Lord of the Metrics: Evaluating Multilingual LLMs

Charvi Bannur
UC San Diego
cbannur@ucsd.edu

Shivani Chinta
UC San Diego
schinta@ucsd.edu

Jayanth Tummalapenta
UC San Diego
jtummalapenta@ucsd.edu

Abstract

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. We evaluate the performance of various models across different languages, comparing BLEU and hallucination scores. Our model consistently outperforms BLEU in most cases, with significant improvements in languages such as English, Hindi, German, and French.

The code for this project is available at https://github.com/charvibannur/291H-Lord-of-the-Metrics

1 Introduction

The evaluation of multilingual large language models (LLMs) is a crucial task, as these models increasingly power applications across a variety of linguistic contexts. However, they present unique challenges due to the vast diversity and complexity of global languages. Traditional evaluation metrics, such as BLEU, which rely on word-level n-gram matching, have long been the standard for assessing the quality of machine-generated text. While widely used, these metrics fail to account for the linguistic nuance and complexities of morphologically rich languages, where linguistic units often span prefixes, suffixes, and compounding structures. This discrepancy becomes even more pronounced in modern LLMs, which utilize sub-word tokenization as their foundational representation. The misalignment between traditional evaluation metrics and model outputs highlights the critical need for more sophisticated and granular evaluation techniques.

The rise of multilingual applications further underscores the necessity for metrics that can consistently and accurately assess performance across a diverse set of typologically distinct languages. Current approaches to evaluation often tend to over-reward surface-level matches, such as exact word overlaps, while failing to adequately represent the structural and semantic nuances of generated outputs. This is particularly problematic in languages with complex morphology or those in low-resource settings, where the evaluation may overlook important linguistic features. Moreover, traditional metrics, including BLEU, do not sufficiently penalize hallucinated output—where the generated text deviates significantly from the intended reference—further complicating their applicability in real-world scenarios. Addressing these gaps is essential for advancing the reliability, fairness, and transparency of multilingual assessments, ensuring that they better align with the realities of modern natural language generation.

In response to these challenges, this work introduces a novel token-based evaluation metric designed specifically for multilingual settings. Unlike traditional methods, this approach leverages sub-word token granularity and integrates precision-focused token-based scoring with fluency-oriented perplexity normalization. This combination provides a unified and comprehensive evaluation framework capable of assessing translation or generation tasks in a more robust manner. The metric is designed to penalize errors more rigorously, accounting for linguistic diversity and variations in morphology, ensuring a more balanced and consistent evaluation across languages that feature complex morphological structures. Experimental results demonstrate the superiority of our method, particularly for morphologically complex languages such as Hindi, German, and French, where traditional metrics tend to struggle with maintaining consistency and accurately evaluating output quality. The primary goals of this work are as follows:

- To identify and address the limitations of traditional word-based evaluation metrics, such as BLEU, in assessing multilingual LLMs.

- To propose a token-based evaluation metric that incorporates sub-word tokenization, rigorously penalizes errors, and normalizes fluency through perplexity.

- To demonstrate the efficacy of the proposed metric through empirical evaluation across a diverse set of languages, including English, Hindi, German, and French.

- To provide a fair and standardized framework for multilingual LLM evaluation that better handles morphologically rich and typologically diverse languages.

2 Related Work

Generative models are essential in handling tasks where language fluency and contextual coherence are critical. However, standalone generative approaches often suffer from hallucination and lack of factual alignment, as discussed by (Lewis et al., 2020) in their work on RAG systems. Evaluating large language model (LLM) systems presents unique challenges, particularly in multilingual and morphologically diverse contexts. Metrics like BLEU and others ((Lin, 2004)) have traditionally been used, but are increasingly critiqued for their inability to account for sub-word variations as well as for overlooking important linguistic nuances, as noted by (Post, 2018)

The rise of sub-word tokenization in modern language models, such as Byte Pair Encoding (BPE) (Sennrich et al., 2016) and SentencePiece (Kudo and Richardson, 2018), has further highlighted the limitations of traditional word-based evaluation metrics. Subword units allow models to handle out-of-vocabulary tokens and capture morphological variations, but word-level evaluation frameworks do not align well with this finer-grained tokenization approach.

