Multilingual Compression Parity: How Efficiently Large Language Models Represent Information Across Languages?

Alexander Tsvetkov [1] **Alon Kipnis** [1]

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 Compression Parity (CP) that can predict an LLM's capabilities across multiple languages in a task-agnostic manner. CP has a solid motivation from an information theory perspective: it is associated with the ability of the LLM to compress text in a given language compared to the same ability in a reference language. We evaluate CP 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, CP 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 CP may be useful for ranking multilingual LLMs' capabilities regardless of the downstream task.

1. Introduction

1.1. Background

LLMs have become ubiquitous, powering applications like email generation, virtual assistants, and machine translation in our daily lives. Trained on massive datasets, they can comprehend and generate human language across various domains and tasks. As LLMs become more widely used globally, it is necessary to assess their capabilities in processing and understanding a specific language.

Current evaluation methods for multilingual LLMs typically focus on specific tasks like cross-lingual question answering (Artetxe et al., 2020), cross-lingual NLI (Conneau et al., 2018), or machine translation. While informative, these approaches have limitations. Task-specific datasets can be limited or biased, the number of languages considered might be restricted, and the metrics used can be difficult to compare or interpret across different tasks and languages. Additionally, they often fail to capture the underlying linguistic factors that influence multilingual ability, such as variations in grammar, vocabulary, semantics, and pragmatics (Rajaee & Monz, 2024). This is further complicated by the observation that, unlike English metrics scores which often correlate with model size, multilingual metrics can exhibit performance drops with size and low correlation between each other (Ali et al., 2024; Ahuja et al., 2024). To make matters worse, existing 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.

Another common problem with LLM evaluation benchmarks is that most of them are prompt-based. Namely, the LLM is given a natural language query or instruction as the prompt, and expected to produce a natural language response or answer. However, the way the prompt is phrased can significantly impact performance, 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 not have the required expertise to craft optimal prompts. Additionally, prompts might only assess a narrow aspect of its language understanding or generation, overlooking its broader potential or limitations.

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 make prompt-based evaluations insufficient to assess the multilingual capabilities of LLMs, and emphasize the need

[1]Department of Computer Science, Reichman University, Herzlia, Israel. Correspondence to: Alexander Tsvetkov <alexander.tsvetkov@post.runi.ac.il>, Alon Kipnis <alon.kipnis@runi.ac.il>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Table 1. Pearson correlation (absolute values) between metrics and downstream tasks performance under the LLM Llama 2 7B. Only correlation values that are statistically significant at level 0.05 are shown.

Metric/Task	MMLU	ARC	HellaSwag	xnli	pawsx	xcopa	xquad	mlqa
CP Flores 200	**0.95**	**0.93**	**0.96**	**0.93**	0.91	0.89	**0.84**	0.82
CP Tatoeba	0.89	0.87	0.94	0.92	**0.96**	**0.96**	0.82	0.83
Training Lang Distribution	-	0.62	0.68	-	0.88	-	-	-
Tokenizer Fertility	0.72	0.66	0.71	0.86	-	-	0.83	**0.84**
Tokenizer Parity	0.80	0.76	0.79	0.69	0.94	-	0.81	-

for a more standardized evaluation method.

Previous work suggested assessing an LLM's multilingual capabilities via tokenization metrics such as Tokenizer Parity (Petrov et al., 2023) and Tokenizer Fertility (Rust et al., 2021). Such metrics might also be motivated by the intimate connection between language modeling and data compression (Shannon, 1951). However, (Ali et al., 2024) found no clear correlation between these metrics and downstream task performance, and argued that they have limited explanatory power for multilingual LLMs. Moreover, newer tokenizers such as Gemma (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 more comprehensive approach that examines the core information compression capabilities of multilingual LLMs beyond tokenization.

1.2. Contribution

We propose a novel metric called Compression Parity (CP) for assessing the multilingual capabilities of LLMs in a task-independent manner. For a language L and some multilingual text, CP is the ratio between the English text's negative log-likelihood (NLL) and the L text's NNL. This definition has a clear information-theoretic interpretation as the efficiency relative to English of compressing the L text generated by the LLM. Therefore, we may motivate CP from the concept of a language-agnostic compressor, a theoretical ideal that encodes the same information with identical efficiency regardless of the language. Since such a compressor is not practically available, we use English as a proxy for the most efficient encoding an LLM can achieve. Practically, the LLM's ability to represent information in English best approximates the theoretical ideal due to the prevalence of English text in the training data. By measuring how efficiently an LLM represents the same information in a different language relative to English, we aim to capture its performance potential in that different language.

