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-evaluate 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 on LS tasks.

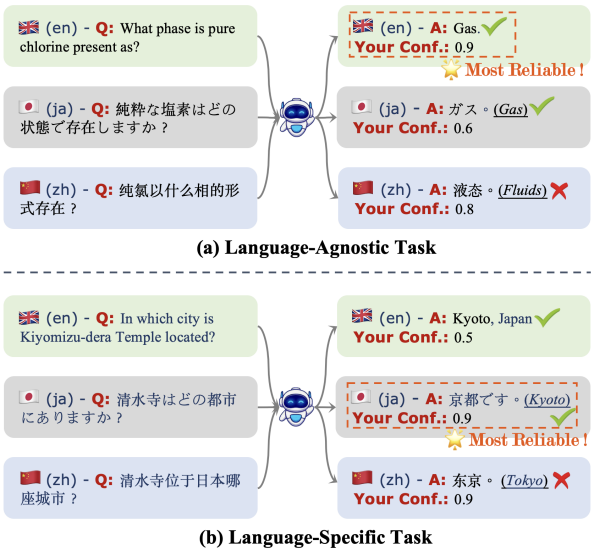


Figure 1: Examples of generations and confidence scores of *Llama-3.1* given the same inputs in three languages in LA and LS scenarios derived from SciQ and LSQA datasets respectively.

1 Introduction

LLMs' susceptibility to generating hallucinated contents in their generations 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 of LLMs has been confined exclusively in English (Kadavath et al., 2022; Lin et al., 2022; Geng et al., 2023; Tian et al., 2023b). Considering the LLMs' global deployment, the dearth of confidence estimations in languages other than English hinders users from ascertaining the reliability of generations in non-English scenarios. Owing to the disparities in the quantity and domain coverage of pre-training corpora across different languages, the confidence estimation ability may also presumably vary. The performance of confidence estimation methods mainly employed in English remains a subject worthy of exploration when generalized to other languages.

Additionally, linguistic dominance referring to that one language holds a superior status over others within a specific social or cultural context

(Blommaert, 2010; Treffers-Daller, 2019; Heller, 2007), also exist in confidence estimation ability on different languages. To achieve more reliable confidence estimations in multilingual contexts over various tasks, we study and divide the tasks into language-agnostic (LA) (Zhao et al., 2020) and language-specific (LS) scenarios as in Figure 1. 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 task 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* on both LA and LS tasks respectively. First, we meticulously construct the MlingConf dataset for this study including five high-quality multilingual datasets in LA and LS scenarios. The LA involves four tasks including closed-book factual knowledge question-answering (QA) derived from TriviaQA (Joshi et al., 2017), math reasoning from GSM8K (Cobbe et al., 2021), common sense reasoning from CommonsenseQA (Talmor et al., 2019), and professional scientific knowledge QAs from (Johannes Welbl, 2017). These four datasets 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) for the LS scenario, which is comprised of specific social, cultural, and geographical QA pairs for the investigated five languages respectively. To ensure the data quality, we conduct rigorous translation consistency checks to filter the failed samples and finally employ linguistic experts for human evaluation.

Experiments are conducted on three major LLM confidence estimation methods including probability-based (Vazhentsev et al., 2023; Varshney et al., 2023), $p(\text{True})$-based (Kadavath et al., 2022), and self-verbalized (Lin et al., 2022; Xiong et al., 2024) confidence estimations 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 all the confidence estimation methods demonstrate comparable performance of reliability across different languages, while the calibration performance in English significantly outperforms others. 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 for the response. Compared to the use of any single language, this approach leads to significant improvements in both accuracy and confidence estimation.

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 [1].
- We conduct experiments on MlingConf datasets and the results reveal that language dominance affects confidence estimation performance in *English* on LA tasks and query-related languages on LS tasks respectively.
- 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

2.1 Data Source

Owing to the lack of multilingual resources to comprehensively exhibit confidence estimation across diverse languages, our study leverages several datasets in *English* (**`en`**) and translates into **four** languages, encompassing *Japanese* (**`ja`**), *Chinese* (**`zh`**), *French* (**`fr`**), and *Thai* (**`th`**). We evaluate the quality of confidence estimates on our proposed MlingConf dataset including four datasets for the language-agnostic (LA) scenario and one dataset for the language-specific (LS) scenario.

Language-Agnostic (LA) Tasks For LA tasks, we employ the following four datasets that are widely used for confidence estimation research

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

in *English* in Kuhn et al. (2023); Xiong et al. (2024). We translate the datasets into other four languages. 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.

