Form and Meaning in Intrinsic Multilingual Evaluations

Wessel Poelman and **Miryam de Lhoneux**
L$^{\text{A}}$G$_{\text{O}}$M·NLP, Department of Computer Science, KU Leuven
{firstname.lastname}@kuleuven.be

Abstract

Intrinsic evaluation metrics for conditional language models, such as perplexity or bits-per-character, are widely used in both mono- and multilingual settings. These metrics are rather straightforward to use and compare in monolingual setups, but rest on a number of assumptions in multilingual setups. One such assumption is that comparing the perplexity of CLMs on parallel sentences is indicative of their quality since the *information content* (here understood as *the semantic meaning*) is the same. However, the metrics are inherently measuring *information content* in the *information-theoretic* sense. Consistency in meaning does not neutralize different forms (paraphrases). We make such assumptions explicit and discuss their implications. We perform experiments with six metrics on two multi-parallel corpora both with mono- and multilingual models. We find that current metrics are not universally comparable and look at the form-meaning debate to provide some explanation for this.

1 Introduction

Intrinsic evaluation metrics, such as perplexity (PPL; Jelinek et al., 1977), are often the first step in evaluating conditional language models (CLMs). PPL is the most ubiquitous, but other transformations of the loss (or negative log-likelihood) are also used. Intrinsic evaluations are sometimes the only option for certain languages due to resource availability (Joshi et al., 2020).

There are two multilingual setups where these metrics are used: *a single multilingual model* or *multiple monolingual models*. In the multilingual model setting the intrinsic metrics indicate how well a single model learns multiple languages. Comparing metrics is used for model design: *model A is better than B since A achieves a lower PPL.* For the setting with multiple monolingual models, the metrics are used to study the interaction of language characteristics and language modeling: *which language is hard to model?* The comparisons are used to describe languages: *the PPL for language A is lower than B, A is therefore easier to model.*

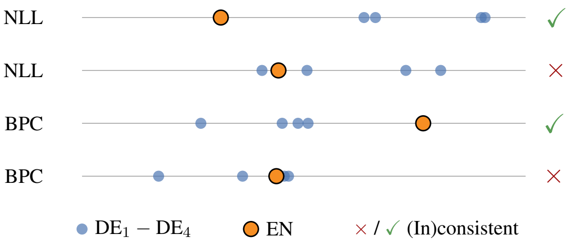

Figure 1 – Each line represents a row in a parallel dataset, each dot is an individual sentence; four German and one English. Common multilingual evaluations center around comparing intrinsic metrics such as the negative log-likelihood (NLL) or bits per character (BPC) with the assumption that these comparisons are fair since the *semantic meaning* is consistent. However, these metrics measure *information content* in the information-theoretic sense. This results in (1) differences within a language ($\text{DE}_i \leftrightarrow \text{DE}_k$), and (2) inconsistency across languages: if the EN sentence falls outside the range of the DE sentences it is consistent. If it falls within the range it is inconsistent, meaning conclusions can flip depending on the DE sentence we choose in our test set.

A common approach in both setups is to calculate the loss (or transformations of it) on multi-parallel datasets (*e.g.*, Cotterell et al., 2018; Gerz et al., 2018; Mielke et al., 2019; Park et al., 2021; Wan, 2022; Chang et al., 2024a,b; Limisiewicz et al., 2024; Arnett and Bergen, 2025). A popular choice is the FLORES-200 machine translation dataset (NLLB Team et al., 2022). Multi-parallel datasets are presumed to allow for a fair evaluation: samples have the same *semantic meaning* in all languages. However, the fairness of comparing metrics using multi-parallel datasets is less clear. CLMs are optimized using an *information-theoretic* objective, namely to minimize the cross-entropy loss, which is explicitly not about semantics. We make such assumptions explicit and ask:

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2503–2521
March 24-29, 2026 ©2026 Association for Computational Linguistics

1. How does the distinction between *semantic meaning* and *information content* apply to intrinsic evaluation metrics?
2. What are the factors at play when comparing metrics, within and across languages?
3. What are assumptions about comparing metrics when using a single multilingual model or multiple monolingual models?

To answer these questions, we first discuss six existing intrinsic metrics and show how they relate to each other. Next, we formalize the problem of comparing intrinsic metrics and relate it to the form-meaning debate. Third, we look at the behavior of the metrics on two parallel corpora for mono- and multilingual models. Fourth, we perform experiments with paraphrases to test the *consistency* of the metrics (Figure 1). Finally, we take a broader view and tie our findings to the larger question of intrinsic CLM evaluations, the form-meaning debate, and linguistic information content.

2 Related Work

Multilingual Language Modeling. A CLM ($\mathcal{M}$) is trained to model the distribution of a token w_i conditioned on its preceding context:

$$\text{NLL}(\mathcal{M}) = -\sum_{i=1}^{S} \log P(w_i | w_1, \ldots, w_{i-1}; \mathcal{M}). \tag{1}$$

The model has to maximize the log-probability, or equivalently, minimize the negative log-likelihood (NLL), of the cross-entropy of the model's distribution and the true distribution of the next token.

As mentioned, multilinguality for CLMs can be tackled from two angles: a multilingual model that includes many languages (*e.g.,* Scao et al., 2022; Üstün et al., 2024; Chang et al., 2024a) or comparing multiple monolingual models (*e.g.,* Mielke et al., 2019; Park et al., 2021; Chang et al., 2024b). Our study is closer to the second, although we discuss implications for both.

Intrinsic Evaluation. Intrinsic metrics for CLMs are based on how well a model predicts tokens in an unseen test sequence; perplexity being the most common. In order to make the evaluation fair, multi-parallel data is used. However, this type of data can have issues, such as translation effects or other artifacts. This has resulted in a number of slight variations or transformations of NLL that attempt to address these issues (*e.g.,* Cotterell et al., 2018; Wan, 2022; Tsvetkov and Kipnis, 2024). We discuss the metrics in detail in §3.1.

Consistency. The ability of language models to equivalently handle paraphrases is known as *consistency* (*e.g.,* Elazar et al., 2021; Ohmer et al., 2024). Recently, Poelman et al. (2025) noted that comparing intrinsic metrics in multilingual settings could be unreliable due to being dependent on the form of parallel data. We repeat their hypothetical example in Table 1.

Model	Sequence	PPL
A	Sabe jugar al ajedrez	20
B	Do you know how to play chess	22
B	Can you play chess	18

Table 1 – Two valid parallel English sentences and a Spanish sentence with hypothetical PPLs. If we select the first parallel English option, model A is "better" than B, and vice-versa for the second.

Linguistic Universals. Instead of intrinsic metrics, we can also sample text from a CLM to evaluate how close it is to human language. Human languages seem to have a distinct "statistical fingerprint".[1] Perhaps the most famous metric related to this fingerprint is Zipf's law (1949): words in a corpus are inversely proportional to their frequency rank. This has been shown to hold across a number of human languages (Piantadosi, 2014). The type-token ratio (or Heaps' law; Herdan, 1960) is another law that characterizes vocabulary growth. Since the true distribution of human language is not known outside of careful experiments (Jumelet and Zuidema, 2023), we have to evaluate models by using test sets or by sampling text from models.

In the context of CLMs, Meister and Cotterell (2021) measure how well generated output from models obeys Zipf's law and other statistical tendencies for English. Takahashi and Tanaka-Ishii (2017, 2019) perform similar studies for English and Chinese. However, the usefulness of linguistic laws when evaluating CLMs is debatable. Zipf's law can be found in many unexpected places (Piantadosi, 2014). Similarly, a model can produce gibberish while fully adhering to either law. For these reasons, we choose to focus on intrinsic evaluation metrics. We come back to the laws in §7.

[1] We refer to Bentz (2023) for recent empirical work and to Sproat (2023) for a broader overview.

