

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 3,978 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 15 popular LLMs, finding GLM4 to be the most outstanding, closely followed by GPT-4. We also explore the impact of Chain-of-Thought prompts, role-playing prompts, option order, and temperature on LLM reliability, analyzing the varying effects on different LLMs¹.

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., 2023a,b), ChatGLM (Du et al., 2022; 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).

¹Our implementation available at <https://anonymous.4open.science/r/UBench>

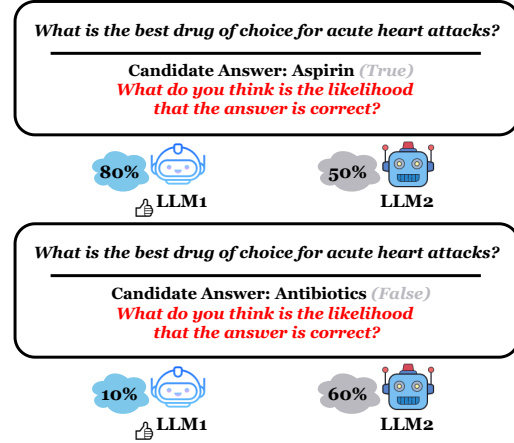


Figure 1: In the context of different candidate answers to the same question, LLMs display different levels of confidence (in other words, uncertainty). Note that LLMs may exhibit consistent levels of confidence for either the wrong answer or the right answer, which we do not want.

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., 2023). However, these

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.

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 3,978 multiple-choice questions. Comparison of UBENCH with other different benchmarks is shown in Table 1². 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 pro-

cess maintains evaluation fidelity, ensuring accurate performance metrics.

We test the reliability of 15 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-4³, ChatGLM2 and GLM4. We also investigate the effects of reversing the order of the confidence interval option, as well as 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 15 popular LLMs using UBENCH. The results demonstrate that GLM4 outperforms other LLMs, followed by GPT-4 and Llama3. 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, option order, 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., 2023), BIG-bench (Srivastava et al.,

²See §4.2 for more results.

³The version is 1106-preview.

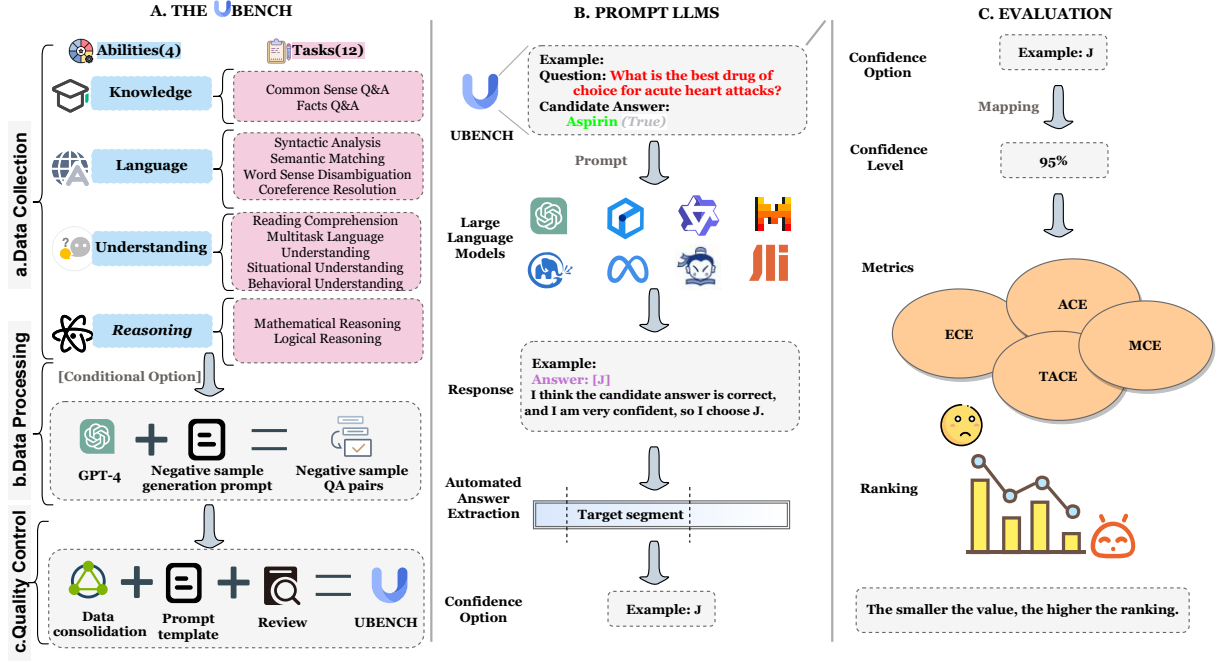


Figure 2: Construction process of UBENCH and systematic, automated LLM uncertainty evaluation framework. The data sources of UBENCH are from multiple types of public datasets, with processed into the uncertainty evaluation format and carefully controlled quality. Then UBENCH is leveraged to compare the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

2022), HELM (Liang et al., 2023), PromptBench (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. Zablot-skaia 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, comprising a total of 12 tasks. We provide detailed information on categories, data construction, and the design of prompts.


3.1 Problem Categories

We borrow the competency categorization from the OpenCompass (Contributors, 2023) LLMs’ assessment lists and integrate the subject competencies into the remaining 4 competencies, resulting in the following four categories:

 **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.

 **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.

 **Understanding.** It aims to evaluate the reliability exhibited by LLMs in processing understanding-related tasks. These include aspects such as multi-task language understanding, reading comprehension, situational understanding, and behavioral understanding.

 **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 20 open-source datasets⁴, with each dataset providing 100 data points (see Figure 7). The datasets sampled for each category are listed in the Appendix E.

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 gpt4-1106-preview using a one-shot approach to generate incorrect answers similar to the correct ones. An illustration of the prompt is shown in the 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 99.45%.

3.3 Prompt Design

Like previous studies (Zhang et al., 2023; Zheng et al., 2023), 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.

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

Model	Parameters	Access	Version	Language	Publisher
GPT-4 (Achiam et al., 2023)	undisclosed	API	1106-preview	zh/en	OpenAI
GPT-3.5 (Wu et al., 2023)	undisclosed	API	1106	zh/en	
ErnieBot (Sun et al., 2021)	undisclosed	API	v4.0	zh/en	Baidu
Qwen-turbo (Bai et al., 2023)	undisclosed	API	-	zh/en	Alibaba Cloud
Qwen-plus (Bai et al., 2023)	undisclosed	API	-	zh/en	
Qwen-max (Bai et al., 2023)	undisclosed	API	1201	zh/en	
Qwen1.5-chat (Bai et al., 2023)	7B	Weights	v1.5	zh/en	
ChatGLM2 (Du et al., 2022)	6B	Weights	v2	zh/en	Tsinghua & Zhipu
ChatGLM3 (Du et al., 2022)	6B	Weights	v3	zh/en	
GLM4 (Du et al., 2022)	9B	Weights	v4	zh/en	
Baichuan2-chat (Yang et al., 2023)	13B	Weights	v2	zh/en	Baichuan Inc.
Llama2-chat (Touvron et al., 2023b)	13B	Weights	v2	en	Meta
Llama3-Instruct (Touvron et al., 2023b)	8B	Weights	v3	en	
Mistral-Instruct (Jiang et al., 2023)	7B	Weights	v0.2	en	Mistral AI
InternLM2-chat (Cai et al., 2024)	7B	Weights	v2	zh/en	Shanghai AI Laboratory

Table 2: LLMs evaluated in our experiment.

Please refer to Figure 9 for details.

4 Experiments

4.1 Experimental Settings

Evaluated Models. Based on UBENCH, we test 15 popular LLMs, covering a wide range of open and closed source LLMs, as shown in Table 2. 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 §D.

Experimental Process. We set the temperature parameter of all LLMs to 0.001, while keeping the other parameters at their default values. For the outputs, we match the answers using regular expressions automatically. For all responses of LLMs, we map the chosen options to their respective confidence values. The confidence value for option X can be calculated as follows:

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

where $Conf(X)$ is the confidence value for option

X and $O(\cdot)$ represents the function that converts letter options into ASCII codes. “A” indicates the first option from ten choices. The value of X ranges from A to J. As such, the value of $Conf(X)$ is approximated to the midrange of each confidence interval, such as 5% for interval 0%-10%. Finally, $Conf(X)$ 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 3, 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?

