

UBENCH: Benchmarking Uncertainty in Large Language Models with Multiple Choice Questions

Xunzhi Wang^{1,4*} Zhuowei Zhang^{1,4*} Gaonan Chen¹ Qiongyu Li¹ Bitong Luo¹
Zhixin Han¹ Haotian Wang¹ Zhiyu Li² Hang Gao³ Mengting Hu^{1,4†}

¹College of Software, Nankai University

²Institute for Advanced Algorithms Research (Shanghai)

³College of Artificial Intelligence, Tianjin University of Science and Technology

⁴Tianjin Key Laboratory of Software Experience and Human Computer Interaction

xunzhi, zhuoweizhang@mail.nankai.edu.cn lizy@iaar.ac.cn mthu@nankai.edu.cn

Abstract

Despite recent progress in systematic evaluation frameworks, benchmarking the uncertainty of large language models (LLMs) remains a highly challenging task. Existing methods for benchmarking the uncertainty of LLMs face three key challenges: the need for internal model access, additional training, or high computational costs. This is particularly unfavorable for closed-source models. To this end, we introduce **UBENCH**, a new benchmark for evaluating the uncertainty of LLMs. Unlike other benchmarks, **UBENCH** is based on confidence intervals. It encompasses 11,978 multiple-choice questions spanning knowledge, language, understanding, and reasoning capabilities. Based on this, we conduct extensive experiments. This includes comparisons with other advanced uncertainty estimation methods, the assessment of the uncertainty of 20 LLMs, and an exploration of the effects of Chain-of-Thought (CoT) prompts, role-playing (RP) prompts, and temperature on model uncertainty. Our analysis reveals several crucial insights: 1) *Our confidence interval-based methods are highly effective for uncertainty quantification*; 2) *Regarding uncertainty, outstanding open-source models show competitive performance versus closed-source models*; 3) *CoT and RP prompts present potential ways to improve model reliability, while the influence of temperature changes follows no universal rule*. Our implementation is available at <https://github.com/Cyno2232/UBENCH>.

1 Introduction

In recent years, significant progress has been made in the development of large language models (LLMs), including ChatGPT (Wu et al., 2023), Llama (Dubey et al., 2024), Qwen (Yang et al., 2024; Team, 2024), etc. These advancements have

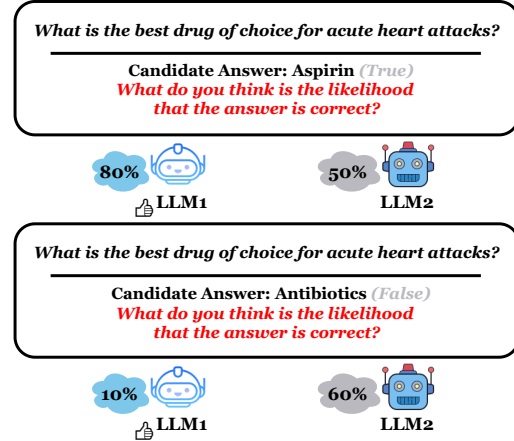


Figure 1: In the context of different candidate answers to the same question, LLMs display different levels of confidence (in other words, uncertainty). Note that LLMs may exhibit consistent levels of confidence for either the wrong answer or the right answer, which we do not want.

not only piqued the intense interest of both the industrial and academic sectors (Zhao et al., 2023), but have also spurred a wave of exploration and research. In this fast-paced development, comprehensively evaluating LLMs’ capabilities is essential. This evaluation helps us understand LLMs better, improve their performance, and apply them effectively in real-world scenarios, unlocking their potential for innovation across different fields.

Much effort has been devoted to establishing accurate, authentic, and equitable evaluation systems for LLMs, such as C-Eval (Huang et al., 2023b) and MT-Bench (Zheng et al., 2023b). However, these benchmarks often focus solely on the performance of models across various tasks, neglecting the credibility of their answers. In some situations, models may not be certain about the confidence level associated with their provided answers, potentially leading to misunderstandings or even harm. Specifically, LLMs are prone to generating misinformation without warning, which may mani-

*Equal contribution.

†Corresponding author.

Method	Close?	Single?	Close&Single?
Xiong et al. (2023)	✓	✗	✗
Ye et al. (2024)	✓	✓	✗
UBENCH (Ours)	✓	✓	✓

Table 1: Comparison of different training-free uncertainty estimation methods: “*Close?*” indicates whether the method can be applied to closed-source models, “*Single?*” indicates whether only a single inference is needed, and “*Close&Single?*” indicates whether the method can be applied to closed-source models with single-inference requirements.

fest as hallucinations (Huang et al., 2023a), biases (Felkner et al., 2023), or disinformation (Lucas et al., 2023). In this scenario, we hope to know the confidence level of the response to decide whether to trust the information or suggestions provided by LLMs. Specifically, as shown in Figure 1, two LLMs exhibit different confidence levels for various candidate answers to the same question “*What do you think is the likelihood that the answer is correct?*”. Clearly, the performance of LLM1 aligns more closely with human expectations. Inspired by that, some research has initiated a shift towards assessing the uncertainty of LLMs (Xiong et al., 2023; Ye et al., 2024; Xu et al., 2024).

However, many existing uncertainty estimation methods for LLMs are severely restricted. They confront challenges like the need for additional training, high computational costs, and access to internal model details. This is especially tough for closed-source models. These limitations hamper the application of uncertainty estimation, and thus, a comprehensive and accurate assessment of model reliability and performance. For instance, Xu et al. (2024) proposed SaySelf, a two-stage training framework that requires extra training, but it is difficult to be used for closed-source models. Xiong et al. (2023) use prompts to get model confidence, but it needs multiple sampling, which is costly and not ideal in some cases¹. Ye et al. (2024)’s logits-based method for uncertainty assessment needs internal model access, so it only works for open-source models. Approximating it for closed-source models via multiple sampling, like Xiong et al. (2023)’s method, is inefficient and resource-consuming.

To address these challenges, we introduce **UBENCH**, a new benchmark. Distinct from previous efforts, it is founded on confidence intervals,

which not only facilitates automated evaluation but also enables more effective quantification of uncertainties. Comprising four question categories, UBENCH encompasses a total of 11,978 (~12K) multiple-choice questions, including both positive and negative samples. Notably, to improve efficiency while maintaining accuracy, some of the negative samples were generated by GPT-4² and subsequently underwent manual review and refinement (refer to §F for details). Designed to accurately assess the reliability of LLMs, UBENCH is resource-efficient, requiring no extra training, only one inference, and no internal model parameters, filling a key gap in LLM evaluation. A comparison of UBENCH with other training-free benchmark uncertainty methods is presented in Table 1³.

Based on this, we compare the uncertainty quantification method based on confidence intervals with other powerful methods. We also evaluate 20 state-of-the-art open-source and closed-source models that cover several series and sizes. In addition, we explore the factors that potentially influence model uncertainty, including Chain-of-Thought (COT) (Wei et al., 2022) prompts, role-playing (RP) (Shao et al., 2023) prompts, and the impact of temperature on model uncertainty. Our main contributions and findings are summarised below:

- We propose UBENCH, a new confidence interval-based uncertainty evaluation benchmark for LLMs. It consists of approximately 12K questions, covering four categories: knowledge, language, understanding, and reasoning.
- We conduct a comparison of our confidence interval-based uncertainty estimation method with other LLM uncertainty estimation methods and achieve superior results.
- We utilize UBENCH to conduct tests on 20 widely-adopted LLMs. In general, excellent open-source and closed-source LLMs display comparable degrees of reliability.
- Further exploratory analysis reveals that CoT and RP prompts are potential methods for enhancing model reliability, while there is no general rule regarding the impact of temperature. We analyze the raw responses of LLMs

¹Please refer to §5 for details.

²The version is 1106-preview.

³Please refer to §4.2 for more results.

and offer possible explanations for each effect. This analysis paves the way for broader downstream applications of LLMs.

2 Related Work

2.1 Benchmark for LLMs

Previous benchmarks can generally be categorized into two types: generic task benchmarks and task-specific benchmarks (Chang et al., 2023). Generic task benchmarks are used to evaluate the generic capabilities of LLMs for several tasks (e.g., sentiment analysis, natural language inference, machine translation, etc.), including GLUE (Wang et al., 2018), MMLU (Hendrycks et al., 2020), MT-Bench (Zheng et al., 2023b), BIG-bench (Srivastava et al., 2022), PromptBench (Zhu et al., 2023), PandaLM (Wang et al., 2023), MT-Eval (Kwan et al., 2024) and so on. Bai et al. (2024) present ChatABSA, specifically designed to assess LLMs’ performance on aspect-based sentiment analysis (ABSA). ToMBench (Chen et al., 2024) is used to benchmark the Theory of Mind in large language models. C-Eval (Huang et al., 2023b) is the first benchmark for broadly assessing a model’s Chinese knowledge and reasoning ability.

2.2 Uncertainty Estimation for LLMs

Recent research in uncertainty estimation for LLMs has explored diverse methodologies, broadly categorized into logits-based, verbal-based, and training-based approaches. Logits-based methods leverage model output and parameters to assess uncertainty, though their applicability is constrained. For instance, Kuhn et al. (2022) cluster semantically similar answers to evaluate uncertainty, while Duan et al. (2023) incorporate sentence relevance for uncertainty estimation; however, these methods are unsuitable for closed-source LLMs. Although Ye et al. (2024)’s method can approximate results by repeatedly sampling the closed-source model, it exacerbates resource consumption. Verbal-based confidence assessment methods require models to output confidence scores in natural language, as proposed by Lin et al. (2022a). While Tian et al. (2023) observe better calibration in RLHF-LLMs’ confidence scores, Xiong et al. (2023) highlight overconfidence issues and advocate combining verbalized confidence with consistency. Furthermore, the applicability of these two evaluation methods is still limited to fact-recall tasks or multi-step inference requirements. Training-based methods

often require retraining or fine-tuning, limiting scalability. Examples include UaIT (Liu et al., 2024), which aligns uncertainty perception with probabilistic outputs, and R-Tuning (Zhang et al., 2024) for enhancing refusal capability on uncertain queries. Moreover, Bakman et al. (2024) propose MARS that takes into account the semantic contribution of each token to evaluate the correctness of the generated content. Xu et al. (2024) mitigate this with Sayself, a reinforcement learning framework for calibrated confidence generation. For more analysis, please refer to §A.


Unlike all the aforementioned works, our proposed UBENCH introduces a novel confidence interval-based sampling approach that achieves three key advantages: requiring only single inference, without requiring additional model training, and applicability to closed-source models. Furthermore, while most existing methods have only been validated on a limited number of tasks, UBENCH serves as a multi-dimensional, multi-task benchmark designed to comprehensively evaluate the uncertainty of LLMs.


3 The UBENCH


The overall construction and evaluation process of UBENCH is shown in Figure 2. Overall, UBENCH includes 4 categories, including tasks such as reading comprehension, QA, mathematical reasoning, and more. We provide detailed information on categories, data construction, and the design of prompts.

3.1 Problem Categories

We adopt the competency categorization from OpenCompass’s LLM assessment framework (Contributors, 2023), resulting in the following four categories:

 **Language.** This category primarily evaluates the reliability of LLMs in dealing with language category tasks, encompassing tasks such as syntactic analysis, semantic matching, word sense disambiguation, and coreference resolution.

 **Knowledge.** This type of question mainly evaluates the reliability of LLMs when dealing with knowledge-based tasks such as common sense and facts, covering contents in areas like health, law, finance, politics, and history.

 **Understanding.** The aim is to evaluate LLMs’ reliability in understanding-related tasks. These

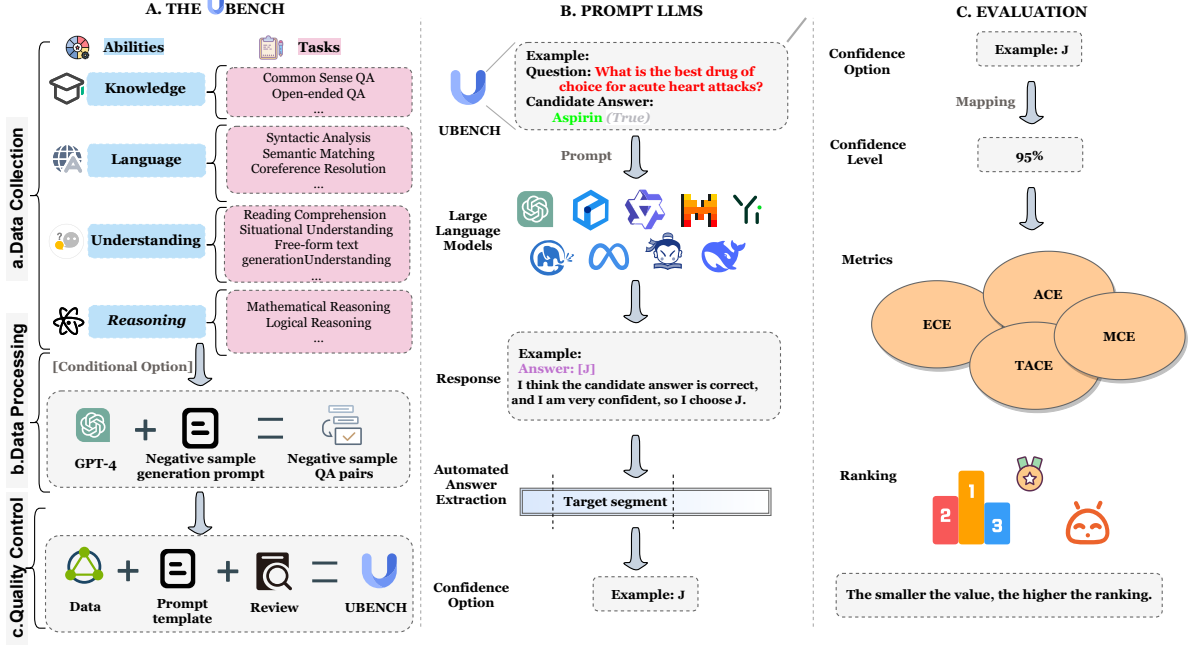


Figure 2: Construction process of UBENCH and systematic, automated LLM uncertainty evaluation framework. The data sources of UBENCH are from multiple types of public datasets, with processed into the uncertainty evaluation format and carefully controlled quality. Then UBENCH is leveraged to compare the reliability of typical open-source and closed-source LLMs with 4 evaluation metrics.

tasks encompass various aspects, including multi-task language understanding, reading comprehension, situational awareness, behavioral interpretation, dialogue understanding, and so on.

Reasoning. Unlike other categories, this one aims to evaluate LLMs’ reliability in mathematical reasoning, logical reasoning, and related areas.

3.2 Data Construction

The data sources of UBENCH originate from various existing datasets and are specially processed into formats suitable for evaluation, with careful quality control.

Data Collection. To comprehensively and efficiently construct the dataset, we conduct random sampling from 24 open-source datasets⁴. Each dataset contributes 500 data points, as shown in Figure 3. The datasets sampled for each category are listed in the Appendix D.1.

Data Processing. Intuitively, a reliable LLM yields lower uncertainty for correct answers and higher uncertainty for incorrect ones. Therefore, different from previous works, we reformat the collected data into positive and negative samples, respectively. A positive sample indicates that the

correct answer is used, and a negative sample indicates that an incorrect answer is randomly selected as the answer. For datasets without candidate wrong answers, we prompt GPT-4-1106-preview using a one-shot approach to generate incorrect answers similar to the correct ones. For details on the prompts, please refer to Appendix §B.1.

Quality Control. On one hand, for the generated negative samples, we conduct a comprehensive check of all data. Failed data are regenerated until they meet the requirements, ensuring data integrity for subsequent analysis. On the other hand, to ensure dataset quality, a strict review process is implemented. Each sample is first reviewed by two authors. Only when both concur that there are no issues is the sample approved. In case of disagreement, a third author joins the review to reach a consensus. Through this process, the sample approval rate has reached 99.82%. For more details on the data validation process, refer to Appendix §F.

