Multilingual Evaluation of Generative AI

Abhivir Singh
Imperial College London
as9422@ic.ac.uk

1 Introduction

The advent of Large Language Models (LLMs) has revolutionized natural language processing, showcasing unprecedented capabilities across diverse tasks. However, ensuring these models serve global audiences equitably necessitates rigorous evaluation of their performance across multiple languages. This challenge is amplified by the linguistic diversity and disparity in resource availability among languages.

This review examines the state of multilingual evaluation for LLMs, focusing on advancements in benchmarking methodologies, metrics, and datasets. Papers were selected based on their prominence in top-tier venues, their citation impact, and their depth in addressing multilingual evaluation challenges. Special attention is given to issues such as data contamination, cross-task and cross-language performance variations, and emerging solutions like the Compression Parity (CP) metric. By synthesizing insights from these works, this review aims to provide a comprehensive understanding of the field's current state and future directions.

2 MEGA

2.1 Datasets and Languages

This paper (Ahuja et al., 2023) considers five families of NLP tasks across 16 datasets, spanning a wide variety of classification, question answering, sequence labelling, natural language generation, and responsible AI tasks. The datasets are listed in Figure 3. The study includes 70 languages across 21 language families, with Indo-Aryan (eg. Hindi, Bengali, Punjabi) and Afro-Asiatic (eg. Somali, Arabic) languages being the most represented (see Figure 4).

2.2 Evaluation Methodology

To assess the potential of Large Language Models (LLMs) across diverse linguistic tasks, this paper leverages two key properties of LLMs:

1) In-context learning: The ability to learn new tasks from a few examples without explicit fine-tuning.

2) Instruction following: The capability to perform tasks based on textual instructions.

This methodology involves providing the LLM with both a few-shot prompt (containing examples) and a task-specific instruction. The LLM is then tasked with predicting the output for a test example, which is always presented in the target language being evaluated across all three strategies.

2.2.1 Multilingual Prompting Strategies

The prompt structure remains consistent across all three strategies, comprising an instruction, a few-shot prompt, and a final test example (as shown in Figure 5). However, the strategies differ in how languages are incorporated into the prompt, enabling insights into the most effective language combinations: **Monolingual Prompting:** The k-shot examples and the test example are both in the same target language being evaluated. **Zero-Shot Cross-Lingual Prompting:** The k-shot examples are drawn from a pivot language (English, the best-performing language in our experiments), while the test example remains in the target language. Although this strategy involves a few-shot prompt in the pivot language, it is termed 'zero-shot' because no examples in the target language are provided to the model. This setup evaluates the model's ability to transfer task knowledge learned from one language to another without target-language-specific examples. **Translate-Test:** Similar to Zero-Shot Cross-Lingual, the k-shot examples are sampled from English. However, the test example is translated into English before being presented to the LLM for prediction.

Choice of Few-Shot Examples. In all experiments, the few-shot examples are randomly selected from the training or validation set of the dataset, depending on availability. The number of examples is fixed at $k = 8$, as this was empirically determined to be the optimal value. Performance was observed to improve with increasing k up to $k = 8$, after which it plateaued.

2.2.2 Evaluating Across Language Families & Tasks

As shown in the MEGAVERSE paper (Ahuja et al., 2024), to assess the performance of models across diverse language families and tasks, we introduce a deviation-based metric, $\Delta(i,j)$, which measures how well a model's performance on a given language family or task deviates from the average performance across all models. This allows us to identify which language families and tasks are well-supported or underrepresented.

The metric is calculated as the difference between the penalised score $p\_score(i,j)$ for each model on a given language family or task, and the average penalised score across all models:

$$\Delta(i,j) = p\_score(i,j) - \frac{1}{N}\sum_{i=1}^{N} p\_score(i,j)$$

The penalised or normalised score $p\_score(i,j)$ is computed by accounting for the data imbalance, using the following formula:

$$p\_score(i,j) = \left(\frac{|X_j|}{\sum_i |X_j|}\right) \times score_i$$

where $|X_j|$ represents the number of data instances for a given task or language family, and $score_i$ is the normalised performance score of the model for that task or family. This penalisation prevents the overrepresentation of tasks or languages with fewer data points, providing a more balanced evaluation of the models' capabilities.

