

Figure 4: Tokenizer Fertility for OpenAI models, mBERT, and BLOOM for different languages

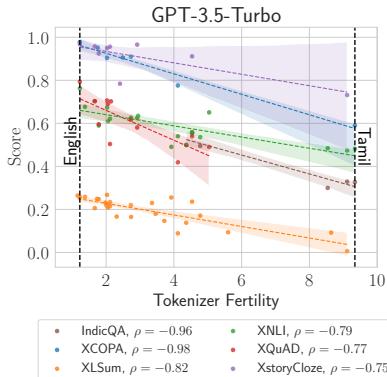


Figure 5: Correlation between the performance of GPT-3.5-Turbo with the tokenizer fertility. We report the curves for the cases where the person coefficient $|\rho| > 0.7$ with a p-value of 0.05. We have combined Indic-XNLI and XNLI for a better coverage of languages. Similar plots for GPT-4 can be found in Figure 7b of Appendix.

tically significant (negative) correlations between the tokenizer’s fertility and dataset-specific performance i.e. the models obtain worse performance on languages for which the tokenizer is of poor quality, and vice-versa.

We also study the effect that the amount of data available for each language during pre-training (Wu and Dredze, 2020; Lauscher et al., 2020) has on the multilingual performance of these models. We measure the correlations between the language-wise number of tokens present in the pre-training data with language-wise performance on each dataset. While the exact language-wise pre-training data distribution for GPT-3.5 and GPT-4 models is not available, we use the GPT-3’s language-wise pretraining distribution as a proxy. We observe that for four tasks (PAWS-X, XNLI, XCOPA, and XQuAD) statistically significant positive correlations between the pre-training data size and performance. Note that, the amount of pre-training data and tokenizer fertility are highly likely to be cor-

related with each other. However, we do see that using pre-training data we are able to explain some trends that are not explained by tokenizer fertility alone. For example, even though the OpenAI models have similar tokenizer fertilities for both French and Japanese, these models perform much better in French than they do for Japanese (72.1% accuracy vs 67% accuracy for GPT-3.5-Turbo) for PAWS-X. However, when we take into consideration the amount of pre-training data for these languages: roughly 3.5 B French tokens in the pre-training data versus 214M for Japanese, we can partially explain this discrepancy.

However, we must note that these two factors correlate well with only a subset of the tasks and what we are measuring is the correlation which might not imply causation. Investigating different factors that together more holistically explain multilingual capabilities is an important direction that we leave for future work. Please check Appendix §A.6 for detailed results from this section.

4 Challenges in Multilingual Evaluation

In this section, we examine some of the challenges and consequent limitations of a large-scale multilingual evaluation like ours.

4.1 A Kaleidoscope of Choices.

There are various moving parts when evaluating LLMs using prompting-based approaches, including the choice of prompt templates, instructions, and few-shot examples (Liu et al., 2022; Lu et al., 2022; Zhao et al., 2021), different prompting strategies (Wei et al., 2023; Nye et al., 2021; Ye and Durrett, 2022a), using external tools (Schick et al., 2023), the language of prompts (Shi et al., 2022; Lin et al., 2022a), as well as different decoding specific hyper-parameters (Shih et al., 2023), which can have varying degrees of impact on the performance, sometimes in unexpected ways. Holisti-

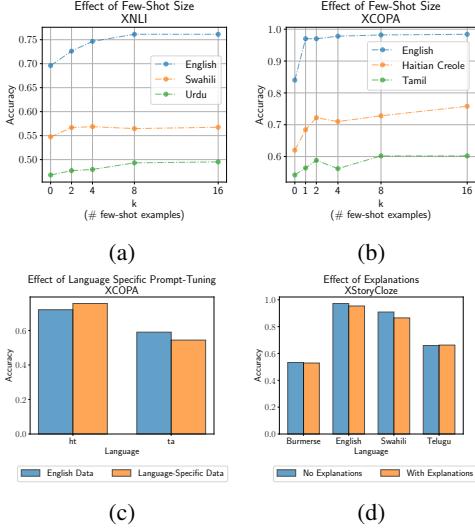


Figure 6: Analysing the effect on GPT-3.5-Turbo’s performance given different evaluation factors. To obtain explanations we use Super-Natural Instructions (Wang et al., 2022).