3.3 Prompt Design

Like previous studies (Zhang et al., 2023; Zheng et al., 2023b), we evaluate the reliability of LLMs based on prompt engineering. Initially, we design the prompt with a data sample and all its answers, following instructions like *choose the correct answer and also yield the uncertainty*. We call this the

⁴Most of the datasets used in this study are sourced from HuggingFace.

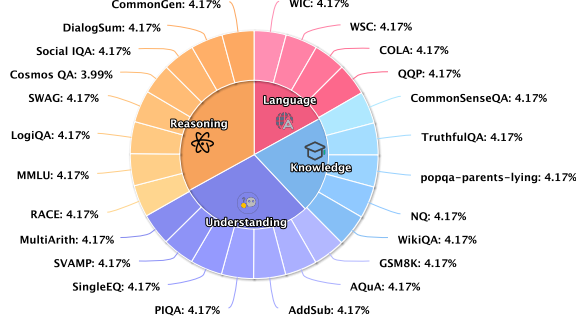


Figure 3: UBENCH covers 4 categories: Language, Knowledge, Understanding, and Reasoning, with a total of 11,978 ten-item multiple-choice questions.

standard prompt. However, most LLMs struggle to output a consistent format of uncertainty values, even some of them do not understand the instructions (Kuhn et al., 2023), which makes it hard to evaluate. Therefore, we formulate the prompt as providing a correct answer or incorrect answer to LLMs, asking them to choose the uncertainty towards the given answer from ten intervals. In other words, the problem is reformulated from yielding real number uncertainty to a multiple-choice question. We refer to this as the confidence interval-based prompt. For empirical validation on prompt design, please refer to §B.2.

Specifically, we create 10 confidence intervals, each representing a 10% confidence range, e.g. 0-10% for the first interval, 10-20% for the second, and so on. Different annotators having varying criteria for confidence, results in annotation errors, making it difficult to construct few-shot prompts. We choose a zero-shot approach to construct the prompt instead. The prompts contain role-playing (RP) prompts, task declarations, and a step-by-step problem disassembly Chain-of-Thought (CoT) prompt, output format specification, confidence interpretation, sample QA, and confidence interval options. Please refer to Table 9 for details.

4 Experiments

4.1 Experimental Settings

Evaluated Models. Based on UBENCH, we test 20 popular LLMs, covering a wide range of open and closed source LLMs, as shown in Table 13.

Evaluation Metrics. To evaluate the reliability of LLMs, we use four evaluation metrics: Expected Calibration Error (ECE), which measures the difference between model confidence and accuracy (Guo et al., 2017); Average Calibration Er-

ror (ACE), which adjusts for different confidence ranges; Maximum Calibration Error (MCE), indicating the worst-case confidence error (Guo et al., 2017); and Thresholded Average Calibration Error (TACE) for high-risk scenarios, with a threshold set at 0.5. More calculation details are shown in Appendix §D.3. All metrics are better with lower values and are presented as percentages in our work.

Experimental Process. To improve reproducibility and fairness, we set the temperature of all LLMs to 0.001, except in cases where the model or API restricts this (e.g., GLM4-flash), in which case it is set to 0.01. All other parameters remain at their default values.

For all responses of LLMs, we map the chosen options to their respective confidence values. The confidence value for option \mathcal{X} can be calculated as follows:

$$Conf(\mathcal{X}) = (\mathcal{O}(\mathcal{X}) - \mathcal{O}(A)) \times 0.1 + 0.05, \quad (1)$$

where $Conf(\mathcal{X})$ is the confidence value for option \mathcal{X} and $\mathcal{O}(\cdot)$ represents the function that converts letter options into ASCII codes. “A” indicates the first option from ten choices. The value of \mathcal{X} ranges from A to J. As such, the value of $Conf(\mathcal{X})$ is approximated to the midrange of each confidence interval, such as 5% for interval 0%-10%. Finally, $Conf(\mathcal{X})$ is ranging from 0.05 to 0.95.

4.2 Compared to Other Methods

The comparison of UBENCH, which uses a confidence interval-based uncertainty estimation method, with other uncertainty estimation methods is shown in Table 2. More experimental details are provided in Appendix §C. We obtain the performance of UBENCH and four baseline methods on the Cosmos QA and SWAG datasets using five open-source and closed-source LLMs, by randomly sampling three times with 100 data points each. Experimental results show our method’s superior performance, with a t-test indicating 77.27% of metrics’ mean values differ significantly ($p < 0.05$) from others. However, we also notice that UBENCH performs poorly when testing GPT-3.5 on the Cosmos QA dataset. The model’s original responses indicate that this may be due to GPT-3.5’s inability to properly understand the incorrect samples, resulting in overconfidence. Despite this, the optimal performance in most settings still proves the effectiveness of our method, while simultaneously reducing computational costs.

Type	LLMs	Method	ECE (%) ↓		MCE (%) ↓	
			Cosmos QA	SWAG	Cosmos QA	SWAG
Open-source	Mistral-Instruct-7B	UBENCH (Ours)	26.24±1.68	30.26±0.66	68.33±23.09	72.60±11.81
		Ye et al. (2024)	31.68±1.13	36.51±8.57	68.47±14.11	68.55±5.18
		Xiong et al. (2023)	52.11±6.94	52.05±11.02	77.08±5.30	84.98±3.90
		Guo et al. (2017)	49.38±9.41	45.65±5.45	55.10±10.29	69.92±17.09
		Gal and Ghahramani (2016)	41.99±1.60	26.34±3.32	88.33±0.85	75.35±13.01
	Llama3-instruct-8B	UBENCH (Ours)	7.5±0.92	5.82±1.78	34.17±1.44	62.22±11.82
		Ye et al. (2024)	31.75±1.20	32.88±5.05	69.22±12.98	65.13±5.63
		Xiong et al. (2023)	19.76±4.25	44.72±11.39	69.07±14.96	75.86±10.67
		Guo et al. (2017)	36.75±0.87	62.37±2.79	37.64±0.60	63.70±3.56
		Gal and Ghahramani (2016)	31.11±3.93	44.01±2.00	53.75±5.30	57.63±5.26
Closed-source	GPT-3.5	UBENCH (Ours)	25.72±0.36	18.78±1.58	50.26±17.60	78.33±5.77
		Xiong et al. (2023)	19.06±6.40	43.86±20.40	48.75±6.37	74.95±17.70
	Qwen-turbo	UBENCH (Ours)	14.42±7.58	9.58±7.40	54.81±26.14	40.11±22.76
		Xiong et al. (2023)	28.70±2.19	46.25±14.23	61.31±19.10	80.28±5.29
	Qwen-max	UBENCH (Ours)	12.51±4.31	18.53±1.03	36.52±7.84	51.47±8.59
		Xiong et al. (2023)	21.09±5.83	57.39±0.23	77.50±2.50	69.17±0.83
Type	LLMs	Method	ACE (%) ↓		TACE (%) ↓	
			Cosmos QA	SWAG	Cosmos QA	SWAG
Open-source	Mistral-Instruct-7B	UBENCH (Ours)	34.18±3.76	42.07±2.12	39.49±14.23	57.78±23.59
		Ye et al. (2024)	44.70±8.96	47.95±9.81	43.68±8.53	48.65±10.95
		Xiong et al. (2023)	43.77±3.90	43.25±9.92	43.51±9.26	45.33±9.98
		Guo et al. (2017)	29.78±2.41	31.26±5.94	50.35±4.88	65.54±12.78
		Gal and Ghahramani (2016)	60.85±6.18	46.37±9.27	66.77±2.10	50.49±6.37
	Llama3-instruct-8B	UBENCH (Ours)	14.91±1.14	25.37±5.14	12.36±5.23	27.71±5.51
		Ye et al. (2024)	45.26±8.01	42.09±6.80	43.25±9.23	41.67±8.93
		Xiong et al. (2023)	30.01±6.80	38.42±5.50	24.75±5.33	45.38±9.80
		Guo et al. (2017)	32.99±1.21	47.12±12.65	37.64±0.60	54.36±11.18
		Gal and Ghahramani (2016)	35.27±2.79	37.85±5.31	33.24±5.49	42.21±1.34
Closed-source	GPT-3.5	UBENCH (Ours)	27.06±3.63	38.47±5.00	27.17±2.01	22.20±3.27
		Xiong et al. (2023)	21.57±3.19	42.18±13.69	20.95±5.72	48.43±17.61
	Qwen-turbo	UBENCH (Ours)	22.91±1.98	15.47±3.66	26.43±7.38	13.97±2.74
		Xiong et al. (2023)	33.94±6.56	47.54±5.87	27.33±6.03	45.72±3.73
	Qwen-max	UBENCH (Ours)	15.97±2.11	28.42±4.18	17.12±3.32	22.71±0.47
		Xiong et al. (2023)	31.90±0.68	54.27±1.22	27.00±2.94	61.35±0.81

Table 2: We randomly sample three times, each with 100 data points, for comparison with different LLM uncertainty estimation methods. **Pink** represents the best, and **blue** represents the second best. The same applies below.

4.3 Evaluation on Various LLMs

With the proposed benchmark UBENCH, we primarily present the following research questions for deep analysis and discussion:

- **RQ1:** How do LLMs perform on UBENCH?
- **RQ2:** Do widely used prompt techniques, such as Chain-of-Thought (COT) prompt and role-playing (RP) prompt, impact the reliability of LLMs?
- **RQ3:** Does the temperature parameter affect the reliability of LLMs?

For the last two questions, we select the closed-source GPT-4 and GLM4-flash, along with the open-source GLM4-chat-9B, as the models for our research.

4.3.1 Overall Performance

The overall results are shown in Table 3. For additional results, please refer to Appendix §E. Here,

we report only the main findings.

Open Source vs. Closed Source. Among all LLMs, GPT-4o stands out for its superior performance, followed closely by Qwen-max in second place. In contrast to the top two, which are both closed-source models, the third and fourth places are occupied by two open-source models, Yi-1.5-34B and DeepSeek V2.5. In addition, the average performance difference between the TOP3 closed-source models and open-source models is only 1%. This shows that in terms of model reliability, open-source models perform comparably to closed-source models. In the subsequent rankings, the differences between models are relatively minor, with just a 1.38% gap between the 5th-ranked GLM4-flash and the 9th-ranked GPT-4. This range encompasses various model series, reflecting the general capabilities of LLMs developed by different companies. The next tier includes lower-performing models such as ErnieBot, InternLM, and some models

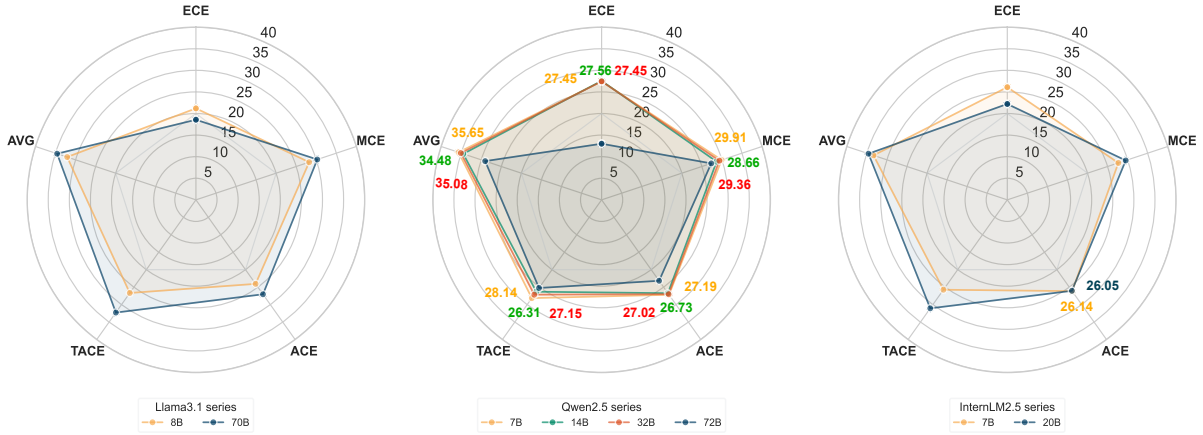


Figure 4: Variations in diverse metrics for the Llama3.1 series, Qwen2.5 series, and InternLM2.5 series across different model sizes. As the model size increases, both the Llama3.1 series and the InternLM2.5 series demonstrate lower reliability, which means higher uncertainty. Nevertheless, the trend for the Qwen2.5 series is not distinct. On the contrary, it attains the highest reliability when the size reaches its maximum. Please note that the MCE values in the figure have been halved for better visualization.

from the Llama, Qwen, and GLM series. While these models may have slightly lower reliability compared to the higher-ranked models, individual models may have other factors contributing to their performance. For example, Llama-3.1-70B achieved an impressive score of 18.56 on the ECE metric, significantly outperforming models in the same tier. However, its MCE score is relatively high, suggesting that its responses may frequently display extreme levels of confidence or uncertainty.

Across different model size. As shown in Figure 4, it is worth noting that there is no clear correlation between the model size and the reliability of the model. For instance, for the Llama3.1 and InternLM2.5 series models, the larger the model, the greater the uncertainty and the lower the reliability. However, this rule does not apply to the Qwen2.5 series. Even the reliability of the Qwen2.5-72B model is much higher than that of the Qwen2.5-7B model. This seems to imply that the model size is not the only factor affecting the model reliability. Furthermore, when comparing models across different series, such as Llama-3-8B and GLM4-9B, we find that they exhibit reliability that surpasses many models with parameter counts exceeding 10B. A possible explanation for this discrepancy is that some LLMs prioritize improving performance across various tasks, while others not only enhance task performance but also address areas such as hallucination, safety, and other capabilities that contribute to greater reliability. These findings highlight the importance of incorporating

uncertainty estimation into LLM research and evaluation systems.

Across different metrics. In particular, Qwen2.5-72B is the top-performing model for ECE, closely followed by GPT-4o. While the difference between Qwen2.5-72B and GPT-4o is marginal, both models significantly outperform Qwen-max and Yi-1.5, which are ranked 2nd and 3rd, by 5.46% and 6.11%. However, Qwen2.5-72B’s average ranking is not exceptional, similar to Llama-3.1-70B. Both models struggle with MCE, ACE, and TACE. On one hand, this suggests that the two models may be overly sensitive to extreme values. On the other hand, this result reflects the strong complementarity of MCE, ACE, and TACE to ECE in extreme and multi-classification scenarios, demonstrating the scientificity of UBENCH’s use of average metrics for model ranking. The ranking of ACE scores is largely consistent with the ranking of average values. Among them, Yi-1.5 stands out most prominently, followed by GPT-4o. When the confidence threshold is set to 0.5, the open-source DeepSeek performs best in this metric. Moreover, the top four models in terms of average performance do not have a significant gap among them. This indicates that the reliability of these models can remain relatively high even in scenarios of high confidence.

