MLINGCONF: A Comprehensive Study of Multilingual Confidence Estimation on Large Language Models

Anonymous ACL submission

Abstract

The tendency of Large Language Models (LLMs) to generate hallucinations raises concerns regarding their reliability. Therefore, confidence estimations indicating the extent of trustworthiness of the generations become essential. However, current LLM confidence estimations in languages other than English remain underexplored. This paper addresses this gap by introducing a comprehensive investigation of **M**ulti**ling**ual **Conf**idence estimation (MLINGCONF) on LLMs, focusing on both language-agnostic (LA) and language-specific (LS) tasks to explore the performance and language dominance effects of multilingual confidence estimations on different tasks. The benchmark comprises four meticulously checked and human-evaluated high-quality multilingual datasets for LA tasks and one for the LS task tailored to specific social, cultural, and geographical contexts of a language. Our experiments reveal that on LA tasks *English* exhibits notable linguistic dominance in confidence estimations than other languages, while on LS tasks, using question-related language to prompt LLMs demonstrates better linguistic dominance in multilingual confidence estimations. The phenomena inspire a simple yet effective native-tone prompting strategy by employing language-specific prompts for LS tasks, effectively improving LLMs' reliability and accuracy in LS scenarios.

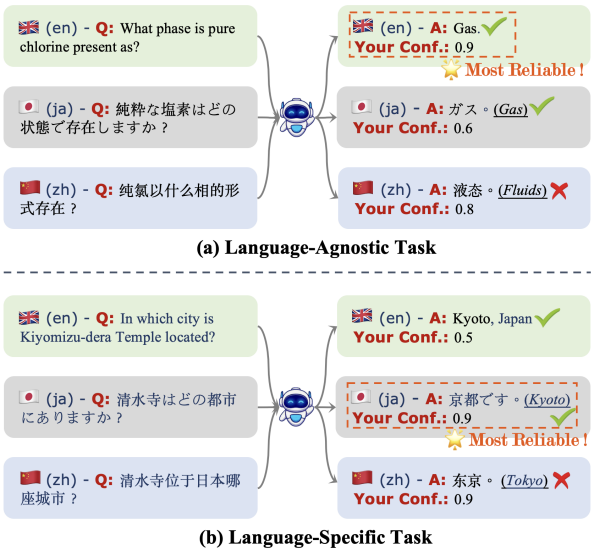


Figure 1: Examples of generations and confidence scores of Llama-3.1 given the same inputs in three languages in LA and LS scenarios derived from SciQ and LSQA datasets respectively.

1 Introduction

Large Language Models' (LLMs) susceptibility to generating hallucinated contents incurs concerns about unreliability in real-world applications (Ji et al., 2023; Rawte et al., 2023). Therefore, it becomes increasingly crucial for users to directly ascertain how much they can trust a model's response. Assessing the confidence or uncertainty of a model's output can immediately indicate the level of reliability to users, thereby playing a key role in developing trustworthy AI systems (Geng et al., 2023; Kadavath et al., 2022).

However, existing research on confidence or uncertainty estimations for LLMs has been predominantly limited to English (Kadavath et al., 2022; Lin et al., 2022; Geng et al., 2023; Tian et al., 2023b). The dearth of confidence estimations in languages other than English hinders users from assessing the reliability of LLMs in non-English scenarios, restricting the LLMs' global deployment. Due to the variations in the quantity and domain coverage of pre-training corpora across different languages, the confidence estimation ability may also presumably vary. Therefore, the performance of confidence estimation methods primarily developed for English remains a crucial subject for explorations when applied to other languages.

Additionally, to finely-grained investigate multilingual confidence estimations over various fine-grained tasks, we divide the tasks into language-

agnostic (LA) (Zhao et al., 2020) and language-specific (LS) scenarios as in Figure 1 considering the effects of linguistic dominance. Linguistic dominance refers to that one language holds a superior status over others within a specific social or cultural context (Blommaert, 2010; Treffers-Daller, 2019; Heller, 2007), can also exist in confidence estimation ability on different languages. In this study, the LS refers to the tasks that hold linguistic dominance caused by the knowledge domain coverage varying in different language training corpora, such as questions pertaining to social, cultural, or geographical contexts for a specific language, while the LA involves linguistic dominance mainly caused by the quantities of training corpora, such as general knowledge, common sense, and reasoning (Basaj et al., 2018; Sánchez et al., 2024).

To this end, we propose a benchmark called MLINGCONF (**Multi**l**ing**ual **Conf**idence) to investigate the performance of several LLM confidence estimation methods on five languages including *English*, *Japanese*, *Chinese*, *French*, and *Thai*. First, we meticulously constructed a high-quality multilingual dataset called the MlingConf dataset for the benchmark including five datasets of different tasks in LA and LS scenarios respectively. The LA involves four different tasks that are widely used in confidence estimation in *English* (Kuhn et al., 2023; Xiong et al., 2024) are translated into other four languages. We also create a language-specific QA (LSQA) dataset for the LS scenario, including five subsets for the investigated five languages respectively. Each subset comprises QA pairs about social culture, history, geography, and celebrity pertaining to the specific language. To ensure the data quality, we conduct rigorous translation consistency checks to filter the failed samples and finally employ linguistic experts for human evaluations.

Experiments are conducted on three major LLM confidence estimation methods including probability-based (Vazhentsev et al., 2023; Varshney et al., 2023) and prompt-based confidence estimations ($p(\text{True})$ (Kadavath et al., 2022) and self-verbalize (Lin et al., 2022; Xiong et al., 2024)) using the curated five multilingual datasets on several LLMs. We evaluate the confidence estimation ability and calibration using AUROC and ECE. Results on LA tasks suggest that prompt-based confidence estimations are preferable on LLMs with stronger instruction-following abilities, and English exhibits linguistic dominance. Results on the LS task reveal a pronounced phenomenon of language dominance, indicating that, for questions related to specific linguistic contexts, utilizing the respective languages yields the highest accuracy and confidence estimation performance. This observation inspires a native-tone prompting strategy: whereby, in the LS task, the relevant linguistic background of the question is first assessed, and then the corresponding language is employed to generate the response. Compared to the use of any single language, this approach leads to significant improvements in both accuracy and confidence estimations. Furthermore, we employ and generalize on extended confidence estimation methods and languages for both LS and LA tasks. The results further complete and enhance the above findings and analysis to the benchmark.

The contributions are summarized as follows:

• To the best of our knowledge, the MLINGCONF first propose to investigate multilingual confidence estimations with intricately constructed and expert-checked MlingConf datasets for both LA and LS scenarios, serving as a valuable benchmark for future works of reliable multilingual LLMs [1].

• Experiments conducted on MlingConf datasets present valuable findings about confidence estimation uses in multilingual scenarios, language dominance effects of *English* on LA tasks, and query-related languages on LS tasks respectively.

• Based on the observed linguistic dominance on LS tasks, we propose a native-tone prompting strategy, which significantly enhances the reliability and accuracy compared to the use of any single language prompts for LS tasks.

2 MlingConf Dataset

Owing to the lack of multilingual resources to comprehensively exhibit confidence estimation across diverse languages, we construct a high-quality multilingual dataset called MlingConf dataset encompassing five languages: *English* (**en**), *Japanese* (**ja**), *Chinese* (**zh**), *French* (**fr**), and *Thai* (**th**). Specifications of the language selection in consideration of language family and resource level are demonstrated in Appendix A. The MlingConf dataset includes four tasks for the language-agnostic (LA) scenario and one task for the language-specific (LS) scenario. We specify the data source and construction process of the MlingConf dataset in Sec. 2.1 and 2.2 respectively. Further dataset details and statistics can be referred to Appendix B.

[1]The codes and the MlingConf datasets will be released.

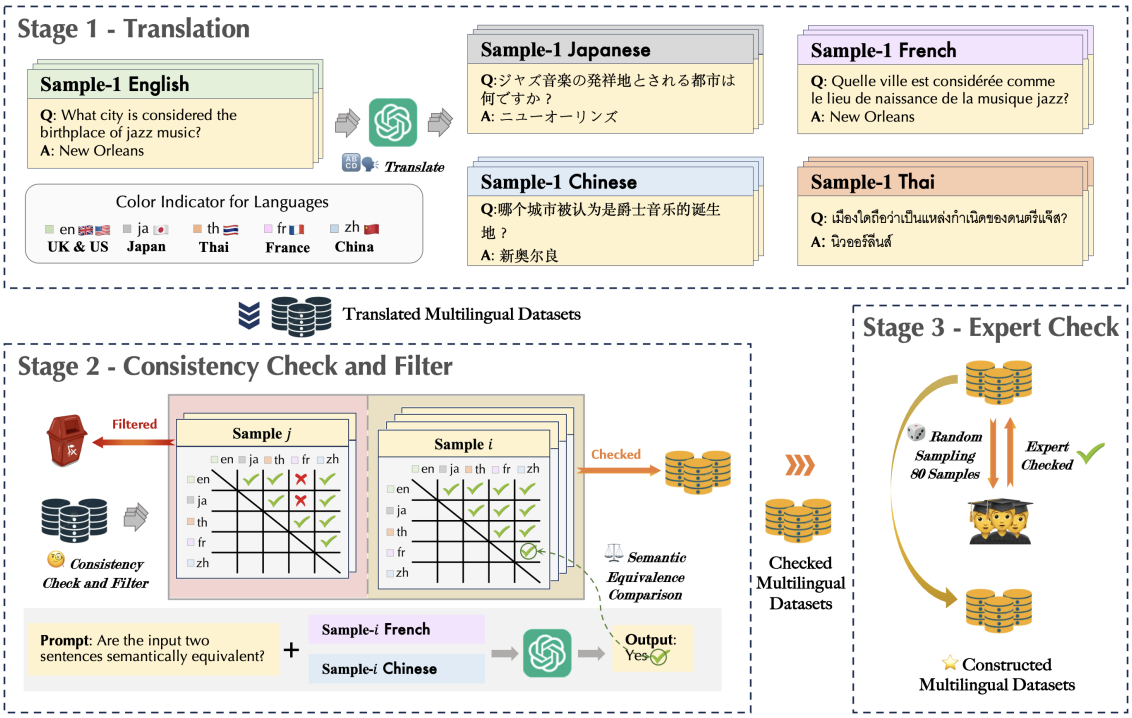


Figure 2: Three stages of MlingConf dataset construction.

