Multilingual Prompting in LLMs: Investigating the Accuracy and Performance

Praneeth Vadlapati
Independent researcher
praneeth.vad@gmail.com
ORCID: 0009-0006-2592-2564

Abstract: Large Language Models (LLMs) such as GPT-3.5 are trained using data from multiple sources, such as web data, which are predominantly in English. Hence, LLMs are commonly hypothesized to exhibit significant variations in response accuracy across multiple languages. This research investigates the hypothesis that the primary language of training data impacts the accuracy of responses to multilingual prompts. The experiments are conducted to evaluate the performance of LLMs across English and several other supported and unsupported languages, with questions structured to measure accuracy quantitatively. The study has been conducted on diverse tasks that include mathematical operations, word manipulation, and linguistic analysis. The results of the experiment demonstrate a clear edge of English prompts over prompts in other languages, with an accuracy of 80% to 100% for English prompts. A significant degradation is observed in accuracy for the same prompts translated into multiple languages other than English. The research underscores the limitations of English-dominated LLM architectures in effectively handling prompts across diverse languages. The study reveals the requirement for support for multiple languages to enable equitable access to AI-powered applications throughout the world. The source code is available at github.com/Pro-GenAI/PromptLang.

Keywords: Large Language Models (LLMs), multilingual prompts, cross-language NLP, response accuracy, Natural Language Processing (NLP)

I. INTRODUCTION

Recent advancements in Artificial Intelligence (AI) include Transformer architecture [1], which has facilitated the development of sophisticated Large Language Models (LLMs) [2] that are based on Transformer architecture. Language Models (LLMs) have rapidly advanced in their ability to process, interpret, and write text similar to humans [3]. However, despite their capabilities in handling multiple languages, LLMs often perform better in the primary language of training data, which is English. A predominance of the English language exists in the training data [4], [5]. Although this bias is attributable to the abundance and predominance of English content across the web, it raises significant concerns regarding equitable multilingual performance and access to AI tools for non-English speakers. The imbalance in languages has a potential to affect the performance of LLMs when processing non-English prompts. While the LLM training data includes non-English languages, their quantity and quality are typically not sufficient enough to ensure a considerable amount of performance and accuracy across those languages.

A. Proposed experiment

Considering the predominance of English in LLM training data, there is a requirement to evaluate LLMs using prompts in other languages to reveal the impact of the language of prompts on model performance. Hence, this research

investigates the LLM responses to the same prompts translated into multiple languages and compares them with the performance of the prompts in English.

B. Related work

Conneau et al. (2020) [6] highlighted that multilingual models often perform comparatively better in resource-rich languages and struggle with low-resource languages. Bender et al. (2021) [7] highlighted the "English-centric" nature of AI training, arguing that it reinforces global inequalities by prioritizing one predominant language over others. The challenges of multilingual support in AI have been documented extensively in NLP literature. Previous research has also explored performance discrepancies in translation tasks, sentence generation, and sentiment analysis across languages. However, studies lack a focus on calculating accuracy in multilingual prompts that include diverse tasks for the LLM and evaluation through quantitative methods and expected answers. This study utilizes prior research by examining multilingual accuracy discrepancies in a broad context through quantitative methodology.

II. METHODS

A. Selecting and loading an LLM

The experiment requires a pre-trained LLM that is trained using vast amounts of data from various sources. The accuracy of an LLM is a critical factor to consider in the selection process. Accordingly, GPT-3.5 [8] is the LLM selected for the experiment, considering the demonstrated performance across diverse tasks. LLMs accept prompts as inputs and return responses.

B. Creating prompts

Four distinct tasks were designed to test the response accuracy of the model across multiple languages. The tasks include mathematical operations, word manipulation, and linguistic analysis. Non-English languages that were selected for the experiment based on popularity are Hindi, Spanish, French, and German. For each task, English prompts were created, and each prompt was translated into the selected languages using Google Translate [9] to ensure accurate translation and consistency of the instructions across multiple languages. Google Translate is proven for its accuracy in translations across multiple languages [10].

