P. Sychev, A. Goncharov, D. Vyazhev, E. Khalafyan,
A. Zaytsev

WHEN AN LLM IS APPREHENSIVE ABOUT ITS ANSWERS – AND WHEN ITS UNCERTAINTY IS JUSTIFIED

Abstract. Uncertainty estimation is crucial for evaluating Large Language Models (LLMs), particularly in high-stakes domains where incorrect answers result in significant consequences. Numerous approaches consider this problem, while focusing on a specific type of uncertainty, ignoring others. We investigate what estimates, specifically token-wise entropy and model-as-judge (MASJ), would work for multiple-choice question-answering tasks for different question topics. Our experiments consider three LLMs: Phi-4, Mistral, and Qwen of different sizes from 1.5B to 72B and 14 topics. While MASJ performs similarly to a random error predictor, the response entropy predicts model error in knowledge-dependent domains and serves as an effective indicator of question difficulty: for biology ROC AUC is 0.73. This correlation vanishes for the reasoning-dependent domain: for math questions ROC-AUC is 0.55. More principally, we found out that the entropy measure required a reasoning amount (amount of reasoning steps). Thus, data-uncertainty related entropy should be integrated within uncertainty estimates frameworks, while MASJ requires refinement. Moreover, existing MMLU-Pro samples are biased, and should balance required amount of reasoning for different subdomains to provide a more fair assessment of LLMs performance.

1. Introduction

LLMs have succeeded in various tasks, demonstrating impressive capabilities in generating human-like text and solving complex problems. However, their ability to accurately assess the uncertainty in their predictions remains a significant challenge [21]. This gap between a model's confidence and its actual correctness, often called the confidence gap [17], poses critical risks, particularly in high-stakes applications where overconfidence in

incorrect answers leads to severe consequences: in domains such as health, law [13], education, or economics, an LLM's overestimation of its accuracy could result in harmful decisions or misinformation. Understanding and decreasing this confidence gap is therefore essential for ensuring the safe and reliable deployment of LLMs.

Addressing this limitation has led to the development of numerous uncertainty estimation techniques. Most of them belong to one of two problem statements. They either focus on binary classification or rely on rich signals from long, free-text responses [16]. These methods often generalize poorly for concise multiple-choice questions.

Moreover, existing methods tend to answer how good an overall confidence estimate is, ignoring why uncertainty appeared in a specific case. While classic decomposition of uncertainty to aleatoric (data) and epistemic (model) [20] also holds for LLMs, the problem is challenging [23]. For example, the authors of [7] consider only epistemic uncertainty.

To address these gaps, we propose a pipeline that enables detailed investigation of domain-specific or complexity-labeled datasets with multiple-choice answers. This framework allows us to explore how commonly used uncertainty estimation methods, particularly entropy-based approaches, perform under varying question topics and levels of required reasoning. We can also leverage an auxiliary LLM to generate labels for reasoning, knowledge requirements, and the number of reasoning steps. The introduction of the pipeline led to a more detailed investigation of datasets commonly used for benchmarking LLMs and their uncertainty estimates. In more detail, our main contributions are the following:

- We present a fully automated pipeline for evaluating uncertainty estimation approaches within the MMLU-Pro setting. It is presented in Figure 1 and allows datasets or tested methods to be varied. The associated work is available on Github[1].
- We propose using data uncertainty estimates based on token-wise entropy and model uncertainty estimates based on MASJ. Thus, our approach covers two main components of uncertainty estimation: data and model one, given that MMLU-Pro answers are typically short and provide limited information on model confidence.
- Within the developed pipeline, we consider different subsets of the MMLU-Pro dataset with splits by domain and reasoning requirement estimated via MASJ proposed in this work.

---

[1]https://github.com/LabARSS/question-complextiy-estimation

- Our experimental evidence that includes consideration of four models (Phi-4, Mistral-Small-24B-Instruct, Qwen-1.5 B, Qwen-72B) suggests that the entropy predicts well errors of an LLM if the amount of required reasoning is small. Moreover, its quality increases if the model size increases. Unlike MASJ provides much weaker results, being unable to identify error patterns. Further progress here would be possible with more reasoning steps during an uncertainty estimate with an entropy-based approach on top of it.

- We also observe that the existing dataset MMLU-Pro has internal biases with the complexity of the question, the same as in [19], and the required reasoning steps amount for them varies significantly depending on the topic. Thus, more advanced and less biased datasets are required for a fair measurement of the progress of LLMs. Additional human annotation of questions in a dataset could identify further gaps in LLMs' capabilities.

2. Related Works

Tests have been a popular way to estimate an individual's proficiency for more than 500 years [10]. After their emergence, LLMs have also been extensively tested and compared with each other or humans [6]. Several papers established benchmarks, with natural ones being QA datasets: an LLM receives a prompt and should generate a response [5]. Different questions exist, but automated evaluation of answers is available only for multiple-choice questions. For LLMs, slight changes in the way the questions are asked can significantly affect the accuracy of a model, probably due to data leakage and format overfitting [4], other options are much more expensive and challenging. One of the recent options that takes into account possible leakages - and, on the other hand, is widely used is an expanded version of the MMLU dataset, called MMLU-Pro [24] with about 0.56 accuracy for Llama-70B model. The presented ~ 12000 questions span 13 diverse topics, including STEM and other areas of scientific studies. The complexity, volume, and diversity of the dataset provide a strong foundation for our study of how different aspects of the questions affect the probability of correctly answering them.

