrics for future investigation into their potential value for detecting incorrect answers through high uncertainty estimates.

9 Conclusion

We presented a systematic evaluation of four UQ methods for LLMs across seven natural science datasets and eleven open-weight model families, including base, instruction-tuned, and reasoning variants. We acquired our results by developing an open-source framework for benchmarking LLMs with a focus on efficiency and reproducibility.

Our results reveal a pronounced polarization effect in token-level confidence distributions induced by instruction-tuning, which diminishes their utility as reliable uncertainty signals. This effect is even stronger in reasoning models, where the generation process commits to a single hypothesis as part of the reasoning chain.

At the sequence level, verbalized UQ methods exhibit consistently poor performance across most models. In particular, **P(True)** is adversely affected by the same confidence polarization previously observed at the token level. **Verbalized Uncertainty** scores are biased toward a narrow range of high-confidence scores and generally fail to correlate with answer correctness. A notable exception is observed in the *gpt-oss* family, which yielded well calibrated scores, underscoring the need for continued benchmarking to disentangle the impact of architectural design choices, fine-tuning data, and training paradigms on UQ behavior. The **Frequency of Answer** method based in semantic consistency showed strong reliability, albeit at high computational cost and with the non-trivial challenge of robust semantic equivalence detection. **Claim-Conditioned Probability (CCP)** suffers from vanishing sequence-level scores as generation length increases, compounded by NLI misclassifications and instability from stop-word aggregation. These issues render it unreliable for sequence-level uncertainty estimation. Yet, its efficiency and informative token-level signals suggest potential for refinement through improved aggregation and entailment modeling.

We advocate that our results have substantial relevance outside of scientific QA tasks given the variety and number of datasets. As such, our findings highlight a critical need for the development of more robust, efficient, and theoretically grounded UQ methods for LLMs. Advancing this field will require not only algorithmic innovation but also sustained empirical benchmarking to isolate the contributions of model architecture, fine-tuning strategy, and training data to uncertainty behavior in real world scenarios.

10 Future Work

Future research on uncertainty estimation in LLMs should pursue several directions. In the short term, benchmarks should be extended with new UQ methods as they are proposed, e.g. building on the benchmarking framework provided alongside this paper. Systematic studies of inference parameters (e.g., temperature, sampling strategies) and prompting strategies are also promising. As new model LLM families continue to emerge, ongoing UQ benchmarking will be essential not only to track improvements in uncertainty estimation but also to trace these gains back to architectural choices or training paradigms that most strongly drive better calibration behavior.

Further, LLM progress will require new dataset designs. Benchmarks that explicitly induce uncertainty (e.g., contradictory or unsolvable questions) or reformulate multiple-choice tasks into open-ended formats could better isolate task-specific sources of uncertainty. Longer-term directions include claim-level uncertainty estimation, which assesses reliability of individual statements or reasoning steps, and linguistic calibration, which studies alignment between epistemic markers and model confidence. Finally, ranking-oriented metrics may prove valuable in practice, guiding model outputs toward more trustworthy generations even without perfect calibration.

Reproducibility Statement

Reproducibility in UQ for LLMs poses unique challenges due to the multistage evaluation pipelines and inherent randomness at each step, which can compound and introduce measurement noise. This stochasticity affects comparisons across UQ methods and model outputs, particularly when methods are evaluated on different generations, potentially contaminating results. To address these issues, we developed a modular benchmarking framework designed to ensure reproducibility and resource efficiency in large-scale LLM experiments. The framework supports extensible and replaceable computation nodes, caches probabilistic outputs such as model generations to ensure consistent re-evaluation, preserves intermediate outputs for qualitative inspection and allows incremental updates so that only affected steps need recomputation. It also enables sharing of intermediate results to reduce compute costs and allow other researchers to rerun the benchmark with their own methods. This is particularly valuable for ongoing research in UQ on models gated by proprietary access or costly hardware, supporting broader collaboration within the scientific community. These features not only enhance the reliability of the present study but also establish a foundation for future research in the field.

To facilitate reproducibility, the repository includes the full framework `async-graph-bench` in its current form, final leaf-node outputs containing confidence scores required to reproduce the plots, a container definition file to build the execution environment, exact pip and driver versions, and instructions to install the framework and run the benchmarks presented in this work. While the framework is in its early stages, it is fully functional for reproducing the experiments reported here, and we plan to continue improving documentation and testing before its broader release in the near future.

LLM Disclosure

LLMs were used in a limited manner to assist with writing, coding, and literature search. Specifically, LLMs provided minor support for language polishing, code-related assistance and identifying relevant references. All outputs were carefully reviewed, tested, and verified by the authors. LLMs did not contribute to the conceptualization of the research, the design of experiments, or the interpretation of results.

References

Neil Band, Xuechen Li, Tengyu Ma, and Tatsunori Hashimoto. Linguistic calibration of long-form generations, 2024. URL https://arxiv.org/abs/2404.00474.

Sourav Banerjee, Ayushi Agarwal, and Saloni Singla. Llms will always hallucinate, and we need to live with this, 2024. URL https://arxiv.org/abs/2409.05746.

Bowen Cao, Deng Cai, Zhisong Zhang, Yuexian Zou, and Wai Lam. On the worst prompt performance of large language models, 2024. URL https://arxiv.org/abs/2406.10248.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms' internal states retain the power of hallucination detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=Zj12nzlQbz.

Youngseog Chung, Ian Char, Han Guo, Jeff Schneider, and Willie Neiswanger. Uncertainty toolbox: an open-source library for assessing, visualizing, and improving uncertainty quantification. *arXiv preprint arXiv:2109.10254*, 2021.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

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. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

André F. Cruz, Moritz Hardt, and Celestine Mendler-Dünner. Evaluating language models as risk scores, 2024. URL https://arxiv.org/abs/2407.14614.

Maxime Darrin, Pablo Piantanida, and Pierre Colombo. RainProof: An umbrella to shield text generator from out-of-distribution data. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5831–5857, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.357. URL https://aclanthology.org/2023.emnlp-main.357/.

Siddartha Devic, Tejas Srinivasan, Jesse Thomason, Willie Neiswanger, and Vatsal Sharan. From calibration to collaboration: Llm uncertainty quantification should be more human-centered, 2025. URL https://arxiv.org/abs/2506.07461.

Shehzaad Dhuliawala, Vilém Zouhar, Mennatallah El-Assady, and Mrinmaya Sachan. A diachronic perspective on user trust in ai under uncertainty, 2023. URL https://arxiv.org/abs/2310.13544.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models, 2024. URL https://arxiv.org/abs/2307.01379.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. Lm-polygraph: Uncertainty estimation for language models, 2023a. URL https://arxiv.org/abs/2311.07383.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. LM-polygraph: Uncertainty estimation for language models. In Yansong Feng and Els Lefever (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 446–461, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.41. URL https://aclanthology.org/2023.emnlp-demo.41.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. Fact-checking the output of large language models via token-level uncertainty quantification, 2024. URL https://arxiv.org/abs/2403.04696.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07421-0. URL https://www.nature.com/articles/s41586-024-07421-0. Publisher: Nature Publishing Group.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Rico Sennrich, and Lucia Specia. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555, 2020. doi: 10.1162/tacl_a_00330. URL https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00330/96475/Unsupervised-Quality-Estimation-for-Neural-Machine.

Yarin Gal et al. Uncertainty in deep learning, 2016.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6577–6595. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.366. URL https://aclanthology.org/2024.naacl-long.366/.

