Evaluating Cross-lingual Consistency of Factual Knowledge in Large Language Models

Anonymous ACL submission

Abstract

Large Language Models (LLMs), exemplified by the likes of ChatGPT, have marked significant strides in the field of Natural Language Processing, earning widespread acclaim for their multitasking prowess. However, as the demand for cross-lingual applications escalates, the issue of response consistency in different linguistic contexts within LLMs becomes increasingly apparent, particularly in terms of knowledge-based queries. This study is committed to a profound evaluation of cross-lingual consistency in the knowledge embedded within LLMs. Existing research on knowledge-based cross-lingual consistency is notably scarce and suffers from conspicuous limitations. To address these shortcomings, we have constructed a factual knowledge dataset based on Wikidata, spanning five domains and twelve languages. Furthermore, we propose a novel set of metrics for evaluating cross-lingual consistency of knowledge, incorporating cross-lingual semantic consistency, cross-lingual accuracy consistency, and cross-lingual timeliness consistency. Leveraging this newly constructed dataset and evaluation metrics, we have undertaken a comprehensive evaluation and analysis of six representative open-source and closed-source models[1].

1 Introduction

In recent years, the rapid development of Large Language Models (LLMs) has led to significant advancements in natural language processing (NLP), e.g., ChatGPT[2], Llama (Touvron et al., 2023b) and Baichuan (Yang et al., 2023). These models have shown remarkable performance across various NLP tasks, including machine translation (Jiao et al., 2023), and question-answering (Bang et al., 2023).

With the increasing demand for global applications and the necessity to accommodate diverse linguistic communities, the multilingual capabilities of LLMs have gained significant importance. Unfortunately, in practical applications, LLMs often generate inconsistent responses to identical questions posed in different languages. For example, as shown in Figure 1, ChatGPT generates the responses “Paris Saint-Germain (PSG)” for the English query “Which team does Lionel Messi play for?" and “巴塞罗那足球俱乐部” (Chinese translation of “FC Barcelona”) for the Chinese query “利昂内尔·梅西效力于什么球队？” respectively.

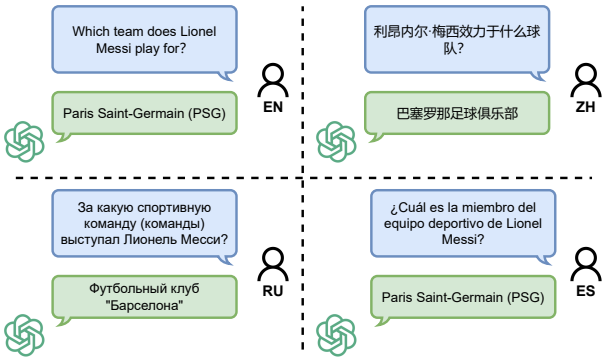

Figure 1: The ChatGPT exhibits variability in outcomes when the identical query is articulated in diverse languages.

Therefore, evaluating the cross-lingual consistency of the knowledge embedded in LLMs has become a crucial task. We need to ensure that these LLMs maintain robust, reliable and consistent performance when processing different languages. This not only helps to enhance the multilingual processing capabilities of LLMs but also has significant implications for meeting the demands of global applications.

However, current research on the cross-lingual consistency of large models is very limited. Qi et al. (2023) first proposed the concept of the cross-lingual consistency of knowledge, constructed a multilingual aligned knowledge dataset BMLAMA based on existing datasets, and proposed a consis-

[1]All code and data released at xxx

[2]https://chat.openai.com/

tency measurement method RankC based on perplexity ranking. However, this research still has some shortcomings that need to be improved: in terms of the dataset, the covered domains and relationships are monotonous, making it difficult to comprehensively measure the actual performance of the models; in terms of evaluation metrics, the model's answer is not autoregressively generated by models, creating a gap between the metric and the practical application, and making it unsuitable for evaluating closed-source models; furthermore, a single ranking metric cannot fully measure the performance of cross-lingual consistency of knowledge in the model.

In light of this, this paper constructs a Multilingual Aligned Knowledge-based Question-Answering dataset (MAKQA) based on Wikidata, which includes 12 languages across 6 domains, and proposes three innovative cross-lingual consistency of knowledge evaluation metrics: cross-lingual semantic consistency (CLSC), cross-lingual accuracy consistency (CLAC), and cross-lingual timeliness consistency (CLTC). We select six prevalent LLMs and conduct a comprehensive evaluation and analysis of them using these metrics.

