Information Parity: Measuring and Predicting the Multilingual Capabilities of Language Models

Alexander Tsvetkov and **Alon Kipnis**
Department of Computer Science, Reichman University, Herzlia, Israel

Abstract

Large Language Models (LLMs) are increasingly deployed in user-facing applications worldwide, necessitating handling multiple languages across various tasks. We propose a metric called Information Parity (IP) that can predict an LLM's capabilities across multiple languages in a task-agnostic manner. IP is well-motivated from an information theoretic perspective: it is associated with the LLM's efficiency of compressing the text in a given language compared to a reference language. We evaluate IP and other popular metrics such as Tokenization Parity (TP) and Tokenizer Fertility (TF) on several variants of open-sourced LLMs (Llama2, Gemma, Mistral). Among all metrics known to us, IP is better correlated with existing task-specific benchmark scores from the literature and thus better predicts such scores in a certain language. These findings show that IP may be useful for ranking multilingual LLMs' capabilities regardless of the downstream task.

Benchmark/Metric	IP (Ours)	TP
MMLU	**0.95**	0.83
ARC	**0.91**	0.74
HellaSwag	**0.89**	0.75

Table 1: Average absolute Pearson correlation of Information Parity (IP) and Tokenization Parity (TP) metrics with multilingual benchmarks performance. Metrics were computed on Flores-200 and correlated to the translated MMLU, ARC, HellaSwag benchmarks from Lai et al. (2023b) for Mistral 7B IT, Gemma 2B IT, Llama2 7B, 13B, 70B chat models.

1 Introduction

LLMs comprehend and generate human language across various domains and tasks, powering applications like virtual assistants and machine translation. As LLMs become more widely used globally, it is necessary to assess their capabilities in processing and understanding a specific language.

1.1 Limitations of Current Evaluation Methods

Standard evaluation metrics for multilingual LLMs focus on specific tasks like cross-lingual question answering (Artetxe et al., 2020), cross-lingual NLI (Conneau et al., 2018), or machine translation. This approach presents challenges. Task-specific datasets can be limited in scope or biased (Huang et al., 2024), the number of languages considered might be restricted, and the metrics used can be difficult to compare or interpret across different tasks and languages (Xu et al., 2024). Additionally, they often fail to capture the underlying linguistic factors that influence multilingual ability, such as variations in grammar, vocabulary, semantics, and pragmatics (Rajaee and Monz, 2024). Word overlap metrics like ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) can be unsuitable for comparison between languages with significant word order and phrasing variations. In addition, these metrics can produce vastly different scores for languages with rich morphology, even if the underlying meaning remains the same. This is further complicated since multilingual task scores sometimes exhibit low correlations between languages and can exhibit unexpected performance drops as models' sizes increase (Ali et al., 2024; Ahuja et al., 2024). This situation is in contrast to English downstream task scores, which often correlate with model size (Brown et al., 2020). Furthermore, existing tasks and benchmarks are often skewed by data contamination (Ahuja et al., 2024), where models are exposed to test data during training or fine-tuning, leading to artificially magnified performance.

1.2 Prompt-Based Evaluation Shortcomings

In LLM evaluation benchmarks, the LLM is given a natural language query or instruction as the prompt, and is expected to produce a natural language re-

Findings of the Association for Computational Linguistics: EACL 2024, pages 7971–7989
November 12-16, 2024 ©2024 Association for Computational Linguistics

Metric/Task	ARC	HellaSwag	MMLU	gen_enid	belebele	xcopa	paws-x	xnli	conv_enid
IP Flores	**0.93**	**0.98**	**0.98**	0.82	0.88	**0.95**	0.92	**0.88**	**0.86**
IP Tatoeba	0.87	0.95	0.97	**0.98**	0.83	0.92	**0.97**	0.79	-
TF	0.54	0.67	0.68	0.84	0.84	0.93	-	0.66	0.83
TP	0.72	0.82	0.84	0.82	**0.94**	0.78	-	0.71	0.79

Table 2: Pearson correlation (absolute values) between metrics and downstream tasks performance under the LLM Mistral 7B IT. Only correlation values that are statistically significant at level 0.05 are shown. Our proposed Information Parity (IP) typically better correlates with downstream tasks/benchmarks than other metrics. IP Flores (respectively, Tatoeba) refer to IP evaluated on the multilingual dataset Flores 200 (Tatoeba), TF and TP refer to the tokenization metrics, gen_enid, conv_enid refer to IN22 dataset.

sponse or answer. However, the way the prompt is phrased can significantly impact performance (Sclar et al., 2023), and different models might require tailored prompts to showcase their strengths. Finding these optimal prompts can be a laborious process that typically depends on human expertise. This situation may lead to irrelevant performance judgment, since in certain applications users may lack the expertise to craft optimal prompts (Zamfirescu-Pereira et al., 2023). Additionally, prompts might only assess a narrow aspect of its language understanding or generation, overlooking its broader potential or limitations (Biderman et al., 2024).

These issues escalate in multilingual performance evaluations. Inefficient tokenization in a certain language can limit the number of examples that can fit into the context window, hindering a model's ability to showcase its strengths (Ahia et al., 2023). Moreover, the need for cross-lingual prompting strategies introduces additional evaluation variations (Lai et al., 2023a; Qin et al., 2023). These limitations emphasize the need for a more standardized evaluation method.

1.3 Evaluation Through the Lens of Tokenization and Perplexity

Another potential way to evaluate multilingual LLMs is to measure their intrinsic ability to model the probability distribution of natural language, via perplexity. Perplexity quantifies how well an LLM can predict the next token given a context, and is often used as a proxy for language modeling quality. However, perplexity is sensitive to the choice of vocabulary and tokenizer (Remy et al., 2024), and can vary significantly across languages and models (Minixhofer et al., 2022), which makes it impractical for multilingual evaluations (Cao and Rimell, 2021). It inherently disadvantages languages with high morphological complexity or languages which suffer from high tokenizer fertility, requiring more tokens to represent the same information, as it averages over tokens.