By calculating $\Delta(i,j)$, we can effectively pinpoint areas where models excel or struggle, and highlight tasks or language families that may need further development in multilingual AI systems.

2.3 Key Findings

2.3.1 Comparison across different prompting Strategies

Zero-shot Cross-Lingual Performance. While the Zero-shot Cross-Lingual strategy performs comparably to Monolingual prompting for DV003, its effectiveness diminishes with GPT-3.5-Turbo, particularly for tasks involving extremely low-resource languages such as Quechua and Haitian Creole. This decline highlights the limitations of the model's ability to transfer knowledge from the pivot language (English) to highly underrepresented target languages.

Benefits of Monolingual Prompting. Grounding the model through Monolingual prompting enhances its comprehension of these low-resource languages, leading to noticeably better performance. This suggests that providing examples in the same language as the test instance offers the model a stronger contextual understanding, particularly for linguistically underrepresented languages.

Translate-Test Strategy. For low-resource languages, the results show that the Translate-Test strategy often outperforms Monolingual prompting. This is particularly evident in Figure 8, where low-resource languages (shown in lighter green) exhibit a significant performance improvement with Translate-Test compared to the Monolingual strategy.

2.3.2 Comparing different Models

The aggregated results comparing different models and prompting strategies are provided in Figure 7

GPT-3.5 (both DV003 and Turbo) underperforms relative to SOTA fine-tuned models, with the performance difference between the two being 20% on almost all datasets. The performance gap between GPT-4 and the SOTA models is narrower. Notably, GPT-4 performs well when queried directly in the target language for many high-resource and Latin script languages but it still lags significantly behind when queried in non-English languages. This is because of a lack of representation of non-english tokens in the training data of GPT-4, thus leading to a worse tokenisation value for them and a worse output. Overall, GPT-4 outperforms GPT-3.5 (Turbo), demonstrating emerging multilingual capabilities and a step in the right direction.

2.3.3 Comparison across Language Families & Tasks

We observe that languages in IE:Germanic Family (languages with Latin script like English), which ranks at the top, attain a significantly higher score than the mean, while at the the opposite end, Afro-Asiatic languages significantly underperform the mean across models and datasets.

We also find that the models tested are significantly better at tasks such as MCQ Reading Comprehension and Parts of Speech Tagging (across all languages), than more open tasks such as Q&A and text Summarisation, as seen in Figure 6.

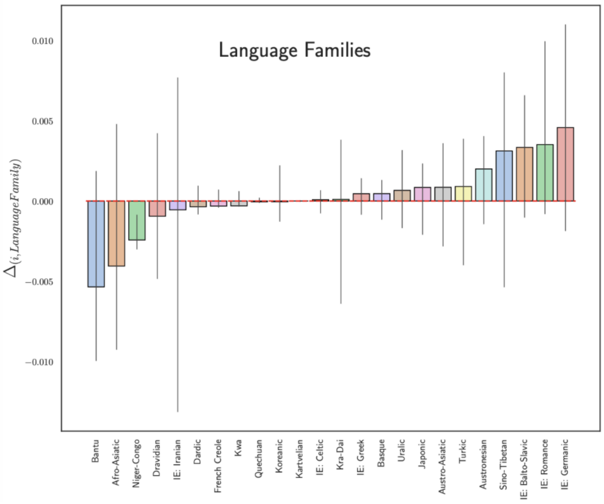


Figure 1: Deviation across language families. The positive scores of the bar-plots denote that the current LLMs are relatively good with those language families

2.4 Limitations

Data contamination issues: Many benchmarks are compromised when models are exposed to test data during training, resulting in inflated performance scores. As highlighted in (Ahuja et al., 2024), data contamination has affected nearly all datasets in the MEGA benchmark, particularly with the recent state-of-the-art models released in the past six months.

Task-specific benchmarks: Narrow focus; fail to generalise; human effort intensive or coarse grained like Summarisation tasks that used ROUGE-L.

