

Semantic Calibration Prevails Where Token Confidence Fails: Benchmarking Long-Form Scientific QA

Philip Müller¹ Nicholas Popović^{2,3} Michael Färber^{2,3} Peter Steinbach¹

¹Helmholtz-Zentrum Dresden-Rossendorf ²Technische Universität Dresden ³ScaDS.AI

p.mueller@hzdr.de nicholas.popovic@tu-dresden.de michael.farber@tu-dresden.de p.steinbach@hzdr.de

Abstract

Reliable uncertainty quantification (UQ) is essential for safe deployment of large language models (LLMs) in scientific question answering, where long-form outputs exceed practical human verification at scale. We introduce the first large-scale benchmark for UQ calibration in long-form, reasoning-demanding scientific QA, evaluating four UQ methods on 685,000 responses across up to 20 LLMs and seven datasets, supported by an extensible open-source framework whose shared-generation design enables reproducible cross-method comparisons. Instruction tuning is shown to associate with systematic token probability polarization, collapsing confidence distributions and undermining the reliability of token-level uncertainty signals. Reasoning model families diverge: some reproduce this polarization while others actively mitigate it, a pattern that clusters by provider and suggests training pipeline design as a key differentiating factor. Verbalized and token-aggregation sequence-level methods fail systematically. Only semantic consistency, as measured by consistency of the final answer, yields well-calibrated outputs, providing the first large-scale evidence that semantic calibration persists in multi-step, dependency-rich reasoning settings.

1 Introduction

Large language models (LLMs) are increasingly deployed for complex question answering tasks spanning knowledge-intensive scientific queries, multi-step reasoning problems, and research question answering. These LLMs now form core components of search-augmented agentic reasoning pipelines (Jin et al., 2025), automated knowledge-discovery systems (Xiong et al., 2024a), and AI-powered search (Kelly et al., 2023). In these settings, long-form outputs with interdependent reasoning steps are generated at a scale that exceeds practical human verification. Automated UQ is therefore essen-

tial for flagging unreliable outputs and mitigating error propagation in downstream pipelines. Undetected propagated uncertainties can lead to severe downstream errors that may remain unnoticed in scientific QA, including in agentic research assistance (Gu and Krenn, 2025), automated research direction discovery (Lu et al., 2024; Kramer et al., 2026), and long-form information synthesis.

Calibration characterizes whether a model’s stated confidence reflects its actual likelihood of being correct (accuracy), allowing confidence scores to support reliable automated decisions and risk-aware system behavior (Guo et al., 2017). This task is orthogonal to selective prediction (Xia, 2026), which focuses on identifying individually unreliable generations for abstention under risk-coverage trade-offs. Selective prediction has received substantially more benchmarking attention (Fadeeva et al., 2023).

Token-level confidence scores are the most commonly deployed UQ signal, but their reliability in instruction-tuned and reasoning models remains poorly understood at scale (OpenAI, 2023). When these scores systematically collapse through instruction tuning, any downstream process relying on confidence-based filtering receives near-constant high-confidence signals regardless of actual answer quality, a failure mode invisible without distribution-level analysis.

Scientific QA as a benchmark domain. Scientific QA spans a full reasoning spectrum, from fact retrieval to multi-step symbolic reasoning, producing long-form answers with interdependent reasoning steps. Answers are often too complex for casual verification, making UQ most needed and least validated precisely here. Prior calibration benchmarks target short-form or classification-style outputs (Guo et al., 2017; OpenAI, 2023; Fadeeva et al., 2023). This benchmark instead evaluates calibration on unrestricted multi-sentence reasoning

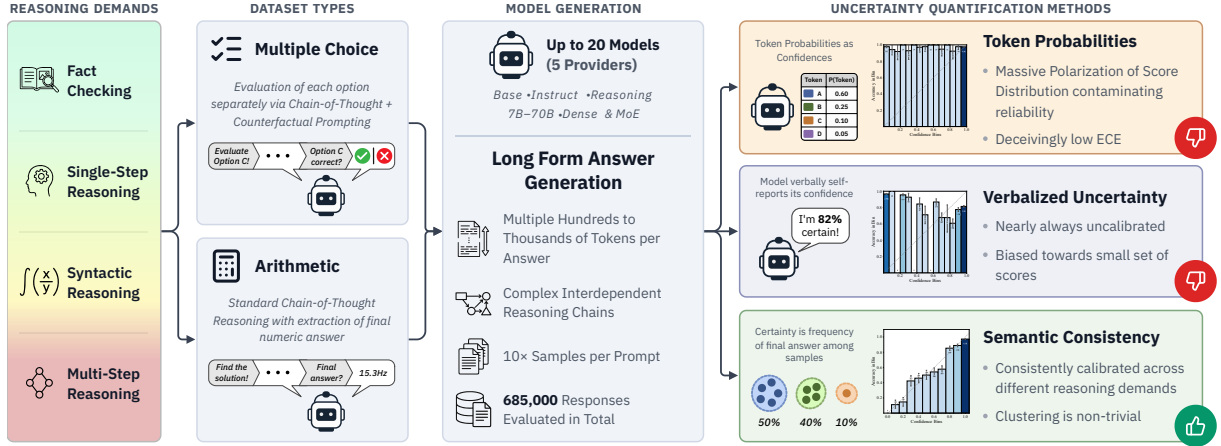


Figure 1: **Overview of the Benchmark and Evaluation Pipeline.** Diverse reasoning demands and QA task formats are used to generate large-scale long-form responses from up to 20 language models. Four uncertainty quantification methods are evaluated under increasing reasoning complexity, revealing consistent calibration only for semantic consistency while token probabilities and verbalized uncertainty remain poorly calibrated.

chains. Verifiable ground truth in the form of multiple choice question answering (MCQA) labels or arithmetic answers enable automatic computation of calibration metrics without costly human annotation.

Research questions. This paper addresses two research questions:

- (RQ1)** To what extent are token-level probabilities a calibrated measure of model confidence, and what effect does instruction tuning or the reasoning process have on calibration?
- (RQ2)** To what extent do sequence-level UQ methods provide reliable uncertainty estimates for long-form answers in scientific QA, across tasks spanning retrieval-based to reasoning-intensive questions?

Contributions. **First**, we introduce the first large-scale calibration benchmark for long-form scientific QA, comprising 685,000 responses across up to 20 LLMs (base, instruction-tuned, and reasoning variants) and seven datasets, with a reproducible shared-generation design in `async-graph-bench`, an open-source framework developed for this paper. **Second**, we characterize instruction-tuning-induced probability mass polarization across five model families and four MCQA datasets; reasoning families diverge by provider in their response to this effect (see Section 4; all findings are observational). **Third**, we show that verbalized and token-aggregation methods fail systematically while semantic consistency via answer consistency yields well-calibrated outputs, extending semantic cali-

bration into multi-step, dependency-rich reasoning settings.

2 Related Work

Token-level UQ and calibration. UQ for LLMs addresses the hallucination problem intrinsic to token-level probabilistic generation (Maynez et al., 2020; Sui et al., 2024; Zhang et al., 2023b; Banerjee et al., 2024). A broad range of UQ methods has been established for LLM-generated text (Geng et al., 2024), though many assume equal token importance and thereby misrepresent sequence-level confidence (Ullrich et al., 2025; Kuhn et al., 2023). Empirical evidence consistently shows that base LLMs are better calibrated than their instruction-tuned counterparts, as measured by Expected Calibration Error (ECE) (OpenAI, 2023; Tian et al., 2023; Wang et al., 2025). The GPT-4 Technical Report (OpenAI, 2023) documents ECE degradation after instruction tuning without analyzing the underlying score distribution shift. Cruz et al. (2024) observed polarization of label probability scores due to instruction tuning in a narrow binary classification setting, leaving generalizability across model families and diverse reasoning tasks open.

Benchmarking calibration in LLMs. LM-Polygraph (Fadeeva et al., 2023) is the primary UQ benchmarking framework for LLMs, evaluating selective prediction via the Prediction Rejection Ratio (PRR), but does not evaluate calibration and targets short-form outputs. Tao et al. (2025) conduct a broad model comparison restricted to a single dataset; Liu et al. (2025) survey UQ methods with-

out empirical calibration comparison. Calibration-focused evaluation across diverse model families, reasoning-demanding tasks, and long-form generation remains poorly characterized at scale.

Semantic calibration. Nakkiran et al. (2026) demonstrate that LLMs exhibit strong calibration at the level of semantic answer clusters despite token-level miscalibration, but restrict their analysis to short, single-hop answers. Whether semantic calibration persists in multi-step, dependency-rich reasoning settings remains open, and is directly addressed by our benchmark. No prior work simultaneously evaluates calibration across long-form scientific QA, multiple UQ paradigms, and large-scale model diversity spanning base, instruction-tuned, and reasoning models across five providers.

3 Methodology

3.1 Benchmark Design

A fundamental challenge in cross-method UQ comparison is that integrating new methods either requires rerunning the entire benchmark across all methods, which is computationally costly, or evaluating the new method on separately generated outputs. In the latter case, stochastic decoding prevents performance differences from being cleanly attributed to the UQ method itself rather than to generation variance, thereby compromising comparability. We therefore evaluate all UQ methods on identical shared generations: outputs are generated once, cached, and reused across all evaluations. We developed `async-graph-bench`, an open-source framework for reproducible multi-stage UQ benchmarking, released alongside the computed UQ scores and full reproduction instructions at <https://github.com/muelphil/llm-uncertainty-bench>.

Scientific QA is chosen as the target domain because its dependency-rich reasoning chains expose compounding uncertainty: errors propagate across dependent steps in ways short-form benchmarks conceal, making it an ideal stress test for UQ methods. Crucially, MCQA structure is used exclusively for automatic answer verification, not to constrain output format; APriCoT prompting (Moore et al., 2024) elicits open-ended multi-sentence reasoning traces and arithmetic tasks impose no length restrictions.

3.2 Dataset Selection

The requirement for verifiable ground truth to compute calibration metrics (ECE, AUROC) drives use of structured tasks: four MCQA datasets and three arithmetic QA datasets, with emphasis on scientific QA covering physics, chemistry, and biology, plus mathematics for a low-complexity arithmetic baseline. Datasets span three reasoning tiers (full details in the appendix):

At the **recall tier**, MMLU (Hendrycks et al., 2021), SciQ (Welbl et al., 2017), and ARC-Easy (Clark et al., 2018) test factual recall and single-step reasoning. At the **multi-step tier**, ARC-Challenge (Clark et al., 2018), GSM8K (Cobbe et al., 2021) with its multiple-choice variant GSM-MC (Zhang et al., 2024b), and SVAMP (Patel et al., 2021) require chained reasoning steps. At the **high-complexity tier**, SciBench (Wang et al., 2024b) requires formula manipulation, unit tracking, and multi-equation operations; GPQA (Rein et al., 2024) demands expert multi-hop reasoning explicitly designed to resist lookup.

Response lengths, as reported in Table A.9, confirm the long-form nature of the generated outputs. The average mean response length per dataset ranges from 286 tokens on SVAMP to 2,403 tokens on GPQA. On the most challenging datasets, this increases substantially for reasoning models, with mean response lengths averaged across models reaching 4,194 tokens on GPQA and 4,086 tokens on SciBench. Even instruction-tuned models exceed 1,000 tokens on average for these datasets (GPQA: 1,123 tokens; SciBench: 1,029 tokens).

3.3 Model Selection

Models are chosen to cover a broad and up-to-date LLM spectrum while ensuring reproducibility via public open-weight models. The selection spans five providers (OpenAI, Mistral, Meta, Qwen, and Google), varying sizes (7B to 70B parameters), three variant types (base, instruction-tuned, and reasoning), and multiple architectures including mixture-of-experts variants. Instruction-tuned and reasoning models are complemented by their base model counterparts to enable controlled comparison of instruction-tuning and reasoning-training effects.

3.4 First Experiment: Token-Level Label Probability Calibration

The first experiment replicates and substantially extends the token-level calibration analysis from the GPT-4 Technical Report (OpenAI, 2023), expanding along three dimensions. First, model coverage is extended to base, instruction-tuned, and reasoning model variants across all five providers, rather than a single model. Second, dataset diversity spans four MCQA datasets covering the full reasoning complexity spectrum: MMLU (factual recall), ARC-Challenge (multi-step reasoning), GSM8K-MC (symbolic arithmetic), and GPQA (expert multi-step reasoning). Third, analysis depth includes calibration plots and score distribution analysis alongside ECE, characterizing what causes calibration degradation rather than only observing that it occurs.

Models are prompted to respond with a single answer label (A, B, C, or D) for each MCQA item. Token-level probabilities assigned to these label tokens serve as confidence scores. Reasoning models produce extended reasoning traces, but the final label token probability is the confidence measure. Normalization of token probabilities over answer labels is applied and validated following Wang et al. (2024a).

3.5 Second Experiment: Sequence-Level UQ Calibration

Our second experiment evaluates sequence-level UQ calibration across all seven datasets. It is restricted to instruction-tuned and reasoning models, as base models exhibit insufficient task comprehension for long-form QA. All four UQ methods are applied to identical cached long-form generations. Metric implementation details can be found in subsection A.11.3.

Method selection. Calibration requires normalized per-instance confidence scores in $[0, 1]$; methods producing rankings, unnormalized scores, or claim-level outputs are excluded. The four selected methods cover three paradigms: token-probability aggregation, self-evaluation, and semantic consistency. A comprehensive method overview, relevance statement and selection rationale are provided in Sections A.7 and A.8.

Verbalized Uncertainty. Tian et al. (2023) propose prompting the model post-generation for a self-assessed probability of correctness. This

method is straightforward but ignores token probability distributions and may be susceptible to training data bias.

P(True). Kadavath et al. (2022) propose prompting the model post-generation to classify its own answer as (A) True or (B) False. Token probabilities of these two labels are used as confidence scores.

Claim-Conditioned Probability (CCP). Fadeeva et al. (2024) propose CCP, which estimates token-level uncertainty by clustering the top probable token alternatives into entailing and contradicting classes using natural language inference (NLI). Sequence confidence is computed as the product of per-token confidences.

Answer Consistency. Wang et al. (2023) propose measuring the proportion of semantically equivalent answers across multiple sampled generations. For MCQA datasets, a follow-up prompt assesses whether the generation concludes that the evaluated answer option is correct; responses are clustered per answer option by this binary label. For arithmetic QA, a follow-up prompt extracts the final numeric answer from each generation; responses are clustered by the extracted value. This design separates the initial long-form generation from the answer-mapping step used for ground-truth comparison. All confidence scores are computed from agreement among the sampled judgments per answer option alone; gold labels are introduced only afterwards to determine the correctness of each judgment and to evaluate calibration (Sections A.11.2 and A.11.3). This method is computationally expensive but proxies self-consistency prompting.

