Apples to Apples? Towards Comparable Crosslingual Language Model Evaluation

Anonymous ACL submission

Abstract

Fair crosslingual evaluation of language models remains a fundamental challenge in multilingual NLP. Existing studies adopt a variety of downstream tasks and intrinsic metrics with different theoretical justifications, yet there has been little empirical investigation into whether these approaches yield meaningful crosslingual conclusions. We systematically examine crosslingual evaluation approaches using well-controlled monolingual language models trained on parallel data with varying tokenizer vocabulary sizes, and validate our findings on multilingual LLMs. We further discuss challenges in achieving comparable downstream evaluation across languages. Our results show that several widely used normalized metrics introduce crosslinguistic biases rooted in tokenization, encoding, and orthographic differences. In contrast, sentence-level negative log-likelihood computed over semantically equivalent sequences provides more meaningful and consistent crosslingual comparisons.[1]

Metric	Cross tokenizer	Cross language	Condition
Perplexity	✗	✗	same tokenizer+ same test set required
BPB/BPC	✓	✗	same test set (i.e., monolingual) only
MRR	✓✗	✓✗	same vocab size required
NLL	✓	✓	semantically equivalent content required

Figure 1: Suitability of intrinsic metrics for cross-tokenizer and cross-lingual comparison (✓ = suitable; ✗ = unsuitable; ✓✗ = conditionally suitable). NLL computed on semantically equivalent parallel content is the only metric suitable for both comparisons.

1 Introduction

Fairly evaluating language models across typologically diverse languages is a fundamental yet underappreciated challenge in multilingual NLP. Despite significant progress in extending language technology to an ever-growing number of languages for both monolingual and multilingual models, substantial performance gaps across languages persist (Chang et al., 2023; Joshi et al., 2020; Shani et al., 2026). Yet accurately measuring these gaps itself is a challenge: different metrics can yield substantially different conclusions about crosslingual model performance (Shani et al., 2026), raising the question of what a *fair* comparison even means.

Evaluation approaches in multilingual NLP broadly fall into two categories. Downstream task evaluations assess a model across a range of linguistic competencies, but fair crosslingual comparison using such evaluations is far from straightforward: conclusions are highly task-dependent (Rust et al., 2021; Limisiewicz et al., 2023), and language-specific benchmarks may not exist for many languages. Intrinsic metrics, by contrast, directly quantify how well a model assigns probability to held-out parallel texts, offering a more feasible and controlled signal for crosslingual comparison. Yet as we discuss below, "more controlled" does not necessarily mean "fair".

Within intrinsic metrics, there is no consensus yet on metric choice. One line of work evaluates the probability assigned to sequences that express the same meaning across languages, typically using negative log-likelihood or surprisal (e.g., Mielke et al., 2019; Cotterell et al., 2018; Wan, 2022). Another common approach relies on unit-normalized variants such as perplexity (PPL; Shliazhko et al., 2024; Arnett and Bergen, 2025; Thakur et al., 2025), bits-per-byte (BPB; Xue et al., 2022), or bits-per-character (BPC; Blevins et al., 2022), which normalize likelihood by tokens, bytes, or characters, respectively. Despite the intuitive justification for

[1] Code is available: [anonymous link].

each of these metrics (e.g., BPB (Gao et al., 2020) is tokenizer-agnostic), empirical evidence examining these claims about which metric to use has remained limited. The few studies that do compare metrics directly raise further concerns: Poelman and de Lhoneux (2026) argue that current evaluation metrics may not be universally comparable for two reasons: (i) different metrics can produce different rankings across languages, and (ii) even within a single metric, paraphrases expressing the same meaning can lead to inconsistent crosslingual rankings.

In this paper, we define a fair metric as one that is *independent of engineering choices*, specifically, tokenization, encoding, and orthography, such that no language gains an artificial advantage over another. To isolate metric behavior from confounds inherent in multilingual models, using 50 strictly-controlled monolingual language models (Anonymous, 2026) trained on parallel corpora with five vocabulary size settings, we conduct a systematic empirical comparison of six intrinsic metrics, identify their sources of bias. We further validate our findings on seven multilingual LLMs. Our results challenge both common practice and recent proposals in multilingual evaluation literature. We show that both token-normalized metrics such as perplexity (e.g., Shliazhko et al., 2024; Arnett and Bergen, 2025; Thakur et al., 2025) and character- and byte-level alternatives, which have been explicitly advocated as more robust (Shani et al., 2026), introduce substantial and systematic biases that favors languages with longer sequences or more tokens across languages. In contrast, we propose sentence-level NLL, normalizing by meaning rather than surface form, as a more principled metric for crosslingual comparison. We show that it is (i) substantially less sensitive to tokenization and encoding choices, (ii) more stable across translation alternatives that express the same meaning, and (iii) generalizable to both monolingual and multilingual settings. We further discuss key considerations for downstream evaluation, noting the challenges of achieving comparable assessment across languages and tasks.

2 Background & Related Work

2.1 Tokenization & Representation Biases

For crosslinguistic evaluation to be meaningful, it is necessary to disentangle intrinsic linguistic complexity from technical artifacts introduced by representation and preprocessing choices (Shani et al., 2026). Prior work has identified several sources of such artifacts, particularly those arising from tokenization and orthographic encoding (e.g., Petrov et al., 2023; Ahia et al., 2024).

Tokenization Language models are trained to predict the next token in a sequence, from a vocabulary of discrete tokens. Consequently, how text is segmented into tokens directly affects how the model represents and learns linguistic structure. Semantically equivalent content across languages can be segmented into very different numbers of tokens, even when the same tokenization algorithm is used (Petrov et al., 2023; Arnett et al., 2025). This disparity is often quantified with *Corpus Token Count* (CTC; Schmidt et al. (2024)), which measures the total number of tokens required to encode a corpus under a given tokenizer.

Differences in token segmentation can influence both training dynamics and evaluation metrics. Even under identical tokenization settings, languages may be split into substantially different numbers of segments, which can affect model performance (Shani et al., 2026). As a result, different languages may benefit from different tokenization strategies (e.g., Fujii et al., 2023; Vemula et al., 2025; Reddy et al., 2025; Toraman et al., 2023). For instance, Toraman et al. (2023) show that morphologically-rich languages like Turkish require a larger vocab size.

Orthography and Encoding Representation disparities also arise from differences in orthographic encoding. The same semantic content written in different scripts may require different numbers of UTF-8 bytes, a phenomenon known as the *byte premium* (Arnett et al., 2024). Similarly, different writing systems may encode equivalent content using substantially different numbers of characters, resulting in what has been termed the *length premium* (Arnett et al., 2024). These representational differences are independent of linguistic meaning but can systematically distort evaluation metrics that normalize by bytes or characters.

2.2 Crosslinguistic Evaluation

The most widely adopted approach relies on negative log-likelihood (NLL) computed over semantically equivalent sequences across languages. Within this framework, studies differ in how they normalize the raw NLL scores. Cotterell et al. (2018) and Mielke et al. (2019) normalize sentence-level surprisals by the number of characters of a

specific language (usually English), using the resulting scores to rank languages by difficulty. Limisiewicz et al. (2024) adopts a similar approach but normalizes by the number of bytes instead. Wan (2022) simply aggregates sentence-level NLL over an entire development set without normalization. A more structured alternative is proposed by Tsvetkov and Kipnis (2024), who introduce Information Parity (IP), defined as the ratio of English NLL to the NLL of another language over parallel text.

Other work evaluates multilingual models using metrics that operate at different granularities. Shliazhko et al. (2024) report perplexity scores across languages, while Blevins et al. (2022) argue for bits per character (BPC) as a more language-neutral unit. Byte-level language models further shift toward bits per byte (BPB; Xue et al., 2022; Zhang and Xu, 2022), which Poelman and de Lhoneux (2026) suggest may yield fairer comparisons when scores are scaled to a shared unit. Despite the apparent justification for each of these metrics, Poelman and de Lhoneux (2026) demonstrate that applying different metrics to the same model on the same data can produce conflicting language rankings. This highlights that metric choice is not merely a technical detail but a substantive decision with interpretive consequences. This work offers concrete empirical guidance for metric selection in crosslingual evaluation and contributes to ongoing efforts toward fair multilingual NLP.

3 Existing Intrinsic Metrics & Their Potential Confounds

In this section, we review mainstream probability-based metrics used in language modeling and evaluate their suitability for crosslinguistic and cross-model comparison. Let t_i be the ith token in sequence $\mathbf{s}$ which has b UTF-8 bytes, c characters, and can be tokenized into n tokens.