3 Metrics

3.1 Intrinsic Metrics

We define an “intrinsic metric” as measuring how well a CLM predicts an unseen test sequence. In Eq. 1, we define how a CLM is trained. Given a trained CLM, we can give it a sequence (S) and get the average loss (NLL; ↓) across the tokens:

$$\text{NLL} = -\frac{1}{S}\sum_{t=1}^{S} \log P(w_t|w_{<t}). \tag{2}$$

Lower values mean the model assigns higher probabilities to the true next tokens. The NLL, either averaged or not, forms the basis for other metrics. *Perplexity (PPL; ↓)*, sometimes referred to as *surprisal*, is the exponential of NLL:

$$\text{PPL} = \exp(\text{NLL}). \tag{3}$$

Depending on the segmentation of the sequence, we can measure the *bits* per segment (e.g., bytes, characters, words). *Bits per Character (BPC; ↓)* adapts cross-entropy to characters:

$$\text{BPC} = -\frac{1}{S}\sum_{t=1}^{S} \log_2 P(c_t|c_{<t}), \tag{4}$$

where c_i is the i-th character in a sequence S. BPC measures how many bits are needed to encode each character. *Bits per English Character* (*BPEC; ↕;* Cotterell et al., 2018) and *Information Parity* (*IP; ↕;* Tsvetkov and Kipnis, 2024) compare BPC across languages. BPEC normalizes a target language's BPC by English's BPC:

$$\text{BPEC} = \frac{\text{BPC}_{\text{Target}}}{\text{BPC}_{\text{EN}}}, \quad \text{IP} = \frac{\text{BPC}_{\text{EN}}}{\text{BPC}_{\text{Target}}}. \tag{5}$$

If BPEC < 1, the model is performing “better” in the target language than English. IP flips this to interpret higher values as better. Finally, *Mean Reciprocal Rank* (*MRR; ↑;* Limisiewicz et al., 2023) evaluates the model's *ranking* of the true token:

$$\text{MRR} = \frac{1}{S}\sum_{t=1}^{S}\frac{1}{R_t}, \tag{6}$$

where R_t is the rank of the correct token in the model's predicted distribution over the vocabulary V (with size $|V|$). Higher MRR means the model places the correct token near the top of its predictions. Intuitively, MRR is a more lenient metric than the NLL-based metrics since it does not matter which probability the model assigns: MRR stays the same if w_t is at the top rank with 0.9 or 0.02.

3.2 Metric Usage

As mentioned, intrinsic metrics are often used with multi-parallel datasets. This keeps the semantic meaning constant across languages. There are a number of reasons for using the metrics in this way.

Park et al. (2021) mention that *“because each [sequence] is intended to express the same meaning across languages, differences in (. . .) surprisal primarily indicate differences in cross-linguistic language model quality.”*

Regarding potential translation issues, Mielke et al. (2019) outline that *“different surprisals on the translations of the same sentence reflect quality differences in the language models, unless the translators added or removed information. (. . .) if we find NLL(A) > NLL(B), we must assume A contains more information, or that our language model was simply able to predict it less well.”*

Tsvetkov and Kipnis (2024) discuss the effect of compression: *[IP] measures how efficiently the LLM represents information provided by a text in the language L compared to the same information provided in English. A higher IP indicates a higher representation efficiency hence a closer alignment with the ideal language-agnostic compressor.*

Compression is also considered by Wan (2022): *[We use] the total number of bits needed to encode the [parallel] dev set [BPC without averaging]. [This is a] more general and flexible way of evaluating data that has not been or cannot be perfectly segmented or aligned line by line.*

These are all valid concerns. However, all metrics use NLL and transform it in some way. Since the dataset is multi-parallel (consistent meaning), a common assumption is that metrics are therefore directly fair to compare. However, this may not hold: consistent meaning does not “neutralize” different forms with the same meaning. The core of training a CLM is an information-theoretic objective, which is not about *meaning content* as we know it from semantics, it is instead about *information content*: the probability of an event occurring (for CLMs, we can say predicting a token is the event, with the vocabulary of the model as the options). Information content in this sense is explicitly not about semantics (Shannon, 1948, 1951). We outline this problem in more detail in the next section.

4 Theoretical Outline

In order to compare the metrics, we first need to know what factors are involved. We outline the tokenization, dataset, and modeling setups we consider. Let $\mathcal{L} = \{L_1, L_2, \ldots, L_N\}$ be a set of N languages in our experiment (and $\mathcal{L}^* \cup \mathcal{L} = \mathcal{L}^\dagger$ are "all human languages"). For each language L_i, we have a *Monolingual Model* ($\mathcal{M}_i^{\text{Mono}}$), which uses a *Monolingual Tokenizer* ($\mathcal{T}_i^{\text{Mono}}$), both trained on a *Monolingual Corpus* ($\mathcal{C}_i^{\text{Mono}}$). The *Multilingual Model* ($\mathcal{M}^{\text{Multi}}$) and *Multilingual Tokenizer* ($\mathcal{T}^{\text{Multi}}$) are trained on $\mathcal{C}_i^{\text{Mono}} \in \mathcal{C}^{\text{Multi}}$. We consider $\mathcal{C}^{\text{Multi}}$ to be a *multi-parallel* corpus: $\mathcal{C}_{i,j}^{\text{Mono}} \leftrightarrow \mathcal{C}_{k,j}^{\text{Mono}}$ are parallel translations of the sample j for L_i and L_k. Multi-parallel means $|\mathcal{C}_i^{\text{Mono}}| = |\mathcal{C}_k^{\text{Mono}}|$ for any pair in $\mathcal{L}$ and that the rows are all consistent in their meaning. We indicate the test split of a corpus as *corpus-prime*: $\mathcal{C}^{\text{Mono'}}$. Each tokenizer-model combination has a vocabulary V. When evaluating any model $\mathcal{M}$, we assume any of the metrics from §3.1 are used; for our running example we use NLL.

Monolingual. Inherently, the corpora $\mathcal{C}_i^{\text{Mono}}$ and $\mathcal{C}_k^{\text{Mono}}$ are not the same. This means the models $\mathcal{M}_i^{\text{Mono}} \neq \mathcal{M}_k^{\text{Mono}}$, regardless of the choice of segmentation. For example, if two languages share an alphabet and we train character-based models,[2] we end up with $\mathcal{T}_i^{\text{Mono}} = \mathcal{T}_k^{\text{Mono}}$ and $V_i = V_k$. Since the corpora are not the same, this means the distribution learned by the models will be different. Often times, segmentations and vocabularies are not the same between languages, leading to $\mathcal{T}_i^{\text{Mono}} \neq \mathcal{T}_k^{\text{Mono}}$. Still, because $\mathcal{C}_i^{\text{Mono}} \neq \mathcal{C}_k^{\text{Mono}}$, we can say for the respective models $\mathcal{M}_i^{\text{Mono}}$ and $\mathcal{M}_k^{\text{Mono}}$ that the distribution they learn is not the same, even if the sequences express the same meaning. All metrics outlined in §3.1 are affected by this in the monolingual setting. This means different distributions are being used and compared, both in training and testing. Whether this is desirable comes down to the question of the existence of a universal statistical fingerprint to any human language (see §2 and §7), whether this distribution can be accessed by any single language, and whether the metrics are the right tool for this.

Multilingual. Describing the distributions for $\mathcal{M}^{\text{Multi}}$ and $\mathcal{T}^{\text{Multi}}$ is more nuanced. Starting from the corpus $\mathcal{C}^{\text{Multi}}$, we can identify how many tokens are overlapping between $\mathcal{C}_i^{\text{Mono}}$ and $\mathcal{C}_k^{\text{Mono}}$ (or, V_i and V_k), regardless of the segmentation choice. For overlapping tokens, we explicitly combine the probability distributions of L_i and L_k in $\mathcal{M}^{\text{Multi}}$. When doing this for all languages ($\mathcal{L}$), we essentially "merge" (parts of) their distributions. Even when disallowing overlapping tokens in V^{Multi}, or assuming $\mathcal{C}_i^{\text{Mono}}$ and $\mathcal{C}_k^{\text{Mono}}$ share no tokens, it is often assumed a multilingual model can learn from this combined or "universal"[3] distribution across languages. This is the basis of *zero-shot* language modeling, where a multilingual model can process one or more $L^* \in \mathcal{L}^*$. In practice, there are many factors involved: writing systems, (cross-lingual) homographs, polysemy, the aforementioned allocation of $\mathcal{L}$ in V^{Multi} or $\mathcal{T}^{\text{Multi}}$, and more.

At least in principle, we can assume $\mathcal{M}^{\text{Multi}}$ is able to process[4] tokens of $\mathcal{L}$. So, with some caveats, we can assume:

$$P(\mathcal{L}) \approx P(\mathcal{L}^\dagger). \tag{7}$$

This means metrics from §3.1 can, to some extent, be meaningfully compared between any pair in $\mathcal{L}$ since the model and tokenizer both have seen the distributions through $\mathcal{C}^{\text{Multi}}$.

Paraphrases. As mentioned, a multi-parallel corpus removes confounds such as corpus size imbalances between languages and ensures consistent meaning across languages. We now ask the question "are all metrics comparable across all languages when we use the multilingual components we have discussed?" For this, we look at the *consistency* of intrinsic metrics across paraphrases. We define $\mathcal{C}_k^{\text{Alt'}}$ as alternatives or paraphrases of $\mathcal{C}_k^{\text{Mono'}}$ for L_k. If we take a parallel sequence from either set and if the metrics are comparable *and* if they measure a "universal" distribution (either in meaning, or in information content), we should see:

$$\text{NLL}(\mathcal{C}_{k,j}^{\text{Mono'}}) \approx \text{NLL}(\mathcal{C}_{k,j}^{\text{Alt'}}). \tag{8}$$

This is *within* a language L_k. There is also the question of comparing *across* languages. A strict interpretation of consistency across languages is that the NLL of parallel sequences has to be similar for any two L_i and L_k:

$$\text{NLL}(\mathcal{C}_{i,j}^{\text{Mono'}}) \approx \text{NLL}(\mathcal{C}_{k,j}^{\text{Mono'}}), \tag{9}$$

but as noted before by several works in §3.2, this adds factors: how good is the language model? Is one language harder to model than another?

