Facts Do Care About Your Language: Assessing Answer Quality of Multilingual LLMs

Shmuel Berman*, **Yuval Kansal***, **Lydia Liu**
Princeton University, USA
{shmuel, yuvalkansal, ltliu}@princeton.edu

Abstract

Factuality is a necessary precursor to useful educational tools. As adoption of Large Language Models (LLMs) in education continues of grow, ensuring correctness in all settings is paramount. Despite their strong English capabilities, LLM performance in other languages is largely untested. In this work, we evaluate the correctness of the Llama3.1 family of models in answering factual questions appropriate for middle and high school students. We demonstrate that LLMs not only provide extraneous and less truthful information, but also exacerbate existing biases against rare languages.

Keywords: AI in Education, LLMs, Question Answering, Factuality

Introduction

LLMs have shown exciting human-like capabilities in tasks such as text generation and question answering. Tools such as ChatGPT and LLama are being used in educational learning contexts and are employed directly by students without expert guidance. As they become ubiquitous learning tools, it is vital to ensure they reduce rather than reinforce existing inequities. Equitable access to education is essential for the betterment of society. The inconsistencies in LLMs for providing information to simple questions across multiple languages are concerning. In this study, we assess the factuality and consistency of LLM responses to middle and high school level questions in non-English languages.

About 55% of the content on the internet is in English[11]. All other languages make up a fraction of the content, with some accounting for only 0.1-0.2% of the internet websites. It has been shown that a LLM's effectiveness is heavily dependent on the distribution and quality of training data[12]. LLMs trained on poorly distributed data can undergo model collapse, gradually degenerate and produce increasingly homogeneous outputs[4].

Recent studies have shown that "dominant" languages exhibit superior performance due to their substantial representation in training data, while low-resource languages demonstrate significantly degraded capabilities[12, 7]. Insufficient training data in a language can result in the model losing valuable information about that language[4, 1]. While prior work has evaluated language fluency and translation quality of LLMs [12, 7], our aim is to measure their factuality in different languages, especially in educational settings.

We present three initial contributions in evaluating LLM factuality: (1) A benchmark of 54 factual questions and corresponding keyword-level answers spanning various subjects (2) A manual evaluation (IRB approved) of translated answer quality conducted by bilingual individuals (3) A statistical evaluation of several LLMs' responses across languages which shows poorer performance across several metrics in non-English languages.

---

* These authors contributed equally. The order was determined by a coin toss.

Related Work

Prior research has shown that LLMs tend to perform better in English and French compared to Bengali, Georgian, Nepali, and Maithili[2]. The paper highlights safety alignment issues, the discrepancy in multilingual research, and proposes a new safety multilingual benchmark."Factcheck-Bench" introduces a similar English-only dataset to evaluate LLM factual accuracy and proposes a pipeline to get better results[10] These datasets limit their evaluation to factuality and do not take into account extraneous information.

Researchers at Georgia Tech found that LLMs gave richer and higher-quality responses in English than in other languages[5]. Wang et al. showed that ChatGPT performs much better at translating English to German than to Chinese[9]. Li et al. showed that there is a direct relationship between a specific LLM's performance in a language and the percentage of data belonging that language in the training corpus[6]. Zhu et al. showcased inconsistency across non-English languages and successfully boosted LLM performance in instruction following and translation through multilingual instruction tuning [13]. These studies quantify translation quality and contribute valuable benchmarks but do not discuss the factual information content.

Methodology

To assess LLM factuality, we compiled a list of 54 questions across different subjects covering middle and high school curriculum. Each question was designed to be answerable in a single sentence and is accompanied by manually curated target words essential for a correct response. For example, for the question "What is photosynthesis?", target words included "sunlight","carbon dioxide", and "glucose".

These questions were presented to Llama 3.1 (8B & 70B). The evaluation was conducted in 12 languages—Japanese, Arabic, French, Chinese, Persian, Hebrew, Hindi, Nepali, Haitian Creole, Tulu, Māori, and English. All responses were machine-translated back into English using the Google Translate API for analysis. While machine translation introduces potential limitations, prior research has demonstrated that Google Translate achieves adequate accuracy for tasks requiring factual content. Taira et al. assessed Google Translate's performance and reported mean Likert scale adequacy scores exceeding 4 across languages[8]. Our study focuses on factual accuracy, not fluency. Google Translate is sufficient for this purpose given that literal translations often suffice to convey essential information.

For the manual analysis, bilingual evaluators assessed each response on a multidimensional rubric: counting the presence of each target keyword, factual accuracy (1-3 scale), and extraneous information (1-5 scale). For our automated analysis, we search for the presence of target words in both the original language and in the response translated to English, perform keyword extraction, and measure per-language response properties such as keyword repetition. We use KeyBERT[3] to identify semantically important terms and extract them as keywords, and include metrics for the number of keywords and how many times they repeat. Our question bank and code can be found here.

