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

Anonymous ACL submission

Abstract

The rapid development of large language models (LLMs) has shown promising practical results. However, their low interpretability often leads to errors in unforeseen circumstances, limiting their utility. Many works have focused on creating comprehensive evaluation systems, but previous benchmarks have primarily assessed problem-solving abilities while neglecting the response's uncertainty, which may result in unreliability. Recent methods for measuring LLM reliability are resource-intensive and unable to test black-box models. To address this, we propose UBENCH, a comprehensive benchmark for evaluating LLM reliability. UBENCH includes 11,978 multiple-choice questions covering knowledge, language, understanding, and reasoning abilities. Experimental results show that UBENCH has achieved state-of-the-art performance, while its single-sampling method significantly saves computational resources compared to baseline methods that require multiple samplings. Additionally, based on UBENCH, we evaluate the reliability of 20 popular LLMs, finding GPT-4o to be the most outstanding, closely followed by Qwen-max. We also explore the impact of Chain-of-Thought prompts, role-playing prompts and temperature on LLM reliability, analyzing the varying effects on different LLMs[1].

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 (Touvron et al., 2023), ChatGLM (Zeng et al., 2023), etc. These models demonstrate strong abilities and impressive performance in tasks like conversation and code generation, attracting significant attention from both industry and academia (Zhao et al., 2023).

[1]Our implementation available at https://anonymous.4open.science/r/UBench

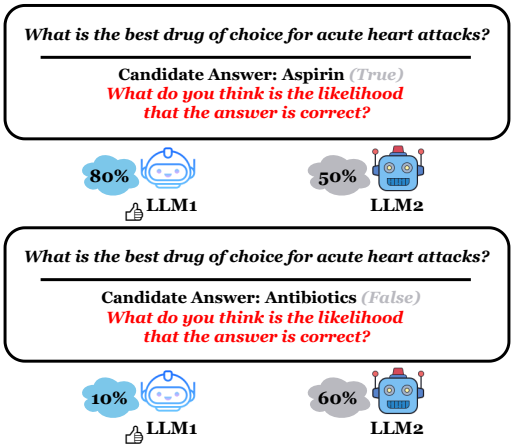

Figure 1: In the context of different candidate answers to the same question, LLMs display different levels of confidence (in other words, uncertainty). Note that LLMs may exhibit consistent levels of confidence for either the wrong answer or the right answer, which we do not want.

Despite their strong capabilities, the opacity of LLMs' internal mechanisms leads to low interpretability, raising questions about their credibility. Specifically, LLMs are prone to generating misinformation without warning, which may manifest as hallucinations (Huang et al., 2023a), biases (Felkner et al., 2023), or disinformation (Lucas et al., 2023). In this scenario, while obtaining LLMs' response to the question, we also hope to know the confidence level of the response to decide whether to trust the information or suggestions provided by LLMs. 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.

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,

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

Table 1: Comparison of different uncertainty estimation methods, where **"Close?"** indicates whether the method is applicable to closed-source models, **"Open?"** indicates whether the method is applicable to open-source models, and **"Single?"** indicates whether only a single inference is required.

these benchmarks often focus solely on the accuracy of the models, 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. Inspired by that, some research has initiated a shift towards assessing the quality of LLMs' output content, considering factors like safety (Zhang et al., 2023) and hallucinations (Li et al., 2023).

On the other hand, uncertainty estimation, as an effective risk assessment method, can reflect the calibration of the model and provide a basis for understanding the reliability of the model's responses. However, traditional uncertainty estimation methods in LLMs are limited due to the challenges in acquiring training data and intermediate outputs, particularly for closed-source LLMs. Therefore, some studies have begun to explore uncertainty estimation methods and benchmarks that are adapted to LLMs. For instance, Ye et al. (2024) work on constructing benchmarks for uncertainty assessment. However, their methodology is limited to white-box LLM. Xiong et al. (2023) utilize prompt to elicit model output confidence, but the work requires multiple sampling and doesn't work well in some circumstances.

To better assess the confidence extent of LLMs' outputs, we propose UBENCH, a new benchmark consisting of four categories of questions, totaling 11,978 multiple-choice questions. Comparison of UBENCH with other different benchmarks is shown in Table 1[2]. UBENCH is designed to support a wide range of open-source and closed-source models, with a focus on efficient inference and scalability. Its novel approach requires only a single sampling instance, thereby significantly reducing the computational cost compared to multiple sampling methods. Meanwhile, this streamlined process maintains evaluation fidelity, ensuring accurate performance metrics.