3.6 Prompting Strategy

MCQA introduces selection bias and positional bias, including systematic preferences for certain label positions (Zheng et al., 2024) and token and formatting frequency bias (Myrzakhan et al., 2024). APriCoT prompting (Moore et al., 2025) addresses both biases by evaluating each answer choice independently as correct or incorrect via Chain-of-Thought combined with counterfactual prompting. This preserves a verifiable label format while eliciting genuine long-form reasoning responses. APriCoT produces four separate prompts per MCQA item, one per answer option. The MCQA structure constrains ground truth verification only, not

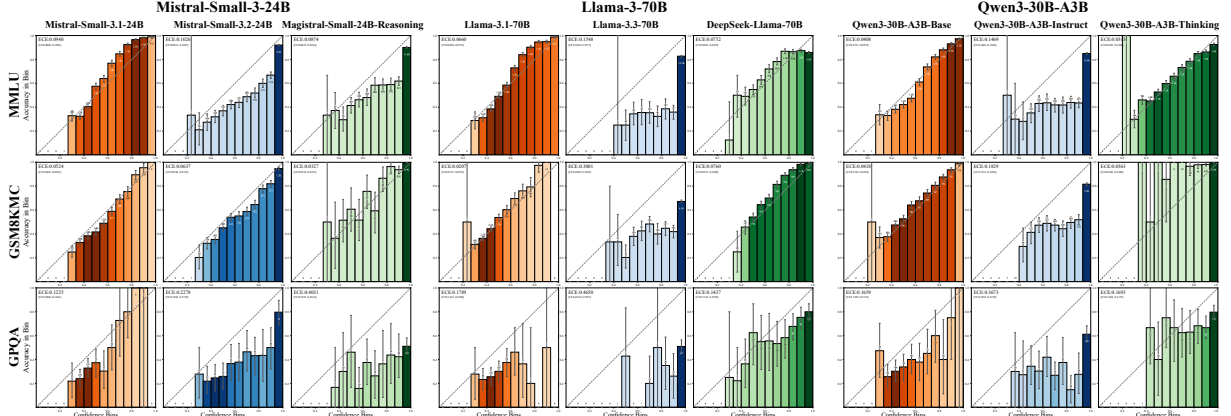


Figure 2: **Calibration Plots Using Label Probabilities as Uncertainty Scores.** Columns correspond to three selected model families: base variants in orange, instruction-tuned in blue, reasoning in green. Rows refer to MMLU, GSM8KMC, and GPQA. Darker shading indicates more items per confidence bin; each bin is labelled with its sample count. Results use Prompt Design 1 with structured decoding enabled.

response format; responses are multi-sentence reasoning chains, not single label tokens. For Arithmetic QA we use Chain-of-Thought (CoT) prompting.

Each dataset is subsampled to 250 items with 10 generations per prompt, yielding 2,500 responses per arithmetic dataset and 10,000 per MCQA dataset via APriCoT ($250 \times 4 \times 10$); full counts (685,000 total) are detailed in Section A.6.

For the first experiment, four alternative 3-shot prompts are designed to address task comprehension differences between base and instruction-tuned models. The prompt maximizing average probability mass on label tokens is selected per model, as a proxy for task comprehension. Structured decoding is used to reliably obtain relative label probabilities and prevent invalid answers.

3.7 Evaluation Metrics

Expected Calibration Error (ECE) is the most widely used aggregate metric for calibration evaluation (Guo et al., 2017; Hendrycks and Gimpel, 2017). However, as a coarse summary statistic, it can obscure important distributional pathologies. In particular, when confidence estimates collapse into a narrow or highly polarized range, ECE becomes dominated by overall model accuracy rather than reflecting whether confidence scores remain informative at the instance level. This limitation is consistent with the broader observation that summary statistics can conceal substantial structural differences in the underlying data distribution (Matejka and Fitzmaurice, 2017). Consequently, aggregate calibration metrics alone are insufficient

for a comprehensive assessment of calibration behavior.

ECE failure under polarization. When probability mass collapses, as documented in the binary classification setting by Cruz et al. (2024) and extended here across model families, confidence scores become effectively degenerate. If the collapsed score range coincides with the model’s average accuracy, ECE may remain deceptively low despite the confidence estimates no longer supporting meaningful thresholding, selective prediction, or risk-aware decision making. We therefore pair calibration plots with ECE throughout this benchmark to preserve visibility into the underlying score distributions and calibration structure.

To assess robustness, we estimate bootstrap confidence intervals at the 95% confidence level over calibration curves and representative ECE values using 1000 resamples. AUROC values are reported in the full result plots and tables for completeness and to make our benchmark results amenable to the selective prediction community. As the primary focus of this work is calibration, AUROC is not used as an evaluation criterion.

4 Results: Token-Level Calibration

4.1 Main Findings

Figure 2 reports a subset of representative calibration plots. All results shown use Prompt Design 1 (see Section A.9.1), which was found to optimize task comprehension across model types (see Section A.9.2). Structured decoding has been employed to generate the final result, after normal-

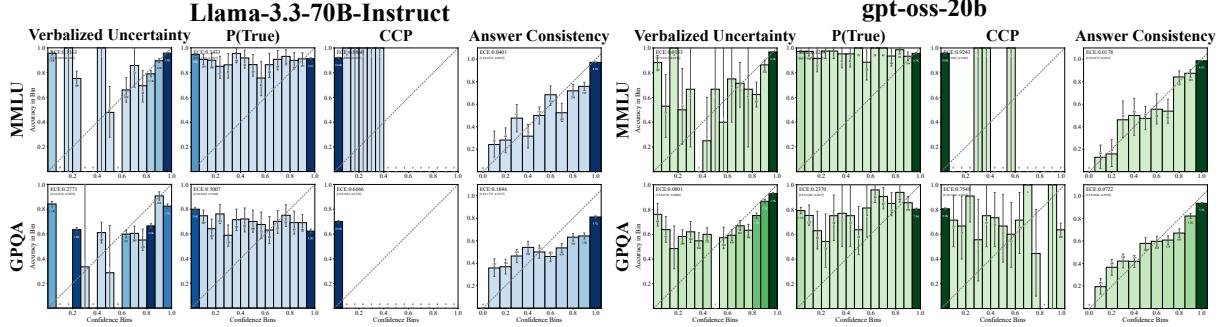


Figure 3: **Selected Calibration Plots for the Four Evaluated Sequence-Level UQ Methods.** Results for *Llama-3.3-70B* (left) and *gpt-oss-20b* (right) on MMLU and GPQA.

ization has been validated (see Section A.10.2). Comprehensive results for all models, datasets and prompt designs are included in Section A.10.1.

ECE increases with reasoning complexity: symbolic and multi-step tasks such as GSM8K and GPQA exhibit substantially higher ECE than factual retrieval tasks such as MMLU. Token-level probabilities capture aleatoric uncertainty in fact retrieval but become unreliable for reasoning tasks requiring epistemic uncertainty estimation across multiple dependent steps.

Instruction tuning degrades ECE across most model families. Calibration plots reveal systematic polarization: many instruction-tuned models concentrate nearly all probability mass on a single label, rendering confidence scores near-constant across items and destroying their value as per-instance reliability signals. This polarization is characterized across five model families and four MCQA datasets spanning the full reasoning complexity spectrum; Mistral models are a notable exception.

4.2 Reasoning Model Divergence

Reasoning models diverge markedly across providers. *gpt-oss* and *Magistral* commit to a single answer option early in the reasoning process, and token probabilities collapse as in instruction-tuned models. The reasoning trace does not distribute probability mass across alternatives, resolving uncertainty prematurely rather than exploring multiple possibilities.

DeepSeek-R1 and *Qwen3*, by contrast, actively explore and assess multiple answer options in their reasoning traces, externalizing uncertainty via epistemic markers across alternatives. Token probabilities remain distributed, and calibration is substantially preserved or improved relative to base models, at the cost of considerably longer reason-

ing traces.

These differences cluster by provider rather than by scale or architecture. Training pipeline design, including RLHF targets, CoT supervision, and reward shaping, is the most plausible differentiating factor. Without access to training configurations, however, this remains an observational hypothesis rather than a causal claim. The benchmark provides a foundation for further investigation into how individual training stages influence model calibration and polarization.

5 Results: Sequence-Level Calibration

Figure 3 reports a representative subset of calibration plots. Extensive visualizations for all models and datasets across UQ methods are provided in Section A.12.2. Figure 4 reports the distribution of calibration correlation across all (model, dataset) pairs for each UQ method. Additional aggregate comparisons based on ECE and AUROC are reported in Figures A.14 and A.15. The following subsections provide method-wise analyses of these calibration behaviors.

5.1 Verbalized Uncertainty

Models default to a small set of confidence scores, with high-confidence scores dominating the output distribution, as seen in Figure 5a. This biased distribution is associated with training data composition and instruction tuning. Calibration plots show no correlation between verbalized scores and answer accuracy across instruction-tuned models.

A slight improvement is observed in reasoning models, most notably the *gpt-oss* family, which achieves better correlation with correctness on challenging datasets such as GPQA and SciBench. Without access to training configurations and workflow setups, we can only speculate that this reflects

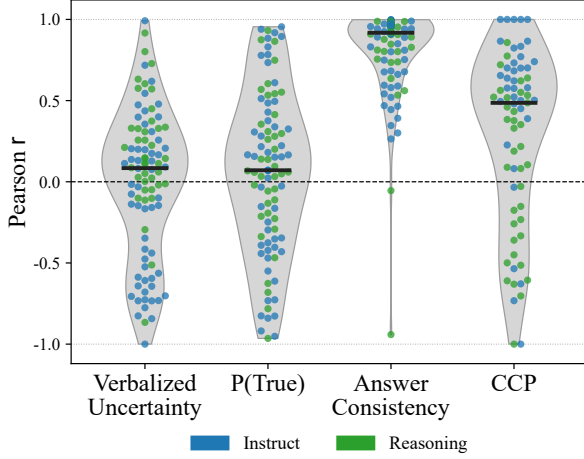


Figure 4: **Calibration Correlation by UQ Method.** Distribution of Pearson calibration correlation (r) across all (model, dataset) pairs for each UQ method. Correlations are computed between the bin-confidence midpoints and the empirical bucket accuracies of the corresponding calibration curves, excluding empty bins and weighting all non-empty bins equally. Higher values indicate stronger alignment between confidence and empirical accuracy. Violins summarize the overall distribution and horizontal bars indicate the median.

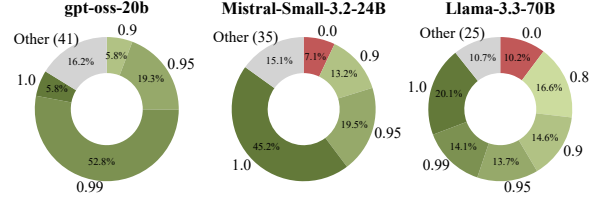
training-related factors specific to this model family.

The failure mode is characterized by output distribution collapse driven by training data bias: instruction-tuned models default to high-confidence scores regardless of actual answer correctness. This is not a calibration mismatch in the statistical sense but a fundamental collapse of the output distribution, making verbalized uncertainty unreliable as a per-instance uncertainty signal.

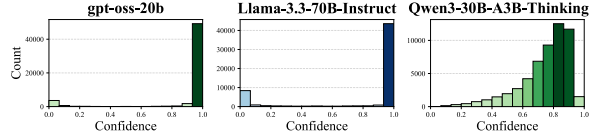
5.2 P(True)

Many models assign near-1.0 confidence scores with little use of intermediate values, resulting in a polarized distribution, as seen in Figure 5b. Calibration plots show no meaningful correlation between P(True) scores and actual correctness across models.

P(True) relies on token-level probabilities for its confidence signal. The polarization documented in Section 4 directly propagates: when token probabilities collapse, P(True) assigns near-constant high confidence regardless of answer correctness. The failure of P(True) is thus a downstream consequence of token-level polarization, providing a direct cross-experiment link between the findings of Section 4 and those of this section.



(a) **Verbalized Uncertainty.** Confidence scores seen in less than 5% of total responses are grouped into “Other”, with the number of distinct scores in brackets.



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

Figure 5: **Distribution of Confidence Scores Across Selected Representative Models.** Scores aggregated across all datasets (57,500 prompts total). See Section A.12.4 and Section A.12.5 for extensive plots across all models.

5.3 Claim-Conditioned Probability (CCP)

CCP suffers from vanishing sequence-level confidence scores: multiplying many per-token confidences drives scores toward zero regardless of correctness, yielding uninformative near-zero signals for all long-form outputs. Calibration plots show no meaningful alignment with correctness and high calibration error across models or datasets. Three structural failure modes are identified: (1) vanishing scores due to cumulative probability multiplication; (2) NLI misclassification instability, where a single incorrect entailment decision propagates multiplicatively and dominates the sequence score; and (3) stop-word contamination, where grammatical tokens dilute the semantic signal during aggregation. Each represents a concrete direction for future improvement.

5.4 Answer Consistency

Higher answer frequencies strongly correlate with correctness across both MCQA and arithmetic tasks. More challenging datasets such as GPQA and SciBench exhibit greater answer diversity across sampled generations, reflecting higher model uncertainty on these tasks.

Calibration plots confirm well-aligned confidence estimates across model families and dataset types, as seen in Section A.12.3, making Answer Consistency the most reliable approach evaluated

in this benchmark. Semantic calibration persists in long-form, multi-step, dependency-rich reasoning settings where all other methods collapse. This extends and validates the finding of Nakkiran et al. (2026) into a substantially more challenging reasoning regime: answer consistency remains the most reliable and well-calibrated uncertainty indicator even in complex scientific QA requiring extended reasoning chains.

How Answer Consistency differs from single-trajectory methods. Answer Consistency differs in kind from the other evaluated methods: while token probabilities and verbalized confidence read uncertainty off a single generation trajectory, it measures agreement across independently generated semantic outcomes, i.e., whether the model consistently converges to the same semantic answer rather than how confident it appears along one generation path. We interpret this contrast as a plausible reason for the observed pattern, in which single-trajectory signals collapse after instruction tuning while Answer Consistency remains calibrated; the benchmark is observational, and we do not claim to have established this mechanism.

Answer Consistency is not universally reliable. Three conditions can undermine it: *systematic shared errors*, in which a model repeatedly produces the same incorrect answer and thus assigns high confidence to an incorrect response; *ambiguous questions* that admit multiple valid interpretations, in which case diverse answers reflect genuine task uncertainty rather than model miscalibration; and *errors in semantic equivalence clustering*, in which the answer-mapping step merges distinct answers or splits equivalent ones. Our results indicate that none of these conditions is systematically in force on the structured tasks evaluated here, which is consistent with the well-calibrated behavior observed across models and datasets; the clustering condition nonetheless becomes harder to guarantee for open-ended QA. See Section A.11.2 for the formal definitions and the Limitations section for a discussion of scope.

The primary limitation is computational cost, which scales linearly with the number of samples per prompt (10 samples per prompt in this benchmark; Section A.6). Answer Consistency is restricted to settings where cluster formation is tractable. For practitioners requiring calibrated confidence estimates in scientific or reasoning-intensive QA, Answer Consistency is the most re-

liable option among currently evaluated methods; token-based methods (CCP, P(True)) and verbalized uncertainty should not be used as calibration quality signals without further validation.