In related works, datasets such as ARC-AI2 [2] and GPQA [3] have been widely used to evaluate multiple-choice QA in the scientific domain. However, ARC-AI2 is often criticized for its relatively low difficulty, while

GPQA, though highly challenging, suffers from a dataset size of fewer than 500 samples, which restricts its utility for robust evaluation. To address these limitations and ensure a comprehensive assessment, we focus our final experiments on MMLU-Pro, which offers a larger scale and a more balanced difficulty level for evaluating model performance in scientific reasoning tasks.

MMLU-Pro is suited for evaluating a model's ability to identify uncertainty in the generated output and signal potential errors, which addresses a fundamental question in AI safety. In a broader context, we can refer to this effect as hallucination [23]; In a more classic context, it is a prediction error [26]. Another related approach is assess the question difficulty, provided in [25], but suitable there only for an RAG scenario. Universal Model-as-Judge (MASJ) approaches use available models or a larger and more powerful model to evaluate LM output [11] and can be applied in a broader context.

What makes the problem more difficult is the diverse nature of possible uncertainty sources. Classic works decompose the uncertainty into the model part, related to lack of knowledge of a model about a specific output, and the data part, related to the complexity of the data itself [20]. For natural language, model uncertainty refers to the familiarity of a specific domain and the presence of relevant content in a training sample. Data uncertainty refers to noise in the example or the overall complexity of a question at hand. While broader classification exists with five different features for a generated question [12], we follow this approach while identifying the considered parts of the uncertainty. Using entropy-based measures, such as semantic entropy [14] and word-sequence entropy [18], is highly effective for evaluating uncertainty in LLMs because these methods capture the variability and confidence of model predictions in a structured way. Semantic entropy, as proposed, quantifies uncertainty by considering the meaning of generated text, making it robust to paraphrasing and linguistic variations. Similarly, word-sequence entropy demonstrates that word-sequence entropy effectively measures uncertainty in open-ended tasks such as medical QA, providing insight into model confidence and reliability. In our case, we use the token-wise entropy approach [21] due to its simplicity and suitability for MMLU cases with short answers with multiple-choice. For model uncertainty, we consider a variation of the MASJ approach with a specific prompt [8, 11].
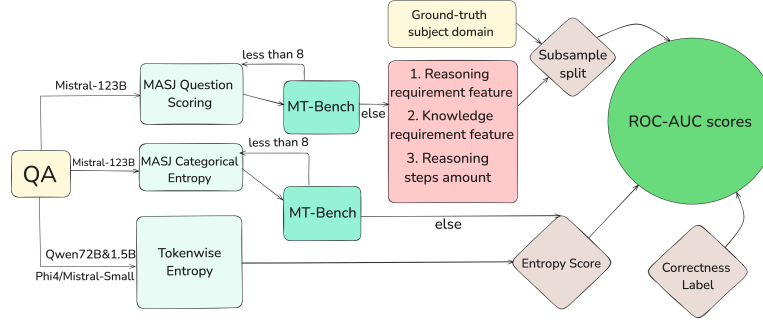
Figure 1. Question complexity evaluation pipeline. See more details in the methods section

The survey paper [9] provides a comprehensive taxonomy of uncertainty estimation methods in LLMs, highlighting state-of-the-art approaches such as evaluation through LLM or token-wise entropy. For an MMLU-type dataset, we would be able to identify how different types of uncertainties are combined for typical questions in different subjects and how our measurement corresponds to the estimate of the required amount of reasoning steps for specific questions. These results would help identify hallucinations and errors in language model outputs and provide insights into how to reduce the frequency of these effects by training the model using additional data and varying the model size.

3. Methods

3.1. General Pipeline. Input. We consider a set of questions $Q = \{q_i\}_{i=1}^n$ and a set of answers $A = \{a_i\}_{i=1}^n$ of size n with generated answers $G = \{g_i\}_{i=1}^n$. The questions are augmented with true topic labels $T = \{t_i\}_{i=1}^n$ and correctness of the answers $Y = \{y_i\}_{i=1}^n$, where y_i is binary value which means equality of ground truth answer a_i and generated answer g_i.

Uncertainty estimation. Our procedures estimate the uncertainty of a model response given a question and a corresponding answer $U = \{u_i\}_{i=1}^n$, each $u_i = f(q_i, g_i)$ and depend on the LLM used. We additionally prompt models to obtain an estimation of the reasoning requirements $R = \{r_i\}_{i=1}^n$.