Main Contributions:

- We construct a multilingual aligned knowledge-based question answering dataset (MAKQA) covering 5 domains and 12 languages, providing effective support and assistance for research on the cross-language consistency of knowledge in LLMs.
- We design a set of evaluation metrics aimed at assessing the cross-lingual consistency of knowledge in LLMs, including cross-lingual semantic consistency, cross-lingual accuracy consistency, and cross-lingual timeliness consistency.
- Through the dataset and evaluation metrics, we conduct evaluations and analyses on multiple open-source and closed-source LLMs. We find that: (i) The knowledge embedded in LLMs exhibits a significant clustering phenomenon based on language families in terms of cross-lingual consistency; (ii) The cross-lingual consistency of knowledge shows distinct language distribution rules and imbalance phenomena, and this imbalance does not get compensated with the increase in model size; (iii) The cross-lingual consistency of knowledge remains stable, unaffected by prompt variations; (iv) There is a correlation between the cross-lingual consistency of knowledge in LLMs and their multilingual translation capabilities.

Domain	#Entity	#Rel	#QA pairs
Sports	50	9	253
Movie	49	17	432
Science	49	12	492
History	45	12	389
Geography	94	6	286
Literature	50	5	165
Timeliness	129	2	136

Table 1: Satistics of the MAKQA dataset used in our analysis.

2 Dataset

Before embarking on the construction of a new dataset, we conduct an in-depth evaluation of existing factual knowledge datasets. We observe that, despite the multilingual alignment achieved by BMLAMA (Qi et al., 2023), a multilingual factual knowledge dataset, the knowledge it encompasses is predominantly concentrated in the field of geography. This bias limits its utility for comprehensively assessing the cross-lingual consistency of knowledge in LLMs. On the other hand, factual knowledge datasets that have not achieved multilingual alignment are unable to effectively measure the cross-lingual consistency of knowledge. Attempts to extend these datasets to multiple languages using automatic translation engines may introduce translation errors, thereby impacting the reliability of the results. Given these factors, we decide to develop a new multilingual aligned knowledge question-answering dataset to more accurately and comprehensively evaluate the cross-lingual knowledge consistency of LLMs.

We utilize Wikidata as the fundamental data source for establishing our dataset. We collect entity names in English from diverse sources and subsequently, through Wikipedia, we acquire knowledge triplets associated with these entities. From these triplets, we selectively retained those knowledge triples that contained key relations. In addition, we capitalized on the feature that every entity in Wikipedia is logged with its multilingual names, thereby expanding English knowledge triples to multilingual aligned knowledge triples. Notably,

we only employed translation engines as supplements for specific language names missing from some entities in Wikipedia when necessary. Finally, we transformed knowledge triples into knowledge question-answer pairs using GPT-4 (OpenAI et al., 2023), yielding our Knowledge QA dataset.

Using this methodology, we construct a Multilingual Aligned Knowledge-based Question-Answering dataset (MAKQA) that encompasses twelve languages: English (En), German (De), Dutch (Nl), French (Fr), Spanish (Es), Italian (It), Portuguese (Pt), Greek (El), Russian (Ru), Chinese (Zh), Japanese (Ja), and Korean (Ko). Concurrently, the dataset covers knowledge from six fields: sports, movie, science, history, geography, and literature, as detailed in Table 1.

To fulfill the need for evaluating the cross-lingual timeliness consistency in LLMs, we construct a timeliness dataset. To ensure the reliability of the dataset and respect for privacy, we choose to use the clubs and leagues that well-known athletes participate in as the background for the questions. All the information used is publicly available and can be found on Wikipedia. The methodology for data construction as previously described is employed in the creation of this dataset. Within the dataset, the answers are systematically arranged in chronological order, reflecting the sequence of the events.

This dataset serves not only for evaluating the cross-lingual consistency of LLM in the domain of knowledge but also aids in delving deeply into the disparities in common knowledge and question-answering abilities of LLM under different language environments, and their strengths and weaknesses. We will release the dataset in the hope of fostering research in related fields.