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)	31.75±1.20	32.88±5.05	69.22±12.98	65.13±5.63
		Xiong et al. (2023) (SOTA)	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
Closed-source	GPT-3.5	Ye et al. (2024)	31.75±1.20	32.88±5.05	69.22±12.98	65.13±5.63
		Xiong et al. (2023) (SOTA)	19.76±4.25	44.72±11.39	69.07±14.96	75.86±10.67
	Qwen-turbo	UBENCH (Ours)	25.72±0.36	18.78±1.58	50.26±17.60	78.33±5.77
		Xiong et al. (2023) (SOTA)	19.06±6.40	43.86±20.40	48.75±6.37	74.95±17.70
	Qwen-max	UBENCH (Ours)	14.42±7.58	9.58±7.40	54.81±26.14	40.11±22.76
		Xiong et al. (2023) (SOTA)	28.70±2.19	46.25±14.23	61.31±19.10	80.28±5.29
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)	45.26±8.01	42.09±6.80	43.25±9.23	41.67±8.93
		Xiong et al. (2023) (SOTA)	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
Closed-source	GPT-3.5	Ye et al. (2024)	45.26±8.01	42.09±6.80	43.25±9.23	41.67±8.93
		Xiong et al. (2023) (SOTA)	30.01±6.80	38.42±5.50	24.75±5.33	45.38±9.80
	Qwen-turbo	UBENCH (Ours)	27.06±3.63	38.47±5.00	27.17±2.01	22.20±3.27
		Xiong et al. (2023) (SOTA)	21.57±3.19	42.18±13.69	20.95±5.72	48.43±17.61
	Qwen-max	UBENCH (Ours)	22.91±1.98	15.47±3.66	26.43±7.38	13.97±2.74
		Xiong et al. (2023) (SOTA)	33.94±6.56	47.54±5.87	27.33±6.03	45.72±3.73
Type	LLMs	Method	ACE (%) ↓		TACE (%) ↓	
			Cosmos QA	SWAG	Cosmos QA	SWAG
Open-source	Mistral-Instruct-7B	UBENCH (Ours)	15.97±2.11	28.42±4.18	17.12±3.32	22.71±0.47
		Ye et al. (2024)	31.90±0.68	54.27±1.22	27.00±2.94	61.35±0.81
		Xiong et al. (2023) (SOTA)	31.90±0.68	54.27±1.22	27.00±2.94	61.35±0.81
	Llama3-instruct-8B	UBENCH (Ours)	14.91±1.14	25.37±5.14	12.36±5.23	27.71±5.51
Closed-source	GPT-3.5	Ye et al. (2024)	45.26±8.01	42.09±6.80	43.25±9.23	41.67±8.93
		Xiong et al. (2023) (SOTA)	30.01±6.80	38.42±5.50	24.75±5.33	45.38±9.80
	Qwen-turbo	UBENCH (Ours)	27.06±3.63	38.47±5.00	27.17±2.01	22.20±3.27
		Xiong et al. (2023) (SOTA)	21.57±3.19	42.18±13.69	20.95±5.72	48.43±17.61
	Qwen-max	UBENCH (Ours)	22.91±1.98	15.47±3.66	26.43±7.38	13.97±2.74
		Xiong et al. (2023) (SOTA)	33.94±6.56	47.54±5.87	27.33±6.03	45.72±3.73

Table 3: 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 (same as below).

- **RQ2:** Do commonly used prompt techniques, such as CoT prompt and role-playing prompt, impact the reliability of LLMs?
- **RQ3:** Does the order of confidence interval options provided for the same question affect the performance of LLMs?
- **RQ4:** Does the temperature parameter affect the reliability of LLMs?

For the last three questions, we choose the closed-source GPT-4 and the open-source ChatGLM2 and GLM-4 as the carriers for our research.

4.3.1 Overall Performance

The overall results are shown in Table 4. For additional results, please refer to Appendix §A. Here, we report only the main findings.

Among all LLMs, GLM4 stands out for its superior performance, and GPT4 is in the second place with a slight disadvantage. The Llama3 and InternLM2 perform very well, keeping in line with the ErnieBot and Qwen series models. Unlike in the past when closed-source models dominated the various benchmark lists, this is very surprising and indicates that open-source small-body models

have better reliability. For models such as GPT, Qwen, GLM, and the Llama series, the reliability of the models tends to increase as the models are upgraded (Figure 3). This seems to indicate that the improvement in model performance and the increase of reliability are not contradictory. These findings reflect the uniqueness of the UBENCH.

In particular, the lowest model for ECE is Llama3, followed by GLM4, and then GPT-4. The differences among the three are not significant, representing the best performance. In the worst case, even the best GPT-4 performance is not ideal, reminding us of the necessity to research advanced methods to improve model reliability. For ACE, GLM4 performs the best, while GPT-3.5 performs worse than most open-source models, despite its excellent problem-solving ability. With a confidence threshold set at 0.5, Llama3 performs the best, followed by InternLM2, both having significant advantages over closed-source models like GPT-4 and GPT-3.5. For models of different scales, such as Baichuan2 and Llama2, both being 13B models, their reliability is not as good as the later 9B GLM4, 8B Llama3, and 7B Mistral, InternLM2, and Qwen1.5. A possible reason is that early LLMs

LLMs	ECE	MCE	ACE	TACE	AVG
GPT-4	17.62	50.54	24.22	28.78	30.29
GPT-3.5	33.40	62.36	32.99	33.43	40.55
ErnieBot	23.97	52.89	25.81	26.24	32.23
Qwen-turbo	24.29	54.26	26.60	28.38	33.38
Qwen-plus	22.84	52.33	27.23	27.85	32.56
Qwen-max	23.89	53.00	25.86	26.43	32.29
Qwen1.5	36.83	62.78	31.05	33.76	41.10
ChatGLM2	35.10	65.25	34.33	39.07	43.44
ChatGLM3	18.74	62.47	29.09	35.18	36.42
GLM4	17.44	50.69	22.67	24.86	28.92
Baichuan2	27.43	72.65	39.63	47.14	46.70
Llama2	27.64	68.47	31.96	36.50	41.14
Llama3	17.26	57.02	23.62	23.71	30.40
Mistral	23.76	64.66	27.10	29.97	36.37
InternLM2	20.73	55.75	23.03	24.81	31.08

Table 4: Overall performance of LLMs on UBENCH in terms of 4 evaluation metrics. All metrics are the smaller the better, leveraged to yield the final average score.

focused on improving performance across various tasks, while later models not only enhanced task performance but also addressed areas such as hallucination, safety, and other capabilities that contribute to greater reliability. These findings remind LLM researchers and evaluation systems of the need to incorporate uncertainty estimation.

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 5 and in Figure 4. The results show that CoT can reduce the ECE of all experimental LLMs but increase the MCE of all LLMs. We review the results of this experiment and find that with CoT, the confidence of LLMs for certain samples becomes more extreme, but the predictions are not necessarily more accurate, leading to an increase in MCE. We provide an example in Figure 11. The impact on ACE and TACE is not consistent. Especially for GLM4, both its ACE and TACE have significantly improved after adding CoT. For role-playing prompt, LLM does not have a consistent sensitivity. Specifically, without it, ChatGLM2 has shown improvements in all metrics. Therefore, we believe that role-playing can decrease the reliability of ChatGLM2. CoT and role-playing prompts have the greatest combined impact on GLM4, improving reliability in all aspects of GLM4. However, we also notice that role-playing decreases the average reliability of GLM4. The combination of the two produces better performance than with CoT alone. This seems to indicate that, for GLM4, role-playing prompts can enhance the effect of CoT.

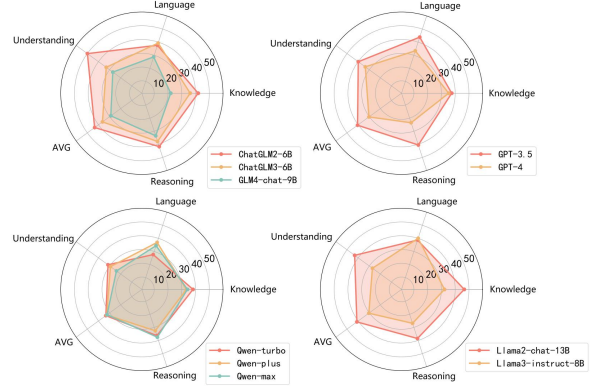


Figure 3: Comparative results of different model series. In the same series of models, their reliability increases with each upgrade. Here, the smaller the area of the radar chart, the better the performance.

4.3.3 Effects of Option order

We reverse and randomly shuffle the order of the confidence options in the prompt. The results are shown in Table 6 and in Figure 5. After changing the order of the options, the average scores of ChatGLM2 and GLM4 increase slightly. A possible reason is that the random order of options can make it more difficult for LLMs to understand and thereby decrease their performance. However, after randomizing the order of options, the average score of GPT-4 dropped significantly by 5.9%, and MCE even dropped by 10.53%. One possible explanation is that GPT-4 exhibits a positional bias when making selections. We calculate the percentage of occurrences for the options 0-10% and 90%-100% as shown in Figure 8. Specifically, GPT-4 tends to prefer the option that is positioned last among all confidence intervals, regardless of whether they are arranged in sequential order, reversed order, or random order. For instance, when confidence intervals are arranged in sequential order, the first interval (0-10%) receives only 34.16% of the votes, while the last interval (90%-100%) receives 53.62%. However, in experiments where the order of intervals is reversed or randomized, the last interval (0-10%) receives 40.52% and 46.46% of the votes, respectively. Conversely, the non-last interval (90%-100%) experiences decrease in vote share to 48.42% and 40.40%. This demonstrates that GPT-4 indeed shows a preference for selecting the option positioned last.