We test the reliability of 20 popular LLMs on UBENCH. In addition, we further explore the effects of two prompt methods, Chain-of-Thought (CoT) (Wei et al., 2022) and role-playing (Shao et al., 2023), on the reliability of LLMs using GPT-4[3], GLM4-flash and GLM4. We also investigate the impact of changing the temperature parameter. Our main contributions and findings are summarised below:

- We propose UBENCH, a new systematic and automated uncertainty evaluation benchmark for LLMs. We categorize all questions into 4 categories, covering the most common use in practical applications.
- We conduct a comparison of UBENCH with other LLM uncertainty estimation methods and achieve superior results.
- We conduct tests on 20 popular LLMs using UBENCH. The results demonstrate that GPT-4o outperforms other LLMs, followed by Qwen-max and Yi-1.5-34B. Open-source and closed-source LLMs generally exhibit comparable levels of reliability, and the performance tends to increase as the models are upgraded.
- Comparative experiments show that the effects of CoT, role-playing and temperature parameter are various on different LLMs. We analyze the raw responses of LLMs, providing possible explanations for each effect, which helps to clear the way for broader downstream applications of LLMs.

2 Related Work

2.1 Evaluation 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), HELM (Liang et al., 2023), PromptBench

[2]See §4.2 for more results.

[3]The version is 1106-preview.

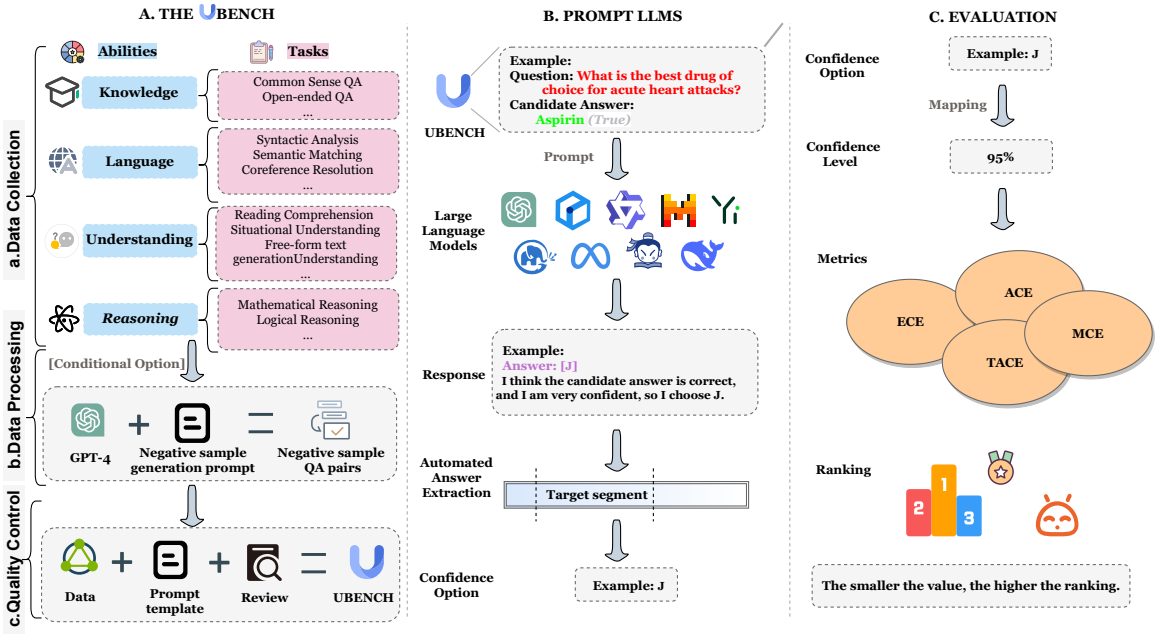

Figure 2: Construction process of UBENCH and systematic, automated LLM uncertainty evaluation framework. The data sources of UBENCH are from multiple types of public datasets, with processed into the uncertainty evaluation format and carefully controlled quality. Then UBENCH is leveraged to compare the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

(Zhu et al., 2023), PandaLM (Wang et al., 2023), and so on. 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. Zhang et al. (2023) present SafetyBench, specifically designed to assess the safety capabilities of LLMs.