Results

We conducted a manual evaluation of Llama-3.1-8B and 70B responses on English, Hindi, Hebrew, and present our averaged findings in the table below:

Language	Model	# Speakers (millions)	Keyword Coverage (%)	Incorrectness [1]	Extraneous Score [2]
Hebrew	LLaMA 8B	9	27.28	1.93	3.48
	LLaMA 70B	9	69.25	1.44	2.11
Farsi	LLaMA 70B	110	87.36	1.06	1.13
Hindi	LLaMA 8B	345	51.17	1.50	2.13
	LLaMA 70B	345	77.25	1.13	1.30
English	LLaMA 8B	1500	94.44	1.04	1.35
	LLaMA 70B	1500	97.48	1.09	1.26

Table 1: *Manual Analysis Results for LLaMA 8B and LLaMA 70B. Keyword Coverage (higher is better), Incorrectness Score[1], and Extraneous Score[2] (lower is better).*

To analyze the relationship between language representation and response quality, we present the Spearman Correlation coefficients for target word detection correlating the number of speakers of the language of each response with its features. We determine there is a statistically significant correlation between the presence of our hand-picked target words. For the larger model, there is also a strong correlation between the raw number of KeyBERT keywords, which are question-agnostic. We conjecture this is because smaller models give terser responses across all languages, but leave formal analysis to future work.

Feature	Model	Correlation	p-Value
Original Target Words Count	LLaMA 8B	0.426310	1.535e-27
	LLaMA 70B	0.463306	6.1005e-33
Translated Target Words Count	LLaMA 8B	0.372825	5.841e-21
	LLaMA 70B	0.400362	2.8294e-24
KeyBERT Keywords Count	LLaMA 8B	0.074542	6.993e-02
	LLaMA 70B	0.348306	2.2022e-18
KeyBERT Keyword Repetition	LLaMA 8B	0.063026	1.256e-01
	LLaMA 70B	0.345928	3.8637e-18
Word Repetition	LLaMA 70B	0.372388	5.6222e-21
English Answer Length	LLaMA 70B	0.423705	2.8087e-27

Table 2: *Correlation Analysis Results between # of Speakers and Features for LLaMA 8B and 70B.*

Conclusion

Our experiments confirm our hypothesis that LLMs are less factual in languages that have fewer speakers and are less-represented in available training material. Future work will focus on extending this study to a broader range of LLMs and refining the analysis by utilizing professional human translators rather than machine translation. By identifying potential linguistic biases embedded within the training data of these models, we aim to enable researchers to create more equitable language models.

---

1. **Incorrectness Score**: Scale 1-3 — 1 if no incorrect information, 2 if some incorrect information, and 3 if clearly incorrect information.
2. **Extraneous Score**: Scale 1-5 — 1 if no extraneous information, 5 if most of the answer is unnecessary.

References

[1] Alham Fikri Aji Chenxi Whitehouse Monojit Choudhury. "LLM-powered Data Augmentation for Enhanced Crosslingual Performance". In: *EMNLP Proceedings* (2023).

[2] Guoliang Dong, Haoyu Wang, Jun Sun, and Xinyu Wang. *Evaluating and Mitigating Linguistic Discrimination in Large Language Models*. en. arXiv:2404.18534 [cs]. May 2024. URL: http://arxiv.org/abs/2404.18534 (visited on 10/09/2024).

[3] Maarten Grootendorst. "KeyBERT: Minimal keyword extraction with BERT". In: *arXiv preprint arXiv:2004.13226* (2020).

[4] Abdul Waheed Hao Chen et al. "On the Diversity of Synthetic Data and its Impact on Training Large Language Models". In: *arXiv preprint arXiv:2410.15226* (2024).

[5] Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. *Better to Ask in English: Cross-Lingual Evaluation of Large Language Models for Healthcare Queries*. en. arXiv:2310.13132 [cs]. Oct. 2023. URL: http://arxiv.org/abs/2310.13132 (visited on 10/09/2024).

[6] Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. *Quantifying Multilingual Performance of Large Language Models Across Languages*. en. arXiv:2404.11553 [cs]. June 2024. URL: http://arxiv.org/abs/2404.11553 (visited on 10/09/2024).

[7] Edoardo Ponti et al. "Modeling Language Variation and Universals: A Survey on Morphosyntactic Information in Language Models". In: *Computational Linguistics* (2019).

[8] Breena R. Taira, Lisa C. Diamond, Vanessa Kreger, and Aristides Orue. "A Pragmatic Assessment of Google Translate for Emergency Department Instructions". In: *Olive View-UCLA Medical Center; Memorial Sloan Kettering Cancer Center* (2021).

[9] Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. *Document-Level Machine Translation with Large Language Models*. en. arXiv:2304.02210 [cs]. Oct. 2023. URL: http://arxiv.org/abs/2304.02210 (visited on 10/09/2024).

[10] Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. *Factcheck-Bench: Fine-Grained Evaluation Benchmark for Automatic Fact-checkers*. en. arXiv:2311.09000 [cs]. Apr. 2024. URL: http://arxiv.org/abs/2311.09000 (visited on 10/09/2024).