Prompt-based evaluation drawbacks: Most benchmarks rely on natural language prompts, which can be highly sensitive to phrasing. This approach may not reflect real-world usage scenarios where users lack prompt engineering expertise

Thus, we see that these datasets are hard to create and are not very robust at all. *The challenges and limitations that we see in a benchmark like this one are directly addressed by* a new metric called *Compression Parity* that we will look at next.

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

To evaluate the efficiency of language models (LMs) across various languages, (Tsvetkov and Kipnis, 2024) introduce the metric **Compression Parity (CP)** in ICML 2024. CP quantifies the efficiency of a language model's representation of a document in a given language L, relative to its English translation, under the same language model. This metric helps assess the model's ability to compress and represent information across different languages.

3.1 Evaluation Methodology: Compression Parity

The Compression Parity of a document D in a language L is defined as the ratio of the negative log-likelihood (NLL) of the document's English translation to the NLL of the document in language L, as calculated by the same language model (LM):

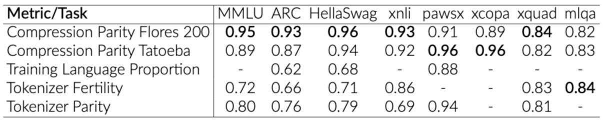
$$CP(L; D) \triangleq \frac{NLL(\text{English}(D))}{NLL(L(D))}$$

where: - $D = (w_1, \ldots, w_n)$ is a document represented as a sequence of tokens, - $NLL(D) = \sum_{i=1}^{n} - \log_2 \text{PLM}(w_i | w_{i-1})$ is the negative log-likelihood of the document D under the language model PLM, - $L(D)$ denotes the representation of D in language L.

In words, CP measures the relative compression efficiency between a document in language L and its English translation. A higher CP indicates better alignment of the language model's representation with the language-agnostic ideal, where the model can efficiently compress text across languages, capturing semantic content more effectively.

3.2 Key Findings

As seen in Figure 2, we find that CP correlates strongly with existing task-specific metrics, making it a valuable tool for ranking and comparing LLMs' multilingual capabilities.



Metric/Task	MMLU	ARC	HellaSwag	xnli	pawsx	xcopa	xquad	mlqa
Compression Parity Flores 200	**0.95**	**0.93**	**0.96**	**0.93**	0.91	0.89	**0.84**	0.82
Compression Parity Tatoeba	0.89	0.87	0.94	0.92	**0.96**	**0.96**	0.82	0.83
Training Language Proportion	-	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	-

Figure 2: Pearson correlation (absolute values) between metrics and downstream tasks/benchmarks performance under the LLM Llama 2 7B.

This metric allows us to compare how different languages perform in terms of compression efficiency, offering insights into the relative effectiveness of the language model in diverse linguistic contexts. By evaluating CP, we gain a clearer understanding of how well a model supports various languages, particularly in its ability to compress and represent them comparably to English.

3.3 Limitations of Compression Parity

While Compression Parity (CP) is a useful metric for evaluating multilingual LLMs, it has several limitations:

- **Influence of Instruction Tuning:** CP assumes ideal multilingual LLMs are language-agnostic, but instruction-tuned models—such as those trained with RLHF or DPO—may show different compression behavior due to changes in internal architecture (as noted in Ouyang:2022).

- **Focus on Compression, Not Task-Specific Performance:** CP evaluates compression efficiency but does not assess task-specific performance, such as the ability to follow instructions or perform tasks like translation and summarization, which are crucial for some applications.

These limitations suggest the need for refining Compression Parity, particularly considering instruction-tuning and language-specific factors.

3.4 Future Directions

Future research could explore the following areas to enhance the understanding of LLM performance across languages:

- **Impact of Instruction Tuning on Compression Parity:** Investigating how instruction tuning affects CP values could reveal how task-specific optimization influences multilingual efficiency.

- **Tokenization and Training Language Distribution:** Studying the effects of tokenization and language distribution during training on CP will help assess how these factors impact multilingual performance.

- **Language-Specific Variations in Compression Efficiency:** Further research could explore how syntactic and morphological differences across languages affect compression efficiency and how CP might be adapted to reflect these variations.

