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 and exhibit overconfidence in predictions raises concerns regarding their reliability. Confidence or uncertainty estimations indicating the extent of trustworthiness of a model's response are essential to developing reliable AI systems. Current research primarily focuses on LLM confidence estimations in English, remaining a void for other widely used languages and impeding the global development of reliable AI applications. This paper introduces a comprehensive investigation of **M**ulti**ling**ual **Conf**idence estimation (MLINGCONF) on LLMs. First, we introduce an elaborated and expert-checked multilingual QA dataset. Second, we delve into the performance of confidence estimations and examine how these confidence scores can enhance LLM performance through self-refinement across diverse languages. Finally, we propose a cross-lingual confidence estimation method to achieve more precise confidence scores. The experimental results showcase the performance of various confidence estimation methods across different languages as well as present that our proposed cross-lingual confidence estimation technique significantly enhances confidence estimation and outperforms several baseline methods.

1 Introduction

Although LLMs have exhibited noteworthy proficiencies across a range of NLP tasks, their susceptibility to generating hallucinated contents and being overconfident 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; Hu et al., 2023; Anonymous, 2024; Kuhn et al., 2023; Lin et al., 2023; Chen and Mueller, 2023; Lin et al., 2022).

However, existing research on confidence or uncertainty estimations of LLMs is confined exclusively in English (Lin et al., 2023, 2022; Chen and Mueller, 2023; Duan et al., 2023; Varshney et al., 2023; Tian et al., 2023b). With the accelerated pace of globalization, it is essential to develop worldwide reliable AI applications. There is a dearth of research dedicated to extending the confidence estimation methods of LLMs reported in *English* to encompass other widely used languages such as *Chinese* and *French*. The limitation leaves the open problem of utilizing current confidence estimation methods in diverse linguistic contexts, raising the first research question: **i)** *How does the confidence estimation performance vary across various languages?* By performing several confidence estimation methods, the optimal approach applicable within multilingual environments can determined. When viewing different languages, we can discern if different languages have consistent confidence scores on LLMs. In addition, we can further explore the potential variations in different languages when using the obtained scores as feedback for LLM self-refinement (Madaan et al., 2023).

Furthermore, with the confidence scores in various languages, we can further explore the potential to ensemble the scores in the cross-lingual paradigm to improve confidence estimation. In prior studies, temperature sampling and prompt strategies have been extensively employed for confidence or uncertainty estimations (Chen and Mueller, 2023; Kuhn et al., 2023; Lin et al., 2023; Anonymous, 2024). These methods involve inducing LLMs to consider various generative pathways to acquire more accurate confidence scores akin to self-consistency (Wang et al., 2023c). However, temperature sampling entails sampling multiple times given a fixed context, where the input rep-

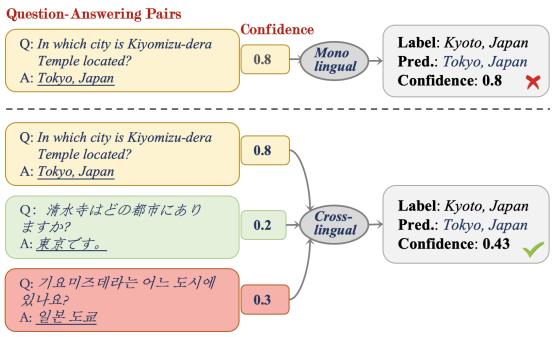

Figure 1: Comparison of monolingual and cross-lingual confidence estimation.

resentation lacks diversity. Prompt strategies fail to ensure that a fixed input after prompt rewriting is still strictly semantically equivalent as before (Kuhn et al., 2023), potentially exacerbating the dispersion of various outputs and posing challenges in uncertainty estimation (Lin et al., 2023). In contrast, the exploration of cross-lingual techniques offers a novel perspective on "thinking in different pathways in multilingual spaces" as shown in Figure 1, where we can maintain identical semantics of multilingual inputs with entirely distinct word embedding representations. This motivates the second question: **ii)** *How can we ensemble multilingual confidence scores to achieve more precise cross-lingual confidence estimates for LLM?*

To this end, we introduce MLINGCONF, a comprehensive investigation of multilingual confidence estimation for LLMs. First, we meticulously construct a high-quality benchmark for multilingual confidence estimation, named Mling-TriviaQA, derived from TriviaQA (Joshi et al., 2017). The dataset encompasses **ten** diverse languages and is rigorously checked by linguistic experts. We then dive into the above research questions as follows.

For question **i)**, we employ several confidence estimation methods on LLMs including likelihood-based confidence (Varshney et al., 2023), true probability confidence (Kadavath et al., 2022), and linguistic confidence (Lin et al., 2022) across a spectrum of languages. Then we treat the obtained confidence scores as feedback for self-refinement (Madaan et al., 2023) experiments to observe performance improvements across diverse languages.

Regarding question **ii)**, we propose a cross-lingual confidence estimation method. Languages selected for the cross-lingual technique are categorized by **language family**[1] wherein languages sharing the same language family exhibit analogous cognitive patterns (Bouckaert et al., 2012). Then we ensemble and average the monolingual confidence scores of selected languages for cross-lingual confidence estimations. Experimental results indicate our proposed cross-lingual confidence estimation outperforms the preceding prompting and temperature sampling methods in monolingual space.