Bits per Byte BPB is proposed by Gao et al. (2020) because of "its invariance to different tokenization schemes and the ambiguity of measuring characters in Unicode."

$$\mathrm{BPB}(\mathbf{s}) = -\frac{1}{b}\sum_{i=1}^{n}\log_2(P(t_i \mid \mathbf{t}_{<i})) \qquad (1)$$

It has been used in cross-model evaluations (e.g., Rae et al., 2021) for the same language. However, we argue that BPB introduces bias when comparing across languages due to differences in orthographic encoding. The same semantic content written in different scripts may require different numbers of UTF bytes (i.e., different *byte premiums*). For example, Arnett et al. (2024) report that characters in scripts such as Khmer typically require three bytes per character (excluding diacritics), whereas Latin characters require only one byte. As a result, if two models assign similar probabilities to semantically equivalent sequences in two languages, the language with the higher byte premium will necessarily obtain a lower BPB score because the total surprisal $\sum_{i=1}^{n}\log_2(P(t_i \mid \mathbf{t}_{<i}))$ is normalized by a larger number of bytes, b.

Bits per Character Bits per Character (BPC) measures the average sequence NLL per character:

$$\mathrm{BPC}(\mathbf{s}) = -\frac{1}{c}\sum_{i=1}^{n}\log_2(P(t_i \mid \mathbf{t}_{<i})) \qquad (2)$$

It is mainly used in evaluations of character-level models (e.g., Al-Rfou et al., 2019). While it mitigates tokenization bias, it may still introduce orthographic bias like BPB. For example, Chinese translations are usually represented with fewer characters than English and thus it has a lower length premium. Normalizing over the number of characters means that Chinese will almost always have a higher BPC than English, as c is smaller.

Mean Reciprocal Rank MRR (Limisiewicz et al., 2023) evaluates models on the rank of the next token rather than its probability, making it probability-agnostic. However, since ranking is computed over the entire vocabulary, a larger vocabulary raises the upper bound of possible ranks, such that models with larger vocabularies tend to achieve lower MRR regardless of actual performance.

$$\mathrm{MRR}(\mathbf{s}) = \frac{1}{n}\sum_{t=1}^{n}\frac{1}{\mathrm{rank}(t_i, P(\cdot \mid \mathbf{t}_{<i}))} \qquad (3)$$

Perplexity Held-out perplexity measures the average number of nats (or bits) required to predict tokens in a corpus. It has been the default metric for evaluating language models, with lower values typically indicating better performance (e.g., Brown et al., 2020):

$$\mathrm{PPL}(\mathbf{s}) = \exp(-\frac{1}{n}\sum_{i=1}^{n}\ln(P(t_i \mid \mathbf{t}_{<i}))) \qquad (4)$$

Some previous studies report perplexity as the measure to compare models with different tokenizers and training data (e.g., Thakur et al., 2025). Because perplexity is normalized by the number of tokens, it is highly sensitive to how the input is tokenized. Consider a language model trained on Language A and evaluated on a parallel corpus containing both Language A and Language B, where the two languages use different scripts. Since the model has not been exposed to Language B's script, its tokenizer will decompose Language B's text into smaller units (down to bytes under standard BPE) and produce longer token sequences. As the denominator increases, the normalized log-likelihood decreases, artificially deflating perplexity for Language B despite the model having no genuine knowledge of it.[2]

Token-level Negative log-likelihood Token-level NLL (Token NLL) measures the NLL of the sequence normalized by the number of tokens:

$$\text{Token NLL}(\mathbf{s}) = -\frac{1}{n}\sum_{i=1}^{n} \ln(P(t_i \mid \mathbf{t}_{<i})) \quad (5)$$

Token NLL shares the same tokenization sensitivity as perplexity, as discussed above.

Sentence-level Negative log-likelihood Sentence-level NLL (Sent-NLL) is the sum of the NLL or surprisal per token (See 6).

$$\text{Sent-NLL}(\mathbf{s}) = -\sum_{i=1}^{n} \ln(P(t_i \mid \mathbf{t}_{<i})) \quad (6)$$

Because Sent-NLL sums over the entire sequence without dividing by token count, it is not directly affected by how the sequence is segmented, making it less sensitive to vocabulary size than perplexity, Token NLL, BPB, or BPC. For this reason, Sent-NLL of parallel sentences has been more widely adopted in crosslinguistic comparisons (Lin et al., 2024; Costa-Jussà et al., 2022; Chang et al., 2026), and we argue that it provides the most meaningful basis for crosslingual evaluation among the metrics considered here.

Other metrics There are other metrics such as Information Parity (IP; Tsvetkov and Kipnis, 2024) and Bits per English Character (BPEC; Cotterell et al., 2018). As they are transformations of Sent-NLL covered here, they should align with the results of Sent-NLL.

[2]We present a case study using `GPT2` (Radford et al., 2019) in Appendix E to illustrate this point.

Summary As summarized in Figure 1, normalization over tokens, characters, or bytes introduces systematic biases and is unsuitable for cross-tokenizer or cross-lingual comparison. BPB and BPC are appropriate for cross-tokenizer comparison within the same language, but not cross-lingually, as different writing systems yield different byte and character counts. Perplexity is unsuitable for cross-lingual comparison because different tokenizers segment the same content into different numbers of tokens. MRR requires constant vocabulary size across models. Only NLL over semantically equivalent content is suitable for both cross-model and cross-lingual comparison, as it normalizes by meaning rather than representation units.

4 Experiment 1: All metrics except for Sent-NLL are biased in different ways

To ensure that observed metric differences reflect the metrics themselves rather than model-specific confounds,[3] we use monolingual models trained on parallel corpora with varying vocabulary sizes.

Experiment Setup. We use `GPT2` models from concurrent work (Anonymous, 2026):[4] monolingual models, each trained exclusively on one language's portion of a parallel corpus. This design ensures that cross-linguistic differences are attributable to the languages themselves rather than to shared parameters or training data imbalances.

Two model sizes are used. `GPT2-small` models are trained on Parallel-10, a ten-language parallel corpus spanning six language families (∼20M words per language),[5] with five vocabulary sizes,[6] yielding 50 models in total. `GPT2-medium` models are trained on Parallel-3, a trilingual corpus (Chinese, English, Arabic; ∼400M words per language) with two vocabulary sizes (32k and 65k), yielding 6 models, allowing us to test whether findings hold at scale.

Evaluation We evaluate the models using the metrics discussed in Section 3 on three parallel data

[3]For example, a multilingual model trained predominantly on Chinese may assign lower BPC to Chinese text, masking BPC's inherent bias against logographic writing systems: Chinese characters encode more information per character than alphabetic scripts, artificially shortening sentences and improving BPC scores.

[4]An anonymized draft is provided in the Supplementary Materials.

[5]Arabic, Chinese, English, French, German, Finnish, Polish, Russian, Korean, and Turkish.

[6]8192, 32768, 49152, 65536, 81920.

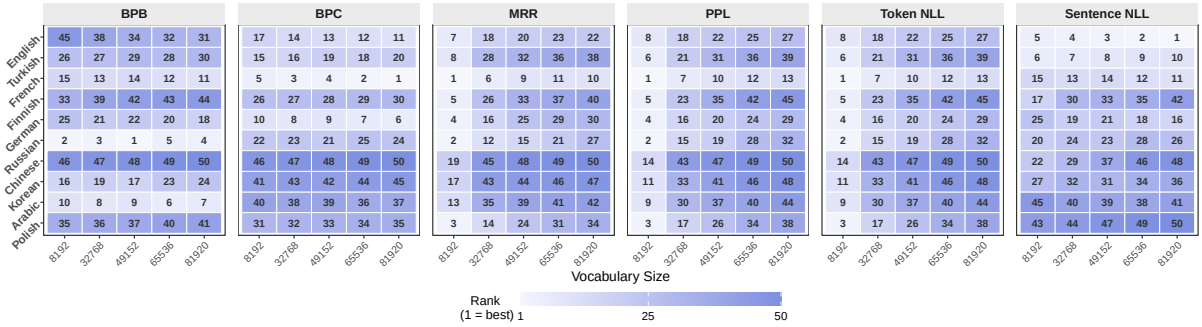

Figure 2: Language model rankings across intrinsic metrics. Each cell shows the rank of a language-vocabulary size combination among all 50 settings (10 languages * 5 vocabulary sizes) within each metric. Lighter colors indicate better performance (rank 1 = best).

sources in the corresponding languages: FLORES-200 (NLLB Team et al., 2022), a multi-parallel machine translation benchmark covering more than 200 languages; the in-domain test split of our training corpus (40k sentences, see Table 3); and 1,000 sentences from the Parallel Universal Dependencies treebanks (PUD; Zeman et al., 2017).

Results Figure 2 reports rankings across 10 languages and 5 vocabulary sizes on FLORES-200 (results on other datasets are in Appendix C). Different metrics produce strikingly different crosslingual rankings. While Poelman and de Lhoneux (2026) attribute this to the general unreliability of probability-based metrics, we argue the inconsistency stems from specific, identifiable biases in each metric's normalization unit.