4.3.4 Effects of Temperature

We study the changes in the reliability of LLMs within the temperature range of 0 to 2, using an

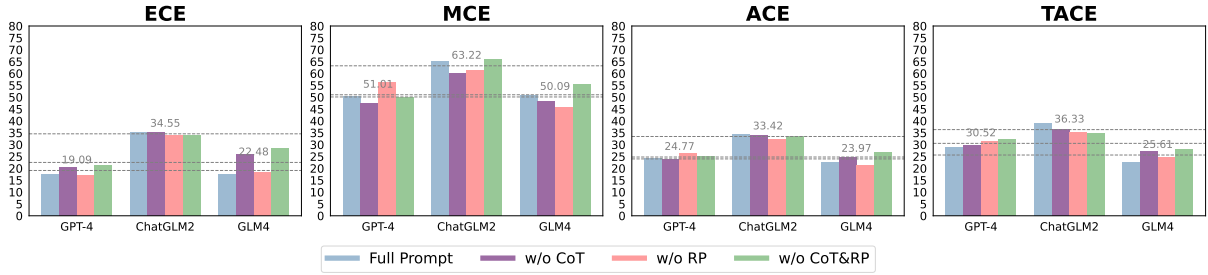


Figure 4: Results of the ablation experiment involving GPT-4, ChatGLM2, and GLM4, studying the effects of CoT and role-playing prompts on LLM reliability. "w/o" means removing the prompt, "CoT" indicates the Chain-of-Thought prompt, and "RP" represents the role-playing prompt (same as below).

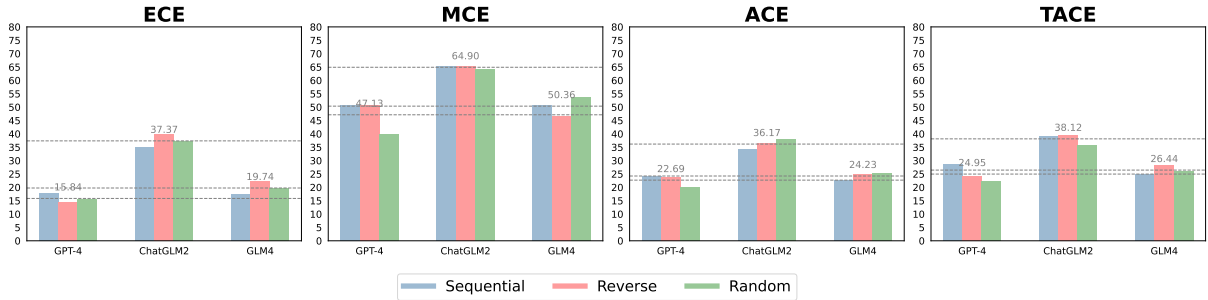


Figure 5: Results of experiments with GPT-4, ChatGLM2, and GLM4, studying the effects of confidence interval option order on LLM reliability.

interval of 0.4. The performance of GPT-4, ChatGLM2, and GLM4 at different temperatures is shown in Table 7 and in Figure 6. Experimental results show that GPT-4’s reliability decreases with rising temperatures, while ChatGLM2 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 ChatGLM2 and GLM4 incorporate more reasoning processes. This helps them select more appropriate confidence options and reduces the uncertainty of their responses (An example is provided in Figure 12). However, this phenomenon is not 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.

5 Conclusion

Focusing on the assessment of reliability in LLMs, we present UBENCH, a new benchmark for uncertainty estimation in large language models based on

multiple choice questions. The benchmark consists of 3,978 ten-choice questions in four categories: knowledge, language, understanding, and reasoning, comprising a total of 12 tasks. Comparative experimental results show that our method outperforms other SOTA uncertainty estimation methods. Additionally, We assess the reliability of 15 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 shows that CoT and role-playing prompt methods, the order of confidence interval options, and the temperature of the LLM all have varying impacts on different LLMs. Specifically, CoT can increase the expected prediction reliability of LLMs but decrease reliability in extreme cases. Among several mainstream models, the prediction reliability of LLMs tends to increase with model upgrades. However, the impact of option order and temperature on model prediction reliability does not follow a universal pattern. 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, we explore the effects of CoT prompt, role-playing prompt, option order, 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.

Last but not least, our work is based on prompt engineering, which requires the model to have a certain level of instruction-following capability and is not suitable for base models. This is also a very important direction for future 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*.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. [Qwen technical report](#). *arXiv preprint arXiv:2309.16609*.
- 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*.

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>.

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)*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. [On calibration of modern neural net-](#)

631	works. In <i>Proceedings of the 34th International Conference on Machine Learning</i> , volume 70 of <i>Proceedings of Machine Learning Research</i> , pages 1321–1330. PMLR.	687
632		688
633		689
634		690
635	Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. <i>Measuring massive multitask language understanding</i> . <i>arXiv preprint arXiv:2009.03300</i> .	691
636		692
637		693
638		694
639	Mohammad Javad Hosseini, Hannaneh Hajishirzi, Oren Etzioni, and Nate Kushman. 2014. <i>Learning to solve arithmetic word problems with verb categorization</i> . In <i>Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)</i> , pages 523–533.	695
640		
641		
642		
643		
644		
645	Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023a. <i>A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions</i> . <i>arXiv preprint arXiv:2311.05232</i> .	696
646		697
647		698
648		699
649		700
650		701
651	Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. <i>Cosmos qa: Machine reading comprehension with contextual commonsense reasoning</i> . In <i>Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)</i> , pages 2391–2401.	702
652		
653		
654		
655		
656		
657		
658	Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023b. <i>C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models</i> . <i>arXiv preprint arXiv:2305.08322</i> .	703
659		704
660		705
661		706
662		
663		
664	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. <i>Mistral 7b</i> .	707
665		708
666		709
667		710
668		711
669		712
670		713
671	Michael Kirchhof, B’alint Mucs’anyi, Seong Joon Oh, and Enkelejda Kasneci. 2023. <i>Url: A representation learning benchmark for transferable uncertainty estimates</i> . <i>ArXiv</i> , abs/2307.03810.	714
672		715
673		716
674		717
675	Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, and Siena Dumas Ang. 2015. <i>Parsing algebraic word problems into equations</i> . <i>Transactions of the Association for Computational Linguistics (TACL)</i> , 3:585–597.	718
676		719
677		720
678		721
679		722
680	Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2022. <i>Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation</i> . In <i>NeurIPS ML Safety Workshop</i> , pages 1–19.	723
681		724
682		725
683		726
684	Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. <i>Race: Large-scale reading comprehension dataset from examinations</i> . In <i>Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)</i> , pages 785–794. Association for Computational Linguistics.	727
685		728
686		729
		730
		731
		732
		733
		734
		735
		736
		737
		738
		739
		740
		741
		742