Xuemei Gu and Mario Krenn. Forecasting high-impact research topics via machine learning on evolving knowledge graphs. *Machine Learning: Science and Technology*, 6(2):025041, may 2025. doi: 10.1088/2632-2153/add6ef. URL https://dx.doi.org/10.1088/2632-2153/add6ef.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. Does prompt formatting have any impact on llm performance?, 2024. URL https://arxiv.org/abs/2411.10541.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks, 2018. URL https://arxiv.org/abs/1610.02136.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. Look before you leap: An exploratory study of uncertainty measurement for large language models. *IEEE Transactions on Software Engineering*, 51(2):413–429, 2025. ISSN 0098-5589, 1939-3520, 2326-3881. doi: 10.1109/TSE.2024.3519464. URL http://arxiv.org/abs/2307.10236.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *CoRR*, abs/2202.03629, 2022. URL https://arxiv.org/abs/2202.03629.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning, 2025. URL https://arxiv.org/abs/2503.09516.

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. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Dominique Kelly, Yimin Chen, Sarah E Cornwell, Nicole S Delellis, Alex Mayhew, Sodiq Onaolapo, and Victoria L Rubin. Bing chat: The future of search engines? *Proceedings of the Association for Information Science and Technology*, 60(1):1007–1009, 2023.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision?, 2017. URL https://arxiv.org/abs/1703.04977.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation, 2023a. URL https://arxiv.org/abs/2302.09664.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023b. URL https://openreview.net/forum?id=VD-AYtP0dve.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression, 2018. URL https://arxiv.org/abs/1807.00263.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/abdeb6f575ac5c6676b747bca8d09cc2-Abstract.html.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*, 2023. URL https://arxiv.org/abs/2305.19187.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. Uncertainty quantification and confidence calibration in large language models: A survey, 2025. URL https://arxiv.org/abs/2503.15850.

Xinyi Lu and Xu Wang. Generative students: Using llm-simulated student profiles to support question item evaluation. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, L@S '24, pp. 16–27, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706332. doi: 10.1145/3657604.3662031. URL https://doi.org/10.1145/3657604.3662031.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*, 2020. URL https://arxiv.org/abs/2002.07650.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization, 2020. URL https://arxiv.org/abs/2005.00661.

Kyle Moore, Jesse Roberts, Thao Pham, and Douglas Fisher. Chain of thought still thinks fast: Apricot helps with thinking slow, 2025. URL https://arxiv.org/abs/2408.08651.

Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. Open-llm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena, 2024. URL https://arxiv.org/abs/2406.07545.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *arXiv preprint arXiv:2405.20003*, 2024. URL https://arxiv.org/pdf/2405.20003.

OpenAI. Gpt-4 technical report, 2023. URL https://arxiv.org/abs/2303.08774. arXiv:2303.08774 [cs.CL].

Theodore Papamarkou, Maria Skoularidou, Konstantina Palla, Laurence Aitchison, Julyan Arbel, David Dunson, Maurizio Filippone, Vincent Fortuin, Philipp Hennig, José Miguel Hernández-Lobato, et al. Position: Bayesian deep learning is needed in the age of large-scale ai. *arXiv preprint arXiv:2402.00809*, 2024.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are nlp models really able to solve simple math word problems?, 2021. URL https://arxiv.org/abs/2103.07191.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training, 2018. URL https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark, 2023. URL https://arxiv.org/abs/2311.12022.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J. Liu. Out-of-distribution detection and selective generation for conditional language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=kJUS5nD0vPB.

Jonatan Reyes, Anil Ufuk Batmaz, and Marta Kersten-Oertel. Trusting AI: does uncertainty visualization affect decision-making? *Frontiers in Computer Science*, 7, February 2025. ISSN 2624-9898. doi: 10.3389/fcomp.2025.1464348. URL https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2025.1464348/full. Publisher: Frontiers.

Mike S Schäfer. The notorious gpt: science communication in the age of artificial intelligence. *JCOM: Journal of Science Communication*, 22(2):Y02, 2023.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting, 2024. URL https://arxiv.org/abs/2310.11324.

Peiqi Sui, Eamon Duede, Sophie Wu, and Richard Jean So. Confabulation: The surprising value of large language model hallucinations, 2024. URL https://arxiv.org/abs/2406.04175.

Junya Takayama and Yuki Arase. Relevant and informative response generation using pointwise mutual information. In Yun-Nung Chen, Tania Bedrax-Weiss, Dilek Hakkani-Tur, Anuj Kumar, Mike Lewis, Thang-Minh Luong, Pei-Hao Su, and Tsung-Hsien Wen (eds.), *Proceedings of the First Workshop on NLP for Conversational AI*, pp. 133–138, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4115. URL https://aclanthology.org/W19-4115/.

Linwei Tao, Yi-Fan Yeh, Minjing Dong, Tao Huang, Philip Torr, and Chang Xu. Revisiting uncertainty estimation and calibration of large language models, 2025. URL https://arxiv.org/abs/2505.23854.

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. URL https://arxiv.org/abs/2305.14975.

Herbert Ullrich, Tomáš Mlynář, and Jan Drchal. Claim extraction for fact-checking: Data, models, and automated metrics, 2025. URL https://arxiv.org/abs/2502.04955.

Liam van der Poel, Ryan Cotterell, and Clara Meister. Mutual information alleviates hallucinations in abstractive summarization. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5956–5965, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.399. URL https://aclanthology.org/2022.emnlp-main.399/.

Artem Vazhentsev, Gleb Kuzmin, Akim Tsvigun, Alexander Panchenko, Maxim Panov, Mikhail Burtsev, and Artem Shelmanov. Hybrid uncertainty quantification for selective text classification in ambiguous tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11659–11681, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.652. URL https://aclanthology.org/2023.acl-long.652/.

Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. Llms may perform mcqa by selecting the least incorrect option, 2024a. URL https://arxiv.org/abs/2402.01349.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models, 2024b. URL https://arxiv.org/abs/2307.10635.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023. URL https://arxiv.org/abs/2203.11171.

Ziming Wang, Zeyu Shi, Haoyi Zhou, Shiqi Gao, Qingyun Sun, and Jianxin Li. Towards objective fine-tuning: How llms' prior knowledge causes potential poor calibration?, 2025. URL https://arxiv.org/abs/2505.20903.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions, 2017. URL https://arxiv.org/abs/1707.06209.

Haoyi Xiong, Jiang Bian, Yuchen Li, Xuhong Li, Mengnan Du, Shuaiqiang Wang, Dawei Yin, and Sumi Helal. When search engine services meet large language models: visions and challenges. *IEEE Transactions on Services Computing*, 2024a.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms, 2024b. URL https://arxiv.org/abs/2306.13063.

KiYoon Yoo, Jangho Kim, Jiho Jang, and Nojun Kwak. Detection of adversarial examples in text classification: Benchmark and baseline via robust density estimation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 3656–3672, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.289. URL https://aclanthology.org/2022.findings-acl.289/.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. Luq: Long-text uncertainty quantification for llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5244–5262, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.299. URL https://aclanthology.org/2024.emnlp-main.299/.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. Enhancing uncertainty-based hallucination detection with stronger focus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 915–932, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.58. URL https://aclanthology.org/2023.emnlp-main.58/.

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

Ziyin Zhang, Zhaokun Jiang, Lizhen Xu, Hongkun Hao, and Rui Wang. Multiple-choice questions are efficient and robust llm evaluators, 2024b. URL https://arxiv.org/abs/2405.11966.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors, 2024. URL https://arxiv.org/abs/2309.03882.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models, 02 2023. URL https://arxiv.org/abs/2302.13439.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, and Maarten Sap. Relying on the unreliable: The impact of language models' reluctance to express uncertainty, 2024. URL https://arxiv.org/abs/2401.06730.

A Appendix

A.1 Repository

The repository accompanying this paper is available online at https://anonymous.4open.science/r/llm-uncertainty-bench-9B2B/.

A.2 Model Selection

Table A.1: Overview of LLMs used in the experiments of this paper by provider, type, size, and release date. Model families of corresponding base, instruct and reasoning variants have been grouped together.