6 Discussion

6.1 Implications of Token-Level Polarization

Polarization is not identical to overconfidence. Overconfidence denotes a systematic tendency for confidence scores to exceed actual accuracy; polarization denotes a full collapse of the score distribution. A polarized model has lost any discriminative signal, and post-hoc recalibration cannot recover information no longer present in the scores: standard procedures such as temperature scaling or Platt scaling presuppose that scores retain some ordinal relationship with correctness. Practitioners relying on token-level confidence thresholds for abstention or routing in reasoning-intensive domains should therefore verify that scores retain discriminative value before deployment.

6.2 Reasoning Models and Calibration

The provider-clustered divergence documented in Section 4.2 suggests that training pipeline design, not the mere presence of a reasoning process, is the key determinant of calibration quality. The behavioral contrast between polarization-mitigating models (*DeepSeek-R1*, *Qwen3*) and polarization-reproducing ones (*gpt-oss*, *Magistral*) is consistent with differences in reasoning supervision: exploratory multi-option deliberation during the reasoning trace appears to preserve probability mass distribution. This interpretation is observational; controlled ablation of specific training stages is a direct open problem uncovered by this benchmark.

7 Conclusion

Our large-scale evaluation of four UQ methods across 20 models demonstrates that semantic calibration persists in complex reasoning tasks, yet token-level probabilities are largely uninformative due to polarization induced by instruction tuning. Although reasoning processes can attenuate this effect, the strongest differences are observed across providers rather than architectures. Verbalized uncertainty and CCP suffer from distribution collapse and instability, whereas Answer Consistency proves reliable despite higher computational cost. Our findings underscore the necessity of evaluating UQ at the semantic outcome level (Nakkiran et al.,

2026) and call for sustained benchmarking to disentangle the contributions of architecture, fine-tuning, and training data to calibration behavior.

Future work should extend the evaluation to additional UQ methods and study the effects of inference parameters and prompting strategies on calibration. Reformulating MC tasks as open-ended would isolate task-specific uncertainty sources. Controlled ablation of individual training stages (SFT, RLHF, RLAI, and inference-time scaling) would determine whether the provider-clustered behavioral differences observed here are driven by specific pipeline components, transforming an observational finding into an actionable mechanism for improving model calibration.

Limitations

Despite approximating open question answering for MCQA via APriCoT prompting and complementing with arithmetic QA, the benchmark scope remains restricted to structured task formats. This is a deliberate design choice: calibration metrics such as ECE require unambiguous correctness labels, and introducing LLM-judged or human-evaluated open-ended correctness would add a second source of uncertainty to the measurement pipeline, making it impossible to disentangle miscalibration in the model’s self-assessment from noise in the evaluation method itself. Even so, the structured-task scope limits generalizability to some extent, e.g. to fully open-ended generation and other task types. All datasets are in English, leaving open questions about behavior in multilingual or code-generation settings.

Despite APriCoT mitigation, the MC setting may introduce steering effects on model responses. Arithmetic QA provides complementary evidence with fully open-ended numerical answers, and the consistency of findings across both task types and model variants supports generalizability. Explicit cross-domain validation nonetheless remains necessary.

Answer Consistency validity is contingent on the quality of semantic equivalence clustering. This is tractable for MCQA and arithmetic QA but non-trivial for open-ended QA.

Our benchmark cannot establish why instruction tuning induces polarization or why certain reasoning models mitigate it. We speculate that provider-level behavioral clustering is due to training pipeline differences. Controlled ablations were

not feasible as training pipeline setups are undisclosed by model providers.

Inference parameters are held constant, including fixed temperature and fixed decoding settings across all models and datasets. Variation in these parameters is not examined in this study.

References

- Sourav Banerjee, Ayushi Agarwal, and Saloni Singla. 2024. [LLMs Will Always Hallucinate, and We Need to Live With This](#). *Preprint*, arXiv:2409.05746.
- Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. [INSIDE: LLMs’ Internal States Retain the Power of Hallucination Detection](#). In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Jiuhai Chen and Jonas Mueller. 2024. [Quantifying uncertainty in answers from any language model and enhancing their trustworthiness](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5186–5200, Bangkok, Thailand. Association for Computational Linguistics.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. [Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge](#). *Preprint*, arXiv: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. 2021. [Training Verifiers to Solve Math Word Problems](#). *Preprint*, arXiv:2110.14168.
- André F. Cruz, Moritz Hardt, and Celestine Mendler-Dünner. 2024. [Evaluating language models as risk scores](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Maxime Darrin, Pablo Piantanida, and Pierre Colombo. 2023. [Rainproof: An umbrella to shield text generator from out-of-distribution data](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5831–5857, Singapore. Association for Computational Linguistics.
- DeepSeek-AI. 2025. [DeepSeek-R1 Incentivizes Reasoning in Large Language Models through Reinforcement Learning](#). *Nature*.
- Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. [Shifting Attention to Relevance: Towards the Predictive Uncertainty Quantification of](#)

- Free-Form Large Language Models.** In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 5050–5063. Association for Computational Linguistics.
- Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsybalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. 2024. **Fact-Checking the Output of Large Language Models via Token-Level Uncertainty Quantification.** In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Findings of ACL, pages 9367–9385. Association for Computational Linguistics.
- 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. 2023. **LM-polygraph: Uncertainty estimation for language models.** In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 446–461, Singapore. Association for Computational Linguistics.
- Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Rico Sennrich, and Lucia Specia. 2020. **Unsupervised Quality Estimation for Neural Machine Translation.** *Transactions of the Association for Computational Linguistics*, 8:539–555.
- Gemma Team. 2025. **Gemma 3 Technical Report.** *arXiv preprint arXiv:2503.19786*.
- Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koepl, Preslav Nakov, and Iryna Gurevych. 2024. **A Survey of Confidence Estimation and Calibration in Large Language Models.** In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595. Association for Computational Linguistics.
- Xuemei Gu and Mario Krenn. 2025. **Forecasting high-impact research topics via machine learning on evolving knowledge graphs.** *Machine Learning: Science and Technology*, 6(2):025041.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.
- Haixia Han, Tingyun Li, Shisong Chen, Jie Shi, Chengyu Du, Yanghua Xiao, Jiaqing Liang, and Xin Lin. 2024. **Enhancing Confidence Expression in Large Language Models Through Learning from Past Experience.** *Preprint*, arXiv:2404.10315.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. **Measuring Massive Multitask Language Understanding.** In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Dan Hendrycks and Kevin Gimpel. 2017. **A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks.** In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.
- Juyeon Heo, Miao Xiong, Christina Heinze-Deml, and Jaya Narain. 2025. **Do LLMs estimate uncertainty well in instruction-following?** In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Mykyta Ielanskyi, Kajetan Schweighofer, Lukas Aichberger, and Sepp Hochreiter. 2025. **Addressing Pitfalls in the Evaluation of Uncertainty Estimation Methods for Natural Language Generation.** In *ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI*.
- Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. **Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning.** In *Second Conference on Language Modeling*.
- 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, and 17 others. 2022. **Language Models (Mostly) Know What They Know.** *Preprint*, arXiv:2207.05221.
- Sanyam Kapoor, Nate Gruver, Manley Roberts, Katie Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. 2024. **Large Language Models Must Be Taught to Know What They Don’t Know.** In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Dominique Kelly, Yimin Chen, Sarah E Cornwell, Nicole S Delellis, Alex Mayhew, Sodiq Onaolapo, and Victoria L Rubin. 2023. Bing chat: The future of search engines? *Proceedings of the Association for Information Science and Technology*, 60(1):1007–1009.
- Stefan Kramer, Mattia Cerrato, Jannis Brugger, Sašo Džeroski, and Ross D. King. 2026. **Automated Scientific Discovery: From Equation Discovery to Au-**

- tonomous Discovery Systems. *Machine Learning*, 115(5):109.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. **Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation**. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. **A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks**. In *Advances in Neural Information Processing Systems*, volume 31.
- Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. **Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models**. *Trans. Mach. Learn. Res.*, 2024.
- Xiaouu Liu, Tiejun Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. 2025. **Uncertainty Quantification and Confidence Calibration in Large Language Models: A Survey**. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V2, KDD 2025, Toronto ON, Canada, August 3-7, 2025*, pages 6107–6117. ACM.
- Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. **The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery**. *arXiv preprint arXiv:2408.06292*.
- Matéo Mahaut, Laura Aina, Paula Czarnowska, Momchil Hardalov, Thomas Müller, and Lluís Marquez. 2024. **Factual Confidence of LLMs: on Reliability and Robustness of Current Estimators**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 4554–4570. Association for Computational Linguistics.
- Andrey Malinin and Mark J. F. Gales. 2021. **Uncertainty Estimation in Autoregressive Structured Prediction**. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Justin Matejka and George Fitzmaurice. 2017. **Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing**. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1290–1294. Association for Computing Machinery.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. **On Faithfulness and Factuality in Abstractive Summarization**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1906–1919. Association for Computational Linguistics.
- Meta AI. 2024. **The Llama 3 Herd of Models**. *arXiv preprint arXiv:2407.21783*.
- Meta AI. 2025. **The Llama 4 Herd: The Beginning of a New Era of Natively Multimodal AI Innovation**. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>. Accessed: 2026-05-22.
- Mistral AI. 2024a. **Mistral NeMo**. <https://mistral.ai/news/mistral-nemo/>. Accessed: 2026-05-22.
- Mistral AI. 2024b. **Un Ministral, des Ministraux**. <https://mistral.ai/news/ministraux>. Accessed: 2026-05-22.
- Mistral AI. 2025a. **Magistral: A Reasoning Model Family**. *arXiv preprint arXiv:2506.10910*.
- Mistral AI. 2025b. **Mistral Small 3**. <https://mistral.ai/news/mistral-small-3>. Accessed: 2026-05-22.
- Kyle Moore, Jesse Roberts, Thao Pham, and Douglas Fisher. 2024. **Reasoning Beyond Bias: A Study on Counterfactual Prompting and Chain of Thought Reasoning**. *Preprint*, arXiv:2408.08651.
- Kyle Moore, Jesse Roberts, Thao Pham, and Douglas H. Fisher. 2025. **Chain of Thought Still Thinks Fast: APriCoT Helps with Thinking Slow**. In *Proceedings of the 47th Annual Meeting of the Cognitive Science Society, CogSci 2025, San Francisco, CA, USA, July 30 - August 2, 2025*. cognitivesciencesociety.org.
- Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. 2024. **Open-LLM-Leaderboard: From Multi-choice to Open-style Questions for LLMs Evaluation, Benchmark, and Arena**. *Preprint*, arXiv:2406.07545.
- Preetum Nakkiran, Arwen Bradley, Adam Golinski, Eugene Ndiaye, Michael Kirchhof, and Sinead Williamson. 2026. **Trained on Tokens, Calibrated on Concepts: The Emergence of Semantic Calibration in LLMs**. In *The Fourteenth International Conference on Learning Representations*.
- Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. **Kernel Language Entropy: Fine-grained Uncertainty Quantification for LLMs from Semantic Similarities**. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- OpenAI. 2023. **GPT-4 Technical Report**. ArXiv:2303.08774 [cs.CL].
- OpenAI. 2025. **GPT-OSS: Open-Weight Language Models**. <https://openai.com/index/introducing-gpt-oss/>. Accessed: 2026-05-22.

- Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. [Are NLP Models really able to Solve Simple Math Word Problems?](#) In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2080–2094. Association for Computational Linguistics.
- Qwen Team. 2025. [Qwen3 Technical Report](#). *arXiv preprint arXiv:2505.09388*.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. [GPQA: A graduate-level google-proof q&a benchmark](#). In *First Conference on Language Modeling*.
- Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J. Liu. 2023a. [Out-of-Distribution Detection and Selective Generation for Conditional Language Models](#). In *International Conference on Learning Representations (ICLR)*.
- Jie Ren, Yao Zhao, Tu Vu, Peter J. Liu, and Balaji Lakshminarayanan. 2023b. [Self-Evaluation Improves Selective Generation in Large Language Models](#). In *Proceedings on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops, 16 December 2023, New Orleans, Louisiana, USA*, Proceedings of Machine Learning Research, pages 49–64. PMLR.
- Vaishnavi Shrivastava, Percy Liang, and Ananya Kumar. 2023. [Llamas Know What GPTs Don't Show: Surrogate Models for Confidence Estimation](#). *Preprint*, arXiv:2311.08877.
- Peiqi Sui, Eamon Duede, Sophie Wu, and Richard Jean So. 2024. [Confabulation: The Surprising Value of Large Language Model Hallucinations](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 14274–14284. Association for Computational Linguistics.
- Junya Takayama and Yuki Arase. 2019. [Relevant and informative response generation using pointwise mutual information](#). In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 133–138, Florence, Italy. Association for Computational Linguistics.
- Linwei Tao, Yi-Fan Yeh, Minjing Dong, Tao Huang, Jialin Yu, Philip Torr, and Chang Xu. 2025. [Revisiting Uncertainty Estimation and Calibration of Large Language Models](#). In *Workshop on Scaling Environments for Agents*.
- Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. [Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5433–5442. Association for Computational Linguistics.
- Herbert Ullrich, Tomáš Mlynář, and Jan Drchal. 2025. [Claim Extraction for Fact-Checking: Data, Models, and Automated Metrics](#). *Preprint*, arXiv:2502.04955.
- Liam van der Poel, Ryan Cotterell, and Clara Meister. 2022. [Mutual information alleviates hallucinations in abstractive summarization](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5956–5965, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Artem Vazhentsev, Gleb Kuzmin, Akim Tsvigun, Alexander Panchenko, Maxim Panov, Mikhail Burtsev, and Artem Shelmanov. 2023. [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)*, pages 11659–11681, Toronto, Canada. Association for Computational Linguistics.
- Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. 2024a. [LLMs May Perform MCQA by Selecting the Least Incorrect Option](#). *Preprint*, arXiv:2402.01349.
- Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2024b. [SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models](#). In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 50622–50649. PMLR / OpenReview.net.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-Consistency Improves Chain of Thought Reasoning in Language Models](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Ziming Wang, Zeyu Shi, Haoyi Zhou, Shiqi Gao, Qingyun Sun, and Jianxin Li. 2025. [Towards Objective Fine-tuning: How LLMs' Prior Knowledge Causes Potential Poor Calibration?](#) In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 14830–14853. Association for Computational Linguistics.
- Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. [Crowdsourcing Multiple Choice Science Questions](#).