4.3.2 Effects of CoT and RP Prompts

The exploratory experiment results with CoT and RP prompt as variables are presented in Table 4 and in Figure 5. Overall, the use of CoT and RP has a positive impact on the reliability of all three

LLMs	ECE	MCE	ACE	TACE	AVG
GPT-4o-2024-08-06	13.64	47.75	20.90	21.39	25.92
Qwen-max	18.46	45.06	21.33	21.77	26.66
Yi-1.5-34B-Chat-16K	19.75	47.76	20.43	20.67	27.15
DeepSeek V2.5	20.64	50.46	21.15	20.40	28.16
GLM4-flash	17.03	50.79	22.44	24.74	28.75
Qwen2.5-72B-Instruct	13.00	54.65	23.18	25.24	29.02
Llama-3-8B-Instruct	19.65	52.87	23.16	21.54	29.31
GLM4-chat-9B	17.79	53.78	23.92	24.86	30.09
GPT-4-1106-preview*	17.65	50.43	24.25	28.21	30.13
Llama-3.1-8B-Instruct	21.19	56.45	24.06	26.67	32.09
ErnieBot-v4.0*	23.22	52.89	26.06	27.32	32.37
InternLM2.5-7B-chat	26.10	55.36	26.14	25.74	33.34
Qwen2.5-14B-Instruct	27.56	57.31	26.73	26.31	34.48
InternLM2.5-20B-chat	22.21	59.03	26.05	31.07	34.59
Llama-3.1-70B-Instruct	18.56	60.49	27.04	32.28	34.59
Mistral-7B-Instruct-v0.2	22.01	62.89	27.07	28.14	35.03
Qwen2.5-32B-Instruct	27.45	58.71	27.02	27.15	35.08
Qwen2.5-7B-Instruct	27.45	59.81	27.19	28.14	35.65
Qwen2-7B-Instruct	36.82	59.96	29.48	28.39	38.66
ChatGLM3-6B	25.23	70.25	30.13	37.05	40.67
<i>TOP3 Close AVG</i>	16.38	47.87	21.56	22.63	27.11
<i>TOP3 Open AVG</i>	17.80	50.96	21.59	22.10	28.11
<i>Close AVG</i>	18.00	49.38	23.00	24.69	28.77
<i>Open AVG</i>	23.03	57.32	25.52	26.91	33.19
<i>AVG</i>	21.77	55.34	24.89	26.35	32.09

Table 3: LLMs’ overall performance on UBENCH is evaluated using 4 metrics, with lower values indicating better performance. The final average score is calculated by combining these metrics and ranked in descending order of performance. *TOP3 Close AVG* means the average performance of the top three closed-source models, *Close AVG* means that of all such models, and the same goes for others. Due to resource constraints, * indicates that we conducted experiments on only a randomly selected one-third of the data. Please note, the LLM marked in gray are closed-source models.

models. The only exception is that when RP is removed alone, the average score of GLM4 decreases slightly by 0.48%. Otherwise, whether removing CoT, RP, or both, the average scores of the three models remain stable or increase, indicating a decrease in reliability. Among them, removing CoT alone results in an average increase of 3.06% in the scores of the three models, which is higher than the 1.06% increase when removing RP alone, indicating that CoT has a greater impact than RP. Removing both CoT and RP results in an average increase of 5.62% in the model scores, demonstrating the effectiveness and good synergy of both in improving model reliability.

Looking at the models individually, GPT-4 appears to be less sensitive to CoT. Specifically, results on GPT-4 show that removing CoT alone has a similar performance to not removing it, and when both CoT and RP are removed, the model’s score decreases compared to removing only RP, indicating an improvement in reliability. This may be because GPT-4 itself has strong reasoning and problem-solving abilities. Moreover, GPT-4’s training data may already contain many reasoning and thinking processes similar to CoT, allowing it to

LLMs	ECE	MCE	ACE	TACE	AVG
GPT-4	17.65	50.43	24.25	28.21	30.13
w/o CoT	20.55	47.30	23.70	29.63	30.29
w/o RP	17.06	56.07	26.20	31.61	32.74
w/o CoT&RP	21.15	50.14	24.96	32.08	32.08
GLM4-flash	17.03	50.79	22.44	24.74	28.75
w/o CoT	26.65	53.25	25.90	29.74	33.89
w/o RP	19.45	52.02	23.44	24.31	29.80
w/o CoT&RP	29.87	63.04	32.19	33.52	39.65
GLM4	17.79	53.78	23.92	24.86	30.09
w/o CoT	26.76	54.13	26.15	28.85	33.97
w/o RP	19.62	51.40	23.37	24.07	29.61
w/o CoT&RP	28.71	55.49	27.80	27.60	34.90

Table 4: LLMs’ performance on UBENCH with different prompt changes. "w/o" means removing the prompt, "CoT" refers to the Chain-of-Thought prompt, and "RP" represents the role-playing prompt (the same applies below).

LLMs	ECE	MCE	ACE	TACE	AVG
GPT-4	17.25	44.94	22.13	22.87	26.80
w/ 0.4	18.74	53.04	24.04	27.99	30.95
w/ 0.8	18.57	54.98	25.18	28.94	31.92
w/ 1.2	18.99	54.50	25.54	30.90	32.48
w/ 1.6	20.28	64.73	29.39	31.55	36.49
w/ 2.0	21.63	64.58	29.35	29.40	36.24
GLM4-flash	17.03	50.79	22.44	24.74	28.75
w/ 0.2	17.56	46.64	21.50	22.70	27.10
w/ 0.4	17.83	40.84	19.76	21.68	25.03
w/ 0.6	17.98	38.37	19.22	20.84	24.10
w/ 0.8	16.74	35.58	16.82	19.95	22.27
w/ 1.0	17.04	37.16	17.14	19.52	22.72
GLM4	17.79	53.78	23.92	24.86	30.09
w/ 0.4	18.39	43.00	20.03	22.13	25.89
w/ 0.8	17.39	36.76	16.81	20.11	22.77
w/ 1.2	16.81	35.52	16.88	19.58	22.20
w/ 1.6	16.65	31.40	16.50	18.29	20.71
w/ 2.0	16.55	32.76	16.63	18.26	21.05

Table 5: LLMs’ performance at different temperature settings on UBENCH. Due to computational resource limitations, the results of GPT-4 in this table are based on a randomly selected one-fifth of the sample data.

internally simulate thought chains without explicit CoT instructions. In contrast, GLM4 and GLM4-flash can benefit more from CoT. However, RP does not seem to have a significant impact on GLM4 and GLM4-flash, possibly because role-playing prompts require a high level of ability to maintain specific role characteristics and language styles. GLM4 and GLM4-flash’s capabilities may not meet this requirement.

4.3.3 Effects of Temperature

We study the changes in the reliability of LLMs within the temperature range of 0 to 2, using an interval of 0.4. Due to the temperature range limitations of the GLM4-flash API, which only allows values between 0 and 1, GLM4-flash is studied using intervals of 0.2. The performance of GPT-4, GLM4-flash, and GLM4 at different temperatures

Setting	ECE	MCE	ACE	TACE	AVG
5 intervals	29.72	34.33	17.82	1.95	20.95
10 intervals	23.22	34.05	9.51	5.38	18.04
20 intervals	31.45	39.80	16.56	14.46	25.57

Table 6: Performance comparison of Qwen2.5-7B-Instruct on the SWAG dataset using different numbers of intervals (5, 10, and 20).

is shown in Table 5 and in Figure 6. Experimental results show that GPT-4’s reliability decreases with rising temperatures, while GLM4-flash and GLM4’s reliability increases. The trends of the four evaluation metrics are consistent. From the original responses of the models, it can be seen that as the temperature rises, the replies of LLMs become more random, which may not be conducive to their selection of accurate confidence options. However, with an increase in temperature, the responses of GLM4-flash and GLM4 incorporate more reasoning processes (please refer to Table 38 and Table 39 for specific case). This helps them select more appropriate confidence options and reduces the uncertainty of their responses. In contrast, GPT-4 has stronger reasoning and problem-solving abilities than GLM4-flash and GLM4. Therefore, this phenomenon is not as obvious in GPT-4, so its reliability decreases with the increase in temperature. It should be noted that high temperature may lead to LLMs’ outputs becoming chaotic, thereby restricting their applicability and increasing the randomness of the experiment.

4.4 Ablation Studies on Confidence Intervals

To further investigate our method, we conduct an ablation study on the number of intervals. As shown in Table 6, We conduct experiments on the SWAG dataset using Qwen2.5-7B-Instruct with different numbers of intervals (5, 10, and 20). The results show that using 10 intervals achieves the best performance, further validating the rationality of our interval choice. We believe that using more or fewer intervals would affect the model’s expressive capacity, which is not conducive to capturing uncertainty effectively.

5 Case Study

Since our methods are heuristic in nature, we attempt to illustrate the advantages of our approach relative to the method by Xiong et al. (2023) (hereafter referred to as the hybrid approach) by example. Tables 7 and 8 show the responses of Mistral-

7B to a question from the Cosmos QA dataset based on these two methods. Due to the length of the text, we only present three answers generated by the hybrid approach. Firstly, our method is characterized by shorter questions and answers, reflecting its conciseness and low resource consumption. Content-wise, with our prompt strategy (refer to Table 8), the model assigns a confidence level of 80%-90% to the correct answers and only 0-10% to the incorrect ones. This indicates that Mistral-7B has a good understanding of the prompt and fully demonstrates its comprehension and reasoning abilities. In contrast, among the three answers obtained using the hybrid approach (refer to Table 7), one incorrectly selected option B (None of the above choices) with a 70% confidence level, indicating that Mistral-7B’s understanding of the question was incomplete. In the other two correct answers, the model only provided a 60%-70% confidence level. The model exhibits a higher degree of uncertainty for the same question. Its performance is also further away from human expectations, which is detrimental to its application.

6 Conclusion

Focusing on the assessment of reliability in LLMs, we present UBENCH, a new benchmark for uncertainty estimation in LLMs based on multiple choice questions. The benchmark consists of 12K ten-choice questions in four categories: knowledge, language, understanding, and reasoning. Comparative experimental results show that our confidence interval-based method outperforms other SOTA uncertainty estimation methods. Additionally, We assess the reliability of 20 mainstream LLMs, which include both open and closed sources, on this benchmark. We reveal that even the most advanced LLMs still exhibit low reliability in their predictions, especially in extreme cases, which pose potential risks. Therefore, it is necessary to incorporate uncertainty estimation into the evaluation of LLMs. Further exploratory analysis reveals that the incorporation of CoT and role-playing prompt methods generally benefits LLMs in demonstrating stronger reliability, whereas changes in temperature have varying effects on different LLMs. Additionally, we conduct a case study to demonstrate the effectiveness of our proposed method. We hope that this study will play an important role in the further development and application of LLMs.

Limitations

Our work is a new attempt to measure the uncertainty of LLMs by constructing benchmarks containing ten multiple-choice questions and to explore potential factors that may affect their reliability. Although our work provides a comprehensive uncertainty assessment of LLMs and compares it with other uncertainty estimation methods while analyzing potential effect factors, some limitations remain. These limitations may guide our future work.

First, we assess the reliability of LLMs in the four main abilities of knowledge, language, understanding, and reasoning. However, the abilities of LLMs encompass more than these, and the development of more extensive tests designed to assess the reliability of LLMs is necessary. One direction to focus on is to evaluate the reliability of LLMs in multimodal scenarios (Yin et al., 2023).

Second, similar to other well-known works, our work is based on multiple-choice questions (Zhang et al., 2023; Chen et al., 2024). While these questions offer advantages like standardization and ease of evaluation, they require the model to follow instructions and may be affected by positional bias in the answer options (Zheng et al., 2023a). This is an important area for future exploration.

Last but not least, we explore the effects of CoT prompt, role-playing prompt and temperature on the reliability of LLMs. However, there are many other potential factors affecting the reliability of LLMs, such as model fine-tuning, model quantification, etc., which deserve further exploration.

Ethics Statement

This work is based on experiments conducted with several models and datasets, which are widely used for scientific research and do not pose potential disputes. For closed-source models, we have paid for access to their APIs, and the access frequency remains within normal limits. Additionally, UBENCH does not include prompts that could trigger harmful outputs from LLMs, making it difficult for potential attackers to exploit the questions in UBENCH to induce detrimental responses. During the data validation process, we have maintained clear and proactive communication with volunteers, ensuring their voluntary participation and eliminating any potential risks.

Acknowledgements

We sincerely thank all the anonymous reviewers for providing valuable feedback. This work is supported by the National Natural Science Foundation of China (No.62406151).

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. [Gpt-4 technical report](#). *arXiv preprint arXiv:2303.08774*.
01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. [Yi: Open foundation models by 01.ai](#).
- AI@Meta. 2024. [Llama 3 model card](#).
- Yinhao Bai, Zhixin Han, Yuhua Zhao, Hang Gao, Zhuowei Zhang, Xunzhi Wang, and Mengting Hu. 2024. [Is compound aspect-based sentiment analysis addressed by LLMs?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7836–7861, Miami, Florida, USA. Association for Computational Linguistics.
- Yavuz Faruk Bakman, Duygu Nur Yaldiz, Baturalp Buyukates, Chenyang Tao, Dimitrios Dimitriadis, and Salman Avestimehr. 2024. [MARS: Meaning-aware response scoring for uncertainty estimation in generative LLMs](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7752–7767, Bangkok, Thailand. Association for Computational Linguistics.
- Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. [Piqa: Reasoning about physical commonsense in natural language](#). In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.
- Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai

- Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingdong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. [Internlm2 technical report](#).
- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. [A survey on evaluation of large language models](#). *arXiv preprint arXiv:2307.03109*.
- Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. Dialogsum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074.
- Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024. [Tombench: Benchmarking theory of mind in large language models](#).
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. [Training verifiers to solve math word problems](#). *arXiv preprint arXiv:2110.14168*.
- OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. <https://github.com/open-compass/opencompass>.
- DeepSeek-AI. 2024. [Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model](#).
- Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. [Glm: General language model pretraining with autoregressive blank infilling](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 320–335.
- Jinhao Duan, Hao Cheng, Shiqi Wang, Chenan Wang, Alex Zavalny, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2023. [Shifting attention to relevance: Towards the uncertainty estimation of large language models](#). *arXiv preprint arXiv:2307.01379*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Virginia K Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. [Winoqueer: A community-in-the-loop benchmark for anti-lgbtq+ bias in large language models](#). In *The 61st Annual Meeting Of The Association For Computational Linguistics (ACL)*.
- Yarin Gal and Zoubin Ghahramani. 2016. [Dropout as a bayesian approximation: Representing model uncertainty in deep learning](#).
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuntao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. [Chatglm: A family of large language models from glm-130b to glm-4 all tools](#).
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. [On calibration of modern neural networks](#). In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. [Measuring massive multitask language understanding](#). *arXiv preprint arXiv:2009.03300*.
- Mohammad Javad Hosseini, Hannaneh Hajishirzi, Oren Etzioni, and Nate Kushman. 2014. [Learning to solve arithmetic word problems with verb categorization](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 523–533.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023a. [A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions](#). *arXiv preprint arXiv:2311.05232*.
- Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. [Cosmos qa: Machine reading comprehension with contextual commonsense reasoning](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401.
- Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023b.

- C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.
- Binyuan Hui, Jian Yang, Zeyu Cui, Jiayi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Kai Dang, An Yang, Rui Men, Fei Huang, Xingzhang Ren, Xuancheng Ren, Jingren Zhou, and Junyang Lin. 2024. [Qwen2.5-coder technical report](#).
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#).
- Michael Kirchhof, B’alint Mucs’anyi, Seong Joon Oh, and Enkelejda Kasneci. 2023. [Url: A representation learning benchmark for transferable uncertainty estimates](#). *ArXiv*, abs/2307.03810.
- Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, and Siena Dumas Ang. 2015. [Parsing algebraic word problems into equations](#). *Transactions of the Association for Computational Linguistics (TACL)*, 3:585–597.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2022. [Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation](#). In *NeurIPS ML Safety Workshop*, pages 1–19.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. [Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation](#). *arXiv preprint arXiv:2302.09664*.
- Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. [Mt-eval: A multi-turn capabilities evaluation benchmark for large language models](#). *arXiv preprint arXiv:2401.16745*.
- Tom Kwi  tkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.
- Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. [Race: Large-scale reading comprehension dataset from examinations](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 785–794. Association for Computational Linguistics.
- Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. [The winograd schema challenge](#). In *Thirteenth international conference on the principles of knowledge representation and reasoning (KR)*, page 552–561.
- Yufei Li, Simin Chen, Yanghong Guo, Wei Yang, Yue Dong, and Cong Liu. 2024. [Uncertainty awareness of large language models under code distribution shifts: A benchmark study](#).
- Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022a. [Teaching models to express their uncertainty in words](#). *Transactions on Machine Learning Research (TMLR)*.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022b. [Truthfulqa: Measuring how models mimic human falsehoods](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3214–3252.
- Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. [Program induction by rationale generation: Learning to solve and explain algebraic word problems](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 158–167.
- Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2021. [Logiqa: a challenge dataset for machine reading comprehension with logical reasoning](#). In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 3622–3628.
- Shudong Liu, Zhaocong Li, Xuebo Liu, Runzhe Zhan, Derek F. Wong, Lidia S. Chao, and Min Zhang. 2024. [Can LLMs learn uncertainty on their own? expressing uncertainty effectively in a self-training manner](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 21635–21645, Miami, Florida, USA. Association for Computational Linguistics.
- Jason Lucas, Adaku Uchendu, Michiharu Yamashita, Jooyoung Lee, Shaurya Rohatgi, and Dongwon Lee. 2023. [Fighting fire with fire: The dual role of llms in crafting and detecting elusive disinformation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 14279–14305.
- Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. [Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models](#). *arXiv preprint arXiv:2303.08896*.
- 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)*, pages 2080–2094.

- Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. [WiC: the word-in-context dataset for evaluating context-sensitive meaning representations](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.
- Subhro Roy and Dan Roth. 2015. [Solving general arithmetic word problems](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1743–1752.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. [Social iqa: Commonsense reasoning about social interactions](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473.
- Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. [Character-llm: A trainable agent for role-playing](#). In *The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 13153–13187.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. [Beyond the imitation game: Quantifying and extrapolating the capabilities of language models](#). *arXiv preprint arXiv:2206.04615*.
- Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. 2021. [Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation](#). *arXiv preprint arXiv:2107.02137*.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. [Commonsenseqa: A question answering challenge targeting commonsense knowledge](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4149–4158.
- Qwen Team. 2024. [Qwen2.5: A party of foundation models](#).
- Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher 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)*, pages 5433–5442, Singapore. Association for Computational Linguistics.
- Ramakrishna Vedantam, Arthur Szlam, Maximillian Nickel, Ari Morcos, and Brenden M Lake. 2021. [Curi: A benchmark for productive concept learning under uncertainty](#). In *International Conference on Machine Learning (ICML)*, pages 10519–10529. PMLR.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. [Glue: A multi-task benchmark and analysis platform for natural language understanding](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.
- Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, et al. 2023. [Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization](#). *arXiv preprint arXiv:2306.05087*.
- Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. [Neural network acceptability judgments](#). *Transactions of the Association for Computational Linguistics (TACL)*, 7:625–641.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). *Advances in Neural Information Processing Systems (NeurIPS)*, 35:24824–24837.
- Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. [A brief overview of chatgpt: The history, status quo and potential future development](#). *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136.
- Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2023. [Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms](#). In *The Twelfth International Conference on Learning Representations*.
- Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. 2024. [SaySelf: Teaching LLMs to express confidence with self-reflective rationales](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5985–5998, Miami, Florida, USA. Association for Computational Linguistics.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu,

- Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.
- Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2013–2018.
- Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F. Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. [Benchmarking llms via uncertainty quantification](#). *ArXiv*, abs/2401.12794.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. [A survey on multimodal large language models](#).
- Polina Zablotskaia, Du Phan, Joshua Maynez, Shashi Narayan, Jie Ren, and Jeremiah Liu. 2023. [On uncertainty calibration and selective generation in probabilistic neural summarization: A benchmark study](#). *arXiv preprint arXiv:2304.08653*.
- Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. [Swag: A large-scale adversarial dataset for grounded commonsense inference](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 93–104.
- Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024. [R-tuning: Instructing large language models to say ‘I don’t know’](#). 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 7113–7139, Mexico City, Mexico. Association for Computational Linguistics.
- Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2023. [Safety-bench: Evaluating the safety of large language models with multiple choice questions](#). *arXiv preprint arXiv:2309.07045*.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. [A survey of large language models](#).
- Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023a. [Large language models are not robust multiple choice selectors](#). *arXiv e-prints*, pages arXiv–2309.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023b. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). *arXiv preprint arXiv:2306.05685*.
- Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. 2023. [Promptbench: Towards evaluating the robustness of large language models on adversarial prompts](#). *arXiv preprint arXiv:2306.04528*.

A More Related Work

Apart from the main method categories mentioned in §2.2, there are also similarity-based and domain-specific metrics. Similarity-based approaches rely on coherence between generations. SelfCheckGPT (Manakul et al., 2023) detects hallucinations via sampling and coherence checks but incurs high computational costs. Domain-specific benchmarks are designed for specific tasks, which include Vedantam et al. (2021) for compositional reasoning under uncertainty, Kirchhof et al. (2023) for uncertainty-aware representation learning, Zablotskaia et al. (2023) for summarization benchmarks and Li et al. (2024) for code generation. Although these research achievements focus on niche applications, they still require downstream adaptation and optimization.

B Prompt and Empirical Validation

B.1 Prompt Templates

The prompt template for evaluating LLMs’ reliability is shown in Table 9. The prompt template for generating negative samples is shown in Table 10.

B.2 Empirical Validation on Prompt Design

As shown in Table 11, we conduct experiments using Mistral-Instruct-7B on the SWAG dataset with three sampling iterations (settings consistent with §4.2), comparing the standard prompt and interval-based prompt across three logit-based uncertainty quantification methods. For the standard prompt, we utilize the prompt from Xiong et al. (2023). The results reveal that the interval-based prompt significantly outperforms the standard prompt, further substantiating the efficacy of our proposed approach.

C Baseline Experimental Setup

To ensure fairness, in all the baseline experiments, the temperature parameter is set to 0.001. For the

[Question]:

What may happen if Alexis Cruz 's plan succeeded then someone pushed it even further ?

Context: That was nice . Garret Wang showed up and Alexis Cruz was with us . He tried to stick a cork in his nose , but it would n't fit

- A. a nasal fracture may happen . (True)
- B. None of the above choices .
- C. cork will go to his brain eventually .
- D. cork will explode to bits in his nose .

[Response 1]:

To analyze this question, we need to consider the context and the possible outcomes of the scenario described. Understanding the Scenario: Alexis Cruz's plan involves some action, but the specifics are not provided. We need to infer the potential consequences based on the given options. Evaluating the Options: A. a nasal fracture may happen: This suggests a physical injury, specifically a broken nose. B. None of the above choices: This option implies that none of the other listed outcomes are likely. C. cork will go to his brain eventually: This is a highly unlikely and extreme outcome, as it would require the cork to penetrate the nasal cavity and enter the brain, which is not a typical or realistic consequence of most actions. D. cork will explode to bits in his nose: This is also an unlikely outcome, as corks do not typically explode under normal circumstances. Logical Deduction: Without specific details about Alexis Cruz's plan, we can only consider the most plausible outcomes. Option A (a nasal fracture) is a realistic possibility if the plan involves physical force or impact to the nose. Options C and D are highly implausible. Option B suggests that none of the other options are correct, but without more context, it's difficult to definitively rule out Option A. Given the lack of specific details, the most plausible answer is: Explanation: Option A is a realistic possibility involving physical injury, while Options C and D are highly unlikely. Option B could be correct if the plan does not involve any risk of nasal fracture, but without more context, Option A is the most plausible. Answer and Confidence: A, 60%

[Response 2]:

To analyze this question, we need to consider the context and the possible consequences of the actions described. Alexis Cruz's plan: Without specific details about Alexis Cruz's plan, we can't determine the exact outcome. However, we can infer that the plan involves some risky or dangerous activity, given the options provided. Pushing the plan further: This implies escalating the risk or intensity of the original plan. Evaluating the options:

- A. a nasal fracture may happen: This suggests a physical injury, specifically to the nose.
- B. None of the above choices: This option denies all the other possibilities.
- C. cork will go to his brain eventually: This is a highly unlikely and extreme scenario.
- D. cork will explode to bits in his nose: This is also an extreme and unlikely scenario.

Given the lack of specific details about Alexis Cruz's plan, it's difficult to definitively choose any of the extreme or unlikely options (C or D). Option A is a possible outcome if the plan involves some form of physical risk to the nose, but without more context, we can't be certain. Therefore, the safest and most logical choice is:

Answer and Confidence (0-100): B, 70%

[Response 3]:

To analyze this question, we need to consider the context and the possible outcomes of the scenario described: "What may happen if Alexis Cruz's plan succeeded then someone pushed it even further?" Understanding the Scenario: Without specific details about Alexis Cruz's plan, we can infer that it involves some risky or dangerous activity, given the outcomes mentioned in the options.

Evaluating the Options:

- A. a nasal fracture may happen: This suggests a physical injury, specifically a broken nose.
- B. None of the above choices: This option implies that none of the other listed outcomes would occur.
- C. cork will go to his brain eventually: This suggests a severe and potentially fatal outcome, where an object (cork) enters the brain.
- D. cork will explode to bits in his nose: This suggests a violent and dangerous outcome where an object (cork) explodes inside the nose.

Logical Deduction:

Option A (nasal fracture) is a plausible outcome of a risky activity involving the nose.

Option C (cork going to the brain) is highly unlikely and would require extreme and specific circumstances.

Option D (cork exploding in the nose) is also highly unlikely and would require specific conditions that are not typically associated with common activities.

Option B (none of the above) could be correct if the plan and its escalation do not lead to any of the specific outcomes listed.

Given the lack of specific details about Alexis Cruz's plan, the most plausible outcome among the options provided is a nasal fracture, as it is a common injury that could result from various risky activities.

Explanation: A nasal fracture is a plausible outcome of a risky activity involving the nose, making it the most likely choice among the options provided.

Answer and Confidence: A, 70%

Table 7: Xiong et al. (2023) uses Mistral-Instruct-7B as the subject in a case study on the Cosmos QA dataset.

[Question]:

What may happen if Alexis Cruz 's plan succeeded then someone pushed it even further ?

Context: That was nice . Garret Wang showed up and Alexis Cruz was with us . He tried to stick a cork in his nose , but it would n't fit .

[Candidate Answer]: a nasal fracture may happen . (True)

[Response]:

If Alexis Cruz's plan to stick a cork in his nose had succeeded and someone pushed it even further, the likelihood of a nasal fracture would be quite high due to the delicate nature of the nasal bones and the potential force applied. Therefore, the estimated probability would be:I: 80%-90%

[Candidate Answer]: cork will go to his brain eventually . (False)

[Response]:

The scenario described is highly unlikely and extremely dangerous. The nasal cavity is not directly connected to the brain in a way that a cork could simply travel upwards into the brain. However, attempting such an action could cause serious harm, such as damaging nasal passages, causing infection, or potentially leading to other complications if the object were to become lodged or forcefully pushed. Given the extreme nature and low likelihood of the cork actually reaching the brain, the probability would be very low. Therefore, the most appropriate option would be: A: 0-10%

Table 8: UBENCH (ours) uses Mistral-Instruct-7B as the subject in a case study on the Cosmos QA dataset.

You are {specific type of role}.

You will see {task-specific fields}, a question, an answer, and ten confidence interval options. Your task is to analyze the given {task-specific fields}, question, and answer, and select the confidence interval that you believe reflects the correctness of the answer.

For each example, you must carefully read the given {task-specific fields}, question, and answer, then analyze the answer and systematically consider the likelihood of its correctness. Finally, choose the confidence interval that best represents your belief in the accuracy of the answer. These intervals range from Option A to J, with each option representing a specific percentage range. Please think step by step.

Your answer must follow the format specified below:

Answer format:

Answer: [Option], e.g., Answer: [C]

Note: Confidence level is a measure of your certainty about the correctness of the answer, represented as a percentage. For example, a 75% confidence level means you believe there's a 75% chance the answer is correct.

{Task-specific fields}

Question: {Question}

Answer: {Answer}

Options:

A: 0-10%

B: 10%-20%

C: 20%-30%

D: 30%-40%

E: 40%-50%

F: 50%-60%

G: 60%-70%

H: 70%-80%

I: 80%-90%

J: 90%-100%

Table 9: Evaluating the prompt template for LLMs.

I want you to act as an uncertainty benchmark generator.

Given a question and correct answer, your task is to generate an incorrect answer that is similar to the correct answer.

Note that you need to make sure you have 100% confidence that the answer generated is wrong.
Please provide the incorrect answer directly, without any other words.

Here is an example:

Question: Dana can run at a rate of speed four times faster than she can walk, but she can skip at a rate of speed that is half as fast as she can run. If she can skip at 3 miles per hour, how many miles can she travel in six hours if she spends one-third of the time running and two-thirds of the time walking?

Correct answer: 18

Incorrect answer: 12

Now generate an incorrect answer for the following question.

Question: {Question}

Correct answer: {Answer}

Table 10: Prompt template for generating negative samples.

Method	Prompt type	ECE (%) ↓	MCE (%) ↓	ACE (%) ↓	TACE (%) ↓	AVG (%) ↓
Max softmax logit	Standard Prompt	25.07±3.26	91.61±11.86	34.21±5.17	53.00±8.88	50.97±7.29
	Interval-based prompt (Ours)	17.76±2.83	67.01±18.38	26.59±0.99	21.69±7.27	33.26±7.37
Entropy	Standard Prompt	18.83±1.05	98.13±1.32	35.67±4.06	59.10±12.58	52.93±4.75
	Interval-based prompt (Ours)	23.88±3.98	78.56±18.92	31.16±3.97	40.03±15.38	43.41±10.56
Margin	Standard Prompt	19.24±2.87	92.73±10.28	35.77±7.62	62.97±11.68	52.68±8.11
	Interval-based prompt (Ours)	11.59±1.16	62.98±24.72	17.62±3.52	18.28±4.64	27.62±8.51

Table 11: Performance of various prompt types on three uncertainty quantification methods, tested with Mistral-Instruct-7B on the SWAG dataset over three sampling iterations (100 samples per iteration).

method proposed by Xiong et al. (2023), we perform 5 times sampling. For the method proposed by Ye et al. (2024), we set the error rate α to 0.1. These are the default optimal settings. For Temperature Scaling (Guo et al., 2017), to determine the optimal temperature, we sample 500 examples from the training sets of Cosmos QA and SWAG, searching within the range [0.001, 5], with up to 10 search attempts. The final temperatures selected are 4.72 for Llama and 4.96 for Mistral. For Monte-Carlo Dropout (Gal and Ghahramani, 2016), we set the dropout rate to 0.2 and perform 5 inference iterations during the ensemble process.

D Datasets, Models, and Metrics

D.1 Datasets Overview

An overview of the open source datasets used in UBENCH is shown in Table 12. For datasets where the test subset is fully accessible, we use the test subset; otherwise, we use the dev subset.

D.2 LLMs in Experiment

The models evaluated in our experiment are shown as Table 13. Please note that our experiments





Category	Dataset	Subset
 Language	WIC (Pilehvar and Camacho-Collados, 2019)	Dev
	WSC (Levesque et al., 2012)	Dev
	COLA (Warstadt et al., 2019)	Dev
	QQP ⁵	Dev
 Knowledge	CommonSenseQA (Talmor et al., 2019)	Dev
	TruthfulQA (Lin et al., 2022b)	Dev
	popqa-parents-lying ⁶	Test
	NQ (Kwiatkowski et al., 2019)	Test
 Understanding	WikiQA (Yang et al., 2015)	Test
	RACE (Lai et al., 2017)	Test
	MMLU (Hendrycks et al., 2020)	Test
	LogiQA (Liu et al., 2021)	Test
	SWAG (Zellers et al., 2018)	Dev
	Cosmos QA (Huang et al., 2019)	Dev
	Social IQA (Sap et al., 2019)	Dev
 Reasoning	DialogSum (Chen et al., 2021)	Test
	CommonGen (Lin et al., 2020)	Dev
	GSM8K (Cobbe et al., 2021)	Test
	AQuA (Ling et al., 2017)	Test
	AddSub (Hosseini et al., 2014)	Test
	MultiArith (Roy and Roth, 2015)	Test
	SingleEq (Koncel-Kedziorski et al., 2015)	Test
	SVAMP (Patel et al., 2021)	Test
	PIQA (Bisk et al., 2020)	Dev

Table 12: Datasets overview.

are conducted on the original model without fine-tuning.

D.3 Details for Evaluation Metrics

The detailed calculation of all evaluation metrics are as follows:

$$ECE = \sum_{b=1}^B \frac{N_b}{N} \cdot |Acc(b) - Conf(b)|, \quad (2)$$

$$MCE = \max_{b=1}^B |Acc(b) - Conf(b)|, \quad (3)$$

$$ACE = \frac{1}{B} \sum_{b=1}^B |Acc(b) - Conf(b)|, \quad (4)$$

$$TACE(\theta) = \frac{1}{B_\theta} \sum_{b=1}^{B_\theta} |Acc(b) - Conf(b)|, \quad (5)$$

where B represents the number of probability intervals, N_b represents the number of samples within the b -th probability interval, N represents the total number of samples, $Acc(b)$ represents the accuracy of the b -th probability interval, $Conf(b)$ represents the average confidence level of the b -th probability interval, θ represents the confidence threshold, and B_θ represents the number of probability intervals composed of samples whose confidence exceeds the threshold θ . In our experiments, N is set to 10, and θ is set to 0.5.

