From Entropy to Calibrated Uncertainty: Training Language Models to Reason About Uncertainty

Azza Jenane* , **Nassim Walha***, **Lukas Kuhn, Florian Buettner**
German Cancer Research Center (DKFZ)
German Cancer Consortium (DKTK)
Goethe University Frankfurt, Germany

Abstract

Large Language Models (LLMs) that can express interpretable and calibrated uncertainty are crucial in high-stakes domains. While methods to compute uncertainty post-hoc exist, they are often sampling-based and therefore computationally expensive or lack calibration. We propose a three-stage pipeline to post-train LLMs to efficiently infer calibrated uncertainty estimates for their responses. First, we compute fine-grained entropy-based uncertainty scores on the training data, capturing the distributional variability of model outputs in embedding space. Second, these scores are calibrated via Platt scaling, producing reliable and human-interpretable uncertainty signals. Finally, the target LLM is post-trained via reinforcement learning to align its policy with these calibrated signals through a verifiable reward function. Unlike post-hoc uncertainty estimation methods, our approach provides interpretable and computationally efficient uncertainty estimates at test time. Experiments show that models trained with our pipeline achieve better calibration than baselines and generalize to unseen tasks without further processing, suggesting that they learn a robust uncertainty reasoning behavior.

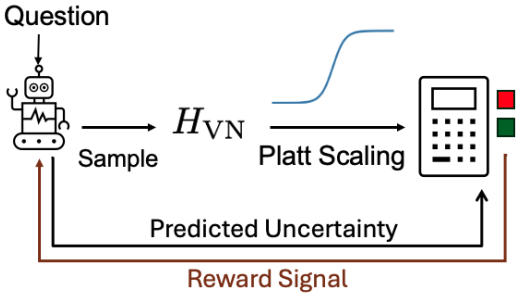

Figure 1: Overview of the reward signal generation pipeline

1 Introduction

Large Language Models (LLMs) have achieved strong performance across a broad range of Natural Language Processing (NLP) tasks, including question answering, summarization, and natural language understanding (Touvron et al., 2023; Chiang et al., 2023; Achiam et al., 2023). Despite these advances, they remain prone to generating confident yet incorrect outputs, commonly referred to as hallucinations (Hadi et al., 2023).

As LLMs are increasingly deployed in high-stakes domains such as healthcare (Wang and Zhang, 2024; Bani-Harouni et al., 2024), finance (Cao, 2022), and legal decision support (Zhong et al., 2020), reliable and interpretable uncertainty estimates are essential to enable risk-aware decision-making, appropriate human oversight, and safe system integration. A central requirement for such reliability is to have calibrated confidence; predicted uncertainty must align with empirical correctness (Guo et al., 2017).

[1]* Equal contribution.

Workshop "Towards Trustworthy Predictions: Theory and Applications of Calibration for Modern AI" at AISTATS 2026, Tangier, Morocco.

Current approaches to uncertainty estimation in LLMs are predominantly post-hoc, relying on sampling-based statistics and entropy measures (Farquhar et al., 2024; Walha et al., 2025; Nikitin et al., 2024). Given a user query, these methods generate multiple responses from the model and quantify uncertainty through the semantic variability among the sampled outputs. While effective, they incur substantial computational overhead due to repeated sampling. Moreover, they yield scale-free uncertainty values that perform well on ranking-based metrics such as AUROC and AUPR, but are inherently uncalibrated, as they do not map directly to probabilities. A separate line of work leverages verbalized uncertainty, employing various prompting strategies to elicit explicit confidence scores from the model (Kadavath et al., 2022; Xiong et al., 2023). These methods are generally more computationally efficient, but their reliability is strongly dependent on model size and capacity. In particular, smaller models deployed on-device in privacy-sensitive settings have been shown to produce poorly calibrated and unreliable confidence estimates (Xiong et al., 2023; Yang et al., 2024). More recent reinforcement learning approaches introduce verifiable rewards to encourage alignment between predicted confidence and actual correctness; however, they often rely on coarse supervision signals or computationally expensive optimization schemes (Bani-Harouni et al., 2025; Damani et al., 2025).