3 Experiments

To evaluate the performance of current state-of-the-art LLMs, we selected five highly acclaimed LLMs and examined their variants of different scales. Specifically, we chose the closed-source model GPT-3.5 (Ouyang et al., 2022), as well as the open-source models Bloomz (Muennighoff et al., 2022) and Llama2 (Touvron et al., 2023a) (which claim to support multiple languages), Baichuan2 (Baichuan, 2023) and Mistral (Jiang et al., 2023, 2024) (which claim to only support a few high-resource languages). To assess the impact of models of different scales on cross-lingual consistency, we measured variants of each open-source model with parameter sizes less than 70b. We utilized the LLAMA-factory (Hiyouga, 2023) to develop an API that faithfully reproduces the models' performance in real-world usage scenarios.

Model	CLSC	CLAC	CLTC
GPT-3.5	**0.7712**	**0.4555**	**0.4798**
Bloomz-560m	0.6217	0.2031	0.0655
Bloomz-1b	0.6267	0.2669	0.1015
Bloomz-3b	0.6339	0.2830	0.1196
Bloomz-7b	0.6229	0.3110	0.1433
Llama2-7b	0.6891	0.2172	0.2236
Llama2-13b	0.6796	0.3179	0.2072
Baichuan2-7b	0.695	0.3360	0.2115
Baichuan2-13b	0.7154	0.3404	0.2426
Mistral-7b	0.6676	0.2683	0.2381
Mixtral-8x7B	0.7655	0.4059	0.297

Table 2: The main result of assessing the cross-lingual consistency of knowledge in LLMs.

We used the 5-shot in-context learning strategy to guide the models in providing responses, in order to mitigate the impact of different models' instruction-following abilities on the answers. Specifically, in each domain, we meticulously select 20 cases from the dataset to serve as examples. During each inference process, we would choose 5 of these examples to serve as cases for in-context learning. The prompts used during the inference process are provided in the appendix. All our experiments were conducted on four A100-PCIE-40GB GPUs.

4 Evaluation

In order to comprehensively evaluate the cross-lingual consistency of the model, we introduce three progressively hierarchical metrics, namely cross-lingual semantic consistency (CLSC), cross-lingual accuracy consistency (CLAC), and cross-lingual timeliness consistency (CLTC). These metrics impose higher requirements on the cross-lingual consistency of the model. In this section, we will provide a detailed description of each metric and compare the performance of different models on these three metrics.

4.1 Cross-Lingual Semantic Consistency

The Cross-Lingual Semantic Consistency (CLSC) aims to measure the consistency of knowledge across different languages in LLMs. In other words, we intend to assess whether a model provides con-

sistent answers when faced with the same questions in different languages, to determine the consistency of knowledge stored in the model across different languages.

4.1.1 Method

To evaluate the semantic consistency of model responses to identical questions posed in various languages, we employ LASER (Heffernan et al., 2022), a multilingual semantic encoding model, to encode the responses generated by the model in different languages. We systematically examine all possible language pair combinations, computing the cosine similarity of the semantic vectors for each pair. Subsequently, we derive an average of these similarities, which provides us with a cross-lingual semantic consistency score for the model. This computation process is detailed in Formula 1.

$$\text{CLSC} = \frac{1}{L(L-1)} \sum_{i=1}^{L} \sum_{\substack{j=1 \\ j \neq i}}^{L} \text{consist}_{i,j}$$

$$\text{consist}_{i,j} = \frac{1}{N} \sum_{s=1}^{N} \text{cos\_similarity}(V(\text{ans}_s^i), V(\text{ans}_s^i)) \quad (1)$$

In Formula 1, ans_s^i denotes the answer provided by the model for the s-th question in language i. L and N respectively denote the number of languages and the total number of question-answer pairs in the dataset. $V(.)$ signifies the vector representation post LASER encoding, and $cos\_sim(.)$ represents the computation of cosine similarity.

4.1.2 Result

The CLSC scores for each model are presented in the first column of Table 2. Firstly, we observe variations among different models. The closed-source model GPT-3.5 performs the best in CLSC, with a score of 0.7712, surpassing all open-source models. Among the open-source models, Mixtral-8x7b performs the best with a score of 0.7655, significantly outperforming other open-source models. Despite Mixtral claiming to only support a limited number of high-resource languages, it exhibits better performance in CLSC.