2.2 Uncertainty Estimation for LLMs

In general, uncertainty estimation in LLMs measures the confidence level of their predictions. Kuhn et al. (2022) evaluate the semantic uncertainty of language models by clustering answers with similar meanings. Duan et al. (2023) propose to incorporate sentence relevance of other answers when assessing a model's confidence in a specific answer. However, these methods are all logit-based and do not apply to black-box LLMs. Lin et al. (2022a) propose the notion of spoken confidence by prompting LLMs to generate answers and confidence levels. However, the evaluation is specifically tailored to pre-trained language models fine-tuned on a particular dataset, and its generalizability has yet to be considered. Mielke et al. (2022) propose training an external calibrator, but the method is limited by the difficulty of obtaining model representations. SelfCheckGPT (Manakul et al., 2023) introduces a simple, sampling-based approach to identifying potential instances of hallucinations using coherence between generations, but it incurs high computational costs. The above methods require additional computational overhead, whereas ours does not.

2.3 Uncertainty Benchmarks

Existing benchmarks focus on evaluating model-specific scenario uncertainty. Vedantam et al. (2021) propose CURI, which is mainly used to evaluate the performance of models in combination reasoning tasks under uncertainty conditions. Zablotskaia et al. (2023) specifically focus on the uncertainty estimation of models in summary-based benchmarks. Li et al. (2024) is dedicated to quantifying uncertainty in large models of code. Kirchhof et al. (2023) propose a URL benchmark that focuses on evaluating the uncertainty transfer in the representation learning process of models. This benchmark can be used to assess the uncertainty of models on unseen downstream data. However, the benchmark requires additional training, which incurs significant costs for LLM. Unlike the above work, UBENCH aims to provide a comprehensive and accurate uncertainty assessment of the model without incurring additional costs.

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 goal is to assess the reliability of large language models (LLMs) in performing tasks related to comprehension. These tasks encompass various aspects, including multitask 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. We randomly extract samples from 24 open-source datasets[4], with each dataset providing 500 data points (see Figure 3). The datasets sampled for each category are listed in the Appendix F.

Data Processing. Intuitively, a reliable LLM yields lower uncertainty for correct answers and higher uncertainty for incorrect ones. Therefore,

[4]Most of the datasets we get are based on HuggingFace.

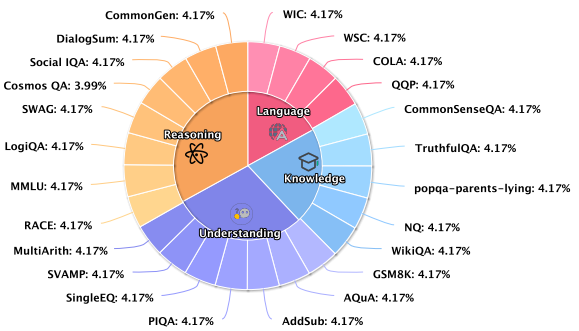

Figure 3: UBENCH covers 4 categories: Language, Knowledge, Understanding, and Reasoning, with a total of 11,978 ten-item multiple-choice questions.

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 gpt4-1106-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. We find that the answers generated by GPT-4 align with our expectations. For the negative samples generated, we regenerate all failed data until they meet the requirements.

Quality Control. To ensure the quality of the datasets, each sample has to be reviewed by two authors, and the sample is only deemed satisfactory if both agree that there are no problems. In case of disagreement, a third author is required to participate in the review to reach a consensus. In this manner, the sample approval rate reaches 99.82%.

3.3 Prompt Design

Like previous studies (Zhang et al., 2023; Zheng et al., 2023b), we evaluate the reliability of LLMs using the prompt engineering method. Initially, we design the prompt with a data sample and all its answers, following instructions to *choose the correct answer and also yield the uncertainty.* However, most LLMs struggle to output a consistent format of uncertainty values, even some of them do not understand the instructions, 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.

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 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 7 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 9. Please note that our experiments are conducted on the original dialogue model without fine-tuning.

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 Error (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 §E.

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

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

where $\mathcal{C}onf(\mathcal{X})$ 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 $\mathcal{C}onf(\mathcal{X})$ is approximated to the midrange of each confidence interval, such as 5% for interval 0%-10%. Finally, $\mathcal{C}onf(\mathcal{X})$ is ranging from 0.05 to 0.95.

4.2 Compared to Other Methods

The comparison of UBENCH with other uncertainty estimation methods is shown in Table 2, with more experimental details provided in Appendix §C. We randomly sample three times, each with 100 data points. We obtain the performance of UBENCH and two baseline methods on the Cosmos QA and SWAG datasets for five open-source and closed-source LLMs. The experimental results indicate that UBENCH achieves the best performance in most settings. 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.