Uncertainty validation. Finally, given all this information, we compare the uncertainty estimates U with the correctness labels Y to obtain the ROC-AUC values and other quality measures. Furthermore, we examine how good specific uncertainty estimates U are for separate topics T or the required amount of reasoning steps R.

Technical details. For each question in the dataset, we prompt a considered model for an answer and record the entropy for the answer. We discard answers with incorrect formatting that appear in less than 5% of the questions.

In model-as-judge (MASJ), we use a large 123B Instruct Mistral model `Mistral-Large-Instruct-2411` to obtain estimated complexity according to the level of education. The same MASJ technique with the same large model was utilized to obtain the reasoning estimate. We ask a binary question of the deep reasoning required and how many steps it would take: low, medium, high. Specific prompts are available in the Appendix A.

For a more detailed analysis, we split the whole set of questions to topics using available labels or reasoning and educational level complexity obtained via MASJ.

3.2. Uncertainty Estimations.

We consider two reliable approaches for uncertainty estimation: the entropy-based procedure [21] and the model-as-a-judge [11]. Details on both are provided below.

3.2.1. *Entropy Uncertainty Estimation.*

Entropy-based uncertainty estimation is a fundamental approach for evaluating LLMs confidence in their predictions [22]. In MMLU, the expected answer is a number with one or two tokens length. So, we use token-wise entropy from the model's output logits.

Our implementation computes entropy through a two-step procedure. Firstly, we obtain the logits $\mathbf{z} = (z_1, \ldots, z_k)$ from the LLM's final layer for each token position, where k equals the vocabulary size. Secondly, we convert these logits into a probability distribution using the softmax function:

$$p_i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}}.$$

Next, we calculate the entropy of this distribution:

$$u = -\sum_{i=1}^{k} p_i \log p_i.$$

This approach aligns with the work [21], which showed that token-level entropy correlates well with uncertainty and performance of the model. When a model is confident in its next-token prediction, the probability mass concentrates on a small subset of tokens, resulting in a low entropy. Conversely, when uncertain, the model distributes probabilities more evenly across the vocabulary, producing higher entropy values.

We hypothesize that high-entropy responses generally correspond to questions where the model lacks the necessary knowledge or reasoning capacity. By analyzing the relationship between entropy and model performance across domains and difficult reasoning complexity, we can identify specific conditions under which LLMs are most prone to overconfidence or appropriate uncertainty calibration.

3.2.2. *MASJ Uncertainty Estimation.* The model-as-judge (MASJ) paradigm represents an alternative approach to uncertainty estimation that leverages an LLM's ability to evaluate a given response. Unlike entropy-based methods that rely on a probability distribution across output tokens, MASJ employs prompt-based techniques.

Our implementation of MASJ follows a methodology similar to the approach described in the MT-Bench evaluation framework [11]. This approach extracts explicit confidence assessments that may capture uncertainty beyond what is reflected in the raw probability distribution: including broader reasoning about the model's knowledge boundaries, the ambiguity of questions, and the model's awareness of its own limitations. In instances where MT-Bench assigned a score below 8 on a scale of 1 to 10, we decided to regenerate the estimate.

4. Results

4.1. Experimental setup.

4.1.1. Datasets.
MMLU-Pro. In this study, we utilize the MMLU-Pro dataset [24], a more challenging and robust variant of the well-known Massive Multitask Language Understanding (MMLU) benchmark. The original MMLU dataset was introduced by Hendrycks et al. [1] and has been widely used to evaluate language model performance across diverse domains. MMLU-Pro builds upon this benchmark by incorporating a filtered and refined subset of the original questions, also introducing newly generated question-answer pairs to enhance difficulty and robustness.

Category	Sample Count
Mathematics	1351
Physics	1299
Chemistry	1132
Law	1101
Engineering	969
Economics	844
Health	818
Psychology	798
Business	789
Biology	717
Philosophy	499
Computer Science	410
History	381
Other	924

Table 1. MMLU-Pro Category Distribution.

The dataset consists of approximately 12000 multiple-choice questions, with 10 answer choices per question in the majority of cases (approximately 10000 instances), while the remaining 2000 instances contain between 3 and 9 answer options. The dataset spans 14 distinct scientific categories, covering a wide range of disciplines with the number of topics by discipline provided in Table 1. Examples of questions from different categories are available in the Appendix B.

The motivation for using MMLU-Pro for this problem is its increased difficulty and improved robustness compared to the original MMLU dataset. The official reports for Phi-4 and Mistral-Small-24B also used the extended dataset to mitigate data leakage. Thus, MMLU-Pro is a reliable and appropriate benchmark for evaluating our models' generalization and reasoning capabilities.

4.1.2. Models. For QA and evaluation of the entropy, we use decoder-only LLMs Mistral (*Mistral-Small-24B-Base-2501*) [28], Phi-4 [27] and Qwen models [15].