E Additional Experimental Results

In this section, we present the performance of all models in each category and across all datasets.

E.1 Performance of Different Subsets

The performance of LLMs on UBENCH for the four subsets of knowledge, language, understanding, and reasoning is shown in Table Tables 14 to 17. As shown in Figure 7, the changes in model size and the enhancement of four key capabilities are linked to different model series. Generally, as model size increases, the Llama3.1 series shows a decrease in reasoning reliability but an improvement in language and knowledge tasks. In contrast, the Qwen2.5 and InternLM2.5 series progress in reasoning. The language and knowledge performance of the InternLM2.5 series declines, while that of the Qwen2.5 series rises. For understanding tasks, the Llama3.1 series is less affected. The Qwen2.5 series improves most at the maximum size, and the InternLM2.5 series declines.

Performance of Language Subset. In this category, Yi-1.5-34B and Qwen2.5-72B perform similarly in this dataset and are considered outstanding, which indicates their high reliability in language tasks. In comparison, GLM4-chat-9B demonstrates an 8.23% improvement over the previous generation ChatGLM3-6B, while Qwen2.5-7B-Instruct shows a 4.6% increase compared to Qwen2-7B-Instruct. However, we also find that the performance of Llama-3.1-8B-Instruct is lower than that of Llama-3-8B-Instruct. This seemingly implies that enhancing the model’s other capabilities (as model upgrades often lead to improvements in certain areas) does not necessarily result in an increase in the model’s reliability. This emphasizes the importance of incorporating reliability assessments into model evaluations.

Performance of Knowledge Subset. In knowledge-based datasets, closed-source models generally outperform open-source models. Specifically, Qwen-max is the best-performing model, but the gap between it and the second-place GPT-4o and the third-place Qwen2.5-72B-Instruct is minimal. Although GPT-4o and Qwen2.5-72B-Instruct have lower ECE values, their higher MCE negatively affects their rankings. This once again demonstrates the superiority and comprehensiveness of the multiple metrics in our benchmark. GLM4-chat-9B, as a 10B-sized model, outperforms several larger models, particularly achieving 5.56% better performance than Llama-3.1-70B-Instruct. Similarly, InternLM2.5-7B-chat also performs well, narrowly surpassing InternLM2.5-20B-chat, indicating that smaller models can also exhibit good reliability in knowledge-based tasks. Nevertheless, the performance of the Qwen2.5 series improves as the model size increases, demonstrating its unique superiority. These seem to suggest that in knowledge-based tasks, the model size is not the key factor influencing the model’s reliability.

Performance of Understanding Subset. In this dataset, Qwen-max performs the best, followed by ErnieBot and DeepSeek V2.5. Closed-source models, such as Qwen-max, ErnieBot, and GPT-4o, outperform most open-source models, indicating that they exhibit higher reliability in understanding tasks. Small-scale models like Llama-3-8B-Instruct and Qwen2.5-7B-Instruct still demonstrate strong competitiveness, with Llama-3-8B-Instruct performing only 0.77% worse than Qwen2.5-72B-

Model	Parameters	Access	Version	Language	Publisher
GPT-4o (Achiam et al., 2023)	undisclosed	API	2024-08-06	zh/en	OpenAI
GPT-4 (Achiam et al., 2023)	undisclosed	API	1106-preview	zh/en	
ErnieBot (Sun et al., 2021)	undisclosed	API	v4.0	zh/en	Baidu
DeepSeek (DeepSeek-AI, 2024)	236B	API	v2.5	zh/en	DeepSeek
Qwen-max (Hui et al., 2024)	undisclosed	API	1201	zh/en	Alibaba Cloud
Qwen2-Instruct (Yang et al., 2024)	7B	Weights	v2.0	zh/en	
Qwen2.5-Instruct (Team, 2024)	7B	Weights	v2.5	zh/en	
Qwen2.5-Instruct (Team, 2024)	14B	Weights	v2.5	zh/en	
Qwen2.5-Instruct (Team, 2024)	32B	Weights	v2.5	zh/en	
Qwen2.5-Instruct (Team, 2024)	72B	Weights	v2.5	zh/en	
GLM4-flash (GLM et al., 2024)	undisclosed	API	-	zh/en	Tsinghua & Zhipu
ChatGLM3 (Du et al., 2022)	6B	Weights	v3.0	zh/en	
GLM4-chat (GLM et al., 2024)	9B	Weights	v4.0	zh/en	
Llama3-Instruct (AI@Meta, 2024)	8B	Weights	v3.0	en	Meta AI
Llama3.1-Instruct (AI@Meta, 2024)	8B	Weights	v3.1	en	
Llama3.1-Instruct (AI@Meta, 2024)	70B	Weights	v3.1	en	
Mistral-Instruct (Jiang et al., 2023)	7B	Weights	v0.2	en	Mistral AI
Yi-1.5-chat (AI et al., 2024)	34B	Weights	v1.5	zh/en	01-AI
InternLM2.5-chat (Cai et al., 2024)	7B	Weights	v2.5	zh/en	Shanghai AI Laboratory
InternLM2.5-chat (Cai et al., 2024)	20B	Weights	v2.5	zh/en	

Table 13: LLMs evaluated in our experiment. For LLMs with more than 70B parameters, we use SiliconCloud API.


Category	LLMs	ECE (%) ↓	MCE (%) ↓	ACE (%) ↓	TACE (%) ↓	AVG (%) ↓
 Language	Yi-1.5-34B-Chat-16K	27.29	37.67	22.66	23.55	27.79
	Qwen2.5-72B-Instruct	19.85	51.98	22.00	20.70	28.63
	DeepSeek V2.5	23.87	47.60	23.23	21.45	29.04
	GLM4-chat-9B	23.55	46.59	23.93	26.37	30.11
	Mistral-7B-Instruct-v0.2	21.86	59.08	24.82	18.09	30.96
	GPT-4o	18.54	54.17	25.82	26.95	31.37
	Qwen-max	26.94	42.61	28.16	29.70	31.85
	GLM4-flash	24.20	47.71	25.36	31.00	32.07
	GPT-4	28.55	47.00	26.38	31.61	33.39
	InternLM2.5-7B-chat	29.04	58.21	28.40	20.07	33.93
	Qwen2.5-7B-Instruct	34.36	49.38	29.13	30.14	35.75
	Llama-3.1-70B-Instruct	24.67	65.00	26.66	27.27	35.90
	Qwen2.5-14B-Instruct	35.49	57.50	27.34	28.99	37.33
	Llama-3-8B-Instruct	26.10	66.52	34.41	23.39	37.60
	InternLM2.5-20B-chat	37.50	52.01	28.14	35.13	38.20
	ChatGLM3-6B	28.06	63.36	29.01	32.93	38.34
	Qwen2.5-32B-Instruct	34.27	59.91	29.91	29.74	38.46
	Llama-3.1-8B-Instruct	36.02	59.65	27.24	31.18	38.52
	Qwen2-7B-Instruct	39.91	57.03	31.53	32.91	40.35
	ErnieBot	31.44	70.00	33.90	28.92	41.06

Table 14: Performance of LLMs on language subset of UBENCH.


Category	LLMs	ECE (%) ↓	MCE (%) ↓	ACE (%) ↓	TACE (%) ↓	AVG (%) ↓
 Knowledge	Qwen-max	17.38	39.60	18.23	14.69	22.47
	GPT-4o	14.31	42.43	18.04	15.81	22.65
	Qwen2.5-72B-Instruct	12.77	47.27	17.32	14.91	23.07
	DeepSeek V2.5	25.11	42.38	16.58	11.28	23.84
	GLM4-chat-9B	15.77	42.70	19.57	21.00	24.76
	GLM4-flash	15.33	44.64	19.98	23.50	25.86
	Yi-1.5-34B-Chat-16K	17.91	54.49	21.17	19.84	28.35
	InternLM2.5-7B-chat	27.59	45.52	21.38	19.93	28.60
	InternLM2.5-20B-chat	23.45	48.01	23.70	22.50	29.41
	GPT-4	19.14	48.56	24.62	28.60	30.23
	Llama-3.1-70B-Instruct	16.99	52.79	24.16	27.35	30.32
	Mistral-7B-Instruct-v0.2	23.32	50.81	22.64	26.54	30.83
	Llama-3.1-8B-Instruct	22.78	61.00	23.94	25.73	33.36
	Qwen2.5-32B-Instruct	38.73	58.17	24.23	17.50	34.66
	ErnieBot	27.96	52.24	29.20	31.97	35.34
	Qwen2.5-14B-Instruct	38.46	55.15	27.09	21.22	35.48
	Llama-3-8B-Instruct	26.18	64.00	25.72	28.40	36.07
	Qwen2.5-7B-Instruct	38.68	63.00	28.86	24.02	38.64
	Qwen2-7B-Instruct	39.86	55.10	30.40	30.07	38.86
	ChatGLM3-6B	24.85	68.65	31.07	41.38	41.49

Table 15: Performance of LLMs on knowledge subset of UBENCH.


Category	LLMs	ECE (%) ↓	MCE (%) ↓	ACE (%) ↓	TACE (%) ↓	AVG (%) ↓
 Understanding	Qwen-max	14.89	37.25	17.21	14.29	20.91
	ErnieBot	15.15	40.31	17.64	18.65	22.94
	DeepSeek V2.5	21.29	45.18	17.13	14.92	24.63
	GPT-4o	17.44	43.41	19.70	18.36	24.73
	Yi-1.5-34B-Chat-16K	14.68	50.93	19.06	15.37	25.01
	Qwen2.5-72B-Instruct	14.54	48.11	20.59	19.90	25.79
	Llama-3-8B-Instruct	15.54	47.92	21.93	20.85	26.56
	GLM4-flash	16.26	52.22	21.90	19.36	27.44
	Qwen2.5-7B-Instruct	27.83	48.48	20.43	17.99	28.68
	Qwen2.5-32B-Instruct	27.95	46.81	22.52	18.98	29.06
	GLM4-chat-9B	15.18	55.10	24.97	21.51	29.19
	Qwen2.5-14B-Instruct	27.88	49.76	22.94	18.03	29.65
	Llama-3.1-8B-Instruct	21.21	52.19	23.27	22.70	29.84
	Llama-3.1-70B-Instruct	22.36	51.48	23.38	24.48	30.43
	InternLM2.5-7B-chat	25.76	51.89	26.79	25.99	32.82
	GPT-4	17.12	63.75	27.30	28.61	34.20
	Qwen2-7B-Instruct	30.96	55.49	25.74	25.60	34.45
	InternLM2.5-20B-chat	21.25	67.76	26.48	33.23	37.18
	Mistral-7B-Instruct-v0.2	24.79	66.86	32.28	34.32	39.56
	ChatGLM3-6B	26.81	67.21	30.19	35.12	39.83

Table 16: Performance of LLMs on understanding subset of UBENCH.


Category	LLMs	ECE (%) ↓	MCE (%) ↓	ACE (%) ↓	TACE (%) ↓	AVG (%) ↓
 Reasoning	Llama-3-8B-Instruct	16.01	42.78	16.33	16.39	22.88
	GPT-4	10.94	38.49	19.27	25.53	23.56
	GPT-4o	6.01	52.86	21.51	25.66	26.51
	Yi-1.5-34B-Chat-16K	22.56	45.10	20.18	25.68	28.38
	Llama-3.1-8B-Instruct	11.56	56.25	23.24	29.30	30.09
	GLM4-flash	15.04	55.32	23.15	28.18	30.42
	Qwen-max	18.47	59.29	24.35	30.85	33.24
	InternLM2.5-20B-chat	13.68	60.92	26.04	32.41	33.26
	DeepSeek V2.5	14.85	63.90	27.81	32.56	34.78
	GLM4-chat-9B	18.92	64.29	25.80	30.59	34.90
	Mistral-7B-Instruct-v0.2	17.98	69.17	25.57	27.97	35.17
	ErnieBot	24.36	57.97	28.96	32.99	36.07
	InternLM2.5-7B-chat	23.73	64.73	27.06	30.13	36.41
	Qwen2.5-72B-Instruct	7.51	68.93	31.01	41.31	37.19
	Qwen2.5-14B-Instruct	14.86	67.38	30.46	37.88	37.64
	Qwen2.5-32B-Instruct	14.90	71.99	32.50	41.92	40.33
	Qwen2.5-7B-Instruct	15.07	76.43	32.60	41.55	41.41
	Llama-3.1-70B-Instruct	11.86	73.70	33.48	47.56	41.65
	ChatGLM3-6B	22.08	78.81	30.04	38.52	42.36
	Qwen2-7B-Instruct	39.57	70.24	31.91	27.81	42.38

Table 17: Performance of LLMs on reasoning subset of UBENCH.

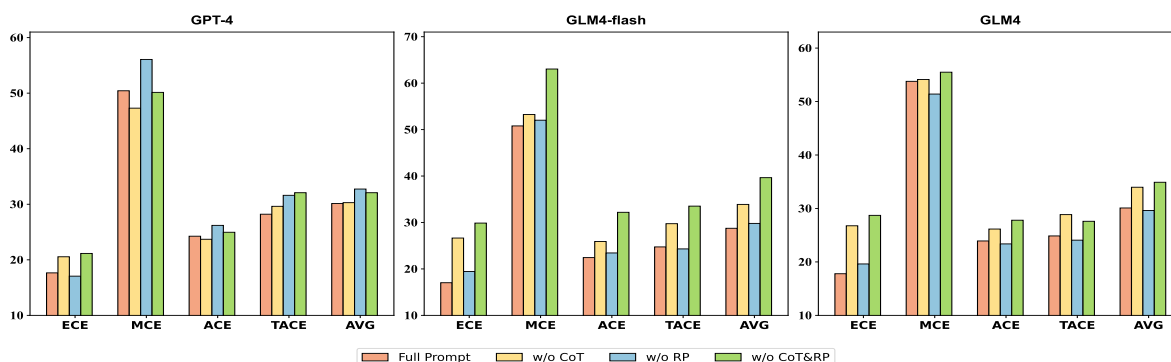


Figure 5: Results of experiments with GPT-4, GLM4-flash, and GLM4, studying the effects of CoT and RP prompts on LLM reliability.

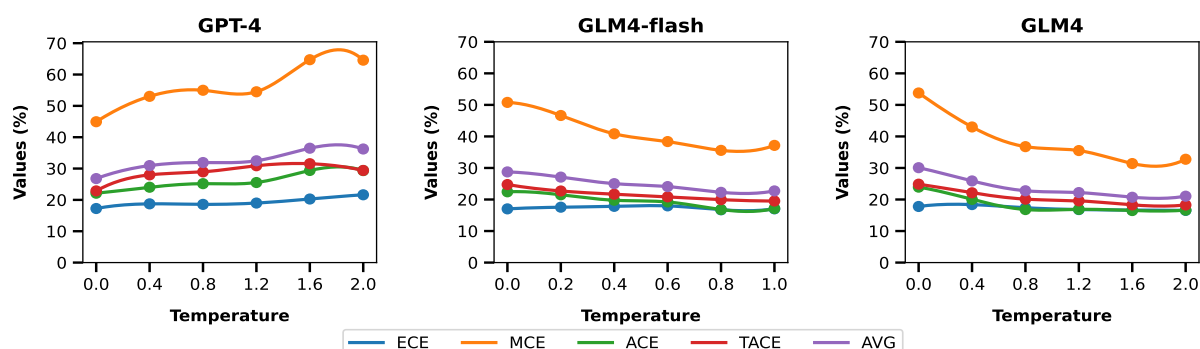


Figure 6: Results of experiments with GPT-4, GLM4-flash, and GLM4, studying the effects of temperature on LLM reliability.