2.1 Data Source

Language-Agnostic (LA) Tasks For LA tasks, we employ the following four datasets that are widely used for confidence estimations in *English* (Kuhn et al., 2023; Xiong et al., 2024). The datasets include 1) **TriviaQA** (**TVQA**) (Joshi et al., 2017) of closed-book trivia question-answering pairs to gauge models' factual knowledge; 2) **GSM8K** (Cobbe et al., 2021) for arithmetic reasoning task of math problems; 3) **CommonsenseQA** (**CSQA**) (Talmor et al., 2019) of multiple-choice QA pairs requiring different types of commonsense knowledge; 4) **SciQ** (Johannes Welbl, 2017) requiring scientific professional knowledge. All the datasets are pre-processed in standard QA format.

Language-Specific (LS) Tasks We create **Language-Specific QA** (**LSQA**) dataset pertaining to language-dominant knowledge covering specific social, geographical, and cultural language contexts for the UK & US, France, China, Japan, and Thailand respectively. We prompt *GPT-4* (OpenAI, 2023) [2] to generate 200 questions pertaining to only one specific language contexts as a language-specific subset. As demonstrated in Figure 11, for example, all questions in *Japanese* subset pertain to Japanese social culture, history, geography, celebrities, etc.

[2] https://platform.openai.com/docs/api-reference

2.2 Dataset Construction

The construction of the MlingConf dataset in this study follows an elaborate three-stage procedure as delineated in Figure 2.

Stage 1 : The QA samples derived from the above datasets are translated into four languages (**ja**, **th**, **zh**, and **fr**) through *GPT-4* (OpenAI, 2023).

Stage 2 : We check the consistency of five translated results by comparing the semantic equivalence in pairs in $C_5^2 = 10$ times for each sample. The samples with more than 2 times semantic inequivalence are treated as noisy data and then filtered. The changes of samples before and after consistency check and filter are in Table 5 and more clean multilingual datasets are obtained. Moreover, we present the number of samples for each language-specific LSQA subset as in Table 6.

Stage 3 : Finally, we employ several experts majoring in linguistics to examine the translation performance across 50 randomly selected samples as shown in Table 1. Human evaluation results suggest the obtained multilingual datasets are high-quality for further experiments.

For all generations of MlingConf dataset construction, the temperature T is set to 0. The translation and semantic equivalence comparison prompts are presented in Appendix C.

Lang.	TVQA	GSM8K	CSQA	SciQ	LSQA
zh	96%	100%	100%	98%	100%
ja	98%	100%	98%	96%	100%
fr	100%	100%	100%	98%	100%
th	96%	100%	98%	94%	100%

Table 1: Translation accuracy evaluated by linguistic experts on 50 randomly selected samples.

3 Experimental Settings

3.1 Confidence Estimation Methods

In this part, we investigate three confidence estimation methods primarily used in *English* for LLMs as in Figure 14. These methods will be also conducted in our multilingual settings. Specifically, we denote $\mathsf{Conf}(\boldsymbol{x}, \boldsymbol{y})$ as the confidence score associated with the output sequence $\boldsymbol{y} = [y_1, y_2, \ldots, y_N]$ given the input context $\boldsymbol{x} = [x_1, x_2, \ldots, x_M]$.

Probability-based Confidence (Prob.): The probability-based confidence is estimated by calculating the joint token-level probabilities over $\boldsymbol{y}$ conditioned on $\boldsymbol{x}$. As longer sequences are supposed to have lower joint likelihood probabilities that shrink exponentially with length, we calculate the geometric mean by normalizing the output token probabilities which are represented as:

$$\mathsf{Conf}(\boldsymbol{x}, \boldsymbol{y}) = \left(\prod_{i}^{N} p(y_i | \boldsymbol{y}_{<i}, \boldsymbol{x}) \right)^{\frac{1}{N}} \quad (1)$$

p(True)-based Confidence (p(True)): The $p(\text{True})$ confidence score is implemented by simply asking the model itself if its first proposed answer $\boldsymbol{y}$ to the question $\boldsymbol{x}$ is true (Kadavath et al., 2022), and then obtaining the probability $p(\text{True})$, which can implicitly reflect self-reflected certainty.

Self-verbalized Confidence (Verb.): As LLMs possess good self-reflection and instruction-following abilities, recent works pay particular attention to linguistic confidence via prompting LLMs to express certainty in verbalized numbers or words (Lin et al., 2022; Xiong et al., 2024). We adopt verbalized numerical probability in token-level space as LLM's confidence estimation.

The multilingual prompts for p(True) and Verb. are in Appendix C.

3.2 Evaluation Metrics

Accuracy (Acc.) We employ a string-matching approach to evaluate the accuracy of generated answers $\boldsymbol{y}$ and compare them with the ground truth $\hat{\boldsymbol{y}}$. Although exact matching (EM) of $\boldsymbol{y} \equiv \hat{\boldsymbol{y}}$ is widely used on GSM8K and CSQA, it always misjudges some correct answers with slight differences on closed-book QA tasks, to better assess the result accuracy (Accu.), we replace EM with a variant called positive-recall exact matching (PREM) of $\boldsymbol{y} \in \hat{\boldsymbol{y}} \vee \hat{\boldsymbol{y}} \in \boldsymbol{y}$. Comparisons of several EM variants we tested as well as human evaluations are presented in Appendix D.

Area Under the Receiver Operator Characteristic Curve (AUROC) AUROC assesses the effectiveness of confidence estimation (Filos et al., 2019) by quantifying how likely a randomly chosen correct answer possesses a higher confidence score than an incorrect one, yielding a score in range of [0, 1], implemented by `sklearn` toolkit [3].

Expected Calibration Error (ECE) ECE gauges the calibration performance which indicates how well a model's predicted confidence matches its actual accuracy (Guo et al., 2017a). For an expected well-calibrated AI system, samples $\boldsymbol{x}$ with confidence of q should also have an average accuracy of q on predictions $\boldsymbol{y}$ where $P(\boldsymbol{y} = \hat{\boldsymbol{y}} | \mathsf{Conf}(\boldsymbol{x}, \boldsymbol{y}) = q) = q$ with ECE=0. ECE is essential for reliable AI systems on prediction tasks like weather forecasting. The smaller the ECE value, the better. Details of the ECE calculation are presented in Appendix D.

3.3 Implementation Details

Experiments are conducted on GPT-3.5-Turbo-0125 (GPT-3.5) and Llama-3.1-8B-Instruct (Llama-3.1) [4] (AI@Meta, 2024). We only present the results of the current most commonly used commercial GPT-3.5 and open-source Llama-3.1 in the main part and leave the results on some other LLMs in Appendix E. Few-shot prompts containing N_f examples are utilized for answer generation with greedy decoding which outperforms temperature decoding on knowledge tasks (Song et al., 2024). N_f is set to 8 for GSM8K and 4 for others.

4 Experiments on LA Tasks

To comprehensively investigate LLMs' multilingual confidence estimations on LA tasks, as presented in Table 2 and Figure 3, experiments are conducted to observe performances varying in different

[3] https://github.com/scikit-learn/scikit-learn/blob/main/sklearn/metrics/_ranking.py