By addressing these gaps, future work can refine the metric and enhance its application in evaluating and improving multilingual LLM performance, ultimately leading to more accurate and fair assessments.

4 Conclusion

The reviewed literature highlights significant advancements in multilingual evaluation for LLMs, with diverse approaches addressing the challenges of linguistic diversity. Benchmarks like MEGA emphasize cross-language task performance, while innovative metrics like Compression Parity (CP) provide task-agnostic insights into model efficiency across languages. However, persistent issues such as data contamination, task-specific evaluation limitations, and the complexities of instruction-tuned models reveal gaps in current methodologies.

Overall, the field is evolving toward more inclusive and robust evaluation frameworks, yet further refinement is needed. Future work should focus on mitigating biases in training data, understanding the impact of instruction tuning on multilingual performance, and accounting for language-specific characteristics in evaluation metrics. These steps will enhance the fairness, accuracy, and applicability of multilingual AI systems.

References

Siddharth Ahuja, Divyansh Aggarwal, Venkatesh Gumma, Isaac Watts, Aashish Sathe, Moses Ochieng, Rohit Hada, Parth Jain, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2024. Megaverse: Benchmarking large language models across languages, modalities, models and tasks.

Alexander Tsvetkov and Alon Kipnis. 2024. *Multilingual Compression Parity: How Efficiently Large Language Models Represent Information Across Languages?* In *ICML 2024 Workshop on Theoretical Foundations of Foundation Models*. Available at: `https://openreview.net/forum?id=LZt1WbkCiT`.

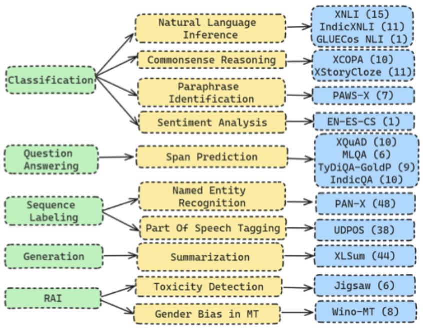
Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI.

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

Sanchit Ahuja, Varun Gumma, and Sunayana Sitaram. 2024. Contamination Report for Multilingual Benchmarks. Microsoft Research.

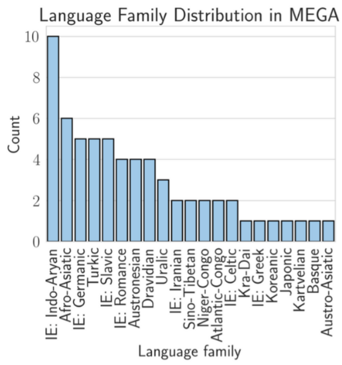
@articleOuyang:2022, author = Long Ouyang and Jeff Wu and Xu Jiang and Diogo Almeida and Carroll L. Wainwright and Pamela Mishkin and Chong Zhang and Sandhini Agarwal and Katarina Slama and Alex Ray and John Schulman and Jacob Hilton and Fraser Kelton and Luke Miller and Maddie Simens and Amanda Askell and Peter Welinder and Paul Christiano and Jan Leike and Ryan Lowe, title = Training Language Models to Follow Instructions with Human Feedback, journal = OpenAI, year = 2022, url = https://arxiv.org/abs/2203.02155,

5 Appendix



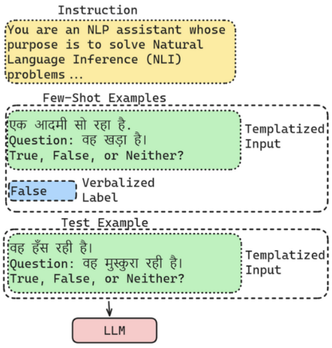
(a) Tasks and Datasets included in MEGA.

Figure 3: Tasks & Datasets included in the MEGA dataset.



(b) Language Family Distribution

Figure 4: Language families included in the MEGA dataset.



(c) Example of multilingual prompting

Figure 5: Example of a prompt in the MEGA dataset.