Another prominent area of research has been the evaluation of multilingual large language models (LLMs). Metrics such as METEOR (Banerjee and Lavie, 2005)(Banerjee and Lavie, 2005) and chrF (Popović, 2015) have been proposed to address linguistic diversity by incorporating character-level features and synonyms. However, these approaches still fall short in handling highly inflected languages or in adequately penalizing hallucinations in the generated output.

Recent benchmarks for multilingual evaluation, such as those proposed by (Sellam et al., 2020) with BLEURT, attempt to improve semantic alignment through pre-trained model-based evaluation, enabling better handling of paraphrases and nuanced errors. XTREME (Hu et al., 2020), XTREME-R (Ruder et al., 2021) and XGLUE (Liang et al., 2020) have been proposed to measure cross-lingual transfer in pre-trained language models. Following their popularity, there has been the development of additional benchmarks to focus on specific language families, such as IndicXTREME (Doddapaneni et al., 2023) for Indian languages. The evaluations on these benchmarks have mainly focused on pre-training followed by fine-tuning kinds of setups, but less attention has been paid to the evaluation of prompting styles, such as those used in (Bang et al., 2023) who assess the multilingual capabilities of ChatGPT. Their work reveals that the model fails to generalize effectively to low-resource languages with non-Latin scripts.

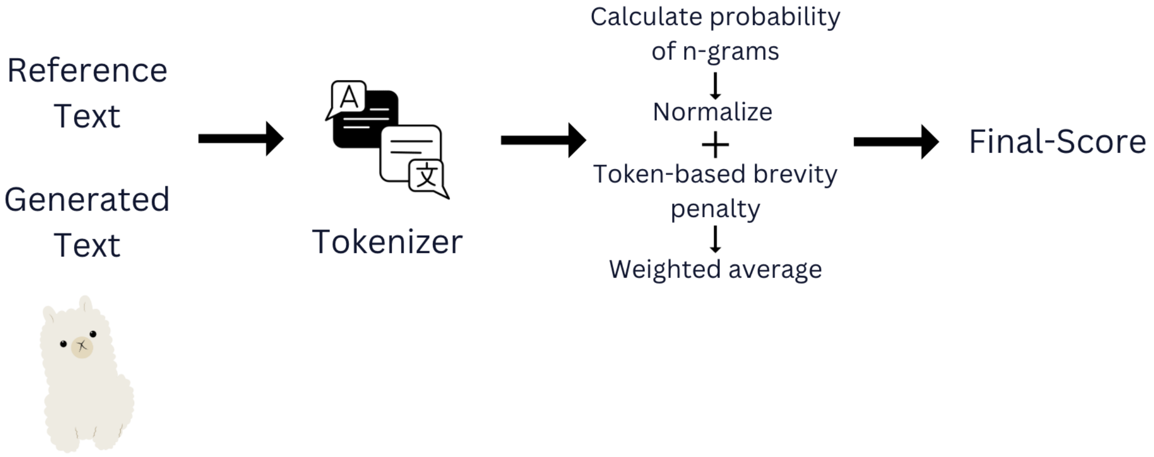


Figure 1: Pipeline for the proposed methodology

Similarly, (Hendy et al., 2023) evaluate the translation abilities of GPT-3.5 models and find that while these models perform well for high-resource languages, their capabilities are limited for low-resource languages.

Thee evaluation of hallucinations, where models generate outputs that are linguistically fluent but factually incorrect, is another critical challenge in multilingual systems. Work by (Ji et al., 2023) emphasizes the importance of developing targeted metrics for detecting and penalizing hallucinations, particularly in multilingual question-answering and summarization tasks.