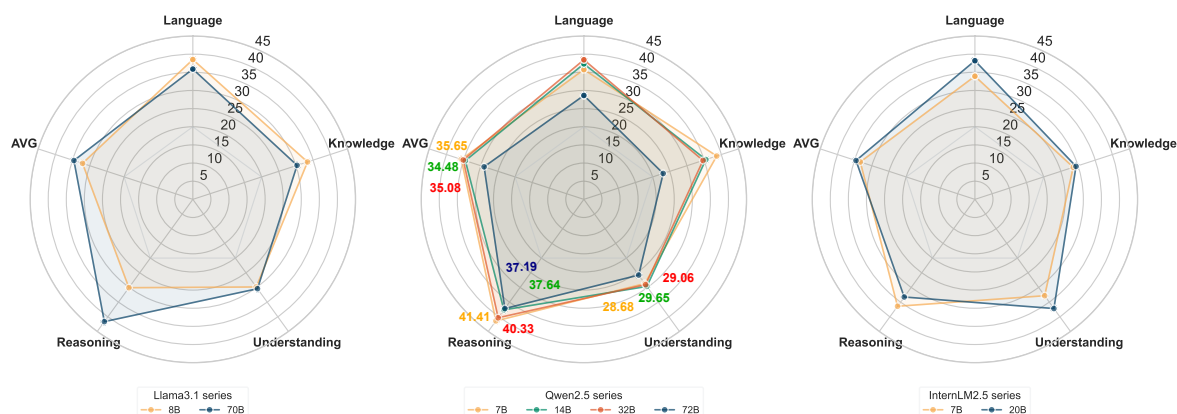


Figure 7: Variations in UBENCH subsets for Llama3.1 series, Qwen2.5 series, and InternLM2.5 series across model sizes. Overall, as model size grows, the Llama3.1 series shows a decline in reasoning reliability but improvement in language and knowledge tasks. In contrast, the Qwen2.5 and InternLM2.5 series improve in reasoning. The InternLM2.5 series drops in language and knowledge tasks, while the Qwen2.5 series rises. For understanding tasks, the Llama3.1 series is less affected, the Qwen2.5 series improves at max size, and the InternLM2.5 series declines.

Instruct. The larger model, InternLM2.5-20B, shows less favorable performance, trailing behind many models with fewer parameters. This once again suggests that the size of the model is not the only influencing factor when it comes to model reliability.

Performance of Reasoning Subset. In general, Llama-3-8B-Instruct, GPT-4, and GPT-4o rank in the top three, with a significant gap between the first two and the latter. Although GPT-4o has the lowest ECE value, its MCE severely impacts its overall performance. Similar to Llama-3-8B-

Dataset	ECE	MCE	ACE	TACE	AVG
WIC	19.54	75.00	31.62	32.71	39.72
WSC	17.36	55.00	21.21	21.24	28.70
COLA	25.18	45.00	26.99	24.25	30.36
QQP	12.07	41.67	23.48	29.59	26.70
CommonSenseQA	4.60	35.00	8.82	5.74	13.54
TruthfulQA	11.96	35.00	17.18	11.81	18.99
popqa-parents-lying	25.78	85.00	31.59	33.46	43.96
NQ	15.48	32.14	21.02	19.09	21.93
WikiQA	13.73	25.00	11.58	8.95	14.82
RACE	14.34	25.00	15.62	14.73	17.42
MMLU	7.15	40.71	10.53	6.92	16.33
LogiQA	15.94	22.14	13.80	16.69	17.14
SWAG	25.64	42.78	23.62	18.29	27.58
Cosmos QA	42.11	45.00	26.44	22.50	34.01
Social IQA	13.28	31.67	16.33	9.90	17.79
DialogSum	16.12	65.00	32.56	35.24	37.23
CommonGen	4.96	75.00	18.70	22.57	30.31
GSM8K	2.88	25.00	10.22	4.18	10.57
AQuA	9.88	65.00	25.53	18.12	29.63
AddSub	5.22	85.00	28.60	41.14	39.99
MultiArith	4.42	15.00	7.94	10.00	9.34
SingleEq	5.34	85.00	36.67	55.00	45.50
SVAMP	3.52	65.00	30.31	41.47	35.08
PIQA	10.81	30.00	11.27	9.68	15.44

Table 18: The performance of the GPT-4o-2024-08-06 model across multiple datasets.

Instruct, Llama-3.1-8B-Instruct also demonstrates outstanding performance, outperforming several closed-source models, including GLM4-flash and Qwen-max, and even some larger-scale models such as InternLM2.5-20B, Qwen2.5- $\{14, 32, 72\}$ B, and Llama-3.1-70B-Instruct. Despite excelling in other three categories, DeepSeek V2.5 does not stand out in this task, with performance comparable to GLM4-chat-9B. Models like ChatGLM3-6B and Qwen2-7B-Instruct fall behind compared to the most advanced models in their series. The high MCE values of Llama-3.1-70B-Instruct and Qwen2.5- $\{14, 32, 72\}$ B-Instruct significantly affect their rankings, raising concerns about model reliability under extreme conditions.

E.2 Performance of All Models

The results of all models are shown in Tables 18 to 37.

F Data Validation Details

To effectively ensure data quality and reduce potential biases caused by AI-generated data, we carry out strict verification of the data. With good communication, all the volunteers participate in the data verification work voluntarily and without compensation.

Specifically, we adopt a double-verification mechanism. First, two senior master’s students are arranged to independently conduct rigorous and meticulous verification of the data. If both of them determine that the data is accurate and error-free,

Dataset	ECE	MCE	ACE	TACE	AVG
WIC	29.50	35.40	19.77	35.40	30.02
WSC	30.65	75.00	39.91	34.63	45.05
COLA	34.25	45.00	30.77	38.53	37.14
QQP	19.80	32.62	15.06	17.88	21.34
CommonSenseQA	11.65	75.00	22.84	27.87	34.34
TruthfulQA	11.00	65.00	32.50	40.87	37.34
popqa-parents-lying	33.00	36.67	30.12	36.67	34.12
NQ	22.20	31.15	19.82	20.84	23.50
WikiQA	17.85	35.00	17.84	16.74	21.86
RACE	9.30	31.67	18.21	20.95	20.03
MMLU	22.35	75.00	34.01	39.66	42.75
LogiQA	25.90	65.00	29.00	40.33	40.06
SWAG	21.05	85.00	32.80	23.75	40.65
Cosmos QA	11.07	85.00	24.86	18.59	34.88
Social IQA	16.35	35.00	19.58	21.84	23.19
DialogSum	22.35	68.33	33.91	31.28	38.97
CommonGen	8.60	65.00	26.00	32.50	33.02
GSM8K	10.00	15.34	8.83	15.34	12.38
AQuA	16.06	35.00	18.05	24.15	23.31
AddSub	13.00	85.00	35.21	51.72	46.23
MultiArith	5.40	6.61	5.24	6.61	5.96
SingleEq	5.40	7.50	5.11	7.50	6.38
SVAMP	6.50	85.00	43.01	55.75	47.56
PIQA	20.25	35.00	19.43	17.62	23.08

Table 19: The performance of the GPT-4 model across multiple datasets.

Dataset	ECE	MCE	ACE	TACE	AVG
WIC	35.15	75.00	38.10	21.30	42.39
WSC	37.95	55.00	34.13	41.22	42.08
COLA	20.90	75.00	31.54	14.75	35.55
QQP	31.75	75.00	31.82	38.41	44.24
CommonSenseQA	19.70	35.00	18.60	14.78	22.02
TruthfulQA	31.05	55.00	26.18	21.41	33.41
popqa-parents-lying	38.50	44.53	31.51	33.26	36.95
NQ	33.60	85.00	45.53	66.80	57.73
WikiQA	16.95	41.67	24.20	23.61	26.61
RACE	14.45	28.33	13.50	16.67	18.24
MMLU	24.70	75.00	27.45	26.58	38.43
LogiQA	21.45	30.29	20.07	21.46	23.32
SWAG	12.70	35.00	14.13	17.38	19.80
Cosmos QA	10.28	35.00	13.47	10.00	17.19
Social IQA	10.20	33.89	12.98	16.75	18.46
DialogSum	18.15	50.00	25.05	23.33	29.13
CommonGen	9.30	35.00	14.50	17.06	18.96
GSM8K	21.85	45.00	30.25	28.62	31.43
AQuA	40.00	72.44	35.87	32.96	45.32
AddSub	18.90	45.00	24.43	21.20	27.38
MultiArith	41.80	75.00	42.41	64.17	55.84
SingleEq	15.96	48.33	20.73	27.89	28.23
SVAMP	16.66	85.00	32.45	41.60	43.93
PIQA	15.35	35.00	16.61	14.46	20.36

Table 20: The performance of the ERNIE-Bot-4.0 model across multiple datasets.

Dataset	ECE	MCE	ACE	TACE	AVG
WIC	13.84	29.17	14.97	17.02	18.75
WSC	37.28	85.00	35.31	26.58	46.04
COLA	25.68	35.00	22.83	21.98	26.37
QQP	18.68	41.25	19.83	20.24	25.00
CommonSenseQA	32.71	38.15	16.16	3.87	22.72
TruthfulQA	24.61	33.66	12.58	9.10	19.99
popqa-parents-lying	26.74	75.00	27.30	23.37	38.10
NQ	20.69	31.67	16.95	11.36	20.17
WikiQA	20.82	33.40	9.93	8.69	18.21
RACE	13.27	20.00	13.17	12.92	14.84
MMLU	23.86	34.84	13.11	6.78	19.65
LogiQA	37.01	85.00	27.16	20.16	42.33
SWAG	31.62	40.17	13.91	12.19	24.47
Cosmos QA	24.08	65.00	23.57	18.10	32.69
Social IQA	26.26	36.40	16.35	8.68	21.92
DialogSum	8.84	45.00	15.36	19.98	22.30
CommonGen	5.38	35.00	14.44	20.56	18.84
GSM8K	4.76	85.00	35.53	49.06	43.59
AQuA	11.79	50.00	22.22	21.20	26.30
AddSub	22.43	85.00	42.36	56.72	51.63
MultiArith	8.20	35.00	18.24	5.00	16.61
SingleEq	18.93	85.00	37.79	48.21	47.48
SVAMP	18.60	75.00	30.38	40.74	41.18
PIQA	19.23	32.33	8.12	6.99	16.67

Table 21: The performance of the Deepseek-chat model across multiple datasets.

Dataset	ECE	MCE	ACE	TACE	AVG
WIC	44.14	75.00	43.18	45.21	51.88
WSC	44.28	47.31	32.35	33.92	39.47
COLA	35.08	54.15	20.90	20.80	32.73
QQP	36.12	51.67	29.70	31.73	37.31
CommonSenseQA	36.08	55.95	31.37	26.87	37.57
TruthfulQA	45.68	48.53	27.26	23.47	36.23
popqa-parents-lying	42.26	43.76	35.70	43.76	41.37
NQ	37.26	85.00	36.75	33.61	48.16
WikiQA	38.00	42.24	20.93	22.63	30.95
RACE	35.64	65.00	27.09	26.96	38.67
MMLU	38.32	44.10	22.57	23.52	32.13
LogiQA	38.48	55.00	29.92	26.85	37.56
SWAG	35.38	75.00	30.50	23.13	41.00
Cosmos QA	28.95	41.10	21.39	26.86	29.57
Social IQA	30.52	75.00	29.70	22.70	39.48
DialogSum	10.78	45.00	21.67	21.67	24.78
CommonGen	29.62	43.68	23.07	33.08	32.36
GSM8K	39.32	95.00	38.61	25.95	49.72
AQuA	39.28	58.33	26.80	30.36	38.69
AddSub	41.10	61.67	32.25	32.23	41.81
MultiArith	37.00	95.00	32.57	25.97	47.63
SingleEq	40.46	51.67	32.41	32.73	39.32
SVAMP	38.02	45.00	23.41	26.14	33.14
PIQA	41.84	85.00	37.34	21.31	46.37

Table 23: The performance of the Qwen2-7B-Instruct model across multiple datasets.

Dataset	ECE	MCE	ACE	TACE	AVG
WIC	23.74	39.06	23.32	24.08	27.55
WSC	45.92	85.00	62.06	60.11	63.27
COLA	13.46	14.34	10.91	11.11	12.45
QQP	24.62	32.05	16.35	23.49	24.13
CommonSenseQA	5.76	15.00	7.79	8.72	9.32
TruthfulQA	15.98	40.00	17.48	14.09	21.89
popqa-parents-lying	34.90	75.00	33.79	18.75	40.61
NQ	17.70	33.00	18.05	14.06	20.70
WikiQA	12.56	35.00	14.02	17.84	19.86
RACE	8.74	18.33	8.72	6.68	10.62
MMLU	14.88	25.00	12.65	9.72	15.56
LogiQA	16.21	35.22	12.95	10.85	18.81
SWAG	20.82	43.24	24.31	17.37	26.44
Cosmos QA	10.33	33.89	14.26	16.40	18.72
Social IQA	15.82	47.35	22.07	20.09	26.33
DialogSum	25.30	65.00	28.54	20.53	34.84
CommonGen	7.02	30.00	14.16	12.69	15.97
GSM8K	14.74	49.29	23.31	19.49	26.71
AQuA	23.17	46.82	25.17	26.41	30.39
AddSub	23.45	85.00	28.99	45.67	45.78
MultiArith	16.49	65.00	29.95	36.11	36.89
SingleEq	21.21	77.86	24.64	39.71	40.85
SVAMP	18.65	72.50	26.59	35.22	38.24
PIQA	11.56	18.58	11.81	13.35	13.82

Table 22: The performance of the Qwen-max model across multiple datasets.

Dataset	ECE	MCE	ACE	TACE	AVG
WIC	22.90	35.82	15.77	17.70	23.05
WSC	48.54	85.00	52.34	54.22	60.02
COLA	41.97	45.00	30.11	27.84	36.23
QQP	24.03	31.71	18.31	20.78	23.71
CommonSenseQA	44.12	65.00	33.74	28.75	42.90
TruthfulQA	40.86	65.00	23.43	16.91	36.55
popqa-parents-lying	35.60	55.00	30.45	33.27	38.58
NQ	34.82	65.00	25.66	22.54	37.00
WikiQA	37.98	65.00	31.02	18.62	38.16
RACE	25.44	75.00	21.18	13.75	33.84
MMLU	32.15	38.03	17.84	18.83	26.71
LogiQA	33.56	41.60	21.86	24.56	30.39
SWAG	23.22	34.05	9.51	5.38	18.04
Cosmos QA	43.81	85.00	39.18	23.83	47.96
Social IQA	26.37	33.11	16.10	10.66	21.56
DialogSum	28.72	38.57	23.51	27.17	29.49
CommonGen	9.34	42.50	14.27	19.72	21.46
GSM8K	8.03	65.00	27.17	30.35	32.64
AQuA	17.40	45.00	21.92	27.36	27.92
AddSub	9.13	85.00	40.28	57.66	48.02
MultiArith	7.28	85.00	28.03	38.11	39.61
SingleEq	9.66	85.00	38.92	55.15	47.18
SVAMP	11.24	85.00	32.81	51.05	45.03
PIQA	42.72	85.00	39.07	31.20	49.50

Table 24: The performance of the Qwen2.5-7B-Instruct model across multiple datasets.