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 commonly used prompt techniques, such as CoT prompt and role-playing 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 GLM-4, 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 §A. Here, we report only the main findings.

Among all LLMs, GPT-4o stands out for its superior performance, followed closely by Qwen-max in second place. GPT-4o particularly excels in the ECE metric, ranking just behind Qwen2.5-72B and leading Qwen-Max by nearly 5 percentage points. In contrast to the top two, which are both

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
	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
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
	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
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 the second best.

closed-source models, the third and fourth places are occupied by two open-source models, Yi 1.5-34B and DeepSeek V2.5. This demonstrates that open-source models can also achieve considerable reliability. In the subsequent rankings, the gap between models is relatively small, with only a 2 percentage point difference between the 5th place GLM4-flash and the 10th place Mistral-7B. This range also includes various model series, representing the general level of LLMs developed by different companies. The next tier includes lower-performing models such as ErnieBot, InternLM, and some models 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.

It is also worth noting that there is no clear correlation between the number of parameters and the reliability of the model. For instance, in the Qwen2.5 series, reliability appears to be roughly positively correlated with the number of parameters, but this trend does not hold for the Llama and InternLM series. Furthermore, when comparing models across different series, such as Llama-3-8B and GLM-4-9B, 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.

In particular, the top-performing model for ECE is Qwen2.5-72B, followed by GPT-4o. Although the difference between Qwen2.5-72B and GPT-4o is marginal, they both significantly outperform Qwen-max and Llama-3.1-70B, which are ranked 3rd and 4th, by around 5 percentage points. 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,

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
GLM-4-9b-chat	17.79	53.78	23.92	24.86	30.09
GPT-4-1106-preview*	17.62	50.54	24.22	28.78	30.29
Mistral-7B-Instruct-v0.2	27.15	40.07	26.81	23.71	30.33
Llama-3.1-8B-Instruct	21.19	56.45	24.06	26.67	32.09
ErnieBot-v4.0*	23.97	52.89	25.81	26.24	32.23
internlm2.5-7B-chat	26.10	55.36	26.01	24.90	33.18
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
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	25.23	70.25	30.13	37.05	40.67

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. Due to resource constraints, * indicates that we conducted experiments on only a randomly selected one-third of the data.

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. The top four positions in both rankings coincide, and their ACE values are not significantly different from each other. When the confidence threshold is set to 0.5, the TACE ranking is generally in line with the ACE ranking, although there are some minor variations. For example, Llama-3-8B and Mistral-7B perform slightly better than their peers in TACE, while GPT-4 performs slightly worse than its peers.

LLMs	ECE	MCE	ACE	TACE	AVG
GPT-4	17.62	50.54	24.22	28.78	30.29
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.

4.3.2 Effects of CoT and Role-Playing Prompts

The ablation experiment results with CoT and role-playing prompt as variables are shown in Table 4 and in Figure 4. Overall, the use of CoT and RP has a positive impact on the reliability of all three 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 3.01% in the scores of the three models, which is higher than the 1.01% 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 5.83% 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. 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. GPT-4's training data may already contain many reasoning and thinking processes similar to CoT, allowing it to 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. GLM-4 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 is shown in Table 5 and in Figure 5. 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

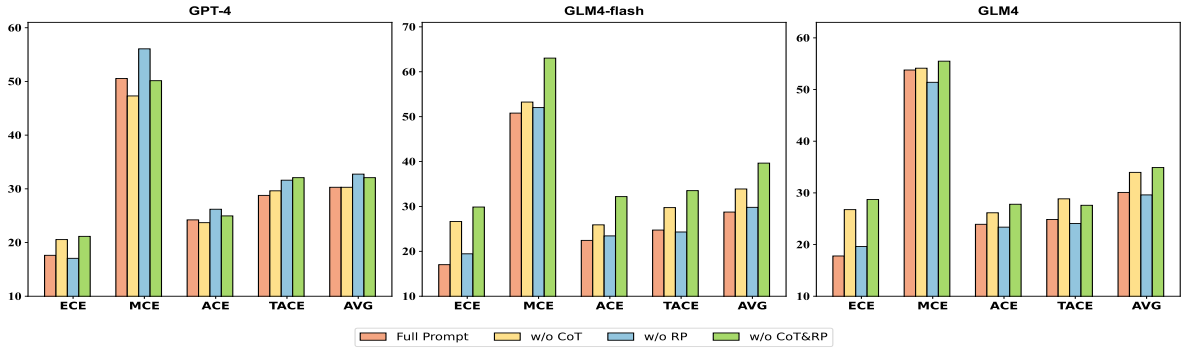

Figure 4: The results of the ablation experiment involve GPT-4, GLM4-flash, and GLM4, studying the effects of CoT and role-playing prompts on LLM reliability. **"w/o"** means removing the prompt, **"CoT"** refers to the Chain-of-Thought prompt, and **"RP"** represents the role-playing prompt (the same applies below).