To address the aforementioned limitations, we propose a three-stage framework that post-trains LLMs to express calibrated uncertainty directly. In the first stage, we compute fine-grained uncertainty scores using von Neumann entropy (Von Neumann, 2018; Walha et al., 2025) over embedding representations derived from multiple sampled generations, capturing distributional structure beyond binary correctness. In the second stage, these scores are calibrated via Platt scaling (Platt et al., 1999) to obtain interpretable probability targets. In the third stage, the model is trained using Group Relative Policy Optimization (GRPO) with a verifiable reward (Wen et al., 2025; Shao et al., 2024) that aligns the model's predicted uncertainty with these calibrated signals. This framework integrates uncertainty estimation directly into model behavior while remaining computationally efficient at inference time. The contributions of this work are as follows:

- We introduce a novel uncertainty calibration reward that aligns the model's verbalized uncertainty with a state-of-the-art sampling-based measure, while explicitly targeting calibrated probability outputs.
- We demonstrate that our reward yields verbalized uncertainties with high rank-correlation with the sampling-based measure, thus inheriting its strong performance on ranking-based metrics, while also achieving state-of-the-art calibration and high efficiency at inference time.
- We compare our reward against a Brier-score-based reward (Glenn et al., 1950) commonly used in the literature (Damani et al., 2025), and demonstrate superior performance both in-distribution and out-of-distribution.

2 Background

2.1 Fine-Grained Entropy-Based Uncertainty

Walha et al. (2025) propose a fine-grained uncertainty measure based on spectral entropy in embedding space. For each input, multiple generations are sampled and mapped to embedding vectors. A kernel matrix is constructed over the embeddings to capture pairwise similarity between generated responses.

The eigenvalues $\{\lambda_i\}_{i=1}^{N}$ of the normalized kernel matrix are then used to compute the von Neumann entropy (Von Neumann, 2018),

$$H_{\mathrm{VN}} = -\sum_{i=1}^{N} \lambda_i \log \lambda_i,$$

which quantifies dispersion in representation space. By operating at the representation level and aggregating across multiple samples, this approach provides a continuous uncertainty signal that captures distributional variability in semantic space, beyond token-level predictive entropy or binary correctness signals.

2.2 Platt Scaling for Calibration

Calibrated confidence (Guo et al., 2017). Let $\hat{Y}$ and $\hat{P}$ denote the model's predicted answer and its associated confidence, respectively, and let Y denote the ground truth answer to query X. A calibrated model satisfies

$$\mathbb{P}(\hat{Y} = Y \mid \hat{P} = p) \approx p, \quad \forall p \in [0, 1],$$

meaning that among all predictions for which the model assigns confidence p, the fraction of correct predictions is approximately p.

Platt scaling (Platt et al., 1999). Platt scaling is a post-hoc calibration method that maps arbitrary real-valued scores to calibrated probability estimates via a parametric logistic transformation. Given an uncalibrated score $s \in \mathbb{R}$, Platt scaling fits a sigmoid

function $p = \sigma(As+B)$, where the parameters A and B are learned by minimizing the negative log-likelihood on a held-out validation set with binary correctness labels. This transformation preserves the ranking of the original scores while rescaling their values so that the resulting probabilities better reflect empirical accuracy.

2.3 Reinforcement Learning for Uncertainty Estimation

Reinforcement Learning (RL) provides a framework for optimizing model behavior with respect to a reward signal rather than supervised targets. It has recently been explored for uncertainty estimation by defining rewards that encourage alignment between predicted confidence and empirical correctness (Damani et al., 2025; Bani-Harouni et al., 2025). Such formulations allow uncertainty calibration to be integrated directly into the training objective, rather than computed post-hoc.

