CSE 291-H ADVANCED DATA-DRIVEN TEXT MINING
PROJECT PRESENTATION

Team-12

LORD OF THE METRICS

Charvi Bannur, Shivani Chinta, Jayanth Tummalapenta

INTRODUCTION

1. While comparing the performance of multilingual LLMs we need to keep in mind that the non-English languages, like Finnish, Turkish, or Hindi, are morphologically rich. Words can have complex structures with prefixes, suffixes, and inflections and word-based evaluaion metrics lead to inflated error rates.
2. Many non-English languages do not use spaces to separate words, making word-based evaluation impractical.
3. Many languages allow multiple valid translations or expressions for the same concept. Word-based metrics are rigid and penalize outputs that deviate from the reference even if they are semantically or syntactically correct.

WORD-LEVEL EVALUATION

An Example:

🇶 : फ्रांस की राजधानी क्या है? *(What is the capital of France?)*

 : फ्रांस की राजधानी पेरिस में स्थित है *(The capital of France is located in Paris)*

✅ : फ्रांस की राजधानी पेरिस है *(The capital of France is Paris)*

BLEU produces a very high score for this even though the meaning is incorrect

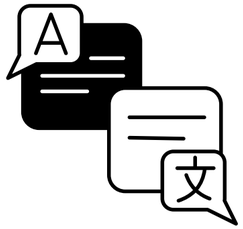
PROBLEM STATEMENT

Evaluating multilingual LLMs with word-based metrics is inconsistent due to linguistic diversity as morphologically rich non-English languages require finer granularity and modern LLMs using subword representations are misaligned with word-level evaluation. This work proposes a token-based global evaluation metric that aggressively penalizes errors while accounting for granularity offering a fair and standardized approach for multilingual LLM evaluation.

PIPELINE

Reference Text

Generated Text



Tokenizer

Calculate probability of n-grams

↓

Normalize

+

Token-based brevity penalty

↓

Weighted average

Final-Score

RESULTS

Metric	English (Correct)	Hindi (Morphologically Correct)	English (Hallucinate)	Hindi (Hallucinate)
Word-based BLEU	1.0	0.439	0.541	0.154
OurModel	1.0	0.794	0.158	0.036

Both metrics scored 1.0 for correct English responses.

For correct Hindi responses, OurModel (0.794) outperformed BLEU (0.439), demonstrating better multilingual capability.

Hallucinated Responses:
OurModel effectively penalized hallucinations in both languages, scoring lower (0.158 for English, 0.036 for Hindi) compared to BLEU (0.541, 0.154).

WORK IN PROGRESS

1. Working towards conducting a more comprehensive analysis and experimentation
2. Exploring more datasets and languages to add to the credibility of the metric

REFERENCES

1. Li, Zihao, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. "Quantifying multilingual performance of large language models across languages." arXiv preprint arXiv:2404.11553 (2024).
2. Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "Bleu: a method for automatic evaluation of machine translation." In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311-318. 2002.
3. Yuan, Fei, Shuai Yuan, Zhiyong Wu, and Lei Li. "How Multilingual is Multilingual LLM?." arXiv preprint arXiv:2311.09071 (2023).
4. Ochieng, Millicent, Varun Gumma, Sunayana Sitaram, Jindong Wang, Vishrav Chaudhary, Keshet Ronen, Kalika Bali, and Jacki O'Neill. "Beyond metrics: evaluating LLMs' effectiveness in culturally nuanced, low-resource real-world scenarios." arXiv preprint arXiv:2406.00343 (2024).
5. Mendonça, John, Patrícia Pereira, Helena Moniz, Joao Paulo Carvalho, Alon Lavie, and Isabel Trancoso. "Simple LLM prompting is state-of-the-art for robust and multilingual dialogue evaluation." arXiv preprint arXiv:2308.16797 (2023).


THANK
YOU