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. 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.

4.3.4 Case Study

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. Figures 14 and 15 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 (Figure 15), 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 (see Figure 14), 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.

5 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 11,978 ten-choice questions in four categories: knowledge, language, understanding, and reasoning. Comparative experimental results show that our 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 use an example to show the effectiveness of our uncertainty evaluation 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 reliability 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 reliability 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.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, 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.

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, Yingtong 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*.

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)*.

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, Yuantao 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, Jiaxi 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, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée 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.

Tom Kwiatkowski, 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.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language

models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6449–6464, Singapore. Association for Computational Linguistics.

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.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models.

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.

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*.

Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics (TACL)*, 10:857–872.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

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*.

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.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130B: an open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, pages 1–56. OpenReview.net.

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 Additional Experimental Results

A.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 10 to 13.

Performance of Language Subset. In this category, Mistral-7B and Yi-1.5-34B-Chat-16K 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 is lower than that of Llama-3-8B-Instruct. 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. 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.

LLMs	ECE	MCE	ACE	TACE	AVG
GPT-4	19.84	43.91	9.85	6.12	19.93
w/ 0.4	18.95	50.15	22.87	27.95	29.98
w/ 0.8	18.74	52.48	24.23	29.10	31.14
w/ 1.2	19.25	51.90	24.67	31.45	31.82
w/ 1.6	20.79	64.18	29.28	32.22	36.62
w/ 2.0	21.58	63.42	28.34	29.21	35.64
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, where the numbers represent the values of Temperature. Due to computational resource limitations, the results of GPT-4 in this table are tested based on 50 randomly selected positive and negative sample data.

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 gpt4o, outperform most open-source models. Their smaller errors indicate 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% higher than Qwen2.5-72B-Instruct. The larger model, internlm2-5-20b-chat, shows less favorable performance, trailing behind many models with fewer parameters.

Performance of Reasoning Subset. In general, Llama-3-8B-Instruct, GPT-4, and Mistral-7B-Instruct-v0.2 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-Instruct, Llama-3.1-8B-Instruct also performs remarkably well, surpassing several closed-source models or larger models such as GLM4-flash and Qwen-max. Despite excelling in other three categories, DeepSeek V2.5 does not stand out in this task, with performance compa-

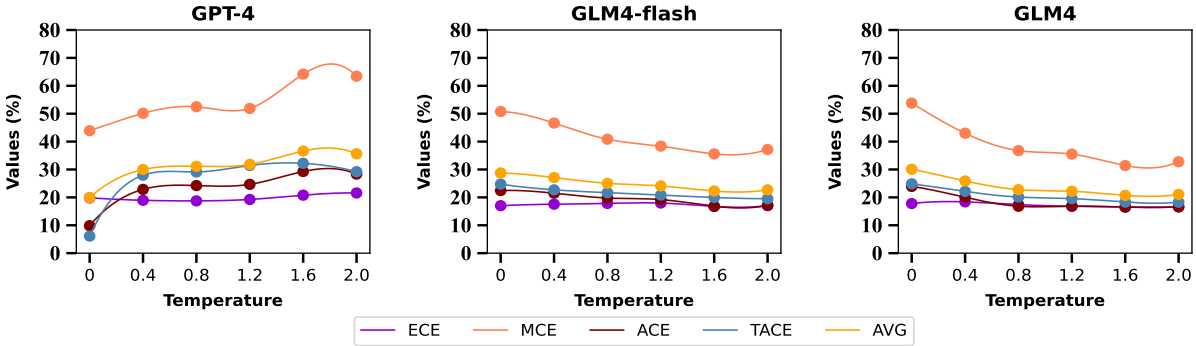

Figure 5: Results of experiments with GPT-4, ChatGLM2, and GLM4, studying the effects of temperature on LLM reliability.

rable 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 Qwen2.5-32B-Instruct significantly affect their rankings, raising concerns about model reliability under extreme conditions.