Group Relative Policy Optimization (GRPO) has been proposed as a computationally efficient reinforcement learning algorithm particularly suited for training LLMs (Shao et al., 2024). Unlike Proximal Policy Optimization (PPO), which requires maintaining and updating a large critic network throughout training, GRPO performs relative comparisons within groups of sampled responses, using group-normalized rewards to guide policy updates. This substantially reduces both memory usage and training complexity while preserving training stability. As a result, GRPO enables scalable reinforcement learning for LLMs at a lower computational cost compared to standard PPO-based pipelines.

3 Methodology

Our pipeline, illustrated in Figure 1, aims to improve the LLM's ability to accurately estimate its own uncertainty given an input question and the model's corresponding answer. Formally, let $\mathcal{X}$ denote the space of questions and $\mathcal{Y}$ denote the space of answers. Given a question $x \in \mathcal{X}$ with ground truth answer $y^* \in \mathcal{Y}$, we sample an answer $\hat{y}$ from the base LLM $\pi_{\theta_0}(y \mid x)$. We then fine-tune the model to predict a scalar uncertainty estimate $u_\theta(x, \hat{y}) \in [0, 1]$, interpreted as the probability that the answer $\hat{y}$ is incorrect, i.e., $u_\theta(x, \hat{y}) \approx \mathbb{P}(\hat{y} \neq y^* \mid x)$.

Entropy-Based Uncertainty Signal. Following the work of Walha et al. (2025), for a fixed base model π_{θ_0}, we generate K stochastic samples $\{y^{(k)}\}_{k=1}^{K} \sim \pi_{\theta_0}(\cdot \mid x)$ and compute a semantic dispersion score

$$S(x) = H_{VN}(x)$$

using kernel-based von Neumann entropy over their embedding representations. This score serves as a continuous proxy for uncertainty.

Calibration Mapping. Since $S(x)$ is not inherently probabilistic and raw values are not directly interpretable or calibrated, we learn a calibration function $g : \mathbb{R}_{\geq 0} \to [0, 1]$ using Platt scaling (Platt et al., 1999) on held-out data with binary correctness labels $z \in \{0, 1\}$, where $z = 1$ indicates an incorrect response. The calibrated uncertainty is then defined as $u_{\text{cal}}(x) = g(S(x))$, which estimates $\mathbb{P}(\text{incorrect} \mid x)$ under the validation distribution.

Reinforcement Learning for Calibrated Uncertainty. Similar to Bani-Harouni et al. (2025), we decouple answer generation from uncertainty estimation during training: answers are generated first and treated as fixed, while uncertainty is produced in a separate step and optimized independently. This ensures that answer quality remains unaffected by the uncertainty calibration objective. Unlike Damani et al. (2025), which fine-tunes all model weights, we adopt Low-Rank Adaptation (LoRA) (Hu et al., 2022) as a parameter-efficient alternative. Beyond reducing memory overhead and mitigating catastrophic forgetting, LoRA naturally supports the decoupling of answer generation and uncertainty prediction at inference time, as the learned adapters can be selectively applied after answer generation to produce uncertainty estimates.

To further optimize training efficiency, we use reinforcement learning with the Group Relative Policy Optimization (GRPO) algorithm. Our entropy-based reward function encourages alignment between the predicted uncertainty $u_\theta(x, \hat{y})$ and the calibrated target $u_{\text{cal}}(x)$ and is defined as follows:

$$R_{\textbf{entropy}}(u_\theta, u_{\text{cal}}) = 1 - \max\big(0.05,\ |u_\theta - u_{\text{cal}}|\big).$$

During training, the model is provided with the question x and its pre-generated answer $\hat{y}$, and is prompted to first produce a reasoning trace about its uncertainty in a chain-of-thought (CoT) format, followed by a scalar uncertainty prediction $u_\theta(x, \hat{y})$. By optimizing the scalar output directly, the model is implicitly encouraged to develop a useful reasoning trace that supports reliable uncertainty estimation.