[4] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

Conf.	en		zh		ja		fr		th		*Avg.*	
	ARC. ↑	ECE ↓	ARC. ↑	ECE ↓	ARC. ↑	ECE ↓	ARC. ↑	ECE ↓	ARC. ↑	ECE ↓	ARC. ↑	ECE ↓
TVQA on *GPT-3.5*												
Prob.	76.51	24.36	78.39	32.95	76.90	**28.14**	72.14	27.39	74.30	40.17	75.65	30.60
p(True)	79.64	18.25	**82.34**	**22.94**	84.50	29.06	80.59	**20.90**	81.22	40.87	81.66	**26.40**
Verb.	**80.32**	**16.52**	81.76	24.32	**84.61**	34.49	**83.47**	26.53	**86.72**	**38.19**	**83.38**	28.01
TVQA on *Llama-3.1*												
Prob.	**80.74**	**10.72**	**80.41**	40.39	**88.75**	**26.24**	**79.05**	**20.38**	**89.59**	**36.77**	**83.71**	**26.90**
p(True)	68.98	18.35	68.10	38.19	52.69	37.85	62.00	22.04	60.55	37.01	62.66	30.69
Verb.	77.18	24.73	63.50	**37.64**	68.91	34.27	69.90	25.44	73.22	40.19	70.54	32.45
GSM8K on *GPT-3.5*												
Prob.	54.79	26.48	58.49	**27.19**	57.09	29.46	57.38	37.29	**61.73**	**41.77**	57.90	32.44
p(True)	**65.25**	31.88	**62.74**	28.65	**69.75**	**19.03**	60.14	39.08	61.45	49.88	**63.87**	33.70
Verb.	62.34	**22.17**	59.25	28.91	58.34	26.71	**66.65**	**25.14**	54.02	45.63	60.12	**29.71**
GSM8K on *LLama-3.1*												
Prob.	**65.69**	22.33	**66.37**	21.92	69.73	35.69	**61.07**	29.56	**63.22**	28.51	**65.22**	27.60
p(True)	61.64	**14.49**	65.83	**17.39**	**71.40**	**11.26**	57.04	**8.02**	57.31	**12.90**	62.64	**12.81**
Verb.	57.00	50.05	63.04	42.89	58.93	45.33	54.45	35.31	55.30	34.94	57.75	41.70
CSQA on *GPT-3.5*												
Prob.	59.06	24.45	55.92	38.30	48.01	50.60	55.33	31.12	48.21	41.71	53.31	48.18
p(True)	67.13	19.65	**58.64**	27.08	65.23	**19.24**	**66.33**	23.43	59.96	34.47	63.46	24.77
Verb.	**69.60**	**16.84**	54.30	**21.54**	**68.34**	19.84	61.87	**21.81**	**68.93**	**21.71**	**64.73**	**24.35**
CSQA on *Llama-3.1*												
Prob.	**78.06**	**13.64**	**64.91**	36.33	**75.65**	**30.11**	**66.12**	**19.72**	**77.65**	42.18	**72.47**	**28.40**
p(True)	56.25	34.04	64.24	36.82	66.60	40.32	59.34	27.72	58.07	**29.91**	60.90	33.76
Verb.	62.42	28.16	54.61	**28.84**	61.06	37.85	57.97	23.96	71.91	37.15	61.39	31.19
SciQ on *GPT-3.5*												
Prob.	69.50	32.23	71.29	35.63	78.28	47.06	**72.66**	34.85	**75.13**	56.17	73.37	41.19
p(True)	**72.06**	23.15	**76.18**	**30.44**	**80.16**	36.18	71.30	37.85	68.29	41.25	**73.60**	33.77
Verb.	70.18	**20.80**	75.50	37.59	77.89	**30.33**	69.31	**32.47**	74.85	**41.15**	73.55	**32.47**
SciQ on *Llama-3.1*												
Prob.	**74.14**	**13.40**	**72.09**	32.26	**74.48**	**34.21**	**77.45**	22.76	**77.61**	**36.10**	**75.15**	**27.75**
p(True)	62.38	19.28	64.89	**37.01**	58.92	36.47	61.01	**10.72**	51.90	41.06	59.82	28.91
Verb.	62.65	24.10	52.90	32.94	69.10	39.15	59.30	24.92	65.93	40.67	61.98	38.36
Avg. (TVQA, GSM8K, CSQA, SciQ) on *GPT-3.5*												
Prob.	64.96	26.88	66.02	33.51	65.07	38.71	64.37	32.66	64.88	44.95	65.06	35.34
p(True)	70.38	23.23	**69.97**	**27.27**	**74.91**	**25.87**	69.59	30.31	67.73	41.61	**70.65**	29.66
Verb.	**70.61**	**19.01**	67.70	28.09	72.29	27.84	**70.32**	**26.48**	**71.13**	**36.67**	70.45	**28.64**
Overall	68.68	23.04	67.90	29.65	70.75	30.84	68.10	29.82	67.90	41.08	68.66	30.89
Avg. (TVQA, GSM8K, CSQA, SciQ) on *Llama-3.1*												
Prob.	**74.88**	**15.02**	**70.94**	32.72	**77.15**	31.56	**70.92**	23.10	**77.01**	35.89	**74.14**	27.66
p(True)	62.31	21.54	65.76	**32.21**	62.40	**31.47**	59.59	**17.12**	56.95	**30.22**	61.51	**26.54**
Verb.	64.81	31.76	58.51	35.57	64.50	39.15	60.40	27.40	66.59	38.23	62.92	35.93
Overall	67.26	22.77	65.10	33.55	67.97	34.06	63.72	22.58	66.85	34.78	66.19	29.38

Table 2: Experimental results of AUROC (ARC.) and ECE of three confidence estimation methods on four LA datasets on GPT-3.5 and Llama-3.1.

confidence estimation methods and languages in Sec. 4.1 and 4.2 respectively.

4.1 Results regarding Different Confidence Estimations on LA Tasks

Findings: Applying prompt-based confidence estimations is preferable in multilingual tasks for LLM with stronger instruction-following ability. The probability method performs better confidence estimations on relatively weak LLM. The findings provide a direct takeaway about selecting optimal confidence estimations of LLMs in multilingual scenarios. In Table 2, we highlight the supreme performance in bold among the three methods for each column of each dataset. On GPT-3.5, both p(True) and Verb. yield the superior performances than Prob. across all languages averaged on four datasets (ARC.: 70.65, 70.45 vs. 66.02; ECE: 29.66, 28.64 vs. 35.34). p(True) and Verb. have comparable performance in ARC. scores, while Verb. is better calibrated. In contrast, Prob. shows superior performance than p(True) and Verb. and performs more stable on Llama-3.1 (ARC.: 74.14 vs. 61.51 and 62.92; ECE: 27.66 vs. 26.54 and 35.93). p(True) demonstrates better calibration results on languages other than *English*.

Analysis: In Table 2, the performance differences between two LLMs can be attributed to that GPT-3.5's strong instruction-following abilities benefit the prompt-based multilingual confidence estimation methods Verb. and p(True), but leading to over-confidence in output token probabilities. In contrast, Llama-3.1 cannot stably generate verbalized confidence scores and perform relatively weak instruction-following abilities, but maintain well-calibrated likelihood probabilities during the pre-training stage for all languages.

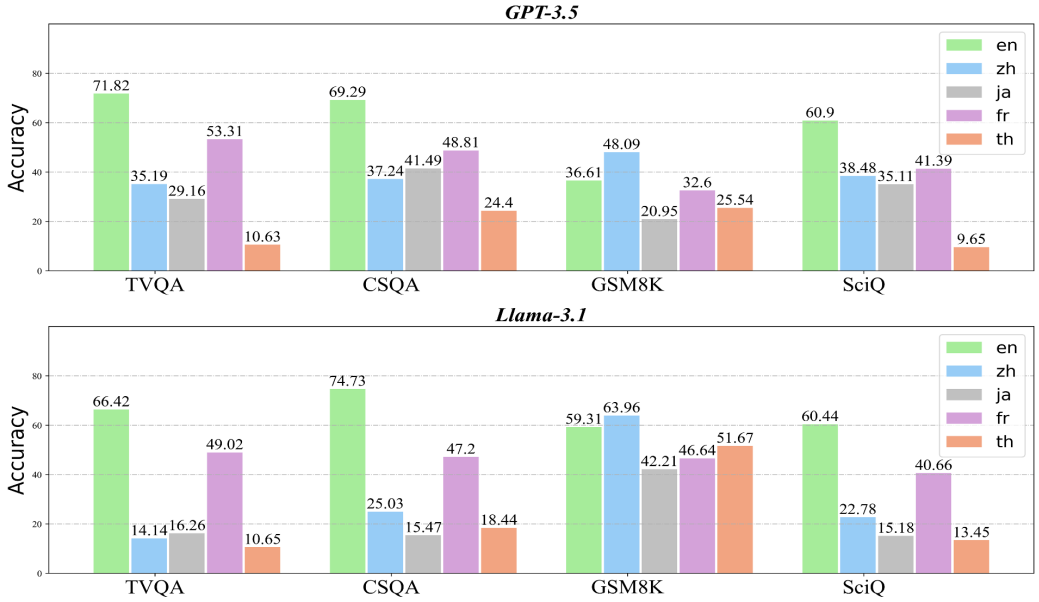


Figure 3: Experimental results of Accuracy on four LA datasets on GPT-3.5 and Llama-3.1.

4.2 Results regarding Different Languages on LA Tasks

Findings: Linguistic dominance is manifested in *English* with superior confidence estimation performances on LA tasks for multilingual LLMs. Prior works only validate the efficacy of prompt-based confidence estimations in *English*. Our findings indicate that the methods are also preferable in other languages and performances fluctuate in different languages. In Table 2, ARC. scores are less fluctuating across different languages while ECE in *English* (23.04 and 22.77) performs better than in other languages on both GPT-3.5 and Llama-3.1. We also report the accuracy on LA datasets in Figure 3. *English* consistently performs better across all datasets exclusively GSM8K. Generally, prompting in *English* outperforms others, hence responding in *English* on LA tasks can be adequately credible and accurate where linguistic dominance is leading in *English*.

Analysis: Despite the powerful multilingual capacity of LLMs, discrepancies exist in the quantity of distinct linguistic training corpora available for each language. Results in Table 2 suggest that ARC. is a metric not significantly related to language usage in LLMs, while the strong performance of ECE in *English* can be attributed to the extensive training corpus or calibrations conducted during training in *English*. As the only middle-resource language, *Thai* exhibits a notably lower level of reliability compared to the other high-resource languages.

Considering consistency check in Table 5, the lowest filter rate in GSM8K translation indicates that mathematical reasoning tasks are minimally affected by language bias. As a result, accuracy fluctuations across different languages on GSM8K are relatively small. For that *Chinese* exhibits slightly superior mathematical capabilities compared to English on GSM8K on both GPT-3.5 and Llama-3.1 (Accu. 48.09 and 63.96), it is hypothesized that pre-training corpora contain a substantial amount of Chinese mathematical problems.