The contributions are summarized as follows:

- To the best of our knowledge, this study is the first to investigate multilingual confidence estimations on LLMs for reliable global AI developments.
- We introduce an intricately constructed and expert-checked multilingual QA dataset, denominated Mling-TriviaQA, to serve as the multilingual confidence estimation benchmark.
- We present a comprehensive study on various languages with several confidence estimation methods and identify the most effective confidence estimation method across a spectrum of languages, as well as elucidate the optimal strategy for employing confidence scores as self-refinement feedback.
- We propose a cross-lingual confidence estimation methodology designed to procure more accurate confidence scores, demonstrating superior performance compared to antecedent prompting and temperature sampling methods.

2 Related Works

Confidence Estimation in Generative LMs Distinguished from previous work on confidence estimation on discriminative models (Jiang et al., 2021; Desai and Durrett, 2020; Guo et al., 2017), the *confidence* or *uncertainty* estimation on generative LMs poses greater challenges due to the variable sequence lengths and the inherent difficulty in aligning uncertainty in semantic space with human intentions (Geng et al., 2023; Kuhn et al., 2023). Numerous methodologies employing logit-based uncertainty estimation techniques on **white-box LLMs** have been extensively utilized in (Tian et al., 2023a; Vazhentsev et al., 2023; Xiao et al., 2022). Kadavath et al. (2022) instructs the LLM to self-evaluate the correctness of the generated answer by directly accessing $p(\text{True})$. Kuhn et al. (2023) introduces the concept of semantic entropy, aggregating invariant token sequences by shared meanings. Recent investigations pay particular attention to confidence estimation for the **black-box LLMs**. Lin et al. (2023) extends a range of easily implementable confidence measures. Lin

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

Lang.	Accu.	Lang.	Accu.	Lang.	Accu.
it	98%	ko	96%	ja	96%
fr	100%	zh	100%	de	100%
ar	98%	id	98%	th	96%

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

et al. (2022) instructs LLMs to produce a level of confidence in natural language. Chen and Mueller (2023) proposes BSDETECTOR to improve confidence estimation on LLMs. Anonymous (2024) focuses on confidence elicitation on LLMs to accurately express confidence.

The previous works only consider the confidence estimation in English and most are limited on black-box LLMs. Our work systematically investigates multilingual confidence estimations on both black-box and white-box LLMs. Typically, note that both *confidence* and *uncertainty* are occasionally regarded interchangeably (Chen and Mueller, 2023; Xiao et al., 2022), 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. Consequently, we utilize *confidence estimation* in this work albeit specific uncertainty quantification methodologies are still applicable.

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 languages remains limited, hindering the global proliferation of LLM deployment and provoking a surge in research interest. Cross-lingual learning paradigms can enhance monolingual performance on LLMs (Huang et al., 2023; Zeng et al., 2023; Qin et al., 2023). 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 delves into an underexplored domain of confidence estimation across different languages, alongside the development of cross-lingual confidence estimation methodologies on multilingual LLMs.

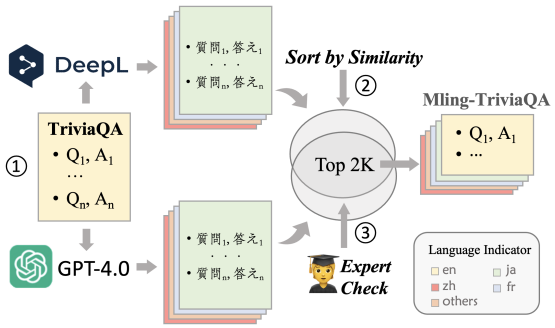

Figure 2: An illustration of data collection of Mling-TriviaQA.

3 Experimental Setting

3.1 Dataset Collection

Owing to the lack of multilingual resources to comprehensively exhibit confidence estimation across diverse languages, our study leverages an extensive multilingual dataset, namely Mling-TriviaQA derived from TriviaQA (*English*) (Joshi et al., 2017) across ten languages, encompassing *English, Japanese, Chinese, Korean, French, Arabic, German, Indonesian, Thai* and *Italian* on which we have employed linguistic experts to check the translation quality. These languages are selected according to the language family category as in Section 5.1. We extract and translate numerous closed-book question-answer pairs from TriviaQA which is widely used in confidence estimation (Chen and Mueller, 2023; Kuhn et al., 2023; Lin et al., 2023). The language family and ISO code information of the investigated languages and why we employ these languages for research are listed in Appendix B.

The construction of the Mling-TriviaQA dataset in this study follows a three-stage procedure as delineated in Figure 2. ①: the QA examples extracted from the TriviaQA dev set are translated into different languages through both GPT-4 (OpenAI, 2023) [2] and DeepL [3]; ②: we gauge the similarity scores of the results by two translators using the uni-gram F1-measure, retaining only the top 2k samples for further experiments; ③: after filtering, 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 translations are high-quality for experiments. The dataset details are presented in Appendix A.