Unlike other metrics, CP is prompt-agnostic, task invariant, and resilient to language and tokenization biases (Ali et al., 2024). This allows for a more direct comparison of multilingual capabilities across different languages on the same model.

We evaluate our metric on publicly available LLMs like Llama2 (Touvron et al., 2023)[1], Gemma (Team et al., 2024), and Mistral (Jiang et al., 2023). We correlate CP with downstream tasks and benchmarks including MMLU (Hendrycks et al., 2021), ARC (Clark et al., 2018) and HellaSwag (Zellers et al., 2019), which exhibit a high correlation to human preference as seen on Chiang et al. (2024). In addition, we compare our metric with existing tokenization-based metrics of tokenizer parity and fertility.

Our results show that CP consistently exhibits strong correlations with various downstream tasks and benchmarks, especially those that require natural language understanding and commonsense reasoning across multiple domains and that align well with human preferences. These findings suggest that CP captures an LLM's multilingual capabilities better than any single tokenization metric or task-specific/benchmark scores. Consequently, CP emerges as a direct and standardized approach for comparing the multilingual capabilities of LLMs.

2. Method

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 LM by

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

where $\mathrm{P}_{\mathrm{LM}}(w_1, ..., w_n)$ is the probability the LM assigns to $w_{1:n}$. We use logarithm in base 2 so that $I(x_{1:n})$ is measured in bits. In the context of data compression, $I(w_{1:n})$ is roughly the length of the binary string produced by a compression scheme employing the language model probabilities and an entropy encoder (Izacard et al., 2019; Bellard, 2021; Mao et al., 2022; Levin & 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 vector of tokens sampled from the law P_{LM}, $\mathbb{E}\left[I(W_{1:n})\right]$ is Shannon's entropy of the distribution the LM induces on n-length sequences. In this case, $\mathbb{E}\left[I(W_{1:n})\right]$ bounds from below the expected length of the binary string

[1]Training language distribution is taken from the Llama2 paper

Table 2. Pearson correlation (absolute values) between metrics and downstream tasks performance under the LLM Gemma 2B. Compression Parity (CP) is our proposed metric. Xrisawoz refers to the dialogue action accuracy benchmark subset. Only statistically significant values at level 0.05 are shown.

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

representation of text sampled from the model under any lossless compression scheme (Cover & Thomas, 2006). If in addition P_{LM} defines an ergodic information source, then the Shannon-McMillan-Brieman theorem says that the entropy rate of P_{LM} is well-defined as the limit of $I(W_{1:n})/n$ as $n \to \infty$ (Algoet & Cover, 1988). In this case, the entropy rate bounds from below the number of bits per token in any binary representation of the text (Cover & Thomas, 2006). 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.

Compression Parity (CP): Computes 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 express the basic proportion of how well the LLM compresses the same information in different languages. A higher CP indicates a higher compression efficiency and a closer alignment with the language-agnostic compressor ideal. The CP is computed as follows:

$$\mathrm{CP}(L) = \frac{I(E)}{I(L)} \tag{2}$$

where $I(E)$ and $I(L)$ are the NLL of (1) of the text in English and the translated text in language L, respectively.

3. Experimental Setup

3.1. Datasets

- **Tatoeba** (Tiedemann, 2020) a multilingual dataset of 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.
- **Floress-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].

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

Table 3. Pearson correlation (absolute values) between CP and tokenization metrics computed on Flores 200 and training language distribution of Llama2. Only statistically significant values at level 0.05 are shown.

Model/Metric	**TP**	**TF**	**TLD**
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

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,13B,70B chat variants (Touvron et al., 2023), and Gemma-2B-it (Team et al., 2024), the smallest open-sourced instruction-tuned model from Google, 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 the CP metric on the datasets in Section 3.1 and report their mean values in A. To conduct further evaluations 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 on MMLU, a 25-shot prompt on ARC, and a zero-shot prompt for HellaSwag.

4. Results

To showcase the advantage of our proposed metric, we compute Pearson's correlation values with results of downstream tasks reported in MEGAVERSE (Ahuja et al., 2024) and from our results on the translated variants of MMLU ARC and HellaSwag from (Lai et al., 2023b). We compare our

[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 Gemma, and 13B Llama models.

Table 4. 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 Compression Parity (CP) typically better correlates with downstream tasks/benchmarks than other metrics. CP Flores (respectively, Tatoeba) refer to CP evaluated on the multilingual dataset Flores 200 (Tatoeba), gen_enid, conv_enid refer to IN22 dataset.

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

results to the tokenizer-based metrics as well, to see whether CP is better suited to predict downstream task performance. We compute the tokenizer 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).