Provider	Model Name	Type	Size	Release Date
OpenAI	gpt-oss-20b	Reasoning	20B	Aug 2025
	gpt-oss-120b	Reasoning	120B	Aug 2025
Mistral AI	Mistral-Nemo-Base-2407	Base	8B	Jul 2024
	Mistral-Nemo-Instruct-2407	Instruct	8B	Jul 2024
	Ministral-8B-Instruct-2410	Instruct	8B	Oct 2024
	Mistral-Small-3.1-24B-Base-2503	Base	24B	Mar 2025
	Mistral-Small-3.2-24B-Instruct-2506	Instruct	24B	Jun 2025
	Magistral-Small-2507	Reasoning	24B	Jul 2025
Meta LLaMA	Llama-3.1-70B	Base	70B	Jul 2024
	Llama-3.3-70B-Instruct	Instruct	70B	Dec 2024
	Llama-4-Scout-17B-16E	Base	109B	Apr 2025
	Llama-4-Scout-17B-16E-Instruct	Instruct	109B	Apr 2025
Qwen	Qwen3-30B-A3B-Base	Base	30B	Jul 2025
	Qwen3-30B-A3B-Instruct-2507	Instruct	30B	Jul 2025
	Qwen3-30B-A3B-Thinking-2507	Reasoning	30B	Jul 2025
DeepSeek AI	DeepSeek-R1-Distill-Llama-70B	Reasoning	70B	Jan 2025
	DeepSeek-R1-Distill-Qwen-32B	Reasoning	32B	Jun 2024
Google	gemma-3-27b-pt	Base	27B	Mar 2025
	gemma-3-27b-it	Instruct	27B	Mar 2025

A.2.1 Model Configuration and Exceptions

All models were evaluated in their default configurations. No system prompts were employed, with the sole exception of *Magistral-Small-2507*, which requires a system prompt to enable reasoning prior to generating a final answer[2]. Without the system prompt, the model will behave like an instruction-tuned model. For the label-probability experiments, this model was tested in two variants: with reasoning enabled (*Magistral-Small-2507-Reasoning-Enabled*) and without reasoning (*Magistral-Small-2507*).

The base (pre-trained) model *gemma-3-27b-pt*, corresponding to the instruction-tuned *gemma-3-27b-it*, was excluded from the label-probability calibration experiments. In preliminary evaluations, the base model exhibited insufficient task comprehension, resulting in negligible probability mass assigned to label tokens within the top-20 most probable tokens. Consequently, label probabilities could not be retrieved through `vllm`, preventing the computation of confidence scores.

All models were used in their base configuration. With the exception of *Magistral-Small-2507*, no system prompts were used. *Magistral-Small-2507* uses the system prompt to elicit reasoning steps before providing a final answer. It was included in the label probability experiment both with (*Magistral-Small-2507-Reasoning-Enabled*) and without (*Magistral-Small-2507*) the system

[2] see https://huggingface.co/mistralai/Magistral-Small-2506

prompt. *gemma-3-27b-pt*, the pre-trained or base variant of *gemma-3-27b-it* was excluded from the experiment researching the label probability calibration due to lack in task comprehension. This resulted in no probability mass being assigned to label tokens within the 20 most probable tokens, rendering probabilities unaccessible via `vllm`. Therefore no certainties could be retrieved.

For the exact configuration parameters supplied to `vllm`, please refer to the models.py files located in the experiment subdirectories of the repository accompanying this paper.

A.3 Dataset Selection

Table A.2: **Scientific QA datasets surveyed.** Selection criteria emphasized natural-science domains (especially physics) and inclusion of different reasoning requirements for answering, while providing a verifiable ground-truth.

Dataset	Size	Task Format	Domain
MMLU	15,908	Multiple Choice	General (57 topics, including Physics)
ARC	7,787	Multiple Choice	Science (Physics, Chemistry, Biology, Earth Science)
SciQ	13,679	Multiple Choice	Science (Physics, Chemistry, Biology)
GPQA	448	Multiple Choice	Physics (Graduate-level)
GSM8K	8,792	Arithmetic	Mathematics
GSM-MC	8,787	Multiple Choice	Mathematics
SVAMP	1,000	Arithmetic	Mathematics
SciBench	2,229	Arithmetic	Science (Physics, Chemistry, Biology, Medicine, Earth Science)

For the exact dataset configuration parameters, please refer to the `data_source` subdirectories located in the experiment directories of the repository accompanying this paper.

A.4 Uncertainty Quantification Method Exclusion

For evaluating the calibration of long form generations, several sequence level uncertainty methods exist. Our selection of methods follows the taxonomy from Fadeeva et al. (2023a) into five different categories: information-based, ensemble-based, density-based, reflexive and meaning diversity approaches. We have conducted an extensive literature review on available UQ methods. The results are contained in Table A.3.

Table A.3: **Reasons for Exclusion of Uncertainty Metrics in Benchmark.** List of UQ Methods was adapted from the comprehensive evaluation and classification of uncertainty methods by Fadeeva et al. (2023a).

Method	Category	Exclusion Reasons
Perplexity (Fomicheva et al., 2020)	Information-based	Produces unnormalized scores
Mean/max token entropy (Fomicheva et al., 2020)	Information-based	Produces unnormalized scores
Monte Carlo sequence entropy (Kuhn et al., 2023b)	Information-based	Produces unnormalized scores
Pointwise mutual information (PMI) (Takayama & Arase, 2019)	Information-based	Produces unnormalized scores
Conditional PMI (van der Poel et al., 2022)	Information-based	Produces unnormalized scores
Rényi divergence (Darrin et al., 2023)	Information-based	Produces unnormalized scores

Method	Category	Exclusion Reasons
Fisher-Rao distance (Darrin et al., 2023)	Information-based	Produces unnormalized scores
Focus (Zhang et al., 2023a)	Information-based	Produces unnormalized scores
Semantic entropy (Kuhn et al., 2023b)	Meaning diversity	Designed to work at the individual claim level rather than on entire sequences Very high computational cost
TokenSAR (Duan et al., 2024)	Meaning diversity	Alters sentences in a way that violates autoregressive assumptions Relies on NLI models whose performance in scientific contexts is inadequate
SentenceSAR (Duan et al., 2024)	Meaning diversity	Alters sentences in a way that violates autoregressive assumptions Relies on NLI models whose performance in scientific contexts is inadequate
SAR (Duan et al., 2024)	Meaning diversity	Alters sentences in a way that violates autoregressive assumptions Relies on NLI models whose performance in scientific contexts is inadequate
EigenScore (Chen et al., 2024)	Meaning diversity	Produces unnormalized scores
Sentence-level ensemble-based measures (Malinin & Gales, 2020)	Ensembling	Requires running multiple independent models, leading to high computational cost Introduces extra variability that complicates comparison to single-model methods
Token-level ensemble-based measures (Malinin & Gales, 2020)	Ensembling	Requires running multiple independent models, leading to high computational cost Introduces extra variability that complicates comparison to single-model methods
Mahalanobis distance (MD) (Lee et al., 2018)	Density-based	Density-based approach that requires proprietary training data
Robust density estimation (RDE) (Yoo et al., 2022)	Density-based	Density-based approach that requires proprietary training data
Relative Mahalanobis distance (RMD) (Ren et al., 2023)	Density-based	Density-based approach that requires proprietary training data
Hybrid Uncertainty Quantification (HUQ) (Vazhentsev et al., 2023)	Density-based	Density-based approach that requires proprietary training data
Number of semantic sets (NumSets) (Lin et al., 2023)	Meaning Diversity	Produces count-based outputs that are not normalized to [0,1]
Sum of eigenvalues of the graph Laplacian (EigV) (Lin et al., 2023)	Meaning Diversity	Produces unnormalized scores
Degree matrix (Deg) (Lin et al., 2023)	Meaning Diversity	Produces unnormalized scores
Eccentricity (Ecc) (Lin et al., 2023)	Meaning Diversity	Produces unnormalized scores
Lexical similarity (LexSim) (Fomicheva et al., 2020)	Meaning Diversity	Relies on surface-level token overlap instead of semantic meaning
Kernel Language Entropy (Nikitin et al., 2024)	Meaning Diversity	Produces unnormalized scores
LUQ ((Zhang et al., 2024a))	Meaning diversity	Produces unnormalized scores

Ensemble- and density-based methods are excluded from our analysis due to high computational cost, dependence on multiple models or inaccessible training data, and limited uncertainty coverage. Claim-level methods from the meaning diversity category (Farquhar et al., 2024) are omitted due to the complexity and unreliability of claim extraction as a required intermediary processing step.