Previous work suggested assessing an LLM's multilingual capabilities via tokenization metrics such as Tokenization Parity (Petrov et al., 2023) and Fertility (Rust et al., 2021). However, (Ali et al., 2024) found no correlation between these metrics and some downstream task performance, and argued that they have limited explanatory power for multilingual LLMs. Moreover, newer tokenizers such as Gemma's (Team et al., 2024) mitigate some of the multilingual tokenization issues, potentially reducing the relevance of tokenization-based metrics in some cases. This motivates a performance evaluation approach that captures the information representation capabilities of multilingual LLMs beyond tokenization.

1.4 Information Parity

In this paper, we propose to measure an LLM's general language capabilities using a novel metric called IP. Roughly speaking, for text in language L, IP is the ratio between the English variant of the text's negative log-likelihood and the L text's negative log-likelihood. As we explain below, IP has an interesting information-theoretic interpretation as the efficiency relative to English of losslessly compressing the L text using the LLM's probabilities followed by an entropy encoder (Izacard et al., 2019; Bellard, 2021; Mao et al., 2022; Levin and Kipnis, 2024). Such compression strategy attains state-of-the-art performance on large texts (Mahoney, 2023). Therefore, we may motivate IP from the concept of an ideal language-agnostic compressor that encodes text in any language with optimal efficiency. Since such a compressor is not attainable, we view English as a proxy for the most efficient encoding an LLM can achieve as measured in bits per token. This view is motivated

Metric/Task	MMLU	ARC	Hellaswag	mlqa	belebele	ind-xnli	xsotrycloze	xrisawoz
IP Flores 200	**0.96**	**0.82**	**0.73**	**0.94**	0.52	0.87	**0.94**	**0.97**
IP Tatoeba	0.90	0.81	0.67	0.89	**0.77**	-	0.84	-
TF	-	0.52	0.61	-	-	-	0.74	-
TP	0.95	0.52	0.53	0.84	-	**0.90**	-	-

Table 3: Pearson correlation (absolute values) between metrics and downstream tasks/benchmarks performance under the LLM Gemma 2B. The Information Parity (IP) metric we propose typically better correlates with downstream tasks/benchmarks than the other metrics. Xrisawoz refers to the dialogue action accuracy benchmark subset. Only statistically significant values at level 0.05 are shown.

by the overwhelming prevalence of English text in the training corpus of popular LLMs (Touvron et al., 2023; Team et al., 2023; Achiam et al., 2023; Jiang et al., 2023). By measuring how efficiently an LLM represents the same information across different languages, we capture its potential for multilingual performance relative to a reference. In our case, the reference is the efficiency of its English representation.

Since IP measures the total amount of information/uncertainty in a sequence as seen by the LLM, it is less affected by the tokenizer. This makes IP more robust to variations in tokenization across different languages and models compared to similar metrics like perplexity (Wang et al., 2023).

1.5 Contributions

We define IP and provide extensive evaluations of it and other metrics on publicly available LLMs like Llama2 (Touvron et al., 2023), Gemma (Team et al., 2024), and Mistral (Jiang et al., 2023). We demonstrate the usefulness of IP by analyzing its ability to predict downstream tasks and benchmark scores including MMLU (Hendrycks et al., 2021), ARC (Clark et al., 2018), and HellaSwag (Zellers et al., 2019), which exhibit high correlation to human preference as seen on Chiang et al. (2024). We compare IP with existing tokenization-based metrics like Tokenization Parity and Fertility, and the proportions of a language text in training data[1] (PTD).

Our results show that IP consistently exhibits strong correlations with the most popular downstream tasks and benchmarks. Especially those that require natural language understanding and commonsense reasoning across multiple domains and that align well with human preferences. Standard analysis of variance shows that IP has superior predictive power compared to other metrics we tried.

[1]PTD is taken from the Llama2 paper (Touvron et al., 2023).

These findings suggest that IP captures an LLM's multilingual capabilities better than any single tokenization metric or task-specific/benchmark scores.

Our findings imply that IP is useful as a standardized approach for comparing capabilities across languages and models which is direct, prompt-agnostic, task-invariant, and resilient to language and tokenization biases. Due to its computational efficiency and predictive prowess, IP emerges as a straightforward method to evaluate multilingual capabilities, reducing the need for inconsistent and complex downstream task evaluations.

1.6 Structure

The remainder of this paper is as follows: We define the IP metric in Section 2. We define the experimental setup and analysis methods in Section 3. We discuss the results in Section 4. We discuss limitations and challenges associated with the IP metric in Section 5. Concluding remarks are in Section 6.

2 Information Parity: Theoretical Background and Definition

For a given text $w_{1:n} = (w_1, .., w_n)$ where w_i is the i-th token, denote its negative log-likelihood under a language model (LM) by

$$\begin{aligned} I(w_{1:n}) &= -\log_2 \mathrm{P}_{\mathrm{LM}}(w_1, ..., w_n) \\ &= \sum_{i=1}^{n} -\log_2 \mathrm{P}_{\mathrm{LM}}(w_i | w_{1:i-1}) \end{aligned} \quad (1)$$

where $\mathrm{P}_{\mathrm{LM}}(w_1, ..., w_n)$ is the probability the LM assigns to $w_{1:n}$. In the discussion below, we use logarithm in base 2 so that $I(w_{1:n})$ is measured in bits. A lower value of $I(w_{1:n})$ indicates that the LM assigns a higher probability to the observed text, implying better prediction accuracy (Jurafsky and Martin, 2024). In the context of data compression, $I(w_{1:n})$ is roughly the length of the binary string produced by a compression scheme