[2] So assuming no training is necessary for the tokenizers.

[3] As mentioned, the true distribution of human language is not known. Previous research has referred to multilingual approaches as "universal" (*e.g.,* Yang et al., 2020).

[4] There are many caveats regarding tokenization and the handling of out-of-vocabulary items. We take a broad view since we are mainly talking about the metrics.

We can loosen this interpretation by taking inspiration from MRR and measure the relative *ranking* of sequences between $\mathcal{C}^{\text{Mono'}}_{i,j} \leftrightarrow \mathcal{C}^{\text{Mono'}}_{k,j} \leftrightarrow \mathcal{C}^{\text{Alt'}}_{k,j}$. Differences in individual values due to the causes listed above should not affect the ranking if the metric is at least consistent. For example, assume we have one English sentence and four German sentences, all with the same meaning. If the values for an intrinsic metric for the four German sentences are *all lower* or *all higher* than the English sentence, the ranking of English and German is consistent for that sample. We can test if this holds per sample in the test data (Figure 1). We can also average across the splits (so $\overline{\mathcal{C}^{\text{Mono'}}_{k}}$ and $\overline{\mathcal{C}^{\text{Alt'}}_{k}}$) and see if those are consistent when comparing with $\overline{\mathcal{C}^{\text{Mono'}}_{i}}$. This ranking of averages is the most lenient interpretation of consistency. Coming back to our example: it allows room for German being potentially harder to model or a model being worse at modeling German, as well as some inconsistencies from sample to sample. If the ranking on the level of samples or averages turns out to be inconsistent (e.g., two German sentences score lower and two score higher compared to the English sentence, or the average of the English set is in between two German sets), the interpretation of any comparison of the metric is ambiguous and the question arises of what we are comparing and if this comparison is meaningful.

Recap. To summarize, we outline three potential issues with the intrinsic metrics:

1. We are inherently comparing different distributions with the metrics for any $\mathcal{M}^{\text{Mono}}_{i}$ and $\mathcal{M}^{\text{Mono}}_{k}$, or we have to assume there is a universal distribution to all human languages which we can access using a metric through a single language: $L \in \mathcal{L}^{\dagger}$.

2. We are comparing a shared distribution when using $\mathcal{M}^{\text{Multi}}$ since it has seen the combined distributions of $\mathcal{L}$. If we take zero-shot capabilities of $\mathcal{M}^{\text{Multi}}$ into account, we can more reliably assume we are evaluating some idea of a universal distribution: the larger or the more diverse $\mathcal{L}$ gets,[5] the closer we get to modeling $\mathcal{L}^{\dagger}$.

3. If we assume the metrics consistently measure semantic meaning of parallel sequences (or are impervious to it), we should see roughly the same values for paraphrases within a single language. Similarly, across languages we should also see roughly the same values in a strict interpretation. With a less strict interpretation, the ranking of $\mathcal{C}^{\text{Mono'}}_{i,j} \leftrightarrow \mathcal{C}^{\text{Mono'}}_{k,j} \leftrightarrow \mathcal{C}^{\text{Alt'}}_{k,j}$ should stay consistent, either on a sample-level (Figure 1) or when using averages across splits (Table 3).

5 Method

We perform the following experiments: (1) We train $\mathcal{M}^{\text{Mono}}$ and $\mathcal{M}^{\text{Multi}}$ models and their tokenizers to study the behavior of the metrics. (2) We test paraphrase consistency as outlined in §4.

Data. All corpora we use are multi-parallel ($\mathcal{C}^{\text{Multi}}$). For training tokenizers and models, we use EuroParl (Koehn, 2005) and the UN Parallel Corpus (UNPC, Ziemski et al., 2016). We train $\mathcal{M}^{\text{Mono}}$ and $\mathcal{M}^{\text{Multi}}$ variants. EuroParl covers 21 languages and UNPC six. The sizes of the datasets are smaller than what is commonly used to train state-of-the-art LLMs. However, we specifically need to train and evaluate on multi-parallel data for our experiments, which limits us in our choice of datasets. Full details are in §A.

For evaluation, we use the aforementioned FLORES-200 dataset (NLLB Team et al., 2022). For testing paraphrase consistency, we use translation data from Freitag et al. (2020a,b), which we discuss further in §6.

Models and Tokenization. We aim to keep the tokenization and model choices as "mainstream" as possible. We train $\mathcal{T}^{\text{Mono}}$ and $\mathcal{T}^{\text{Multi}}$ variants on the same training data as the models. We use BPE (Gage, 1994; Sennrich et al., 2016) with byte-level fallback. All $\mathcal{T}^{\text{Mono}}$ receive a vocabulary size of 32k and $\mathcal{T}^{\text{Multi}}$ 150k, based on previous estimates (*e.g.,* Conneau et al., 2020; Xue et al., 2021; Üstün et al., 2024). We train a number of tiny Llama-3-style (Grattafiori et al., 2024) models using Huggingface `transformers` (Wolf et al., 2020). We apply insights about training tiny models from the MobileLLM project (Liu et al., 2024). Small models are sometimes deemed unreliable since some phenomena only appear at larger scales. However, small models are suitable for specific investigations (*e.g.,* Chang et al., 2024a; Tatariya et al., 2025; Wilcox et al., 2025), such as ours. While the absolute values for the intrinsic metrics might

[5] Multilingual generalizability comes either from a *large* set of languages, or from a *representative* set (Bender, 2011; Ploeger et al., 2024).

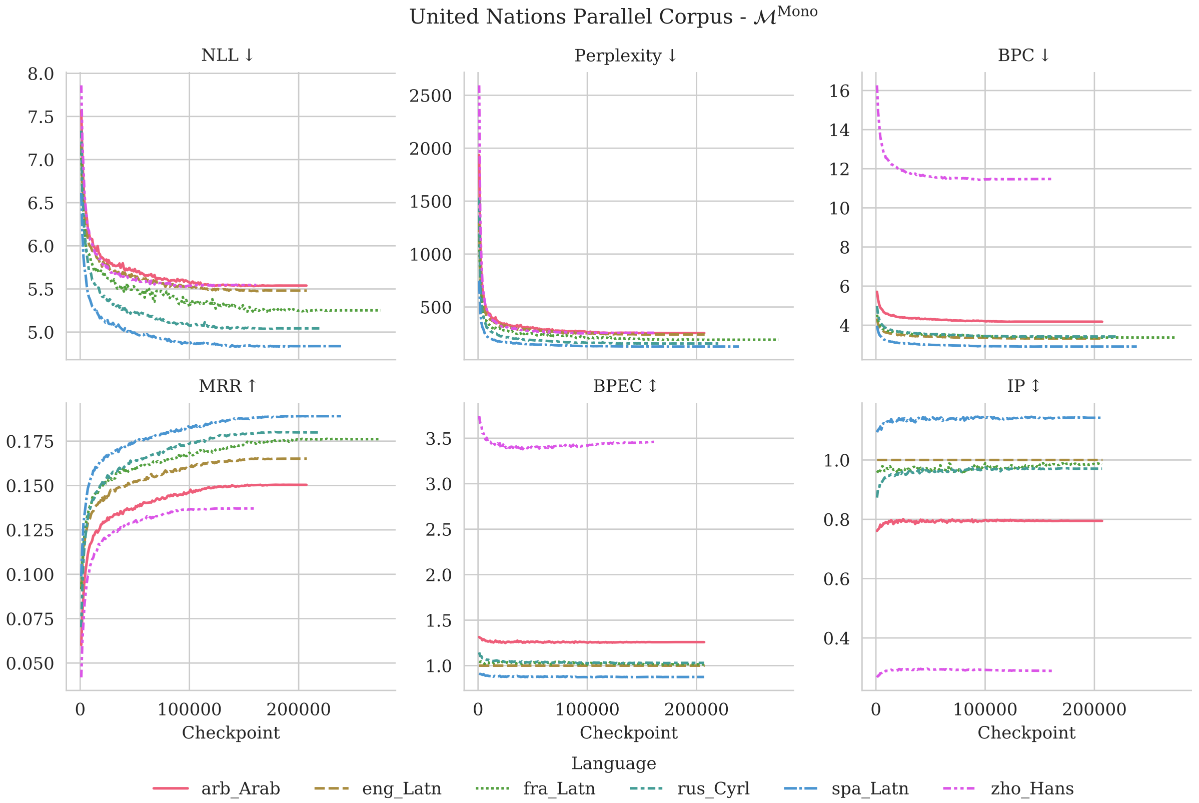

Figure 2 – Metrics across checkpoints during training when evaluated on FLORES-200. The different lengths of the lines is due to tokenization differences between languages, resulting in shorter or longer sequences. The models have seen the same amount of parallel data when measured in number of lines.

change, the potential inconsistency of the comparisons will not suddenly disappear at larger scales. Additionally, large models do not add much when our datasets are this small. Full training details are listed in §B.