cally exploring these choices for all the datasets and languages can quickly get out of hand, especially given the excessive computational cost of running these models. In order to understand the sensitivity of our observations to the choices we make in §3, we re-evaluate our setups on a subset of datasets and languages for a varying set of parameters. Our findings are summarized in Figure 6, where we see that having a large few-shot size generally helps improve performance, however, the performance is often stable beyond $k = 8$. Further, language-specific fine-tuning can help improve the performance like we see for Haitian Creole in XCOPA, but for Tamil we actually observe the accuracy to go down which might be attributed to the small size of the validation set (100 in the case of XCOPA). Finally, on XStoryCloze dataset (also for XCOPA), we see using explanations to prompt the models have negligible impact on the performance. Overall, these experiments indicate that the existing prompting approaches might not be sufficient to address the performance gap that exists for non-English languages (especially mid-to-low resource languages) and there is an imminent need to propose new methods as well as improve the representation of different languages in these model’s pre-training (and instruction-tuning) data.

4.2 Test data contamination

Given the massive amount of online data that LLMs are trained with, it is critical to factor in the possibility of contamination of test datasets (Sainz et al.,

2023). Accordingly, we attempt to verify if the performances we observed are in fact, representative of the capabilities of these models or merely a result of memorization. Given the lack of transparency in the training distribution of recent models like GPT-4, we perform some preliminary investigations against this phenomenon. Specifically, we consider three factors: i) LLM’s knowledge of the dataset, ii) availability of test datasets on the internet, and iii) dataset release date.

To measure the LLM’s (we do this for GPT-4) memory of the dataset, we prompt it to fill the dataset cards for each of MEGA’s datasets (denoted as Card Fill). This involves filling template-like information like the task’s supported languages, input-output structure and description. If the model fills a dataset card correctly (Full), we note this as suspicion of contamination. If it fills the card partially correct (Partial) i.e. detecting either the correct task structure or correct set of languages, we mark it as partial evidence, and if it succeeds in neither, we mark it as no suspicion (None). For test dataset availability, we check if the test dataset can be accessed online directly without downloading either as part of the official release from the authors or via other sources such as Hugging Face dataset viewer (Data Acc. w/o Down.). For release date, we check if the dataset was made public after the cut-off date of September 2021.

The overall results from this analysis are provided in Table 2. We see that for a majority of datasets, GPT-4 can fill in the dataset card correctly; On the more recent datasets like XLSum and XStoryCloze it is only partially successful, while on Jigsaw and code-mixing datasets it fails to correctly fill the cards. Note that except XStoryCloze, Jigsaw and the Code-mixing datasets, evaluation sets for all other datasets are directly accessible online. Collectively, this connotes that for tasks like XStoryCloze and IndicQA there is a weak suspicion against contamination. While all other tasks are highly likely contaminated (except Jigsaw, and Code-Mixed datasets).

Implications. Our analysis implies a notable chance of the test data appearing in the training datasets of these LLMs. The contamination of test datasets is a serious problem for works centered around LLM evaluation (including ours), as they might lead to an overestimation of the capabilities of these models. However, we would like to highlight that despite the possibility of contamina-