Only methods producing normalized sequence-level uncertainty scores are included in our analysis to enable reliability UQ validation by calibration. Subsequently, we focus our work on the following UQ methods: *Verbalized Uncertainty* (Tian et al., 2023), *P(True)* (Kadavath et al., 2022), *Frequency of Answer* (Wang et al., 2023), *Claim-Conditioned Probability (CCP)* (Fadeeva et al., 2024). More details of these methods are discussed in Section 7.1.

A.5 Label Probability Calibration

A.5.1 Prompt Designs

Listing 1: **Prompt Design 1 for Experiment 1.** The prompt design uses 3-shot prompting. <QUESTION> and <ANSWER CHOICE X> are replaced with the individual questions and answer choices from the benchmarked dataset. Prompt Design 1 features markdown construct.

```
You are a highly capable language model trained for multiple-choice question answering.
Below are three examples of multiple-choice questions with labeled answer choices. Each example includes the correct answer.
After the examples, you will be given a new question with four labeled answer choices (A, B, C, D).

Your task is to select the answer choice you believe is correct by responding with only the corresponding label: A, B, C, or D.
Do not include any explanation or additional text.

### Example 1:
**Question:** What is the capital of France?
A) Berlin
B) Madrid
C) Paris
D) Rome

**Correct Answer:** C

<Two More Examples Omitted for Readability>

---

**Question:** <QUESTION>
A) <ANSWER CHOICE A>
B) <ANSWER CHOICE B>
C) <ANSWER CHOICE C>
D) <ANSWER CHOICE D>

**Correct Answer:**
```

Listing 2: **Prompt Design 2 for Experiment 1.** The prompt design uses 3-shot prompting. <QUESTION> and <ANSWER CHOICE X> are replaced with the individual questions and answer choices from the benchmarked dataset. Prompt Design 2 features no introductary text, role or task description. The format is designed to represent natural language without special formatting.

```
Question: What is the capital of France?
A) Berlin
B) Madrid
C) Paris
D) Rome

The correct answer is C

<Two More Examples Omitted for Readability>

Question: <QUESTION>
A) <ANSWER CHOICE A>
B) <ANSWER CHOICE B>
C) <ANSWER CHOICE C>
D) <ANSWER CHOICE D>

The correct answer is
```

Listing 3: **Prompt Design 3 for Experiment 1.** The prompt design uses 3-shot prompting. <QUESTION> and <ANSWER CHOICE X> are replaced with the individual questions and answer choices from the benchmarked dataset. The format of Prompt 3 is designed to represent natural language without special formatting. The format of the answer specifically requests the label, not the answer in general.

```
You are a highly capable language model trained for multiple-choice
    question answering. In the following examples, you will see questions
     with answer choices. The answer choices are preceded by the phrase "
    Answer Choices:". Each answer choice is annotated with one of the
    labels A, B, C or D. The correct answer to the question is given by
    the sentence "The label of the correct answer choice is" followed by
    the corresponding label. Your task is to answer the new question in
    the same format, outputting only the label of the correct answer to
    the question you are provided. Do not output anything other than one
    of the labels A, B, C or D.

Question: What is the capital of France?
Answer Choices:
A) Berlin
B) Madrid
C) Paris
D) Rome

The label of the correct answer choice is C

<Two More Examples Omitted for Readability>

Question: <QUESTION>
Answer Choices:
A) <ANSWER CHOICE A>
B) <ANSWER CHOICE B>
C) <ANSWER CHOICE C>
D) <ANSWER CHOICE D>

The label of the correct answer choice is
```

Listing 4: **Prompt Design 4 for Experiment 1.** The prompt design uses 3-shot prompting. <QUESTION> and <ANSWER CHOICE X> are replaced with the individual questions and answer choices from the benchmarked dataset. Prompt Design 2 features special tags to mark the answer given as a label.

```
You are a highly capable multiple-choice question answering model. Below
    are three examples that show the format you must follow. Each
    question has four answer choices labeled A, B, C, and D. Your task is
     to answer a new question by outputting the correct answer in the
    following format: <ANSWER>X<ANSWER>, where X is the label
    corresponding to the correct answer, A, B, C or D. Do not add any
    extra text or explanation.

Example 1:
Question: What is the capital of France?
Answer Choices:
A) Berlin
B) Madrid
C) Paris
D) Rome

<ANSWER>C<ANSWER>

<Two More Examples Omitted for Readability>

Now, please answer the following question in the same format.

Question: <QUESTION>
```

```
Answer Choices:
A) <ANSWER CHOICE A>
B) <ANSWER CHOICE B>
C) <ANSWER CHOICE C>
D) <ANSWER CHOICE D>

<ANSWER>
```

A.5.2 Task Comprehension per Prompt Design

Table A.4: **Task Comprehension Measured by Probability Mass Assigned to Answer Labels.** Mean over the sum of label probabilities per question for base, instruction-tuned, and reasoning models. On average, task comprehension is highest under Prompt 1.

Model Category	Prompt 1	Prompt 2	Prompt 3	Prompt 4
Base Models	0.2368	0.5928	0.4135	0.1632
Instruction-Tuned Models	0.9912	0.3597	0.9561	0.2646
Reasoning Models	0.7002	0.0895	0.4475	0.0022
Average	**0.6427**	**0.3474**	**0.6057**	**0.1434**

A.5.3 Invalid Answer Counts

Table A.5: **Number of Invalid Answers Given by Models for Different Prompt Designs Across All Datasets ($n = 25{,}316$).** Invalid answers assign no probability mass to any answer-choice labels.

Model	Prompt 1	Prompt 2	Prompt 3	Prompt 4
gpt-oss-20b	307	3356	498	25304
gpt-oss-120b	38	3323	27	24873
Ministral-8B-Instruct-2410	0	0	0	0
Mistral-Nemo-Base-2407	0	0	12574	0
Mistral-Nemo-Instruct-2407	0	0	0	0
Mistral-Small-3.1-24B-Base-2503	0	0	0	0
Mistral-Small-3.2-24B-Instruct-2506	0	0	0	0
Magistral-Small-2507	0	0	0	0
Magistral-Small-2507-Reasoning-Enabled	8604	18716	4749	21232
Llama-3.1-70B	0	13	0	6636
Llama-3.3-70B-Instruct	0	2527	19	125
Llama-4-Scout-17B-16E	2	3	0	5048
Llama-4-Scout-17B-16E-Instruct	0	15	0	0
Qwen3-30B-A3B-Base	2	1	0	55
Qwen3-30B-A3B-Instruct-2507	0	6270	0	722
Qwen3-30B-A3B-Thinking-2507	6242	11746	8656	21840
DeepSeek-R1-Distill-Llama-70B	425	7538	725	7477
DeepSeek-R1-Distill-Qwen-32B	914	7465	1170	3337
gemma-3-27b-it	0	431	0	2

A.5.4 Comprehensive Plots per Prompt with Tables

Detailed calibration plots of label probabilities for all prompt designs and configurations are available in the project repository at `label_prob_calibration/resources/figures/full_plots`. Each file follows the naming convention `cal_plot_prompt<id>_table<t>_chosenonly<c>_norm<n>.svg`, where the placeholders encode the following settings:

- `prompt`: Identifier of the prompt design used to generate the plot (see Section A.5.1).