Metrics. The metrics are calculated for all sentences in FLORES-200 for a specific language for $\mathcal{M}^{\text{Mono}}$ and for all supported languages ($\mathcal{L}$) for $\mathcal{M}^{\text{Multi}}$. We do this every 1000 steps. For BPEC and IP, we need the English BPC to scale the BPC of other languages with. For $\mathcal{M}^{\text{Mono}}$, we first calculate BPC for $\mathcal{M}^{\text{Mono}}_{\text{EN}}$, which we use for calculating BPEC and IP for other $\mathcal{M}^{\text{Mono}}_{i}$. For $\mathcal{M}^{\text{Multi}}$ we also first calculate BPC for English, but the BPC for other languages comes from the same $\mathcal{M}^{\text{Multi}}$.

6 Results

Figure 2 shows the behavior of the metrics when using monolingual models trained on UNPC. We can see English, Arabic, and Chinese have similar NLL values, all being quite high (lower is better). Perplexity is a linear transformation of NLL, so it does not show anything new. BPC shows Chinese as a clear outlier, likely due to tokenization and script differences. Interestingly, Russian is similar to English and French. MRR shows a similar ranking as NLL (albeit inverse; higher is better). However, the difference between English, Arabic, and Chinese is consistent, whereas these three traded places for NLL. BPEC and IP are linear transformations of BPC, so while the patterns show nothing new, their interpretations is potentially interesting: *Chinese is more than three times "harder to model" than English.* This interpretation is meant as an example, not as a conclusion (see the following section).

The multilingual results are similar (Figure 4), but the NLL is noticeably higher for all languages, likely due the reasons listed in §3.2. All languages using a Latin script are performing the "best" for $\mathcal{M}^{\text{Multi}}$, whereas Russian is the second best for $\mathcal{M}^{\text{Mono}}$. This could be due to the idea of a shared distribution outlined in §3.1. The question of whether we are evaluating how well this particular set of technical decisions is suited to model these languages or how difficult these languages are to model is not straightforward (see §7).

Finally, the EuroParl results (§C.1) are similar to the UNPC results. One interesting aspect of EuroParl is that all 21 languages in the dataset except Bulgarian and Greek use the Latin script. However, their values for the metrics do not stand out like Chinese or Arabic do for UNPC.

Paraphrase Consistency. As shown in the previous section, it is tempting to draw conclusions about a language as a whole when looking at the results. Regardless of whether the metrics measure information content or semantics, are they consistent when presented with paraphrases? To test this, we use the EN→DE set of Freitag et al. (2020a,b). Every source sequence has *four* human translated references. This results in four parallel splits:

1. The original reference translation (DE_1).
2. An alternative reference translation (DE_2).
3. A "paraphrased as-much-as-possible" version of the original reference (DE_3).
4. A "paraphrased as-much-as-possible" version of the alternative reference (DE_4).

All four German sentence and the English sentence convey the same semantic meaning.[6] The dataset is similar in domain and size to FLORES-200: high quality news articles and about 2000 samples. We use $\mathcal{M}^{\text{Mono}}_{\text{EN}}$ and $\mathcal{M}^{\text{Mono}}_{\text{DE}}$, as well as $\mathcal{M}^{\text{Multi}}$ trained on EuroParl. Additionally, we track the ranking of the source and four targets as a more lenient alternative to having similar NLL values. Table 2 summarizes the consistency of the metrics. We can see that all metrics are highly inconsistent.

Metric	$\mathcal{M}^{\text{Mono}}$	$\mathcal{M}^{\text{Multi}}$
NLL	47%	51%
BPC	50%	52%
MRR	50%	51%

Table 2 – Sample-level inconsistency. For instance: 47% of all samples are inconsistent for NLL with $\mathcal{M}^{\text{Mono}}$. Figure 1 shows an intuitive explanation.

Figure 3 shows the NLL consistency of $\mathcal{M}^{\text{Mono}}_{\text{EN}}$ and $\mathcal{M}^{\text{Mono}}_{\text{DE}}$ EuroParl models. The two issues outlined before seem to be correct: the difference *within* a language is notable (the light red dots should be close together vertically if the NLL values were similar), and the comparison *across* languages is ranked consistently for only about half the samples.

Model	Metric	DE^{F}	EN	DE^{S}
$\mathcal{M}^{\text{Mono}}$	NLL ↓	6.43–6.64 ✓	6.91	5.89–7.23 ×
	BPC ↓	2.09–2.20 ✓	2.27	1.82–2.44 ×
	MRR ↑	19.6–21.2 ✓	18.6	14.6–26.0 ×
$\mathcal{M}^{\text{Multi}}$	NLL ↓	6.81–7.02 ✓	7.05	6.25–7.60 ×
	BPC ↓	2.40–2.51 ✓	2.54	2.13–2.77 ×
	MRR ↑	22.5–23.7 ✓	21.4	17.4–28.7 ×

Table 3 – Split-level averages. The original splits from Freitag et al. (2020a,b) are in DE^{F} and the row-wise sorted splits in DE^{S}. We show the maximum and minimum of the four splits since this is enough to tell us whether the averages are consistent or not: if EN falls outside the DE split range, it is consistent, if it falls within, it is inconsistent. *All* metrics are consistent for the DE^{F} subset and *all* become inconsistent for DE^{S}.

A logical question now is *does it average out?* So far we kept the splits (columns) from Freitag et al. (2020a,b), and with these splits the averages across DE_1, DE_2, DE_3, and DE_4 are consistent for all metrics. However, since the *rows* are all parallel in meaning (e.g., the sample with index 8 has the same meaning in all splits: $\text{DE}_{1,8} \leftrightarrow \text{DE}_{2,8} \leftrightarrow \text{DE}_{3,8} \leftrightarrow \text{DE}_{4,8}$), there is no reason we cannot make new splits, as long as we do not change the ordering of the rows. If we sort the German samples row-wise across the four splits, we end up with "easiest" to "hardest" splits. The averages of these new splits are *not* consistent. The original and sorted splits and are shown in Table 3.

This has practical implications. If we were to do a study on the difficulty of modeling German versus English, and pick the "easiest" split from the sorted splits, we would conclude that German is easier to model than English. If we were to pick the "hardest" split, we would arrive at the *opposite* conclusion. This seems to confirm the hypothetical example by Poelman et al. (2025) in Table 1.

7 Discussion

Our theoretical analysis and empirical results raise questions about the comparability of intrinsic metrics in multilingual language modeling. We cover some interpretations in this section.

Technical Decisions or Languages. An unsolved question in multilingual NLP is which set of technical decisions results in the "best" language model in terms of equal performance across languages. On the one hand are the modeling decisions, ranging from tokenization to hyperparameters. On the other hand are intrinsic differences

[6] The caveats mentioned in §3.2 about adding or removing information in translations remain. Although this should not matter since we use this dataset in the exact same way FLORES-200 and other parallel datasets are commonly used.

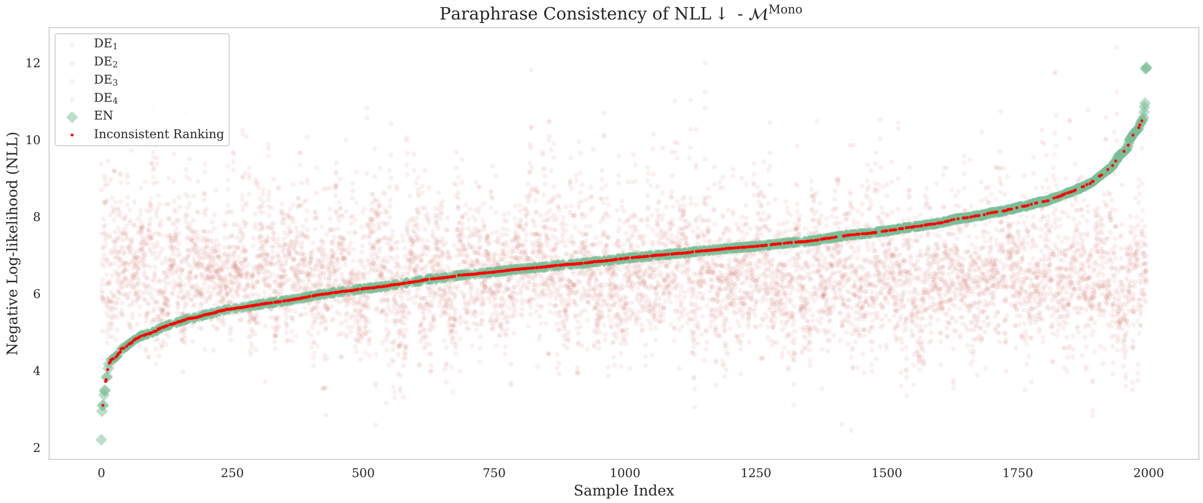

Figure 3 – Paraphrase consistency of NLL values for the EN source and four parallel DE paraphrases using monolingual EuroParl models. On the x-axis, we list the sample index. Each sample consists of the EN sentence (green) and the four DE paraphrases (light red). For the sake of visual clarity, we sort the results by the English NLL. We show the ranking consistency: if all DE paraphrases are above or below the EN source, it means the ranking is consistent. Inconsistent rankings are marked with a red dot inside the green diamond for English. We see similar results for the multilingual model, as well as for the other metrics, see §C.2.

between written languages, ranging from writing systems to corpus-specific characteristics.