As shown in Figure 2, metrics normalized by surface forms are systematically biased by encoding or vocabulary size. BPB and BPC consistently rank Chinese last, as Chinese has shorter character and byte sequences. MRR, PPL, and Token NLL rankings are dominated by vocabulary size: smaller vocabularies yield better rankings regardless of language. The same pattern holds in our scaling experiments (Figure 6). These results suggest that rankings under surface-normalized metrics are largely predictable from engineering choices alone, rather than reflecting genuine differences in language modeling quality. In contrast, Sent-NLL shows no sensitivity to either encoding or vocabulary size.

To substantiate this claim, we conduct Spearman's rank correlation tests between each metric and the corresponding potential bias factors discussed in Section 3. For BPC and BPB, the number of characters and bytes for a given language remains constant regardless of tokenization, so we average model performance across tokenizers before computing correlations. In contrast, other metrics are confounded by CTC. As CTC varies across tokenization settings because different tokenizers produce different numbers of tokens for the same language, we therefore compute correlations between each model's metric value and its corresponding CTC directly, without averaging across tokenizers. For Sent-NLL, which is not affected by normalization over tokens, characters, or bytes, we compute correlations with all three factors.

Metric	Potential Confounds	ρ	p
BPC	#characters	-0.81	0.0082
BPB	#bytes	-0.94	<0.001
PPL	CTC	-0.93	<0.001
MRR	CTC	0.89	0.001
Token NLL	CTC	-0.93	<0.001
Sent-NLL	#characters	-0.38	0.28
Sent-NLL	#bytes	0.14	0.71
Sent-NLL	CTC	-0.19	0.19

Table 1: Correlation Analysis between metrics and potential confounds. PPL is scaled by log10.

Both the correlation coefficient ρ and p-value p of the correlation analysis are reported in Table 1. We find that the potential confounds are strongly and significantly correlated with language rankings for all metrics (ρ>0.8, p<0.01) except Sent-NLL, which shows no significant correlation with sequence length measured in bytes, characters, or tokens.

5 Experiment 2: Sent-NLL is more robust for translation alternatives

One potential objection to Sent-NLL is that different translations of the same text may yield different probability distributions, leading to inconsistent crosslingual rankings (Poelman and de Lhoneux, 2026). However, Poelman and de Lhoneux (2026)

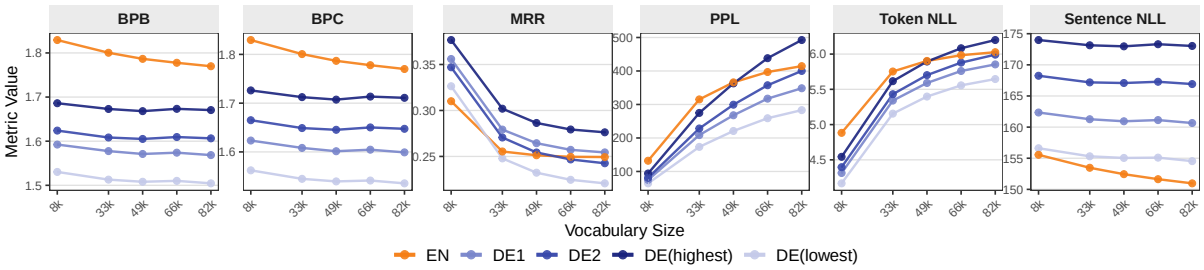

Figure 3: Metric values for English and four German translation alternatives across six metrics.

demonstrate this instability using both translations and paraphrases within the same language rather than translations only. Because paraphrases allow more flexible wording as long as the gist of the meaning is kept (Neusner, 1986; Bhagat and Hovy, 2013), they are more likely to encode different amounts of information. In this case, their findings speak to within-language variation rather than the validity of crosslingual comparison over parallel translations. We provide additional discussion about this assumption in Section D.2

Experiment Setup. To test each metric's sensitivity to meaning-equivalent paraphrases, we re-evaluate Poelman and de Lhoneux's (2026) experiment using translations only and additionally evaluate Sent-NLL. Specifically, we use the English-German translation pairs from WMT2019 (Barrault et al., 2019) with an additional translation reference source created independently from Freitag et al. (2020a,b), denoted as DE1 and DE2 separately. Following Poelman and de Lhoneux (2026), for these two alternative German translations, we construct two derived datasets: DE_{lowest} and $\text{DE}_{highest}$, consisting of sentences that receive the lower or higher metric value in each pairwise comparison, respectively.

We use the same sets of models trained on English and German in Experiment 1 and compare Sent-NLL for English and German across the five data sources: EN (English source text), DE1 (German translation), DE2 (German alternative translation), DE_{lowest}, and $\text{DE}_{highest}$.

Results. The results are shown in Figure 3 and reveal three major findings. First, metrics including MRR, PPL, and Token NLL vary monotonically by vocabulary size, which is consistent with our findings in Section 4. Second, for these metrics, even with the same vocab size setting, using different translations in German yields different language rankings between English and German. This pattern is especially salient when vocab size $\geq$ 33k. This suggests that these metrics are more sensitive to surface-level differences. Finally, we find that Sent-NLL produces consistent language rankings across translation variants (i.e., English scores lower than all four German datasets). If we accept these results, they support the notion that paraphrases do differ in meaningful ways. We argue that we should not consider paraphrases in the same way as translation-equivalent pairs for this reason. The same applies to BPB and BPC, but as we show in Section 4, if we consider more typologically diverse language with different scripts, the results will change. We further show that the rankings across languages become stable when the sample size $\geq$ 500. Details of the experiment can be found in Appendix D.1.

6 Experiment 3: Sent-NLL generalizes to multilingual language models

The metric biases we identify are mathematical in nature and therefore model-agnostic: they arise from how metrics normalize probability, not from properties of any particular model.

Experiment Setup. To verify this, we evaluate nine multilingual LLMs on four languages (Arabic, English, Russian, Chinese) from FLORES-200. We select two groups of models based on different criteria. First, `BLOOM` (Workshop et al., 2022), `mGPT` (Shliazhko et al., 2022), and `XGLM` (Lin et al., 2022) are selected because their training data composition is well-documented, allowing us to derive expected language rankings. Second, `Teuken` (Ali et al., 2024), `Qwen` (Yang et al., 2025), and `Llama` (Grattafiori et al., 2024) are included due to their wide usage in multilingual settings. Training data composition is reported in Table 7. For the first group, we compare expected language rankings, derived from training data proportions under the assumption that more data leads to better performance

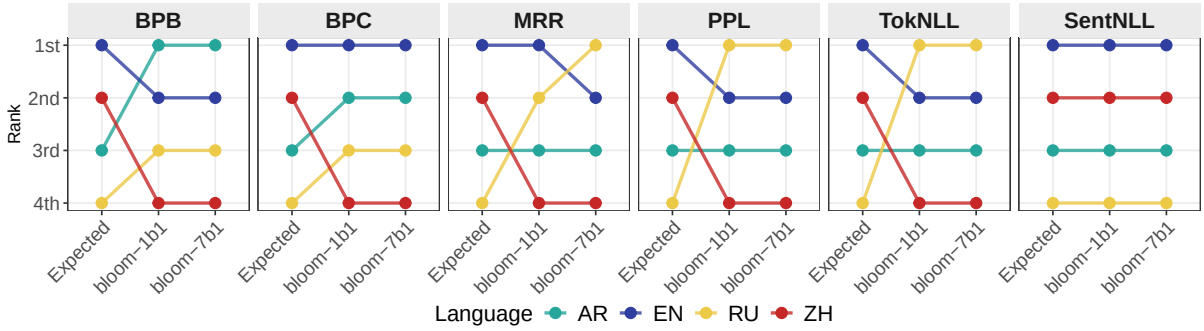

Figure 4: Expected ranking and actual ranking across 6 metrics for BLOOM-7b1 and BLOOM-1b1.

(Hoffmann et al., 2022), against actual metric rankings. For the second group, whose full training data composition is undisclosed, we verify whether the dominant language reported in the technical documentation achieves the best ranking under each metric. A reliable metric should produce rankings consistent with these expectations.

Results. Results for BLOOM-7B and BLOOM-1B are shown in Figure 4 and confirm our predictions. The leftmost column shows the expected ranking based on training data composition; the right two columns show actual rankings under each metric. BPB and CPB are distorted by script differences: Chinese consistently ranks worst despite comprising the second largest portion of BLOOM's training data. PPL and Token-NLL are distorted by tokenization: Russian, which receives little training data in BLOOM, paradoxically ranks best because unseen languages are over-segmented into more tokens, artificially deflating token-normalized scores. Only Sent-NLL produces rankings that reflect actual training data composition. We find the same pattern with other LLMs; raw results for all models are reported in Figure 13, and ranking results for XGLM-4.5B and mGPT in Figure 12.