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 12, for example, all questions in *Japanese* subset pertain to Japanese social culture, history, or geography. All the subsets are also translated into other four languages.

The language selection, dataset details, prompts, and samples extracted from the LSQA are demonstrated in Appendix A, B, and C respectively.

2.2 Dataset Construction

The construction of the MlingConf dataset in this study follows a 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`, `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 shown in Table 1 and more clean multilingual datasets are obtained. Moreover, we present the number of samples for each language-specific subset in the LSQA dataset as in Table 2.

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

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

Table 1: Number of samples before and after consistency check and filter.

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

Table 2: Statistics of samples for each language-specific subset of the LSQA dataset.

Stage 3 : Finally, we employ several experts majoring in linguistics to examine the translation performance across 50 randomly selected samples as shown in Table 3. Human evaluation results suggest the obtained multilingual datasets are high-quality for further experiments. Temperature T is set to 0 for all the generations. 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 3: 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 6. 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:

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

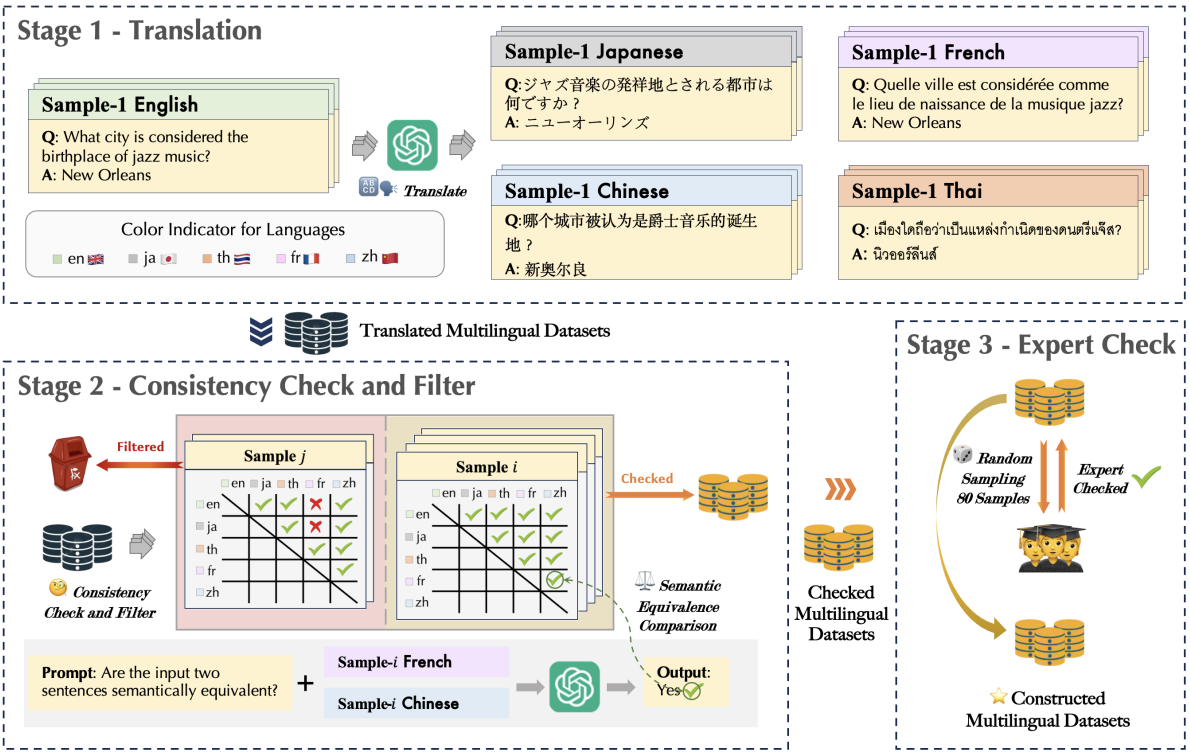


Figure 2: Three stages of MlingConf dataset construction.

p(True)**-based Confidence (*p(True)*):** The p(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(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 (Accu.) 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 within the range of [0, 1], implemented by `sklearn` toolkit [3]. A higher AUROC score implying higher reliability is preferred.