Our findings (and similar studies) are limited by the technical decisions made and languages used. It could be that there exists a set of technical decisions and languages (or even corpora) that *do* show consistent values for the metrics and that *are* robust towards paraphrases. However, our requirement for multi-parallel corpora for both training and evaluation restricts us to the data and languages we used. In this controlled setting the issue of consistency and comparability is apparent. To summarize:

- Model-to-model comparisons are fair if the same test set and the same segmentation are used. Different segmentations are potentially fair if scores are scaled to a shared unit such as characters (Mielke, 2019; Bauwens, 2024).

- Language-to-language comparisons come with a number of assumptions and issues. (1) Is there a "universal" human language distribution and can monolingual and multilingual models access it? If so, how reliable are the metrics to measure this? (2) Parallel semantic meaning in datasets does not "neutralize" the effect of form, both within and across languages. Intrinsic metrics are inherently sensitive to this. (3) Inconsistency of metrics can lead to opposite conclusions, depending on the choice of test samples or splits, even with multi-parallel datasets.

Form and Meaning. The distinction between form and meaning is the basis of a number of analyses in various fields (*e.g.,* Frege, 1892; Quine, 1960; Searle, 1980; Harnad, 1990). The "octopus paper" by Bender and Koller (2020) is probably the most pertinent discussion of the topic in the context of language modeling. Their main argument is that a language model inherently only learns *form* and not *meaning*. In some way this is a revisit of the Chinese Room argument (Searle, 1980) with more focus on discourse, grounding, and intentionality.

We do not have to go that far, our findings and argument are purely about the usage and assumptions about intrinsic metrics in multilingual evaluations. Similar empirical results have been found by Ohmer et al. (2024) who find inconsistency in the answers of LLMs to paraphrased questions. Our argument is based on the distinction of information content in the information-theoretic or semantic meaning sense. The metrics (and CLM training) operate on the former, even when we use data that is parallel for the latter. Having parallel data does not neutralize different forms and the metrics, and thus the conclusions we draw based on them, are inherently sensitive to this. However, as mentioned, it is certainly possible there exists a set of languages and technical decisions that result in consistent metrics. This shows the metrics might be unsuitable for questions like *which language is easier to model?* We may need to look elsewhere. Sproat et al. (2014) propose an external approach:

> *"Which languages convey the most information in a given amount of space? This is a question often (...) asked by engineers who have some information theoretic measure of 'information' in mind, but rarely define how they would measure that information. If one had a database of close translations between a set of typologically diverse languages, with detailed marking of morphosyntactic and morphosemantic features, one could hope to quantify differences [in] information."*

This database was unfortunately never fully completed it seems. How we would use such a database for evaluating CLMs is not obvious, but issues regarding form, meaning, and language characteristics would largely be "solved".[7]

Another area to look for alternatives to the intrinsic metrics are the statistical linguistic laws. Even though the ones listed in §2 might not be that useful, if we could formulate a law that accounts for paraphrases or semantics, it would solve the problem. How to do this is another question entirely, one that we do not have an answer to. Until then, the best we can do is to keep paraphrases and potential (in)consistencies in mind.

8 Conclusion

We investigate intrinsic evaluation metrics for multilingual conditional language models. Such evaluations are often done with multi-parallel datasets, where the *meaning* of samples is consistent across languages. However, the metrics are designed to measure *information content*. We (1) introduce existing intrinsic metrics, (2) formalize the problem of comparing the metrics in mono- and multilingual setups, (3) look at the behavior of the metrics on two parallel corpora for both setups, (4) perform experiments with *paraphrases* to test the *consistency* of the metrics, and (5) discuss what metric we might be looking for instead of existing ones. We ultimately find that using intrinsic metrics to comparing languages requires some strong assumptions.

Limitations

Generation Quality. Intrinsic metrics can be useful, but they are not perfect. Lower perplexity values have been shown to sometimes correlate with more unnatural text (Kuribayashi et al., 2021). Scoring well on these metrics does not necessarily mean that a model is of higher quality. We discuss the issue of *comparability*, the question of *quality* is outside the scope of our work.

Language Interactions. Language distributions are not as cleanly distinct as they are generally treated in NLP and (to some extent) our current study. Distributions of many languages are inherently mixed due to loan words, colexification, homographs, code-mixing, and so on. This is even more prevalent with web-based datasets that can contain "unnatural" language mixing due to problems with data quality (Blevins and Zettlemoyer, 2022) or with entirely new ways of mixing languages in prompt-based evaluations of language models (Poelman and de Lhoneux, 2025). Strictly dividing languages is necessary for furthering our understanding of language modeling, but caveats remain.

Acknowledgements

We thank Esther Ploeger for pointing us to the WMT paraphrase dataset, Thomas Bauwens for early discussions, and Coleman Haley for the valuable suggestion of looking into the question of *does it average out?* WP is funded by a KU Leuven Bijzonder Onderzoeksfonds C1 project with reference C14/23/096. The computational resources and services used were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government - department EWI.

References

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The Parallel Meaning Bank: Towards a Multilingual Corpus of Translations Annotated with Compositional Meaning Representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 242–247. Association for Computational Linguistics.

Catherine Arnett and Benjamin Bergen. 2025. Why do language models perform worse for morphologically

[7] Other cross-lingual meaning representations such as UMR or DRS could be a way to test this (Van Gysel et al., 2021; Abzianidze et al., 2017), but at the time of writing, these do not have the coverage of a dataset like FLORES-200.

complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6607–6623. Association for Computational Linguistics.

Thomas Bauwens. 2024. Bits-per-character and its relation to perplexity.

Emily M. Bender. 2011. On Achieving and Evaluating Language-Independence in NLP. *Linguistic Issues in Language Technology*, 6.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198. Association for Computational Linguistics.

Christian Bentz. 2023. The Zipfian Challenge: Learning the statistical fingerprint of natural languages. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 27–37. Association for Computational Linguistics.

Terra Blevins and Luke Zettlemoyer. 2022. Language Contamination Helps Explains the Cross-lingual Capabilities of English Pretrained Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574. Association for Computational Linguistics.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Ben Bergen. 2024a. When Is Multilinguality a Curse? Language Modeling for 250 High- and Low-Resource Languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096. Association for Computational Linguistics.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2024b. Goldfish: Monolingual Language Models for 350 Languages.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and Improving Consistency in Pretrained Language Models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Gottlob Frege. 1892. Über Sinn Und Bedeutung [On Sense and Meaning]. *Zeitschrift für Philosophie Und Philosophische Kritik*, 100(1):25–50.

Markus Freitag, George Foster, David Grangier, Colin Cherry, et al. 2020a. Human-Paraphrased References Improve Neural Machine Translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1183–1192. 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. Association for Computational Linguistics.

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38.

Daniela Gerz, Ivan Vulić, Edoardo Ponti, Jason Naradowsky, Roi Reichart, and Anna Korhonen. 2018. Language Modeling for Morphologically Rich Languages: Character-Aware Modeling for Word-Level Prediction. *Transactions of the Association for Computational Linguistics*, 6:451–465.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, et al. 2024. The Llama 3 Herd of Models.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346.

Gustav Herdan. 1960. *Type-Token Mathematics*. Mouton.

F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1).

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. Association for Computational Linguistics.

Jaap Jumelet and Willem Zuidema. 2023. Transparency at the Source: Evaluating and Interpreting Language Models With Access to the True Distribution. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4354–4369. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86.

Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower Perplexity is Not Always Human-Like. 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 5203–5217. Association for Computational Linguistics.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, et al. 2021. Datasets: A Community Library for Natural Language Processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184. 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. 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. Association for Computational Linguistics.

Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong, Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, Liangzhen Lai, and Vikas Chandra. 2024. MobileLLM: Optimizing Sub-billion Parameter Language Models for On-Device Use Cases. In *Proceedings of the 41st International Conference on Machine Learning*, pages 32431–32454. PMLR.

Clara Meister and Ryan Cotterell. 2021. Language Model Evaluation Beyond Perplexity. 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 5328–5339. Association for Computational Linguistics.

Sabrina J. Mielke. 2019. Can you compare perplexity across different segmentations?

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. Association for Computational Linguistics.

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, et al. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation.

Xenia Ohmer, Elia Bruni, and Dieuwke Hupke. 2024. From Form(s) to Meaning: Probing the Semantic Depths of Language Models Using Multisense Consistency. *Computational Linguistics*, 50(4):1507–1556.

Hyunji Hayley Park, Katherine J. Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2021. Morphology Matters: A Multilingual Language Modeling Analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276.

Steven T. Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21(5):1112–1130.

Esther Ploeger, Wessel Poelman, Miryam de Lhoneux, and Johannes Bjerva. 2024. What is "Typological Diversity" in NLP? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5681–5700. Association for Computational Linguistics.

Wessel Poelman, Thomas Bauwens, and Miryam de Lhoneux. 2025. Confounding Factors in Relating Model Performance to Morphology. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7273–7298. Association for Computational Linguistics.

Wessel Poelman and Miryam de Lhoneux. 2025. The Roles of English in Evaluating Multilingual Language Models. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 492–498. University of Tartu Library.

Willard Van Orman Quine. 1960. *Word and Object [2013 Reissue]*. The MIT Press.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, et al. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.

John R. Searle. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–424.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Claude Elwood Shannon. 1951. Prediction and entropy of printed English. *The Bell System Technical Journal*, 30(1):50–64.

Richard Sproat. 2023. *Symbols: An Evolutionary History from the Stone Age to the Future*. Springer Nature Switzerland.

Richard Sproat, Bruno Cartoni, HyunJeong Choe, David Huynh, Linne Ha, Ravindran Rajakumar, Evelyn Wenzel-Grondie, et al. 2014. A Database for Measuring Linguistic Information Content. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 967–974. European Language Resources Association (ELRA).

Shuntaro Takahashi and Kumiko Tanaka-Ishii. 2017. Do neural nets learn statistical laws behind natural language? *PLOS ONE*, 12(12):e0189326.

Shuntaro Takahashi and Kumiko Tanaka-Ishii. 2019. Evaluating Computational Language Models with Scaling Properties of Natural Language. *Computational Linguistics*, 45(3):481–513.

Kushal Tatariya, Artur Kulmizev, Wessel Poelman, Esther Ploeger, Marcel Bollmann, Johannes Bjerva, Jiaming Luo, Heather Lent, and Miryam de Lhoneux. 2025. How Good is Your Wikipedia? Auditing Data Quality for Low-resource and Multilingual NLP.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218. European Language Resources Association (ELRA).

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. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, et al. 2024. Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939. Association for Computational Linguistics.

Jens E. L. Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, Jan Hajič, James H. Martin, Stephan Oepen, Martha Palmer, James Pustejovsky, et al. 2021. Designing a Uniform Meaning Representation for Natural Language Processing. *KI - Künstliche Intelligenz*, 35(3):343–360.

Ada Wan. 2022. Fairness in Representation for Multilingual NLP: Insights from Controlled Experiments on Conditional Language Modeling. In *International Conference on Learning Representations*.

Ethan Gotlieb Wilcox, Michael Y. Hu, Aaron Mueller, Alex Warstadt, Leshem Choshen, Chengxu Zhuang, Adina Williams, Ryan Cotterell, and Tal Linzen. 2025. Bigger is not always better: The importance of human-scale language modeling for psycholinguistics. *Journal of Memory and Language*, 144:104650.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, et al. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498. Association for Computational Linguistics.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual Universal Sentence Encoder for Semantic Retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94. Association for 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. European Language Resources Association (ELRA).

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press.

A Datasets

A.1 Pre-processing

The training datasets (UNPC and EuroParl) are aligned in a multi-parallel way. We first determine the language with the biggest overlap with all other languages; this is our pivot language (Arabic for UNPC and Italian for EuroParl). Afterwards, we collect sequences that align with the pivot, resulting in our final datasets. The test datasets (FLORES+ and WMT-19 paraphrases) are already multi-parallel for the languages we use, so did not require special pre-processing. FLORES+ is a community effort to keep improving the FLORES test set. We combine the `dev` and `devtest` splits from FLORES+ and use them as our test set. Our development set consists of 10% randomly sampled lines per language. For monolingual models, this is just one language, for multilingual models, it depends on the languages included in the specific corpus. Table 4 describes the datasets and Table 5 lists their language coverage.

Dataset	Link	$\|\mathcal{L}\|$	$\|\mathcal{C}\|$
EuroParl (Koehn, 2005)	HF	21	211521
United Nations Parallel Corpus (Ziemski et al., 2016)	HF	6	11290186
WMT-19 Paraphrases (Freitag et al., 2020a,b)	GH	2	1997
FLORES-200 (FLORES+) (NLLB Team et al., 2022)	HF	221*	2009

Table 4 – Datasets used in our analyses. EP and UNPC are taken from the OPUS (Tiedemann, 2012) collection on the Huggingface hub (Lhoest et al., 2021). The number of languages is listed, as well as the number of samples (rows) after multi-parallel alignment. *The listed number is unique language-script combinations, not languages.

Dataset	Languages (ISO-639-3)
EuroParl	bul, ces, dan, deu, ell, eng, est, fin, fra, hun, ita, lvs, lit, nld, pol, por, ron, slk, slv, spa, swe
UNPC	arb, eng, fra, rus, spa, zho
WMT-19 Paraphrases	deu, eng
FLORES+	Coverage for all languages in listed above.

Table 5 – Language coverage of the datasets we use. Note that Latvian is listed as `lat` (Latin) in FLORES+, while it should be either `lav` (inclusive code) or `lvs` (Standard Latvian). We use the latter.

B Model Setup and Hyperparameters

Setting	Value
Architecture	LlamaForCausalLM
Attention Bias	False
BOS Token ID	1
EOS Token ID	2
Hidden Act	`silu` (swish)
Hidden Size	576
Initializer Range	0.02
Intermediate Size	1536
Max Position Embeddings	2048
Model Type	`llama`
Attention Heads	9
Hidden Layers	30
Key-Value Heads	3
Pretraining TP	1
RMS Norm ϵ	1×10^{-5}
ROPE Scaling	False
ROPE θ	10000
Tie Word Embeddings	False
Torch Data Type	bfloat16

(a) Model architecture.

Parameter	Value
Epochs	3
Learning Rate	3×10^{-4}
LR Warmup Steps	2
LR Warmup Style	linear
LR Decay Style	cosine
Minimum Decay LR	1×10^{-5}
Zero Stage	0
Weight Decay	0.01
Clip Grad	1.0
Accumulate Grad in FP32	True
Optimizer	AdamW
Adam ϵ	1×10^{-8}
Adam β_1	0.9
Adam β_2	0.95

Setting	$\mathcal{M}^{\textbf{Mono}}$	$\mathcal{M}^{\textbf{Multi}}$
Vocabulary Size	32k	150k
Parameters	$\approx$ 143M	$\approx$ 279M

(b) Training hyperparameters (top) and differences between monolingual and multilingual models (bottom).

Table 6 – Model architecture and hyperparameters partially based on Liu et al. (2024)'s 125M Llama-style model.

Step	Hardware	Cost (hours)
Dataset alignment	CPU	10 total
Tokenizer training	CPU	4 total
Monolingual training	H100 GPU	Less than 2 hours per model (27 models $\approx$ 50)
Multilingual training	H100 GPU	About 12 for UN-PC and 18 for EuroParl

Table 7 – Rough breakdown of compute used in the study. Estimates are somewhat exaggerated to cover experimentation and debugging.