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
	WikiQA (Yang et al., 2015)	Test
Understanding	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
	DialogSum (Chen et al., 2021)	Test
	CommonGen (Lin et al., 2020)	Dev
Reasoning	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 6: Datasets Overview

B Prompt Templates

The prompt template for evaluating LLMs' reliability is shown in Table 7. The prompt template for generating negative samples is shown in Table 8.

C Baseline Experimental Setup

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

D LLMs in Experiment

The models evaluated in our experiment are shown as Table 9.

E Details for Evaluation Metrics

The detailed calculation of all evaluation metrics are as follows:

$$ECE = \sum_{b=1}^{B} \frac{N_b}{N} \cdot |\mathcal{A}cc(b) - \mathcal{C}onf(b)|, \quad (2)$$

$$MCE = \max_{b=1}^{B} |\mathcal{A}cc(b) - \mathcal{C}onf(b)|, \quad (3)$$

$$ACE = \frac{1}{B} \sum_{b=1}^{B} |\mathcal{A}cc(b) - \mathcal{C}onf(b)|, \quad (4)$$

$$TACE(\theta) = \frac{1}{B_\theta} \sum_{b=1}^{B_\theta} |\mathcal{A}cc(b) - \mathcal{C}onf(b)|, \quad (5)$$

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, $\mathcal{A}cc(b)$ represents the accuracy of the b-th probability interval, $\mathcal{C}onf(b)$ 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.

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 7: 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 8: Prompt template for generating negative samples.

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

F Datasets Overview

An overview of the open source datasets used in UBENCH is shown in Table 6. For datasets where the test subset is fully accessible, we use the test subset; otherwise, we use the dev subset.

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 17 to 23. Additionally, there are two examples of our comparative analysis as shown in Table Tables 14 and 15.

Category	LLMs	ECE (%) ↓	MCE (%) ↓	ACE (%) ↓	TACE (%) ↓	AVG (%) ↓
Language	Mistral-7B-Instruct-v0.2	24.98	41.30	21.62	13.30	26.10
	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
	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-nstruct	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 10: 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
	Llama-3.1-70B-Instruct	16.99	52.79	24.16	27.35	30.32
	ErnieBot	29.75	44.84	25.43	23.15	30.79
	Llama-3.1-8B-Instruct	22.78	61.00	23.94	25.73	33.36
	Mistral-7B-Instruct-v0.2	26.18	48.00	27.58	34.70	34.05
	Qwen2.5-32B-Instruct	38.73	58.17	24.23	17.50	34.66
	GPT-4	18.55	58.89	28.49	35.14	35.27
	Qwen2.5-14B-Instruct	38.46	55.15	27.09	21.22	35.48
	Qwen2.5-7B-Instruct	27.45	59.81	27.19	28.14	35.65
	Llama-3-8B-Instruct	26.18	64.00	25.72	28.40	36.07
	Qwen2-7B-Instruct	38.68	63.00	28.86	24.02	38.64
	ChatGLM3-6B	24.85	68.65	31.07	41.38	41.49

Table 11: 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	15.63	39.59	16.93	18.14	22.57
	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	17.67	62.78	26.41	27.52	33.59
	Mistral-7B-Instruct-v0.2	30.70	44.35	30.98	23.78	33.69
	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
	ChatGLM3-6B	26.81	67.21	30.19	35.12	39.83

Table 12: 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
	Mistral-7B-Instruct-v0.2	25.02	28.81	24.45	21.67	25.89
	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
	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 13: Performance of LLMs on reasoning subset of UBENCH.

[Prompt]:
Read the multiple-choice question, analyze step by step, select the correct option and give option letter e.g. A or B as your answer. Use the following format to provide your answer and confidence level: Explanation: [insert step-by-step analysis here] Answer and Confidence (0-100): [Your answer, please only include the capital letter, e.g. B], [Your confidence level, please only include the numerical number, e.g. 80]%
Note: The confidence level indicates the degree of certainty you have about your answer and is represented as a percentage. For instance, if your confidence level is 80%, it means you are 80% certain that your answer is correct and there is a 20% chance that it may be incorrect.
[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 .
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 14: 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 15: UBENCH (ours) uses Mistral-Instruct-7B as the subject in a case study on the Cosmos QA dataset.

[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 16: 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 17: 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 18: 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 19: 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 20: 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 21: 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 22: 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 23: GPT-4o-2024-08-06 responses on the positive sample of the DialogSum dataset.