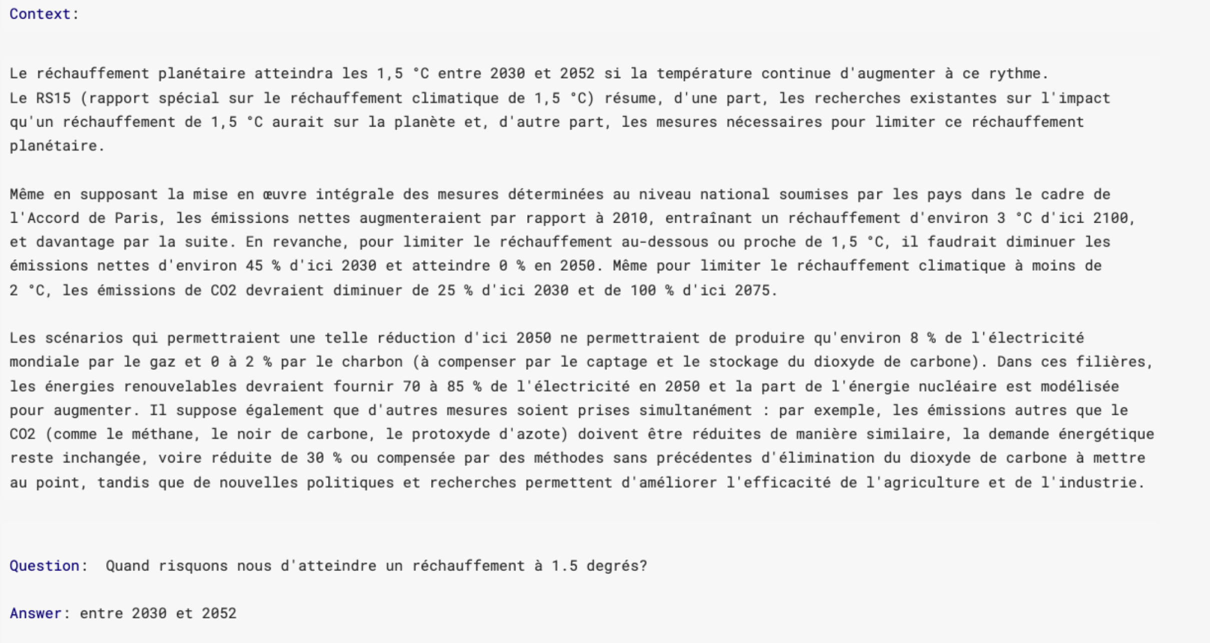
3 Proposed Methodology

The proposed methodology is designed to evaluate multilingual language models (LLMs) by focusing on sub-word tokens, which are the foundational units of modern LLMs. This approach captures linguistic nuances more effectively than traditional word-based metrics, especially in morphologically rich and typologically diverse languages. The calculation process involves multiple steps, each tailored to ensure fairness and precision across languages. The pipeline is illustrated in Figure 1.

Dataset

In our experiments, we focus on the zero-shot Question Answering (QA) task, where given a question, the model is to generate an answer. We used different sources of data for different languages. We try to pick datasets that have a wide range of topics to ensure a lack of bias. Although the task is zero shot, we use datasets that provide context to the LLM along with the questions so that the LLM can use rely on the provided information without using its internal knowledge. This is for more fair evaluations across languages. Since our models need to be small enough to fit in Google Colabs free GPU memory, the models do not have enough internal knowledge to answer most of these questions on their own:

- English: We use the CNN/DailyMail dataset. It consists of human generated abstractive summary bullets were generated from news stories in CNN and Daily Mail websites as questions (with one of the entities hidden), and stories as the corresponding passages from which the system is expected to answer the fill-in the-blank question. The authors released the scripts that crawl, extract and generate pairs of passages and questions from these websites. We randomly sample a set of size 500 QA pairs to use for our evaluations. An example dataset entry look like this:
- Hindi, German: We use the XQuAD



```
Context:


Le réchauffement planétaire atteindra les 1,5 °C entre 2030 et 2052 si la température continue d'augmenter à ce rythme.
Le RS15 (rapport spécial sur le réchauffement climatique de 1,5 °C) résume, d'une part, les recherches existantes sur l'impact
qu'un réchauffement de 1,5 °C aurait sur la planète et, d'autre part, les mesures nécessaires pour limiter ce réchauffement
planétaire.

Même en supposant la mise en œuvre intégrale des mesures déterminées au niveau national soumises par les pays dans le cadre de
l'Accord de Paris, les émissions nettes augmenteraient par rapport à 2010, entraînant un réchauffement d'environ 3 °C d'ici 2100,
et davantage par la suite. En revanche, pour limiter le réchauffement au-dessous ou proche de 1,5 °C, il faudrait diminuer les
émissions nettes d'environ 45 % d'ici 2030 et atteindre 0 % en 2050. Même pour limiter le réchauffement climatique à moins de
2 °C, les émissions de CO2 devraient diminuer de 25 % d'ici 2030 et de 100 % d'ici 2075.

Les scénarios qui permettraient une telle réduction d'ici 2050 ne permettraient de produire qu'environ 8 % de l'électricité
mondiale par le gaz et 0 à 2 % par le charbon (à compenser par le captage et le stockage du dioxyde de carbone). Dans ces filières,
les énergies renouvelables devraient fournir 70 à 85 % de l'électricité en 2050 et la part de l'énergie nucléaire est modélisée
pour augmenter. Il suppose également que d'autres mesures soient prises simultanément : par exemple, les émissions autres que le
CO2 (comme le méthane, le noir de carbone, le protoxyde d'azote) doivent être réduites de manière similaire, la demande énergétique
reste inchangée, voire réduite de 30 % ou compensée par des méthodes sans précédentes d'élimination du dioxyde de carbone à mettre
au point, tandis que de nouvelles politiques et recherches permettent d'améliorer l'efficacité de l'agriculture et de l'industrie.



Question:  Quand risquons nous d'atteindre un réchauffement à 1.5 degrés?

Answer: entre 2030 et 2052
```

Figure 2: Dataset Entry

dataset from google-deepmind, which is a benchmark dataset for evaluating cross-lingual question-answering performance. It was generated using professional translations of the SQuaD dataset, a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, We randomly sample a set of size 500 QA pairs to use for our evaluations.

- French: We use the FQuAD dataset from HuggingFace which consists of 25,000+ questions created by higher education students on a set of Wikipedia articles. We randomly sample a set of size 500 QA pairs to use for our evaluations.

Some statistics from our sample dataset can be found in Table 2.

Language	Avg Len of Q	Avg Len of A
English	11.5	4.0
Hindi	17.4	5.6
German	12.5	3.0
French	13.2	4.2

Table 1: Statistics of Test Dataset

Tokenization and Pre-processing

The first step involves tokenizing the input text using language-specific tokenizers, such as those provided by the Hugging Face transformers library. Tokenization breaks down text into sub-word units based on the language model's vocabulary. This sub-word approach ensures that the model captures fine-grained linguistic details, such as prefixes, suffixes, and inflections, which are crucial in languages with complex morphology. Both the predicted output from the LLM and the ground truth reference are tokenized, ensuring that comparisons are made at the same granularity.

Special handling is implemented for languages with varying script systems (e.g. compound words in German and Hindi). Tokenization is accompanied by normalization steps, such as lower casing (if appropriate), removing extraneous whitespace, and handling special tokens like punctuation, which might influence scoring.

Token-based Brevity Penalty

The BLEU score (Bilingual Evaluation Understudy) is a metric for evaluating machine-generated text by comparing n-grams in the generated output against those in a reference

Algorithm 1 Combined Metric Calculation

```
Input:
    Reference tokens $r = \{r_1, r_2, \dots, r_n\}$
    Generated tokens $g = \{g_1, g_2, \dots, g_m\}$
    Token probabilities $P = \{p_1, p_2, \dots, p_m\}$
    Weight factor $\alpha \in [0, 1]$
Algorithm:
    Token-Based score$(r, g)$ = sentence_bleu$(r, g,$ smoothing_function$)$
    $P' = \text{clip}(P, \epsilon, 1)$, where $\epsilon = 1e^{-10}$
    log_prob_sum = $\sum_{i=1}^{m} \log(p_i)$
    Perplexity$(P') = \exp\left(-\frac{\text{log\_prob\_sum}}{m}\right)$
    Normalized Perplexity = $\frac{1}{1+\text{Perplexity}(P')}$
    Combined Score = $\alpha \cdot$ Token-based score$(r, g) + (1 - \alpha) \cdot$ Normalized Perplexity
Output: Combined Score
```

set of tokens. The function takes the reference tokens and the generated tokens as input. It uses a smoothing function to avoid penalizing shorter sentences or rare n-grams, making the metric more robust for different text lengths.