C Full Results

C.1 Language Modeling Experiments

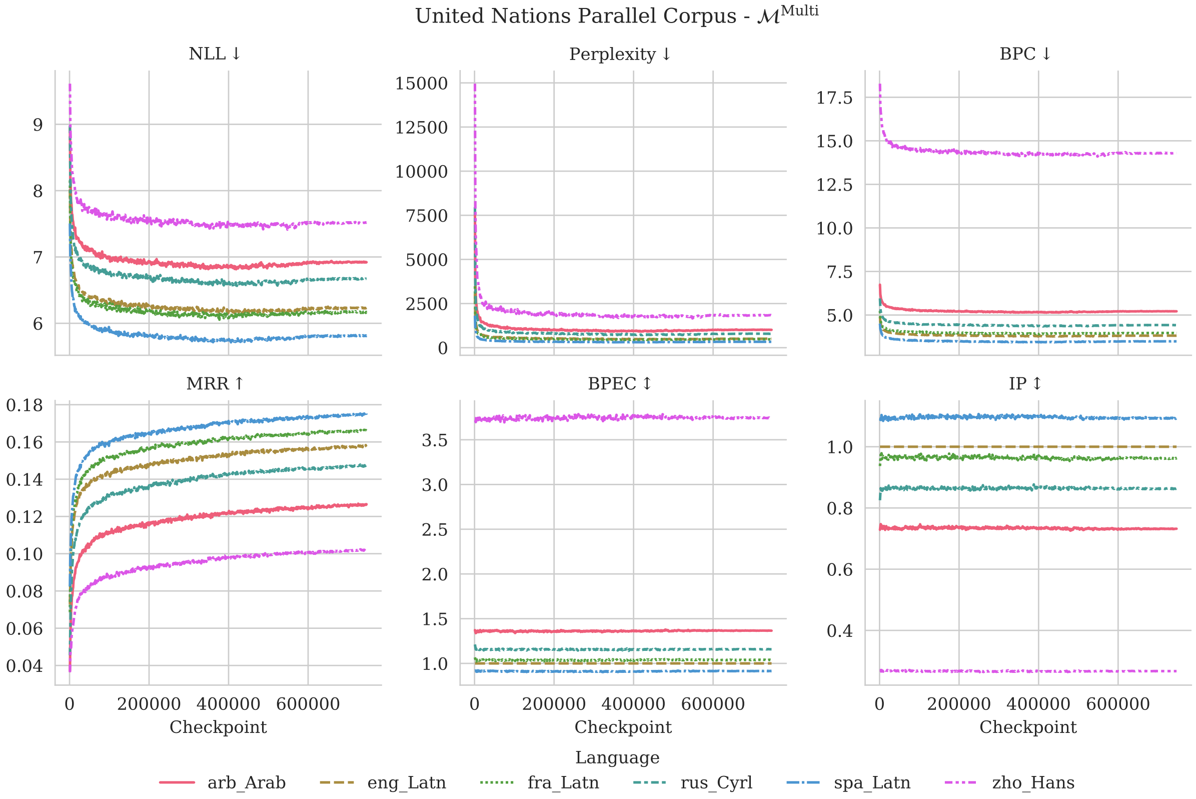

Figure 4 – Metrics across checkpoints evaluated using FLORES-200 for the multilingual UNPC model.

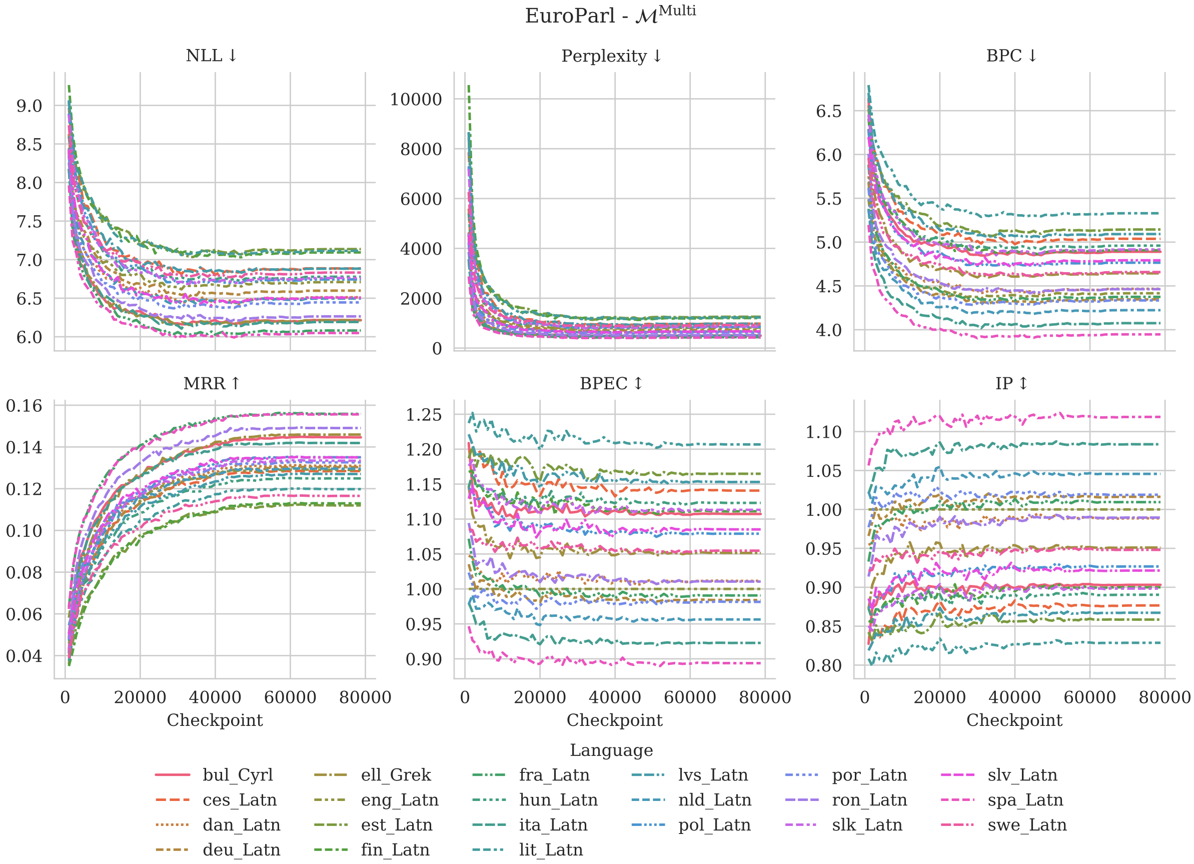

Figure 5 – Metrics across checkpoints evaluated using FLORES-200 for the multilingual EP model.

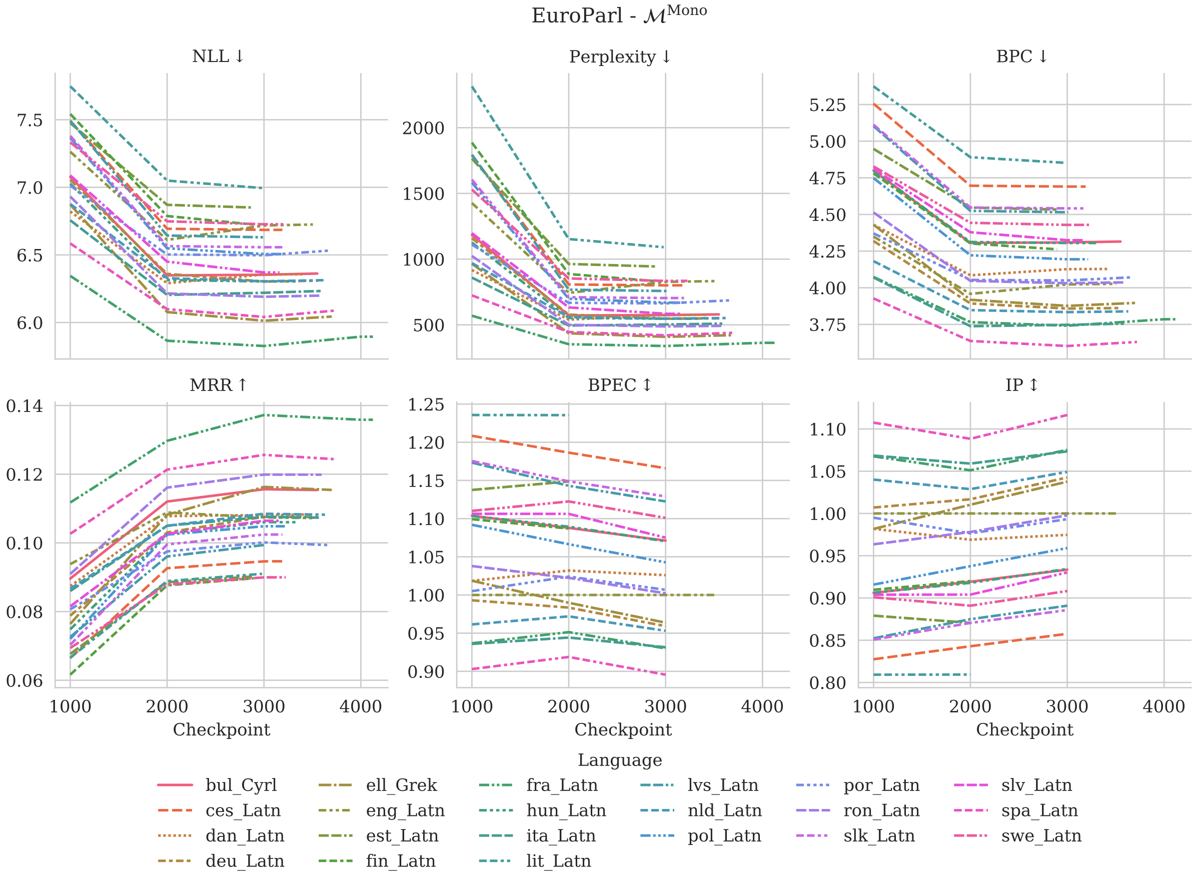

Figure 6 – Metrics across checkpoints evaluated using FLORES-200 for the monolingual EuroParl models.