Secondly, we observe a significant improvement in the performance of Mixtral as the number of model parameters increases. However, it is noteworthy that Mixtral modifies the model structure compared to Mistral by incorporating the MOE (Mixture of Experts) structure (Fedus et al., 2022) in the FeedForward blocks. In Baichuan2 models, we note a minor increase in CLSC scores as the model size grows. Yet, in the Bloomz and Llama2 models, we do not observe the impact of model size on CLSC. Therefore, we infer that merely increasing the size of the model may not effectively enhance the CLSC score.

To enhance our understanding of the distribution of semantic consistency across various language pairs, we conduct a detailed analysis and visualize the results. These heatmaps represent the semantic similarity scores between all language pairs for these four open-source models of 7b size, depicted in Figure 2. The analysis illuminates a notable pattern: the CLSC scores between languages are profoundly influenced by their linguistic families. Specifically, languages within the Germanic language family (English [En], German [De], Dutch [Nl]) and the Indo-European-Romance language family (French [Fr], Spanish [Es], Italian [It], Portuguese [Pt]) demonstrate a pronounced level of semantic consistency amongst themselves. In contrast, their semantic alignment with languages outside these families is markedly lower, thereby illustrating a clustering trend. We further employ hierarchical clustering based on CLSC scores to group languages, and obtain the same conclusion, with the experimental results provided in the appendix.

Finally, we independently compute the scores for five representative models across six domains, as depicted in Table 3. The findings reveal that the CLSC scores of these models fluctuate noticeably across the varied domains. Nevertheless, in a general sense, GPT-3.5 surpasses other models in all evaluated domains. Among the open-source models, Baichun2-7b exhibits superior performance in four out of the six domains, while Bloomz-7b consistently underperforms in all domains. These observations suggest that although variations in knowledge across diverse domains can impact the CLSC in LLMs, they do not act as a definitive determinant.

4.2 Cross-Lingual Accuracy Consistency

This section aims to evaluate the consistency of the accuracy of the model's responses across different languages. Accuracy serves as the most critical and straightforward metric for evaluating the model's performance in diverse languages, given that it mirrors the model's effectiveness in downstream tasks. Moreover, the consistency of accuracy across nu-

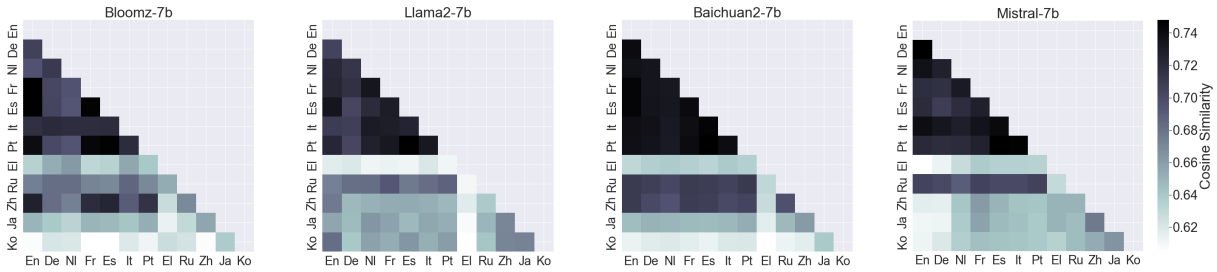

Figure 2: Distribution of average cosine similarity across languages.

Model	Sports	Movie	Science	History	Geography	Literature
GPT-3.5	0.8029	0.738	0.7511	0.7607	0.8241	0.788
Bloomz-7b	0.6326	0.5814	0.6133	0.6411	0.7006	0.6165
Llama2-7b	0.6944	0.6894	**0.7373**	0.6328	0.7053	0.6368
Baichuan2-7b	**0.7241**	0.6466	0.6936	**0.6976**	**0.7451**	**0.692**
Mistral-7b	0.6593	**0.6946**	0.7061	0.6454	0.6607	0.6283

Table 3: CLSC domain result

merous languages necessitates superior standards for the model's cross-lingual consistency of knowledge. This implies that the knowledge embedded in different languages should not only be identical but also accurate. Consequently, we introduce the Cross-Lingual Accuracy Consistency metric (CLAC).

4.2.1 Method

We commence by establishing a metric for accuracy, computed using theFuzz[3] method to determine the partial ratio between the answer and the groundtruth. An answer is deemed correct if the ratio meets or exceeds 75%, thereby receiving a label of 1; otherwise, it is assigned a label of 0. This metric facilitates an evaluation of the model's answer accuracy across diverse languages, and it mitigates the risk of erroneous judgments engendered by exact matching. Subsequently, we employ the Spearman correlation coefficient to ascertain the correlation of accuracy results between every pair of languages. The average value across all language pairs is utilized as an indicator of cross-lingual accuracy consistency.