Table 1: Performance metrics across methods on in-domain (TriviaQA + Natural Questions) and out-of-domain (GSM8k) datasets.

	TriviaQA + NQ (ID)			GSM8k (OOD)	
Method	**ECE (%) ↓**	**AUROC (%) ↑**	**Spearman ↑**	**ECE (%) ↓**	**AUROC (%) ↑**
Base	41.99	51.89	0.03	32.22	53.79
Base+CoT	34.17	66.18	0.17	22.25	62.17
Brier	15.70	**83.36**	0.52	33.28	**66.89**
Entropy-based (ours)	**7.2**	81.53	**0.67**	**3.15**	**66.73**

4 Experiments

4.1 Experimental Setup

We conduct experiments on subsets of TriviaQA (Joshi et al., 2017) and Natural Questions (NQ) (Kwiatkowski et al., 2019), two widely used open-domain question answering benchmarks. We adopt the Qwen2.5-7B-Instruct model, which has demonstrated strong performance in reinforcement learning settings (Hu et al., 2025; Damani et al., 2025), and initialize reinforcement learning training directly from this base model, following recent practice. To assess the performance of our method and the baselines, we report Expected Calibration Error (ECE), Area Under the Receiver Operating Characteristic Curve (AUROC), and Spearman correlation with respect to the calibrated uncertainty targets. Full experimental details, including evaluation protocol, training procedure, and hyperparameter settings, are provided in Appendix A.

Methods. We evaluate the following methods, which differ in their reward design:

- **Base**: The pretrained Qwen2.5-7B-Instruct model without reinforcement learning, serving as a reference.
- **Base + CoT**: The pretrained Qwen2.5-7B-Instruct model without reinforcement learning, aided by chain-of-thought reasoning (Wei et al., 2022).
- **Brier**: Initialized from the base model and trained with GRPO using a reward derived solely from the Brier score between predicted uncertainty and binary correctness labels. This reward is defined as $R_{Brier}(u_\theta) = 1 - (u_\theta - \mathbb{1}_{\hat{y} \neq y^*})^2$ and was used in previous work (Damani et al., 2025).
- **Entropy-based (ours)** : Initialized from the base model and trained with GRPO using our calibrated entropy-based reward signal $R_{entropy}$.

4.2 Results

We evaluate all methods on a held-out in-domain (ID) test set comprising TriviaQA and Natural Questions, and assess out-of-domain (OOD) generalization on GSM8K (Cobbe et al., 2021). Results are reported in Table 1.

In-domain performance. Our entropy-based RL method achieves the strongest overall performance. It reduces ECE from 41.99% (Base) and 34.17% (Base+CoT) to **7.2%**, substantially improving calibration over all baselines. While the Brier variant also improves calibration, reaching 15.70%, it remains notably worse than the entropy-based approach. In terms of ranking quality, both the Brier (83.36%) and entropy-based (81.53%) variants substantially outperform the base models (51.89% and 66.18%, respectively). Furthermore, the entropy-based method achieves the highest alignment with calibrated uncertainty signals, yielding the strongest Spearman correlation (**0.67**) across all methods.

Out-of-domain generalization. On GSM8K, the entropy-based method again achieves the best calibration, reducing ECE to **3.15%**, compared to 32.22% (Base), 22.25% (Base+CoT), and 33.28% (Brier). In terms of AUROC, both the Brier (66.89%) and entropy-based (66.73%) variants substantially outperform the baselines (53.79% and 62.17%), with the Brier variant marginally higher.

Overall, while CoT prompting and Brier-based optimization improve ranking performance, they do not consistently yield well-calibrated uncertainty estimates. In contrast, reinforcement learning with entropy-calibrated uncertainty targets leads to substantial gains in both calibration and ranking quality, and generalizes robustly to out-of-domain settings.