[2] https://platform.openai.com/docs/api-reference
[3] https://www.deepl.com/pro-api

3.2 Generative LLMs

The experiments are carried out using two LLMs: `GPT-3.5 Turbo` and `GPT-4` (OpenAI, 2023). Few-shot prompts are utilized for answer generation, with a fixed temperature setting of $T = 0.8$. We leverage the APIs of `GPT-3.5 Turbo` and `GPT-4` to obtain generations and token-level log probabilities. Conversely, some prominent LLMs like PaLM 540B (Chowdhery et al., 2022) are black-boxed and unable to provide log probabilities, while some other LLMs like BLOOM (Workshop et al., 2023) do not encompass all the languages investigated in this work. Even though we have not conducted the side-by-side experiments on these LLMs as on `GPT-3.5 Turbo` and `GPT-4`, we still exhibit some experiment results of these LLMs in Appendix D.

3.3 Evaluations

Following (Filos et al., 2019; Kuhn et al., 2023; Chen and Mueller, 2023; Lin et al., 2023; Anonymous, 2024), we employ Area Under the Receiver Operator Characteristic Curve (AUROC) to assess the effectiveness of our confidence estimates, implement by `sklearn` toolkit [4]. AUROC quantifies how likely a correct answer possesses a higher uncertainty score than an incorrect one, yielding a score within the range of [0, 1]. A higher AUROC score is preferred.

For evaluating the answer quality of LLMs, we compute the F1 score, exact matching, and the NLI score as accuracy between the generated answer and gold reference, where the NLI score (*Entailment* scores 1 and both *Neutral* and *Contradiction* score 0) is assessed using `GPT-3.5 Turbo` and the multilingual prompts are provided in Appendix C.

4 Confidence Estimation on Various Languages

4.1 Confidence Estimation Methods

In this section, we investigate several commonly used confidence estimation methods for generative LLMs. 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 $\boldsymbol{x} = [x_1, x_2, \ldots, x_M]$.

Likelihood-based Confidence: 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, we calculate the geometric mean by normalizing the product of conditional token probabilities in the output by the sequence length (**Likelihood-Norm**) (Murray and Chiang, 2018; Malinin and Gales, 2021), and the confidence can be represented as:

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

Similarly, we also take the average (**Likelihood-Avg**) of the probabilities of tokens arithmetically:

$$\mathsf{Conf}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{M} \sum_{i}^{M} p(y_i | \boldsymbol{y}_{<i}, \boldsymbol{x}) \quad (2)$$

Furthermore, a low probability associated with even one generated token may provide more informative evidence of uncertainty (Varshney et al., 2023). Hence, we also employ the minimum of token probabilities. (**Likelihood-Min**).

$$\mathsf{Conf}(\boldsymbol{x}, \boldsymbol{y}) = \min \{p(y_1 | \boldsymbol{x}), \ldots, p(y_N | \boldsymbol{y}_{<N}, \boldsymbol{x})\} \quad (3)$$

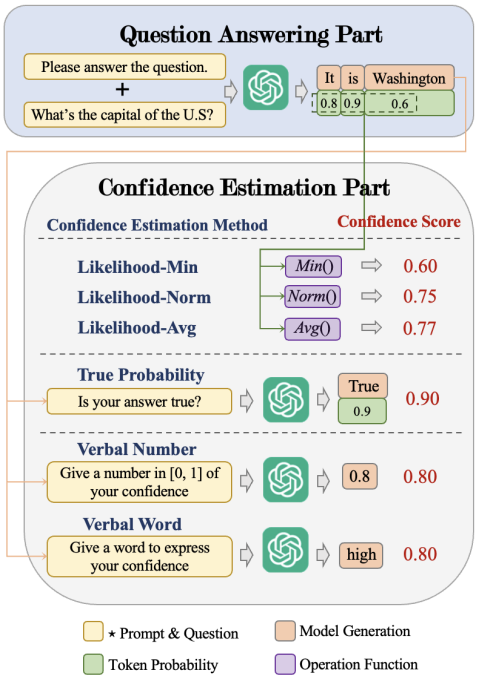

Figure 3: An illustration of several confidence estimation methods on LLMs. ⋆ Note that all complete multilingual prompts used in this work are presented in Appendix C.

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

Languages	en	zh	fr	ko	de	ja	it	ar	id	th	Avg
						GPT-3.5 Turbo					
Avg Length	4.71	15.87	9.22	15.51	8.03	12.96	10.58	14.79	7.75	22.72	-
Likelihood-Min	78.53	83.66	72.33	80.90	79.97	81.30	74.60	82.36	78.37	82.30	79.43
Likelihood-Avg	77.84	83.77	71.65	83.09	77.70	80.77	71.95	80.26	76.51	76.89	78.04
Likelihood-Norm	78.01	84.59	72.14	83.21	78.58	81.32	73.08	81.51	77.37	79.36	79.11
True Probability	83.97	94.55	80.59	**93.20**	81.49	**92.26**	80.77	**93.46**	86.08	**92.90**	87.92
Verbal Number	**87.14**	**94.69**	**83.47**	92.91	**87.26**	92.18	**84.94**	92.09	**88.64**	89.33	**89.24**
Verbal Word	85.60	89.04	82.48	87.08	85.05	88.63	81.48	87.61	86.15	86.39	85.95
						GPT-4					
Avg Length	4.30	7.90	5.95	8.61	5.62	7.72	5.47	14.79	5.25	11.15	-
Likelihood-Min	67.40	77.80	73.63	76.38	70.04	75.85	69.28	78.35	71.90	75.57	73.59
Likelihood-Avg	68.08	76.78	74.75	76.92	70.92	76.45	69.78	79.29	72.40	77.61	74.29
Likelihood-Norm	67.88	76.70	74.32	76.83	70.91	76.35	69.67	79.10	72.36	77.09	74.11
True Probability	87.33	91.23	84.03	92.60	87.44	**95.37**	85.44	89.13	82.66	**91.29**	88.65
Verbal Number	**89.97**	**96.13**	**85.03**	**96.14**	**89.89**	94.92	**88.85**	**92.65**	**87.19**	91.13	**91.18**
Verbal Word	82.24	84.48	82.41	89.04	85.45	92.59	80.83	91.36	86.65	90.25	86.53