7 Considerations for Downstream Tasks

While intrinsic metrics are our primary focus, downstream task evaluations remain widely used in practice. However, fair crosslingual comparison using downstream tasks is far from straightforward. Below, we outline key considerations for researchers seeking to make meaningful comparisons across languages using downstream evaluations.

Translated Benchmarks. Some translated tasks may enable comparable evaluation across languages. For example, a math benchmark translated into another language may allow a researcher to make an inference about the relative performance on math tasks in two languages. Math and reasoning benchmarks are less likely to be influenced by translation quality and are more language-agnostic (Wu et al., 2025). This relies on sufficiently high translation quality, however. Machine translated benchmarks without human validation may contain translation artifacts and introduce noise into the evaluation (Singh et al., 2025). High-quality benchmarks created through professional human translation, e.g. MMMLU (OpenAI, 2024) or FLORES (NLLB Team et al., 2022) and derivatives like Belebele (Bandarkar et al., 2024) or SIB-200 (Adelani et al., 2024), are less likely to suffer from these issues.

Culture-specific tasks. Another consideration is cultural specificity. Even when translated faithfully, some benchmarks may not be appropriate for comparable evaluation if their content is specific to a particular linguistic or cultural context. Localized benchmarks, e.g. IndoMMLU (Indonesian; Koto et al., 2023), CMMLU (Chinese; Li et al., 2024), Turkish MMLU (Yüksel et al., 2024), Arabic MMLU (Koto et al., 2024), and KMMLU (Korean; Son et al., 2025) are evaluate not only target-language performance but also cultural knowledge. Localized benchmarks such as these correlate more strongly with human preferences than simply translated benchmarks (Wu et al., 2025). But achieving the same score on two localized MMLU-style benchmarks does not signal comparable performance, as the questions are not difficulty-matched or otherwise controlled.

Language-specific tasks. Language-specific tasks, such as linguistic knowledge benchmarks like BLiMP (Warstadt et al., 2020) and MultiBLiMP (Jumelet et al., 2026), may not be comparable across languages for other reasons. First, even

benchmarks which evaluate the same relatively narrow tasks, such as subject-verb agreement in MultiBLiMP (Jumelet et al., 2025), are based on non-parallel data. The items for each language differ in quantity and quality. The number of items in MultiBLiMP per language ranges from 7 (Gujarati) to 4615 (Old Russian). For languages with very few items, like Gujarati, MultiBLiMP scores may be less reliable than those with much larger numbers of samples. Additionally, as MultiBLiMP is based on UD Treebanks (de Marneffe et al., 2021; Nivre et al., 2020), which vary in their annotation schema. These inconsistencies may lead to different conclusions about relative linguistic complexity, for example (Rodríguez and López, 2025). Both of these issues can be addressed through the use of balanced parallel data with consistent annotation schema.

These benchmarks are also sensitive to crosslinguistic differences. In the development of CLAMS (Mueller et al., 2020), the authors discuss how differences in pronoun dropping, grammatical person distinction, and distribution of negative polarity items shaped the types of tests they could use in the comparison of agreement across languages.

Subject-verb agreement is inherently different across languages. For some languages, the task is trivial as there is no agreement and the verb form is always the same (e.g. Chinese and Tagalog; Chao, 1968; Reyes, 1969). Other languages have more complex agreement patterns due to higher number of distinctions between number (singular, dual, and plural number, e.g. Arabic; Bettega and D'Anna, 2022), gender/noun class (about 10 noun classes in Swahili; Moxley, 1998), verb class (at least 4 verb classes in Georgian; Makharoblidze, 2012), gender (Tamil; Venkatesan, 2024). It remains an open question how to fairly compare linguistic performance across languages, when the complexity of the phenomenon differs.

8 Discussion & Conclusion

Fair crosslingual evaluation is essential as language models are increasingly deployed beyond English. How we evaluate models determines how we interpret results and allocate resources. Prior work has noted that normalized metrics may be unreliable for cross-lingual comparison in monolingual or multilingual settings (Cotterell et al., 2018; Mielke et al., 2019), but these observations have remained largely theoretical. Problematic metrics continue to be widely used in practice (Shliazhko et al., 2024; Arnett and Bergen, 2025; Thakur et al., 2025) or explicitly advocated in recent proposals (Shani et al., 2026). Our contribution is to provide the first systematic empirical evidence across multiple metrics, languages, and tokenizer settings that confirms these intuitions and quantifies the extent of each bias. Empirical grounding matters precisely because intuition alone has not been sufficient to change community norms.

It is important to note that Sent-NLL is not entirely free from confounds. However, we argue that any remaining sensitivity reflects inherent linguistic complexity rather than arbitrary engineering choices, which is a qualitatively different kind of bias that is arguably unavoidable in any crosslingual comparison. Our proposal implicitly assumes that all languages encode roughly comparable amounts of information, an assumption the literature both supports and challenges (Koplenig et al., 2017; Crystal, 2010; Fenk-Oczlon and Fenk, 2014; Bentz et al., 2023). Even if languages systematically differ in information density, such differences reflect intrinsic linguistic properties rather than metric artifacts. In this sense, Sent-NLL may reflect both how well a model has learned a language and the intrinsic information density of that language. Therefore, even if the linguistic equi-complexity hypothesis is falsified, Sent-NLL may not be a perfect measure of model performance in isolation, but it remains the least biased option among commonly used intrinsic metrics for crosslingual comparison. Any remaining confound due to linguistic complexity should be viewed as a limitation shared by all crosslingual evaluation rather than a systematic bias of the metric itself.

Finally, our results show that Sent-NLL rankings show more variability across datasets than normalized metrics (Figures 7, 8, 2). We argue this reflects a strength rather than a weakness. The apparent stability of metrics such as BPB, BPC, PPL, and Token-NLL is artificial: their rankings are largely determined by engineering choices — script properties, tokenization granularity, and vocabulary size — rather than genuine differences in language modeling quality. Sent-NLL, by contrast, is sensitive to actual test data content, meaning its rankings reflect real differences in how well a model has learned each language. Variability across datasets is therefore expected and desirable.

Limitations

Our evaluation only covers 10 languages due to limited parallel data. A broader language sample would strengthen generalizability, although the significant correlations observed for all other metrics under the same sample size suggest that statistical power is sufficient to detect meaningful relationships.

In the paraphrase experiments, only English and German translation pairs are tested due to the limited translation resources. Extending this analysis to additional bilingual pairs with multiple translation references would further strengthen the conclusions.

More broadly, our approach relies on high-quality parallel sentences. Although datasets such as FLORES provide relatively broad coverage, many languages still lack reliable parallel data, which limits the applicability of this evaluation framework.

AI assistance

We used GPT-5 to assist with polishing the writing. All code was written by the authors and subsequently reviewed with the help of GPT-5.

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Valentin Hofmann, Tomasz Limisiewicz, Yulia Tsvetkov, and Noah A Smith. 2024. Magnet: Improving the multilingual fairness of language models with adaptive gradient-based tokenization. *Advances in neural information processing systems*, 37:47790–47814.

Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. 2019. Character-level language modeling with deeper self-attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3159–3166.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Jan Ebert, Alexander Arno Weber, Richard Rutmann, Charvi Jain, Max Lübbering, Daniel Steinigen, Johannes Leveling, and 1 others. 2024. Teuken-7b-base & teuken-7b-instruct: Towards european llms. *arXiv preprint arXiv:2410.03730*.

Anonymous. 2026. Autoregressive transformer models are systemically biased toward englishlike languages. Under review.

Catherine Arnett and Benjamin Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6607–6623, Abu Dhabi, UAE. Association for Computational Linguistics.

Catherine Arnett, Tyler A. Chang, and Benjamin Bergen. 2024. A bit of a problem: Measurement disparities in dataset sizes across languages. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 1–9, Torino, Italia. ELRA and ICCL.

Catherine Arnett, Tyler A. Chang, Stella Biderman, and Ben Bergen. 2025. Explaining and mitigating crosslingual tokenizer inequities. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The Belebele benchmark: A parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Regina Barzilay and Kathleen R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 50–57, Toulouse, France. Association for Computational Linguistics.

Christian Bentz, Ximena Gutierrez-Vasques, Olga Sozinova, and Tanja Samardžić. 2023. Complexity tradeoffs and equi-complexity in natural languages: a meta-analysis. *Linguistics Vanguard*, 9(s1):9–25.

Simone Bettega and Luca D'Anna. 2022. *Gender and number agreement in Arabic*, volume 109. Brill.

Rahul Bhagat and Eduard Hovy. 2013. Squibs: What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2022. Analyzing the mono- and cross-lingual pretraining dynamics of multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3575–3590, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chutima Boonthum. 2004. iSTART: Paraphrase recognition. In *Proceedings of the ACL Student Research Workshop*, pages 31–36, Barcelona, Spain. Association for Computational Linguistics.