- In *Proceedings of the 3rd Workshop on Noisy User-generated Text, NUT@EMNLP 2017, Copenhagen, Denmark, September 7, 2017*, pages 94–106. Association for Computational Linguistics.
- Guoxuan Xia. 2026. [What \(and what not\) are calibrated probabilities actually useful for?](#) In *ICLR Blogposts 2026*. <https://iclr-blogposts.github.io/2026/blog/2026/useful-calibrated-uncertainties/>.
- Zeguan Xiao, Diyang Dou, Boya Xiong, Yun Chen, and Guanhua Chen. 2026. [Enhancing Uncertainty Estimation in LLMs with Expectation of Aggregated Internal Belief](#). In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pages 34043–34051. AAAI Press.
- Haoyi Xiong, Jiang Bian, Yuchen Li, Xuhong Li, Mengnan Du, Shuaiqiang Wang, Dawei Yin, and Sumi Helal. 2024a. When search engine services meet large language models: visions and challenges. *IEEE Transactions on Services Computing*.
- Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024b. [Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- KiYoon Yoo, Jangho Kim, Jiho Jang, and Nojun Kwak. 2022. [Detection of adversarial examples in text classification: Benchmark and baseline via robust density estimation](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3656–3672, Dublin, Ireland. Association for Computational Linguistics.
- Boxuan Zhang and Ruqi Zhang. 2025. [CoT-UQ: Improving Response-wise Uncertainty Quantification in LLMs with Chain-of-Thought](#). In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pages 26114–26133. Association for Computational Linguistics.
- Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024a. [LUQ: Long-text Uncertainty Quantification for LLMs](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5244–5262, Miami, Florida, USA. Association for Computational Linguistics.
- Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023a. [Enhancing Uncertainty-Based Hallucination Detection with Stronger Focus](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 915–932, Singapore. Association for Computational Linguistics.
- 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. 2023b. [Siren’s Song in the AI Ocean: A Survey on Hallucination in Large Language Models](#). *Preprint*, arXiv:2309.01219.
- Ziyin Zhang, Zhaokun Jiang, Lizhen Xu, Hongkun Hao, and Rui Wang. 2024b. [Multiple-Choice Questions are Efficient and Robust LLM Evaluators](#). *Preprint*, arXiv:2405.11966.
- Xinran Zhao, Hongming Zhang, Xiaoman Pan, Wenlin Yao, Dong Yu, Tongshuang Wu, and Jianshu Chen. 2024. [Fact-and-reflection \(far\) improves confidence calibration of large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8702–8718, Bangkok, Thailand. Association for Computational Linguistics.
- Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. [Large Language Models Are Not Robust Multiple Choice Selectors](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Ziang Zhou, Tianyuan Jin, Jieming Shi, and Li Qing. 2025. [SteerConf: Steering LLMs for Confidence Elicitation](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

A Appendix

A.1 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 (see Section A.2) includes: the complete `async-graph-bench` framework with extensible, replaceable computation nodes; final leaf-node outputs containing all confidence scores required to reproduce every plot and table in this paper; the implementation files of all four UQ methods, with hyperparameters visible in code (e.g., `models.py`, `apricot\_mc\_calc.py`, `ptrue.py`, `arithmetic\_calc.py`); a container definition file to build the execution environment; exact pip and driver versions pinned for deterministic reproduction; and installation and benchmark-execution instructions. All four UQ methods are evaluated on the identical cached generations, eliminating generation variance from cross-method comparisons. The main barrier to full reproduction is therefore the computational cost of generating the 685,000 responses, for which we release the resulting confidence scores instead.

A.2 Repository

The repository accompanying this paper is available online at <https://github.com/muelphil/llm-uncertainty-bench>. It contains the framework instance used for this benchmark, the computed UQ scores (final leaf-node outputs), method implementations, the container definition, and pinned dependency versions, as detailed in the Reproducibility Statement. The underlying benchmarking framework, `async-graph-bench`, is developed and maintained separately at <https://github.com/Helmholtz-AI-Matter/async-graph-bench>.

A.3 Information About the Use of AI Assistants

AI assistants were used to support text refinement, including improvements to wording, phrasing, and overall readability of the manuscript. AI-based tools were also used for coding assistance during implementation and experimentation. All generated content, code, and suggestions were reviewed and verified by the authors before inclusion in this work.

A.4 Model Selection

An overview of the models used in the experiments is given in Table A.1; primary technical references per model family are listed in Table A.2.

Please note that the model *Magistral-Small-2507* requires using a specific system prompt to enable reasoning¹. We have tested the model without a system prompt (acting as an instruction tuned model) and with the system prompt provided, acting as a reasoning model. To highlight whether the reasoning behaviour was enabled through system prompt usage, we added a "Reasoning-Enabled" postfix to the model name in tables and plots.

A.4.1 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 answer². Without the system prompt, the model will behave like an instruction-tuned model. For the label-probability experiments,

¹<https://huggingface.co/mistralai/Magistral-Small-2507>

²see <https://huggingface.co/mistralai/Magistral-Small-2506>

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

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.

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 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 cali-

bration 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 inaccessible 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.5 Dataset Selection

The datasets used in the experiments are listed in Table A.3. 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.6 Prompt Composition

The second experiment evaluates a total of 685,000 long-form responses across models and datasets using sequence-level UQ methods. We include five multiple-choice QA datasets (MMLU, ARC-Easy, ARC-Reasoning, SciQ, GPQA) and three arithmetic QA datasets (GSM8K, SVAMP, SciBench).

For the MCQA datasets, we apply counterfactual prompting using the APriCoT strategy, evaluat-

Provider	Model Family	Reference
OpenAI	GPT-OSS	(OpenAI, 2025)
Mistral AI	NeMo	(Mistral AI, 2024a)
	Ministral	(Mistral AI, 2024b)
	Mistral Small 3	(Mistral AI, 2025b)
	Magistral	(Mistral AI, 2025a)
Meta	LLaMA 3	(Meta AI, 2024)
	LLaMA 4	(Meta AI, 2025)
Qwen	Qwen3	(Qwen Team, 2025)
DeepSeek AI	DeepSeek-R1	(DeepSeek-AI, 2025)
Google DeepMind	Gemma 3	(Gemma Team, 2025)

Table A.2: **Model families used in this study with primary technical references.** We group base, instruction-tuned, and reasoning variants under their respective model families.

Dataset	Size	Task Format	License	Domain
MMLU	15,908	Multiple Choice	MIT	General (57 topics, including Physics)
ARC	7,787	Multiple Choice	CC-BY-NC 4.0	Science (Physics, Chemistry, Biology, Earth Science)
SciQ	13,679	Multiple Choice	CC-BY-SA 3.0	Science (Physics, Chemistry, Biology)
GPQA	448	Multiple Choice	MIT	Science (Physics, Chemistry, Biology) – Graduate-level
GSM8K	8,792	Arithmetic	MIT	Mathematics
GSM-MC	8,787	Multiple Choice	MIT	Mathematics
SVAMP	1,000	Arithmetic	CC-BY 4.0	Mathematics
SciBench	2,229	Arithmetic	CC-BY-SA 4.0	Science (Physics, Chemistry, Biology, Medicine, Earth Science)

Table A.3: **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.

ing each of the four answer options independently. This expands each item into four separate prompts. All datasets are subsampled to 250 items, and we sample 10 responses per prompt. Consequently, each arithmetic dataset yields 2,500 responses per model, and each MCQA dataset yields 10,000 responses per model, totaling 57,500 responses per model.

Due to high inference costs, we reduced the number of generated responses specifically for the *deepseek-ai/DeepSeek-R1-Distill-Llama-70B* model on the GPQA dataset by half, resulting in 5,000 evaluated responses for this model–dataset pair.

Across all 12 evaluated models, this results in a total of 685,000 responses, each assessed indi-

vidually using the four selected UQ methods in Experiment 2. The authoritative per-model, per-dataset accounting is provided in Table A.4.

A.7 Relevance of the Selected Methods to the Field of UQ in LLMs

Obtained from Google Scholar at the time of writing (see Table A.5), Verbalized Uncertainty (as proposed by (Tian et al., 2023)) has been cited 602 times, P(True) (as proposed by (Kadavath et al., 2022)) has been cited 723 times and CCP (as proposed by (Fadeeva et al., 2024)) has been cited 118 times, highlighting their relevance to the field. While Answer Consistency methods have not been directly proposed, it captures semantic consistency and proxies other approaches like self-consistency prompting as detailed in (Wang et al., 2023), which

Model	MMLU	ARC-Easy	ARC-Reasoning	SciQ	GPQA	GSM8K	SVAMP	SciBench	Total
gpt-oss-20b	10k	10k	10k	10k	10k	2.5k	2.5k	2.5k	57.5k
gpt-oss-120b	10k	10k	10k	10k	10k	2.5k	2.5k	2.5k	57.5k
Ministral-8B-Instruct-2410	10k	10k	10k	10k	10k	2.5k	2.5k	2.5k	57.5k
Mistral-Nemo-7B	10k	10k	10k	10k	10k	2.5k	2.5k	2.5k	57.5k
Mistral-Small-3.2-24B-Instruct-2506	10k	10k	10k	10k	10k	2.5k	2.5k	2.5k	57.5k
Llama-3.3-70B-Instruct	10k	10k	10k	10k	10k	2.5k	2.5k	2.5k	57.5k
Llama-4-Scout-17B-16E-Instruct	10k	10k	10k	10k	10k	2.5k	2.5k	2.5k	57.5k
gemma-3-27b-it	10k	10k	10k	10k	10k	2.5k	2.5k	2.5k	57.5k
Qwen3-30B-A3B-Instruct-2507	10k	10k	10k	10k	10k	2.5k	2.5k	2.5k	57.5k
Qwen3-30B-A3B-Thinking-2507	10k	10k	10k	10k	10k	2.5k	2.5k	2.5k	57.5k
DeepSeek-R1-Distill-Qwen-32B	10k	10k	10k	10k	10k	2.5k	2.5k	2.5k	57.5k
DeepSeek-R1-Distill-Llama-70B	10k	10k	10k	10k	5k	2.5k	2.5k	2.5k	52.5k
Total	120k	120k	120k	120k	115k	30k	30k	30k	685k

Table A.4: **Response counts for Experiment 2 per model and dataset.** Values are in thousands of responses (k). Each MCQA dataset contributes 10,000 responses per model (250 items $\times$ 4 APriCoT option prompts $\times$ 10 generations) and each arithmetic dataset 2,500 responses per model (250 items $\times$ 10 generations), yielding 57,500 responses per model. The sole exception is *DeepSeek-R1-Distill-Llama-70B* on GPQA, for which 5,000 responses were generated due to high inference costs. This table is the authoritative accounting of the 685,000 responses evaluated in Experiment 2.

Paper	Citation Count
Verbalized Uncertainty (Tian et al., 2023)	602
P(True) (Kadavath et al., 2022)	723
CCP (Fadeeva et al., 2024)	118
Semantic consistency (Wang et al. (2023), basis of Answer Consistency)	3831

Table A.5: **Citation Counts of employed UQ methods (as of writing).**

has been cited 3831 times. We found different citation counts on OpenAlex³ for entries in Table A.5 but their relative scale is equivalent than reported here.

To underpin the relevance of the four selected UQ methods, a glimpse at the derivative work is instructive. For Verbalized Uncertainty (Tian et al., 2023), the following papers have built on this idea: Xiong et al. (2024b); Zhao et al. (2024); Zhou et al. (2025); Shrivastava et al. (2023); Han et al. (2024). The idea of P(True) was picked up by Mahaut et al. (2024); Chen and Mueller (2024); Ren et al. (2023b); Zhang and Zhang (2025); Xiao et al. (2026); Heo et al. (2025); Kapoor et al. (2024). This demonstrates how essential a rigorous study of the effectiveness and performance of these methods is.

Beyond the concrete P(True) method itself, the

³openalex.org

broader problem of whether token probabilities constitute reliable confidence estimates is highly relevant to LLM-as-a-Judge systems and classifier-style LLM applications in general. Many such systems implicitly rely on token probabilities as confidence signals for labels or decisions, making their calibration directly relevant for the reliability of automated evaluation and downstream decision-making pipelines.

A.8 Uncertainty Quantification Method Exclusion

Our selection of UQ methods reflects principled constraints: feasibility, interpretability, and applicability to realistic scientific reasoning tasks. An overview over exclusion of methods and short form explanation is provided in Table A.6. Only methods producing normalized sequence-level uncertainty scores are included in our analysis to enable relia-

bility 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), *Answer Consistency* (Wang et al., 2023), *Claim-Conditioned Probability (CCP)* (Fadeeva et al., 2024). More details of these methods are discussed in Section A.11.3.

A.8.1 Survey of UQ methods

We used the taxonomy from LM-Polygraph Fadeeva et al. (2023) as a starting point and adapted the categorization of UQ methods into information-based, ensemble-based, density-based, reflexive or verbalized and meaning-diversity approaches. However, it is important to emphasize that LM-Polygraph targets selective generation, whereas our work focuses on calibration, the second core UQ task. Calibration requires normalized confidence scores in $[0,1]$, which immediately excludes many of the 28 methods surveyed in LM-Polygraph. Thus, while we relied on Fadeeva et al. (2023) for a broad overview of UQ approaches, the spectrum of methods applicable to calibration is necessarily narrower.

A.8.2 Density-Based Methods

Density-based methods require access to the model’s training data to estimate likelihoods or information-theoretic quantities. In our benchmark, the evaluated models are open-weight but not open-source, and their training corpora are unavailable. Consequently, density-based approaches cannot be applied, making them unsuitable for our calibration-focused evaluation.

A.8.3 Ensemble-Based Methods

Ensemble methods combine multiple distinct models to estimate uncertainty. While potentially effective, they prevent the assessment of per-model calibration, which is central to our study. Additionally, ensembles introduce substantial design ambiguity regarding which combinations of models to use and incur significant computational costs. In contrast, the Answer Consistency method we benchmark samples a single model multiple times, preserving per-model interpretability while allowing assessment of semantic consensus as an UQ signal.

A.8.4 Claim-Level Methods

Semantic Entropy as a claim-level method was excluded due to methodological and computational limitations. Existing evaluations of this method

are restricted to short-form factual QA datasets and do not assess calibration, offering limited evidence for their suitability in reasoning-intensive scientific QA. Furthermore, scientific answers frequently span hundreds or thousands of tokens. In our internal tests, claim-extraction prompts often produced tens to hundreds of trivial or irrelevant claims, each requiring separate LLM evaluation. This leads to computational costs far exceeding those of the Answer Consistency method, our computationally most expensive method employed, making claim-level approaches impractical for large-scale scientific QA.

A.8.5 Overview Over Excluded Methods

The full list of excluded methods with their exclusion reasons is provided in Table A.6.

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., 2023)	Information-based	Produces unnormalized scores
Pointwise mutual information (PMI) (Takayama and 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
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., 2023)	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 and Gales, 2021)	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 and Gales, 2021)	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., 2023a)	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., 2024)	Meaning Diversity	Produces count-based outputs that are not normalized to [0,1]
Sum of eigenvalues of the graph Laplacian (EigV) (Lin et al., 2024)	Meaning Diversity	Produces unnormalized scores
Degree matrix (Deg) (Lin et al., 2024)	Meaning Diversity	Produces unnormalized scores
Eccentricity (Ecc) (Lin et al., 2024)	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

Table A.6: **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. (2023).

A.9 Label Probability Calibration Experiment Design

For the Label Probability Calibration Experiment, preliminary tests revealed that base models struggled with task comprehension: in many cases, none of the valid label tokens (A/B/C/D) appeared among the top-20 first-token predictions. To address this, we designed four prompt variants with different structural formats to identify the formulation that maximizes task comprehension. We define task comprehension as the average total probability mass assigned to the label tokens (A/B/C/D) in the model’s first generated token.