Moreover, we must also take into account the potential for multiple answer entities within the responses. To manage this scenario, we initially partition the answers and subsequently match each entity with the potential answers. Ultimately, we compute the mean of the scores for all entities predicted by the model to derive the accuracy score for the given question.

[3]https://github.com/seatgeek/thefuzz

4.2.2 Result

Upon examining the experimental results delineated in the second column of Table 2, we observe that different models exhibit similar trends in terms of accuracy consistency and semantic consistency. Specifically, the closed-model GPT-3.5 outperforms all other Language Learning Models (LLMs), and Mixtral demonstrates the best performance among the open-source models.

In contrast to semantic consistency, our findings suggest that accuracy consistency experiences a marked augmentation with the escalation in model size. This trend is particularly conspicuous in the Bloomz series models. We infer that such improvements may be attributable to the enhanced capabilities of the model as a result of the expansion in model parameters, thereby increasing the overall accuracy of the model. Ultimately, it leads to a significant improvement in its cross-lingual consistency in accuracy scores.

To evaluate the preferences of LLMs for CLAC across different languages, we plot the average CLAC scores of each language in relation to other languages (as illustrated in Figure 3).Our investigation reveals that the GPT-3.5 model exhibits a commendable level of consistency in performance across different language pairs, with a relatively uniform distribution of accuracy consistency among various languages. Notably, while Greek displays the lowest average correlation, it nonetheless achieves a correlation coefficient of approximately 0.4. In stark contrast, the open-source

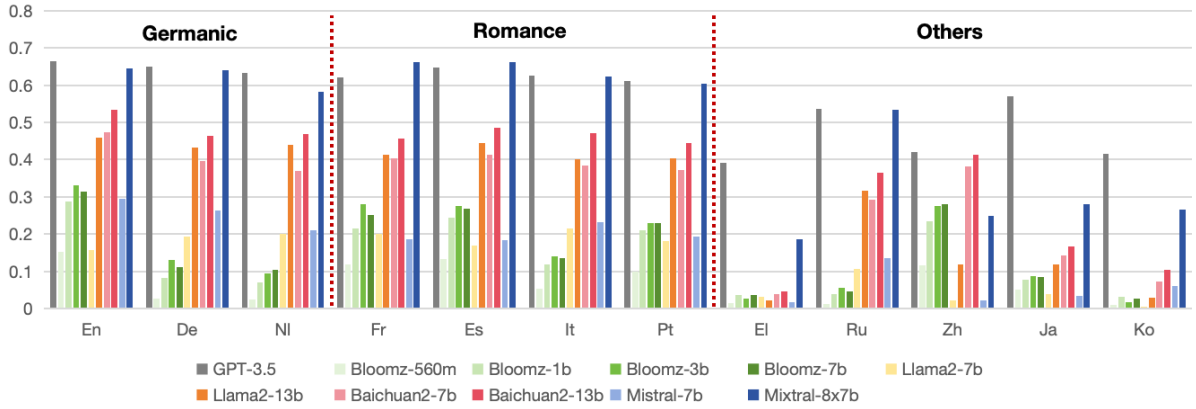

Figure 3: Average cross-lingual accuracy consistency scores of LLMs in different languages.

LLMs under examination, except Mixtral, demonstrate a pronounced disparity in the distribution of accuracy consistency among different languages, with Greek and Korean, for example, registering an average correlation coefficient of less than 0.1. Furthermore, from the figure, we see that across all evaluated models, there is a significantly higher average correlation coefficient with languages belonging to the Germanic and Indo-Romance families as opposed to languages from other families. This observation suggests that the CLAC exhibits a correlation with linguistic families, predominantly manifesting within high-resource language families, more specifically, within the European languages. Lastly, our study also uncovers that while augmenting the size of the model may yield marginal improvements in cross-lingual accuracy consistency, it falls short of addressing the stark imbalances in consistency distribution observed across languages.

4.3 Cross-Lingual Timeliness Consistency

The primary aim of this section is to assess the disparities in the timeliness of responses across various languages. As illustrated in Figure 1, the act of posing time-sensitive queries in distinct languages frequently results in receiving answers with varying degrees of timeliness. To precisely quantify the differences in response timeliness among different languages, we develop a novel metric termed Cross-Lingual Timeliness Consistency (CLTC).