Metric/Task	MMLU	ARC	HellaSwag	xnli	pawsx	xcopa	xquad	mlqa
IP Flores 200	**0.95**	**0.93**	**0.96**	**0.93**	0.91	0.89	**0.84**	0.82
IP Tatoeba	0.89	0.87	0.94	0.92	**0.96**	**0.96**	0.82	0.83
PTD	-	0.62	0.68	-	0.88	-	-	-
TF	0.72	0.66	0.71	0.86	-	-	0.83	**0.84**
TP	0.80	0.76	0.79	0.69	0.94	-	0.81	-

Table 4: Pearson correlation (absolute values) between metrics and downstream tasks/benchmarks performance under the LLM Llama 2 7B. Only correlation values that are statistically significant at level 0.05 are shown. The proposed Information Parity (IP) metric consistently demonstrates a stronger correlation with multilingual downstream task performance compared to other evaluated metrics, indicating its superior ability to predict LLM performance in multilingual settings.

employing the language model probabilities and an arithmetic encoder (Izacard et al., 2019; Bellard, 2021; Mao et al., 2022; Levin and Kipnis, 2024); such a scheme achieves state-of-the-art compression results on large texts (Mahoney, 2023). When the text is seen as a random sequence of tokens sampled from some generating mechanism $\mathrm{P}_{\mathrm{gen}}$ and $(\mathrm{P}_{\mathrm{gen}}, \mathrm{P}_{\mathrm{LM}})$ satisfies some regularity condition, the limit $I(W_{1:n})/n$ almost surely exists and converges to the cross-entropy between $\mathrm{P}_{\mathrm{gen}}$ to P_{LM} (Gray, 2011). This limit also coincides with the limiting number of bits per token attained by an asymptotically optimal implementation of the compression scheme mentioned before (Clarke and Barron, 1990). These well-known characterizations of (1) justify the interpretation of $I(w_{1:n})$ as the "information content" of the text $w_{1:n}$ under the LM.

Information Parity: Suppose we have English text w_E and its translation to another language w_L. We define the IP of w under the LM as

$$\mathrm{IP}(w_L) = \frac{I(w_E)}{I(w_L)} \tag{2}$$

In words, IP is the ratio between the information content of the text in English and the information content of the translated text in another language. It aims to measure 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.

3 Experimental Setup

3.1 Datasets

- **Tatoeba** (Tiedemann, 2020) a multilingual dataset of Machine Translation (MT) benchmarks derived from user-contributed translations. Presents inherent variance and bias between languages since the translation is not multi-parallel across all languages and the dataset is imbalanced between languages. We used a subset of 33 languages in evaluations.
- **Flores-200** (Team et al., 2022) a multilingual MT dataset that covers 200 languages, contains the translated variants of a sentence across all languages, and has the same number of samples across all languages. We used a subset of 50 languages[2].

3.2 Models

We perform our analysis on five open-source LLMs: the instruction-tuned variant of Mistral-7B v0.1 (Jiang et al., 2023), Llama-2-7B-chat, Llama-2-13B-chat, and Llama-2-70B-chat variants (Touvron et al., 2023) and Gemma-2B-it (Team et al., 2024). The latter is the smallest open-sourced instruction-tuned model from Google and is known for low rates of tokenizer fertility across languages. We used the default configuration of each model as provided in the Huggingface platform (Wolf et al., 2020).

3.3 Evaluations

We evaluate IP on all datasets in Section 3.1, per each model variant and across multiple languages. To conduct further evaluations and comparisons of multilingual model performance, we use the multilingual variants of MMLU (Hendrycks et al., 2021), HellaSwag (Zellers et al., 2019), and ARC (Clark et al., 2018), which were translated by Lai et al. (2023b) in 26 languages[3]. We use a 5-shot prompt

[2]We used the test split of the datasets from huggingface: Tatoeba, Flores.

[3]Due to time and compute constraints we evaluate MMLU only on a subset of zh, hi, ko, ar, de, es, ru, vi languages for

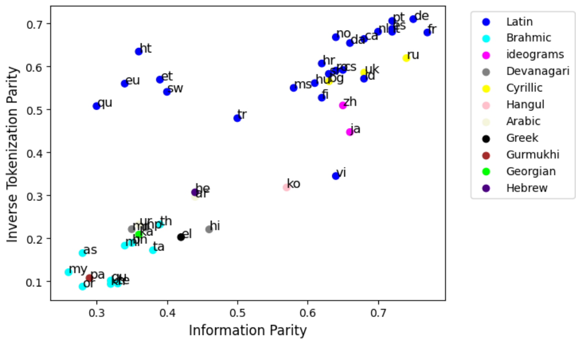

Figure 1: Inverse Tokenization Parity versus Information Parity for Llama 2 7B. Color corresponds to language script.

on MMLU, a 25-shot prompt on ARC, and a zero-shot prompt for HellaSwag. The full evaluation results are available in Appendix A.

3.4 Additional Evaluations from the Literature

We used results of evaluations on downstream tasks reported in MEGAVERSE (Ahuja et al., 2024) and from our results on the translated variants of MMLU ARC and HellaSwag benchmarks[4] from (Lai et al., 2023b), as well as results reported in Liu et al. (2024) for Llama 13B and 70B chat models under EN-BASIC prompt variant.[5] We compute the Tokenization Parity values on the Flores-200 (Team et al., 2022) dataset and use the Fertility values given to us by the authors of Ahuja et al. (2024).