4.2. Reasoning and knowledge-based complexity. We hypothesize that different questions validate different abilities of a model. Some focus
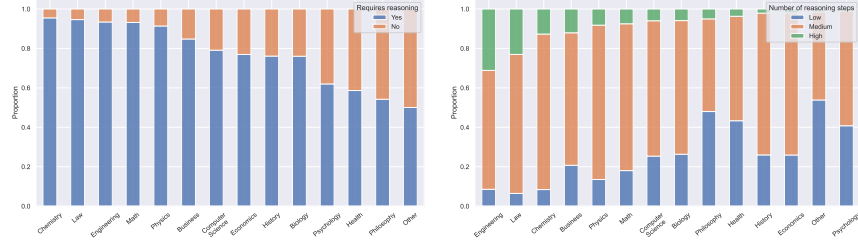
(a) Distribution of questions that require complex reasoning

(b) Distribution of the required number of reasoning steps

Figure 2. Estimation by MASJ of the required reasoning amount. Better to view in zoom.

more on the requirement to know specific things, while others consider a model's ability to reason within a specific topic.

To obtain these properties for specific questions, we obtained estimates via the model-as-Judge. Figures 2 present the question type by the requirement of reasoning and the estimation of the required number of reasoning steps.

The distribution of questions over topics is diverse, with the reasoning requirement attaining a maximum share of about 0.9 for engineering and only 0.5 for philosophy. The number of required steps also varies: again, for engineering, the model estimates the number of questions with a high number of required steps, but for psychology and philosophy, this number is almost zero.

4.3. Distribution of uncertainty measures. For Phi-4 and Qwen, we present the distribution of entropy values in Figures 3. The distribution of entropy is presented for correct and incorrect answers by a model. We can see that near-zero entropy values are observed much more often for correct answers for Phi-4 and Qwen. Thus, entropy only partially explains why the answers from a model are wrong, and the quality of this score depends on the model used.

Both the numerical and nominal Model-as-judge scores also relate little to the complexity, as the corresponding measured ROC AUC values are 0.49 for both models, indicating nearly random predictions. Thus, MASJ
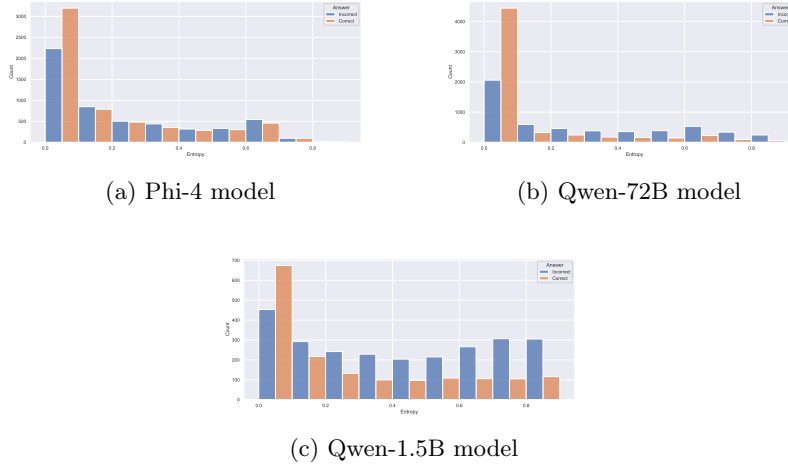
(a) Phi-4 model

(b) Qwen-72B model

(c) Qwen-1.5B model

Figure 3. Entropy distribution of answers. Best viewed when zoomed in.

uncertainty estimates focus on different aspects of an LLM's confidence, which is irrelevant to question complexity.

4.4. Complexity prediction by category. Figure 4 shows the ROC-AUC values for entropy for all four models. Values above or below 0.5 indicate that the prediction is better than random. For most topics, the values are well above this threshold. In psychology and biology, we observe ROC-AUC values of 0.77/0.83 (Qwen 72B) compared to 0.61/0.73 (Phi-4) and 0.65/0.65 (Mistral), with these improvements correlating across models.

4.5. Complexity prediction by reasoning requirement. Figure 5 present similar results, but for the split of the questions by reasoning complexity. Table 2 shows the aggregated ROC AUC scores across categories and reasoning levels.

We see meaningful results of using the reasoning estimates for all questions without categorization by subject. Low entropy is a marginally better predictor of the truthfulness of the answers to questions that did not require complex reasoning and vice versa for models. We also observe the estimate of the number of reasoning steps as a less reliable metric, with
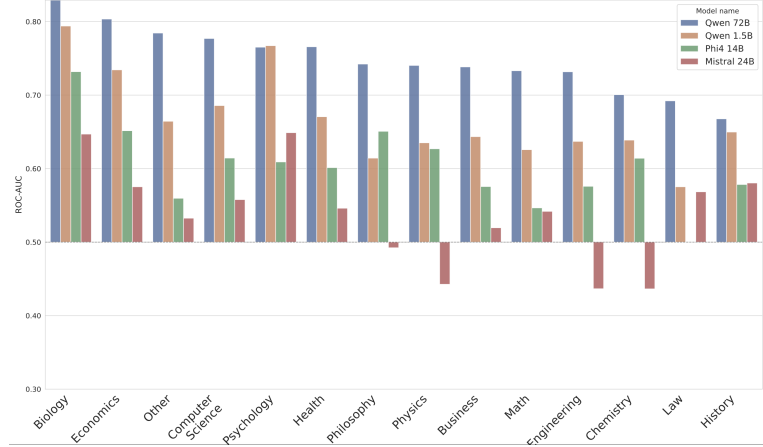
Figure 4. ROC-AUC for error prediction by subject for four different LLMs.