Table 2: Experimental results of AUROC ↑ on Mling-TriviaQA set using several confidence estimation methods.

True Probability Confidence: The **True Probability** 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.

$$\texttt{Conf}(\boldsymbol{x}, \boldsymbol{y}) = p(\text{True}) = p(\boldsymbol{y} \text{ is True}|\boldsymbol{x}) \quad (4)$$

Self-verbalized Confidence: Recent works pay particular attention to linguistic confidence via prompting LLMs to express certainty in verbalized numbers or words (Lin et al., 2022; Mielke et al., 2022; Zhou et al., 2023; Tian et al., 2023b; Xiong et al., 2023). We adopt verbalized numerical probability (**Verbal Number**) and words (**Verbal Word**) in token-level space as LLM's confidence estimations. The verbalized word contains a set of five words (e.g. "lowest", "low", "medium", "high", "highest") indicating the confidence degrees.

4.2 Confidence Scores as Self-refinement Feedback on LLMs

Given the potential variations in confidence estimations across linguistic representations, e.g., the length of generated sequences (Kuhn et al., 2023), we desire to explore how confidence scores from different languages vary in assisting LLM to improve performance. Self-refinement on LLMs is a process that involves generating an initial answer and subsequently refining it based on self-provided feedback (Madaan et al., 2023) received for the previously proposed answer. The confidence scores acquired through this process serve as a fine-grained quantified indicator, guiding the LLM to correct or optimize its answers. Therefore, leveraging these confidence scores as feedback for LLM self-refinement across diverse languages facilitates an exploration of the intricate relationship between confidence estimation and its practical utilization.

Feedback We employ two methods to determine the feedback utilized: *fixed-value* and *random sampling* approaches. The *fixed-value* method introduces a threshold k, and the answers first proposed by the LLM with confidence score $\texttt{Conf}(\boldsymbol{x}, \boldsymbol{y}) \leq k$ will be self-refined by the LLMs. We conduct experiments with four thresholds spanning {0.25, 0.50, 0.75, 1.00}. In contrast, *random sampling* determines feedback by randomly sampling a value k from a uniform distribution $k \sim \text{U}(0, 1)$. Similarly, the self-refinement process is activated for LLM's first proposed answers with confidence scores $\texttt{Conf}(\boldsymbol{x}, \boldsymbol{y}) \leq k$.

4.3 Experimental Results and Analysis

In this section, we present the results of the confidence estimation performance across different languages and further employ the obtained confidence scores as self-refinement feedback on Mling-TriviaQA on both **GPT-3.5 Turbo** and **GPT-4**. Experimental results aim to address the following several sub-questions: a. *Which confidence estimation method consistently yields superior performance in most languages?* b. *What are the differences in the confidence estimation performance with respect to different languages?* c. *What strategies can be employed to optimize self-refinement performance*

utilizing the obtained confidence scores? d. *Can confidence scores as self-refinement feedback enhance the performance of LLMs in all languages?*

The key findings pertaining to the aforementioned four research questions are listed below.

a. The verbalized numerical confidence estimation method emerges as the optimal method among all other estimation methods. As shown in Table 2, self-verbalized numerical confidence attains prominence across most languages, yielding the highest AUROC scores on both **`GPT-3.5 Turbo`** and **`GPT-4`**. Previous studies only validate the efficacy of self-verbalized confidence assessment exclusively within English. Our findings indicate that the **Verbal-Num** approach is also preferable in other languages. Consequently, we have opted to employ this approach for subsequent experiments.

b. The confidence estimation results vary in different languages. This phenomenon can be attributed to several factors. Despite the powerful multilingual capacity of state-of-the-art LLMs **`GPT-3.5 Turbo`** and **`GPT-4`**, discrepancies persist in the quantity of distinct linguistic training corpora available for each language. Furthermore, we report a correlation between confidence estimation performance and the lengths of generated sequences across various languages. Notably, languages characterized by longer generated sequences, such as *Chinese* and *Thai*, demonstrate superior confidence estimation performance compared to those belonging to the Indo-European language family, such as *German*. This trend can be attributed to the notion that an increased number of tokens inherently contains more information, thereby aiding in the reduction of uncertainties.

c. Random sampling outperforms fixed-value threshold as confidence score feedback for LLM self-refinement. As shown in Table 3, we have investigated two methods to determine feedback in three languages encompassing *English*, *Chinese*, and *French*. For fixed threshold, self-refining the outputs of those with self-verbalized numerical confidence scores less than 0.75 yields significant performance enhancements. When the threshold is set to 1.00 which mandates self-refinement for all generated outputs, the performance dramatically decreases as many correct answers are inadvertently modified. Better or comparable performance is obtained using random sampling over the fixed-value method in all tested languages. Therefore, we have elected to employ the random sampling method as confidence estimation feedback for further self-refinement experiments in other languages.