3.5 Statistical Analysis

We analyze our metrics and task/benchmark scores data independently for every model variant. Our analysis is based on standard regression and analysis of variance (c.f. Chatterjee and Hadi (2013, Ch. 3)). Consider the simple regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, ..., n, \quad (3)$$

where $y = \{y\}_{i=1}^n$ is the target score vector, $x = \{x_i\}_{i=1}^n$ is the predictor vector, $\epsilon = \{\epsilon_i\}_{i=1}^n$ is the vector of residuals, and β_0 and β_1 are scalars. For a given (x, y) vector pair, we fit coefficient $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the squared norm of ϵ under the model (3). Denote $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. The squared Pearson correlation ρ^2 between x and y is given by

$$\rho^2(y;x) = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} := 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2},$$

where $\bar{y} := \frac{1}{n}\sum_{i=1}^n y_i$. We check that this correlation is significantly different than zero by testing

$$f(y;x) = \frac{SS_{\text{tot}} - SS_{\text{res}}}{\frac{1}{n-2} SS_{\text{res}}}, \quad (4)$$

against $F_{1,n-2}$, the F distribution with 1 over $n-2$ degrees of freedom. We summarize the result by the P-value

$$p(y;x) := \Pr\left[f(x;y) \geq F_{1,n-2}\right],$$

and reporting that $\rho(y;x)$ is significant if $p(y;x) < 0.05$. The adjusted coefficient of determination is useful to measure the explained variance in predicting y based on x:

$$R^2_{\text{adj}}(x;y) := \frac{n-1}{n-2}\rho^2(y;x). \quad (5)$$

We are typically interested in the ability of one x variable to predict multiple target variables $y_1, ..., y_m$. For example, x is the IP metric, and the y s are the different benchmark/task scores. In this setup, each (x, y_j) pair has a different number of samples n_j. Additionally, the assumption of equal residual variances in (3) underlying many of the existing combination methods does not hold in our case. Arguably, the most reasonable way to summarize prediction errors across multiple independent predictions in this case is by Fisher's combination statistic of F-tests' P-values:

$$\chi^2_{y_1,...,y_m;x} := \frac{1}{m}\sum_{j=1}^m 2\log(1/p_j(y_j;x)), \quad (6)$$

where the j-th F-test is associated with the regression of y_j on x. Note that $\chi^2_{y_1,...,y_m;x}$ has a chi-squared distribution over one degree of freedom when all F statistics $f(y_j;x)$ of (4) are distributed as their null, hence the larger $\chi^2_{y_1,...,y_m;x}$, the better x predicts the targets $y_1, ..., y_m$. Consequently, we treat $\chi^2_{y_1,...,y_m;x}$ as an index of success of x in predicting $y_1, ..., y_m$ in Table 5 .

To compare the predictive power of different metrics, we also performed *competitive regression analysis* for each model variant and downstream task score. In this analysis, we tested whether

Gemma and 13B Llama models, and use the reported results of MMLU on the 70B model in (Bendale et al., 2024).

[4]All evaluations utilized ~280 GPU hours on A100-80GB.

[5]For some combinations of language and benchmark/metric, we do not have values due to the lack of data or translation in the original datasets, hence we indicate the missing values with dashes in the tables.

Result/Metric	IP Flores	IP Tatoeba	TP	TF	PTD
Llama 2 7B Chat					
χ^2	**26.2**	18.62	12.75	10.21	4.73
R^2_{adj}	**0.79**	0.78	0.61	0.56	0.49
# of significant	**8**	8	6	6	3
Llama 2 13B Chat					
χ^2	**12.59**	8.25	10.24	9.77	4.62
R^2_{adj}	0.77	**0.81**	0.66	0.66	0.59
# of significant	9	5	**10**	10	5
Llama 2 70B Chat					
χ^2	**10.64**	-	7.89	6.01	3.00
R^2_{adj}	0.73	-	0.55	0.59	**0.78**
# of significant	**15**	-	12	8	2
Gemma 2B IT					
χ^2	**7.51**	6.27	5.22	2.84	-
R^2_{adj}	**0.73**	0.63	0.62	0.33	-
# of significant	**9**	6	6	3	-
Mistral 7B IT					
χ^2	**16.41**	9.89	12.49	7.79	-
R^2_{adj}	0.74	**0.82**	0.66	0.59	-
# of significant	12	9	**15**	14	-

Table 5: Reported averaged chisquared score (6), averaged R^2_{adj}, and the number of significant correlations, all associated with prediction capabilities under a linear model as explained in 3.5 (higher is better). Missing values indicate the unavailability of data, PTD stands for the proportion of language text in training data.

adding a second metric as a predictor x' to a linear model that already includes a first metric x can significantly reduce the mean squared error (MSE) of the prediction. This is measured by testing

$$f(y; x, x') = \frac{SS_{\text{res}} - SS'_{\text{res}}}{\frac{1}{n-3} SS'_{\text{res}}}$$

against $F_{1,n-3}$, where SS'_{res} is the residual sum of squares of the extended model. We report the results of the competitive regression analysis in 7.

4 Results

4.1 Prediction of Multilingual Performance

The results in Tables 2,3, and 4 show that IP exhibits strong and consistent correlation with downstream tasks performance on all tested models. The results in Table 5 further show that IP is useful in predicting multilingual capabilities across various auto-regressive model families and sizes. Notably, this observation holds for Indic languages, a category of low-resource languages in the sense that their PTD is low. We thus conclude that IP may serve as a reliable predictor of multilingual performance even for low-resource languages, where a lack of data and benchmarks makes it difficult to evaluate language models using conventional methods.

Model/Metric	**TP**	**TF**	**PTD**
Mistral 7B	0.72	0.47	-
Gemma 2B	0.81	0.61	-
Llama 2 7B	0.72	0.62	0.76
Llama 2 13B	0.72	0.62	0.67
Llama 2 70B	0.75	0.63	0.56

Table 6: Pearson correlation (absolute values) between IP and tokenization metrics computed on Flores 200 and PTD. High correlation values indicate a strong relationship between tokenization, PTD and language model effectiveness in encoding multilingual text. Only statistically significant values at level 0.05 are shown.