The four prompt designs are presented in Section A.9.1. An overview of task comprehension scores for each prompt is provided in Section A.9.2, where Prompt 1 emerges as the most effective. Nevertheless, even with improved task comprehension, a substantial fraction of responses remained invalid (i.e. no extractable probability assigned to any label), as detailed in Section A.9.3.

We analyze the causes of these invalid responses in Section A.9.4 and outline the mitigation strategies applied to reduce them in Section A.9.5. Finally, we report the updated invalid-answer statistics for Prompt 1 after applying these adjustments in Section A.9.3. We report the distribution of token counts for the reasoning chains for reasoning models in Section A.9.7.

A.9.1 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 introductory 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

C) <ANSWER CHOICE C>
D) <ANSWER CHOICE D>

<ANSWER>

A.9.2 Task Comprehension per Prompt Design

A.9.3 Invalid Answer Count

Although we selected the prompt that yielded the highest task comprehension for our benchmark, the original experiment exhibited a substantial number of invalid answers, varying with prompt design. These invalid outputs contaminated the results. After performing analysis on the cause of this behaviour and employing mitigation strategies, we were able to significantly decrease the amount of invalid answers, as detailed in Table A.8.

A.9.4 Analysis of Invalid Answers

Our investigation identified two primary causes of invalid responses:

1. **Missing answer labels within top token alternatives:** Models sometimes did not generate any of the expected labels (A/B/C/D) within the top 20 token alternatives that vLLM provides per token. This occurred in:

- (a) *Base models*, which lacked sufficient task comprehension and failed to include any of the answer labels in the top 20 token alternatives.
- (b) *Reasoning models*, which faced the additional challenge of adhering to the expected output format after generating long-form reasoning, as the attention mechanism favours more recent tokens. Reasoning of several thousand tokens appeared to negatively effect task comprehension, leading to invalid answers in extreme cases.

2. **Overflowing generations:** Reasoning models produced outputs exceeding the originally set maximum token limit (4096 tokens), appearing to be stuck in prolonged reasoning cycles without converging to a final answer.

A special case is *Magistral-Small-2507*, which has reasoning enabled via the system prompt provided by MistralAI⁴. For this model, we consistently observed non-converging generations that

⁴<https://huggingface.co/mistralai/Magistral-Small-2506>

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>

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

Table A.7: **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	Prompt 1	Prompt 2	Prompt 3	Prompt 4	Prompt 1 (SD)
gpt-oss-20b	307	3356	498	25304	395
gpt-oss-120b	38	3323	27	24873	0
Ministral-8B-Instruct-2410	0	0	0	0	0
Mistral-Nemo-Base-2407	0	0	12574	0	0
Mistral-Nemo-Instruct-2407	0	0	0	0	0
Mistral-Small-3.1-24B-Base-2503	0	0	0	0	0
Mistral-Small-3.2-24B-Instruct-2506	0	0	0	0	0
Magistral-Small-2507	0	0	0	0	0
Magistral-Small-2507-Reasoning-Enabled	8604	18716	4749	21232	123
Llama-3.1-70B	0	13	0	6636	0
Llama-3.3-70B-Instruct	0	2527	19	125	0
Llama-4-Scout-17B-16E	2	3	0	5048	0
Llama-4-Scout-17B-16E-Instruct	0	15	0	0	0
Qwen3-30B-A3B-Base	2	1	0	55	0
Qwen3-30B-A3B-Instruct-2507	0	6270	0	722	0
Qwen3-30B-A3B-Thinking-2507	6242	11746	8656	21840	58
DeepSeek-R1-Distill-Llama-70B	425	7538	725	7477	148
DeepSeek-R1-Distill-Qwen-32B	914	7465	1170	3337	243
gemma-3-27b-pt	25316	25316	25316	25316	0
gemma-3-27b-it	0	431	0	2	0

Table A.8: **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. This phenomenon is discussed above. Using the prompt design yielding the best task comprehension (as discussed in Section A.9.2), we adjusted the experiment to use structured decoding and increased the maximum token generation limit. After this adjustment, invalid answers may only occur when reasoning chains exceed 10240 tokens. The result is shown in column "Prompt 1 (SD)", drastically decreasing the amount of invalid answers.

produced a very high number of invalid answers, even on the relatively simple MMLU dataset compared to GPQA. Increasing the maximum token generation did not resolve the issue – the model still consistently hit the upper generation limit. Our solution for handling these invalid outputs is detailed in Section A.9.6.

A.9.5 Mitigation Strategies

We addressed these issues as follows:

- Leveraged vLLM’s structured decoding feature to enforce compliance with the expected output format after reasoning using the regex "[ABCD]".
- Increased the maximum token limit for rea-

soning from 4096 to 10240 tokens, a $2.5\times$ increase.

- Adjusted the generation of *Magistral* to handle infinite generations, as detailed in Section A.9.6.

While structured decoding enforces compliance with the expected output format, the prior evaluation of task comprehension per prompt remains essential. Forcing task compliance without adequate comprehension would result in outputs that, although syntactically valid, are semantically incorrect or inconsistent with the question, thereby undermining the reliability of uncertainty quantification and downstream analyses.

Based on the prior evaluation, we chose Prompt 1 for the updated experiment and discarded the other Prompt Designs.

A.9.6 Magistral-Small-2507 Invalid Answer Solution

To evaluate *Magistral-Small-2507* on label probability calibration and score distribution despite persistent infinite-generation issues, we first set the maximum token limit for reasoning chain generation to 4096 tokens. Generated outputs almost always contained unfinished reasoning chains, indicated by a missing end-of-reasoning token "[/THINK]". We removed "[/THINK]" if present, trimmed any trailing incomplete sentence, and appended ". I should now respond with the single label A, B, C, or D associated with the answer I consider most correct.[/THINK]" to artificially terminate the chain and prompt a single-label response. The model was then queried again with this reasoning chain as a prefix, and label probabilities were extracted from the next generated token. Although structured decoding cannot be applied in this prefix-prompting setup, this procedure reduced invalid answers from 8604 to 123, as shown in Table A.8.

A.9.7 Reasoning Token Count Distribution for Reasoning Models

The distribution of reasoning token counts per model and dataset is shown in Figure A.1.

A.10 Label Probability Experiment Results

A.10.1 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` for the initial prompt designs without structured decoding and at `label_prob_calibration/resources_struct_decoding/figures/full_plots` for the chosen Prompt 1 and structured decoding enabled. Each file follows the naming convention `cal_plot_prompt<id>_table<t>_chosenonly<c>_norm<n>.svg`, where the placeholders encode the following settings:

- `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 (Figures A.2 to A.4). The plots are included to visualize systematic patterns in the calibration plots induced by fine-tuning. Full size plots including summary statistics per plot for individual plot assessment can be found in the repository.

- `prompt`: Identifier of the prompt design used to generate the plot (see Section A.9.1).

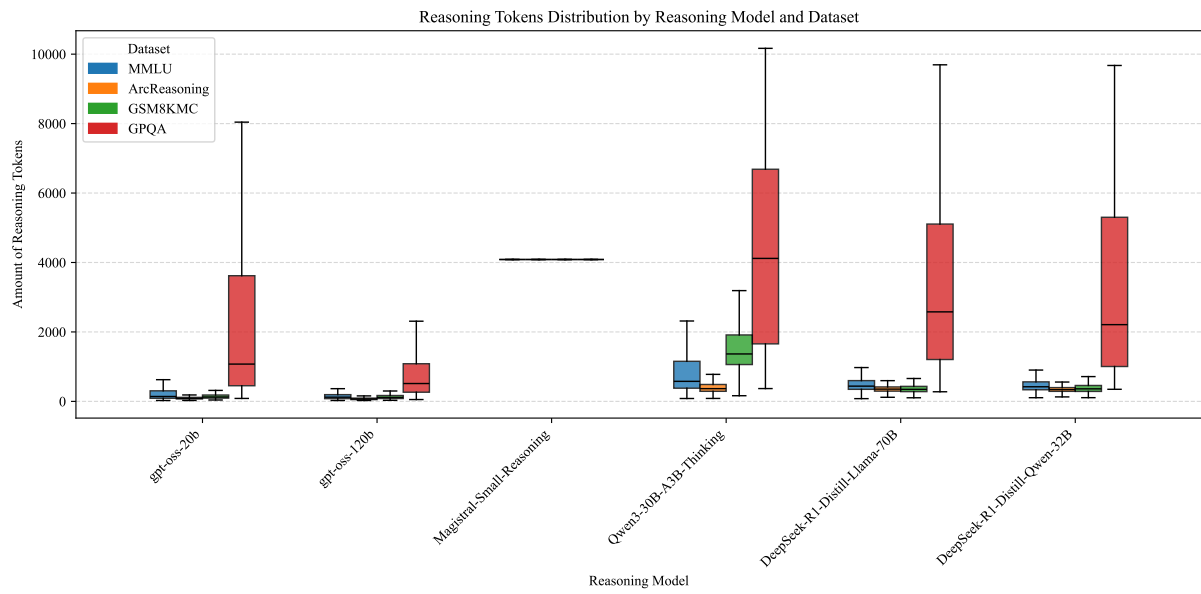


Figure A.1: Distribution of Reasoning Tokens Produced by Models. The figure displays a grouped boxplot illustrating the distribution of the Amount of Reasoning Tokens generated by different Reasoning Models across the individual Datasets. Models are grouped along the X-axis, with each boxplot representing the token count distribution for a specific dataset within that model. The central line in each box represents the median, and the box edges represent the first and third quartiles.

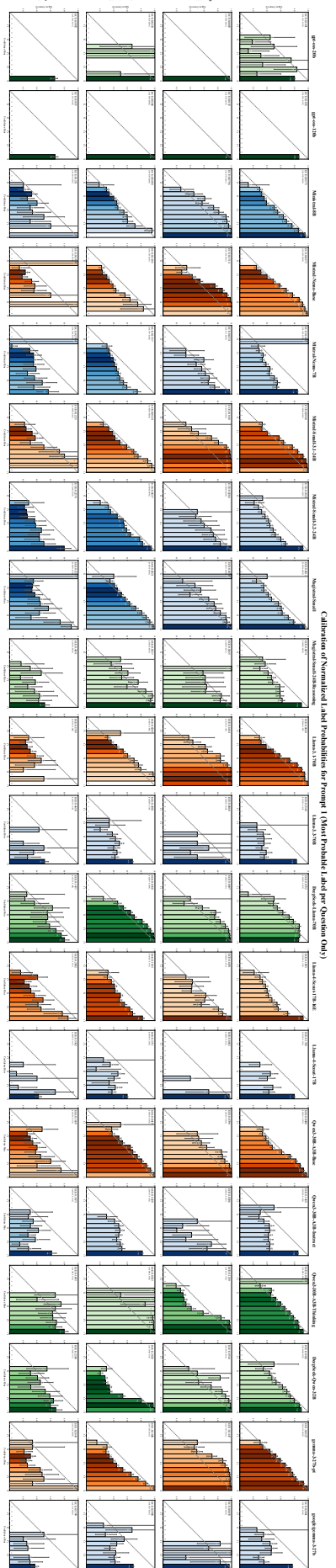


Figure A.2: **Calibration Plots for Prompt 1 with structured decoding enabled, using only the most probable label.** The columns represent different models, while the rows represent datasets. Tables below the plots list summary metrics such as ECE, normalized entropy of bucket counts and AUROC. Base 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.

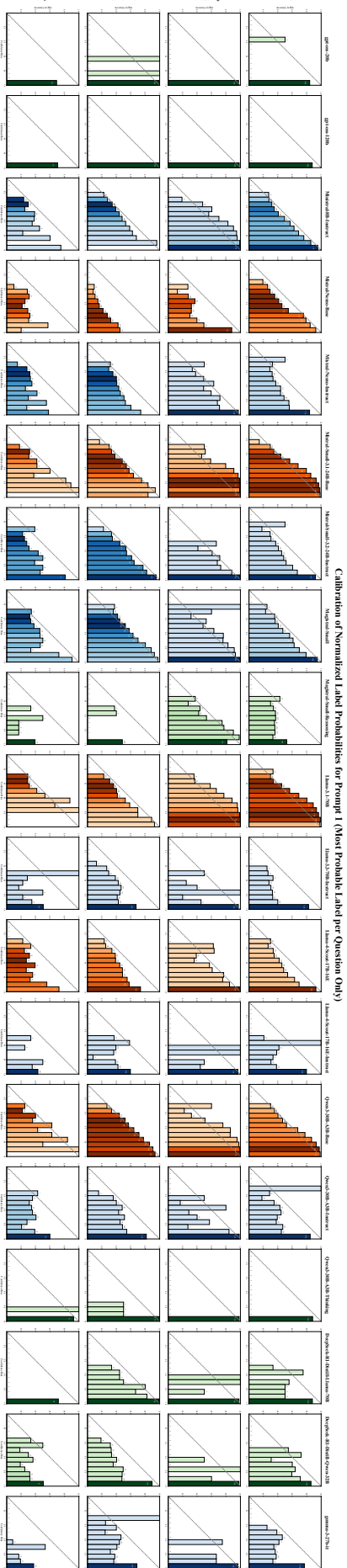
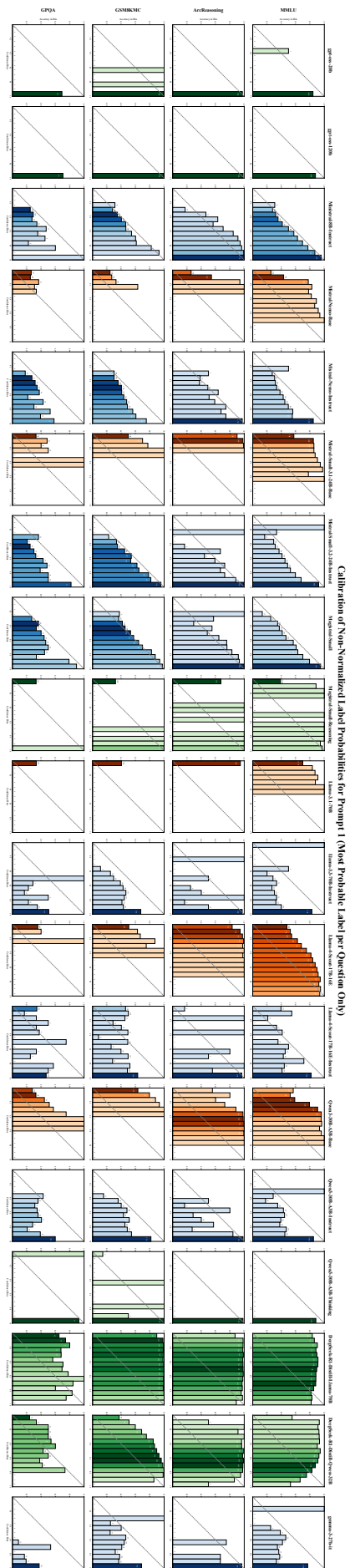


Figure A.3: **Calibration Plots for Prompt 1, with structured decoding disabled, using normalized label probabilities and only the most probable label.** The columns represent different models, while the rows represent datasets. Tables below the plots list summary metrics such as ECE, normalized entropy 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.



A.10.2 Effect of Normalization

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.5).