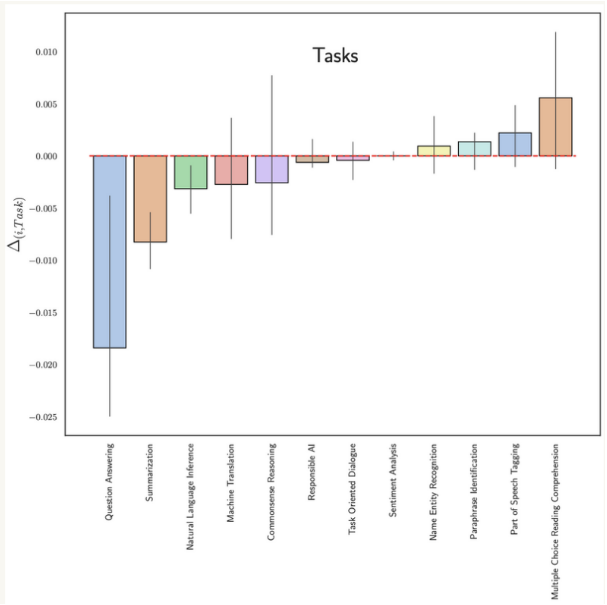
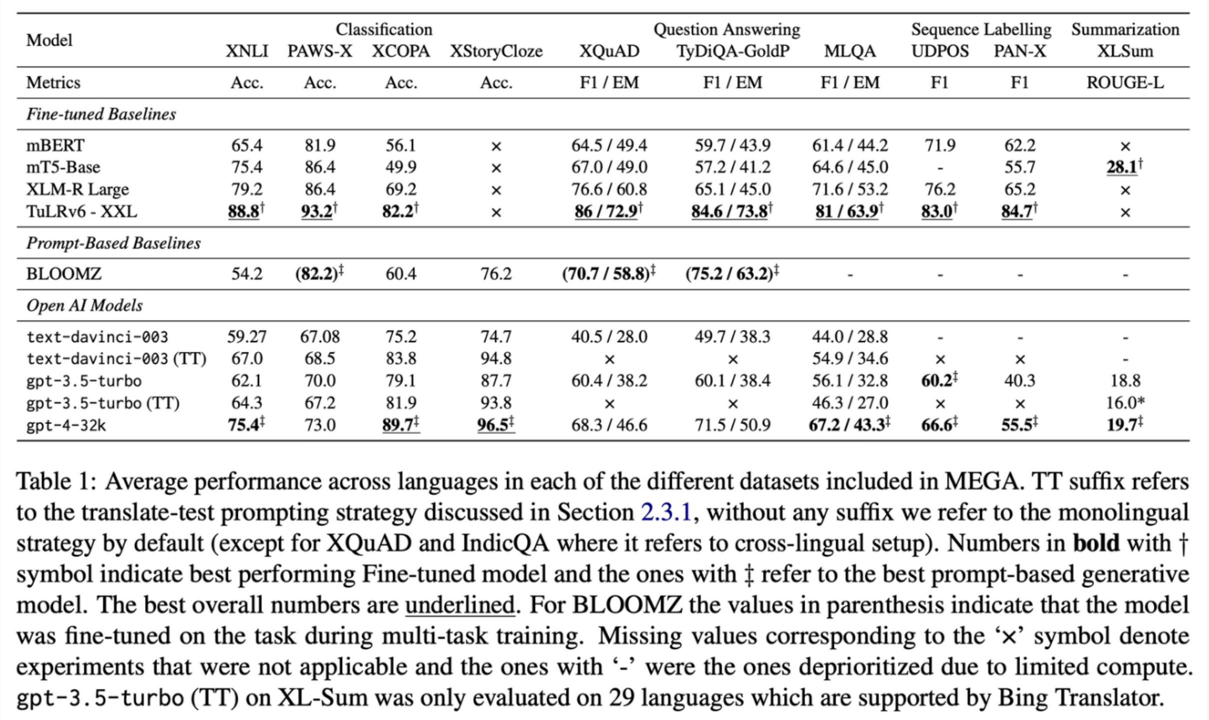