4.2 Relation between Tokenization and Information Parity

Our evaluations reveal that the inverse TP score of languages with Latin script is high, even if those languages do not share many linguistic features with the Indo-European languages; see for example the languages Euskara and Quechua in Figure 1 showing the relation between IP and inverse TP for

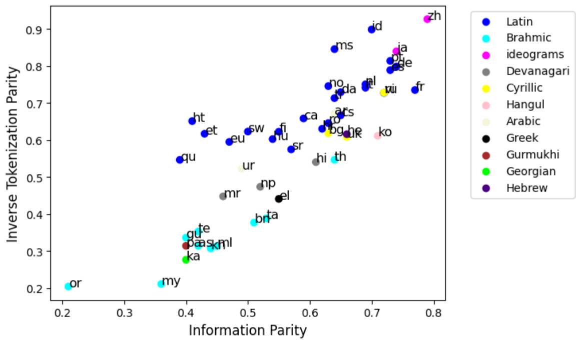

Figure 2: Gemma 2B IT Information Parity against inverse Tokenization Parity colored by language script.

Llama2 7B. On the other hand, it appears that IP manages to capture the linguistic distance between these languages and the Indo-European and thus predicts their performance more realistically.

The relation between IP and TP is evident due to their high correlation in Table 6. This correlation suggests that TP plays a crucial role in LLMs effectiveness at encoding information in these languages, as is expected from the viewpoint of data compression (Ziv and Lempel, 1978). We also observe high correlations between IP and PTD for the Llama models, whereas the correlation between PTD and tokenization metrics is not statistically significant. This supports our claim that IP takes into account the samples seen during the training process which affect the inherent abilities of the model to encode content in different languages.

Our competitive analysis findings in 7 suggest that TP and TF are captured within the explained variance that accounts for IP. This, in turn, could offer insights into when specific aspects of tokenization significantly influence the model's multilingual performance in downstream tasks. While further analysis and experiments are necessary to solidify this hypothesis, it presents promising opportunities for future research.

5 Limitations

Instruction tuning: The ideology behind IP is that good multilingual LLMs act as efficient language-agnostic compressors. Namely, they efficiently encode information in text, regardless of the text's language. However, we evaluated instruction-tuned LLMs via reinforcement learning with human feedback (RLHF) process (Ouyang et al., 2022) and DPO (Rafailov et al., 2023). This tuning may affect their compression efficiency. Additionally, IP does not account for the ability of LLMs to follow instructions in different languages, which may be relevant for some applications or tasks.

Dataset contamination: IP relies on parallel corpora that contain the same information in different languages. However, some of these corpora may have been used in pre-training which may inflate their compression performance.

Machine translation artifacts: Some of the task-specific metrics like MMLU, ARC, and HellaSwag were machine translated by GPT3.5 in Lai et al. (2023b). This translation may add biases and artificats to data that might alter performance measurements on these benchmarks.

Model/Metric	IP Flores	TP
Mistral 7B	10	10
Gemma 2B	**7**	3
Lllama 7B	**7**	2
Lllama 13B	**8**	4
Llama 70B	**9**	5
Model/Metric	IP Flores	TF
Mistral 7B	**9**	1
Gemma 2B	**7**	3
Lllama 7B	**7**	2
Lllama 13B	**7**	2
Llama 70B	**7**	2

Table 7: Competitive regression analysis of metrics as predictors for downstream tasks. The numbers indicate how many tasks benefit significantly in terms of MSE from adding a second predictor to a model that already includes the other predictor.

6 Conclusions

We introduced the Information Parity (IP) metric to provide a task-agnostic evaluation of the multilingual capabilities of LLMs. IP is easy to evaluate and has a natural information-theoretic interpretation as the efficiency of an LLM in representing the same information across different languages. Evaluations with publicly available LLMs reveal strong correlations between IP and a diverse set of downstream tasks, particularly those involving natural language understanding and commonsense reasoning. These properties suggest that IP could enable researchers and practitioners to assess model performance even for low-resource languages, leading to a more comprehensive understanding of LLM behavior across all languages.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R. Mortensen, Noah A. Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. *Preprint*, arXiv:2305.13707.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2024. Megaverse: Benchmarking large language models across languages, modalities, models and tasks. *Preprint*, arXiv:2311.07463.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Schulze Buschhoff, Charvi Jain, Alexander Arno Weber, Lena Jurkschat, Hammam Abdelwahab, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Samuel Weinbach, Rafet Sifa, Stefan Kesselheim, and Nicolas Flores-Herr. 2024. Tokenizer choice for llm training: Negligible or crucial? *Preprint*, arXiv:2310.08754.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Fabrice Bellard. 2021. Nncp v2: Lossless data compression with transformer.

Abhijit Bendale, Michael Sapienza, Steven Ripplinger, Simon Gibbs, Jaewon Lee, and Pranav Mistry. 2024. Sutra: Scalable multilingual language model architecture. *Preprint*, arXiv:2405.06694.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, Charles Lovering, Niklas Muennighoff, Ellie Pavlick, Jason Phang, Aviya Skowron, Samson Tan, Xiangru Tang, Kevin A. Wang, Genta Indra Winata, François Yvon, and Andy Zou. 2024. Lessons from the trenches on reproducible evaluation of language models. *Preprint*, arXiv:2405.14782.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Kris Cao and Laura Rimell. 2021. You should evaluate your language model on marginal likelihood over tokenisations. *Preprint*, arXiv:2109.02550.

S. Chatterjee and A.S. Hadi. 2013. *Regression Analysis by Example*. Wiley Series in Probability and Statistics. Wiley.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *Preprint*, arXiv:2403.04132.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Bertrand S Clarke and Andrew R Barron. 1990. Information-theoretic asymptotics of bayes methods. *IEEE Transactions on Information Theory*, 36(3):453–471.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *Preprint*, arXiv:1809.05053.