A.10.3 Selection Bias

Figure A.6 shows the label probabilities aggregated across all datasets for each model under Prompt 1.

A.10.4 Effect of Finetuning on Uncertainty Score Distributions

The distribution of normalized certainty scores across model variants and datasets is shown in Figure A.7.

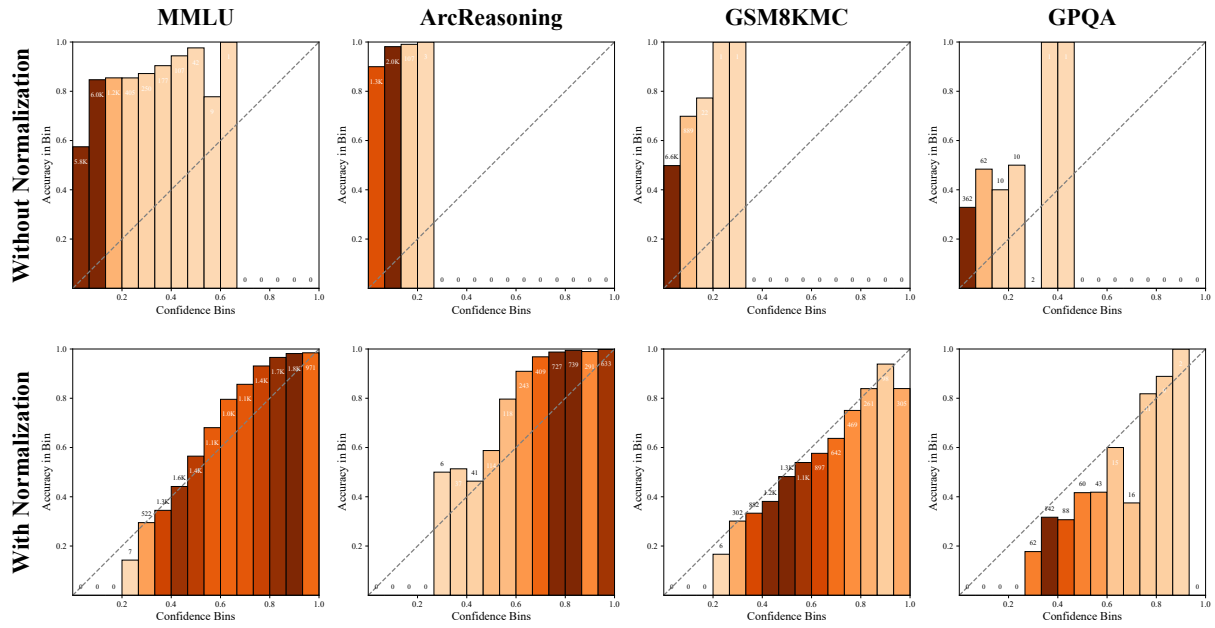


Figure A.5: **Representative Comparison of Calibration Plots for Unnormalized and Normalized Label Probabilities for the Model *Mistral-Small-3.1-24B-Base-2503* and Prompt 1 across all datasets used.** Calibration improves significantly after normalizing label probabilities.

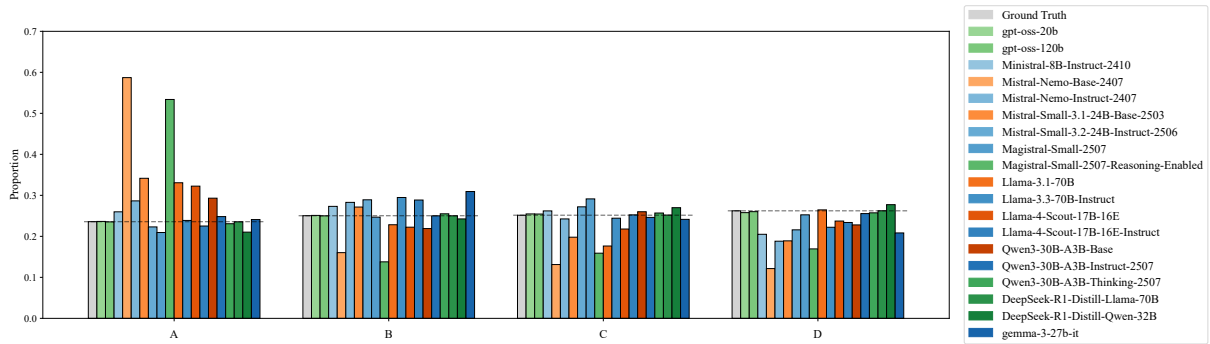


Figure A.6: **Probabilities for the Labels summed across all Datasets for each individual Model using Prompt 1.** The ground truth, represented by the distribution of the labels of the correct answers across all items in the datasets, is visualized by the grey bars and the dashed baselines.

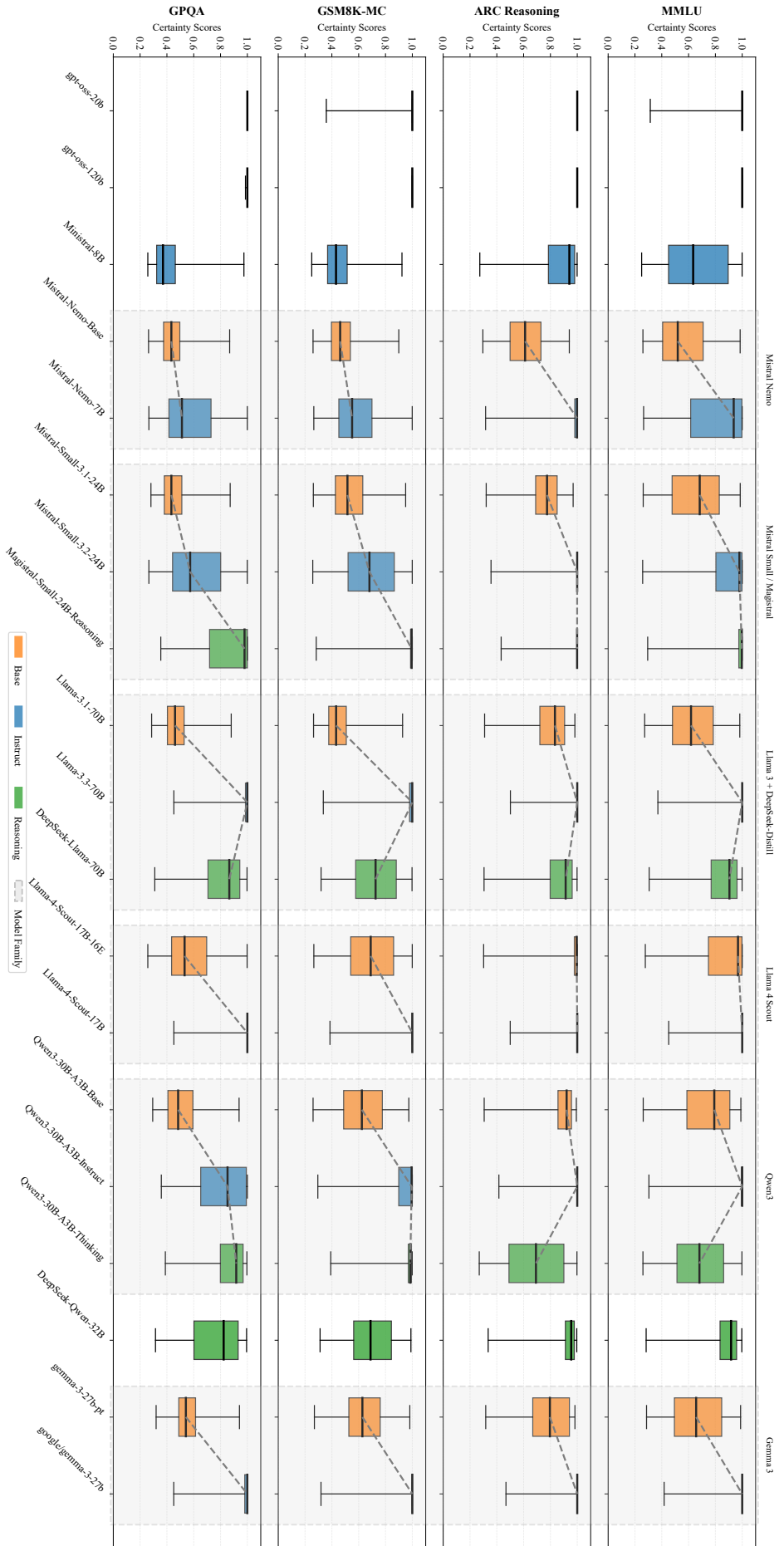


Figure A.7: **Distribution of normalized certainty scores (chosen option only) across model variants and evaluation datasets.** Rows correspond to datasets and columns to models. Box colors encode model variant types: orange = base models, blue = instruction-tuned models, and green = reasoning models. Light gray dashed rectangles group related models belonging to the same model family across all dataset rows. Within each family, dashed gray lines connect the medians of successive variants (base → instruct → reasoning), highlighting shifts in certainty distributions induced by fine-tuning. Family names are shown above the grouped columns.

A.10.5 Task Comprehension per Prompt Design

Task comprehension scores per prompt design and model are shown in Figure A.8.

A.11 Sequence-Level Calibration

A.11.1 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.11.2 Formal Definitions of Answer Consistency

We provide the formal definitions of the Answer Consistency score for both task types. In both cases, the confidence score for a given sampled generation is computed exclusively from the agreement among the sampled generations; ground-truth (gold) labels are not consulted at any point of this computation. Gold labels are introduced only afterwards, solely to determine whether each model judgment was correct and to compute calibration metrics (Section 3.7).

Multiple-Choice QA (MCQA). Under APriCoT prompting, each MCQA item is evaluated once per answer option. For a given item and answer option o , we draw $n = 10$ independent generations and obtain a binary judgment $y_i \in \{0, 1\}$ for each sampled generation $i \in \{1, \dots, n\}$, where $y_i = 1$ iff generation i concludes that option o is the correct answer. The judgment is extracted from the generation by a follow-up prompt and contains no reference to the gold label. The Answer Consistency score of generation i is the empirical frequency with which the remaining generations agree with its judgment:

$$AC_i = \frac{1}{n-1} \sum_{j \neq i} \mathbb{I}[y_j = y_i]. \quad (\text{A.1})$$

Thus the score is determined entirely by the agreement structure of the sampled judgments $\{y_1, \dots, y_n\}$. The gold label is used only post hoc, to establish the correctness of each judgment y_i for the computation of calibration metrics.

Arithmetic QA. For each arithmetic item, we draw $n = 10$ independent generations and extract the final numeric answer $\hat{a}_i$ from generation i via a dedicated extraction prompt (for SciBench, semantic equivalence is established by an LLM-based clustering step; Section A.11.3). Two generations are considered to agree when their extracted answers are semantically equivalent, i.e., strictly numerically equal (with unit-aware equivalence for SciBench). The Answer Consistency score of generation i is

$$AC_i = \frac{1}{n-1} \sum_{j \neq i} \mathbb{I}[\hat{a}_j \equiv \hat{a}_i], \quad (\text{A.2})$$

where $\equiv$ denotes semantic equivalence of the extracted answers. As for MCQA, the score depends only on the sampled generations; the gold answer is introduced only during evaluation to assess correctness.

Discrete score values. With $n = 10$ samples, both scores take values in $\{0.1, 0.2, \dots, 1.0\}$ (and 0.0 for generations that fail to yield a parseable answer in arithmetic QA). Calibration plots and summary statistics for Answer Consistency accordingly use 11 bins centered on these discrete values (Section A.11.3).

A.11.3 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`.

Answer Consistency For this metric, 10 samples were generated per prompt. The *Answer Consis-*

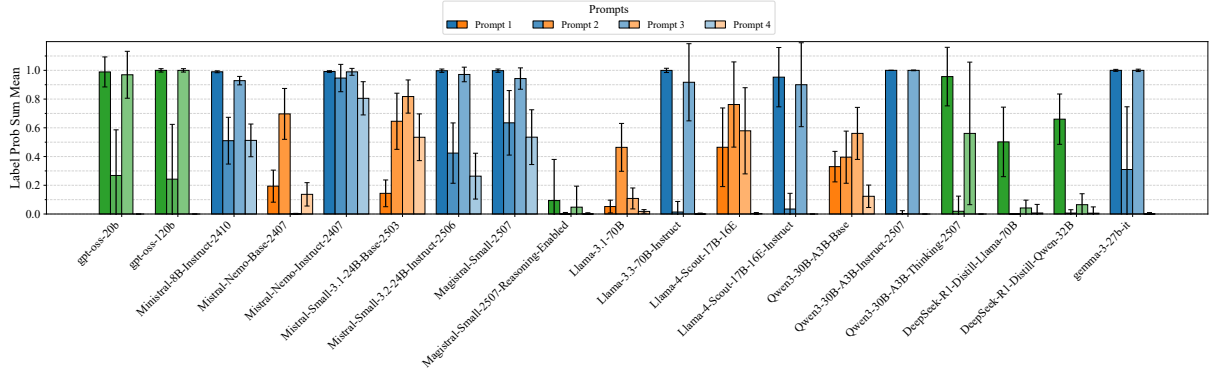


Figure A.8: **Task Comprehension per Prompt Design.** Mean over Sum of Label Probabilities for the Different Prompt Designs and Models. The data has been aggregated across all datasets, spanning 25,316 items per model and prompt design. Base models are shown in orange, instruction-tuned models in blue, and reasoning models in green.

tency 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 Answer Consistency 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 Answer Consistency 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., 2023) and was optimized for improved computational efficiency.

A.12 Empirical Validation of LLM-as-a-Judge Stability via Model Ensembles

We acknowledge that the use of LLMs as judges for verifying answer option correctness requires empirical validation to ensure that our evaluation pipeline is not sensitive to the choice of judge model. Motivated by the ensemble-based evaluation strategy of [Ielanskyi et al. \(2025\)](#), we conduct an additional robustness experiment in which we replace a single judge with an ensemble of three diverse models: Ministral-8B-Instruct, Llama-3.3-70B-Instruct, and Qwen3-30B-A3B-Instruct. Using this setup, we re-evaluate a fixed set of 10,000 cached responses generated by Llama-3.3-70B on MMLU, ensuring that all judge models operate on identical inputs and evaluation prompts.

Across the ensemble, we observe strong agreement on correctness judgments, with consensus in 9,774 out of 10,000 cases (97.74%). This high level of inter-model agreement suggests that the correctness signal used in our benchmark is stable with respect to reasonable variations in judge model selection. Importantly, the small disagreement rate is not systematic and does not indicate any directional bias that would materially affect aggregated results or the comparative conclusions of our study. We therefore conclude that adopting an ensemble-based judging scheme would not meaningfully alter the reported findings.

A.12.1 Response Lengths

Mean response token lengths across models and datasets are reported in [Table A.9](#).

A.12.2 Extensive Calibration Plots

In the following, extensive calibration plots for the evaluation of sequence level uncertainty methods are provided ([Figures A.9 to A.12](#)). 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.

These plots are included in the appendix mainly to visualize systematic patterns in the calibration plots induced by the UQ methods. Full size plots for individual plot assessment are available in the repository.