5. Discussion

The results in 1,2, 4 show that CP exhibits consistently strong correlations with downstream tasks performance.[4] This suggests that CP can be utilized to predict multilingual model capabilities for missing languages on similar tasks or the same datasets which often cover only a small subset of languages.

A comparison of CP values across different languages and models reveals variation in CP values across the models. For instance, we find that Gemma, has consistently higher CP values for non Latin languages than Llama or Mistral. We contribute this phenomenon to the fact that Gemma showed lower tokenization fertility and parity for non-Latin language families. The relation between CP and tokenizer parity is clearly evident via the high absolute correlation values between them as shown in 3. This implies that tokenization parity plays a crucial role in LLMs effectiveness at encoding information in these languages, and thus has an aspect of contribution to higher compression parity.

Our findings suggest that tokenization parity and fertility might be captured within the explained variance that accounts for compression parity. This, in turn, could offer insights into when specific aspects of tokenization significantly influence the model's multilingual performance in downstream tasks.

6. Limitations

Instruction tuning: Our metric is based on the assumption that ideal multilingual LLMs can act as language-agnostic compressors, encoding the same information with the same efficiency across languages. However, we evaluated instruction-tuned LLMs which went through RLHF (Ouyang et al., 2022) and DPO (Rafailov et al., 2023) which may affect their compression behavior to some extent. Moreover, Compression Parity 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: CP relies on parallel corpora that contain the same information in different languages. However, some of these corpora may have been used in pretraining some LLMs, which may inflate their compression performance.

Machine Translation Artifacts: Some of the task-specific metrics like MMLU, ARC, HellaSwag were machine translated by GPT3.5 in Lai et al. (2023b), which may add artifacts and biases to the translations which in turn might alter the actual performance measurements on these benchmarks.

7. Conclusion

We introduced the Compression Parity (CP) metric to provide a task-agnostic evaluation of the multilingual capabilities of LLMs. CP 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 a strong correlation between Compression Parity and a diverse set of downstream tasks, particularly those involving natural language understanding and commonsense reasoning. These properties suggest that CP could enable researchers and practitioners to assess model performance even for low-resource languages, leading to a better understanding of LLM behavior across all languages.

Impact Statement

This paper contributes to the advancement of Machine Learning by focusing on one key area: improving multilingual language models. Our work has the potential to significantly impact society by enabling the development and evaluation of more capable models for under-represented groups and populations. Which will ultimately lead to more inclusive and effective language technology for everyone.

[4]Results for Llama Chat 13B showed similar conclusions to 7B and were omitted for brevity

References

Ahia, O., Kumar, S., Gonen, H., Kasai, J., Mortensen, D. R., Smith, N. A., and Tsvetkov, Y. Do all languages cost the same? tokenization in the era of commercial language models, 2023.

Ahuja, S., Aggarwal, D., Gumma, V., Watts, I., Sathe, A., Ochieng, M., Hada, R., Jain, P., Axmed, M., Bali, K., and Sitaram, S. Megaverse: Benchmarking large language models across languages, modalities, models and tasks, 2024.

Algoet, P. H. and Cover, T. M. A sandwich proof of the shannon-mcmillan-breiman theorem. *The Annals of Probability*, 16(2):899–909, 1988.

Ali, M., Fromm, M., Thellmann, K., Rutmann, R., Lübbering, M., Leveling, J., Klug, K., Ebert, J., Doll, N., Buschhoff, J. S., Jain, C., Weber, A. A., Jurkschat, L., Abdelwahab, H., John, C., Suarez, P. O., Ostendorff, M., Weinbach, S., Sifa, R., Kesselheim, S., and Flores-Herr, N. Tokenizer choice for llm training: Negligible or crucial?, 2024.

Artetxe, M., Ruder, S., and Yogatama, D. 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, 2020. doi: 10.18653/v1/2020.acl-main.421. URL http://dx.doi.org/10.18653/v1/2020.acl-main.421.

Bellard, F. Nncp v2: Lossless data compression with transformer, 2021.

Bendale, A., Sapienza, M., Ripplinger, S., Gibbs, S., Lee, J., and Mistry, P. Sutra: Scalable multilingual language model architecture, 2024.

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., and Stoica, I. Chatbot arena: An open platform for evaluating llms by human preference, 2024.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018.

Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S. R., Schwenk, H., and Stoyanov, V. Xnli: Evaluating cross-lingual sentence representations, 2018.

Cover, T. and Thomas, J. A. Elements of information theory, 2006.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding, 2021.