Mistral favoring a higher number of reasoning steps. However, we can attribute this to the lower precision of the estimation provided by the MASJ approach.

4.6. Calibration assessment via entropy. Figure 6 illustrates calibration curves for four language models, computed using inverted normalized entropy. The analysis reveals systematic deviations between self-reported certainty and empirical accuracy, with distinct patterns across architectures.

Mistral exhibits unstable calibration dynamics, characterized by erratic accuracy in low-confidence regions. Despite moderate confidence estimates, its accuracy fluctuates unpredictably, even peaks anomalously in the lowest entropy bin. This contrasts with its pronounced underperformance in high-confidence regimes, where empirical accuracy consistently trails reported certainty, which is a limitation indicative of architectural weaknesses in uncertainty quantification.

Qwen 72B, despite its large scale, demonstrates inconsistent calibration. High-confidence predictions show weak alignment with accuracy, while mid-range entropy bins display substantial confidence-accuracy gaps.

Category	Phi-4	Mistral	Qwen 1.5B	Qwen 72B
Biology	**0.73**	**0.65**	**0.79**	**0.83**
Economics	0.65	0.58	0.73	0.80
Philosophy	0.65	*0.49*	0.61	0.74
Physics	0.63	**0.44**	0.64	0.74
Computer Science	0.61	0.56	0.69	0.78
Psychology	0.61	**0.65**	0.77	0.77
Chemistry	0.61	0.44	0.64	0.70
Health	0.60	0.55	0.67	0.77
Business	0.58	0.52	0.64	0.74
History	0.58	0.58	0.65	*0.67*
Engineering	0.58	0.44	0.64	0.73
Other	0.56	0.53	0.66	0.78
Math	0.55	0.54	0.63	0.73
Law	*0.50*	0.57	*0.58*	0.69
No Reasoning	**0.59**	**0.59**	**0.72**	**0.79**
Yes Reasoning	0.56	*0.52*	0.68	0.76
Reasoning (Low)	**0.59**	0.58	0.71	**0.79**
Reasoning (Med)	0.57	*0.52*	0.69	0.76
Reasoning (High)	*0.53*	0.53	*0.61*	*0.73*
Overall	0.58	0.52	0.70	0.77

Table 2. ROC AUC performance comparison across categories and models.

Qwen 1.5B emerges as an outlier, surpassing larger counterparts in calibration quality. Its confidence-accuracy relationship progresses monotonically across entropy bins, with minimal extreme deviations.

These results underscore systematic overconfidence as a pervasive issue, especially in high-certainty regimes, where all models significantly overestimate their reliability.
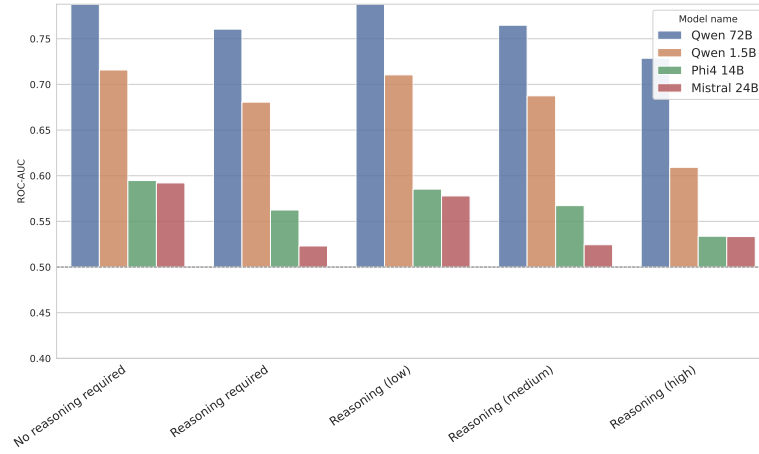
Figure 5. ROC-AUC by reasoning for four different LLMs.

5. Conclusion and discussion

Risks in LLMs involve possible errors in answering questions with high confidence. One possible way to mitigate this problem is by highlighting the questions that, with high probability, would be answered incorrectly. We suggest using uncertainty estimations for LLMs to identify possible errors. For a more detailed analysis, the paper examines various subject areas, question types, and specific requirements for knowledge and reasoning. We explored how different uncertainty estimates for LLMs highlight potential errors of a model and examined various aspects of problems in QA to identify what types of problems a specific uncertainty estimate can highlight. Our focus was on reliable and popular methods, including a model as a judge and an entropy-based approach.