Lang.	Feedback `Conf` $\leq k$	EM ↑	F1 ↑	Accu. ↑
en	First proposed	63.60	73.12	72.30
	$k = 0.25$	63.25	72.85	73.50
	$k = 0.50$	64.50	74.18	74.35
	$k = 0.75$	**64.85**	74.12	**74.65**
	$k = 1.00$	35.25	49.32	60.20
	$k \sim \mathrm{U}(0,1)$	64.60	**74.46**	74.50
zh	First proposed	16.90	17.14	23.55
	$k = 0.25$	16.95	17.32	27.60
	$k = 0.50$	17.25	19.44	32.75
	$k = 0.75$	17.55	21.00	34.85
	$k = 1.00$	11.70	12.82	14.30
	$k \sim \mathrm{U}(0,1)$	**18.40**	**21.36**	**38.70**
fr	First proposed	31.50	47.19	47.20
	$k = 0.25$	31.60	47.06	48.50
	$k = 0.50$	32.45	47.68	50.35
	$k = 0.75$	**32.85**	48.15	51.65
	$k = 1.00$	19.80	24.37	30.20
	$k \sim \mathrm{U}(0,1)$	32.80	**48.24**	**53.85**

Table 3: Experimental results of self-refinement with different confidence estimation feedback on *English*, *Chinese*, and *French* set on **`GPT-3.5 Turbo`**.

d. Confidence scores as self-refinement feedback can consistently improve performance across various languages. As shown in Figure 4, we apply random sampling feedback based on confidence scores in all investigated languages. The results underscore the efficacy of employing confidence scores as feedback for self-refinement to improve performance across all languages using two LLMs. Notably, languages with higher-quality confidence scores such as *Chinese* and *Japanese*, exhibit substantial improvements in 12.50% and 11.25% absolute accuracy gains.

5 Cross-lingual Confidence Estimation

In this section, we propose a cross-lingual confidence estimation aimed at enhancing previously acquired monolingual confidence scores across diverse languages. To this end, we set monolingual English confidence estimation as baselines to further facilitate systematic exploration of potential improvements using cross-lingual methods. For candidate languages to exhibit cross-lingual estimation, we categorize candidate languages into several groups based on their language family affiliations with respect to *English*.

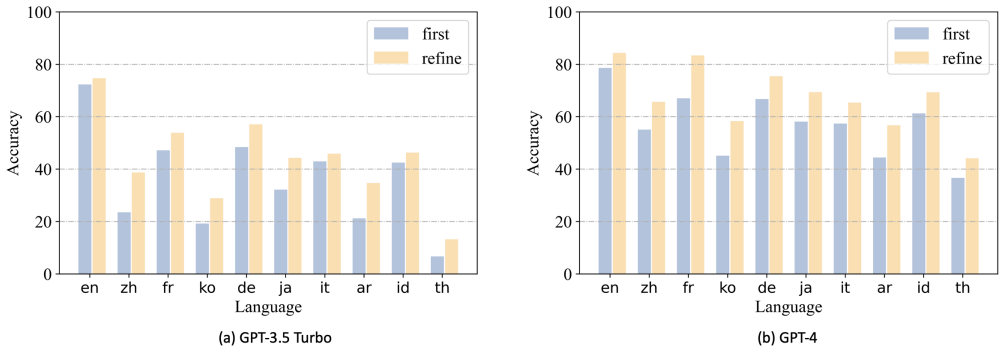

Figure 4: Accuracy ↑ of responses generated by **(a) GPT-3.5 Turbo** and **(b) GPT-4** before and after self-refinement with confidence scores as feedback on Mling-TriviaQA dataset.

5.1 Cross-lingual Confidence Estimation by Language Family

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 [5]. Some well-known language families include Indo-European and Afro-Asiatic. Linguists posit that languages within a shared language family exhibit congruent patterns of cultural cognition and thought processes (Bouckaert et al., 2012; Calude and Pagel, 2011). Motivated by the premise, we develop three distinct categorizations based on language family affiliations to determine the optimal strategy for cross-lingual confidence estimations. First, we take **shared language family** in which the languages used for cross-lingual estimation are all Indo-European languages like *French* and *German*. Then, we employ **distinct language family** comprising languages drawn from families distinct from *English* exemplified by *Chinese* and *Thai*. Finally, the **mixed language family** encompasses languages stemming from both shared and distinct language families.