5 Discussion and Conclusion

We introduced a novel uncertainty calibration reward grounded in an established entropy-based uncertainty measure for LLMs. Our approach not only produces

uncertainty estimates that are well-aligned with this state-of-the-art measure, but also significantly outperforms several baselines in terms of both calibration and uncertainty ranking, with consistent gains observed in-distribution and out-of-distribution. Furthermore, by combining GRPO with LoRA adapters, our framework maximizes training efficiency while remaining computationally lightweight at inference time.

While we provide evaluation across multiple metrics and datasets, extending the experiments to a broader set of models would yield a more comprehensive assessment. Additionally, our evaluation remains purely empirical, which is common practice in LLM research but leaves open the question of theoretical grounding.

Overall, our approach represents a promising direction toward efficient and reliable uncertainty quantification for large language models.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bani-Harouni, D., Navab, N., and Keicher, M. (2024). Magda: Multi-agent guideline-driven diagnostic assistance. In *International workshop on foundation models for general medical AI*, pages 163–172. Springer.

Bani-Harouni, D., Pellegrini, C., Stangel, P., Özsoy, E., Zaripova, K., Keicher, M., and Navab, N. (2025). Rewarding doubt: A reinforcement learning approach to calibrated confidence expression of large language models. *arXiv e-prints*, pages arXiv–2503.

Cao, L. (2022). Ai in finance: challenges, techniques, and opportunities. *ACM Computing Surveys (CSUR)*, 55(3):1–38.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., et al. (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Damani, M., Puri, I., Slocum, S., Shenfeld, I., Choshen, L., Kim, Y., and Andreas, J. (2025). Beyond binary rewards: Training lms to reason about their uncertainty. *arXiv preprint arXiv:2507.16806*.

Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Glenn, W. B. et al. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J., Mirjalili, S., et al. (2023). A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. (2022). Lora:

Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Hu, J., Zhang, Y., Han, Q., Jiang, D., Zhang, X., and Shum, H.-Y. (2025). Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*.

Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. (2017). Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al. (2022). Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al. (2019). Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Nikitin, A., Kossen, J., Gal, Y., and Marttinen, P. (2024). Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *Advances in Neural Information Processing Systems*, 37:8901–8929.

OpenAI (2024). GPT-4o mini: Advancing cost-efficient intelligence. `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`. Accessed: 2024.

Platt, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al. (2024). Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Von Neumann, J. (2018). *Mathematical foundations of quantum mechanics: New edition*. Princeton university press.

Walha, N., Gruber, S. G., Decker, T., Yang, Y., Javanmardi, A., Hüllermeier, E., and Buettner, F. (2025). Fine-grained uncertainty decomposition in large language models: A spectral approach. *arXiv preprint arXiv:2509.22272*.

Wang, D. and Zhang, S. (2024). Large language models in medical and healthcare fields: applications, advances, and challenges. *Artificial intelligence review*, 57(11):299.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Wen, X., Liu, Z., Zheng, S., Ye, S., Wu, Z., Wang, Y., Xu, Z., Liang, X., Li, J., Miao, Z., et al. (2025). Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *arXiv preprint arXiv:2506.14245*.

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., and Hooi, B. (2023). Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Yang, D., Tsai, Y.-H. H., and Yamada, M. (2024). On verbalized confidence scores for llms. *arXiv preprint arXiv:2412.14737*.

Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., and Sun, M. (2020). How does nlp benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5218–5230.

A Implementation Details

A.1 Calibrated uncertainty signals

To compute entropy-based uncertainty signals, we sample model responses at temperature $t = 1.0$, which balances output diversity and determinism for reliable variability estimation. For the correctness labels used in Platt scaling and evaluation, we follow Walha et al. (2025) and Farquhar et al. (2024) and generate answers at a low temperature $t = 0.1$, approximating the model's best-effort prediction. These answers are then compared against the ground truth using GPT-4o-mini (OpenAI, 2024) as an automated judge to obtain binary correctness labels.