- `table`: Indicator of whether a table summarizing key calibration statistics is included.
- `chosenonly`: Flag specifying whether confidence scores are computed for all candidate labels (`0`) or restricted to the chosen (most probable) label (`1`).
- `norm`: Flag denoting whether label probabilities are normalized such that their sum equals one across all candidate labels.

In the following, the plots showing the calibration plots for Prompt 1 (best task comprehension) can be seen.

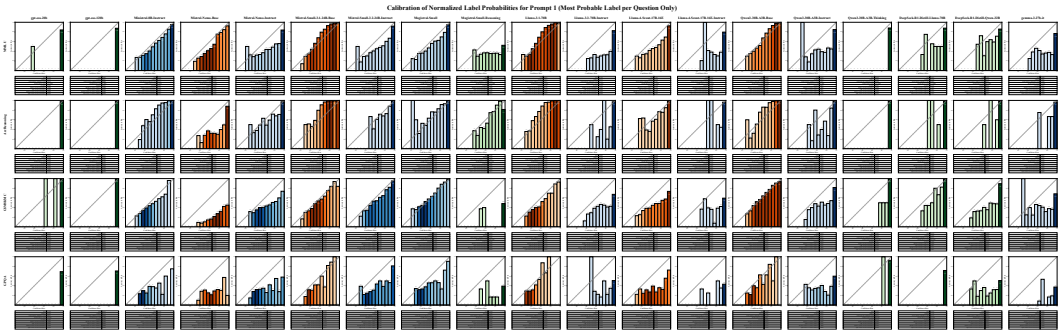


Figure A.1: **Calibration Plots for Prompt 1, using normalized label probabilities and only the most probable label.** The columns represent different models, while the rows represent datasets. Tables below the plots list summary metrics such as ECE, normalized entropy of bucket counts and AUROC. Base models are shown in orange, instruction-tuned models in blue, and reasoning models in green. Darker colors indicate a higher number of items in the bin.

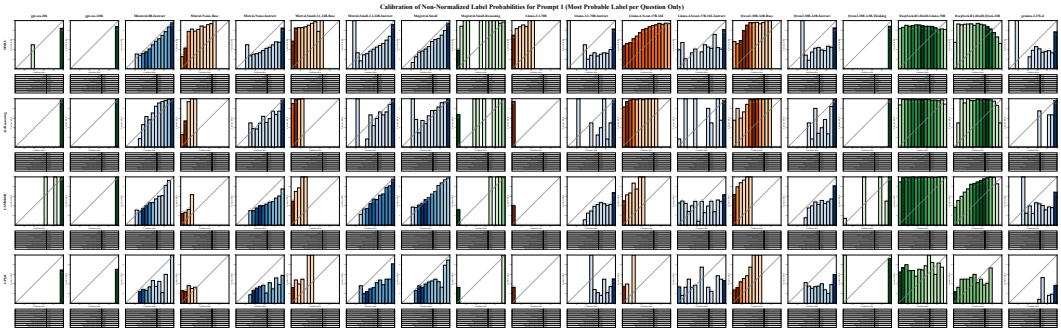


Figure A.2: **Calibration Plots for Prompt 1, using unnormalized label probabilities and only the most probable label.** The columns represent different models, while the rows represent datasets. Tables below the plots list summary metrics such as ECE, normalized entropy of bucket counts and AUROC. Base models are shown in orange, instruction-tuned models in blue, and reasoning models in green. Darker colors indicate a higher number of items in the bin.

A.5.5 EFFECT OF NORMALIZATION

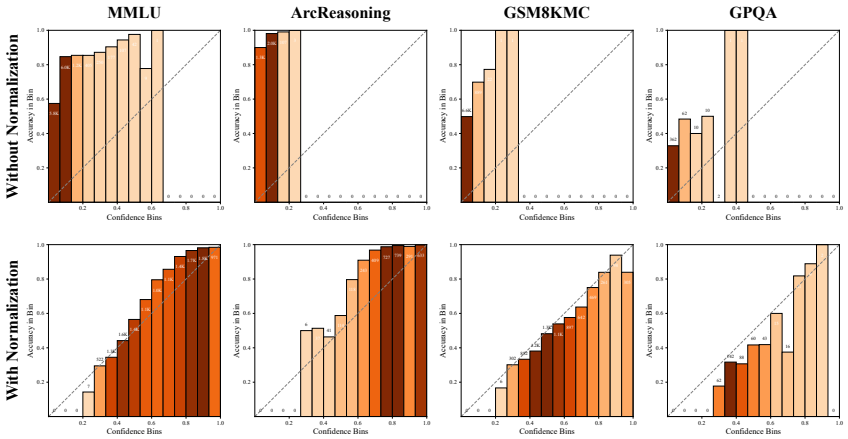


Figure A.3: **Representative Comparison of Calibration Plots for Unnormalized and Normalized Label Probabilities for the Model *Mistral-Small-3.1-24B-Base-2503* and Prompt 1 across all datasets used.** Calibration improves significantly after normalizing label probabilities.

A.5.6 SELECTION BIAS

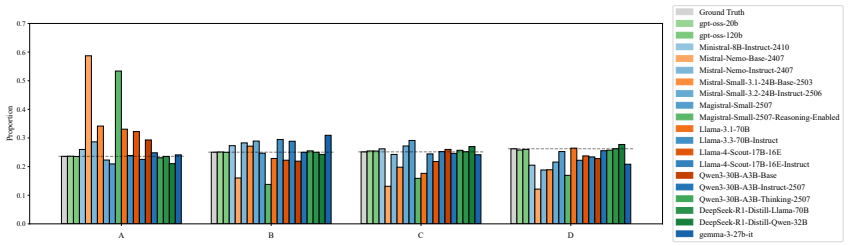


Figure A.4: **Probabilities for the Labels summed across all Datasets for each individual Model using Prompt 1.** The ground truth, represented by the distribution of the labels of the correct answers across all items in the datasets, is visualized by the grey bars and the dashed baselines.

A.5.7 TASK COMPREHENSION PER PROMPT DESIGN

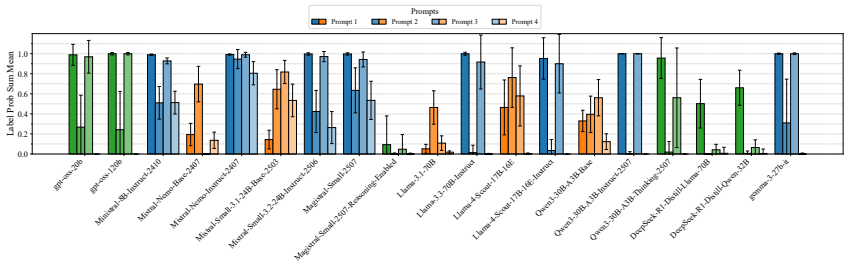


Figure A.5: **Task Comprehension per Prompt Design.** Mean over Sum of Label Probabilities for the Different Prompt Designs and Models. The data has been aggregated across all datasets, spanning 25,316 items per model and prompt design. Base models are shown in orange, instruction-tuned models in blue, and reasoning models in green.

A.6 Sequence-Level Calibration

A.6.1 Prompts Used For Question Answering

The sequence-level experiments employed the APriCoT prompting strategy for MCQA and a standard Chain-of-Thought (CoT) approach for arithmetic question answering (Arithmetic QA), followed by final answer extraction. The exact prompts for both answer generation and final answer extraction are available in the repository accompanying this paper, specifically in `/llm-uncertainty-bench/seq_ue_calibration/nodes/apricot_mc_calc.py` for MCQA and `/llm-uncertainty-bench/seq_ue_calibration/nodes/arithmetic_calc.py` for Arithmetic QA.