Cross-lingual learning has demonstrated efficacy in augmenting the performance of the monolingual approach (Huang et al., 2023; Zeng et al., 2023; Qin et al., 2023). Our cross-lingual confidence estimation is to obtain better confidence scores by leveraging multilingual confidence assessments in Section 4. Previous work for confidence elicitation observed that aggregating confidence scores from temperature sampling by computing the average score can more effectively improve confidence estimation (Anonymous, 2024). In this work, we aggregate the monolingual confidence scores from the candidate language set $\mathcal{L}$ (e.g. $\mathcal{L} = \{\texttt{en}, \texttt{fr}, \texttt{de}\}$) and calculate their average to derive cross-lingual confidence scores as follows.

$$\texttt{Conf} = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} \texttt{Conf}(\boldsymbol{x}^{(l)}, \boldsymbol{y}^{(l)}) \quad (5)$$

where $\boldsymbol{x}^{(l)}$ and $\boldsymbol{y}^{(l)}$ denote the input and output in language l respectively.

5.2 Experimental Results and Analysis

The cross-lingual confidence estimation experiments conducted on the Mling-TriviaQA dataset are presented in this part. The experimental outcomes are structured to address several pertinent sub-questions: a. *How do different language families affect the cross-lingual confidence estimation performance?Does integrating a larger number of languages in cross-lingual confidence estimation lead to better overall performance?* b. *Does cross-lingual confidence estimation exhibit superiority over monolingual estimation for self-refinement purposes?* c. *Does cross-lingual confidence estimation surpass traditional prompting and temperature sampling methodologies?*

The following key findings to the above specific sub-questions are delineated below:

a. Shared language family can significantly improve the cross-lingual confidence estimation while more distinct family languages can not bring more improvements. We explore the relationship between performance and the number of languages integrated based on the language family as in Table 4. We incrementally integrated languages in both shared, distinct, and mixed language family of their respective proportions. The results demonstrate that in shared language family settings, performance improves as more languages are added. These findings emphasize that increasing

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

Family	Language Set	AUROC ↑
GPT-3.5 Turbo		
-	en	87.14
Shared	en de	89.26
	en fr de	91.75
	en fr de it id	**93.43**
Distinct	en zh	88.42
	en zh ko ja	89.27
	en zh ko th ar ja	88.32
Mixed	en zh fr	88.87
	en zh ja fr de	90.05
	all	92.43
GPT-4		
-	en	89.97
Shared	en de	91.43
	en fr de	94.56
	en fr de it id	**95.78**
Distinct	en zh	90.51
	en zh ko ja	92.12
	en zh ko th ar ja	93.45
Mixed	en zh fr	90.68
	en zh ja fr de	92.13
	all	93.33

Table 4: Experimental results of cross-lingual confidence estimation methods in various language family categorization and selected languages on **GPT-3.5 Turbo** and **GPT-4**.

the number of languages in the shared family can effectively improve confidence estimation. We adopt encompassing all languages in the shared family as the optimal cross-lingual confidence estimation setting for the following experiments.

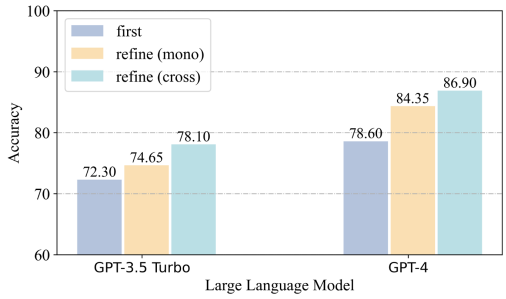

Figure 5: Accuracy of self-refinement using monolingual and cross-lingual confidence scores on Mling-TriviaQA English subset on **GPT-3.5 Turbo** and **GPT-4**.

b. Cross-lingual confidence estimation as self-refinement feedback performs better to improve accuracy. To scrutinize the acquired cross-lingual confidence scores, we perform experiments wherein these scores are introduced as feedback for self-refinement of LLMs in Figure 5, akin to the methodology outlined in Section 4.2. The findings indicate that incorporating cross-lingual confidence scores as feedback yields notable enhancements after self-refinement, surpassing the performance of

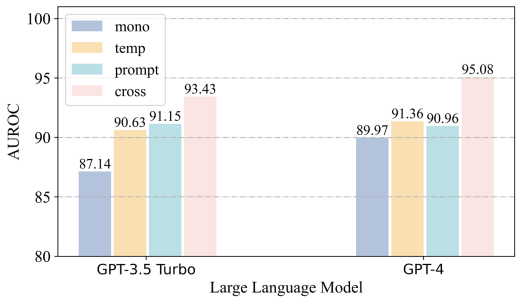

Figure 6: Experimental results of AUROC on Mling-TriviaQA *English* subset using monolingual (mono), temperature sampling (temp), prompt strategy (prompt), and cross-lingual (cross) confidence estimations on **GPT-3.5 Turbo** and **GPT-4**.

monolingual (*English*) confidence scores on both **GPT-3.5 Turbo** and **GPT-4**, suggesting that higher-quality cross-lingual confidence scores are more effective in aiding LLMs to rectify their outputs.

c. Cross-lingual confidence estimation outperforms other baseline confidence estimation methods. In Figure 6, we further conduct experiments by comparing our proposed cross-lingual confidence with prompting and temperature sampling that involve multiple outputs. The experimental results suggest that all confidence estimation methods with multiple outputs outperform monolingual estimations. The proposed cross-lingual method surpasses both temperature sampling and prompting methods on **GPT-3.5 Turbo** and **GPT-4**.