TABLE I. PROMPTS CREATED

Index	Language	Prompt	Correct answer
1	English	What is 2^3? Return the answer in backticks like `100`.	8
	Hindi	2^3 क्या है? `100` जैसे बैकटिक्स में उत्तर लौटाएँ।	
	Spanish	¿Qué es 2^3? Devuelve la respuesta entre comillas simples, como `100`.	
	French	Combien vaut 2^3? Renvoie la réponse entre guillemets inversés, comme `100`.	
	German	Was ist 2^3? Geben Sie die Antwort in Backticks wie `100` zurück.	

Index	Language	Prompt	Correct answer
2	English	If you subtract 23 from twice the number 15, what is the result? Provide it in backticks like `100`.	7
	Hindi	यदि आप संख्या 15 के दुगुने में से 23 घटाते हैं, तो परिणाम क्या होगा? इसे `100` जैसे बैकटिक्स में प्रदान करें।	
	Spanish	Si le restas 23 al doble del número 15, ¿cuál es el resultado? Indícalo entre comillas simples, como `100`.	
	French	Si vous soustrayez 23 du double du nombre 15, quel est le résultat ? Indiquez-le entre guillemets, comme `100`.	
	German	Wenn Sie 23 von der doppelten Zahl 15 abziehen, was ist das Ergebnis? Geben Sie es in Backticks an, z. B. `100`.	
3	English	What is the reverse of the word “strawberry”? Return it as a string in backticks like `example`.	yrrebwarts
	Hindi	“strawberry” शब्द का उल्टा क्या है? इसे `example` जैसे बैकटिक्स में स्ट्रिंग के रूप में लौटाएँ।	
	Spanish	¿Cuál es el reverso de la palabra “strawberry”? Devuélvelo como una cadena entre comillas simples, como `example`.	
	French	Quel est l'inverse du mot “strawberry”? Renvoie-le sous forme de chaîne entre guillemets inversés comme `example`.	
	German	Was ist die Umkehrung des Wortes “strawberry”? Geben Sie es als Zeichenfolge mit Backticks zurück, z. B. `example`.	
4	English	How many vowels are in the word “aeronautics”? Finally, return a number in backticks like `100`.	6
	Hindi	“aeronautics” शब्द में कितने स्वर हैं? अंत में, `100` जैसे बैकटिक्स में एक संख्या लौटाएँ।	
	Spanish	¿Cuántas vocales tiene la palabra “aeronautics”? Por último, devuelve un	

Index	Language	Prompt	Correct answer
		número entre comillas simples, como `100`.	
	French	Combien de voyelles y a-t-il dans le mot “aeronautics”? Enfin, renvoyez un nombre entre guillemets inversés comme `100`	
	German	Wie viele Vokale hat das Wort “aeronautics”? Geben Sie abschließend eine Zahl in Backticks zurück, z. B. `100`.	

C. Generating responses

Prompts include instructions to include answers inside backticks (`) of the responses regardless of the language of the prompt. The response structure ensures a consistent evaluation metric for the accuracy of responses. Responses are generated using the selected LLM using the prompts created in the earlier steps. Ten attempts were conducted to measure the accuracy of each prompt.

D. Calculation of accuracy

The metric of the experiment is the accuracy of the responses. Correct answers are manually defined for each prompt. Accuracy is calculated by measuring the percentage of attempts in which the LLM returns correct answers. The correctness of each response was validated automatically during the experiment. The accuracy of responses generated by the LLM for multiple questions across multiple languages is calculated for comparison.

III. RESULTS

A. Accuracy of prompts in multiple languages

The results indicate that prompts in English achieved an accuracy of 80% to 100%, whereas prompts in other languages exhibited a significantly lower accuracy that is commonly below 50% and often 0% for numerous test cases.

TABLE II. ACCURACY RESULTS ACROSS MULTIPLE LANGUAGES

Index	Accuracy across multiple languages				
	English	***Hindi***	***Spanish***	***French***	***German***
1	**100%**	0%	40%	70%	**100%**
2	**100%**	30%	90%	0%	10%
3	80%	0%	30%	0%	90%
4	90%	0%	0%	0%	70%

B. Average accuracy for each language

The averages are calculated for the accuracy values mentioned in the table mentioned earlier. The average accuracy for multiple languages is mentioned in the table below.