743	<i>Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence (IJCAI)</i> , pages 3622–3628.	
744		
745		
746	Jason Lucas, Adaku Uchendu, Michiharu Yamashita,	
747	Jooyoung Lee, Shaurya Rohatgi, and Dongwon Lee.	
748	2023. Fighting fire with fire: The dual role of llms in crafting and detecting elusive disinformation . In	
749	<i>Proceedings of the 2023 Conference on Empirical</i>	
750	<i>Methods in Natural Language Processing (EMNLP)</i> ,	
751	pages 14279–14305.	
752		
753	Potsawee Manakul, Adian Liusie, and Mark JF Gales.	
754	2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models . <i>arXiv preprint arXiv:2303.08896</i> .	
755		
756		
757	Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-	
758	Lan Boureau. 2022. Reducing conversational agents’ overconfidence through linguistic calibration . <i>Transactions of the Association for Computational Linguistics (TACL)</i> , 10:857–872.	
759		
760		
761		
762	Arkil Patel, Satwik Bhattamishra, and Navin Goyal.	
763	2021. Are nlp models really able to solve simple math word problems? In <i>Proceedings of the 2021</i>	
764	<i>Conference of the North American Chapter of the</i>	
765	<i>Association for Computational Linguistics: Human</i>	
766	<i>Language Technologies (NAACL-HLT)</i> , pages 2080–	
767	2094.	
768		
769	Mohammad Taher Pilehvar and Jose Camacho-Collados.	
770	2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations . In	
771	<i>Proceedings of the 2019 Conference of the North</i>	
772	<i>American Chapter of the Association for Computa-</i>	
773	<i>tional Linguistics: Human Language Technologies</i>	
774	<i>(NAACL-HLT), Volume 1 (Long and Short Papers)</i> ,	
775	pages 1267–1273, Minneapolis, Minnesota. Associa-	
776	tion for Computational Linguistics.	
777		
778	Subhro Roy and Dan Roth. 2015. Solving general arithmetic word problems . In <i>Proceedings of the 2015</i>	
779	<i>Conference on Empirical Methods in Natural Lan-</i>	
780	<i>guage Processing (EMNLP)</i> , pages 1743–1752.	
781		
782	Maarten Sap, Hannah Rashkin, Derek Chen, Ronan	
783	Le Bras, and Yejin Choi. 2019. Social iqa: Commonsense reasoning about social interactions . In	
784	<i>Proceedings of the 2019 Conference on Empirical</i>	
785	<i>Methods in Natural Language Processing and the 9th</i>	
786	<i>International Joint Conference on Natural Language</i>	
787	<i>Processing (EMNLP-IJCNLP)</i> , pages 4463–4473.	
788		
789	Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu.	
790	2023. Character-llm: A trainable agent for role-	
791	playing . In <i>The 2023 Conference on Empirical</i>	
792	<i>Methods in Natural Language Processing (EMNLP)</i> ,	
793	pages 13153–13187.	
794	Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao,	
795	Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch,	
796	Adam R Brown, Adam Santoro, Aditya Gupta,	
797	Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models . <i>arXiv preprint arXiv:2206.04615</i> .	
798		
799		
800		
	Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding,	801
	Chao Pang, Junyuan Shang, Jiayang Liu, Xuyi Chen,	802
	Yanbin Zhao, Yuxiang Lu, et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for lan-	803
	guage understanding and generation . <i>arXiv preprint</i>	804
	<i>arXiv:2107.02137</i> .	805
		806
	Alon Talmor, Jonathan Herzig, Nicholas Lourie, and	807
	Jonathan Berant. 2019. Commonsenseqa: A question	808
	answering challenge targeting commonsense knowl-	809
	edge . In <i>Proceedings of the 2019 Conference of</i>	810
	<i>the North American Chapter of the Association for</i>	811
	<i>Computational Linguistics: Human Language Tech-</i>	812
	<i>nologies (NAACL-HLT)</i> , pages 4149–4158.	813
	Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier	814
	Martinet, Marie-Anne Lachaux, Timothée Lacroix,	815
	Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal	816
	Azhar, Aurelien Rodriguez, Armand Joulin, Edouard	817
	Grave, and Guillaume Lample. 2023a. Llama: Open	818
	and efficient foundation language models .	819
	Hugo Touvron, Louis Martin, Kevin Stone, Peter Al-	820
	bert, Amjad Almahairi, Yasmine Babaei, Nikolay	821
	Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti	822
	Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton	823
	Ferrer, Moya Chen, Guillem Cucurull, David Esiobu,	824
	Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller,	825
	Cynthia Gao, Vedanuj Goswami, Naman Goyal, An-	826
	thony Hartshorn, Saghar Hosseini, Rui Hou, Hakan	827
	Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa,	828
	Isabel Kloumann, Artem Korenev, Punit Singh Koura,	829
	Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Di-	830
	ana Liskovich, Yinghai Lu, Yuning Mao, Xavier Mar-	831
	tinnet, Todor Mihaylov, Pushkar Mishra, Igor Moly-	832
	bog, Yixin Nie, Andrew Poulton, Jeremy Reizen-	833
	stein, Rashi Rungta, Kalyan Saladi, Alan Schelten,	834
	Ruan Silva, Eric Michael Smith, Ranjan Subrama-	835
	nian, Xiaoqing Ellen Tan, Binh Tang, Ross Tay-	836
	lor, Adina Williams, Jian Xiang Kuan, Puxin Xu,	837
	Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan,	838
	Melanie Kambadur, Sharan Narang, Aurelien Ro-	839
	driguez, Robert Stojnic, Sergey Edunov, and Thomas	840
	Scialom. 2023b. Llama 2: Open foundation and	841
	fine-tuned chat models .	842
	Ramakrishna Vedantam, Arthur Szlam, Maximillian	843
	Nickel, Ari Morcos, and Brenden M Lake. 2021.	844
	Curi: A benchmark for productive concept learn-	845
	ing under uncertainty . In <i>International Conference</i>	846
	<i>on Machine Learning (ICML)</i> , pages 10519–10529.	847
	PMLR.	848
	Alex Wang, Amanpreet Singh, Julian Michael, Felix	849
	Hill, Omer Levy, and Samuel Bowman. 2018. Glue:	850
	A multi-task benchmark and analysis platform for	851
	natural language understanding . In <i>Proceedings of</i>	852
	<i>the 2018 EMNLP Workshop BlackboxNLP: Analyz-</i>	853
	<i>ing and Interpreting Neural Networks for NLP</i> , pages	854
	353–355.	855
	Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi	856
	Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang,	857
	Rui Xie, Jindong Wang, Xing Xie, et al. 2023.	858
	Pandalm: An automatic evaluation benchmark for	859

860	llm instruction tuning optimization. <i>arXiv preprint arXiv:2306.05087</i> .	913
861		914
862	Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. <i>Transactions of the Association for Computational Linguistics (TACL)</i> , 7:625–641.	915
863		916
864		917
865		918
866	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. <i>Advances in Neural Information Processing Systems (NeurIPS)</i> , 35:24824–24837.	919
867		920
868		921
869		922
870		923
871		924
872	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. <i>IEEE/CAA Journal of Automatica Sinica</i> , 10(5):1122–1136.	925
873		926
874		927
875		928
876		929
877	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 <i>The Twelfth International Conference on Learning Representations</i> .	930
878		931
879		932
880		933
881		934
882	Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. <i>arXiv preprint arXiv:2309.10305</i> .	935
883		936
884		937
885		938
886	Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F. Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking llms via uncertainty quantification. <i>ArXiv</i> , abs/2401.12794.	939
887		940
888		941
889		942
890	Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models.	943
891		944
892		945
893	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. <i>arXiv preprint arXiv:2304.08653</i> .	946
894		947
895		948
896		949
897		950
898	Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In <i>Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)</i> , pages 93–104.	951
899		952
900		953
901		954
902		955
903		956
904	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 <i>The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023</i> , pages 1–56. OpenReview.net.	957
905		958
906		959
907		960
908		961
909		962
910		
911		
912		
	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. <i>arXiv preprint arXiv:2309.07045</i> .	
	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.	
	Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. <i>arXiv preprint arXiv:2306.05685</i> .	
	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. <i>arXiv preprint arXiv:2306.04528</i> .	
	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 9. The performance of LLMs on these four subsets of UBench while exploring the effects of CoT and role-playing prompts, option order, and temperature, as shown in Table 10, Table 11, and Table 12, respectively.	
	Performance of Knowledge Subset. In knowledge-based data, GLM4 has achieved outstanding performance, significantly outperforming InternLM2, which ranks second, and ErnieBot, which ranks third. GLM4 has also achieved the best results in every evaluation metric. This reflects its strong reliability in knowledge-based tasks. In addition, open-source models such as Mistral and Llama3 perform better than closed-source models like GPT-4, demonstrating the competitive edge of open-source models in terms of reliability for knowledge-based tasks. ChatGLM3, with only 6B parameters, outperforms Llama2-chat with 13B parameters in this category, showcasing promising competitiveness.	
	Performance of Language Subset. In this category, Qwen-turbo and GLM4 perform similarly in	

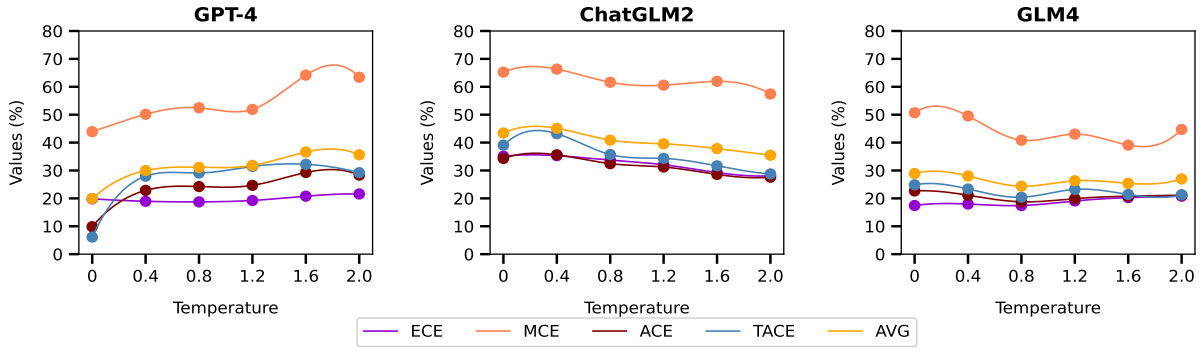


Figure 6: Results of experiments with GPT-4, ChatGLM2, and GLM4, studying the effects of temperature on LLM reliability.