A.6.2 Metric Implementation

Verbalized Uncertainty The prompt for Verbalized Uncertainty was adopted directly from the original work (Tian et al., 2023). The exact prompt can be found in `/llm-uncertainty-bench/seq_ue_calibration/run.py`, where it is provided to the corresponding computation node as `verbalized_prompt`.

P(True) The prompt used for the P(True) metric follows the formulation of (Kadavath et al., 2022). For APriCoT prompting in MCQA, a minor adaptation was applied, while preserving the core semantics of the original design. The exact prompts are available in `/llm-uncertainty-bench/seq_ue_calibration/nodes/ptrue.py`.

Frequency of Answer For this metric, 10 samples were generated per prompt. The *Frequency of Answer* of a given response is defined as the proportion of semantically equivalent answers within the set of 10 samples. Invalid answers are assigned a confidence of 0.0. Semantic equivalence was determined according to the task type:

- **Multiple-Choice QA:** Using the APriCoT prompting strategy, each option is independently evaluated, and the model classifies each option as correct or incorrect. Semantic equivalence is established when different generations reach the same classification decision for a given option.
- **Simple Arithmetic QA:** For datasets such as GSM8K, which primarily involve integers and rarely require floating-point precision, the final numeric result was extracted using a dedicated prompt and parsed into a numeric representation. Semantic equivalence is then determined by strict numeric equality of the extracted results. For details on the final answer extraction prompt, please refer to `/llm-uncertainty-bench/seq_ue_calibration/nodes/arithmetic_calc.py` in the repository.
- **SciBench:** This dataset presents additional complexity due to intricate computations and the inclusion of physical units, rendering the simple numeric matching used for other arithmetic datasets insufficient. To address this, a specialized clustering prompt was developed to group sampled answers into semantically equivalent categories, with `Llama-3.3-70B-Instruct` serving as the judging model. Implementation details are provided in `/llm-uncertainty-bench/seq_ue_calibration/leaf_nodes/answered_correctly_scibench.py`.

For evaluation of the calibration of the Frequency of Answer metric, the binning strategy will be slightly modified. Unlike the other methods, this methods yields only discrete confidence values, determined by the number of sampled generations. With 10 generations per question, the resulting confidence scores can only take on values from 0.1 (indicating that all of the other nine sampled generations resulted in a different answer) to 1.0 (all generations produced the same result), in steps of 0.1. For responses that fail to yield a numeric outcome in arithmetic datasets, a confidence score of 0.0 is assigned. To accommodate these discrete confidence levels, 11 bins centered on the possible certainty scores will be used for generating the calibration plots and summary statistics thereof for the Frequency of Answer metric.

Claim-Conditioned Probability (CCP) The Claim-Conditioned Probability (CCP) metric, proposed by Fadeeva et al. (2024), was originally designed for claim-based uncertainty estimation.

While conceptually valuable, applying CCP to long, complex generations proved challenging: extracting meaningful claims was computationally expensive, often unreliable, and complicated by interdependent claims that hindered aggregation. Nevertheless, CCP was included by aggregating token-level confidence scores through multiplicative composition. The implementation builds upon the authors' implementation of the metric in their LM-Polygraph framework (Fadeeva et al., 2023b) and was optimized for improved computational efficiency.

A.6.3 Extensive Calibration Plots

In the following, extensive calibration plots for the evaluation of sequence level uncertainty methods are provided. Again, instruction tuned models are highlighted in blue, while reasoning models are highlighted in green. Darker shading indicates a higher number of items within each confidence bin.

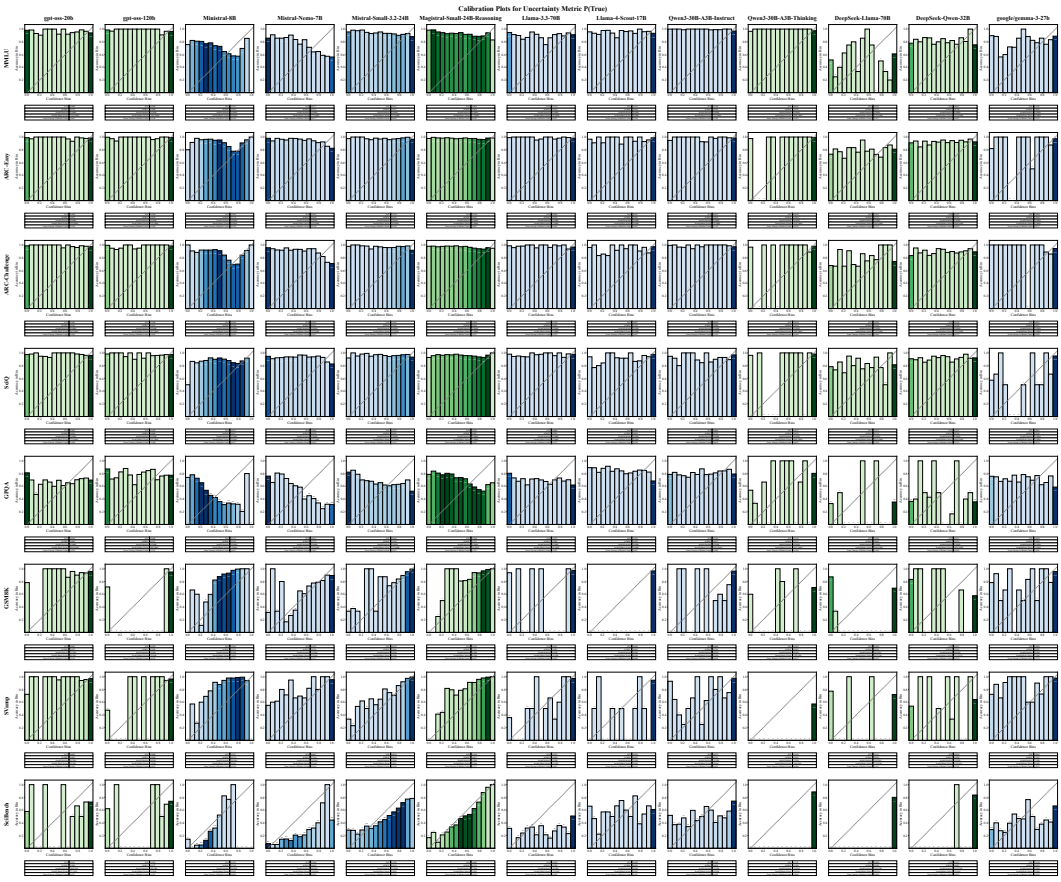


Figure A.6: **Calibration Plots for P(True).** Columns correspond to the models and rows to datasets.