4.3.1 Method

We adopt a similar approach to CLAC to assess the answers generated by the model. We utilize a fuzzy matching technique predicated on the partial ratio to ascertain the correspondence between the entities in the model's responses and those in the pre-established ground truth answer list. the models are scored based on the inverse of the rank assigned to the corresponding entity within the answer list. We calculate the Spearman correlation coefficient across the scores obtained for various language pairs and compute their average to obtain the CLTC score of the model.

4.3.2 Result

The third column in Table 2 presents the CLTC scores of all models. It is evident that GPT-3.5 achieves a score of 0.4798, markedly surpassing the performance of other models. This discrepancy in performance becomes increasingly pronounced as the evaluation criteria shift from CLSC to CLTC, highlighting GPT-3.5's superior capability in cross-lingual tasks. Additionally, for the Bloomz and Baichuan2 models, the CLTC scores exhibit an increasing trend with the increase in model size.

We compute and plot the average correlation coefficient between each language and all other languages, as illustrated in Figure 4. This figure reveals a parallel trend between CLTC and CLAC metrics: (i) All models demonstrate superior cross-lingual consistency between the languages of the Germanic and Indo-Romance families, as compared to other languages; (ii) An increase in model size does not effectively address the issue of imbalanced distribution of CLTC.

5 Discussion

In Section 4, we conduct a comprehensive evaluation of LLMs with a focus on the cross-lingual consistency of knowledge across three distinct dimensions. Next, centering on Cross-Lingual Semantic Consistency (CLSC), we investigate factors that may affect consistency performance: prompt

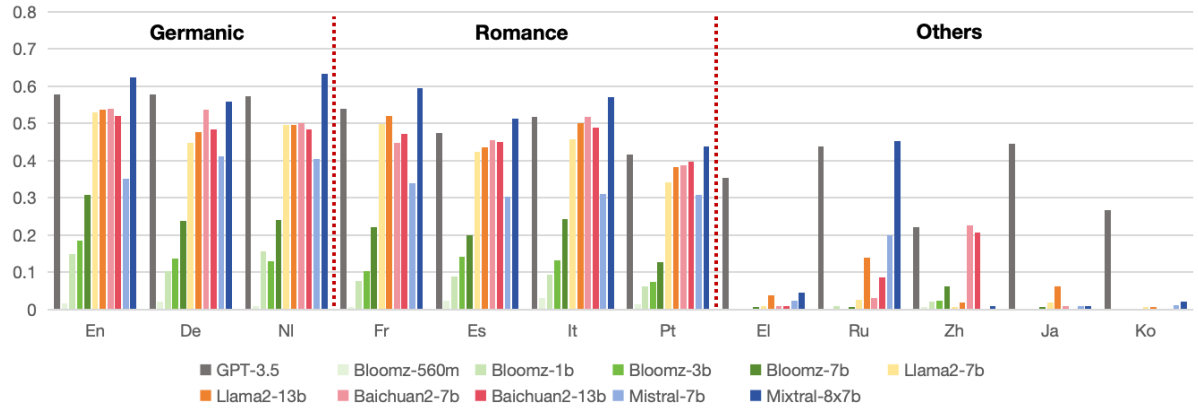

Figure 4: Average cross-lingual timeliness consistency scores of LLMs in different languages.

Model	Prompt1	Prompt2	Prompt3
Bloomz-7b	0.6229	0.6208	0.626
Llama2-7b	0.6891	0.6748	0.6739
Baichuan2-7b	0.695	0.6999	0.6853
Mistral-7b	0.6738	0.6713	0.6676

Table 4: CLSC scores of LLMs using different prompts.

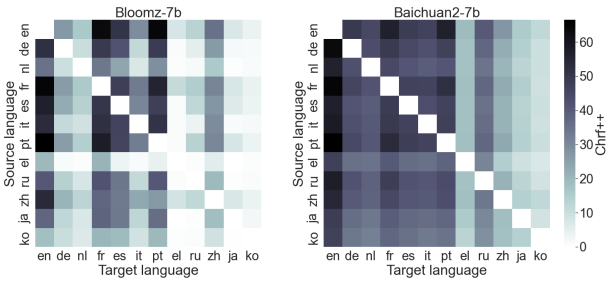

Figure 5: Distribution of Chrf++ scores for translations across languages.