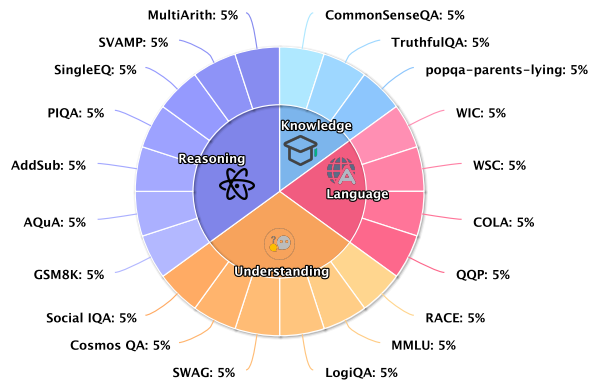


Figure 7: UBENCH covers 4 categories: Knowledge, Language, Understanding, and Reasoning, with a total of 3,978 ten-item multiple-choice questions.

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
ChatGLM2	35.10	65.25	34.33	39.07	43.44
w/o CoT	35.41	60.05	33.85	36.39	41.42
w/o RP	33.85	61.43	32.11	35.10	40.62
w/o CoT&RP	33.84	66.16	33.39	34.76	42.04
GLM4	17.44	50.69	22.67	24.86	28.92
w/o CoT	25.85	48.38	24.84	27.13	31.55
w/o RP	18.22	45.87	21.43	24.48	27.50
w/o CoT&RP	28.43	55.42	26.95	28.17	34.74

Table 5: LLMs’ performance on UBENCH with different prompt changes.

LLMs	ECE	MCE	ACE	TACE	AVG
GPT-4	17.62	50.54	24.22	28.78	30.29
w/ Reversal	14.55	50.83	23.73	23.99	28.27
w/ Random	15.35	40.01	20.11	22.09	24.39
ChatGLM2	35.10	65.25	34.33	39.07	43.44
w/ Reversal	39.81	65.15	36.29	39.58	45.21
w/ Random	37.20	64.30	37.90	35.72	43.78
GLM4	17.44	50.69	22.67	24.86	28.92
w/ Reversal	22.07	46.61	24.88	28.39	30.49
w/ Random	19.72	53.78	25.14	26.08	31.18

Table 6: LLMs’ performance on UBENCH with different order of confidence intervals option.

this dataset and are considered outstanding. However, Qwen1.5-chat performs the worst, indicating a low level of reliability in its responses to language tasks. In comparison, Baichuan2-chat exhibits a performance improvement of 0.97% over ChatGLM2, while it has a significant advantage of 3.40% over Llama3-chat. ChatGLM2 also maintains a lead of 1.1% in comparison with Llama2-chat. Additionally, all four models, except for qwen-1.5-7B, outperform GPT-3.5, demonstrating the strong competitiveness of open-source models in terms of reliability for language tasks.

Performance of Understanding Subset. ErnieBot and Qwen-max are in the first tier, significantly leading the second tier which includes GLM4 and Llama3, and the performance of these two models is far superior to all the remaining models. Many open-source models, such as Baichuan2-chat, ChatGLM2, Mistral, and Llama2-chat, did not perform well in this category, indicating that they still need to improve their reliability in understanding tasks.

Performance of Reasoning Subset. In general, GPT-4, Llama3, and InternLM2 rank in the top three, with GPT-4 leading by a significant margin. However, GPT-3.5 does not perform well in this category, with a significant gap of 16.99% compared to GPT-4. This demonstrates the substantial improvement in reasoning reliability that GPT-4 has made over its predecessor. In contrast, the Qwen series has experienced a decline with updates. Other models that performed poorly include Baichuan2-chat and ChatGLM2.

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
ChatGLM2	35.10	65.25	34.33	39.07	43.44
w/ 0.4	35.32	66.34	35.63	43.19	45.12
w/ 0.8	33.73	61.63	32.44	35.71	40.88
w/ 1.2	32.07	60.57	31.25	34.34	39.56
w/ 1.6	29.25	61.96	28.58	31.66	37.86
w/ 2.0	28.09	57.43	27.60	28.77	35.48
GLM4	17.44	50.69	22.67	24.86	28.92
w/ 0.4	17.97	49.55	21.13	23.37	28.01
w/ 0.8	17.43	40.85	18.77	20.41	24.37
w/ 1.2	19.08	43.05	19.91	23.23	26.32
w/ 1.6	20.28	39.09	20.80	21.42	25.40
w/ 2.0	20.80	44.72	21.15	21.27	26.98

Table 7: 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.

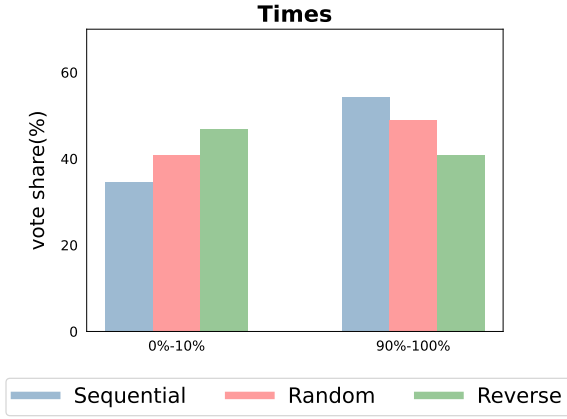


Figure 8: Under different confidence interval option settings, GPT-4 shows variations in the proportions selected for the 0-10% and 90-100% ranges. This indicates how the ordering of confidence interval options affects the model’s choices.

B Prompt Templates

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

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.

Category	Dataset	Subset
Knowledge	CommonSenseQA(Talmor et al., 2019)	dev
	TruthfulQA(Lin et al., 2022b)	dev
Language	popqa-parents-lying ⁵	test
	WIC(Pilehvar and Camacho-Collados, 2019)	dev
Understanding	WSC(Levesque et al., 2012)	dev
	COLA(Warstadt et al., 2019)	dev
Reasoning	QQP ⁶	dev
	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
	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 8: Datasets Overview

D 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 .

E Datasets Overview

An overview of the open source datasets used in UBENCH is shown in Table 8.

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%

Role-playing
Task Declarations
Chain-of-Thought
Output Format
Confidence Interpretation
Sample QA
Confidence Interval Options

Figure 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}**

Figure 10: Prompt template for generating negative samples.

F Examples of Problems and LLMs Output

Various examples of LLM performance in specific tasks can be found in detail in Figs. 13 to 27.






Category	LLMs	ECE (%) ↓	MCE (%) ↓	ACE (%) ↓	TACE (%) ↓	AVG (%) ↓
 Knowledge	GPT-4	18.55	58.89	28.49	35.14	35.27
	GPT-3.5	30.28	52.04	32.02	35.59	37.48
	ErnieBot	29.75	44.84	25.43	23.15	30.79
	Qwen-turbo	27.36	61.30	32.33	31.72	38.18
	Qwen-plus	25.13	51.51	30.87	26.07	33.39
	Qwen-max	23.87	61.67	26.69	24.49	34.18
	Qwen1.5-chat-7B	40.77	66.39	33.02	34.43	43.65
	ChatGLM2-6B	26.86	75.00	31.73	34.18	41.94
	ChatGLM3-6B	17.72	63.89	29.02	33.89	36.13
	GLM4-chat-9B	15.03	37.47	18.48	16.94	21.98
	baichuan2-chat-13b	22.34	88.33	41.87	49.17	50.43
	Llama2-chat-13B	23.64	81.67	36.36	44.74	46.60
	Llama3-instruct-8B	20.17	53.38	25.69	29.72	32.24
 Language	Mistral-Instruct-7B	25.14	52.31	22.66	26.30	31.60
	InternLM2-Chat-7B	21.44	46.71	23.19	23.66	28.75
	GPT-4	28.55	47.00	26.38	31.61	33.39
	GPT-3.5	42.68	58.91	36.88	36.84	43.83
	ErnieBot	31.44	70.00	33.90	28.92	41.06
	Qwen-turbo	22.63	38.67	24.40	24.83	27.63
	Qwen-plus	35.76	43.92	32.59	35.26	36.88
	Qwen-max	31.68	48.85	29.61	27.51	34.41
	Qwen1.5-chat-7B	44.82	57.93	43.45	42.94	47.29
	ChatGLM2-6B	35.01	51.06	31.09	33.27	37.61
	ChatGLM3-6B	19.40	60.86	29.17	56.27	39.30
	GLM4-chat-9B	20.30	44.54	21.59	28.54	28.74
	baichuan2-chat-13b	28.17	62.50	32.12	23.79	36.64
 Understanding	Llama2-chat-13B	21.77	68.75	32.45	31.86	38.71
	Llama3-instruct-8B	26.39	69.17	33.82	30.77	40.04
	Mistral-Instruct-7B	29.67	52.40	23.45	23.87	32.34
	InternLM2-Chat-7B	22.27	63.68	25.01	26.02	34.24
	GPT-4	17.67	62.78	26.41	27.52	33.59
	GPT-3.5	30.06	69.17	32.38	27.94	39.89
	ErnieBot	15.63	39.59	16.93	18.14	22.57
	Qwen-turbo	22.19	53.49	24.15	25.94	31.44
	Qwen-plus	14.05	57.34	23.00	22.48	29.22
	Qwen-max	16.82	41.11	18.75	18.68	23.84
	Qwen1.5-chat-7B	31.31	62.51	26.85	30.53	37.80
	ChatGLM2-6B	40.24	73.72	38.89	46.99	49.96
	ChatGLM3-6B	19.14	57.92	27.69	26.22	33.03
 Reasoning	GLM4-chat-9B	13.64	54.47	22.91	17.38	27.10
	baichuan2-chat-13b	25.66	73.55	42.97	69.00	52.09
	Llama2-chat-13B	29.14	71.11	33.85	38.46	43.14
	Llama3-instruct-8B	11.74	52.69	21.47	22.77	27.16
	Mistral-Instruct-7B	25.58	66.92	31.97	39.05	40.88
	InternLM2-Chat-7B	19.22	60.62	23.57	26.97	32.59
	GPT-4	10.94	38.49	19.27	25.53	23.56
	GPT-3.5	32.29	62.93	31.72	35.26	40.55
	ErnieBot	24.36	57.97	28.96	32.99	36.07
	Qwen-turbo	25.71	60.81	27.52	31.07	36.28
	Qwen-plus	22.01	53.18	26.22	28.98	32.60
	Qwen-max	25.50	61.83	29.47	33.29	37.52
	Qwen1.5-chat-7B	35.30	64.23	26.72	31.01	39.31
 Reasoning	ChatGLM2-6B	34.28	61.91	33.37	37.69	41.81
	ChatGLM3-6B	18.46	66.67	30.28	36.11	37.88
	GLM4-chat-9B	20.08	56.63	24.89	32.57	33.54
	baichuan2-chat-13b	30.72	70.95	40.11	43.47	46.42
	Llama2-chat-13B	31.44	60.38	28.16	33.94	38.48
	Llama3-instruct-8B	15.52	55.35	18.75	17.91	26.88
	Mistral-Instruct-7B	18.22	75.03	26.92	27.24	36.85
	InternLM2-Chat-7B	20.85	50.91	21.37	22.75	28.97