Model	MMLU	ARC Easy	ARC Challenge	SciQ	GPQA	GSM8K	SVamp	SciBench
gpt-oss-20b	1046 ± 431	844 ± 283	851 ± 305	838 ± 273	2838 ± 2274	475 ± 433	309 ± 287	2253 ± 2270
gpt-oss-120b	872 ± 323	695 ± 305	676 ± 313	641 ± 257	1754 ± 1226	347 ± 205	239 ± 182	1320 ± 964
Ministral-8B	433 ± 173	323 ± 102	340 ± 111	303 ± 98	632 ± 196	225 ± 91	127 ± 128	713 ± 234
Mistral-Nemo-7B	346 ± 148	279 ± 78	285 ± 84	248 ± 72	515 ± 202	226 ± 84	134 ± 106	533 ± 281
Mistral-Small-3.2-24B	681 ± 444	471 ± 130	498 ± 169	458 ± 131	1337 ± 904	435 ± 352	284 ± 276	1330 ± 1033
Llama-3.3-70B	484 ± 324	365 ± 83	380 ± 93	357 ± 77	949 ± 536	244 ± 74	198 ± 291	765 ± 384
DeepSeek-Llama-70B	1290 ± 1356	583 ± 253	622 ± 267	548 ± 193	4300 ± 2928	597 ± 537	331 ± 207	4658 ± 3398
Llama-4-Scout-17B	453 ± 144	397 ± 88	413 ± 96	385 ± 78	749 ± 233	231 ± 69	202 ± 269	771 ± 383
Qwen3-30B-A3B-Instruct	675 ± 432	456 ± 140	492 ± 168	456 ± 197	2824 ± 1302	305 ± 303	311 ± 591	2139 ± 1484
Qwen3-30B-A3B-Thinking	1887 ± 1202	1485 ± 632	1558 ± 692	1496 ± 692	6827 ± 3213	1095 ± 701	801 ± 695	6230 ± 3252
DeepSeek-Qwen-32B	1457 ± 1605	649 ± 274	691 ± 310	625 ± 253	5251 ± 3322	537 ± 439	313 ± 120	5966 ± 3267
google/gemma-3-27b	426 ± 184	287 ± 106	299 ± 107	280 ± 100	856 ± 345	248 ± 157	180 ± 187	951 ± 523
Averaged Mean (Instruct)	500	368	387	355	1123	273	205	1029
Averaged Mean (Reasoning)	1310	851	880	830	4194	610	398	4086
Averaged Mean (Total)	838	569	592	553	2403	414	286	2302

Table A.9: **Mean response token lengths across models and datasets.** Each cell reports the mean generated response length in tokens (rounded to integers), with the standard deviation shown beneath in smaller gray text. Rows correspond to individual models and columns to evaluation datasets. The final three rows report dataset-wise averages of mean response lengths over instruction-tuned models, reasoning models, and all models, respectively. These aggregates provide a summary of how dataset design, including differences in reasoning demand, affects response length.

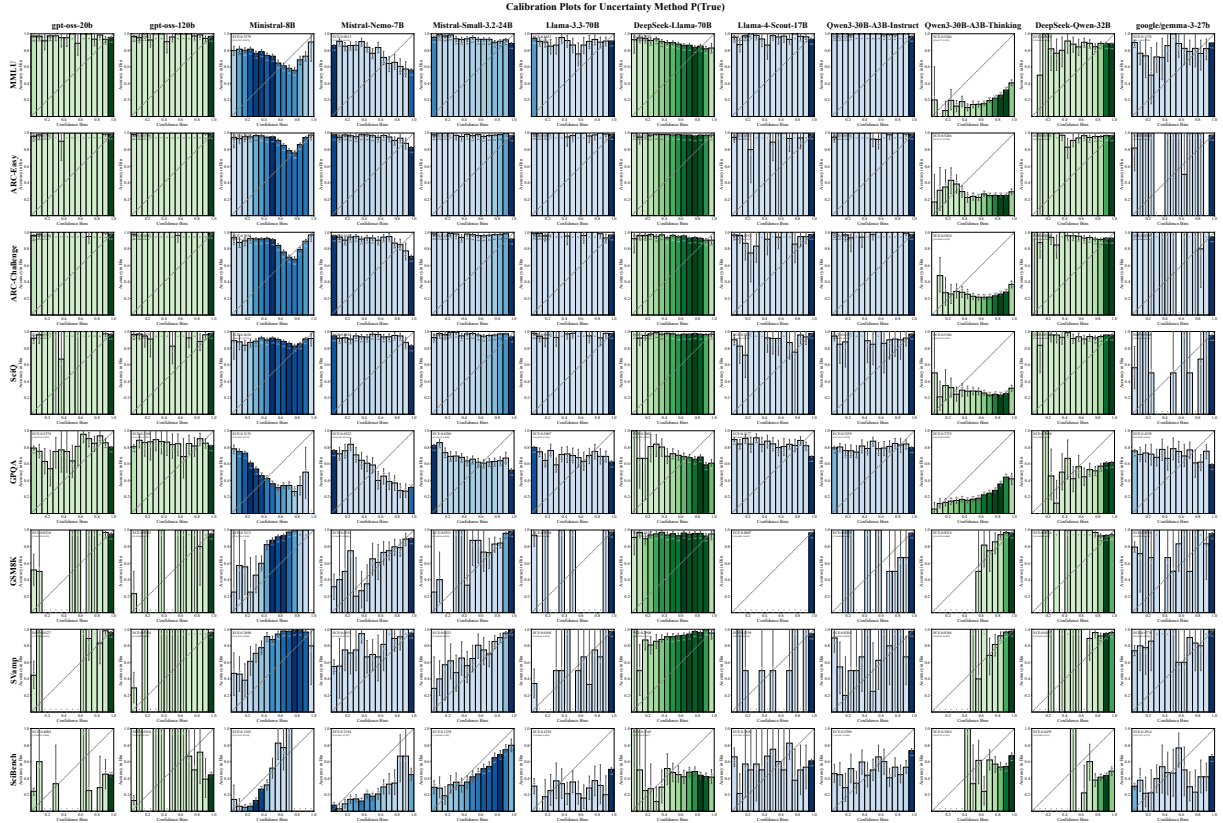


Figure A.9: **Calibration Plots for P(True)**. Columns correspond to the models and rows to datasets.

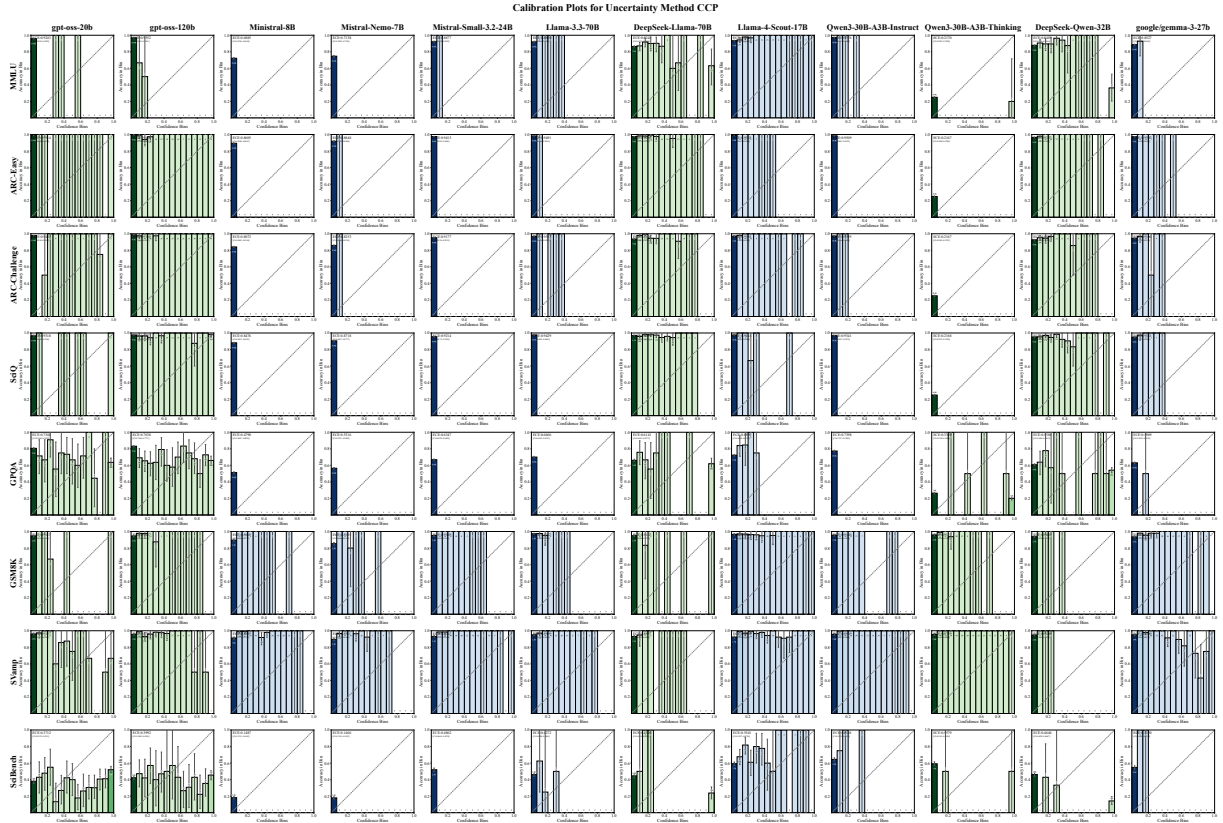


Figure A.10: **Calibration Plots for CCP**. Columns correspond to the models and rows to datasets.

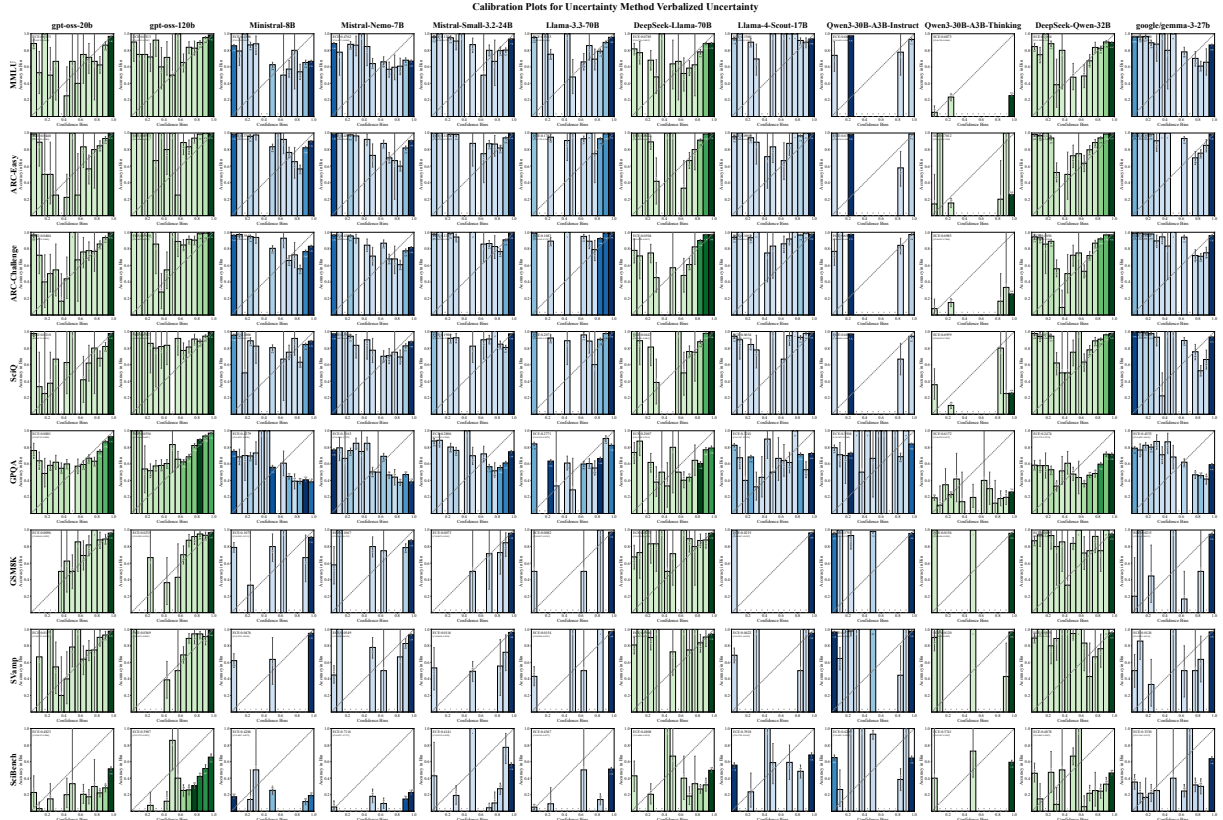


Figure A.11: **Calibration Plots for Verbalized Uncertainty.** Columns correspond to the models and rows to datasets.

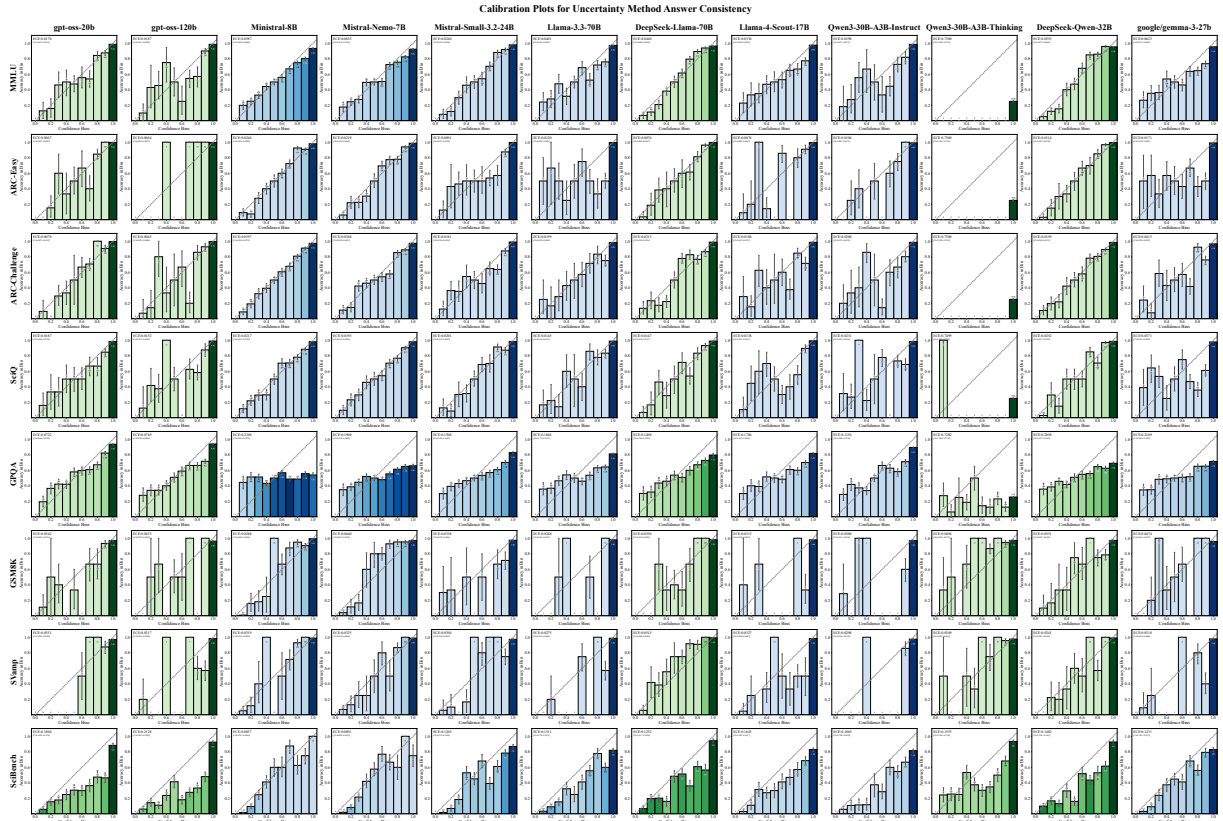


Figure A.12: **Calibration Plots for Answer Consistency.** Columns correspond to the models and rows to datasets.