5 Experiments on LS Task

For the LS task, we present the confidence estimation results on five language-specific subsets of LSQA in Figure 4 and 5 in Sec. 5.1. Based on the findings, we then propose a **Native-Tone Prompting (NTP)** strategy to better leverage linguistic dominance to improve the LLMs' reliability and accuracy on LS task in Sec. 5.2.

5.1 Results of Different Language-specific Subsets on LS Task

Findings: Applying prompts in query-related language demonstrates linguistic dominance with better multilingual confidence estimations on LS task. In Figure 4, the diagonal values of the ECE and ARC. heatmaps of Prob. are more pronounced, indicating that when using Prob. confidence, linguistic dominance is more apparent compared to p(True) and Verb.. Consequently, we have opted Prob. for subsequent experiments in Sec. 5.2.

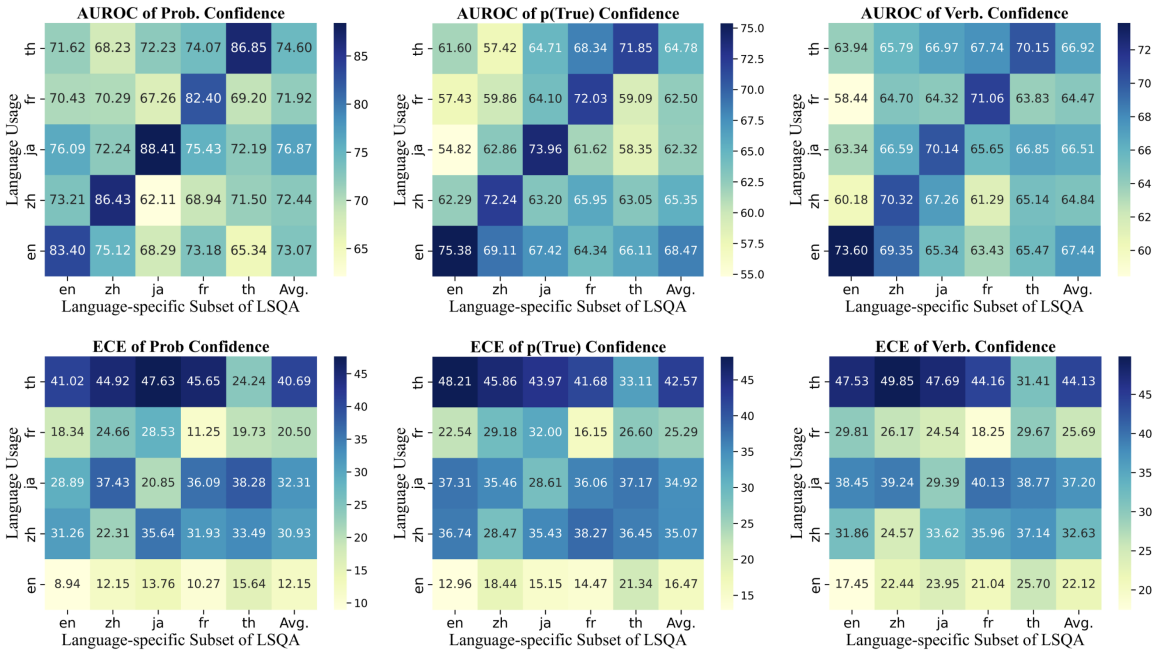


Figure 4: Experimental results of AUROC and ECE of three confidence estimation methods on five language-specific subset of LSQA using Llama-3.1.

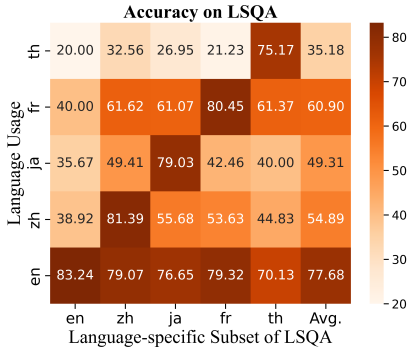


Figure 5: Experimental results of Accuracy on five language-specific LSQA subsets using Llama-3.1.

Additionally, in Figure 5, although prompting in *English* performs well and stable across different subsets, there is a noticeable improvement in accuracy when prompts are related to each subset's language. In comparison with the LA tasks in Sec. 4.1 where linguistic dominance is primarily manifested in *English*, on LSQA, linguistic dominance is determined by specific language of the subset.

Analysis: The linguistic dominance on LSQA can be conjectured to stem from the fact that such data pertaining to the language-specific cultural, geographical, or social contexts are already included in the pre-training corpora of their respective languages with higher certainty or confidence, thereby achieving optimal performance when prompting in their respective specific languages.

5.2 Results of Native-Tone Prompting (NTP) Strategy on LS Task

Confidence estimation performance differences caused by linguistic dominance phenomena on the LS task motivate us to explore the improving method. Inspired by results in Sec. 5.1 on each language-specific LSQA subset, we propose a simple yet effective Native-Tone Prompting (NTP) strategy to achieve better confidence estimation performance on the LS task. NTP first prompts LLMs to identify the language context of the question, and then uses that query-related language to answer the question, effectively exhibiting a "native tone" that is more familiar in that language-related context. We present the results of prompting by any single language versus NTP on LSQA in Table 3. The prompt of NTP is presented in the Appendix C.

Prompt	en	zh	ja	fr	th	NTP
Acc. ↑	77.68	60.16	44.64	60.90	35.18	**79.46**
ARC. ↑	73.07	72.44	76.87	71.92	74.60	**77.25**
ECE ↓	12.15	30.93	32.31	20.50	40.69	**10.28**

Table 3: Experimental results of overall Accu., ARC., and ECE on the LSQA dataset by prompting using different languages and our proposed NTP method.

Findings: Prompting LLMs using the query-related language can enhance the reliability of confidence estimations and accuracy on LS tasks, which provides an insight to improve LLMs'

reliability regarding the prompt language usage. In Table 3, the experiments demonstrate that NTP better leverages the inherent linguistic dominance, thereby yielding more reliable and accurate results than any single language prompt, validating the effectiveness of NTP on the LS task.

Analysis: Results in Table 3 indicate that the multilingual capabilities and reliability of LLMs are still constrained by the imbalanced training corpus among diverse languages. The reliability and accuracy on *English*, serving as the primary training corpora, have not been adequately generalized to other languages. Even for semantically equivalent queries in different languages, the reliability of responses cannot be consistently maintained.

6 Discussion

Extended Confidence Estimations In Appendix E.2, we further investigate three other confidence estimation methods including 1) paraphrasing the questions; 2) sampling multiple responses (Xiong et al., 2024); and 3) introducing Chain-of-Thought (CoT) (Wei et al., 2022) on both LS and LA tasks. As presented in Table 10 and Figure 13, **all findings on extended three confidence estimations are consistent with previous analysis across all languages**. The questions after paraphrasing still maintain semantic equivalence without obvious perturbations for all languages, and LLMs are robust in multilingual confidence estimations to different questions with similar meanings. p(True) and Verb. methods outperform sampling-based methods as the high temperature may incur variability in output spaces which undermines the reliability of QA tasks for all languages. LLMs' CoT ability can be generalized to multilingual domains, thus benefiting multilingual confidence estimations.

Extended Languages In Appendix E.3, we also extend the investigations on other five languages derived and translated from TriviaQA into *Korean*, *Arabic*, *German*, *Indonesian*, and *Italian* as in Sec. 2. As in Table 11, linguistic dominance is still performed in *English* than other languages on the LA task. Low-resource languages demonstrate poor performance in ECE. For the LS task, we also develop a small-size LSQA subset for the above five languages in Table 8 to conduct the NTP method. Experiments suggest that **NTP can also generalize and improve the reliability and accuracy in such middle- or low-resource languages**.

7 Related Works

Confidence Estimation for LLMs Previous confidence estimation methods can be categorized into five classes, as illustrated in Figure 14 and Appendix F. ① **Probability-based**: Vazhentsev et al. (2023) intermediately quantifies sentence uncertainty over token probabilities; ② $p(\text{True})$**-based**: Kadavath et al. (2022) instructs the LLM to self-evaluate the correctness of the generated answer by directly accessing $p(\text{True})$; Both ① and ② require access to token probabilities and thus are limited to **white-box LLMs**. ③ **Self-verbalized**: LLMs' remarkable instruction-following ability provides a view of expressing confidence in words (Lin et al., 2022; Zhou et al., 2023; Tian et al., 2023a; Xiong et al., 2024); ④ **Sampling-based**: By sampling multiple responses to a given question, Xiong et al. (2024) aggregates all the confidence scores as the indicator. ⑤ **Training-based**: Lin et al. (2022); Kadavath et al. (2022) propose to train an external module to improve confidence estimations.

Multlingual LLMs Most recent LLMs primarily pre-trained on English corpora have showcased remarkable capabilities (Pires et al., 2019; Shi et al., 2023; OpenAI, 2023). However, their efficacy in other low-resource languages remains limited. Many research works have extended various tasks in multilingual domains such as claim fact-checking (Pikuliak et al., 2023) and jailbreak problem (Deng et al., 2024). Prior studies have also explored diverse cross-lingual applications (Wang et al., 2023a,b; Qin et al., 2022).

8 Conclusion

This study underscores the necessity of advancing multilingual confidence estimation methods for LLMs to ensure their reliability across diverse linguistic contexts. The proposed MLINGCONF serves as a valuable and noteworthy benchmark to address the gap in multilingual confidence estimation research. Our findings demonstrate the variability of multilingual confidence estimations on both LA and LS scenarios, revealing the influence of linguistic dominance on different tasks. This leads to the NTP strategy, improving accuracy and reliability by aligning the response language with the linguistic context of the query for LS tasks. These insights and the introduction of MlingConf datasets pave the way for future research, enhancing the global applicability and reliability of LLMs.