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

Dataset	Conf.	en		zh		ja		fr		th		*Avg.*	
		ARC. ↑	ECE ↓	ARC. ↑	ECE ↓	ARC. ↑	ECE ↓	ARC. ↑	ECE ↓	ARC. ↑	ECE ↓	ARC. ↑	ECE ↓
GPT-3.5													
TVQA	***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
GSM8K	***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
CSQA	***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
SciQ	***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
Avg.	***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
Llama-3.1													
TVQA	***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	***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	***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	***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.	***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 4: Experimental results of AUROC (ARC.) and ECE of three confidence estimation methods on four LA datasets on *GPT-3.5* and *Llama-3.1*.

3.3 Implementation Details

The experiments are conducted on three LLMs: *GPT-3.5-Turbo-0125* (*GPT-3.5*), *Llama-3.1-8B-Instruct* (*Llama-3.1*) [4] (Dubey et al., 2024), and *Llama-2-13B-Chat* (*Llama-2*) [5] (Touvron et al., 2023), and *Vicuna-7B-v1.5* (*Vicuna-v1.5*) [6] (Zheng et al., 2023). We only present the results of the current most commonly used commercial LLM *GPT-3.5* and open-source LLM *Llama-3.1* in the main part and leave others in Appendix E. Few-shot prompts containing N_f examples are utilized for answer generation with the greedy decoding where N_f is set to 8 for GSM8K and 4 for others.

4 Experimental Results and Analysis

In the experiment part, we aim to address the following three research questions:

a. How do performance (ARC. and ECE) vary in three confidence estimation methods on LA tasks?

b. How do performance and accuracy vary in different languages on LA tasks?

c. How do performance and accuracy vary in language usage on different language-specific LSQA subsets on LS tasks?

4.1 Results and Analysis on LA Tasks

We present the results of the confidence estimation performance of four LA datasets using *GPT-3.5* and *Llama-3.1* as presented in Table 4. The key findings and analysis for research questions *a* and *b* are listed below.

Recalling research question *a*, we highlight the supreme performance using ▒ among the three methods for each column of the dataset in Table 4. **On *GPT-3.5*, both *p(True)* and *Verb.* yield the superior performance 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.*** **show the 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*.

[4] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[5] https://huggingface.co/meta-llama/Llama-2-13b-chat
[6] https://huggingface.co/lmsys/vicuna-7b-v1.5

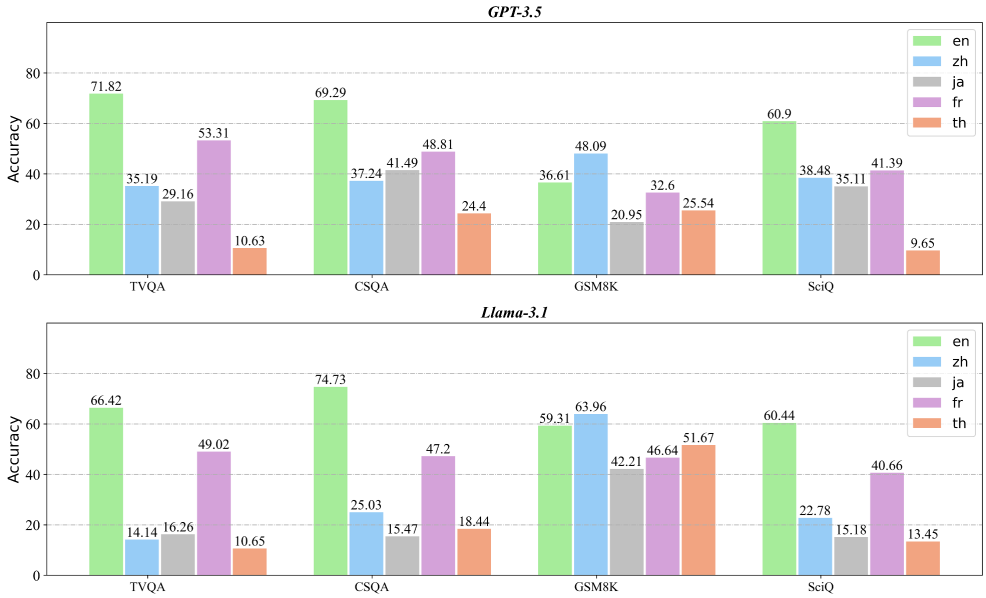


Figure 3: Experimental results of Accuracy on four LA datasets on *GPT-3.5* and *Llama-3.1*.

The performance differences between GPT-3.5 and Llama-3.1 can be attributed to that GPT-3.5's strong instruction-following abilities benefits the prompt-based confidence estimation methods Verb. and p(True)., while Llama-3.1's likelihood probabilities are well-calibrated during training.