Robert M Gray. 2011. *Entropy and information theory*. Springer Science & Business Media.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Kaiyu Huang, Fengran Mo, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, Jian-Yun Nie, and Yang Liu. 2024. A survey on large language models with multilingualism: Recent advances and new frontiers. *Preprint*, arXiv:2405.10936.

Gautier Izacard, Armand Joulin, and Edouard Grave. 2019. Lossless data compression with transformer.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Daniel Jurafsky and James H. Martin. 2024. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Pearson Prentice Hall. Online manuscript released August 20, 2024.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023a. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *Preprint*, arXiv:2304.05613.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023b. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *Preprint*, arXiv:2307.16039.

Dana Levin and Alon Kipnis. 2024. The likelihood gain of a language model as a metric for text summarization. In *'Learn to Compress' Workshop@ ISIT 2024*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2024. Is translation all you need? a study on solving multilingual tasks with large language models. *Preprint*, arXiv:2403.10258.

Matt Mahoney. 2023. Large text compression benchmark.

Yu Mao, Yufei Cui, Tei-Wei Kuo, and Chun Jason Xue. 2022. A fast transformer-based general-purpose lossless compressor. *arXiv preprint arXiv:2203.16114*.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. *Preprint*, arXiv:2305.15425.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. *Preprint*, arXiv:2310.14799.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Sara Rajaee and Christof Monz. 2024. Analyzing the evaluation of cross-lingual knowledge transfer in multilingual language models. *Preprint*, arXiv:2402.02099.

François Remy, Pieter Delobelle, Hayastan Avetisyan, Alfiya Khabibullina, Miryam de Lhoneux, and Thomas Demeester. 2024. Trans-tokenization and cross-lingual vocabulary transfers: Language adaptation of llms for low-resource nlp. *Preprint*, arXiv:2408.04303.

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. *Preprint*, arXiv:2012.15613.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *Preprint*, arXiv:2310.11324.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin

Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

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, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Yequan Wang, Jiawen Deng, Aixin Sun, and Xuying Meng. 2023. Perplexity from plm is unreliable for evaluating text quality. *Preprint*, arXiv:2210.05892.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Yuqi Ye, and Hanwen Gu. 2024. A survey on multilingual large language models: Corpora, alignment, and bias. *Preprint*, arXiv:2404.00929.

JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *Preprint*, arXiv:1905.07830.

Jacob Ziv and Abraham Lempel. 1978. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536.

A Appendix

Code	Llama2 70B	Llama2 13B	Llama2 7B	Gemma 2B	Mistral 7B
ru	0.73	0.74	0.74	0.72	0.75
fr	0.76	0.77	0.77	0.77	0.79
ko	0.57	0.57	0.57	0.71	0.56
ja	0.65	0.65	0.66	0.74	0.55
he	0.44	0.44	0.44	0.66	0.39
hu	0.63	0.62	0.61	0.54	0.61
no	0.65	0.64	0.64	0.63	0.53
hi	0.47	0.46	0.46	0.61	0.38
fi	0.67	0.64	0.62	0.55	0.4
es	0.7	0.7	0.72	0.73	0.74
de	0.75	0.75	0.75	0.74	0.75
it	0.72	0.72	0.72	0.69	0.73
nl	0.72	0.72	0.7	0.69	0.71
zh	0.64	0.63	0.65	0.79	0.65
vi	0.64	0.64	0.64	0.72	0.44
id	0.69	0.69	0.68	0.7	0.58
ro	0.66	0.65	0.64	0.63	0.61
uk	0.68	0.69	0.68	0.66	0.66
sr	0.65	0.64	0.63	0.57	0.58
hr	0.65	0.63	0.62	0.62	0.63
da	0.69	0.67	0.66	0.65	0.65
ca	0.68	0.68	0.68	0.59	0.68
ar	0.45	0.44	0.44	0.64	0.4
tr	0.52	0.51	0.5	0.64	0.49
cs	0.69	0.66	0.65	0.65	0.65
th	0.38	0.39	0.39	0.64	0.32
bn	0.35	0.36	0.35	0.51	0.28
bg	0.66	0.65	0.63	0.63	0.61
el	0.45	0.42	0.42	0.55	0.33
ur	0.37	0.36	0.36	0.49	0.31
mr	0.35	0.35	0.35	0.46	0.28
eu	0.37	0.34	0.34	0.47	0.3
et	0.42	0.4	0.39	0.43	0.34
ms	0.6	0.59	0.58	0.64	0.53
as	0.27	0.28	0.28	0.42	0.18
gu	0.33	0.33	0.32	0.4	0.27
ka	0.37	0.35	0.36	0.4	0.24
kn	0.32	0.32	0.32	0.44	0.28
ml	0.33	0.33	0.34	0.45	0.24
np	0.37	0.37	0.37	0.52	0.3
or	0.29	0.28	0.28	0.21	0.21
pa	0.31	0.3	0.29	0.4	0.25
ta	0.37	0.36	0.38	0.53	0.29
te	0.33	0.33	0.33	0.42	0.26
my	0.27	0.27	0.26	0.36	0.17
sw	0.42	0.41	0.4	0.5	0.36
pt	0.73	0.72	0.72	0.73	0.73
ht	0.38	0.36	0.36	0.41	0.31
qu	0.31	0.3	0.3	0.39	0.3

Table 8: Information Parity (IP) - mean values evaluated on the Flores 200 dataset.