Joan Bresnan and Tatiana Nikitina. 2009. The gradience of the dative alternation. *Reality exploration and discovery: Pattern interaction in language and life*, pages 161–184.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tyler A Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K Bergen. 2023. When is multilinguality a curse? language modeling for 250 high-and low-resource languages. *arXiv preprint arXiv:2311.09205*.

Tyler A Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K Bergen. 2026. Goldfish: Monolingual language models for 350 languages. *Language Resources and Evaluation Conference (LREC)*.

Yuen Ren Chao. 1968. *A Grammar of Spoken Chinese*. University of California Press.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49(2):375–395.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Ryan Cotterell, Sabrina J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 536–541, New Orleans, Louisiana. Association for Computational Linguistics.

David Crystal. 2010. The equality of languages. In *The Cambridge Encyclopedia of Language*, 3rd edition, chapter 2, pages 6–7. Cambridge University Press, Cambridge.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Gertraud Fenk-Oczlon and August Fenk. 2014. Complexity trade-offs do not prove the equal complexity hypothesis. *Poznan Studies in Contemporary Linguistics*, 50(2):145.

Markus Freitag, George Foster, David Grangier, and Colin Cherry. 2020a. Human-paraphrased references improve neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1183–1192, Online. Association for Computational Linguistics.

Markus Freitag, David Grangier, and Isaac Caswell. 2020b. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Takuro Fujii, Koki Shibata, Atsuki Yamaguchi, Terufumi Morishita, and Yasuhiro Sogawa. 2023. How do different tokenizers perform on downstream tasks in scriptio continua languages?: A case study in Japanese. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 39–49, Toronto, Canada. Association for Computational Linguistics.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, and 1 others. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, and 1 others. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Jaap Jumelet, Leonie Weissweiler, and Arianna Bisazza. 2025. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *arXiv preprint arXiv:2504.02768*.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2026. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *Transactions of the Association for Computational Linguistics*, 14:193–216.

Alexander Koplenig, Peter Meyer, Sascha Wolfer, and Carolin Müller-Spitzer. 2017. The statistical trade-off between word order and word structure–large-scale evidence for the principle of least effort. *PloS one*, 12(3):e0173614.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large Language Models Only Pass Primary School Exams in Indonesia: A Comprehensive Test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12359–12374, Singapore. Association for Computational Linguistics.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. ArabicMMLU: Assessing Massive Multitask Language Understanding in Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640, Bangkok, Thailand. Association for Computational Linguistics.

Beth Levin. 1993. English verb classes and alternations: A preliminary investigation. *University of Chicago Press*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285, Bangkok, Thailand. Association for Computational Linguistics.

Tomasz Limisiewicz, Jiří Balhar, and David Mareček. 2023. Tokenization impacts multilingual language modeling: Assessing vocabulary allocation and overlap across languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5661–5681, Toronto, Canada. Association for Computational Linguistics.

Tomasz Limisiewicz, Terra Blevins, Hila Gonen, Orevaoghene Ahia, and Luke Zettlemoyer. 2024. MYTE: Morphology-driven byte encoding for better and fairer multilingual language modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15059–15076, Bangkok, Thailand. Association for Computational Linguistics.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André FT Martins, and Hinrich Schütze. 2024. Mala-500: Massive language adaptation of large language models. *arXiv preprint arXiv:2401.13303*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, and 2 others. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Yikang Liu, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Jialong Tang, Pei Zhang, Baosong Yang, and 1 others. 2026. A systematic assessment of language models with linguistic minimal pairs in chinese. *Transactions of the Association for Computational Linguistics*, 14:755–771.

Nitin Madnani and Bonnie J. Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–387.

Tamar Makharoblidze. 2012. *The georgian verb*. Lincom Europa Munich.

Sabrina J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989, Florence, Italy. Association for Computational Linguistics.

Jeri L Moxley. 1998. Semantic structure of swahili noun classes. *Language history and linguistic description in Africa*, 2:229.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational Linguistics.

Jacob Neusner. 1986. Translation and paraphrase: the differences and why they matter. *Hebrew studies*, pages 26–37.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, 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, and 20 others. 2022. No language left behind: Scaling human-centered machine translation.

OpenAI. 2024. Multilingual Massive Multitask Language Understanding (MMMLU).

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. *Advances in neural information processing systems*, 36:36963–36990.

Wessel Poelman and Miryam de Lhoneux. 2026. Form and meaning in intrinsic multilingual evaluations. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2503–2521, Rabat, Morocco. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, and 1 others. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Varshini Reddy, Craig W Schmidt, Yuval Pinter, and Chris Tanner. 2025. How much is enough? the diminishing returns of tokenization training data. *arXiv preprint arXiv:2502.20273*.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Elena V Reyes. 1969. Some problems of interference in the use of english verbs by native tagalog speakers. *Language Learning*, 19(1-2):87–97.

Antoni Brosa Rodríguez and M. Dolores Jiménez López. 2025. Beyond the data: The impact of annotation inconsistencies in UD treebanks on typological universals and complexity assessment. In *Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 43–51, Vienna, Austria. Association for Computational Linguistics.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Diana Santos and Anabela Barreiro. 2025. Paraphrase and translation: the importance of being close. *Open Research Europe*, 5:52.

Craig W Schmidt, Varshini Reddy, Haoran Zhang, Alec Alameddine, Omri Uzan, Yuval Pinter, and Chris Tanner. 2024. Tokenization is more than compression. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 678–702, Miami, Florida, USA. Association for Computational Linguistics.

Chen Shani, Yuval Reif, Nathan Roll, Dan Jurafsky, and Ekaterina Shutova. 2026. The roots of performance disparity in multilingual language models: Intrinsic modeling difficulty or design choices? *arXiv preprint arXiv:2601.07220*.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mGPT: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mgpt: Few-shot learners go multilingual. *arXiv preprint*.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2025. KMMLU: Measuring massive multitask language understanding in Korean. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4076–4104, Albuquerque, New Mexico. Association for Computational Linguistics.

Aamod Thakur, Ajay Nagpal, Atharva Savarkar, Kundeshwar Pundalik, Siddhesh Dosi, Piyush Sawarkar, Viraj Thakur, Rohit Saluja, Maunendra Sankar Desarkar, and Ganesh Ramakrishnan. 2025. The art of breaking words: Rethinking multilingual tokenizer design. *arXiv preprint arXiv:2508.06533*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Cagri Toraman, Eyup Halit Yilmaz, Furkan Şahinuç, and Oguzhan Ozcelik. 2023. Impact of tokenization on language models: An analysis for turkish. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–21.

Alexander Tsvetkov and Alon Kipnis. 2024. Information parity: Measuring and predicting the multilingual capabilities of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7971–7989, Miami, Florida, USA. Association for Computational Linguistics.

Saketh Reddy Vemula, Sandipan Dandapat, Dipti Sharma, and Parameswari Krishnamurthy. 2025. Rethinking tokenization for rich morphology: The dominance of unigram over BPE and morphological alignment. In *The 14th International Joint Conference on Natural Language Processing and The 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 232–252, Mumbai, India. Association for Computational Linguistics.

Madhusmitha Venkatesan. 2024. Gender agreement in a tamil-hindi bilingual situation: The role of feature valuation. *Formal Approaches to South Asian Languages*.

Ada Wan. 2022. Fairness in representation for multilingual NLP: Insights from controlled experiments on conditional language modeling. In *International Conference on Learning Representations*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, and 1 others. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Minghao Wu, Weixuan Wang, Sinuo Liu, Huifeng Yin, Xintong Wang, Yu Zhao, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. The bitter lesson learned from 2,000+ multilingual benchmarks. *arXiv preprint arXiv:2504.15521*.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Arda Yüksel, Abdullatif Köksal, Lütfi Kerem Senel, Anna Korhonen, and Hinrich Schuetze. 2024. TurkishMMLU: Measuring massive multitask language understanding in Turkish. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7035–7055, Miami, Florida, USA. Association for Computational Linguistics.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, and 43 others. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

Mengjiao Zhang and Jia Xu. 2022. Byte-based multilingual NMT for endangered languages. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4407–4417, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

A Parallel Multilingual Corpus

The training data sources and split can be found in Table 2 and Table 3.

All the sources follow the data license from OPUS (Tiedemann, 2012).

B Model Training Details

The training hyperparameters are reported in Table 4.

C Evaluation Dataset-specific results

The crosslingual rankings for FLORES, Parallel-10 test split, and PUD are shown in Figures 7, 8 separately. The correlation test are visualized in Figures 9, 10, 11 separately. We observe similar patterns as discussed in the main text. In the correlation test, we find significant correlation between each metric and their potential confounds across three datasets, but this is not the case for Sent-NLL.