Recalling research question *b*, previous studies only validate the efficacy of prompt-based confidence assessments exclusively within English. Our findings indicate that the approaches are also preferable in other languages and the performance fluctuates in different languages. Specifically as demonstrated in Table 4, 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*.

Despite the powerful multilingual capacity of LLMs, discrepancies exist in the quantity of distinct linguistic training corpora available for each language. The results 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 the specific calibration 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.

We also report the response accuracy on LA datasets in Figure 3, *English* consistently performs better on the datasets exclusively GSM8K. *Considering consistency check Table 1, the lowest filter rate in GSM8K translation indicates that mathematical reasoning tasks are minimally affected by language bias.* As a result, the performance fluctuations across different languages on GSM8K are relatively small. For *Chinese* exhibit 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.

Additionally, *GPT-3.5* and *Llama-3.1* exhibit notable superiority on multilingual tasks over *Llama-2* and *Vicuna-1.5*. Therefore, we only present the results on *GPT-3.5* and *Llama-3.1* in Table 4 and leave the other two in Appendix E.

4.2 Results and Analysis on LS Tasks

We present the results of the confidence estimation performance on five language-specific subsets of the LSQA dataset using *Llama-3.1* as presented in Figure 4 and 5. The key findings and analysis for research question **c** are listed below.

Recalling research question **c**, in comparison with the LA task in Sec. 4.1 where linguistic dominance is primarily manifested in *English*, on LSQA, linguistic dominance is determined by the language of the subset. As illustrated 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

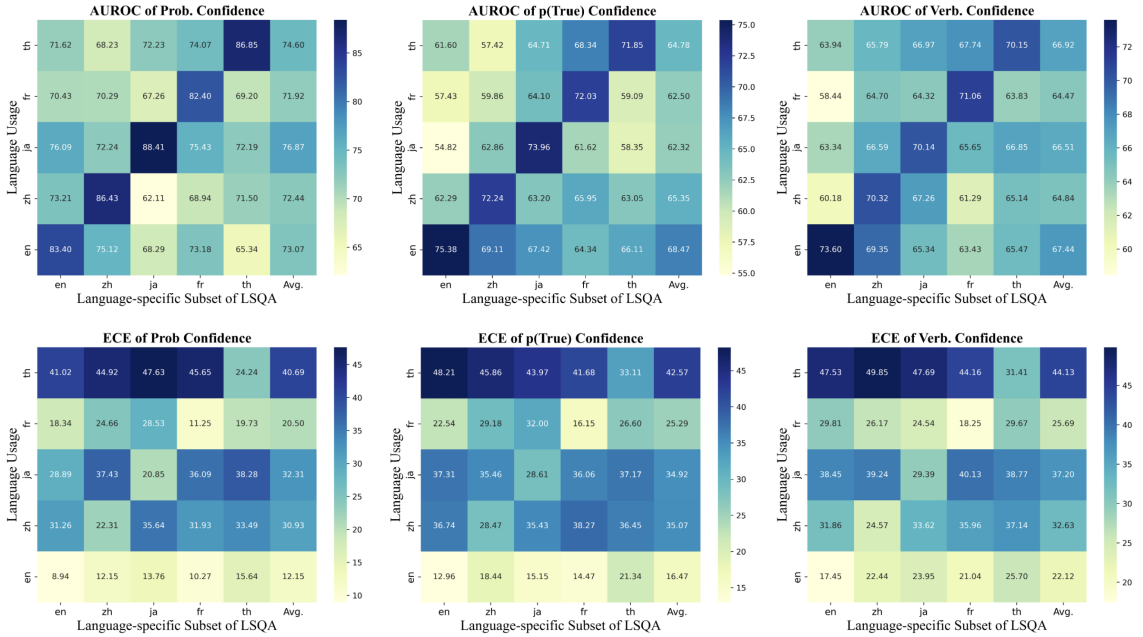


Figure 4: Experimental results of AUROC and ECE of three confidence estimation methods on five language-specific subset of LSQA using *Llama-3.1*.