Language	Llama2	Gemma 2B	Mistral 7B
ru	1.62	1.37	1.85
fr	1.47	1.36	1.6
ko	3.14	1.64	2.44
ja	2.24	1.19	2.15
he	3.26	1.62	3.38
hu	1.78	1.66	2.0
no	1.5	1.34	1.59
hi	4.53	1.85	4.5
fi	1.9	1.61	2.0
es	1.46	1.27	1.58
de	1.41	1.25	1.58
it	1.47	1.35	1.62
nl	1.47	1.33	1.6
zh	1.96	1.08	1.6
vi	2.9	1.37	2.9
id	1.75	1.11	1.84
ro	1.69	1.55	1.81
uk	1.71	1.64	1.93
sr	1.72	1.74	1.89
hr	1.65	1.59	1.77
da	1.53	1.37	1.62
ca	1.51	1.52	1.62
ar	3.37	1.49	3.43
tr	2.09	1.4	2.21
cs	1.69	1.5	1.86
th	4.31	1.83	4.18
bn	5.28	2.65	4.84
bg	1.77	1.62	1.92
el	4.93	2.27	5.19
ur	4.31	1.91	4.26
mr	4.52	2.23	4.6
eu	1.79	1.68	1.89
et	1.76	1.62	1.84
ms	1.82	1.18	1.9
as	6.04	3.19	5.61
gu	9.83	2.98	8.52
ka	4.79	3.62	4.79
kn	10.66	3.26	6.19
ml	5.46	3.2	10.67
np	4.44	2.11	4.4
or	11.39	4.91	11.82
pa	9.3	3.19	10.25
ta	5.8	2.58	5.78
te	10.55	2.84	7.11
my	8.26	4.75	8.09
sw	1.85	1.61	1.94
pt	1.42	1.23	1.55
ht	1.58	1.54	1.67
qu	1.97	1.83	2.06

Table 9: Tokenization Parity evaluated on the Flores 200 dataset

Language	Llama2 7B	Llama2 13B	Gemma 2B	Mistral 7B
ar	0.2724	0.2908	0.292	0.2778
de	0.371	0.4238	0.3046	0.4049
es	0.3928	0.4339	0.3133	0.4183
hi	0.273	0.281	0.2817	0.2714
ru	0.3423	0.3978	0.304	0.3775
vi	0.3178	0.3478	0.3078	0.3052
zh	0.3256	0.3732	0.3221	0.3771
bn	0.2562	-	-	0.2535
ca	0.3721	-	-	0.3997
da	0.3572	-	-	0.3817
fr	0.3814	-	-	0.4153
hr	0.3359	-	-	0.3635
hu	0.3207	-	-	0.3423
id	0.3456	-	-	0.3352
it	0.3696	-	-	0.4005
kn	0.2634	-	-	0.2548
ml	0.2563	-	-	0.2477
mr	0.2628	-	-	0.266
ne	0.2566	-	-	0.2669
nl	0.3643	-	-	0.3981
ro	0.3499	-	-	0.3735
sk	0.32	-	-	0.34
sr	0.3282	-	-	0.3553
ta	0.2564	-	-	0.2524
te	0.2531	-	-	0.2476
uk	0.3348	-	-	0.3629

Table 10: MMLU accuracy evaluated on Llama2 7B, Llama2 13B, Gemma 2B and Mistral 7B using Okapi Evaluation Framework for Multilingual LLMs

Language	Llama2 7B	Llama2 13B	Gemma 2B	Mistral 7B
ar	0.2156	0.2181	0.2275	0.2019
bn	0.1805	-	-	0.1942
ca	0.3834	0.4142	0.2333	0.3602
da	0.3102	0.3573	0.2279	0.3222
de	0.3507	0.4089	0.2515	0.3576
es	0.3744	0.441	0.2897	0.3923
fr	0.3781	0.4183	0.2789	0.3867
hi	0.2286	0.2269	0.2337	0.1978
hr	0.302	0.3182	0.2062	0.3182
hu	0.2834	0.3048	0.1986	0.2688
id	0.3043	0.3316	0.2308	0.2376
it	0.3824	0.4303	0.2429	0.3944
kn	0.2178	-	-	0.2117
ml	0.2215	-	-	0.2172
mr	0.2346	-	-	0.2242
ne	0.2104	-	-	0.2156
nl	0.3584	0.4106	0.2258	0.3447
ro	0.3256	0.3582	0.2099	0.3299
ru	0.349	0.3841	0.2686	0.355
sk	0.2763	0.2806	0.2335	0.2695
sr	0.2917	0.3311	0.2216	0.3131
ta	0.2215	-	-	0.2189
te	0.2088	-	-	0.2096
uk	0.3199	0.3918	0.2618	0.3576
vi	0.2812	0.312	0.2538	0.2427
zh	0.3316	0.3744	0.2821	0.3291

Table 11: ARC accuracy evaluated on Llama2 7B, Llama2 13B, Gemma 2B and Mistral 7B using Okapi Evaluation Framework for Multilingual LLMs

Language	Llama2 7B	Llama2 13B	Gemma 2B	Mistral 7B
ar	0.2867	0.3007	0.2634	0.2793
bn	0.2587	-	-	0.2624
ca	0.389	0.4239	0.2801	0.3848
da	0.3784	0.4135	0.2794	0.3718
de	0.4021	0.431	0.2859	0.3952
es	0.4396	0.4742	0.291	0.4334
fr	0.4263	0.4599	0.2913	0.4261
hi	0.2825	0.289	0.2743	0.2759
hr	0.3438	0.3727	0.2712	0.3444
hu	0.3282	0.3467	0.2672	0.3246
id	0.3546	0.3794	0.2713	0.3268
it	0.4059	0.4394	0.2846	0.402
kn	0.2589	-	-	0.2558
ml	0.2538	-	-	0.2485
mr	0.2593	-	-	0.2579
ne	0.2635	-	-	0.2583
nl	0.3849	0.4195	0.2757	0.3855
ro	0.3653	0.3936	0.282	0.3581
ru	0.3776	0.4111	0.2764	0.3904
sk	0.3068	0.3231	0.2714	0.3026
sr	0.3408	0.3698	0.2739	0.3455
ta	0.2572	-	-	0.2502
te	0.2584	-	-	0.2552
uk	0.3664	0.3909	0.2764	0.3672
vi	0.3457	0.3647	0.2875	0.3107
zh	0.3601	0.3893	0.2954	0.3736