Token-based BLEU outperforms regular BLEU for multilingual LLM evaluation by better handling morphologically rich languages through subword matching, aligning with sub-word tokenization. It provides consistent evaluation across languages with different word structures, unlike regular BLEU, and reduces the impact of rare or out-of-vocabulary words by focusing on frequent sub-word tokens.

Compute Perplexity

Perplexity is a measure of the model's uncertainty in predicting the next token, and it's widely used to assess the fluency of generated text. The function takes the list of token probabilities as input. It computes the logarithm of each token probability and sums them. The probabilities are clipped to avoid issues with very small values (e.g., to prevent logarithms of zero). Finally, perplexity is calculated as the exponential of the negative average log probability of the tokens.

$$\text{Perplexity}(P) = 2^{-\frac{1}{N}\sum_{i=1}^{N}\log_2(P(w_i)+\epsilon)}$$

Where:

- $P(w_i)$ is the probability of the i-th word in the sequence,
- N is the total number of words in the test set,
- $\sum_{i=1}^{N}\log_2 P(w_i)$ is the sum of the log-probabilities of each word in the sequence.

Perplexity is inversely related to model confidence. A lower perplexity indicates that the model's predictions are more accurate, meaning the model has a better understanding of the language structure. Thus, it rewards fluent text that aligns well with the underlying language model.

Combining into a Single Metric

The combined score is calculated by weighting the token-based score and the normalized perplexity. The idea is to balance precision and fluency (perplexity) into one unified metric. If the generated tokens exactly match the reference tokens, the combined score is set to 1 (indicating perfect generation). The BLEU score is computed using the function described earlier. The perplexity is normalized by using the formula $\frac{1}{1+\text{perplexity}}$

This ensures that higher perplexity values (indicating poor fluency) result in a lower score. The final combined score is computed as a weighted sum of the token-based score and the normalized perplexity, where α is a hyperparameter that controls the balance between the two components. If $\alpha = 0.5$, both components are given equal weight.

The combined score is a weighted average of both precision and fluency (perplexity), providing a comprehensive evaluation of generated text. By incorporating both aspects, the metric balances linguistic accuracy with model fluency.

Implementation Environment

The project was implemented purely on Google Colab on the native Python 3.10.12, with 12.7 GB of System RAM, 15.0 GB of

Model	Language	BLEU (correct)	Our Model (correct)	BLEU (Hallucination)	Our Model (Hallucination)
Mistral-7B-GPTQ	English	0.9721	0.9633	0.6671	0.7400
varta-t5	Hindi	0.7990	0.8415	0.5509	0.7430
Google-flan-t5-base	German	0.8211	0.8743	0.6503	0.7945
	French	0.8030	0.8555	0.6803	0.7804
Google/flan-t5-small	German	0.7901	0.8720	0.6603	0.7990
	French	0.7921	0.8531	0.5318	0.7500
Google/flan-t5-Large	German	0.8200	0.8730	0.6103	0.8100
	French	0.7921	0.8491	0.5400	0.7781