Source	Description	parallel-10		parallel-3	
		Words	Sents	Words	Sents
OpenSubtitles (Lison and Tiedemann, 2016)	Movie/TV show subtitles	17.0M	2.0M	64.9M	8.1M
Bible (Christodouloupoulos and Steedman, 2015)	Bible text	1.5M	57K	1.5M	61.2K
WikiTitle	Titles of Wikipedia articles	167K	42K	848.2K	218.3K
Neulab-Ted	Ted talks	356K	19K	3.7M	201.2K
KDE4	KDE4 localization files	117K	13K	280.7K	30.0K
Ted2020 (Reimers and Gurevych, 2020)	Ted talks	347K	19K	3.1M	171.0K
QED	Subtitles for educational videos	10K	0.6K	73.5K	5.0K
Ubuntu	Ubuntu localization files	<1K	<100	2.5K	<1K
Tatoeba	Translated sentences from Tatoeba	<1K	<100	20.2K	2.9K
GNOME	GNOME localization files	<1K	<100	<1K	<100
UNPC (Ziemski et al., 2016)	United Nation documents	NA	NA	306.0M	10.7M
GlobalVoices	News report	NA	NA	156.0K	7.7K
NewsCommentary (Barrault et al., 2019)	news comments	NA	NA	2.5M	55.6K
Ted2013	Ted talks	NA	NA	983.9K	171.0K
Tanzil	Quran translations	NA	NA	14.2K	1.7K
Overall		19.5M	2.4M	384.2M	19.6M

Table 2: Sources of the OPUS-10 and OPUS-3 corpora, from Anonymous (2026).

Dataset	Language	Train		Dev		Test	
		#sent	#word	#sent	#word	#sent	#word
OPUS-10	English	2,338,286	19,089,154	10,000	82,457	40,000	327,724
	Arabic	2,338,286	14,426,353	10,000	61,982	40,000	246,954
	Chinese	2,338,286	30,129,452	10,000	130,205	40,000	516,401
	Turkish	2,338,286	12,540,711	10,000	54,238	40,000	215,345
	German	2,338,286	16,544,740	10,000	71,464	40,000	283,997
	Korean	2,338,286	11,857,382	10,000	51,292	40,000	203,336
	Polish	2,338,286	13,458,472	10,000	57,932	40,000	231,262
	French	2,338,286	17,603,323	10,000	76,087	40,000	302,056
	Russian	2,338,286	14,613,449	10,000	62,951	40,000	251,018
	Finnish	2,338,286	11,975,617	10,000	51,698	40,000	205,766
OPUS-3-380M	English	19,555,763	383,187,466	10,000	198,658	40,000	780,968
	Arabic	19,555,763	332,197,948	10,000	171,796	40,000	676,630
	Chinese	19,555,763	677,898,235	10,000	349,927	40,000	1,380,368

Table 3: Data splits for three experimental settings across ten languages. Word counts are computed using whitespace-separated tokens for all languages except Chinese, for which each character is treated as a word, from Anonymous (2026).

Training	
train_size	10M
num_epoch	10
train_amount	19M
batch_size	128
context_length	128
grad_acc_steps	1
weight_decay	0.1
warmup_steps	10%
lr	5e-4
lr_scheduler	linear

Table 4: Training Hyperparameters

D Paraphrase related experiments and discussion

D.1 Sample size and metric stability

A possible objection is that even for semantically equivalent translations, sentence-level rankings may not align across languages, potentially affecting aggregated language rankings. We argue that such sentence-level mismatches introduce noise, and that their impact diminishes when averaged over a sufficiently large number of parallel sentences. To test this, we conduct a follow-up experiment under the same setting while varying the number of samples from 1 to the full dataset. For each sample size, we repeat the experiment with 15 random seeds. The results are shown in Figure 5. When the sample size is small (1–100), the rankings

are indeed unstable. However, they become stable once the sample size reaches around 500.[7]

D.2 Paraphrases encode different amounts of information

Another concern may arise from our exclusion of paraphrases in Section 5. Including paraphrases from Freitag et al. (2020a) could indeed change the results. For instance, DE_{lowest} might obtain lower values and potentially rank below English , based on the results from Poelman and de Lhoneux (2026). However, this doubt relies on a common, but flawed, assumption that paraphrases within the same language should receive similar probabilities from a language model because they express the *same meaning*, i.e., they are assumed to be semantically identical (Barzilay and McKeown, 2001; Boonthum, 2004; Madnani and Dorr, 2010) and therefore language models should assign similar probabilities to paraphrases of the same text (Poelman and de Lhoneux, 2026).[8]

We argue that this assumption does not necessarily hold. Paraphrases are often semantically similar but not strictly identical, and they encode different amounts of information. Consequently, there is no theoretical reason to expect that language models should assign more similar probabilities to two paraphrases than to translations of either sentence.

To illustrate this point, we use two paraphrases of a sentence from Poelman and de Lhoneux (2026) as an example:

(1) a. Do you know how to play chess?
b. Can you play chess?

'Know how to play chess' entails possessing the procedural knowledge about the rules of chess. 'Can play chess' entails knowledge of the rules, but it is more sensitive to context. It could refer to possession of a chessboard or whether it is socially acceptable to play.

While in many contexts these differences may not be important, paraphrases are rarely semantically equivalent. Even grammatical alternations, e.g. dative alternation (c.f. 'I threw the box to John' and 'I threw John the box.'; Levin, 1993), may not be considered semantically equivalent (Bresnan and Nikitina, 2009). It may be more precise to label the sentences in Ex. (1) as *quasi-paraphrases* (Bhagat and Hovy, 2013) or *near-equivalence* (Santos and Barreiro, 2025), which may encompass a "broader, approximate, equivalence".

Admittedly, both translations and paraphrases may encode different amounts of information for the same source. However, paraphrases typically permit substantially greater variation because they are not constrained by crosslinguistic correspondence (e.g., Neusner, 1986). As a result, inconsistencies observed with paraphrases do not necessarily undermine crosslingual evaluation. Instead, they introduce an additional and less controlled source of variation.

[7] Note that all other experiments in this paper use sample sizes greater than 500.

[8] For example, Poelman and de Lhoneux (2026) assume that paraphrases have the same information content and therefore language models should assign "roughly the same [Token NLL or BPC] values for paraphrases within a single language".

Language	PPL	Token-NLL	MRR	Sent-NLL
English	49.77	3.91	0.43	99.34
Chinese	32.83	3.49	0.36	277.67
Russian	9.45	2.25	0.51	248.56
Arabic	12.22	2.50	0.44	251.00

Table 5: crosslingual evaluation results of an English language model on FLORES.

Language	Accuracy
English	0.97
Chinese	0.54
Russian	0.43
Arabic	0.50

Table 6: MultiBLiMP results of GPT-2 across four languages.

E Case study with `GPT2`

To to illustrate the bias on perplexity introduced by differences in compression, we provide a case study using `openai/GPT-2` model (monolingual English) evaluated on four languages from FLORES. As shown below, token-level metrics (PPL, token-NLL, MRR) all suggest that the model performs best on Russian. This is counterintuitive given that the model is trained only on English, and MultiBLiMP results[9] show that GPT2 performs well only on English (0.97) and near chance on other languages (in fact, below chance in Russian). In contrast, Sentence-level NLL assigns the lowest score to English, aligning better with expectations and the MultiBLiMP evaluations.

[9] For Chinese, it is evaluated on ZhoBLiMP (Liu et al., 2026) VerbPhrase category.