Experimental evidence suggests that different uncertainty estimates perform differently across diverse subject areas. Our findings also demonstrate that the key factor is the varying composition of question sets within a subject area collected for a specific subject area: entropy performs much better when no reasoning or little reasoning is required. Thus, it primarily highlights the lack of the required knowledge. The conclusion also depends on
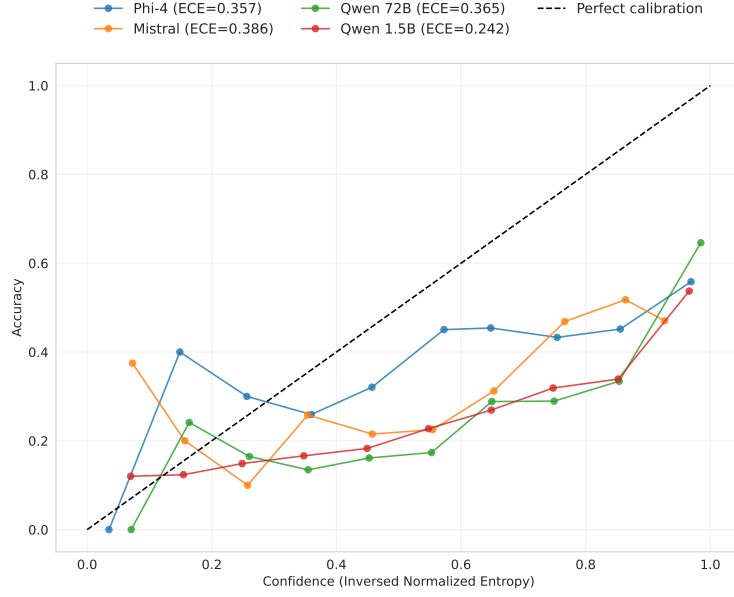
Figure 6. Calibration of LLMs based on their entropy.

the model used: for reasoning-oriented, such as Qwen-72B, entropy better highlights the possibility of an error compared to Mistral.

Appendix A. Prompts

This section provides prompts used for Model-as-judge approaches: for numerical estimation of complexity for MASJ in Table 3, for nominal estimation of complexity for MASJ in Table 4, and for reasoning level estimate in Table 5.

Appendix B. Question examples with answers

This section provides examples of questions included in the MMLU-Pro dataset on Law, Psychology, and Engineering in Tables 6, 7, and 8, correspondingly.

Table 3. Prompt for Numerical model-as-judge

You are an expert in the topic of the question. Please act as an impartial judge and evaluate the complexity of the multiple-choice question with options below. Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, you must not answer the question. You must rate the question complexity as a number from 0 to 1 following the following scale as a reference:
high_school_and_ easier - 0.0-0.22,
undergraduate_easy - 0.2-0.4,
undergraduate_hard - 0.4-0.6,
graduate - 0.6-0.8,
postgraduate - 0.8-1.0.
You must return the complexity by strictly following this format: " [[complexity]]",
for example: "Your explanation... Complexity: [[0.55]]", which corresponds to a hard question at the undergraduate level.

Table 4. Prompt for Nominal model-as-judge

You are an expert in the topic of the question. Please act as an impartial judge and evaluate the complexity of the multiple-choice question with options below. Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, you must not answer the question. You must rate the question complexity by strictly following the scale:
high_school_and_easier,
undergraduate_easy,
undergraduate_hard,
graduate,
postgraduate.
You must return the complexity by strictly following this format: "[[complexity]]",
for example: "Your explanation... Complexity: [[undergraduate]]", which corresponds to the undergraduate level.

Appendix C. Entropy Distributions

This section complements the analysis in Section 4.3 by providing full entropy distributions for the Qwen model family. Figure 7 reveals two key trends:

Table 5. Prompt for Reasoning level estimate

You are an expert in the topic of the question. Please act as an impartial judge and evaluate the complexity of the multiple-choice question with options below.

Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, you must not answer the question. You must rate the question complexity by strictly following the criteria:

1) [[Requires knowledge]] - do we need highly specific knowledge from the domain to answer this question? Valid answers: yes, no;

2) [[Requires reasoning]] - do we need complex reasoning with multiple logical steps to answer this question? Valid answers: yes, no;

3) [[Number of reasoning steps]] - how many reasoning steps do you need to answer this question? Valid answers: low, medium, high.

Your answer must strictly follow this format: "[[Requires knowledge: answer]] [[Requires reasoning: answer]] [[Number of reasoning steps: answer]]".

Example 1: "Your explanation... [[Requires knowledge: yes]] [[Requires reasoning: no]] [[Number of reasoning steps: low]]".

Example 2: "Your explanation... [[Requires knowledge: no]] [[Requires reasoning: yes]] [[Number of reasoning steps: high]]".

Example 3: "Your explanation... [[Requires knowledge: yes]] [[Requires reasoning: yes]] [[Number of reasoning steps: medium]]".

Example 4: "Your explanation... [[Requires knowledge: no]] [[Requires reasoning: no]] [[Number of reasoning steps: low]]".