Limitations

The limitations and prospects for future research are outlined as follows:

Expensive Costs to Obtain High-Quality Low-Resource Languages The present study is constrained by the substantial cost associated with the API cost using ***GPT-4*** for translation as well as linguistic verification. This multilingual research is restricted to five languages in the first version. This initial phase aims to delve into confidence estimation within multilingual domains. Our future endeavors will involve the expansion of the benchmark dataset, encompassing additional languages and data points to enrich our investigations.

Native-Tone Prompting is a Primary Version Although the proposed Native-Tone Prompting method can enhance the accuracy and reliability of LS tasks, it still relies on external prompts to determine which language domain the query pertains to. Moving forward, it is promising to broaden the scope of developing a cross-lingual method that can directly transfer the specific language dominance to other language contexts, thereby facilitating multilingual confidence estimation abilities for LLMs.

Ethics Statement

In this paper, we introduce several self-constructed multilingual datasets derived from the publicly available dataset. The selection of investigated languages in this work depends on whether we can employ appropriate linguistic experts. Most linguistic specialists are M.Phil. or Ph.D. students majoring in linguistics and others are from crowd-sourcing platforms. We meticulously adhered to legal and ethical standards throughout the data collection process, prioritizing privacy and obtaining informed consent. Linguistic experts were furnished with comprehensive details regarding the study's objectives, data collection methodologies, and associated risks or benefits. They were afforded the opportunity to seek clarifications and voluntarily provide consent before their involvement. All collected data were exclusively utilized for research purposes.

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297.

AI@Meta. 2024. Llama 3 model card. *AI@Meta*.

Dominika Basaj, Barbara Rychalska, and Anna Wroblewska. 2018. Las: Language agnostic system for question answering. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 260–263.

J Blommaert. 2010. *The sociolinguistics of globalization*. Cambridge University Press.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jiuhai Chen and Jonas Mueller. 2023. Quantifying uncertainty in answers from any language model via intrinsic and extrinsic confidence assessment. *arXiv preprint arXiv:2308.16175*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2024. Multilingual jailbreak challenges in large language models. In *The Twelfth International Conference on Learning Representations*.

Angelos Filos, Sebastian Farquhar, Aidan N Gomez, Tim GJ Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud de Kroon, and Yarin Gal. 2019. Benchmarking bayesian deep learning with diabetic retinopathy diagnosis. *Preprint at https://arxiv.org/abs/1912.10481*.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.

Yarin Gal et al. 2016. Uncertainty in deep learning. *Ph.D. Thesis*.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2023. A survey of language model confidence estimation and calibration. *Preprint*, arXiv:2311.08298.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017a. 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.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017b. On calibration of modern neural networks. *Preprint*, arXiv:1706.04599.

Haixia Han, Tingyun Li, Shisong Chen, Jie Shi, Chengyu Du, Yanghua Xiao, Jiaqing Liang, and Xin Lin. 2024. Enhancing confidence expression in large language models through learning from past experience. *Preprint*, arXiv:2404.10315.

Monica Heller. 2007. *Bilingualism: A social approach*. Springer.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Matt Gardner Johannes Welbl, Nelson F. Liu. 2017. Crowdsourcing multiple choice science questions. In *arXiv*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. 2024. Calibrating large language models with sample consistency. *Preprint*, arXiv:2402.13904.

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

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*, 10:857–872.

Kenton Murray and David Chiang. 2018. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 212–223, Brussels, Belgium. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Matúš Pikuliak, Ivan Srba, Robert Moro, Timo Hromadka, Timotej Smoleň, Martin Melišek, Ivan Vykopal, Jakub Simko, Juraj Podroužek, and Maria Bielikova. 2023. Multilingual previously fact-checked claim retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16477–16500, Singapore. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Tianbao Xie, Qixin Li, Jian-Guang Lou, Wanxiang Che, and Min-Yen Kan. 2022. GL-CLeF: A global–local contrastive learning framework for cross-lingual spoken language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2677–2686, Dublin, Ireland. Association for Computational Linguistics.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *Preprint*, arXiv:2309.05922.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Yifan Song, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. 2024. The good, the bad, and the greedy: Evaluation of llms should not ignore non-determinism. *Preprint*, arXiv:2407.10457.

Eduardo Sánchez, Belen Alastruey, Christophe Ropers, Pontus Stenetorp, Mikel Artetxe, and Marta R. Costa-jussà. 2024. Linguini: A benchmark for language-agnostic linguistic reasoning. *Preprint*, arXiv:2409.12126.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *Preprint*, arXiv:1811.00937.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023a. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023b. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Jeanine Treffers-Daller. 2019. What defines language dominance in bilinguals? *Annual Review of Linguistics*, 5(1):375–393.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.

Artem Vazhentsev, Akim Tsvigun, Roman Vashurin, Sergey Petrakov, Daniil Vasilev, Maxim Panov, Alexander Panchenko, and Artem Shelmanov. 2023. Efficient out-of-domain detection for sequence to sequence models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1430–1454, Toronto, Canada. Association for Computational Linguistics.

Hao Wang and Dit-Yan Yeung. 2020. A survey on bayesian deep learning. *ACM Comput. Surv.*, 53(5).

Hongru Wang, Boyang Xue, Baohang Zhou, Tianhua Zhang, Cunxiang Wang, Guanhua Chen, Huimin Wang, and Kam fai Wong. 2024. Self-dc: When to retrieve and when to generate? self divide-and-conquer for compositional unknown questions. *Preprint*, arXiv:2402.13514.

Jiaan Wang, Yunlong Liang, Fandong Meng, Beiqi Zou, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023a. Zero-shot cross-lingual summarization via large language models. *Preprint*, arXiv:2302.14229.

Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023b. Towards unifying multi-lingual and cross-lingual summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15127–15143, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In

Advances in Neural Information Processing Systems, volume 35, pages 24824–24837. Curran Associates, Inc.

Yuxin Xiao, Paul Pu Liang, Umang Bhatt, Willie Neiswanger, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. Uncertainty quantification with pre-trained language models: A large-scale empirical analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7273–7284, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*.

Boyang Xue, Shoukang Hu, Junhao Xu, Mengzhe Geng, Xunying Liu, and Helen Meng. 2022. Bayesian neural network language modeling for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2900–2917.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2020. Inducing language-agnostic multilingual representations. *arXiv preprint arXiv:2008.09112*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5506–5524, Singapore. Association for Computational Linguistics.

A Language Information

The basic information of ISO codes and the language family of the investigated languages is presented in Table 4. The investigated languages from widely spoken to lesser-known ones in this work are selected following three principles.

1) Following (Lai et al., 2023; Deng et al., 2024) which determines the resource levels for each language by utilizing the data ratio from the Common-Crawl corpus [5], we select three languages (*Chinese*, *Japanese*, and *French*) in high-resource category whose data ratio exceeds 1%, and one language (*Thai*) from medium-resource class that falls between 0.1% and 1%. To ensure the confidence estimation ability can be observed, the low-resource languages less than 0.1% are omitted and left for future works.

2) This selection ensures coverage of a wide range of linguistic characteristics from different language families as in Table 4. A language family represents a collective of cognate languages stemming from a common ancestral source, serving as a focal point within the domain of linguistics [6].

3) For each selected language, we can employ one linguistic expert for the human check to ensure the data quality;

	ISO 639-1	Family
English	**en**	Indo-European
French	**fr**	Indo-European
Chinese	**zh**	Sino-Tibetan
Japenese	**ja**	Japanese-Ryukyuan
Thai	**th**	Kra–Dai
Indonesian	**id**	Indo-European
German	**de**	Indo-European
Arabic	**ar**	Afro-Asiatic
Korean	**ko**	Koreanic
Italian	**it**	Indo-European

Table 4: List of International Standard Organization (ISO) 639-1 codes and language family information.

B Dataset Details

TriviaQA The TriviaQA dataset (Joshi et al., 2017) is a realistic text-based reading comprehension question-answering dataset containing 650K question-answer-evidence triples from 95K documents collected from Wikipedia and the websites, served as a benchmark for evaluating machine comprehension and question-answering systems, which is more challenging than standard QA benchmark datasets where the answer spans can be directly retrieved and copied.

GSM8K GSM8K (Grade School Math 8K) (Cobbe et al., 2021) is a dataset of 8.5K high quality linguistically diverse grade school math word problems. The dataset was created to support the task of question answering on basic mathematical problems that require multi-step reasoning to solve.

CommonsenseQA CommonsenseQA (Talmor et al., 2019) is a new multiple-choice question answering dataset that requires different types of commonsense knowledge to predict the correct answers. The dataset consists of 12,247 questions with 5 choices each.

[5] http://commoncrawl.org/

[6] https://en.wikipedia.org/wiki/Language_family

Translation Prompt

You are an excellent translator. Please translate the input texts into {language}.

Input ###: {input_sentence}
Output ###:

Figure 6: Translation prompt.

SciQ The SciQ dataset (Johannes Welbl, 2017) contains 13,679 crowdsourced science exam questions about Physics, Chemistry and Biology, among others. The questions are in multiple-choice format with 4 answer options each. For the majority of the questions, an additional paragraph with supporting evidence for the correct answer is provided.

LSQA We present two examples of the LSQA dataset in *English*- and *Japanese*- specific subsets in Figure 11.