[11] Wikipedia. *Languages used on the Internet — Wikipedia, The Free Encyclopedia*. http://en.wikipedia.org/w/index.php?title=Languages%20used%20on%20the%20Internet&oldid=1247975178. [Online; accessed 09-October-2024]. 2024.

[12] Wei Xie, Mengnan Du, et al. "Quantifying Multilingual Performance of Large Language Models Across Languages". In: *arXiv preprint arXiv:2404.11553* (2024).

[13] Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. *Extrapolating Large Language Models to Non-English by Aligning Languages*. en. arXiv:2308.04948 [cs]. Oct. 2023. URL: http://arxiv.org/abs/2308.04948 (visited on 10/09/2024).

Appendix

The correlations we find are more striking when visualized, and reveal nuances to the data that the aggregated figures alone do not provide. Figure 1 shows the relationship between the number of speakers of a language and the average count of target words detected in LLaMA-70B responses. The trend indicates that languages with more speakers tend to have higher target word coverage, suggesting that these languages are better represented in the training data, leading to more accurate responses. The variance in languages with very few speakers probably exists because the online data for these languages, both in terms of quality and quantity, is likely to vary wildly.

Figure 2 illustrates the relationship between the number of speakers and the average response length (translated to English) for LLaMA-70B. The correlation suggests that languages with fewer speakers tend to have shorter responses, which may indicate terseness or under representation in training data. Interestingly, we did not observe the phenomenon for the smaller model, which was terser in general.

For some languages, counting language speakers entails entanglement in complicated social and political issues. Many of these language figures are disputed, which may also contribute to the noisiness of our data. When language figures differed wildly across sources, we averaged them.

In Figure 3 we present sample original questions from our dataset (in English) and their corresponding keywords that the raters were asked to look for and mark in the translated LLM responses. The questions and the keywords were hand-curated and chosen from a breadth of topics relevant to a high/middle school curriculum.

Figure 4 presents a screenshot from an evaluation sheet that were given to the manual evaluators. This particular example is for a Hebrew evaluator. For keyword matching, manual evaluators were asked to look for hand-curated keywords in the translated answers and assign "Keyword", "Incorrectness" and "Extraneous" scores accordingly.

The sample prompt given to the LLM to get its responses in non-English languages was:

You are a helpful assistant. Please answer the question in the same language that it is asked in. <QUESTION>
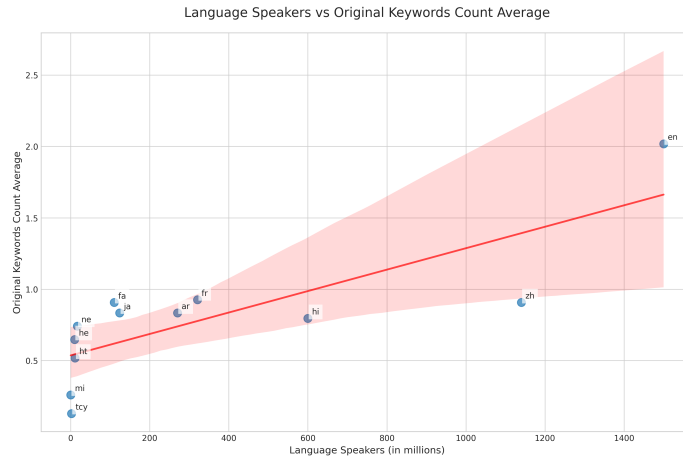
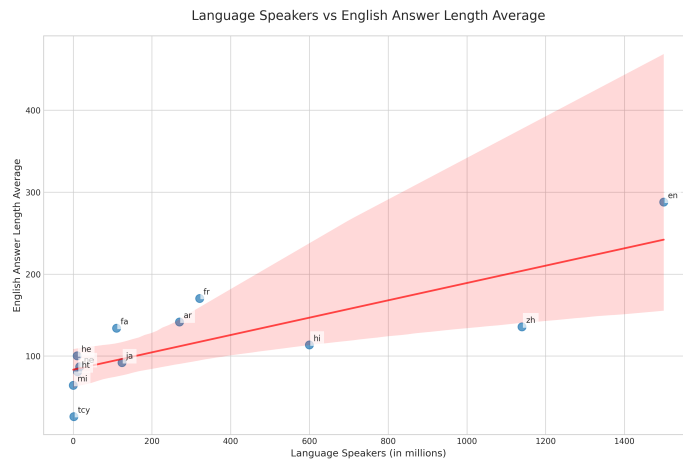Figure 1: Relationship between # of Speakers and Target Words for LLaMA-70B

Figure 2: Relationship between # of Speakers and Response Length (when translated to English) for LLaMA-70B

Figure 3: Sample questions from our dataset (in English) and their corresponding keywords

Figure 4: Sample sheet given to the manual evaluator to rate non-English responses