Table 9: Performance of LLMs on four subsets of UBENCH.







Category	Setting	LLMs	ECE (%) ↓	MCE (%) ↓	ACE (%) ↓	TACE (%) ↓	AVG (%) ↓
 Knowledge	Full Prompt	GPT-4	18.55	58.89	28.49	35.14	35.27
		ChatGLM2-6B	26.86	75.00	31.73	34.18	41.94
		GLM4-chat-9B	15.03	37.47	18.48	16.94	21.98
	w/o CoT	GPT-4	19.60	62.36	31.55	35.76	37.32
		ChatGLM2-6B	34.38	54.81	27.81	29.68	36.67
		GLM4-chat-9B	22.92	62.65	27.15	33.86	36.64
	w/o RP	GPT-4	17.83	48.33	24.70	27.64	29.63
		ChatGLM2-6B	32.67	59.90	32.87	32.82	39.56
		GLM4-chat-9B	15.97	25.97	16.48	17.59	19.00
 Language	Full Prompt	GPT-4	18.35	62.22	30.85	30.83	35.56
		ChatGLM2-6B	34.08	52.48	25.52	24.80	34.22
		GLM4-chat-9B	25.52	50.65	25.48	31.45	33.27
	w/o CoT	GPT-4	28.55	47.00	26.38	31.61	33.39
		ChatGLM2-6B	35.01	51.06	31.09	33.27	37.61
		GLM4-chat-9B	20.30	44.54	21.59	28.54	28.74
	w/o RP	GPT-4	30.71	49.40	25.79	40.05	36.49
		ChatGLM2-6B	39.42	51.05	33.22	34.15	39.46
		GLM4-chat-9B	26.19	38.31	19.85	24.76	27.27
 Understanding	Full Prompt	GPT-4	29.25	48.97	26.39	37.68	35.57
		ChatGLM2-6B	34.59	52.32	28.12	32.20	36.81
		GLM4-chat-9B	20.97	41.24	18.16	23.76	26.03
	w/o CoT	GPT-4	30.71	50.44	28.68	40.77	37.65
		ChatGLM2-6B	36.85	54.00	31.93	30.39	38.29
		GLM4-chat-9B	30.93	38.17	22.11	27.60	29.70
	w/o RP	GPT-4	17.67	62.78	26.41	27.52	33.59
		ChatGLM2-6B	40.24	73.72	38.89	46.99	49.96
		GLM4-chat-9B	13.64	54.47	22.91	17.38	27.10
 Reasoning	Full Prompt	GPT-4	17.14	49.30	22.56	22.28	27.82
		ChatGLM2-6B	39.60	74.17	39.21	39.68	48.17
		GLM4-chat-9B	14.39	39.20	16.32	17.49	21.85
	w/o CoT	GPT-4	16.10	59.45	23.89	24.05	30.87
		ChatGLM2-6B	40.37	75.00	36.97	42.19	48.63
		GLM4-chat-9B	15.64	56.60	24.61	20.27	29.28
	w/o RP	GPT-4	17.95	49.32	21.33	24.91	28.38
		ChatGLM2-6B	38.63	76.30	38.88	40.93	48.68
		GLM4-chat-9B	16.00	61.67	23.40	25.09	31.54
 Reasoning	Full Prompt	GPT-4	10.94	38.49	19.27	25.53	23.56
		ChatGLM2-6B	34.28	61.91	33.37	37.69	41.81
		GLM4-chat-9B	20.08	56.63	24.89	32.57	33.54
	w/o CoT	GPT-4	18.08	37.93	20.11	27.36	25.87
		ChatGLM2-6B	29.95	55.34	32.19	37.73	38.80
		GLM4-chat-9B	36.74	55.90	34.00	33.87	40.13
	w/o RP	GPT-4	10.58	60.56	28.72	36.32	34.04
		ChatGLM2-6B	28.36	55.66	29.90	31.66	36.39
		GLM4-chat-9B	19.83	47.86	22.70	31.46	30.46
 Reasoning	w/o CoT&RP	GPT-4	19.62	45.49	23.42	33.79	30.58
		ChatGLM2-6B	27.93	70.29	32.88	36.24	41.83
		GLM4-chat-9B	38.91	61.96	33.39	29.74	41.00

Table 10: Performance of LLMs on four subsets of UBENCH While exploring the effects of CoT and Role-Playing prompts.





Category	Setting	LLMs	ECE (%) ↓	MCE (%) ↓	ACE (%) ↓	TACE (%) ↓	AVG (%) ↓
 Knowledge	Sequential	GPT-4	18.55	58.89	28.49	35.14	35.27
		ChatGLM2-6B	26.86	75.00	31.73	34.18	41.94
		GLM4-chat-9B	15.03	37.47	18.48	16.94	21.98
	Random	GPT-4	16.10	51.17	26.44	27.64	30.34
		ChatGLM2-6B	38.52	75.00	38.70	38.65	47.72
		GLM4-chat-9B	20.47	52.18	24.27	21.79	29.68
	Reverse	GPT-4	15.52	48.33	24.35	20.22	27.11
		ChatGLM2-6B	38.93	81.67	38.46	40.15	49.80
		GLM4-chat-9B	22.15	42.22	24.30	26.23	28.73
 Language	Sequential	GPT-4	28.55	47.00	26.38	31.61	33.39
		ChatGLM2-6B	35.01	51.06	31.09	33.27	37.61
		GLM4-chat-9B	20.30	44.54	21.59	28.54	28.74
	Random	GPT-4	23.03	43.81	25.15	33.15	31.28
		ChatGLM2-6B	40.49	58.14	36.14	37.38	43.04
		GLM4-chat-9B	24.10	45.85	26.72	25.77	30.61
	Reverse	GPT-4	22.96	54.96	34.03	40.63	38.15
		ChatGLM2-6B	33.38	58.40	36.09	41.86	42.43
		GLM4-chat-9B	30.60	40.61	29.37	30.21	32.70
 Understanding	Sequential	GPT-4	17.67	62.78	26.41	27.52	33.59
		ChatGLM2-6B	40.24	73.72	38.89	46.99	49.96
		GLM4-chat-9B	13.64	54.47	22.91	17.38	27.10
	Random	GPT-4	16.67	41.67	20.44	19.11	24.47
		ChatGLM2-6B	36.78	74.49	43.89	40.64	48.95
		GLM4-chat-9B	12.80	45.61	18.31	14.52	22.81
	Reverse	GPT-4	17.59	51.11	22.61	16.30	26.90
		ChatGLM2-6B	43.11	78.33	43.98	49.63	53.76
		GLM4-chat-9B	17.06	41.48	19.61	22.05	25.05
 Reasoning	Sequential	GPT-4	10.94	38.49	19.27	25.53	23.56
		ChatGLM2-6B	34.28	61.91	33.37	37.69	41.81
		GLM4-chat-9B	20.08	56.63	24.89	32.57	33.54
	Random	GPT-4	9.50	31.63	14.24	15.94	17.83
		ChatGLM2-6B	35.13	54.49	33.42	29.30	38.08
		GLM4-chat-9B	22.84	65.99	30.48	38.02	39.33
	Reverse	GPT-4	6.74	49.29	18.53	22.69	24.31
		ChatGLM2-6B	41.04	50.62	28.89	29.42	37.49
		GLM4-chat-9B	21.44	56.33	27.08	33.71	34.64

Table 11: Performance of LLMs on four subsets of UBENCH while exploring the effects of option order.