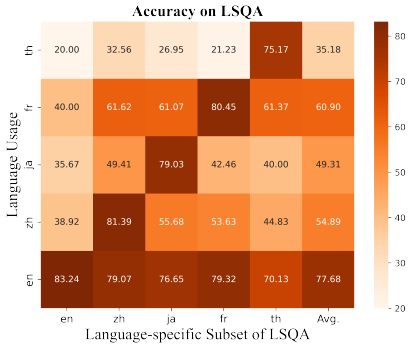


Figure 5: Experimental results of Accuracy on five language-specific LSQA subsets using *Llama-3.1*.

experiments in 5. Additionally, as shown 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.

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, thereby achieving optimal performance when prompting their specific languages. The use of maximum likelihood estimation during LLM pre-training further accounts for* ***Prob.*** *confidence exhibits more prominent linguistic dominance than* ***p(True)*** *and* ***Verb.*** *confidence estimations.*

5 Native-Tone Prompting Method

Confidence estimation performance differences caused by linguistic dominance on LA and LS tasks in Sec. 4 motivate us to further propose a research question *How can we better leverage linguistic dominance to improve the LLMs' reliability and accuracy on LA and LS tasks?*

Method In Sec. 4.1, prompting in *English* generally outperforms others, hence responding in *English* on LA tasks can be adequately credible and accurate where linguistic dominance is leading in *English*. Inspired by the linguistic dominance phenomena on each language-specific LSQA dataset in Sec. 4.2, we propose a simple yet effective Native-Tone Prompting (NTP) strategy to achieve better confidence estimation performance and accuracy on the LS tasks. NTP first prompts LLMs to identify the language context of the question, and then uses that language to answer the question, effectively exhibiting a "native tone" that is more familiar in that language-related context. The prompt of NTP is presented in Appendix C.

Experiments In Table 5, we present the results of prompting by any single language versus NTP. The experiments demonstrate that our proposed 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 LS tasks.

Prompt	en	zh	ja	fr	th	NTP
Accu. ↑	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 5: Experimental results of overall Accu., ARC., and ECE on the LSQA dataset by prompting using different languages and our proposed NTP method.

Discussion Linguistic dominance exerts a significant influence on multilingual confidence estimations on both LA and LS scenarios, where the dominance is manifested in *English* and query-related language respectively. This phenomenon indicates that the multilingual capabilities and reliability of LLMs are still constrained by the imbalanced training corpus among diverse languages. Furthermore, 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, the reliability of responses cannot be consistently maintained, highlighting challenges in enhancing multilingual reliability for LLMs.

6 Related Works

Confidence Estimation for LLMs Both *confidence* and *uncertainty* estimations indicate the level of certainty or assurance of a response $\boldsymbol{y}$ generated by LLMs given a query $\boldsymbol{x}$ and are occasionally regarded interchangeably (Geng et al., 2023). Previous confidence estimation methods can be categorized into five distinct classes, as illustrated in Figure 6. ① **Probability-based method**: Following model calibration on classification tasks (Guo et al., 2017b), Vazhentsev et al. (2023) intermediately quantify sentence uncertainty over token probabilities; ② $p(\text{True})$**-based method**: Kadavath et al. (2022) instructs the LLM to self-evaluate the correctness of the generated answer by directly accessing $p(\text{True})$; However, both ① and ② require access to token probabilities and thus are limited to **white-box LLMs**. ③ **Self-verbalized method**: Recently, 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) especially for **black-box LLMs**; ④ **Sampling-based method**: By sampling multiple responses to a given question, Xiong et al. (2024); Lyu et al. (2024) aggregate all the confidence scores as the indicator. Kuhn et al. (2023) proposes semantic entropy to quantify uncertainty for sequences with shared meanings at the semantic level. ⑤ **Training-based method**: Mielke et al. (2022); Lin et al. (2022); Han et al. (2024); Kadavath et al. (2022) propose to train an external natural language inference (NLI) module or LLM itself to improve confidence estimations.

Current research for LLM confidence estimations is focused on English. We first present a comprehensive study to investigate LLMs' confidence estimations in languages other than English. Note that sampling- and training-based methods are omitted in this work as they are cost-expensive and time-consuming for multilingual tasks.

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; Workshop et al., 2023; Chowdhery et al., 2022; Zhang et al., 2022; Brown et al., 2020). However, their efficacy in other low-resource languages remains limited, hindering the global proliferation of LLM deployment and provoking a surge in research interest. Recently, 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 underscored the advantages of leveraging pre-trained multilingual models in diverse cross-lingual applications (Wang et al., 2023a,b; Bhattacharjee et al., 2023; Qin et al., 2022). Inspired by them, our research systematically delves into an under-explored domain of multilingual confidence estimation.