6 Conclusion

This paper presents a comprehensive investigation of multilingual confidence estimation MLING-CONF on LLMs, encompassing the construction of the self-curated Mling-TriviaQA dataset, experiments with various confidence estimation methods in multiple languages, and the implementation of cross-lingual confidence estimation strategies for more accurate confidence scores. The experimental findings elucidate the performance of confidence estimation across diverse languages as well as demonstrate the improvements using our proposed cross-lingual confidence estimation method. By addressing the limitations of existing confidence estimation research confined to English, this study makes a significant contribution towards the advancement of confidence estimation techniques to generalize across diverse linguistic contexts, thereby enhancing the reliability of the global AI system.

Limitations and Future Work

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

1. The present study is constrained by the substantial cost associated with the API cost using **GPT-4** and **DeepL** for translation as well as human verification checks by linguistics. In order to ensure the data quality, our exploration into multilingual research is restricted to ten languages in the first version of the benchmark. 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.

2. Due to page limitation, the cross-lingual estimation simply averages the confidence scores from different languages. More aggregation methods to ensemble multilingual confidence scores remain under-explored. It is promising to investigate other cross-lingual confidence estimation methods for further improvements.

Ethics Statement

In this paper, we introduce a self-constructed multilingual QA dataset derived from the publicly available TriviaQA dataset named Mling-TriviaQA. 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

Anonymous. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*.

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.

Remco Bouckaert, Philippe Lemey, Michael Dunn, Simon J Greenhill, Alexander V Alekseyenko, Alexei J Drummond, Russell D Gray, Marc A Suchard, and Quentin D Atkinson. 2012. Mapping the origins and expansion of the indo-european language family. *Science*, 337(6097):957–960.

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.

Andreea S Calude and Mark Pagel. 2011. How do we use language? shared patterns in the frequency of word use across 17 world languages. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1567):1101–1107.

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*.

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.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2023. Shifting attention to relevance: Towards the uncertainty estimation of large language models.

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.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Mengting Hu, Zhen Zhang, Shiwan Zhao, Minlie Huang, and Bingzhe Wu. 2023. Uncertainty in natural language processing: Sources, quantification, and applications.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting.

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).

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

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*.

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*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

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

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.

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, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709, Singapore. 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.

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*.

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*.

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.

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.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023c. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

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.

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. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms.

Qingcheng Zeng, Lucas Garay, Peilin Zhou, Dading Chong, Yining Hua, Jiageng Wu, Yikang Pan, Han Zhou, Rob Voigt, and Jie Yang. 2023. Greenplm: Cross-lingual transfer of monolingual pre-trained language models at almost no cost. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6290–6298. International Joint Conferences on Artificial Intelligence Organization. AI for Good.

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.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: Expressions of overconfidence and uncertainty in language models. *arXiv preprint arXiv:2302.13439*.

Language	ISO 639-3	ISO 639-1	Family	Macroarea
English	eng	en	Indo-European	Eurasia
Italian	ind	it	Indo-European	Eurasia
Indonesian	ind	id	Indo-European	Eurasia
French	fra	fr	Indo-European	Eurasia
German	deu	de	Indo-European	Eurasia
Chinese	zhs	zh	Sino-Tibetan	Eurasia
Japenese	jpn	ja	Japanese-Ryukyuan	Eurasia
Arabic	arb	ar	Afro-Asiatic	Eurasia
Korean	kor	ko	Koreanic	Eurasia
Thai	tha	th	Kra–Dai	Eurasia

Table 5: List of International Standard Organization (ISO) 639-1 Codes.

A Dataset Details

A.1 TriviaQA Dataset

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.

B Language Information

The basic information including the language family, ISO code, Macroarea of the investigated languages is presented in Table 5. The languages investigated in this work are selected following two principles.

1. For each selected language, we can employ one linguistic expert for the human check to ensure the data quality.

2. The candidate languages for the cross-lingual method are grouped based on language family. As *English* belongs to Indo-European languages, we introduce the other four Indo-European languages including *Italian, Indonesian, French, German*, and five non-Indo-European languages including *Chinese, Japanese, Arabic, Korean* and *Thai*.

C Prompt Details

Multilingual prompts for answer generation and self-estimated confidence estimation are presented in Figure 7 and 8. Notably, the prompts for self-reflected true probability confidence estimation are followed by previous work (Kadavath et al., 2022; Kuhn et al., 2023).

D Confidence Estimation on PaLM 2

Our primary experiments are confined to the utilization of widely employed models GPT-3.5 and GPT-4, which have recently granted access to the log probabilities. Notably, several existing LLMs like PaLM-2 (Chowdhery et al., 2022) are black-boxed, thereby restricting end-users from accessing token-level probabilities and impeding the deployment of likelihood-based and true probability confidence estimation methodologies. Therefore, we only conduct linguistic confidence estimations on ***PaLM 540B*** as in Table

E Self-refinement Experiments

Experimental results of EM and F1 scores between gold answers and responses generated by **(a) GPT-3.5 Turbo** before and after self-refinement with confidence scores as feedback on Mling-TriviaQA dataset are presented in Figure 9. The trends are similar to the accuracy improvements in Section 4.3.