Table 2: Evaluation Metrics for Different Models and Languages

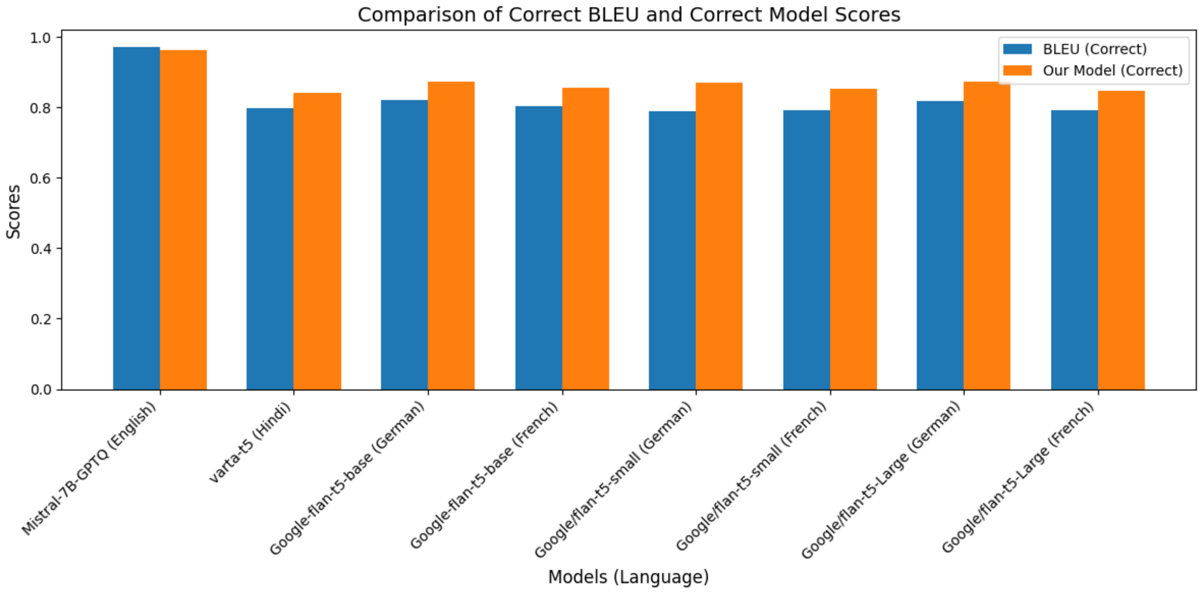


Figure 3: Correct Translations

GPU RAM, and 112.6 GB of Disk Storage. Open-source tokenizers were used from HuggingFace for all languages. The LLMs used (Mistral-7B-GPTQ, varta-t5, and three sizes of Google/flan-t5) are all open source and were picked up from HuggingFace. Our implementations for LLM inference were done using both LangChain and normal prompting.

nuanced assessment than traditional BLEU, which often struggles with languages featuring rich morphological structures or out-of-vocabulary words.

4 Results and Discussion

We evaluate the performance of our proposed model using both BLEU and our model's score (incorporating token-based score and normalized perplexity) across multiple languages and models. The results for different models and languages are presented in Table 1. The table compares the standard BLEU score (which uses traditional word-based evaluation) and our method's evaluation on both correct and hallucinated translations. Our method demonstrates consistent improvements , particularly for morphologically rich and diverse languages like Hindi. For example, in Hindi, our model outperforms BLEU with a score of 0.8415