	TVQA	GSM8K	CSQA	SciQ	LSQA
Before	2000	1319	1221	1000	1000
After	1238	1318	1152	640	857

Table 5: Number of samples before and after consistency check and filter.

LSQA	en	zh	ja	fr	th	Total
	185	172	167	179	154	857

Table 6: Statistics of samples for each language-specific subset of the LSQA dataset.

C Prompt Details

The translation prompt for multilingual dataset construction and semantic equivalence comparison prompt for consistency check in Sec, 2 are presented in 6 and 7 respectively. Standard multilingual Question-Answering prompts are in 8. Multilingual confidence estimations of ***P(True)*** and ***Verb.*** are presented in Fig. 9 and 10. Notably, the prompts for self-reflected true probability confidence estimation are followed by previous work (Kadavath et al., 2022; Kuhn et al., 2023).

D Metric Details

Expected Calibration Error (ECE) We partition the inference results into M disjoint bins $\{B_m\}_{m=1}^{M}$ based on the confidence scores $\{q_i\}$, compute the average confidence score in $\left(\frac{m-1}{M}, \frac{m}{M}\right]$ for the m-th bin B_m, and compare it with the average true accuracy $\mathsf{acc}(B_m)$ of the answers within B_m. The ECE is calculated by:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathsf{acc}(B_m) - \mathsf{conf}(B_m)\right| \quad (2)$$

The average accuracy $\mathsf{acc}(B_m)$ and confidence $\mathsf{conf}(B_m)$ of the answers in B_m is obtained by:

$$\mathsf{acc}(B_m) = \frac{1}{|B_m|} \sum_{a_i \in m} \mathbb{I}(\hat{b}_i = b_i) \quad (3)$$

$$\mathsf{conf}(B_m) = \frac{1}{|B_m|} \sum_{a_i \in m} q_i \quad (4)$$

where a_i, b_i, $\hat{b}_i$, and q_i indicate the input data, label, prediction result, and confidence score respectively for the i-th sample.

Accuracy For closed-book QA evaluation, we observe that simply applying EM may misjudge the correct answers. We compare several variants of EM as in Table 7 and report their successful judgments on responses of 20 selected samples that are misjudged using EM, where PEM, RRM, and PREM indicate Positive-EM, Recall-EM, and Positive-Recall-EM and the mathematical explanations are presented in Table 7. Upon human discrimination, EMPR exhibits the lowest failure rate and is therefore selected as the evaluation metric for this work.

Variant	Explanation	# Fail
EM	$\boldsymbol{y} \equiv \hat{\boldsymbol{y}}$	20
PEM	$\boldsymbol{y} \in \hat{\boldsymbol{y}}$	16
REM	$\hat{\boldsymbol{y}} \in \boldsymbol{y}.$	6
PREM	$\boldsymbol{y} \in \hat{\boldsymbol{y}} \vee \hat{\boldsymbol{y}} \in \boldsymbol{y}.$	2

Table 7: Number of failed judgments by human check for different EM variants.

Semantic Equivalence Comparison Prompt
You are an excellent natural language inference model. You MUST determine whether the provided two Sentences are semantically equivalent. The response you provided MUST be "Yes" or "No". ### Sentence 1 ###: {input_sentence_1} ### Sentence 2 ###: {input_sentence_2} ### Your Response ###: ("Yes" or "No"):

Figure 7: Semantic equivalence comparison prompt.

Multilingual Question-Answering Prompts	
English	You are an excellent question responder. Please correctly answer the following questions. {few-shot-examples} *** Question ***: {question} *** Answer ***:
Chinese	你是一个出色的问题回答者。请正确回答下列问题。 {few-shot examples} *** 问题 ***: {question} *** 答案 ***:
Japanese	あなたは優れた質問回答者です。次の質問に正しく答えてください。 {few-shot examples} *** 質問 ***: {question} *** 答え ***:
French	Vous êtes un excellent répondeur aux questions. Veuillez répondre correctement aux questions suivantes. {few-shot examples} *** Question ***: {question} *** Réponse ***:
Thai	คุณเป็นคนตอบคำถามได้ดีเยี่ยม,กรุณาตอบคำถามต่อไปนี้ให้ถูกต้อง. {few-shot examples} *** คำถาม ***: {question} *** คำตอบ ***:

Figure 8: Multilingual Question-Answering prompts.

E Appendix Experiments

E.1 Experiments on Llama-2 and Vicuna

We present the experimental results on Llama-2-13B-Chat (Llama-2) [7] (Touvron et al., 2023), and Vicuna-7B-v1.5 (Vicuna-v1.5) [8] (Zheng et al., 2023) in Table 9. Results suggest that confidence estimation abilities are relatively weak in both Llama-2 and Vicuna-1.5 across three methods.

E.2 Experiments of Extended Confidence Estimations

E.2.1 Experiments of Multilingual Confidence Estimations with Paraphrasing

Following Xiong et al. (2024), we investigate the prompt sensitivity for multilingual confidence estimation by introducing perturbations in the questions. We utilize GPT-3.5 to paraphrase the questions in different ways to generate different responses. We sample 200 questions from SciQ and prompt GPT-3.5 to paraphrase these questions. We also employ GPT-3.5 to check the semantic equivalence before and after paraphrasing to ensure the meaning is not changed. The AUROC and ECE results are presented in Table 10 and Figure 13. The findings and analysis are in Sec. 4.1.

E.2.2 Experiments of Multilingual Confidence Estimations of Sampling

To make comparisons, we also present the AUROC and ECE results of sampling-based confidence estimation methods on 200 samples from our multilingual SciQ datasets by setting Temperature T=0.8 on GPT-3.5. We cluster the sampled responses in semantic spaces and calculate the consistency score as Xiong et al. (2024) to represent the confi-

[7] https://huggingface.co/meta-llama/Llama-2-13b-chat
[8] https://huggingface.co/lmsys/vicuna-7b-v1.5

Multilingual p(True)-based Confidence Estimation Prompt	
English	You are an excellent referee to judge the answer correct or not. *** Question ***: {question} *** Proposed Answer ***: {first_proposed_answer} Is the proposed answer: True False The possible answer is:
Chinese	你是一个出色的判断者来评判问题的回答正确还是错误。 *** 问题 ***: {question} *** 模型回答 ***: {first_proposed_answer} 请判断这个问题的回答是正确的吗: 正确 错误 你的判断是:
Japanese	あなたは、答えが正しいかどうかを判断する優れた審判です。 *** 質問 ***: {question} *** 提案された回答 ***: {first_proposed_answer} 提案された答えは 真 誤り 可能な答えは:
French	Vous êtes un excellent arbitre pour juger la réponse correcte ou non. *** Question ***: {question} *** Réponse proposée ***: {first_proposed_answer} Est la réponse proposée : Vrai Faux La réponse possible est :
Thai	คุณเป็นผู้ตัดสินที่ยอดเยี่ยมในการตัดสินว่าคำตอบถูกหรือไม่. *** คำถาม ***: {question} *** คำตอบที่เสนอ ***: {first_proposed_answer} คำตอบที่เสนอคือ: จริง เท็จ คำตอบที่เป็นไปได้คือ:

Figure 9: Multilingual ***p(True)***-based Confidence Estimation Prompt.

dence. As presented in Table 10 and 13, the results demonstrate that our employed p(True) and Verb. methods outperform sampling-based methods as the high temperature may incur variability in output spaces which undermines the reliability of QA tasks.

E.2.3 Experiments of Multilingual Confidence Estimations using CoT

We supply the Chain-of-Thought (CoT) (Wei et al., 2022) for prompt-based confidence estimations of p(True) and Verbalized methods as in Table 10 and Figure 10. We present the AUROC and ECE results of p(True) and Verb. using CoT on 400 samples from SciQ and LSQA on GPT-3.5. Results suggest that CoT can marginally enhance the reliability of prompt-based confidence estimations in various languages.

Lang.	ko	id	it	ar	de
Prompt in *English*					
Accu. ↑	24.39	40.60	34.58	22.64	54.78
ARC. ↑	72.40	70.12	75.45	68.22	76.18
ECE ↓	33.55	36.78	33.16	46.78	27.14
NTP Method					
Accu. ↑	28.60	46.54	39.20	27.44	59.65
ARC. ↑	74.66	78.52	77.23	70.17	79.60
ECE ↓	28.10	32.44	30.50	42.76	23.18

Table 8: Experimental results of overall Accu., ARC., and ECE on the LSQA dataset by prompting using *English* and NTP method on other five investigated languages.

Self-verbalized Confidence Estimation Prompt

You are an excellent estimator. Provide your best estimated probability that the proposed answer is correct (0.0 to 1.0) for the following question. Give ONLY the probability, no other words or explanation.

*** Question ***: {question}
*** Proposed Answer ***: {first_proposed_answer}
*** Your estimated probability ***:

Figure 10: Self-verbalized confidence estimation prompt.

E.3 Experiments on Extended Languages

To further validate the observed linguistic dominance in multilingual confidence estimations, we employ five subsets derived and translated from TriviaQA into *Korean* (**ko**), *Arabic* (**ar**), *German* (**de**), *Indonesian* (**id**), and *Italian* (**it**) as in Sec. 2. The LA experiments are conducted on dataset translated from TriviaQA in all investigated languages in Table 11. We also develop small-size LSQA subsets for such languages and conduct LS experiments in 8.