A.12.3 Calibration Correlation by UQ Method

Figure 4 shows the per-method distribution of Pearson calibration correlation across all (model, dataset) pairs.

A.12.4 Verbalized Uncertainty Confidence Score Distribution

The distribution of confidence scores produced by the Verbalized Uncertainty method across all models is shown in Figure A.16.

A.12.5 P(True) Confidence Score Distribution

The distribution of confidence scores assigned by P(True) across all models is shown in Figure A.17.

A.13 Effect of Temperature

The following sections clarify why additional temperature-sweep experiments would not meaningfully affect our findings or improve the interpretability of the results.

A.13.1 Experiment 1: Temperature in Token-Level Calibration

Experiment 1 evaluates calibration using the label probability of the first generated token under a fixed temperature of 1.0 (reasoning traces, when used, are decoded greedily at the same temperature). The central finding is a near-complete polarization of probability mass onto a single token, driven by extremely large logit margins. This collapse reflects a structural distributional issue, not a mild miscalibration. In such regimes, temperature scaling is ineffective: once margins grow this large, relative logit differences become dominated by noise and cease to carry calibration-relevant information. Adjusting temperature therefore cannot meaningfully “repair” the distribution or alter the calibration behaviour observed.

A.13.2 Experiment 2: Temperature During Response Generation

In Experiment 2, temperature affects only the generative content of responses, not the utility of the evaluated UQ methods. Since our goal is not to optimize accuracy or reduce hallucinations, but to identify responses that exhibit uncertain or erroneous reasoning, additional generations at different temperature levels would not strengthen the benchmark. Running a full temperature sweep would require repeating 685,000 generations per temperature value, with negligible methodological benefit:

temperature primarily alters response variability, not the structure or comparability of the UQ scores themselves.

A.14 Experiment 2: Temperature for Verbalized Methods

For verbalized uncertainty methods, we use temperature 1.0 with greedy decoding. This ensures the model provides the reliability judgment it is most confident in. For P(True), the situation mirrors Experiment 1: the method depends on label probabilities, which frequently exhibit polarization and show no reliable alignment with predictive accuracy. Increasing temperature cannot repair this mismatch, nor does it meaningfully alter calibration quality when the underlying distribution is already collapsed.

For these reasons, additional temperature-adjustment experiments would not yield meaningful benefits for either experiment.

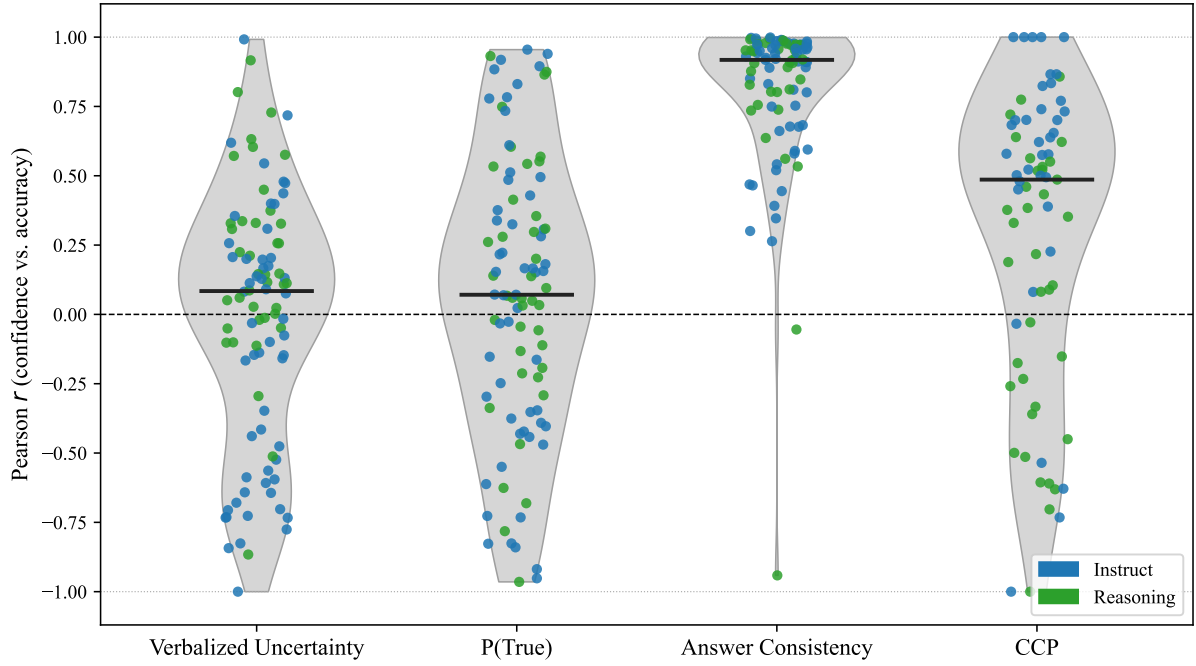


Figure A.13: **Calibration Correlation by UQ Method.** Each point represents one (model, dataset) pair and shows the Pearson correlation (r) between the bin-confidence midpoints and the bucket accuracies of the corresponding calibration curve. Bins with no items are excluded; all non-empty bins contribute equally. A value of $r = 1$ means accuracy increases perfectly in step with confidence; $r = 0$ indicates no linear relationship; negative values indicate miscalibration where higher confidence predicts lower accuracy. Points are coloured by model type (blue = instruct, green = reasoning). The grey violin summarises the distribution across all points for each method; horizontal bars mark the median.

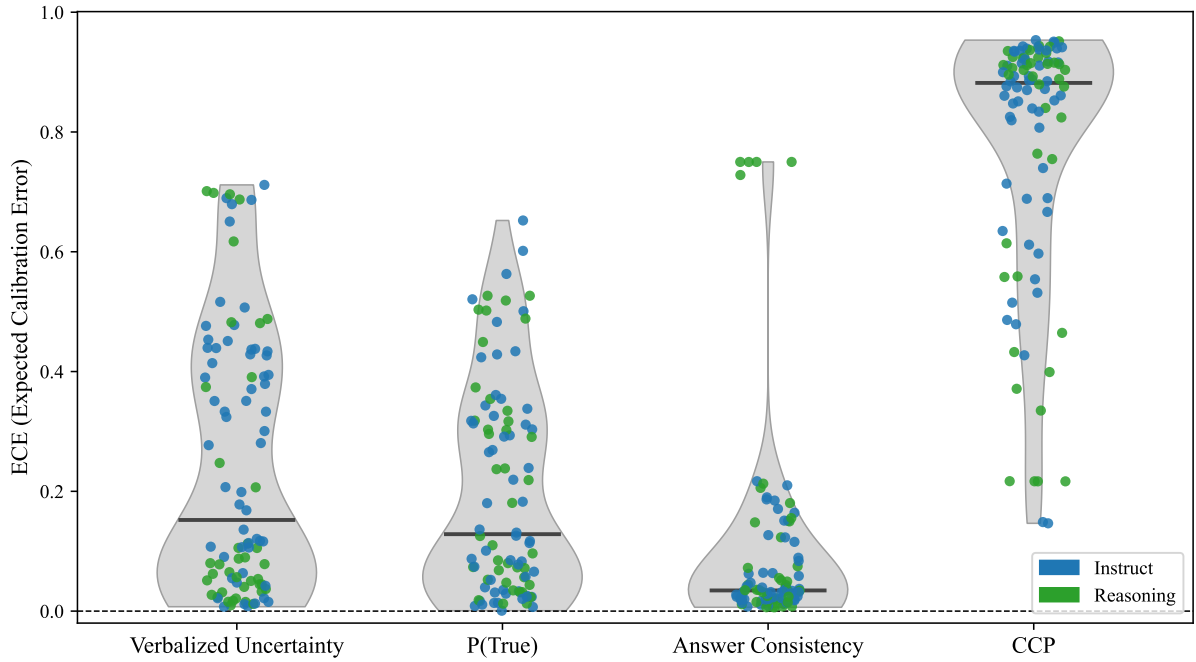


Figure A.14: **Expected Calibration Error (ECE) by UQ Method.** Distribution of ECE values across all (model, dataset) pairs for each UQ method. Lower values indicate better calibration. Points are coloured by model type (blue = instruct, green = reasoning); violins summarize the overall distribution and horizontal bars indicate the median.

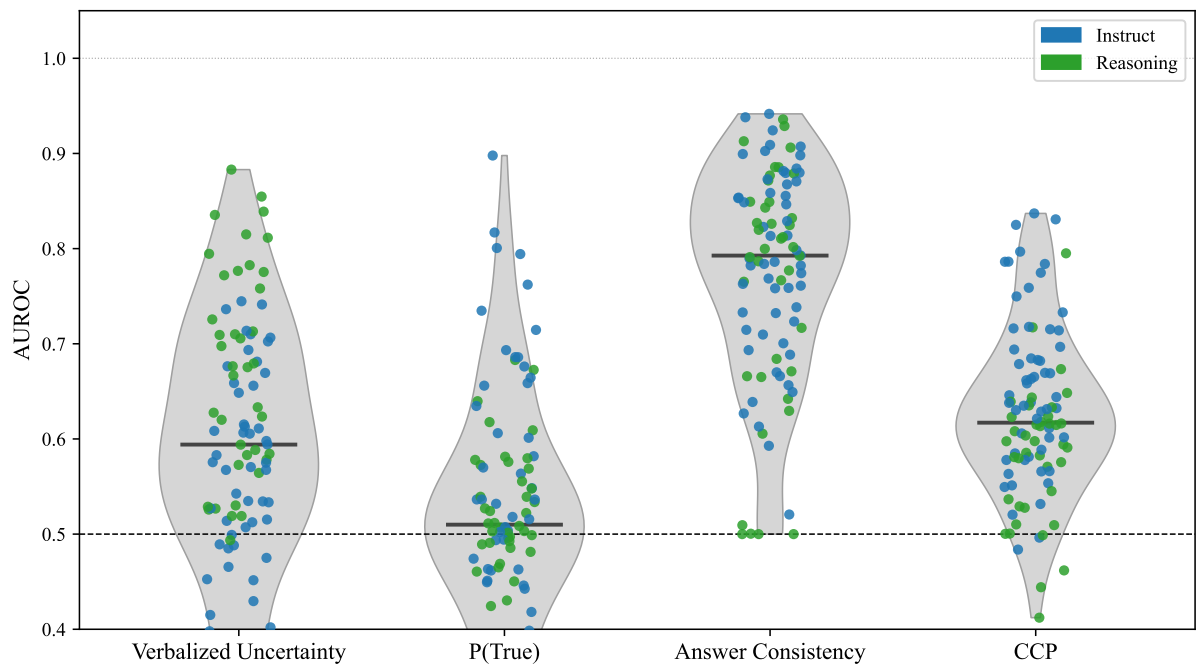


Figure A.15: **AUROC by UQ Method.** Distribution of AUROC values across all (model, dataset) pairs for each UQ method. Higher values indicate better discrimination between correct and incorrect responses. Points are coloured by model type (blue = instruct, green = reasoning); violins summarize the overall distribution and horizontal bars indicate the median.



Figure A.16: **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.

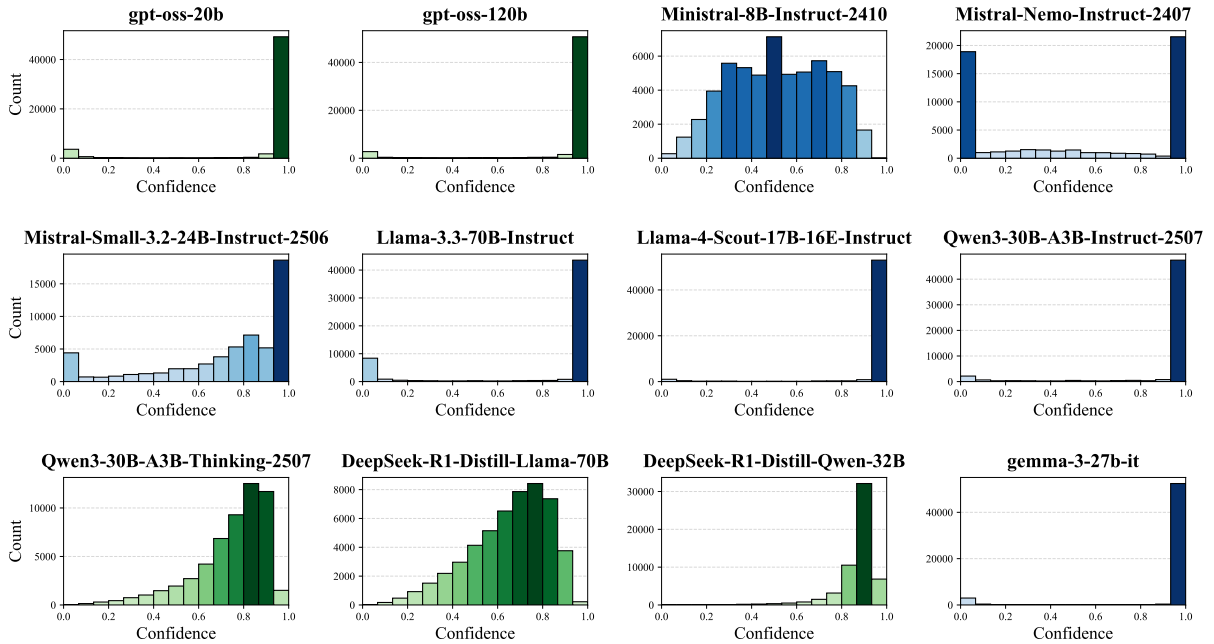


Figure A.17: **Distribution of Confidence Scores Assigned by $P(\text{True})$ Across Models.** Confidence scores have been aggregated across all datasets (57, 500 prompts in total). Most models exhibit a clear polarization towards either (A) (representing the model’s confidence that the answer is true) or (B) (representing the model’s confidence that the answer is false) regarding the token probabilities.