7 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 performance of different confidence estimation methods 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 to refine these methods, enhancing the global applicability and trustworthiness of LLMs.

Limitations and Future Work

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

Extensions on More 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.

Extensions on Native-Tone Prompting 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

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.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat Shahriyar. 2023. CrossSum: Beyond English-centric cross-lingual summarization for 1,500+ language pairs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2541–2564, Toronto, Canada. Association for Computational Linguistics.

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.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *Preprint*, arXiv:2204.02311.

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*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks,

Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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*.

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.

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

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.

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.

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*.

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.

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.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden,

Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

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*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.

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.

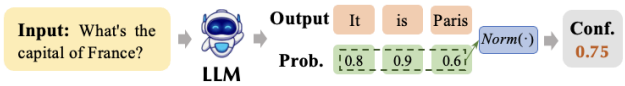
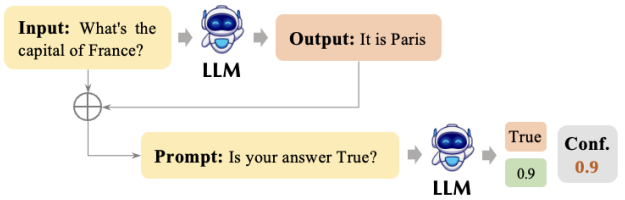
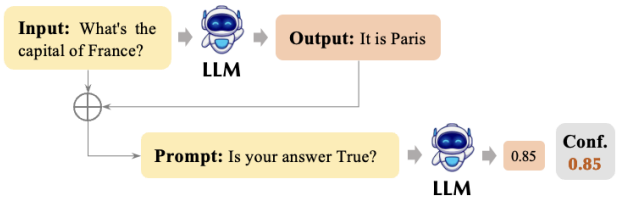
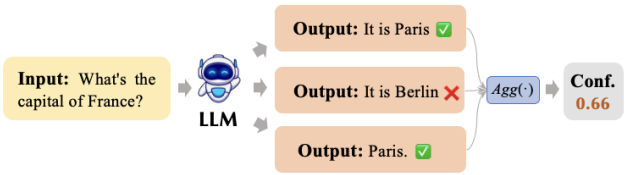
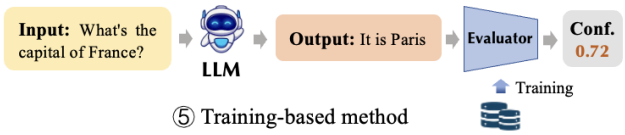
Confidence & Uncertainty Estimation Methods on LLMs	Disadvantages
 ① Probability-based method (***Prob.***)	a. Requires normalization due to variable sequence length; b. Requires access to token-level probabilities, inapplicable to black-box LLMs; c. Fails to capture semantic meaning over token-level 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 6: 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.

A Language Information

The basic information of ISO codes and the language family of the investigated languages is presented in Table 6. 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 CommonCrawl corpus [7], 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 6. 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 [8].

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

Table 6: 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.

[7] http://commoncrawl.org/

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

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.

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 12.

C Prompt Details

The translation prompt for multilingual dataset construction and semantic equivalence comparison prompt for consistency check in Sec, 2 are presented in 7 and 8 respectively. Standard multilingual Question-Answering prompts are in 9. Multilingual confidence estimations of ***P(True)*** and ***Verb.*** are presented in Fig. 10 and 11. 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:

Translation Prompt

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

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

Figure 7: Translation prompt.

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 8: Semantic equivalence comparison prompt.

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

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

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.

E Appendix Experiments

We present the experimental results on *Llama-2* and *Vinuca-1.5* in Table 8. Results suggest that confidence estimation abilities are relatively weak on both *Llama-2* and *Vinuca-1.5* across three methods.

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 9: Multilingual Question-Answering prompts.

Dataset	Conf.	en		zh		ja		fr		th	
		ARC. ↑	ECE ↓	ARC. ↑	ECE ↓	ARC. ↑	ECE ↓	ARC. ↑	ECE ↓	ARC. ↑	ECE ↓
Llama-2											
TVQA	***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	***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	***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	***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
Vicuna-1.5											
TVQA	***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	***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	***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	***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 8: Experimental results of AUROC (ARC.) and ECE of three confidence estimation methods on four LA datasets on *GPT-3.5* and *Llama-3.1*.

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

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 11: Self-verbalized confidence estimation prompt.

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

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 13: Native-tone prompting (NTP).