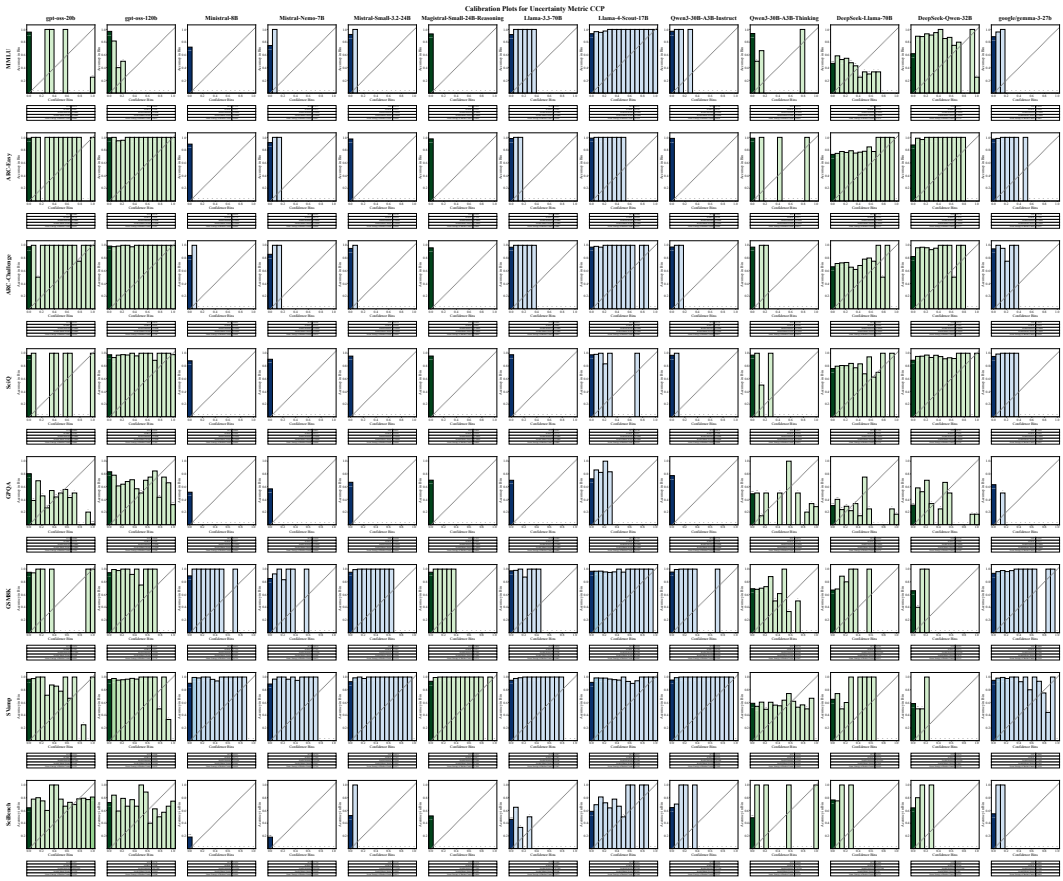


Figure A.7: **Calibration Plots for CCP.** Columns correspond to the models and rows to datasets.

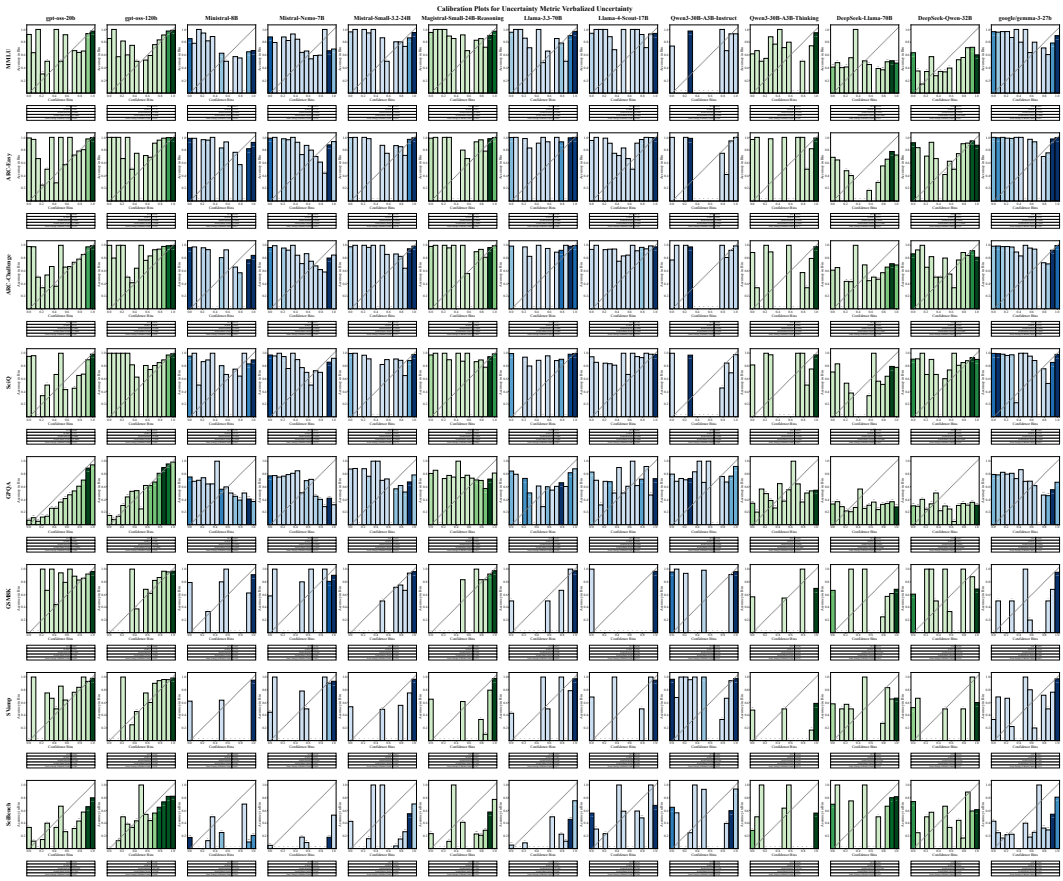


Figure A.8: **Calibration Plots for Verbalized Uncertainty.** Columns correspond to the models and rows to datasets.

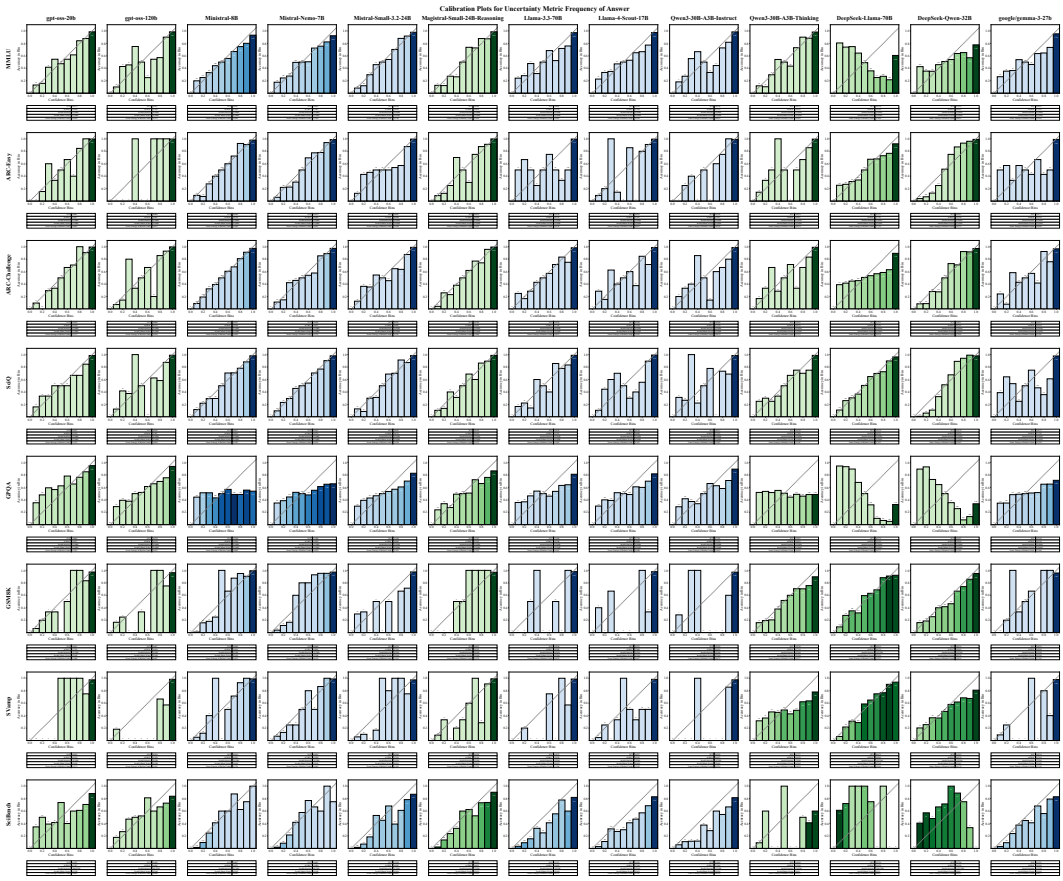


Figure A.9: **Calibration Plots for Frequency of Answer.** Columns correspond to the models and rows to datasets.