Table 6. Example of a question from Law category

Which of the following criticisms of Llewellyn's distinction between the grand and formal styles of legal reasoning is the most compelling?
1. There is no distinction between the two forms of legal reasoning.
2. Judges are appointed to interpret the law, not to make it.
3. It is misleading to pigeon-hole judges in this way.
4. Judicial reasoning is always formal.

Model scale correlates with entropy separation: Larger models (72B, 32B) show clearer divergence between correct (low entropy) and incorrect (higher entropy) predictions, while smaller variants (3B, 0.5B) exhibit overlapping distributions.

Table 7. Example of a question from Psychology category

A 66-year-old client who is depressed, has rhythmic hand movements, and has a flattened affect is probably suffering from:
1. Huntington's disease
2. Creutzfeldt-Jakob disease
3. Multiple Sclerosis
4. Alzheimer's disease
5. Parkinson's disease
6. Vascular Dementia
7. Frontotemporal Dementia
8. Schizophrenia
9. A right frontal lobe tumor
10. Bipolar Disorder

Table 8. Example of a question from Engineering category

A cumulative compound motor has a varying load upon it which requires a variation in armature current from 50 amp to 100 amp.
If the series-field current causes the air-gap flux to change by 3 percent for each 10 amp of armature current, find the ratio of torques developed for the two values of armature current.
1. 2.26
2. 0.66
3. 3.95
4. 1.00
5. 2.89
6. 1.75
7. 4.12
8. 1.15
9. 0.87
10. 3.40

References

1. D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt, *Measuring Massive Multitask Language Understanding*, in: International Conference on Learning Representations, 2021.
2. P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, Ø. Tafjord, *Think You Have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge*, ArXiv preprint arXiv:1803.05457 (2018).
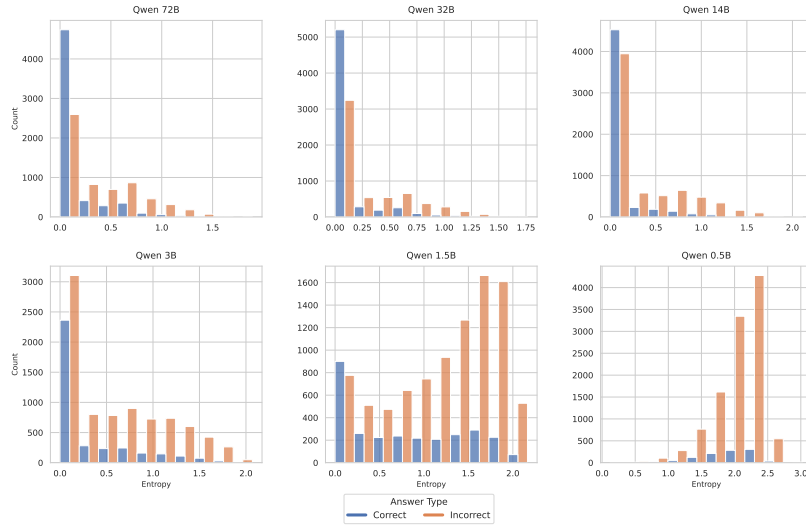
Figure 7. Distribution of entropy values for Qwen models (72B, 32B, 14B, 3B, 1.5B, 0.5B), stratified by answer correctness. For larger models (72B, 32B), correct answers (blue) exhibit a pronounced left skew toward low entropy, indicating higher confidence in accurate predictions. Smaller models (1.5B, 0.5B) show flatter distributions, with less separation between correct and incorrect (orange) entropy values. Notably, the 72B variant demonstrates near-zero entropy peaks for correct responses, aligning with findings from Section 4.2.

3. D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman, *GPQA: A Graduate-Level Google-Proof Q&A Benchmark*, ArXiv preprint arXiv:2311.12022 (2023). Available: `https://arxiv.org/abs/2311.12022`.

4. A. Elhady, E. Agirre, and M. Artetxe, *WiCkeD: A Simple Method to Make Multiple Choice Benchmarks More Challenging*, ArXiv preprint arXiv:2502.18316 (2025).

5. A. Rogers, M. Gardner, I. Augenstein, *QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension.* — ACM Computing Surveys **55**(10) (2023), 1–45.

6. F. Sizov, C. España-Bonet, J. van Genabith, R. Xie, K. Dutta Chowdhury, *Analysing Translation Artifacts: A Comparative Study of LLMs, NMTs, and Human Translations*, in: Proceedings of the Ninth Conference on Machine Translation, 2024, pp. 1183–1199.

7. Y. Abbasi Yadkori, I. Kuzborskij, A. György, C. Szepesvári, *To Believe or Not to Believe Your LLM: Iterative Prompting for Estimating Epistemic Uncertainty*, Advances in Neural Information Processing Systems **37** (2025), 58077–58117.

8. S. Chern, E. Chern, G. Neubig, P. Liu, *Can Large Language Models Be Trusted for Evaluation? Scalable Meta-Evaluation of LLMs as Evaluators via Agent Debate*, ArXiv preprint arXiv:2401.16788 (2024). Available: `https://arxiv.org/abs/2401.16788`.