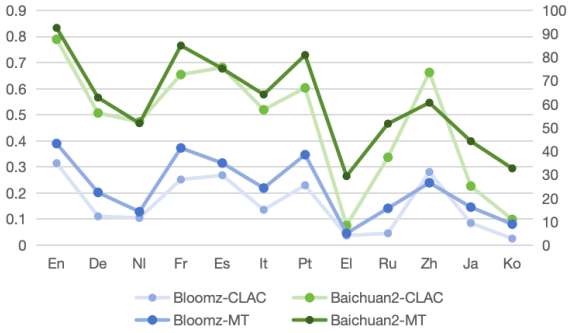

Figure 6: Average cross-lingual accuracy consistency scores and average translation scores for LLMs in different languages.

and multilingual translation.

5.1 Is cross-language consistency prompt-sensitive?

Firstly, we evaluate the robustness of CLSC in LLMs by scrutinizing the impact of varying prompts. To accomplish this, we employ not only the original questions (hereafter referred to as Prompt1) but also devise two distinct sets of new questions, denoted as Prompt2 and Prompt3. Prompt2 follows a standardized question template, incorporating relations and head entities to generate questions. Prompt3 is derived by rephrasing the original questions using GPT-4. By comparing the model's performance on these three types of questions, we can effectively evaluate the extent of variation in cross-lingual consistency of knowledge under disparate prompts. We tabulate the experimental results in Table 4.

We observe that the models display minor variations in performance when subjected to different prompts. Specifically, the Bloomz-7b model registers performance scores of 0.6229, 0.6208, and 0.626 under disparate prompts, respectively. Nevertheless, it is imperative to highlight that despite these minor discrepancies, the overall shift in performance is not statistically significant. This indicates that the assessment of cross-lingual consistency in LLMs is largely impervious to the choice of prompts. These results infer that large language models exhibit a commendable degree of robustness and reliability in CLSC.

5.2 Is cross-language consistency relevant to translation?

Secondly, our research aims to investigate the correlation between CLSC and the multilingual translation capabilities of LLMs. To achieve this, we select 12 languages from the Flores-200 devtest dataset (NLLB Team, 2022), forming a test

set that encompasses a total of 132 translation directions. We select two models, Bloomz-7b and Baichuan2-7b, and evaluate their performance across all translation directions utilizing the Chrf++ metric (Popovic, 2017). Figure 5 delineates the performance distribution of these models.

From the figure, it can be observed that the distribution of LLMs' multilingual translation ability follows a similar pattern to the distribution of their CLSC. More precisely, the models exhibit markedly superior translation performance within the Germanic language family (En, De, Nl) and the Indo-European-Romance language family (Fr, Es, It, Pt). In contrast, their performance is relatively subpar in other translation directions. Furthermore, we have noted that for languages within the Germanic and Indo-European-Romance language families, the models' translation performance is significantly elevated when these languages are used as the target language compared to the source language. However, this particular trend is not observable for languages belonging to other language families.

We plot Figure 6 to show the correlation between the multilingual translation capabilities of LLMs and their CLAC. In the figure, the darker points within each color represent the average translation performance of the model across all translation directions that include the respective language. The lighter points indicate the model's average CLAC score in that language relative to other languages.

Based on the figure, it can be inferred that a discernible positive correlation between the multilingual translation capabilities of LLMs and CLAC can be observed. This correlation is not merely confined to different models, but it also persists within the same model across a variety of languages.

6 Conclusion

Our research focuses on the evaluation and analysis of the cross-lingual consistency of knowledge in LLMs:

- We construct a Multilingual Aligned Knowledge-based Question-Answering dataset (MAKQA), which covers 12 languages and 5 domains. With this dataset, we comprehensively evaluate the cross-lingual consistency of knowledge in LLMs.
- We develop an evaluation metric system grounded in three key aspects: semantic consistency, accuracy consistency, and timeliness consistency. Utilizing this metric system, we carry out evaluations on a range of widely-used LLMs.
- Through our analysis of LLMs, we have unearthed several intriguing phenomena. Firstly, we observed clear language distribution patterns and imbalances in the cross-lingual consistency of knowledge in LLMs. Notably, the imbalances are not mitigated by simply increasing the model size. Secondly, the cross-lingual consistency of LLMs remains relatively stable despite changes in prompts. Lastly, our research reveals a discernible positive correlation between the multilingual translation capabilities of LLMs and their CLAC.