Dataset	Card Fill	Data Acc. w/o Down.	Release Date
XNLI	Full	Yes	September 2019
Indic-XNLI	Full	Yes	April 2022
PAWS-X	Full	Yes	August 2019
XCOPA	Partial	Yes	April 2020
XStoryCloze	Partial	No	May 2023
XQuAD	Full	Yes	October 2019
MLQA	Full	Yes	October 2019
TyDiQA-GoldP	Full	Yes	February 2020
IndicQA	Partial	Yes	September 2022
PAN-X	Full	Yes	July 2017
UDPOS	Full	Yes	March 2020
XLSum	Partial	Yes	June 2021
Jigsaw	None	No	February 2020
GLUECos NLJ	None	No	June 2020
EN-ES-CS	None	No	May 2016

Table 2: Contamination analysis for the datasets that we consider in MEGA. We use red color when there is a strong suspicion of contamination based on these three metrics, green for no suspicion, and yellow for partial.

tion, LLMs still vastly underperform on (especially low-resource) non-English languages . These observations about data contamination indicate that the disparity in performance between English and non-English languages might be even greater than what we observe in our work.

5 Related Work

Evaluation of LLMs. A growing interest in the evaluation of LLMs has harbingered several efforts towards the holistic evaluation of their capabilities. While work like BIG-bench Srivastava et al. (2023) cover a diverse range of tasks, the non-English tasks are mostly translation-oriented which limit the more general task based inferences that for such an evaluation. Similarly, Liang et al. (2022) propose a taxonomy of scenarios and metrics in Holistic Evaluation of Language Models (HELM) to define the space of LLM evaluation, and evaluate 30 language models on 42 scenarios and 7 metrics. However, all the scenarios are focused on datasets in standard English or its dialects.

Multilingual Benchmarks and Evaluation. Benchmarks for multilingual evaluation, such as XTREME (Hu et al., 2020), XTREME-R (Ruder et al., 2021) and XGLUE (Liang et al., 2020) have been proposed to measure cross-lingual transfer in pre-trained language models. Following their popularity, there has been the development of benchmarks covering specific language families, such as IndicXTREME (Doddapaneni et al., 2022) for Indian languages, Adelani et al. (2022) for African Languages, and Wilie et al. (2020) for Indonesian languages, as well. The evaluations on these bench-

marks have mainly focused on pre-train then fine-tune kinds of setups. Particularly for prompting style evaluation, Bang et al. (2023) evaluates the multilingual capabilities of ChatGPT and shows that it fails to generalize to low-resource languages with non-latin scripts. However, multilingual evaluation is performed only on a few tasks, and a subset of 50-100 examples are used for testing the model. Hendy et al. (2023) evaluate the translation abilities of GPT-3.5 models and find that these models, while perform well in translating high-resource languages, their capabilities for low-resource languages are limited. Concurrent work BUFFET (Asai et al., 2023) and Lai et al. (2023) also perform multilingual benchmarking of large language models, however, they evaluate the performance of ChatGPT and BLOOMZ in their work while our evaluation also spans GPT-4.

Multilingual Prompting: While most work on prompting or in-context learning in LLMs focuses on English data, recently, there has been some interest in prompting them with non-English data. Zhao and Schütze (2021), for instance, use discrete and soft prompting techniques to evaluate XLM-RoBERTa and show that prompting can be more effective compared to fine-tuning when the amount of labeled data is limited. Lin et al. (2022a) show that English prompts perform better than prompts written in the target language (both hand-written and translated). Finally, (Shi et al., 2022) show chain-of-thought (CoT) prompting results leads to striking multilingual reasoning capabilities in LLMs, even in under-represented languages especially when prompted when English CoT.

6 Conclusion

In this work, we conduct an evaluation across different prompting strategies, models, tasks, and languages to investigate the multilingual capabilities of LLMs. We also investigate underlying properties like tokenizer quality and size of pretraining data to explain the trends in performance that we observe. Our investigation shows the consistent performance gap between high-resource, Latin script, and under-resourced languages in addition to highlighting the efficacy, yet limited sufficiency of methods like translate-test prompting. Through our evaluation, we present evidence of the need to prioritize automatic benchmarking and human evaluation across as many languages as possible. We hope that this work spurs research in meeting this goal.