9. O. Shorinwa, Z. Mei, J. Lidard, A. Z. Ren, A. Majumdar, *A Survey on Uncertainty Quantification of Large Language Models: Taxonomy, Open Research Challenges, and Future Directions*, ArXiv preprint arXiv:2412.05563 (2024). Available: `https://arxiv.org/abs/2412.05563`.

10. S. Buyske, *Optimal Design in Educational Testing*, in: Applied Optimal Designs, Wiley Online Library, 2005, pp. 1–19.

11. L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, *et al.*, *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*, Advances in Neural Information Processing Systems **36** (2023), 46595–46623.

12. Y. Wan, Y. Liu, A. Ajith, C. Grazian, B. Hoex, W. Zhang, C. Kit, T. Xie, I. Foster, *SciQAG: A Framework for Auto-Generated Science Question Answering Dataset with Fine-Grained Evaluation*, ArXiv preprint arXiv:2405.09939 (2024).

13. M. Dahl, V. Magesh, M. Suzgun, D. E. Ho, *Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. —* Journal of Legal Analysis **16**(1) (2024), 64–93. Available: `http://dx.doi.org/10.1093/jla/laae003`.

14. L. Kuhn, Y. Gal, and S. Farquhar, *Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation*, ArXiv preprint arXiv:2302.09664 (2023).

15. Qwen *et al.*, *Qwen2.5 Technical Report*, ArXiv preprint arXiv:2412.15115 (2025). Available: `https://arxiv.org/abs/2412.15115`.

16. R. Yang, C. Zhang, Z. Zhang, X. Huang, S. Yang, N. Collier, D. Yu, D. Yang, *LoGU: Long-Form Generation with Uncertainty Expressions*, ArXiv preprint arXiv:2410.14309 (2024).

17. P. Chhikara, *Mind the Confidence Gap: Overconfidence, Calibration, and Distractor Effects in Large Language Models*, ArXiv preprint arXiv:2502.11028 (2025).

18. Z. Wang, J. Duan, C. Yuan, Q. Chen, T. Chen, Y. Zhang, R. Wang, X. Shi, K. Xu, *Word-Sequence Entropy: Towards Uncertainty Estimation in Free-Form Medical Question Answering Applications and Beyond*, ArXiv preprint arXiv:2402.14259 (2024).

19. A. P. Gema, J. O. J. Leang, G. Hong, A. Devoto, A. C. M. Mancino, R. Saxena, X. He, Y. Zhao, X. Du, M. R. G. Madani, C. Barale, R. McHardy, J. Harris, J. Kaddour, E. van Krieken, P. Minervini, *Are We Done with MMLU?*, ArXiv preprint arXiv:2406.04127 (2025).

20. E. Hüllermeier and W. Waegeman, *Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods.* — Machine Learning **110**(3) (2021), 457–506.

21. S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, N. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson, *et al.*, *Language Models (Mostly) Know What They Know*, ArXiv preprint arXiv:2207.05221 (2022).

22. E. Fadeeva, A. Rubashevskii, A. Shelmanov, S. Petrakov, H. Li, H. Mubarak, E. Tsymbalov, G. Kuzmin, A. Panchenko, T. Baldwin, *Fact-Checking the Output of Large Language Models via Token-Level Uncertainty Quantification*, in: Findings of the Association for Computational Linguistics: ACL 2024, 2024, pp. 9367–9385.

23. L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, H. Wang, Q. Feng, B. Qin, *et al.*, *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions.* — ACM Transactions on Information Systems **43**(2) (2025), 1–55.

24. Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, *et al.*, *MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark*, Advances in Neural Information Processing Systems **37** (2024), 95266–95290.

25. M. Gabburo, N. Jedema, S. Garg, L. Ribeiro, A. Moschitti, *Measuring Question Answering Difficulty for Retrieval-Augmented Generation*, 2024.

26. E. Fadeeva, R. Vashurin, A. Tsvigun, A. Vazhentsev, S. Petrakov, K. Fedyanin, D. Vasilev, E. Goncharova, A. Panchenko, M. Panov, *LM-Polygraph: Uncertainty Estimation for Language Models*, in: Proceedings of EMNLP 2023: System Demonstrations, 2023, pp. 446–461.

27. M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, *et al.*, *Phi-4 Technical Report*, ArXiv preprint arXiv:2412.08905 (2024).

28. A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, *et al.*, *Mistral 7B*, ArXiv preprint arXiv:2310.06825 (2023). Available: `https://arxiv.org/abs/2310.06825`.

Skolkovo Institute
of Science and Technology (Skoltech)
Moscow, Russia;
National Research University
Higher School of Economics
Moscow, Russia

E-mail: `petr.sychev@skoltech.ru, daniel.vyazhev@skoltech.ru,`
`khalafyan.ea@phystech.edu`

Skolkovo Institute
of Science and Technology (Skoltech)
Moscow, Russia

E-mail: `Andrey.Goncharov@skoltech.ru, a.zaytsev@skoltech.ru`