Table 12: HellaSwag accuracy evaluated on Llama2 7B, Llama2 13B, Gemma 2B and Mistral 7B using Okapi Evaluation Framework for Multilingual LLMs

Language	Llama2 7B	Llama2 13B	Gemma 2B	Mistral 7B
de	0.74	0.75	0.77	0.7
ru	0.75	0.76	0.75	0.69
it	0.69	0.7	0.75	0.68
nl	0.66	0.68	0.73	0.66
da	0.63	0.65	0.7	0.62
zh	0.58	0.55	0.78	0.62
ca	0.59	0.6	0.64	0.59
hr	0.56	0.58	0.69	0.59
cs	0.58	-	0.67	0.58
ko	0.53	0.51	0.72	0.58
no	0.61	0.63	0.68	0.57
uk	0.67	0.67	0.7	0.57
id	0.6	0.62	0.75	0.57
ja	0.58	0.56	0.74	0.57
hu	0.55	0.56	0.62	0.55
ro	0.58	0.61	0.67	0.54
tr	0.52	-	0.67	0.52
bg	0.57	-	-	0.51
sr	0.57	0.57	0.64	0.51
vi	0.56	0.56	0.74	0.49
he	0.47	0.48	0.71	0.48
hi	0.49	0.49	0.69	0.48
th	0.47	-	0.73	0.47
fi	0.54	0.56	0.62	0.46
el	0.45	-	-	0.44
ar	0.46	0.46	0.69	0.44
et	0.43	-	-	0.42
eu	0.35	-	-	0.37
ur	0.4	-	-	0.36
mr	0.39	-	-	0.35
bn	0.42	-	-	0.33

Table 13: Information Parity (IP) evaluated on the Tatoeba dataset.

Task/Metric	IP Flores	TP	TF
HellaSwag	**0.89**	0.82	0.88
ARC	**0.90**	0.82	0.86
MMLU	**0.95**	0.95	0.90
xnli-TIAYN	**0.93**	0.72	0.80
pawsx-TIAYN	**0.98**	0.94	0.84
xnli	0.75	0.90	**0.94**
xquad	**0.82**	0.70	0.78
mgsm-TIAYN	**0.96**	0.69	0.73
xcopa-TIAYN	**0.83**	-	0.77
pawsx	-	**0.83**	-
xcopa	-	**0.91**	0.87

Table 14: Pearson correlation (absolute values) between metrics and downstream tasks/benchmarks performance under the LLM Llama 2 13B. Only correlation values that are statistically significant at level 0.05 are shown. TIAYN refers to results from (Liu et al., 2024)

Task/Metric	IP-Flores	TP	TF
MMLU	**0.89**	0.76	-
xnli	**0.89**	0.81	0.86
pawsx	0.91	**0.99**	0.93
xquad	**0.77**	0.68	0.76
mlqa	**0.87**	-	-
belebele	**0.98**	0.91	0.78
conv-iden	**0.77**	0.64	-
gen-enid	**0.78**	0.61	-
gen-iden	**0.78**	0.67	0.71
xriawoz	**0.96**	-	-
MGSM	**0.96**	0.69	0.73
xnli-TIAYN	**0.86**	0.59	0.74
pawsx-TIAYN	0.85	**0.91**	-
xcopa-TIAYN	**0.85**	-	-
xcopa	-	**0.87**	0.87

Table 15: Pearson correlation (absolute values) between metrics and downstream tasks/benchmarks performance under the LLM Llama 2 70B. Only correlation values that are statistically significant at level 0.05 are shown.Xrisawoz refers to the success rate accuracy benchmark subset, gen-enid,gen-iden, conv-enid refer to IN22 dataset. TIAYN refers to results from (Liu et al., 2024)

Language Name	Code
English	en
Hungarian	hu
Russian	ru
Norwegian	no
Hindi	hi
French	fr
Korean	ko
Japanese	ja
Hebrew	he
Finnish	fi
Spanish	es
German	de
Italian	it
Dutch	nl
Chinese	zh
Vietnamese	vi
Indonesian	id
Romanian	ro
Ukrainian	uk
Serbian	sr
Croatian	hr
Danish	da
Catalan	ca
Arabic	ar
Turkish	tr
Czech	cs
Thai	th
Bengali	bn
Bulgarian	bg
Greek	el
Urdu	ur
Marathi	mr
Basque	eu
Estonian	et
Malay	ms
Assamese	as
Gujarati	gu
Georgian	ka
Kannada	kn
Malayalam	ml
Nepali	np
Odia	or
Punjabi	pa
Tamil	ta
Telugu	te
Burmese	my
Swahili	sw
Portuguese	pt
Haitian Creole	ht
Quechua	qu

Table 16: Flores 200 used languages - language Names to codes

Language Code	**Language Name**
ru	Russian
fr	French
ko	Korean
jp	Japanese
he	Hebrew
hu	Hungarian
no	Norwegian
hi	Hindi
fi	Finnish
es	Spanish
de	German
it	Italian
nl	Dutch
zh	Chinese
vi	Vietnamese
id	Indonesian
ro	Romanian
uk	Ukrainian
sr	Serbian
hr	Croatian
da	Danish
ca	Catalan
ar	Arabic
tr	Turkish
cs	Czech
th	Thai
bn	Bengali
bg	Bulgarian
el	Greek
ur	Urdu
mr	Marathi
eu	Basque
et	Estonian

Table 17: Tatoeba used languages - language Names to codes