A.6.4 VERBALIZED UNCERTAINTY CONFIDENCE SCORE DISTRIBUTION

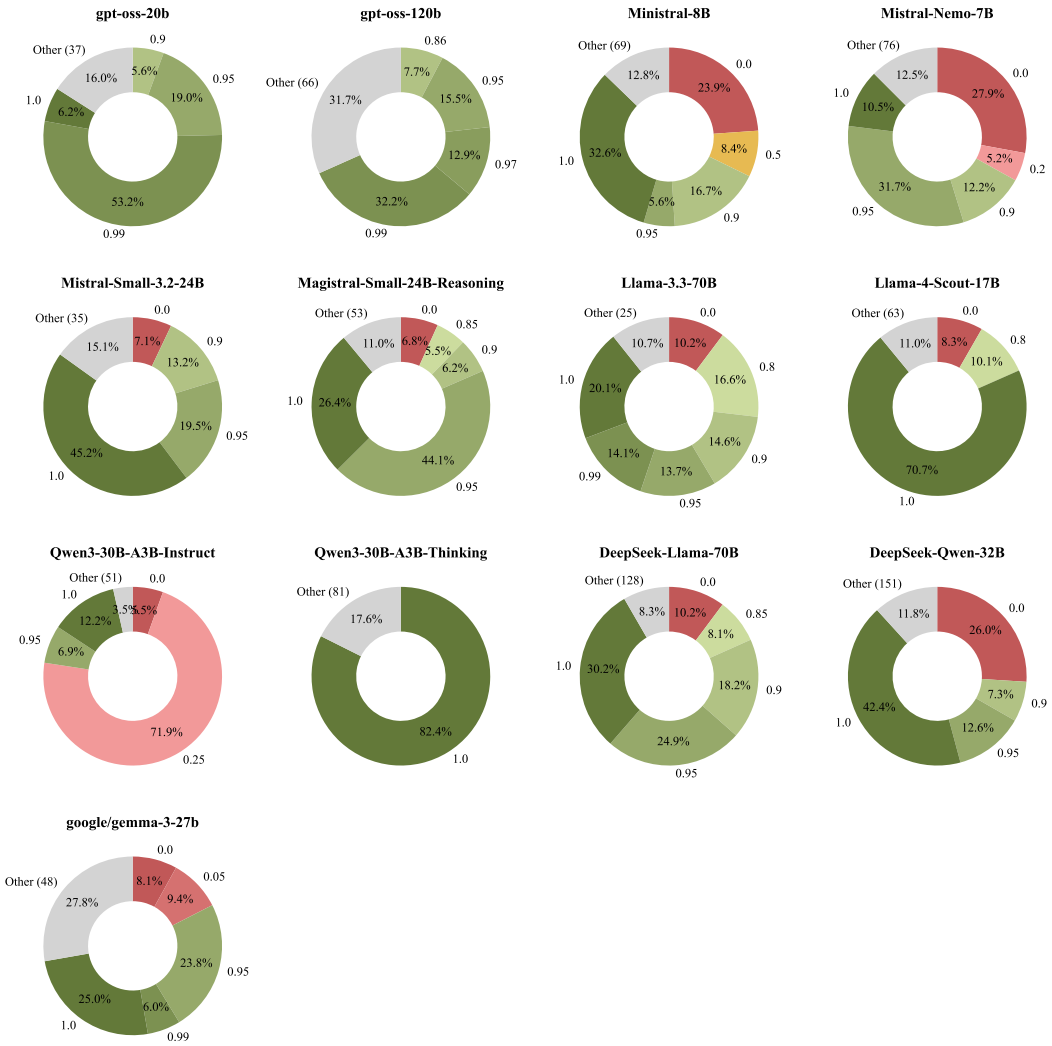


Figure A.10: **Distribution of Confidence Scores Provided During Verbalized Uncertainty Prompting Across Models.** Value counts are aggregated over all datasets (57, 500 prompts in total). Confidence scores seen in less than 5% of the total responses have been grouped into "Other", with the number of distinct confidence scores shown in brackets.

A.6.5 P(TRUE) CONFIDENCE SCORE DISTRIBUTION

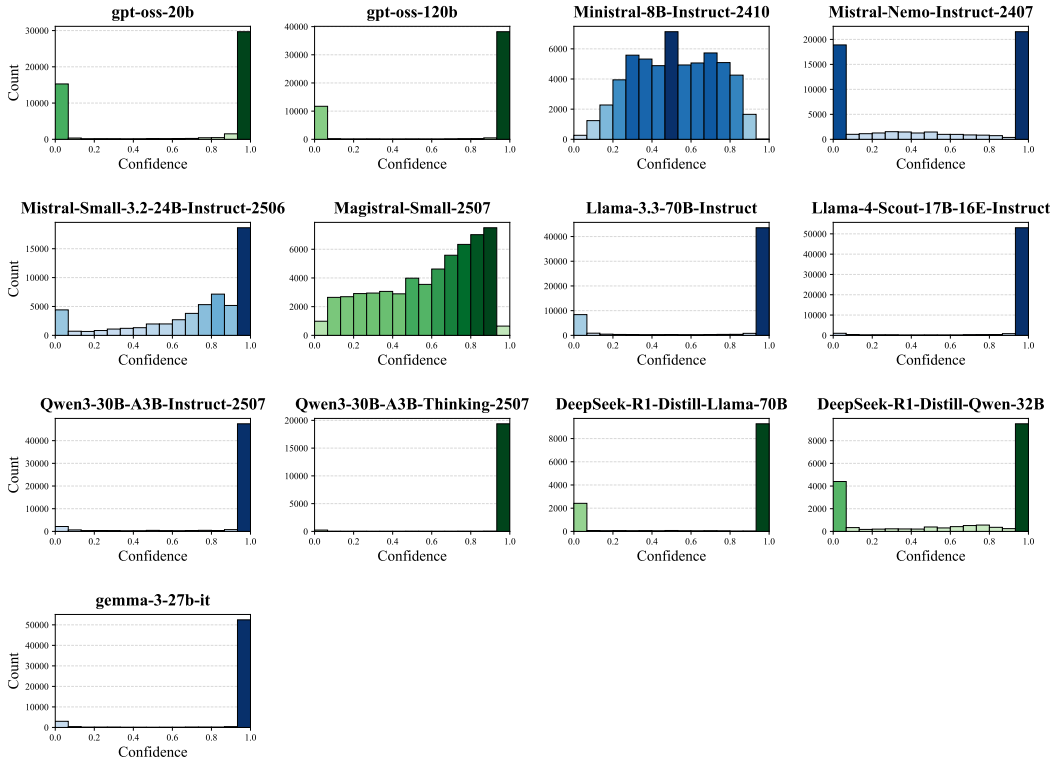


Figure A.11: **Distribution of Confidence Scores Assigned by P(True) Across Models.** Confidence scores have been aggregated across all datasets ($57,500$ prompts in total). Most models exhibit a clear polarization towards either `(A)` (representing the model's confidence that the answer is true) or `(B)` (representing the model's confidence that the answer is false) regarding the token probabilities.