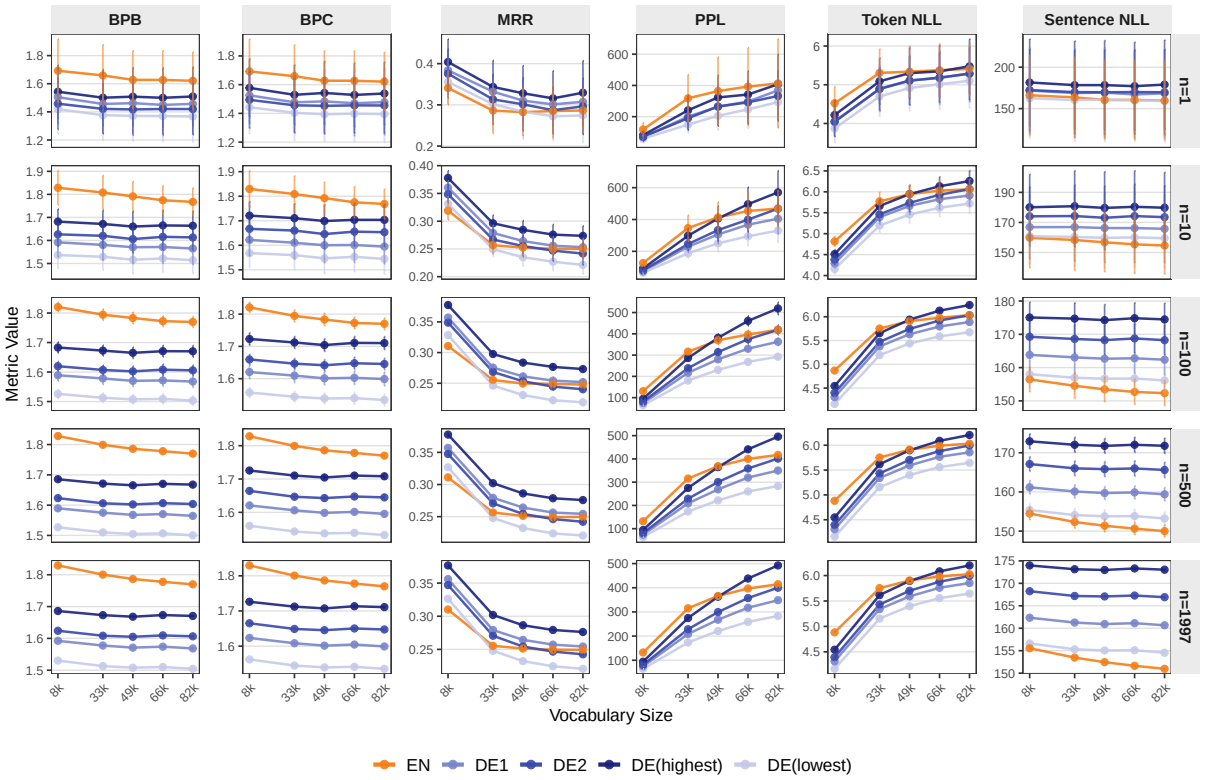

Figure 5: Each row shows results for a different sample size. Error bars represent 95% CI across 15 random seeds.

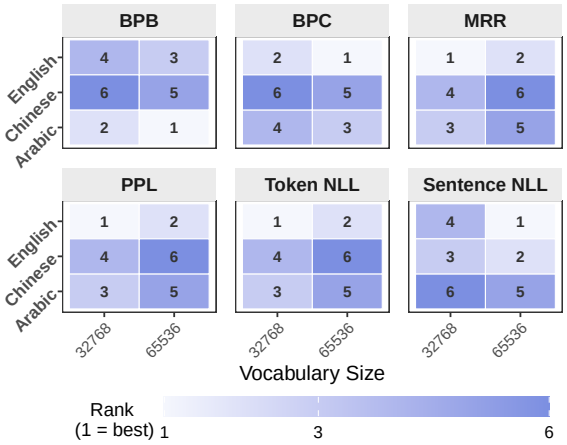

Figure 6: Language model rankings in a scaled setting.

F Scaled Experiment Result

See Figure 6.

G Typological Features of the Languages

H Multilingual Experiment

Model	Training data
Llama-3.1-8B	Primarily English; AR, RU, ZH not officially supported (Grattafiori et al., 2024)
Llama-3.2-1B	Primarily English; AR, RU, ZH not officially supported (Grattafiori et al., 2024)
Llama-3.2-3B	Primarily English; AR, RU, ZH not officially supported (Grattafiori et al., 2024)
Qwen3-8B	36T tokens across 119 languages; primarily in English and Chinese (Yang et al., 2025)
Teuken-7B	Supporting 24 European languages; AR, RU, ZH not officially supported (Ali et al., 2024)
BLOOM-1.1B	45 languages; AR: <1%, EN: 31.3%, RU: 0%, ZH: 18.3% (Workshop et al., 2022)
BLOOM-7.1B	45 languages; AR: AR: <1%, EN: 31.3%, RU: 0%, ZH: 18.3% (Workshop et al., 2022)
XGLM-4.5B	31 languages; training data EN>RU>ZH>AR (Lin et al., 2022)
mGPT	61 languages; training data EN>RU>ZH>AR (Shliazhko et al., 2022)

Table 7: Models used in this study and their pretraining data composition.

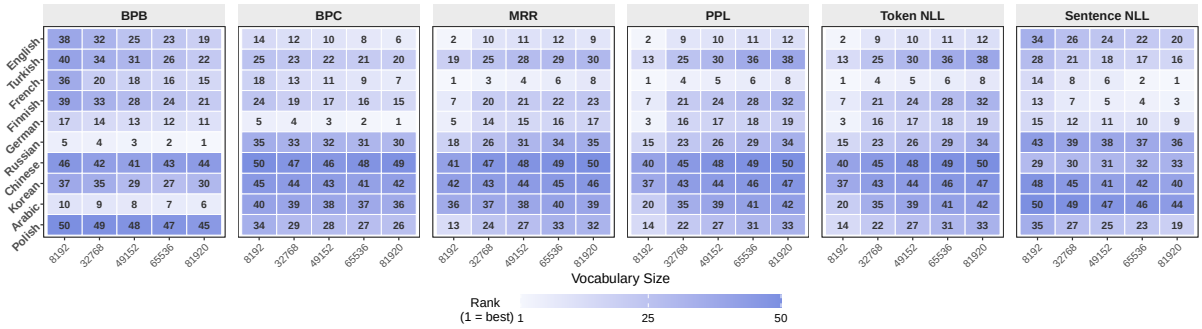

Figure 7: Language rankings across different intrinsic metrics for Parallel-10 test split

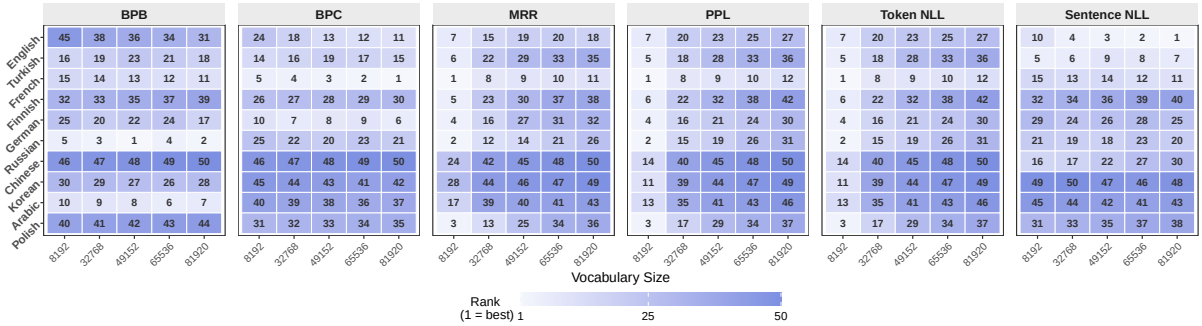

Figure 8: Language rankings across different intrinsic metrics for PUD

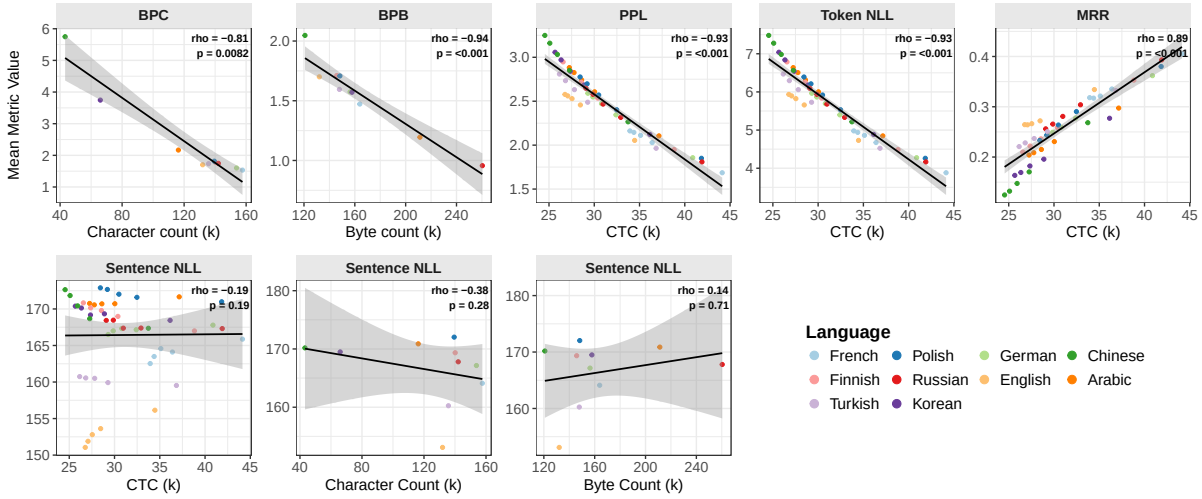

Figure 9: Spearman's Rank Correlation between the intrinsic metrics and their potential confounding factors in FLORES (perplexity is scaled by log10).