Izacard, G., Joulin, A., and Grave, E. Lossless data compression with transformer, 2019. URL https://bellard.org/nncp/nncp.pdf.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b, 2023.

Lai, V. D., Ngo, N. T., Veyseh, A. P. B., Man, H., Dernoncourt, F., Bui, T., and Nguyen, T. H. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning, 2023a.

Lai, V. D., Nguyen, C. V., Ngo, N. T., Nguyen, T., Dernoncourt, F., Rossi, R. A., and Nguyen, T. H. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback, 2023b.

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

Liu, C., Zhang, W., Zhao, Y., Luu, A. T., and Bing, L. Is translation all you need? a study on solving multilingual tasks with large language models, 2024.

Mahoney, M. Large text compression benchmark, 2023. URL http://www.mattmahoney.net/dc/text.html.

Mao, Y., Cui, Y., Kuo, T.-W., and Xue, C. J. A fast transformer-based general-purpose lossless compressor. *arXiv preprint arXiv:2203.16114*, 2022.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback, 2022.

Petrov, A., Malfa, E. L., Torr, P. H. S., and Bibi, A. Language model tokenizers introduce unfairness between languages, 2023.

Qin, L., Chen, Q., Wei, F., Huang, S., and Che, W. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages, 2023.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model, 2023.

Rajaee, S. and Monz, C. Analyzing the evaluation of cross-lingual knowledge transfer in multilingual language models, 2024.

Rust, P., Pfeiffer, J., Vulić, I., Ruder, S., and Gurevych, I. How good is your tokenizer? on the monolingual performance of multilingual language models, 2021.

Shannon, C. E. Prediction and entropy of printed english. *Bell system technical journal*, 30(1):50–64, 1951.

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., Tafti, P., Hussenot, L., Sessa, P. G., Chowdhery, A., Roberts, A., Barua, A., Botev, A., Castro-Ros, A., Slone, A., Héliou, A., Tacchetti, A., Bulanova, A., Paterson, A., Tsai, B., Shahriari, B., Lan, C. L., Choquette-Choo, C. A., Crepy, C., Cer, D., Ippolito, D., Reid, D., Buchatskaya, E., Ni, E., Noland, E., Yan, G., Tucker, G., Muraru, G.-C., Rozhdestvenskiy, G., Michalewski, H., Tenney, I., Grishchenko, I., Austin, J., Keeling, J., Labanowski, J., Lespiau, J.-B., Stanway, J., Brennan, J., Chen, J., Ferret, J., Chiu, J., Mao-Jones, J., Lee, K., Yu, K., Millican, K., Sjoesund, L. L., Lee, L., Dixon, L., Reid, M., Mikuła, M., Wirth, M., Sharman, M., Chinaev, N., Thain, N., Bachem, O., Chang, O., Wahltinez, O., Bailey, P., Michel, P., Yotov, P., Chaabouni, R., Comanescu, R., Jana, R., Anil, R., McIlroy, R., Liu, R., Mullins, R., Smith, S. L., Borgeaud, S., Girgin, S., Douglas, S., Pandya, S., Shakeri, S., De, S., Klimenko, T., Hennigan, T., Feinberg, V., Stokowiec, W., hui Chen, Y., Ahmed, Z., Gong, Z., Warkentin, T., Peran, L., Giang, M., Farabet, C., Vinyals, O., Dean, J., Kavukcuoglu, K., Hassabis, D., Ghahramani, Z., Eck, D., Barral, J., Pereira, F., Collins, E., Joulin, A., Fiedel, N., Senter, E., Andreev, A., and Kenealy, K. Gemma: Open models based on gemini research and technology, 2024.

Team, N., Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K. R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N. F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., and Wang, J. No language left behind: Scaling human-centered machine translation, 2022.

Tiedemann, J. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Graham, Y., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., and Negri, M. (eds.), *Proceedings of the Fifth Conference on Machine Translation*, pp. 1174–1182, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.wmt-1.139.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models, 2023.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Huggingface's transformers: State-of-the-art natural language processing, 2020.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. Hellaswag: Can a machine really finish your sentence?, 2019.

A. Appendix

Table 5. Compression Parity (CP) - mean values evaluated on the Flores 200 dataset.

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 6. Tokenization Parity 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 7. 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	0.2724	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 8. 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	0.2156	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 9. 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
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 10. Compression Parity (CP) evaluated on the Tatoeba dataset.

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 11. 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	CP Floress	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 12. 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), MMLU values are the reported results of MMLU on the 70B model in (Bendale et al., 2024).

Task/Metric	CP-Flores	Tokenizer P	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 13. Flores 200 used languages - language Names to codes

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 14. Tatoeba 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