Figure 6: Deviation across tasks



Model	Classification XNLI	 PAWS-X	 XCOPA	 XStoryCloze	Question Answering XQuAD	 TyDiQA-GoldP	 MLQA	Sequence Labelling UDPOS	 PAN-X	Summarization XLSum
Metrics	Acc.	Acc.	Acc.	Acc.	F1 / EM	F1 / EM	F1 / EM	F1	F1	ROUGE-L
Fine-tuned Baselines										
mBERT	65.4	81.9	56.1	×	64.5 / 49.4	59.7 / 43.9	61.4 / 44.2	71.9	62.2	×
mT5-Base	75.4	86.4	49.9	×	67.0 / 49.0	57.2 / 41.2	64.6 / 45.0	-	55.7	**28.1**†
XLM-R Large	79.2	86.4	69.2	×	76.6 / 60.8	65.1 / 45.0	71.6 / 53.2	76.2	65.2	×
TuLRv6 - XXL	**88.8**†	**93.2**†	**82.2**†	×	**86 / 72.9**†	**84.6 / 73.8**†	**81 / 63.9**†	**83.0**†	**84.7**†	×
Prompt-Based Baselines										
BLOOMZ	54.2	**(82.2)**‡	60.4	76.2	**(70.7 / 58.8)**‡	**(75.2 / 63.2)**‡	-	-	-	-
Open AI Models										
`text-davinci-003`	59.27	67.08	75.2	74.7	40.5 / 28.0	49.7 / 38.3	44.0 / 28.8	-	-	-
`text-davinci-003` (TT)	67.0	68.5	83.8	94.8	×	×	54.9 / 34.6	×	×	-
`gpt-3.5-turbo`	62.1	70.0	79.1	87.7	60.4 / 38.2	60.1 / 38.4	56.1 / 32.8	**60.2**‡	40.3	18.8
`gpt-3.5-turbo` (TT)	64.3	67.2	81.9	93.8	×	×	46.3 / 27.0	×	×	16.0*
`gpt-4-32k`	**75.4**‡	73.0	**89.7**‡	**96.5**‡	68.3 / 46.6	71.5 / 50.9	**67.2 / 43.3**‡	**66.6**‡	**55.5**‡	**19.7**‡

Table 1: Average performance across languages in each of the different datasets included in MEGA. TT suffix refers to the translate-test prompting strategy discussed in Section 2.3.1, without any suffix we refer to the monolingual strategy by default (except for XQuAD and IndicQA where it refers to cross-lingual setup). Numbers in **bold** with † symbol indicate best performing Fine-tuned model and the ones with ‡ refer to the best prompt-based generative model. The best overall numbers are underlined. For BLOOMZ the values in parenthesis indicate that the model was fine-tuned on the task during multi-task training. Missing values corresponding to the '×' symbol denote experiments that were not applicable and the ones with '-' were the ones deprioritized due to limited compute. `gpt-3.5-turbo` (TT) on XL-Sum was only evaluated on 29 languages which are supported by Bing Translator.

Figure 7: Performance analysis by model

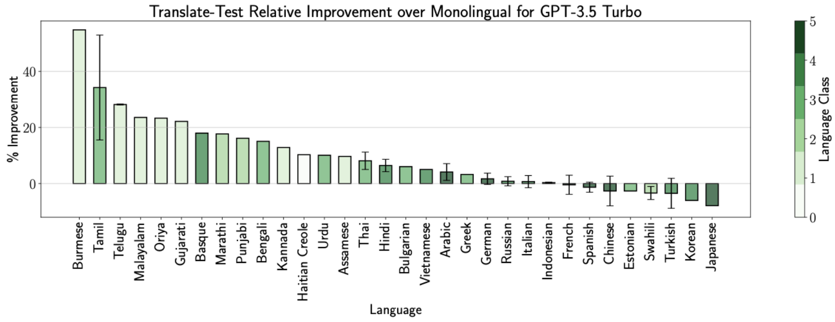


Figure 8: Relative percentage improvement over Monolingual prompting when using Translate-Test for GPT-3.5- Turbo. The bars are colour-coded based on the class taxonomy provided in (Joshi et al., 2020).