Languages	Language Family	Morphology	Orthography	Word Order
English	Indo-European (Germanic)	Analytic	Alphabetic (Latin)	SVO
French	Indo-European (Romance)	Fusional	Alphabetic (Latin)	SVO
German	Indo-European (Germanic)	Fusional	Alphabetic (Latin)	SVO/V2
Polish	Indo-European (Slavic)	Fusional	Alphabetic (Latin)	Flexible (SVO)
Finnish	Uralic	Agglutinative	Alphabetic (Latin)	SVO
Turkish	Turkic	Agglutinative	Alphabetic (Latin)	SOV
Chinese	Sino-Tibetan	Isolating	Logographic	SVO
Arabic	Afro-Asiatic (Semitic)	Nonconcatenative	Abjad	VSO/SVO
Korean	Koreanic	Agglutinative	Featural (Hangul)	SOV

Table 8: Typological properties of the languages used in our experiments.

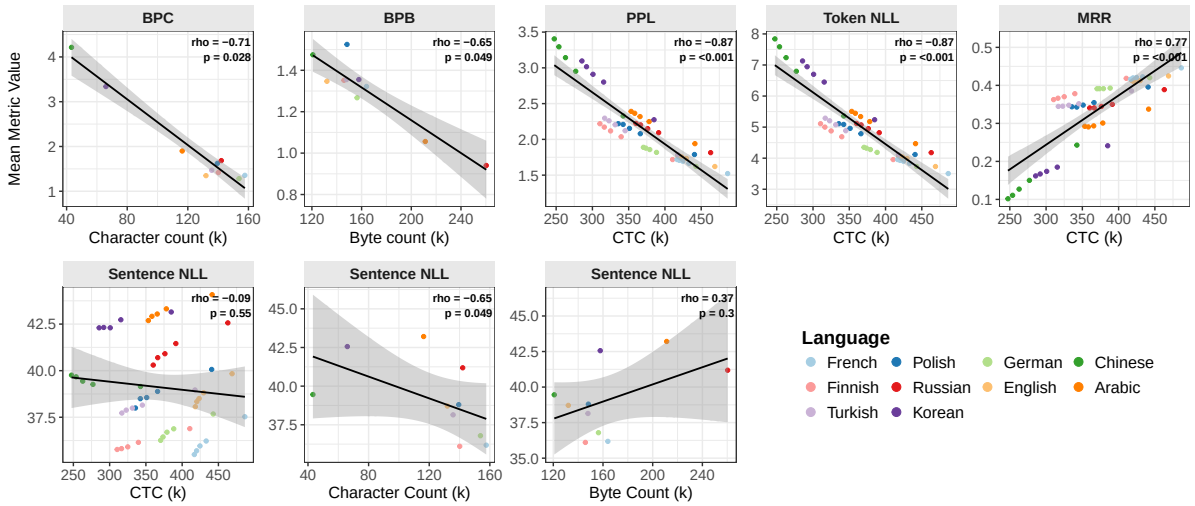

Figure 10: Spearman's Rank Correlation between the intrinsic metrics and their potential confounding factors in Parallel10 (perplexity is scaled by log10).

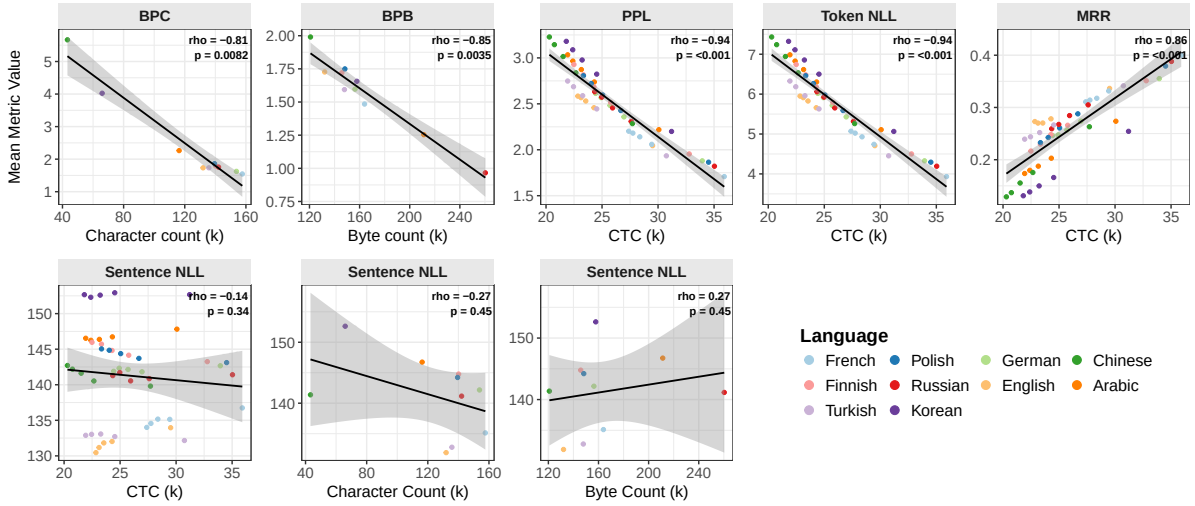

Figure 11: Spearman's Rank Correlation between the intrinsic metrics and their potential confounding factors in PUD (perplexity is scaled by log10).

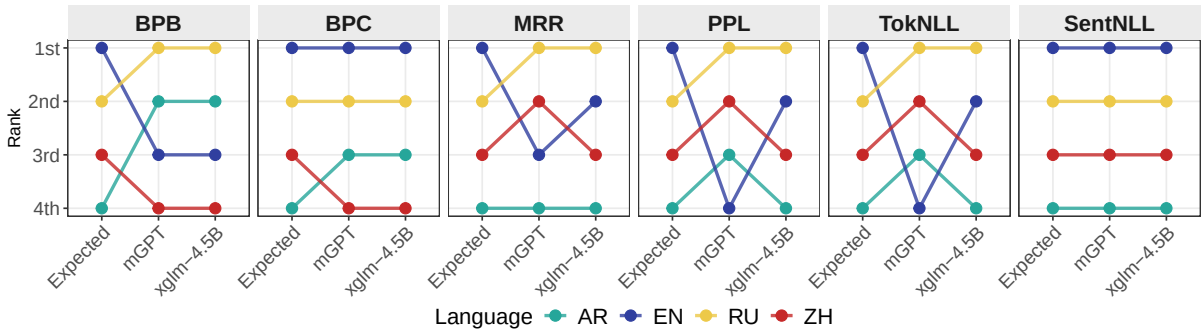

Figure 12: Expected ranking and actual ranking across 6 metrics for `mGPT` and `XGLM-4.5B`.

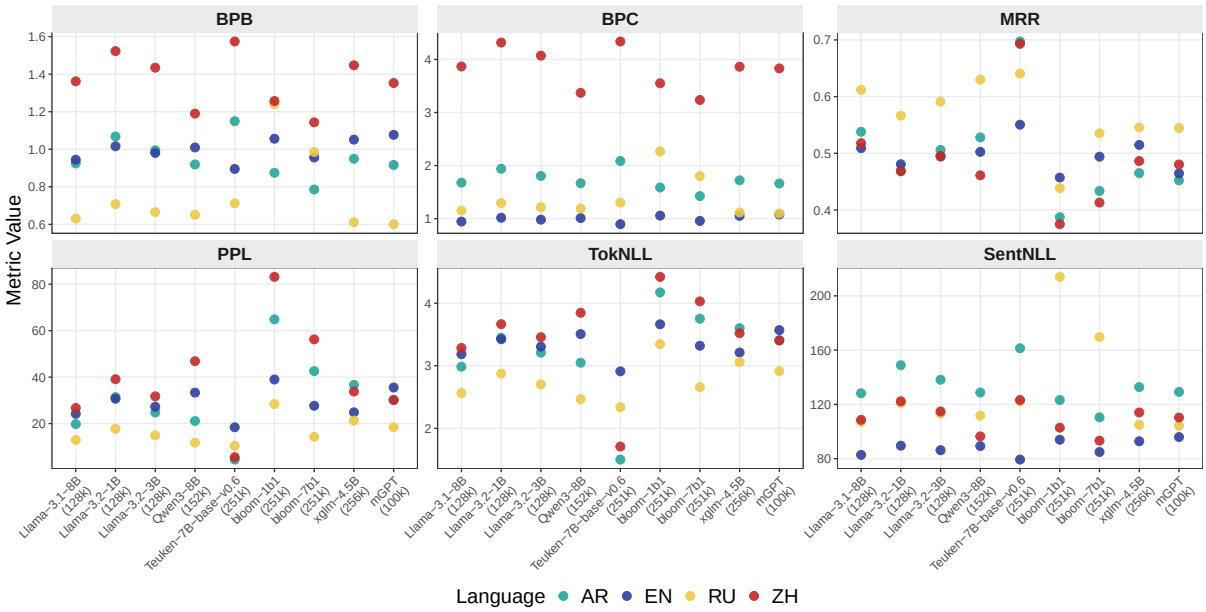

Figure 13: Raw results for all models across six evaluation metrics