A.2 GRPO training

We train on a combined subset of TriviaQA and Natural Questions consisting of 18,000 training samples and evaluate on a held-out set of 2,000 samples. All experiments are conducted on a single NVIDIA H200 NVL GPU, with total training time ranging between 10 and 14 hours depending on the method.

Reinforcement learning is performed using GRPO for 1,000 update steps with a batch size of 32 and a group size of 16. The grouped sampling stabilizes learning under stochastic reward signals derived from calibrated uncertainty targets. Within each group, responses are sampled with temperature 1.5 to encourage diversity and improve exploration during policy optimization.

We employ LoRA-based parameter-efficient fine-tuning with rank $r = 16$ and scaling factor $\alpha = 32$, together with a dropout rate of 0.05. This configuration provides sufficient adaptation capacity while preserving the pretrained model's generation abilities and preventing catastrophic forgetting.

A.3 Evaluation

We evaluate uncertainty quality using the following metrics:

- **Expected Calibration Error (ECE)** $\downarrow$: measures the discrepancy between predicted uncertainty and empirical error rates. Lower values indicate better calibration.
- **AUROC** $\uparrow$: Area Under the Receiver Operating Characteristic curve, which reflects the quality of the uncertainty ranking.
- **Spearman Correlation** $\uparrow$: is a rank correlation measure that describes the monotonic relationship between predicted uncertainty and calibrated uncertainty targets.

For evaluation, we generate standard answers at low temperature $t = 0.1$ to approximate each model's best-effort prediction. These answers are provided as input to the uncertainty inference prompt (Figures 4 and 5) for all evaluated methods, from which the models produce their predicted uncertainties u_θ. Binary correctness labels are obtained by comparing the generated standard answers against the ground truth using GPT-4o-mini as a judge. These labels are then used alongside the predicted uncertainties to compute ECE and AUROC. Spearman correlation is computed by comparing the predicted uncertainties against the calibrated target uncertainties u_{cal} on the test set.

B Prompts

Objective
Answer the question below. Think briefly, then give your final answer. Always provide an answer—if you are unsure or do not know the answer, give your best guess without indicating uncertainty. Your answer has to be short and precise.

Output Format
R: ⟨1-2 sentences of reasoning⟩
A: ⟨your final answer⟩

Task

Figure 2: Answer generation prompt template for entropy sampling and standard answers

Objective
In this task, you will receive a question. You will also receive a ground truth answer to the question and a model generated answer. Your goal is to compare the ground truth answer and the model generated answer in order to decide whether the model generated answer is correct or not.

Important Rules
1. The model generated answer is correct, when it is a valid answer to the question, and semantically equivalent to the ground truth answer. It does not necessarily need to overlap with the ground truth answer lexically.
2. If the model generated answer contains more information (more specific) or less information (less specific) than the ground truth answer, but still correctly answers the question, then you should consider it correct.
3. If you decide that the model generated answer is correct, say yes, otherwise say no.
4. Your output should only contain your decision (yes or no). It should not contain any other text, explanation or reasoning.

Figure 3: Correctness judge prompt template.

You are an assistant that estimates your total uncertainty about your answer to a question.

I previously asked you to answer the question below, and you provided the answer below. Now, I want you to reflect on your answer and estimate how uncertain you are about it.

Uncertainty is the probability that your answer would be wrong.
Before giving the final value, briefly reflect on why your answer might be right and why it might be wrong. Use your current judgment to produce a probability between 0 and 1.

Output Format
Reflection: ⟨1-2 sentences on why the answer might be right or wrong⟩
Total_uncertainty: ⟨float between 0 and 1⟩

Figure 4: Training and inference prompt template for post-trained and CoT methods.

Given the question and your answer below, estimate your uncertainty about the answer. Only output the number, no explanations.
Output Format
Total_uncertainty: ⟨float between 0 and 1⟩

Figure 5: Inference prompt template for the base model.