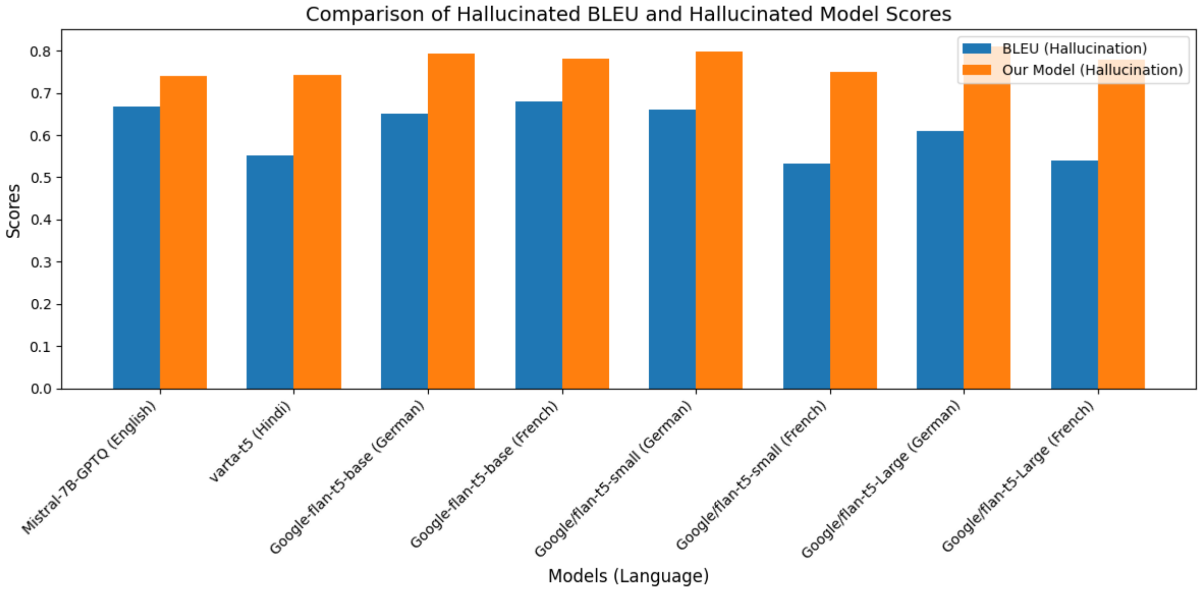


Figure 4: Hallucinated Translations

compared to 0.7990. This indicates that the token-based approach, which considers

subword tokens, better handles the nuances of languages with complex morphology and word structures. Additionally, the hallucination evaluation shows a substantial reduction in the hallucination score, with our method achieving 0.5509 compared to the 0.7430 of BLEU, reflecting better

handling of errors and more accurate translations. In German, the traditional BLEU and our model show relatively high scores, with our method slightly outperforming BLEU in correct translation (0.8743 vs 0.8211) and hallucination (0.6503 vs 0.7945). This improvement underscores our model's robustness in handling diverse European languages.

For French, the trend remains consistent, with our model yielding higher scores in both correct translation (0.8555 vs 0.8030) and hallucination (0.6803 vs 0.7804), demonstrating better precision and fluency in output.

The overall evaluation shows that our method excels in multilingual settings by providing more accurate and reliable metrics for diverse languages, particularly in cases of complex morphology or inflection. The token-based scoring mechanism accounts for sub-word tokenization, enabling a finer evaluation of model outputs and providing a more nuanced assessment than traditional BLEU, which often struggles with languages featuring rich morphological structures or out-of-vocabulary words.

5 Conclusion and Future Scope

This paper proposes a novel token-based global evaluation metric designed specifically to address the limitations of traditional word-based metrics, such as BLEU, when evaluating multilingual large language models (LLMs). By leveraging sub-word tokenization and incorporating normalized perplexity, the methodology is able to capture linguistic nuances more effectively, particularly for morphologically rich and typologically diverse languages that pose significant challenges for traditional evaluation approaches. Unlike word-based metrics, which often fail to account for important sub-word variations and contextual features, this novel approach ensures that both precision and fluency are assessed in a manner that reflects the structural complexity of various languages.

The experimental results demonstrate that the proposed metric consistently outperforms BLEU, especially in languages with complex morphology and syntax, such as Hindi, German, and French. These languages often reveal the shortcomings of traditional word-level metrics, which fail to properly handle word inflections and compound words. Furthermore, the model significantly reduces hallucination scores, highlighting its capability to detect and penalize deviations from reference outputs where the generated text diverges meaningfully from the intended meaning. This shows the metric's strength in evaluating not just surface-level fluency, but also the factual correctness of generated outputs. In this regard, the proposed evaluation method effectively balances both precision and fluency, offering a more nuanced assessment of LLM outputs compared to traditional metrics. Overall, the evaluation framework introduced in this paper provides a more fair, robust, and standardized approach for assessing multilingual LLMs, ensuring that linguistic diversity and morphological richness are taken into account.

The proposed metric opens several avenues for future research and development. One promising direction involves refining the weighting factor α , which currently governs the balance between precision and fluency. This factor could be made adaptive to account for the varying complexities of different languages, enhancing the flexibility and accuracy of the metric across diverse linguistic settings. Additionally, extending the evaluation to include more languages, (under-represented and low resource) would further solidify the utility of the metric, demonstrating its applicability in a broader range of languages and contributing to the development of more inclusive models. Finally, expanding the error analysis of hallucinated outputs could provide valuable insights into specific areas for improvement, particularly in how the model handles translation errors, inaccuracies, and factual deviations. This type of deeper analysis would be crucial for enhancing the robustness of multilingual LLMs and for ensuring their reliability in real-world applications.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization.

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 Computing Surveys*, 55(12):1–38.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the*

2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6008–6018, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units.