F Uncertainty Estimations

Both *confidence* and *uncertainty* estimations indicate the level of assurance of a response generated by LLMs given a query and are occasionally regarded interchangeably (Geng et al., 2023). To alleviate over-confidence and enhance the reliability of LLMs, reliable uncertainty estimation is essential to determine whether a question is known or not to the LLM (Geng et al., 2023). Both *Uncertainty* and *Confidence* estimations can indicate the reliability degree of the responses generated by LLMs, and are generally used interchangeably (Xiao et al., 2022; Chen and Mueller, 2023; Geng et al., 2023). In this part, we investigate several commonly used *confidence* & *uncertainty* estimation methods for generative LLMs as mentioned in Sec. 7. Specifically, we denote $\mathsf{Conf}(\boldsymbol{x}, \boldsymbol{y})$ as the confidence score associated with the output sequence $\boldsymbol{y} = [y_1, y_2, \ldots, y_N]$ given the input context $\boldsymbol{x} = [x_1, x_2, \ldots, x_M]$. We also illustrate the summarized estimation methods as well as their disadvantages in Fig. 14.

Likelihood-based Methods: Following model calibration on classification tasks (Guo et al., 2017b), Vazhentsev et al. (2023); Varshney et al. (2023); Wang et al. (2024) intermediately quantify sentence uncertainty over token probabilities. In traditional discriminative models, except likelihood-based methods, confidence estimations also include ensemble-based and Bayesian methods (Lakshminarayanan et al., 2017; Gal and Ghahramani, 2016; Xue et al., 2022; Wang and Yeung, 2020; Gal et al., 2016; Abdar et al., 2021), and density-based methods (Lee et al., 2018). However, this likelihood-based method requires access to token probabilities and thus being limited to white-box LLMs. The likelihood-based confidence is estimated by calculating the joint token-level probabilities over $\boldsymbol{y}$ conditioned on $\boldsymbol{x}$. As longer sequences are supposed to have lower joint likelihood probabilities that shrink exponentially with length, the product of conditional token probabilities of the output should be normalized by calculating the geometric mean by the sequence length (Murray and Chiang, 2018; Malinin and Gales, 2021), and the confidence score can be represented as:

$$\mathsf{Conf}(\boldsymbol{x}, \boldsymbol{y}) = \left(\prod_{i}^{N} p(y_i | \boldsymbol{y}_{<i}, \boldsymbol{x}) \right)^{\frac{1}{N}} \quad (5)$$

Similarly, the arithmetical average of the token probabilities is adopted in Varshney et al. (2023):

$$\mathsf{Conf}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{N} \sum_{i}^{N} p(y_i | \boldsymbol{y}_{<i}, \boldsymbol{x}) \quad (6)$$

Furthermore, a low probability associated with even one generated token may provide more informative evidence of uncertainty (Varshney et al., 2023). Hence, the minimum of token probabilities is also employed.

$$\mathsf{Conf}(\boldsymbol{x}, \boldsymbol{y}) = \min \{ p(y_1 | \boldsymbol{x}), \ldots, p(y_N | \boldsymbol{y}_{<N}, \boldsymbol{x}) \} \quad (7)$$

Prompting-based Methods: Recently, LLMs' remarkable instruction-following ability (Brown et al., 2020) provides a view of instructing LLMs

Examples of LSQA Dataset in English and Japanese Specific Subsets

English

```
{
    "question": {
        "en": "What is the highest mountain in the United Kingdom?",
        "zh": "英国最高的山是哪座？",
        "th": "ภูเขาที่สูงที่สุดในสหราชอาณาจักรคืออะไร?",
        "ja": "イギリスで最も高い山は何ですか？",
        "fr": "Quelle est la plus haute montagne du Royaume-Uni ?"
    },
    "answer": {
        "en": "The highest mountain in the United Kingdom is Ben Nevis.",
        "zh": "英国最高的山是本尼维斯山。",
        "th": "ภูเขาที่สูงที่สุดในสหราชอาณาจักรคือเบนเนวิส",
        "ja": "イギリスで最も高い山はベンネビスです。",
        "fr": "La plus haute montagne du Royaume-Uni est le Ben Nevis."
    }
}
```

Japanese

```
{
    "question": {
        "en": "Which city in Japan is known for its deer population?",
        "zh": "哪个日本城市以其鹿群闻名？",
        "th": "เมืองใดในญี่ปุ่นที่มีชื่อเสียงจากประชากรกวาง?",
        ja": "日本のどの都市が鹿の数で知られていますか？",
        "fr": "Quelle ville au Japon est connue pour sa population de cerfs?"
    },
    "answer": {
        "en": "The city known for its deer population in Japan is Nara.",
        "zh": "以鹿群闻名的日本城市是奈良。",
        "th": "เมืองที่มีชื่อเสียงจากประชากรกวางในญี่ปุ่นคือนารา",
        "ja": "鹿の数で知られる日本の都市は奈良です。",
        "fr": "La ville connue pour sa population de cerfs au Japon est Nara."
    }
}
```

Figure 11: Examples of the LSQA ataset in *English* and *Japanese* specific subsets.

to self-estimate their confidence level to previous inputs and outputs including expressing uncertainty in words (Lin et al., 2022; Zhou et al., 2023; Tian et al., 2023a; Xiong et al., 2024), or instructing the LLM to self-evaluate its correctness on $p(\text{True})$ (Kadavath et al., 2022). The $P(\text{True})$ confidence score is implemented by simply asking the model itself if its first proposed answer $\boldsymbol{y}$ to the question $\boldsymbol{x}$ is true (Kadavath et al., 2022), and then obtaining the probability $p(\text{True})$ assigned by the model, which can implicitly reflect self-reflected certainty as follows.

$$\mathsf{Conf}(\boldsymbol{x}, \boldsymbol{y}) = p(\text{True}) = p(\boldsymbol{y} \text{ is True?}|\boldsymbol{x}) \quad (8)$$

Another method is to prompt LLMs to linguistically express tokens of confidence scores in verbalized numbers or words (Lin et al., 2022; Mielke et al., 2022; Zhou et al., 2023; Tian et al., 2023b; Xiong et al., 2024).

The sampling-based method refers to randomly sampling multiple responses given a fixed input $\boldsymbol{x}$ using beam search or temperature sampling strategies (Manakul et al., 2023; Xiong et al., 2024; Lyu et al., 2024). Various aggregation methods are adopted on sampled responses to calculate the consistency level as the confidence score. Kuhn et al. (2023) proposes semantic entropy to quantify uncertainty for sequences with shared meanings at the semantic level. Moreover, some uncertainty quantification methods are used to calculate the entropy indicating the dispersion level of multiple outputs (Kuhn et al., 2023; Lin et al., 2023).

Training-based Methods: For training methods, an external evaluator trained on specific datasets is introduced to output a confidence score given an input and an output. The evaluator can be a pre-trained NLI model (Mielke et al., 2022), or a value head connected to the LLM output layer (Lin et al., 2022; Kadavath et al., 2022), or the LLM itself (Han et al., 2024).

Native-Tone Prompting (NTP) Strategy

You are an excellent natural language inference model. You are required to identify the language spoken in the country related to the question.
{few_shot_examples}

*** Question ***: {question}
*** Your Identified Language Category ***: [Output Language]

Answer the following question in {Output Language}.
{few_shot_examples}

*** Question ***: {question}
*** Answer ***:

Figure 12: Native-tone prompting (NTP).

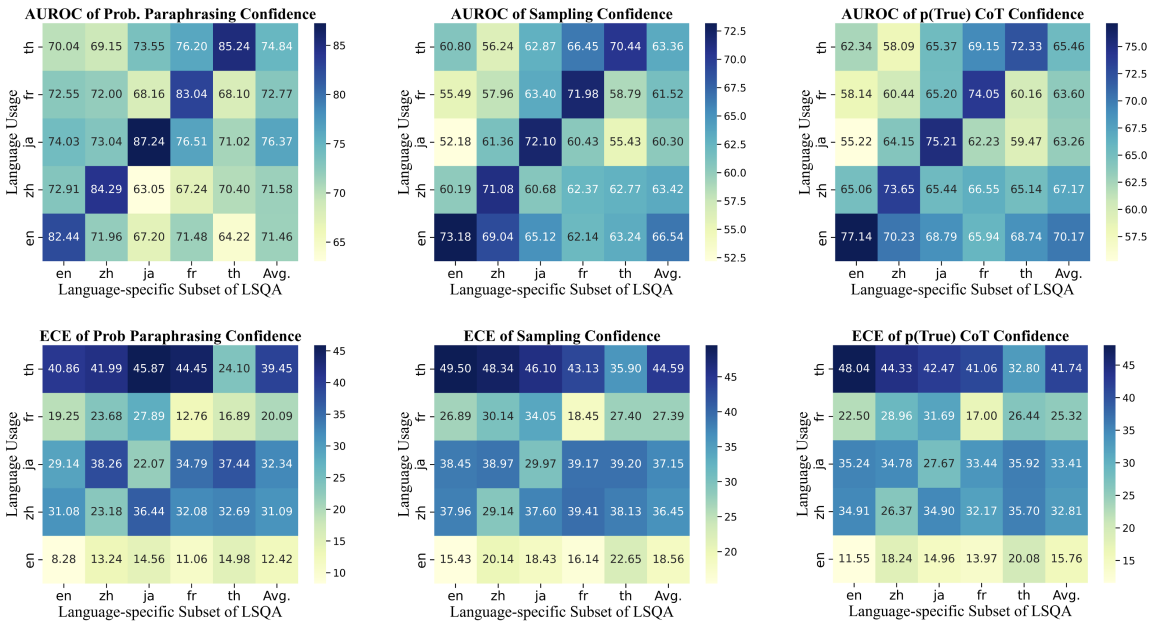


Figure 13: Experimental results of AUROC and ECE of three confidence estimation variants of paraphrasing, sampling, and CoT on LSQA for LS task on GPT-3.5.