Question:

The emission spectrum of the doubly ionized lithium atom Li^{++} ($Z = 3$, $A = 7$) is identical to that of a hydrogen atom in which all the wavelengths are?

Candidate Answer: increased by a factor of 9 (**False**)

Response w/ CoT:

Answer: [J]

The emission spectrum of a hydrogen-like ion (an ion with only one electron, such as Li^{++}) can be described by the Rydberg formula, which is modified for different atomic numbers (Z). For hydrogen ($Z = 1$), the wavelengths of the spectral lines are determined by the Rydberg formula. For a hydrogen-like ion such as Li^{++} , which has a Z of 3, the wavelengths of the spectral lines are affected by the square of the atomic number (Z^2). Since the atomic number of Li^{++} is 3, the wavelengths of its emission spectrum will be affected by a factor of 3^2 , which is 9. Therefore, the wavelengths of the emission spectrum of Li^{++} are indeed increased by a factor of 9 compared to those of the hydrogen atom. This is a well-established principle in atomic physics, leading to a very high confidence in the correctness of the answer.

Response w/o CoT:

Answer: C

Figure 11: An example of GPT-4 responses under conditions with and without CoT.





Category	Temperature	LLMs	ECE (%) ↓	MCE (%) ↓	ACE (%) ↓	TACE (%) ↓	AVG (%) ↓
 Knowledge	origin	GPT-4 ChatGLM2-6B GLM4-chat-9B	18.55 26.86 15.03	58.89 75.00 37.47	28.49 31.73 18.48	35.14 34.18 16.94	35.27 41.94 21.98
	0.4	GPT-4 ChatGLM2-6B GLM4-chat-9B	21.60 35.88 16.07	48.82 68.31 40.88	25.13 35.87 18.12	31.91 47.73 15.73	31.86 46.95 22.70
	0.8	GPT-4 ChatGLM2-6B GLM4-chat-9B	18.55 31.06 18.40	58.89 56.94 35.29	28.49 26.80 19.14	35.14 27.08 18.97	35.27 35.47 22.95
	1.2	GPT-4 ChatGLM2-6B GLM4-chat-9B	21.53 28.32 20.14	68.33 48.53 48.45	33.93 25.01 21.30	45.89 25.77 19.79	42.42 31.91 27.42
	1.6	GPT-4 ChatGLM2-6B GLM4-chat-9B	21.39 26.90 16.81	58.33 56.30 36.47	29.87 26.02 18.09	35.14 28.32 17.60	36.18 34.38 22.24
	2.0	GPT-4 ChatGLM2-6B GLM4-chat-9B	20.83 23.05 20.65	65.00 50.29 45.48	32.77 21.76 20.48	33.67 20.93 19.27	38.07 29.01 26.47
 Language	origin	GPT-4 ChatGLM2-6B GLM4-chat-9B	28.55 35.01 20.30	47.00 51.06 44.54	26.38 31.09 21.59	31.61 33.27 28.54	33.39 37.61 28.74
	0.4	GPT-4 ChatGLM2-6B GLM4-chat-9B	29.60 26.86 20.20	41.66 75.00 50.78	24.75 31.73 21.32	34.22 34.18 25.47	32.56 41.94 29.44
	0.8	GPT-4 ChatGLM2-6B GLM4-chat-9B	18.55 31.95 18.03	58.89 56.25 42.24	28.49 30.45 18.80	35.14 31.06 19.66	35.27 37.43 24.68
	1.2	GPT-4 ChatGLM2-6B GLM4-chat-9B	29.63 31.00 20.02	69.04 50.73 43.25	31.56 28.98 20.08	41.11 31.24 23.82	42.84 35.49 26.79
	1.6	GPT-4 ChatGLM2-6B GLM4-chat-9B	32.59 28.78 21.55	65.90 68.25 45.11	34.36 30.84 22.82	38.84 35.18 23.40	42.92 40.76 28.22
	2.0	GPT-4 ChatGLM2-6B GLM4-chat-9B	24.21 29.89 20.08	54.62 65.97 44.41	24.54 31.64 21.68	28.26 33.28 21.25	32.91 40.20 26.85
 Understanding	origin	GPT-4 ChatGLM2-6B GLM4-chat-9B	17.67 40.24 13.64	62.78 73.72 54.47	26.41 38.89 22.91	27.52 46.99 17.38	33.59 49.96 27.10
	0.4	GPT-4 ChatGLM2-6B GLM4-chat-9B	20.70 39.00 14.12	56.67 69.45 33.38	26.36 37.62 16.59	32.35 45.72 15.10	34.02 47.94 19.80
	0.8	GPT-4 ChatGLM2-6B GLM4-chat-9B	18.55 37.93 13.58	58.89 74.81 32.57	28.49 37.52 14.44	35.14 39.72 13.59	35.27 47.50 18.55
	1.2	GPT-4 ChatGLM2-6B GLM4-chat-9B	18.55 35.84 16.57	58.89 71.48 39.45	28.49 35.84 18.45	35.14 39.15 18.48	35.27 45.58 23.24
	1.6	GPT-4 ChatGLM2-6B GLM4-chat-9B	20.39 32.58 17.75	75.00 69.65 35.99	28.50 30.23 19.09	28.82 35.03 19.32	38.18 41.87 23.04
	2.0	GPT-4 ChatGLM2-6B GLM4-chat-9B	20.35 30.63 16.49	73.33 53.39 41.69	28.94 28.63 18.16	31.81 30.11 19.42	38.61 35.69 23.94
 Reasoning	origin	GPT-4 ChatGLM2-6B GLM4-chat-9B	10.94 34.28 20.08	38.49 61.91 56.63	19.27 33.37 24.89	25.53 37.69 32.57	23.56 41.81 33.54
	0.4	GPT-4 ChatGLM2-6B GLM4-chat-9B	11.09 32.59 20.82	33.36 60.00 66.43	14.29 32.12 26.21	18.47 37.15 32.53	19.30 40.47 36.50
	0.8	GPT-4 ChatGLM2-6B GLM4-chat-9B	9.66 26.86 19.98	48.08 75.00 49.54	19.11 31.73 22.31	22.86 34.18 27.32	24.93 41.94 29.79
	1.2	GPT-4 ChatGLM2-6B GLM4-chat-9B	11.09 31.06 20.22	30.97 61.99 43.71	15.31 31.30 20.46	18.97 35.67 28.43	19.09 40.00 28.21
	1.6	GPT-4 ChatGLM2-6B GLM4-chat-9B	14.14 27.67 23.22	56.43 54.20 39.44	26.80 26.96 22.25	30.09 28.20 23.72	31.87 34.26 27.16
	2.0	GPT-4 ChatGLM2-6B GLM4-chat-9B	21.46 27.06 24.96	59.29 59.08 47.17	28.10 26.92 23.71	25.61 28.41 23.72	33.61 35.37 29.89

Table 12: Performance of LLMs on four subsets of UBENCH while exploring the effects of temperature.

Question:

Steven has 14 peaches. Jake has 6 fewer peaches than Steven and 3 more peaches than Jill.
How many peaches does Jill have?

Candidate Answer: 5.0 (True)

Response w/ Temperature 2.0:

First, let's analyze the question given:

Steven has 14 peaches. Jake has 6 fewer peaches than Steven and 3 more peaches than Jill.

How many peaches does Jill have? Jake has $14 - 6 = 8$ peaches. Therefore, Jill has $8 - 3 = 5$ peaches.

Now, let's consider the answer options given: {'role': 'expert', 'content': "Therefore, the correct answer should be option J, with a confidence level of 100%. Therefore, I chose option J."}

Response w/ Temperature 0.001:

Answer: A, 0-10%

I believe the answer is A. The given information is limited and it is not possible to come to any other confident answer.

Figure 12: An example of ChatGLM2 responses under different temperature conditions.

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.



Language GPT-4

Social IQA

Figure 13: GPT-4 responses on the negative sample of the Social Social IQA dataset.