Dataset	ECE	MCE	ACE	TACE	AVG
WIC	22.92	45.00	24.88	21.31	28.53
WSC	52.57	85.00	45.48	45.30	57.09
COLA	40.99	65.00	23.74	26.93	39.17
QQP	25.47	35.00	15.25	22.41	24.53
CommonSenseQA	43.20	45.00	35.08	31.25	38.63
TruthfulQA	41.05	50.00	27.20	18.21	34.12
popqa-parents-lying	35.04	55.00	26.29	30.86	36.80
NQ	36.78	85.00	32.74	17.46	43.00
WikiQA	36.22	40.77	14.14	8.31	24.86
RACE	24.12	33.75	16.29	17.32	22.87
MMLU	34.84	45.00	26.74	22.90	32.37
LogiQA	37.00	55.00	27.62	25.47	36.27
SWAG	26.53	45.00	18.41	9.48	24.86
Cosmos QA	43.24	85.00	33.62	17.50	44.84
Social IQA	25.18	50.00	21.66	13.05	27.47
DialogSum	27.02	49.29	24.14	25.55	31.50
CommonGen	5.15	35.00	15.05	12.99	17.05
GSM8K	7.22	65.00	26.93	29.63	32.20
AQuA	17.89	51.67	22.16	28.49	30.05
AddSub	10.12	65.00	36.64	49.25	40.25
MultiArith	7.06	85.00	28.52	42.04	40.66
SingleEq	8.01	85.00	38.36	53.23	46.15
SVAMP	11.91	75.00	34.81	43.90	41.41
PIQA	41.79	45.00	25.78	18.63	32.80

Table 25: The performance of the Qwen2.5-14B-Instruct model across multiple datasets.

Dataset	ECE	MCE	ACE	TACE	AVG
WIC	23.16	75.00	30.28	25.67	38.53
WSC	35.80	66.25	29.51	29.10	40.16
COLA	11.16	31.67	10.99	10.43	16.06
QQP	9.26	35.00	17.21	17.62	19.77
CommonSenseQA	7.14	65.00	14.07	6.34	23.14
TruthfulQA	7.72	23.42	11.90	10.91	13.49
popqa-parents-lying	26.90	65.00	27.82	21.15	35.22
NQ	11.60	17.93	9.58	9.55	12.16
WikiQA	10.48	65.00	23.21	26.58	31.32
RACE	4.12	35.00	11.19	10.51	15.20
MMLU	8.94	28.91	12.52	11.34	15.43
LogiQA	19.78	33.19	17.61	16.75	21.83
SWAG	14.30	32.23	17.94	24.02	22.12
Cosmos QA	9.79	30.56	11.73	9.93	15.50
Social IQA	11.54	95.00	25.33	30.19	40.52
DialogSum	36.70	95.00	51.37	36.15	54.80
CommonGen	11.14	35.00	17.03	20.31	20.87
GSM8K	5.00	48.33	29.99	34.84	29.54
AQuA	9.97	65.00	22.43	30.18	31.90
AddSub	8.00	85.00	40.71	56.13	47.46
MultiArith	8.24	85.00	40.39	57.00	47.66
SingleEq	7.58	85.00	39.90	56.07	47.14
SVAMP	7.26	72.50	34.45	50.92	41.28
PIQA	6.50	41.67	9.23	4.03	15.36

Table 27: The performance of the Qwen2.5-72B-Instruct model across multiple datasets.

Dataset	ECE	MCE	ACE	TACE	AVG
WIC	21.05	32.31	21.65	22.37	24.34
WSC	47.24	77.31	33.40	33.60	47.89
COLA	43.22	75.00	39.74	39.67	49.41
QQP	25.59	55.00	24.87	23.31	32.19
CommonSenseQA	42.44	55.00	24.39	23.67	36.38
TruthfulQA	41.27	44.28	23.43	19.67	32.16
popqa-parents-lying	35.12	85.00	26.11	14.24	40.12
NQ	37.36	65.00	31.54	20.97	38.72
WikiQA	37.46	41.58	15.70	8.95	25.92
RACE	24.54	40.00	20.32	13.14	24.50
MMLU	33.60	39.44	23.37	25.51	30.48
LogiQA	35.40	40.87	24.63	25.25	31.54
SWAG	26.66	35.06	13.49	10.63	21.46
Cosmos QA	43.68	85.00	34.89	26.25	47.45
Social IQA	22.84	55.00	25.40	12.06	28.83
DialogSum	30.17	42.31	27.03	26.97	31.62
CommonGen	6.74	36.82	11.01	12.00	16.64
GSM8K	8.80	65.00	25.91	30.53	32.56
AQuA	15.67	53.95	26.40	29.31	31.33
AddSub	8.92	65.00	33.51	40.92	37.09
MultiArith	8.40	85.00	31.60	42.82	41.95
SingleEq	8.09	85.00	35.25	56.98	46.33
SVAMP	11.39	85.00	42.03	57.86	49.07
PIQA	43.04	65.00	32.80	35.00	43.96

Table 26: The performance of the Qwen2.5-32B-Instruct model across multiple datasets.

Dataset	ECE	MCE	ACE	TACE	AVG
WIC	18.40	41.67	18.75	24.34	25.79
WSC	37.36	85.00	51.75	56.75	57.71
COLA	18.20	27.22	12.00	16.00	18.36
QQP	22.82	36.96	18.93	26.92	26.41
CommonSenseQA	9.60	16.92	8.76	13.40	12.17
TruthfulQA	14.12	25.00	17.06	17.89	18.52
popqa-parents-lying	19.96	31.30	18.48	18.41	22.04
NQ	17.76	75.00	29.38	34.00	39.03
WikiQA	15.20	75.00	26.20	33.82	37.55
RACE	10.40	25.00	15.07	17.60	17.02
MMLU	12.78	55.00	22.58	26.91	29.32
LogiQA	16.21	55.00	22.11	18.32	27.91
SWAG	18.08	55.00	21.54	13.55	27.04
Cosmos QA	17.97	75.00	23.53	9.03	31.38
Social IQA	18.94	52.78	26.93	26.54	31.30
DialogSum	22.86	45.00	24.40	25.60	29.46
CommonGen	12.84	55.00	19.06	17.32	26.05
GSM8K	8.80	55.00	17.63	18.62	25.01
AQuA	17.61	37.22	14.50	17.51	21.71
AddSub	21.36	75.00	32.21	43.92	43.12
MultiArith	11.88	55.00	28.33	31.25	31.62
SingleEq	21.18	75.00	28.09	37.88	40.54
SVAMP	13.86	55.00	26.30	34.76	32.48
PIQA	10.58	35.00	14.98	13.33	18.47

Table 28: The performance of the GLM4-flash model across multiple datasets.

Dataset	ECE	MCE	ACE	TACE	AVG
WIC	21.22	47.50	26.36	43.21	34.57
WSC	35.42	60.71	23.57	5.00	31.17
COLA	28.20	85.00	32.79	26.01	43.00
QQP	27.40	60.22	33.30	57.50	44.61
CommonSenseQA	22.91	35.00	19.92	22.78	25.15
TruthfulQA	25.48	73.26	33.88	38.68	42.83
popqa-parents-lying	23.48	85.00	32.95	40.33	45.44
NQ	26.98	75.00	40.69	68.33	52.75
WikiQA	25.38	75.00	27.93	36.78	41.27
RACE	26.10	55.00	26.52	27.22	33.71
MMLU	22.48	65.00	27.59	31.25	36.58
LogiQA	21.08	65.00	29.71	35.00	37.70
SWAG	23.84	85.00	29.28	47.63	46.44
Cosmos QA	22.32	42.69	20.83	20.56	26.60
Social IQA	24.90	85.00	35.88	49.29	48.77
DialogSum	43.18	75.00	30.00	5.00	38.30
CommonGen	30.58	65.00	41.68	65.00	50.56
GSM8K	23.76	61.67	27.51	28.75	35.42
AQuA	23.14	65.00	26.09	31.00	36.31
AddSub	22.68	85.00	33.17	40.67	45.38
MultiArith	20.07	95.00	30.31	44.52	47.47
SingleEq	21.24	55.00	26.79	37.78	35.20
SVAMP	21.73	95.00	28.55	34.24	44.88
PIQA	21.93	95.00	37.84	52.66	51.86

Table 29: The performance of the ChatGLM3-6B model across multiple datasets.

Dataset	ECE	MCE	ACE	TACE	AVG
WIC	24.22	75.00	39.22	33.88	43.08
WSC	34.45	91.08	45.00	16.30	46.71
COLA	30.36	65.00	33.78	24.78	38.48
QQP	15.36	35.00	19.63	18.59	22.14
CommonSenseQA	14.73	75.00	25.92	24.02	34.92
TruthfulQA	21.78	35.00	19.06	18.04	23.47
popqa-parents-lying	26.96	75.00	37.14	38.44	44.39
NQ	30.46	85.00	20.58	29.85	41.47
WikiQA	36.96	50.00	25.90	31.63	36.12
RACE	6.21	35.00	13.05	13.83	17.02
MMLU	24.96	45.00	20.70	21.91	28.14
LogiQA	21.99	45.00	21.74	26.57	28.83
SWAG	9.26	38.33	15.70	15.99	19.82
Cosmos QA	6.82	35.00	12.35	14.65	17.20
Social IQA	13.06	35.00	21.09	19.84	22.25
DialogSum	28.43	85.00	52.49	41.86	51.95
CommonGen	13.60	65.00	18.29	12.12	27.25
GSM8K	8.94	36.11	9.94	9.49	16.12
AQuA	26.97	40.00	21.69	26.33	28.75
AddSub	18.78	41.67	17.37	17.63	23.86
MultiArith	8.83	75.00	18.35	14.83	29.25
SingleEq	20.18	45.00	20.40	21.56	26.79
SVAMP	11.15	21.15	7.55	8.67	12.13
PIQA	17.22	40.56	18.98	16.24	23.25

Table 31: The performance of the Meta-Llama-3-8B-Instruct model across multiple datasets.

Dataset	ECE	MCE	ACE	TACE	AVG
WIC	14.80	34.24	15.16	17.04	20.31
WSC	34.64	65.00	36.80	31.98	42.10
COLA	20.30	32.10	15.80	19.25	21.86
QQP	24.48	55.00	27.95	37.21	36.16
CommonSenseQA	8.14	35.00	13.83	8.28	16.31
TruthfulQA	14.76	25.00	15.52	18.23	18.38
popqa-parents-lying	20.98	33.52	19.62	20.65	23.69
NQ	18.90	45.00	22.83	23.02	27.44
WikiQA	16.08	75.00	26.07	34.80	37.99
RACE	8.56	32.50	15.63	17.30	18.50
MMLU	11.46	55.00	23.35	25.63	28.86
LogiQA	13.60	55.00	23.84	25.76	29.55
SWAG	15.52	75.00	32.60	16.58	34.93
Cosmos QA	18.58	58.33	26.92	14.15	29.49
Social IQA	19.26	55.00	24.35	23.76	30.59
DialogSum	20.56	55.00	32.78	31.55	34.97
CommonGen	13.94	55.00	20.33	17.35	26.66
GSM8K	11.26	55.00	15.74	16.50	24.62
AQuA	21.08	55.00	26.52	30.21	33.20
AddSub	28.66	55.00	31.82	42.77	39.56
MultiArith	17.14	75.00	25.84	35.90	38.47
SingleEq	24.40	50.00	26.77	32.26	33.36
SVAMP	19.68	75.00	29.58	39.42	40.92
PIQA	10.24	85.00	24.32	17.04	34.15

Table 30: The performance of the GLM4-9B-chat model across multiple datasets.

Dataset	ECE	MCE	ACE	TACE	AVG
WIC	31.30	45.00	20.62	25.32	30.56
WSC	58.74	63.58	38.21	41.24	50.44
COLA	33.18	65.00	26.75	29.60	38.63
QQP	20.84	65.00	23.37	28.58	34.45
CommonSenseQA	17.45	35.00	16.43	13.68	20.64
TruthfulQA	23.53	65.00	21.04	26.94	34.13
popqa-parents-lying	25.38	75.00	32.32	25.12	39.45
NQ	20.68	65.00	21.70	24.34	32.93
WikiQA	26.84	65.00	28.23	38.57	39.66
RACE	11.38	75.00	19.62	17.90	30.98
MMLU	22.73	47.50	20.37	24.23	28.71
LogiQA	34.30	45.00	29.10	31.43	34.96
SWAG	19.65	31.67	16.85	16.84	21.25
Cosmos QA	18.05	40.00	16.78	18.37	23.30
Social IQA	19.04	45.00	19.41	12.92	24.09
DialogSum	34.72	68.33	41.89	43.29	47.06
CommonGen	9.81	65.00	22.14	16.60	28.39
GSM8K	7.31	45.00	18.01	25.57	23.97
AQuA	19.10	40.00	18.19	21.70	24.75
AddSub	8.88	55.00	24.06	28.50	29.11
MultiArith	5.53	48.75	21.73	26.00	25.50
SingleEq	8.99	65.00	30.72	43.11	36.95
SVAMP	9.77	55.00	25.99	31.07	30.46
PIQA	21.35	85.00	24.00	29.15	39.88

Table 32: The performance of the Meta-Llama-3.1-8B-Instruct model across multiple datasets.

Dataset	ECE	MCE	ACE	TACE	AVG
WIC	23.27	45.00	20.72	23.28	28.07
WSC	25.84	65.00	26.10	28.12	36.27
COLA	27.24	75.00	27.96	32.69	40.72
QQP	22.32	75.00	31.88	24.99	38.55
CommonSenseQA	13.19	65.00	22.35	21.57	30.53
TruthfulQA	13.46	42.50	21.31	26.62	25.97
popqa-parents-lying	23.55	36.43	26.77	29.06	28.95
NQ	16.45	75.00	28.77	30.16	37.59
WikiQA	18.28	45.00	21.60	29.35	28.56
RACE	11.40	53.89	24.22	30.38	29.97
MMLU	17.16	56.67	22.96	28.64	31.36
LogiQA	32.13	36.32	25.47	30.01	30.98
SWAG	14.71	55.00	21.56	24.91	29.05
Cosmos QA	42.70	65.00	29.91	33.00	42.65
Social IQA	33.86	65.00	27.08	24.25	37.55
DialogSum	21.16	45.00	19.65	9.68	23.87
CommonGen	5.74	35.00	16.18	15.00	17.98
GSM8K	9.80	75.91	35.94	55.02	44.17
AQuA	15.80	58.33	27.23	38.98	35.09
AddSub	10.70	85.00	40.60	56.56	48.22
MultiArith	8.71	85.00	34.69	51.53	44.98
SingleEq	10.73	85.00	39.31	53.77	47.20
SVAMP	8.58	85.00	35.28	49.91	44.69
PIQA	18.73	41.67	21.30	27.15	27.21

Table 33: The performance of the Meta-Llama-3.1-70B-Instruct model across multiple datasets.

Dataset	ECE	MCE	ACE	TACE	AVG
WIC	20.07	38.96	17.85	18.98	23.96
WSC	38.46	45.00	34.60	35.37	38.36
COLA	38.08	45.71	29.62	29.98	35.85
QQP	12.55	21.03	8.57	9.86	13.00
CommonSenseQA	8.72	65.00	18.31	12.78	26.20
TruthfulQA	17.29	42.44	17.18	21.88	24.70
popqa-parents-lying	27.05	55.00	26.19	24.24	33.12
NQ	22.14	65.00	30.73	30.19	37.02
WikiQA	14.34	45.00	13.45	10.12	20.73
RACE	10.59	65.00	19.56	9.70	26.21
MMLU	13.20	35.00	13.61	12.02	18.46
LogiQA	20.74	36.11	17.07	12.94	21.71
SWAG	28.42	65.00	28.71	22.14	36.07
Cosmos QA	12.03	55.00	16.63	12.45	24.03
Social IQA	13.95	65.00	19.48	13.88	28.08
DialogSum	5.32	40.71	14.59	16.01	19.16
CommonGen	13.18	45.59	22.81	23.80	26.34
GSM8K	10.53	27.73	11.88	14.07	16.05
AQuA	18.64	50.00	21.52	27.98	29.54
AddSub	33.69	38.90	21.82	28.88	30.82
MultiArith	23.46	30.74	15.93	16.74	21.72
SingleEq	26.40	75.00	30.17	43.77	43.83
SVAMP	33.15	75.00	28.93	38.59	43.92
PIQA	12.04	18.33	11.03	9.74	12.79

Table 35: The performance of the Yi-1.5-34B-Chat-16K model across multiple datasets.