Conf.	en		zh		ja		fr		th	
	ARC. ↑	ECE ↓	ARC. ↑	ECE ↓	ARC. ↑	ECE ↓	ARC. ↑	ECE ↓	ARC. ↑	ECE ↓
TVQA on *Llama-2*										
Prob.	51.92	20.56	51.24	34.12	51.92	32.04	49.51	29.33	49.63	48.72
p(True)	55.89	17.09	82.65	46.11	81.76	43.47	65.79	29.59	70.62	55.99
Verb.	59.78	21.10	53.20	45.71	51.95	39.33	61.66	38.98	54.23	59.80
GSM8K on *Llama-2*										
Prob.	42.72	32.18	50.46	33.59	50.35	51.67	43.60	36.25	55.11	52.87
p(True)	60.82	49.30	59.86	58.87	62.89	67.39	59.51	62.36	47.91	75.22
Verb.	59.39	43.65	53.29	54.59	53.40	49.27	53.26	37.61	54.53	56.96
CSQA on *Llama-2*										
Prob.	49.30	30.40	49.95	31.72	50.28	43.28	49.72	27.40	50.23	40.84
p(True)	56.53	26.05	55.34	45.65	53.46	46.01	59.76	25.49	50.21	63.09
Verb.	53.64	19.54	51.74	24.06	50.36	34.03	52.93	15.08	50.73	62.01
SciQ on *Llama-2*										
Prob.	55.40	24.65	76.39	44.42	74.97	45.56	62.32	39.76	51.93	59.05
p(True)	48.60	32.18	52.02	40.44	51.60	30.19	49.53	32.50	45.26	43.75
Verb.	56.34	19.89	55.17	41.36	55.58	37.20	60.27	39.14	71.17	54.95
TVQA on *Vicuna-1.5*										
Prob.	45.45	35.34	48.43	47.07	51.75	36.63	46.13	35.19	53.17	40.73
p(True)	47.45	23.58	78.96	42.86	79.71	42.40	60.38	28.89	76.58	53.43
Verb.	55.74	21.41	52.98	57.36	50.94	54.89	55.46	42.76	45.86	71.83
GSM8K on *Vicuna-1.5*										
Prob.	50.90	53.91	51.07	49.00	50.51	53.73	50.42	49.08	50.19	55.40
p(True)	65.40	68.30	67.28	59.33	51.09	60.70	66.78	55.69	52.86	60.83
Verb.	55.66	46.26	53.90	48.75	54.60	48.03	53.70	45.62	61.81	51.87
CSQA on *Vicuna-1.5*										
Prob.	48.88	26.04	50.01	43.65	49.67	45.64	45.94	31.53	49.78	51.78
p(True)	65.00	27.06	57.39	35.80	57.62	38.54	48.21	25.95	50.53	55.54
Verb.	52.32	29.80	52.29	38.59	51.08	44.49	58.68	35.15	51.90	61.77
SciQ on *Vicuna-1.5*										
Prob.	38.10	50.94	48.69	44.44	50.07	42.19	38.65	37.26	49.75	49.70
p(True)	45.78	31.29	73.55	40.17	66.85	45.15	59.20	42.18	74.16	58.34
Verb.	55.13	36.47	51.66	55.27	51.92	56.98	56.33	57.93	52.89	57.74

Table 9: Experimental results of AUROC (ARC.) and ECE of three confidence estimation methods on four LA datasets on *Llama-2* and *Vicuna-1.5*.

Conf.	en		zh		ja		fr		th	
	ARC. ↑	ECE ↓	ARC. ↑	ECE ↓	ARC. ↑	ECE ↓	ARC. ↑	ECE ↓	ARC. ↑	ECE ↓
SciQ on *GPT-3.5*										
Prob.	69.58	30.04	67.14	36.77	81.44	45.40	74.35	36.98	72.55	51.34
p(True)	72.80	23.86	77.56	31.99	82.44	38.27	72.00	40.13	63.45	40.80
Verb.	71.43	22.18	72.50	36.47	72.95	31.43	74.16	31.97	73.40	42.34
Re-Prob.	67.47	28.16	72.86	33.43	75.69	41.05	71.40	34.88	80.37	48.96
Re-p(True)	74.14	25.14	82.66	32.04	76.96	36.70	71.48	42.13	64.44	42.05
Re-Verb.	73.80	21.96	73.40	35.13	79.49	30.60	66.16	32.65	73.19	40.44
Sampling	67.55	27.40	71.69	37.97	74.07	42.09	67.94	40.04	66.50	48.65
CoT-p(True)	73.65	22.95	80.05	29.90	82.16	37.10	71.92	30.86	65.90	40.19
CoT-Verb.	73.64	20.60	75.73	32.79	74.61	27.50	72.62	31.26	74.96	40.33

Table 10: Experimental results of AUROC and ECE of several confidence estimation variants of paraphrasing the questions, sampling multiple responses, and adding CoT on SciQ for LA task on GPT-3.5.

Conf.	en		ko		it		ar		de		id	
	ARC. ↑	ECE ↓	ARC. ↑	ECE ↓	ARC. ↑	ECE ↓	ARC. ↑	ECE ↓	ARC. ↑	ECE ↓	ARC. ↑	ECE ↓
TriviaQA on *GPT-3.5*												
Prob.	69.58	30.04	73.21	46.37	73.08	28.60	71.51	46.78	72.48	33.74	77.37	50.12
p(True)	72.80	23.86	63.19	40.66	70.67	35.47	63.24	50.55	78.49	26.16	66.08	49.81
Verb.	71.43	22.18	72.41	34.80	72.19	41.54	76.65	28.68	68.75	47.14		

Table 11: Experimental results of AUROC and ECE of confidence estimations on other languages on TriviaQA for LA task on GPT-3.5.

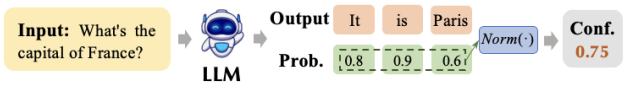
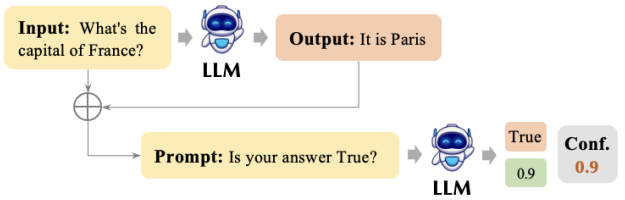
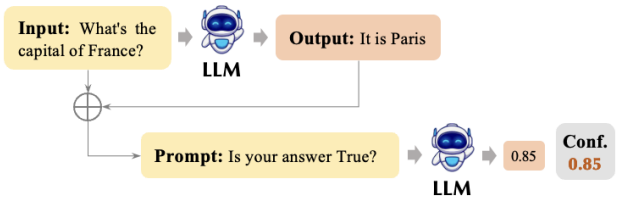
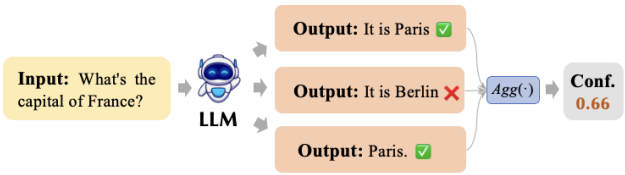
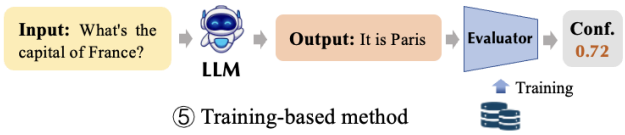
Confidence & Uncertainty Estimation Methods on LLMs	Disadvantages
 ① Probability-based method (***Prob.***)	a. Requires normalization due to variable sequence length; b. Requires access to token-level probabilities, inapplicable to black-box LLMs; c. Fails to capture semantic meaning over token-level probabilities.
 ② p(True)-based method (***p(True)***)	a. Relies on prompting strategies to elicit confidence estimation, varying in different prompts; b. Cannot improve LLM's intrinsic confidence estimation ability. c. Requires access to token-level probabilities, inapplicable to black-box LLMs; d. Prone to be over-confident.
 ③ Self-verbalized method (***Verb.***)	a. Relies on prompting strategies to elicit confidence estimation, varying in different prompts; b. Cannot improve LLM's intrinsic confidence estimation ability. c. Prone to be over-confident.
 ④ Sampling-based method	a. Requires additional inference time cost; b. Varying in different aggregation methods; c. Cannot improve LLM's intrinsic confidence estimation ability.
 ⑤ Training-based method	a. Requires training an additional evaluator; b. Difficult to learn LLM's intrinsic confidence estimation on unseen domains.

Figure 14: An illustration of several confidence estimation methods as well as their drawbacks. Note that sampling- and training-based methods are omitted in this work as they are cost-expensive and time-consuming for multilingual confidence estimations. All complete multilingual prompts used in this work are presented in Appendix C. In addition, although *confidence* and *uncertainty* are always used interchangeably, the former *confidence* pertains to the model's certainty regarding a specific generation, while the latter *uncertainty* denotes the "dispersion" of potential predictions for a given context. In this work, the semantically equivalent inputs in various languages are thoroughly distinct in token space. Therefore, we utilize *confidence estimation* in this work albeit specific uncertainty quantification methodologies are still applicable.