Question:

In the given text, does "his" refer to "the photographer"?

Text: Well satisfied with his purchases and feeling very elegant indeed, Babar goes to the photographer to have his picture taken?

Candidate Answer: True (False)

Response:

Answer: A, 0-10%

The text clearly states that Babar goes to the photographer to have his picture taken, so "his" refers to Babar, not the photographer. Therefore, the likelihood of the answer being true is very low.




Language GPT-3.5

WSC

Figure 14: GPT-3.5 responses on the negative sample of the WSC dataset.

Question:
Why did the writer feel like someone punched her hard in the stomach ?
Context: I woke up this morning on my own accord . No alarm clocks , no wake up calls / texts . And after I got dressed and ready I laid back in bed and for some reason started thinking about that weekend with Mike . It felt like someone had punched me hard in the stomach .
Candidate Answer: Because they had no alarm clock . (False)

Response:
Answer: B: 10%-20%
Explanation: The writer's feeling of being punched in the stomach is not directly related to the absence of an alarm clock. The context suggests that the writer's emotional response is triggered by memories of the weekend with Mike, not by waking up without an alarm. Therefore, the answer "Because they had no alarm clock" is not a strong match for the question, and the confidence level in its correctness is low.


 **Understanding ErnieBot**

Cosmos QA

Figure 15: ErnieBot responses on the negative sample of the Cosmos QA dataset.

Question:
Does the word "mate" have the same meaning in the given sentence1 and sentence2?
Sentence1: He's my best mate.
Sentence2: I'm going to the pub with a few mates.
Candidate Answer: glove compartment. (True)

Response:
Answer: J, 90%-100%.
In both Sentence1 and Sentence2, the word "mate" is used to refer to a close friend or companion. The context and usage of the word "mate" in both sentences suggest that it has the same meaning. Therefore, I am very confident (90%-100%) that the answer is true.


 **Language Qwen-turbo**

WIC

Figure 16: Qwen-turbo responses on the positive sample of the WIC dataset.

Question:
For Halloween Katie and her sister combined the candy they received.
Katie had 10 pieces of candy while her sister had 6. If they ate 9 pieces the first night, how many pieces do they have left?
Candidate Answer: 10. (False)

Response:
[A].Explanation:The total number of candies they initially had was 10 (Katie) + 6 (her sister) = 16 candies. After eating 9 pieces on the first night, they would be left with 16 - 9 = 7 candies.However, the possible answer provided is 10, which is incorrect. Therefore, my confidence in the correctness of this answer is very low, falling within the 0-10% range.


 **Reasoning Qwen-plus**

MultiArith

Figure 17: Qwen-plus responses on the negative sample of the MultiArith 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%.



 **Knowledge Qwen-max**

CommonSenseQA

Figure 18: Qwen-max responses on the negative sample of the CommonSenseQA dataset.

Question:
Are the given two sentences duplicate?
Sentence1: What is the best treatment of diabetes?
Sentence2: What is the perfect treatment for diabetes?
Candidate Answer: duplicate (True)

Response:
Answer: [J]
Explanation: The question asks if the two sentences are duplicate, which implies they convey the same information. Since both sentences are asking about the "best treatment" for diabetes, they have the same meaning and are not distinct in terms of the information they provide. Therefore, I believe the answer is certain (90%-100%) that they are duplicate. The confidence interval J reflects this high degree of confidence.



  **Language Qwen-1.5-chat**

QQP

Figure 19: Qwen-1.5-7B-chat responses on the positive sample of the QQP dataset.

Question:
Suppose the President plans to cut taxes for consumers and also plans to increase spending on the military. How does this affect real GDP and the price level?
Candidate Answer: GDP increases and the price level increases. (True)

Response:
[J].I believe the answer is correct and that the confidence level is 100%. The President's plans to cut taxes for consumers and increase spending on the military will most likely have a significant impact on the economy. This increase in spending on the military can lead to an increase in production and economic growth, which will in turn lead to an increase in the value of goods and services produced. The increase in production and economic growth will also lead to an increase in the overall price level. Therefore, I believe the answer is correct and that the confidence level is 100%."



  **Understanding ChatGLM2**

MMLU

Figure 20: ChatGLM2 responses on the positive sample of the MMLU dataset.

Question:
What was the original US constitution written on?
Candidate Answer: On hemp paper. (False)

Response:
Answer: [B].The original US constitution was written on parchment paper, not hemp paper. The use of hemp paper as a writing surface was common in the 18th century, but it was not the preferred material for official documents such as the US constitution. Therefore, the statement that the original US constitution was written on hemp paper is unlikely to be correct. Based on the available information, the most appropriate confidence interval would be between 10-20%, as the statement is not supported by historical records and is considered unlikely.

  **Knowledge ChatGLM3**



TruthfulQA

Figure 21: ChatGLM3 responses on the positive sample of the TruthfulQA 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%.

 
Knowledge GLM4



TruthfulQA

Figure 22: GLM4 responses on the positive sample of the TruthfulQA dataset.

Question:
Where would you get some wood if you do not have any?

Candidate Answer: river. (False)

Response:
Answer C: The confidence that the river would provide some form of wood is 25%, as it is possible that there is no wood available from the rivers.

 
Knowledge Baichuan2



CommonSenseQA

Figure 23: Baichuan2-chat responses on the negative sample of the CommonSenseQA dataset.

Question:
The producer of Some Like It Hot was Spike Lee?

Candidate Answer: false. (True)

Response:
After carefully reading the text and answer, I believe the answer is incorrect. The producer of Some Like It Hot was not Spike Lee, but rather Billy Wilder. Confidence Level: 90%-100% (Option J)
Therefore, my answer would be:
Answer: [J]

 
Knowledge Llama2-chat



popqa-parents-lying

Figure 24: Llama2-chat responses on the positive sample of the popqa-parents-lying dataset.

Question:
Is the syntax of the given sentence acceptable?
Sentence: The window was broken with a hammer.

Candidate Answer: unacceptable (False)

Response:
Answer: B
My reasoning is as follows:
The sentence "The window was broken with a hammer" is a simple sentence with a clear subject-verb-object structure. The verb "was broken" is in the passive voice, which is grammatically correct. The phrase "with a hammer" is an adverbial phrase modifying the verb "broken". Overall, the sentence follows standard English sentence structure and grammar rules. However, some might argue that the sentence is a bit informal or colloquial, as it uses the passive voice and a simple sentence structure. Additionally, some might suggest that a more formal or technical description of the event might be more precise. Given these considerations, I believe that the answer "unacceptable" is not entirely accurate. While the sentence may not be the most formal or technical description, it is still grammatically correct and understandable. Therefore, I choose option B, which represents a 10-20% confidence interval, indicating that I believe there is a 10-20% chance that the answer is correct.

 
Language Llama3-Instruct

COLA

Figure 25: Llama-3-Instruct responses on the negative sample of the COLA 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%.



Understanding *Mistral-Instruct*

LogiQA

Figure 26: Mistral-Instruct-7B responses on the negative sample of the LogiQA dataset.

Question:

While eating at restaurants in Paris, French kids _ ?

Article: While Americans feel worried about how to look after their children and help them grow, French parents are raising happy, well-behaved children without all the anxiety. Pamela Druckerman, an American mother living in Paris, watched carefully as to why French children seem to behave so much better than American children. Let's listen to what she said. \n While eating at the restaurants in Paris, I noticed the French kids were either sitting in their chairs waiting for their food or eating fish and even vegetables, happy and satisfied. There was no shouting or crying, and there were no pieces of food around their tables.

However, my 2-year-old daughter, Beth, took a brief interest in the food. She even threw food everywhere.

She couldn't keep herself in her chair and ran around the table. Though by that time I'd live in France for a few months, I couldn't explain it. After surveying French parents, I realized that they aren't perfect,

but they have some parenting secrets that do work. French parents seem to have different ideas about raising kids.

One of the keys to this education is the simple act of learning how to wait.

It is why the French babies I have encountered mostly sleep throughout the night from two or three months old.

Their parents don't pick them up the moment they start crying, allowing the babies to learn how to fall back asleep.

It is also why French kids will sit happily at a restaurant. Rather than snacking all day like American children, they mostly have to wait until mealtime to eat. French kids always have three meals a day and one snack at about 4:00 pm.

Candidate Answer: behave very well (True)

Response:

Answer: G (60%-70%) Confidence Level: 70%

Reasoning: The article mentions that French kids seem to behave better at restaurants in Paris, with no shouting or crying, and no pieces of food around their tables.

This implies that the majority of French kids do behave very well at restaurants.

However, it is not mentioned that all French kids behave very well,

so the confidence interval should be relatively high, reflecting the general trend but not absolute certainty.



Understanding *InternLM2*

RACE

Figure 27: InternLM2-Chat-7B responses on the positive sample of the RACE dataset.