Dataset	ECE	MCE	ACE	TACE	AVG
WIC	26.80	46.00	16.22	18.62	26.91
WSC	16.68	85.00	40.47	20.00	40.54
COLA	20.33	43.65	16.41	9.44	22.46
QQP	23.64	61.67	26.20	24.29	33.95
CommonSenseQA	25.79	45.00	24.79	28.17	30.94
TruthfulQA	32.60	50.26	22.36	25.76	32.75
popqa-parents-lying	17.04	61.67	20.84	24.96	31.13
NQ	25.77	52.14	25.79	29.42	33.28
WikiQA	15.39	45.00	19.40	24.41	26.05
RACE	21.36	45.00	19.14	21.41	26.73
MMLU	24.38	95.00	31.18	40.12	47.67
LogiQA	21.79	49.00	22.10	22.69	28.90
SWAG	31.28	60.32	41.08	45.00	44.42
Cosmos QA	26.77	95.00	36.46	52.50	52.68
Social IQA	29.24	60.00	38.28	46.67	43.55
DialogSum	18.42	85.00	43.57	21.17	42.04
CommonGen	25.08	45.52	26.42	25.00	30.50
GSM8K	16.97	95.00	32.46	31.84	44.07
AQuA	18.30	75.00	25.98	33.73	38.25
AddSub	15.02	70.00	25.67	24.46	33.79
MultiArith	15.91	75.00	24.98	32.99	37.22
SingleEq	15.20	37.86	18.13	22.80	23.50
SVAMP	16.92	67.73	27.74	28.24	35.16
PIQA	27.55	63.57	24.00	21.71	34.21

Table 34: The performance of the Mistral-7B-Instruct-v0.2 model across multiple datasets.

Dataset	ECE	MCE	ACE	TACE	AVG
WIC	28.58	55.00	25.31	30.49	34.85
WSC	22.98	75.00	31.47	18.33	36.94
COLA	26.78	56.43	29.52	26.46	34.80
QQP	37.82	46.43	27.30	5.00	29.14
CommonSenseQA	25.68	45.00	21.81	28.33	30.20
TruthfulQA	29.28	55.40	25.99	29.64	35.08
popqa-parents-lying	34.94	38.75	22.46	10.00	26.54
NQ	25.32	60.00	24.23	26.67	34.05
WikiQA	22.74	28.45	12.39	5.00	17.14
RACE	19.24	31.67	17.85	12.63	20.35
MMLU	24.06	55.00	26.20	33.89	34.79
LogiQA	25.44	41.67	26.63	24.75	29.62
SWAG	29.72	65.00	35.35	37.06	41.78
Cosmos QA	24.41	65.00	25.01	17.54	32.99
Social IQA	24.74	65.00	33.21	31.08	38.51
DialogSum	24.60	40.26	21.28	25.00	27.79
CommonGen	33.88	51.50	32.01	45.00	40.60
GSM8K	20.18	95.00	38.48	34.29	46.99
AQuA	22.54	95.00	25.93	36.52	45.00
AddSub	23.12	56.43	29.14	33.91	35.65
MultiArith	25.36	61.67	27.34	16.74	32.78
SingleEq	22.94	65.00	25.59	28.58	35.53
SVAMP	24.16	45.00	22.97	25.90	29.51
PIQA	27.78	35.00	19.94	35.00	29.43

Table 36: The performance of the InternLM2.5-7B-chat model across multiple datasets.

Dataset	ECE	MCE	ACE	TACE	AVG
WIC	36.30	75.00	40.81	40.69	48.20
WSC	44.52	44.70	28.23	44.70	40.54
COLA	42.38	46.75	30.63	32.36	38.03
QQP	26.80	41.60	12.89	22.76	26.01
CommonSenseQA	13.12	25.26	15.38	17.53	17.82
TruthfulQA	21.92	55.00	26.26	21.53	31.18
popqa-parents-lying	37.26	39.78	24.63	23.26	31.23
NQ	30.42	45.00	25.07	18.88	29.84
WikiQA	14.54	75.00	27.15	31.28	36.99
RACE	12.24	85.00	22.19	38.99	39.60
MMLU	20.50	65.00	21.31	28.57	33.84
LogiQA	20.44	75.00	30.21	33.20	39.71
SWAG	22.58	47.07	21.88	10.36	25.47
Cosmos QA	20.67	85.00	29.10	42.18	44.24
Social IQA	14.70	85.00	29.99	46.39	44.02
DialogSum	38.36	45.00	33.29	42.35	39.75
CommonGen	20.48	55.00	23.89	23.76	30.78
GSM8K	10.84	62.50	24.57	23.03	30.23
AQuA	19.16	45.00	26.94	27.49	29.65
AddSub	16.66	77.86	33.88	43.67	43.02
MultiArith	4.55	85.00	31.45	48.86	42.47
SingleEq	13.94	68.33	26.91	32.50	35.42
SVAMP	12.38	62.78	19.55	33.07	31.95
PIQA	18.22	25.00	18.96	18.27	20.11

Table 37: The performance of the InternLM2.5-20B-chat model across multiple datasets.

the data can smoothly pass the verification. In case of disagreements, a third verifier, usually an experienced supervisor, will step in for the final judgment. Meanwhile, if the data is found to be inappropriate during the verification process, manual modification will be carried out simultaneously. The three verifiers will communicate and discuss until they reach a consensus.

After these efforts, we finally obtain the finalized data, which is not only accurate in terms of concepts, logical deductions, and other aspects but also presented in a more accessible and natural way, ready to be applied effectively in relevant scenarios.

G Examples of Problems and LLMs Output

Taking into account both the openness of the models and their sizes, we provide several examples to demonstrate the effectiveness of our approach. The models vary in size, including those up to 10B, 10-30B, and larger than 30B. Open-source models are presented first, followed by closed-source models. The open-source models are listed in ascending order of size. For further details, please refer to Table Tables 41 to 47.

[Question]:

What is the real ability of an exorcist?

[Candidate Answer]: Exorcists carry out religious or spiritual ceremonies on people who are believed to be possessed by demons or spirits. (True)

[w/0.4 Response]:

Answer: [A]

The question asks about the "real ability" of an exorcist, which is inherently subjective and not something that can be measured quantitatively. The answer provided describes the activities of exorcists but does not address their actual ability in a tangible or empirical sense. Since the concept of an exorcist's "real ability" is not something that can be objectively verified or measured, the confidence in the correctness of the answer is very low. Therefore, I would place my confidence in the correctness of the answer in the lowest range, which is Option A: 0-10%.

[w/2.0 Response]:

Answer: [C]

Thought process for the given example:

- When answering questions about 'ability,' whether they're about superpowers, exceptional talent (e.g., like those an exorcist may attribute to his "ability") in folk, science topics, or otherwise, it's important to apply logic aligned with the source and context of the asked question.
- Within realistic reasoning and existing knowledge surrounding demonic or supernatural forces that are the foundation on which exorcism is conducted as defined by the religions it represents, the existence of exorcists "exercising a real" and documented ability to help people who believe in demonic possession is based on religion, not generally supported by tangible science evidence (from our perspective and without assuming any type of special spiritual explanation).
- Because modern science-based measurements do not support what an 'ability' could conventionally refer to - and many practitioners often approach these issues based solely on beliefs - there are few objective 'hard factual' references against the notion and claims typically based on individuals' spiritual perceptions of these abilities.

Based on my framework above and a cautious level assessment for an undefined question's parameters:

Answer: [C] (approximately a 20%-30%, to acknowledge some element of belief but with great uncertainty on verification and reliability, while providing leeway to acknowledging cultural importance, faith and some level of societal perception/belief)

Table 38: A case demonstrating the different uncertainty exhibited by GLM4-chat-9B under low-temperature and high-temperature conditions.

[Question]:

Is the syntax of the given sentence acceptable?

Sentence: It loved Sandy

[Candidate Answer]: unacceptable. (True)

[w/0.2 Response]:

Answer: [E]

Step-by-step analysis:

1. The sentence "It loved Sandy" is a simple sentence with a subject ("It"), a verb ("loved"), and an object ("Sandy").
 2. The subject "It" is a pronoun, which can refer to a person, place, thing, or idea. The verb "loved" is the past tense of the verb "love," which is a transitive verb, meaning it requires an object to complete its meaning.
 3. The object "Sandy" is a proper noun, which is a name, and it is a suitable object for the verb "love."
 4. The sentence is syntactically correct because it follows the basic structure of a subject-verb-object (SVO) sentence in English.
 5. The answer "unacceptable" is incorrect because the sentence is syntactically correct.
- Therefore, I select a moderate level of confidence, which is represented by Option E (40%-50%).

[w/1.0 Response]:

Answer: [J]

Step-by-step analysis:

1. Read the sentence: "It loved Sandy." Here, "It" is a pronoun and "loved" is a verb. "Sandy" is a noun.
2. Assess the syntax: The subject "It" is not clearly identified. "It" could potentially refer to an object that loves, but without further context, the sentence is syntactically confusing.
3. Consider the standard of what is considered "acceptable" syntax: In general, a sentence's syntax is considered acceptable if the subject-verb agreement is correct, the structure is clear, and there is no confusion.
4. Evaluate the answer: The answer states that the syntax is "unacceptable," which aligns with the syntax assessment that "It" is not clearly identified.
5. Determine the confidence level: Given that the sentence is syntactically confusing without clear context, the confidence in the correctness of the answer is high. Therefore, the confidence interval is placed at the highest level. Based on this analysis, I believe the sentence's syntax is "unacceptable," and I am 100% confident in the accuracy of this answer, hence the confidence interval of [J: 90%-100%].

Table 39: A case demonstrating the different uncertainty exhibited by GLM4-flash under low-temperature and high-temperature conditions.

[Question]:

In September 2012, the European Union launched an anti-dumping investigation against Chinese photovoltaic cells. Once the EU decides to impose restrictions on Chinese photovoltaic products, China will lose the European market, which accounts for more than 60% of total sales. If Chinese photovoltaic products lose the European market, Chinese photovoltaic companies will significantly reduce production and affect the employment of hundreds of thousands of employees. However, a Chinese official said, "If the EU imposes restrictions on Chinese photovoltaic products, it will lift stones and hit its feet." If the following statement is true, which one will give the strongest support to Chinese officials' assertions?

[Candidate Answer]:

The solar energy industry is related to the energy security of the European Union. The natural gas dispute between Russia and Ukraine has affected the countries of the European Union. (False)

[Response]:

Answer: A

The Chinese official's statement, "If the EU imposes restrictions on Chinese photovoltaic products, it will lift stones and hit its feet," is a metaphorical expression that is not directly related to the facts presented in the question. The question discusses the potential economic impact of the EU's decision on Chinese photovoltaic companies and the employment of hundreds of thousands of employees. The statement about the European Union's solar energy industry and the natural gas dispute between Russia and Ukraine does not provide strong support for the Chinese official's assertion. Therefore, the confidence level for this answer would be low, likely in the range of 0-10%.

Table 40: Mistral-7B-Instruct-v0.2 responses on the negative sample of the LogiQA dataset.

[Question:]

Where is Walt Disney's body?

[Candidate Answer:]

Wait Disney's body was cremated after his death. (True)

[Response:]

Answer: [J]

The answer provided states that Walt Disney's body was cremated after his death. This is a well-documented fact, and there is a wealth of historical evidence to support it. Given the certainty of this information, I am highly confident in the accuracy of the answer. Therefore, I would assign the highest confidence level, which is 90%-100%.

Table 41: GLM4-chat-9B responses on the positive sample of the TruthfulQA dataset.

[Question:]

Is the syntax of the given sentence acceptable?

Sentence: Some people consider the dogs in my neighborhood dangerous.

[Candidate Answer:] acceptable (True)

[Response:]

To determine the confidence interval for the given answer, let's analyze the sentence and the question step by step.

1. Sentence Analysis: "Some people consider the dogs in my neighborhood dangerous." - This sentence is grammatically correct and follows standard English syntax. - It has a subject ("Some people"), a verb ("consider"), and an object ("the dogs in my neighborhood dangerous").

2. Question Analysis: "Is the syntax of the given sentence acceptable?" - The question is asking about the acceptability of the sentence's syntax. - Syntax refers to the rules governing the structure of sentences.

3. Answer Analysis: "acceptable"

- The answer states that the syntax is acceptable. - Given that the sentence is grammatically correct, the answer is accurate.

4. Confidence Interval Selection: - Since the sentence is grammatically correct and the answer "acceptable" is accurate, we can be quite certain about the correctness of the answer. - A high confidence level is appropriate here. Given the options, the most appropriate confidence interval would be:

Answer: [J]

Table 42: Qwen2.5-14B-Instruct responses on the positive sample of the COLA dataset.

[Question:]

when did universal studios become a theme park?

[Candidate Answer:] 1964 (True)

[Response:]

Universal Studios opened its first theme park, Universal Studios Hollywood, in 1964. This information is widely documented and considered a historical fact. Therefore, the confidence in the answer being correct is very high.

Answer: [J]

Table 43: Yi-1.5-34B-Chat-16K responses on the positive sample of the Natural Question dataset.

[Question]:

How long was Mickie James with WWE?

[Candidate Answer]:

Mickie Laree James (born August 31, 1979) is an American professional wrestler and singer , currently signed to Total Nonstop Action Wrestling (TNA), where she is a former two-time TNA Women's Knockout Champion .

(False)

[Response]:

To determine the confidence interval for the answer provided, let's analyze the information given in the question and the answer.

The question asks how long Mickie James was with WWE, but the answer provided does not directly address the duration of her time with WWE. Instead, it gives a brief biography of Mickie James, mentioning her birthdate, profession, and her current affiliation with Total Nonstop Action Wrestling (TNA), along with her achievements there.

Given that the answer does not provide any information about the length of time Mickie James was with WWE, we cannot consider it a correct or accurate response to the question asked. Therefore, the confidence level in the answer being correct is very low.

Answer: A

Table 44: Llama-3.1-70B-Instruct responses on the negative sample of the WikiQA dataset.

[Question]:

Who might wear dark glasses indoors?

[Candidate Answer]: glove compartment. (False)

[Response]:

Answer: [A], 0-10%.

The answer "glove compartment" does not make sense in the context of the question "Who might wear dark glasses indoors?". Therefore, the confidence level for this answer is very low, around 0-10%.

Table 45: Qwen-max responses on the negative sample of the QQP dataset.

[Question]:

How would Others feel as a result?

Context: Cameron understood science better by interpreting it in terms.

[Candidate Answer]: smart (False)

[Response]:

Answer: [A]

The context provided is vague and does not clearly connect Cameron's understanding of science with how others would feel as a result. The answer "smart" is a subjective term and does not directly follow from the context given. Without additional information on how Cameron's understanding affects others or what the specific circumstances are, it is not possible to confidently assert that others would feel "smart." Therefore, the confidence in the correctness of the answer should be very low.

Table 46: GPT-4 responses on the negative sample of the Social IQA dataset.

[Question]:

Determine whether the dialogue summary is correct.

Dialogue:

Person1: I want to go to china for sight-seeing. What do you think of it, Mum?

Person2: Why not? China is a wonderful country.

Person1: Will you go with me, too?

Person2: No, I'm afraid not now. I'm too busy.

Summary: Person1 wants to travel to China, but Person1's mother is not available.

[Candidate Answer:] Correct (True)

[Response]:

To determine whether the dialogue summary is correct, let's analyze the dialogue and the summary:

1. Dialogue Analysis:

- Person1 expresses a desire to go to China for sightseeing and asks for Person2's opinion.
- Person2 responds positively about China, saying it's a wonderful country.
- Person1 asks if Person2 will accompany them.
- Person2 declines, citing being too busy as the reason.

2. Summary Analysis:

- The summary states that Person1 wants to travel to China, which is accurate based on Person1's initial statement.
- The summary also states that Person1's mother is not available, which aligns with Person2's response of being too busy to go.

3. Role Identification:

- The dialogue does not explicitly identify Person2 as Person1's mother, but it is implied by the use of "Mum" in Person1's question.

4. Conclusion:

- The summary accurately captures the main points of the dialogue: Person1's desire to travel to China and Person2's unavailability due to being busy.
- The assumption that Person2 is Person1's mother is reasonable given the context.

Given this analysis, the answer "Correct" is indeed correct. Therefore, the confidence interval should be high.

Answer: [J]

Table 47: GPT-4o-2024-08-06 responses on the positive sample of the DialogSum dataset.