TABLE III. AVERAGE ACCURACY ACROSS MULTIPLE LANGUAGES

Language	Average accuracy
English	**92.50%**
Hindi	7.50%
Spanish	40.00%
French	17.50%
German	67.50%

IV. DISCUSSION

The dominance of English training data correlates with the model's multilingual performance results. While LLMs are multilingual and possess the ability to respond in multiple languages, their understanding of non-English languages is limited. The discrepancy is particularly evident in Hindi, which is a widely used language that recorded the lowest average accuracy across tasks. German performed relatively better than other non-English languages despite being relatively less popular than other selected non-English languages. The lower quality of training data in non-English languages emerges as a predominant factor that contributes to the inaccuracy and inconsistency in performance. The discrepancies in the performance of multilingual prompts lead to concerns regarding equitable access to AI applications and tools. Users who are non-English speakers or lack expertise in English have a high probability of encountering errors or inconsistencies, thereby undermining trust and usability among diverse groups of users. Addressing such issues remains crucial for the adoption of AI across multiple fields across multiple regions of the world.

V. CONCLUSION

The study demonstrates that LLMs possess significant accuracy with English prompts, with a discrepancy of accuracy with prompts in other languages. The findings underscore the limitations of the English-centric training process and highlight the requirement for diverse datasets that contain high-quality content in multiple languages. As LLMs become popular across the industry and are integrated into numerous AI-based applications across the world, the improvement in multilingual performance is crucial to ensure equal access, trustability, and inclusiveness of AI-based applications for diverse audiences.

Future work should address the linguistic gaps by improving training data quantity and quality for non-English languages. Future work should explore the implications of linguistic gaps further through additional investigations. Challenges in improving linguistic diversity and multilingual accuracy include the collection and improvement of training data by hiring workers from linguistic expertise in diverse languages. Improvements in non-English responses could lead to societal benefits by addressing multilingual discrepancies, reducing digital inequality, and enhancing AI adoption globally by numerous more users and organizations.

REFERENCES

[1] A. Vaswani et al., "Attention is all you need," in Proceedings of the 31st International Conference on Neural Information Processing Systems, in NIPS'17. Red Hook, NY, USA: Curran Associates Inc., Dec. 2017, pp. 6000–6010. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[2] J. Uszkoreit, "Transformer: A Novel Neural Network Architecture for Language Understanding," Google Research. Accessed: Dec. 03, 2022. [Online]. Available: https://research.google/blog/transformer-a-novel-neural-network-architecture-for-language-understanding/

[3] B. A. y Arcas, "Do Large Language Models Understand Us?," Daedalus, vol. 151, no. 2, pp. 183–197, May 2022, doi: 10.1162/daed_a_01909.

[4] J. Kreutzer et al., "Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets," Transactions of the Association for Computational Linguistics, vol. 10, pp. 50–72, Jan. 2022, doi: 10.1162/tacl_a_00447.

[5] L. M. Alkwai, M. L. Nelson, and M. C. Weigle, "Comparing the Archival Rate of Arabic, English, Danish, and Korean Language Web Pages," ACM Trans. Inf. Syst., vol. 36, no. 1, Jun. 2017, doi: 10.1145/3041656.

[6] A. Conneau et al., "Unsupervised Cross-lingual Representation Learning at Scale," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 8440–8451. doi: 10.18653/v1/2020.acl-main.747.

[7] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?," in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, in FAccT '21. New York, NY, USA: Association for Computing Machinery, Mar. 2021, pp. 610–623. doi: 10.1145/3442188.3445922.

[8] "Introducing ChatGPT," OpenAI. Accessed: Dec. 03, 2022. [Online]. Available: https://openai.com/index/chatgpt/

[9] "Google Translate." Accessed: Dec. 03, 2022. [Online]. Available: https://translate.google.com/

[10] K. Turner, "Google Translate is getting really, really accurate," The Washington Post. Accessed: Dec. 03, 2022. [Online]. Available: https://www.washingtonpost.com/news/innovations/wp/2016/10/03/google-translate-is-getting-really-really-accurate/