Language	Prompt
en	Please correctly answer the following questions.\n\n{few-shot demos}# Question #: {}\n# Answer #:
zh	请正确回答下列问题。\n\n{few-shot demos}# 问题 #: {}\n# 答案 #:
fr	Veuillez répondre correctement aux questions suivantes.\n\n{few-shot demos}# Question #: {}\n# Réponse #:
ja	次の質問に正しく答えてください。\n\n{few-shot demos}# 質問 #: {}\n# 答え #:
de	Bitte beantworten Sie die folgenden Fragen richtig.\n\n{few-shot demos}# Frage #: {}\n# Antwort #:
ko	다음 질문에 올바르게 답해 주십시오.\n\n{few-shot demos}# 질문 #: {}\n# 답변 #:
it	Si prega di rispondere correttamente alle seguenti domande.\n\n{few-shot demos}# Domanda #: {}\n# Risposta #:
th	กรุณาตอบคำถามต่อไปนี้ให้ถูกต้อง\n\n{few-shot demos}# คำถาม #: {}\n# คำตอบ #:
id	Jawablah pertanyaan-pertanyaan berikut dengan benar.\n\n{few-shot demos}# Pertanyaan #: {}\n# Jawaban #:
ar	الرجاء الإجابة بشكل صحيح على الأسئلة التالية.\ \n\n{few-shot demos} \n{} # سؤال #: {}\n# إجابة #:

Figure 7: Prompt templates for answering questions for each language.

Language	Prompt
en	# Predicted Answer #: {}\n# Gold Answer #: {}\n# Your Response #: (\"Yes\" or \"No\"):
zh	# 预测答案 #: {}\n# 标准答案 #: {}\n# 你的判断 #: (\"是\"或\"否\"):
fr	# Réponse prédite #: {}\n# Réponse en or #: {}\n# Votre réponse #: (\"Oui\" ou \"Non\"):
ja	# 予想される答え #: {}\n# ゴールドの答え #: {}\n# あなたの答え #: (\"はい\" または \"いいえ\"):
de	# Vorausgesagte Antwort #: {}\n# Goldene Antwort #: {}\n# Ihre Antwort #: (\"Ja\" oder \"Nein\"):
ko	# 예상 답변 #: {}\n# 골드 답변 #: {}\n# 귀하의 답변 #: (\"예\" 또는 \"아니요\"):
it	# Risposta pronosticata #: {}\n# Risposta Oro #: {}\n# La tua risposta #: (\"Sì\" o \"No\"):
th	# คำตอบที่ทำนายไว้ #: {}\n# คำตอบระดับทอง #: {}\n# คำตอบของคุณ #: (\"ใช่\" หรือ \"ไม่ใช่\"):
id	# Prediksi Jawaban #: {}\n# Jawaban Emas #: {}\n# Jawaban Anda #: (\"Ya\" atau \"Tidak\"):
ar	# الإجابة المتوقعة #: {}\n# رقم الإجابة الذهبية #: {}\n# إجابتك #: ("\"نعم\" أو \"لا\"\) :

Figure 8: Prompt templates for Self-aware confidence estimation for each language.

Languages	en	zh	fr	ko	de	ja	it	ar	id	th	**Avg**
						PaLM 540B					
Avg Length	5.18	12.44	7.32	11.05	6.93	10.86	10.34	15.22	5.84	18.05	-
Verbal Number	82.31	86.65	80.44	87.49	83.42	88.60	81.77	86.50	79.37	86.12	84.26
Verbal Word	79.36	83.60	77.44	81.26	78.45	82.19	77.50	78.36	78.45	80.69	79.73

Table 6: Experimental results of AUROC ↑ on Mling-TriviaQA set using several confidence estimation methods.

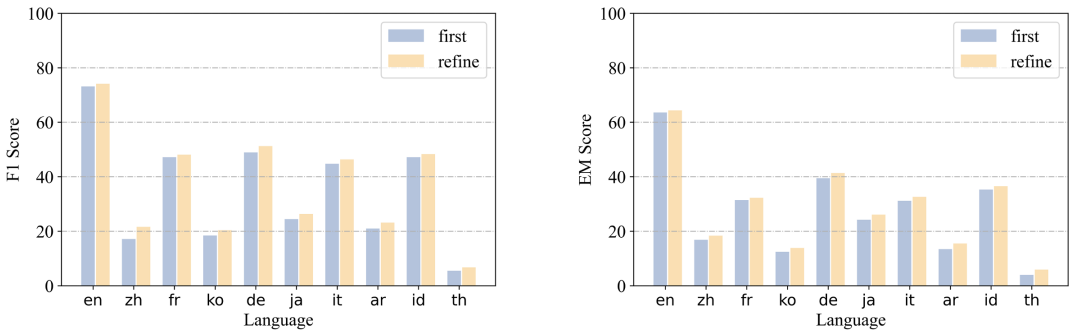

Figure 9: EM and F1 scores of different languages before and after self-refinement using confidence scores as feedback.