C.2 Metric Consistency Experiments

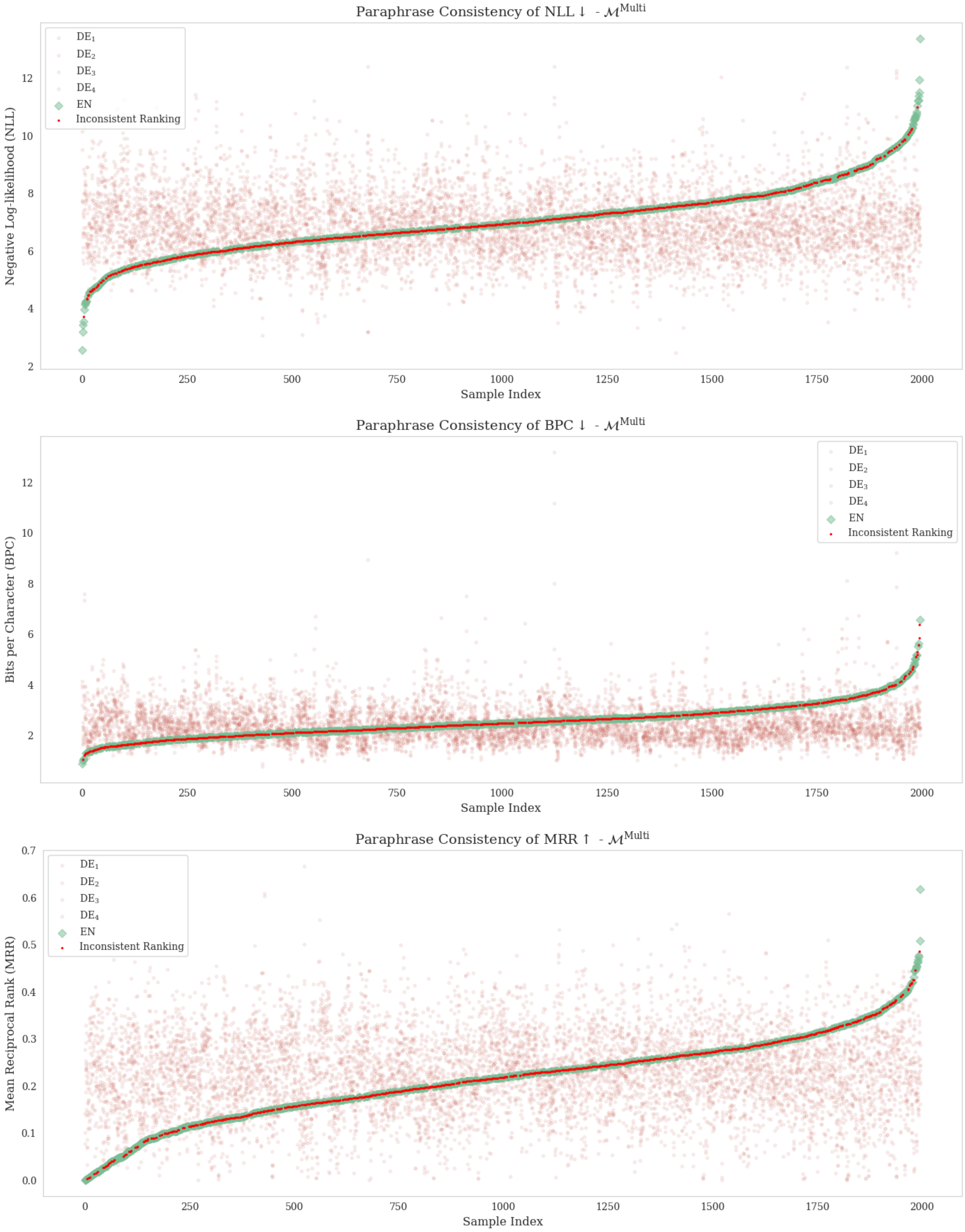

Figure 7 – Sensitivity of NLL, BPC, and MRR using $\mathcal{M}^{\text{Multi}}$ trained on EuroParl.

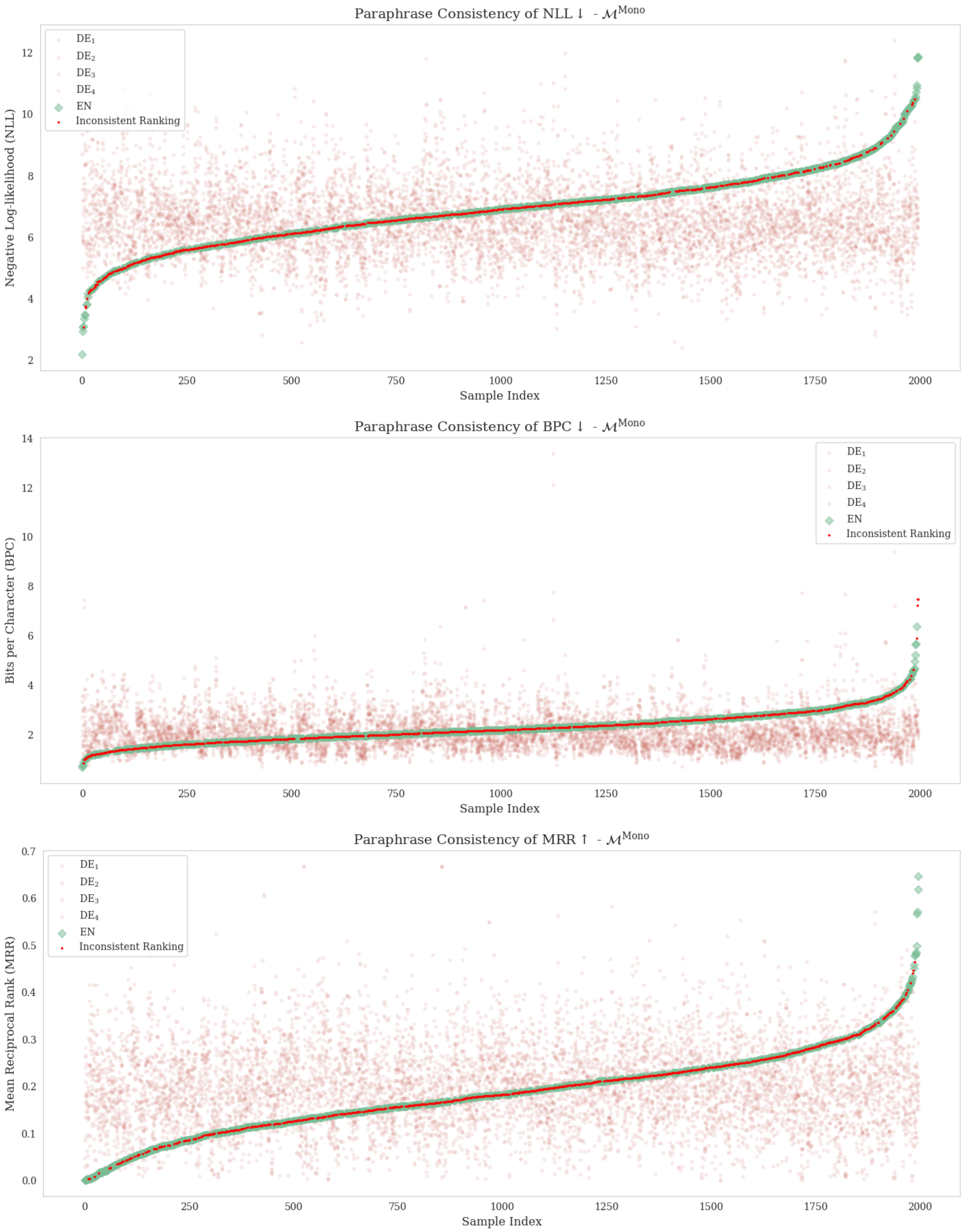

Figure 8 – Consistency of NLL, BPC, and MRR using $\mathcal{M}_{\text{DE}}^{\text{Mono}}$ and $\mathcal{M}_{\text{EN}}^{\text{Mono}}$ trained on EuroParl.