7 Limitations

In this paper, we conduct experiments on 12 languages and 5 LLMs to evaluate the cross-lingual consistency of knowledge in LLMs. It is, however, crucial to acknowledge that the implications drawn from our study may not be universally applicable to all LLMs. Therefore, to ensure the validity and generalizability of our findings, further research needs to be conducted on a wider range of languages and models.

Furthermore, it is noted that this paper is exclusively dedicated to the evaluation and analysis of the cross-lingual consistency of knowledge. Our future research will primarily focus on exploring how to improve the cross-lingual consistency in LLMs at a lower cost. This will help to address inconsistency issues that currently exist between different languages in LLMs and provide a more reliable foundation for practical applications.

References

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

William Fedus, Jeff Dean, and Barret Zoph. 2022. A review of sparse expert models in deep learning.

Kevin Heffernan, Onur elebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages.

Hiyouga. 2023. Llama factory. https://github.com/hiyouga/LLaMA-Factory.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L é lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth é e Lacroix, and William El Sayed. 2023. Mistral 7b.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L é lio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Th é ophile Gervet, Thibaut Lavril, Thomas Wang, Timoth é e Lacroix, and William El Sayed. 2024. Mixtral of experts.

Wenxiang Jiao, Wenxuan Wang, JT Huang, Xing Wang, and ZP Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

James Cross Onur elebi Maha Elbayad Kenneth Heafield Kevin Heffernan Elahe Kalbassi Janice Lam Daniel Licht Jean Maillard Anna Sun Skyler Wang Guillaume Wenzek Al Youngblood Bapi Akula Loic Barrault Gabriel Mejia Gonzalez Prangthip Hansanti John Hoffman Semarley Jarrett Kaushik Ram Sadagopan Dirk Rowe Shannon Spruit Chau Tran Pierre Andrews Necip Fazil Ayan Shruti Bhosale Sergey Edunov Angela Fan Cynthia Gao Vedanuj Goswami Francisco Guzm á n Philipp Koehn Alexandre Mourachko Christophe Ropers Safiyyah Saleem Holger Schwenk Jeff Wang NLLB Team, Marta R. Costa-juss à . 2022. No language left behind: Scaling human-centered machine translation.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Sim ó n Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, ukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, ukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David M é ly, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cer ó n Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang,

Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Maja Popovic. 2017. chrf++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 612–618. Association for Computational Linguistics.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore. Association for Computational Linguistics.

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, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023a. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

A Appendix

A.1 CLSC languages' results

A.2 System prompt

In this section, we will present the system prompt used in the evaluation process of LLMs :

> You are a helpful assistant. Please respond to user questions about factual knowledge, following four rules:
> 1. Provide direct answers without explaining or repeating the question.
> 2. Ensure your answers are as concise as possible.
> 3. If the answer involves multiple entities, separate them with ", ".
> 4. Use the same language as the user.
> 5. If you don't know or can't answer the question, strictly respond with "I don't know"; do not provide any other response.

A.3 CLSC Experiments

We adopt the hierarchical clustering method to divide all languages into four clusters based on their average cross-lingual consistency scores. The clustering results are shown in Table 5. The clustering results align with our observations: there is a clear clustering phenomenon in the cross-lingual consistency of the models. For the Germanic language family (En, De, Nl) and the Indo-European-Romance language family (Fr, Es, It, Pt), the semantic consistency between them is very high, while the consistency with other languages is relatively low.

Model	Cluster1	Cluster2	Cluster3	Cluster4
Bloomz-7b	'En', 'De', 'Nl', 'Fr', 'Es', 'It', 'Pt', 'Ru', 'Zh'	'El'	'Ja'	'Ko'
Llama2-7b	'En', 'De', 'Nl', 'Fr', 'Es', 'It', 'Pt', 'Ru'	'El'	'Zh'	'Ja', 'Ko'
Baichuan2-7b	'En', 'De', 'Nl', 'Fr', 'Es', 'It', 'Pt', 'Ru', 'Zh'	'El'	'Ja'	'Ko'
Mistral-7b	'En', 'De', 'Nl', 'Fr', 'Es', 'It', 'Pt', 'Ru'	'El'	'Zh', 'Ja'	'Ko'

Table 5: CLSC languages cluster