MEGAVERSE: Benchmarking Large Language Models Across Languages, Modalities Models and Tasks

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, Sunayana Sitaram

Microsoft Corporation

Motivation

There has been a surge in LLM evaluation research to understand LLM capabilities and limitations. However, much of this research has been confined to English, leaving LLM building and evaluation for non-English languages relatively unexplored. Several new LLMs have been introduced recently, necessitating their evaluation on non-English languages. There has also been a need to detect and handle contamination of the current benchmarks that are used for evaluation.

Contributions

- Expanded the MEGA suite to include 6 new datasets and benchmarked nine new text LLMs such as PaLM2, Llama2, Mistral, GPT-4, Gemini etc. along with multimodal LLMs such as LLaVA family as well
- Provided a methodology to analyze and study the overall picture of this exercise
- Presented a thorough contamination analysis of both open-source and commercial LLMs

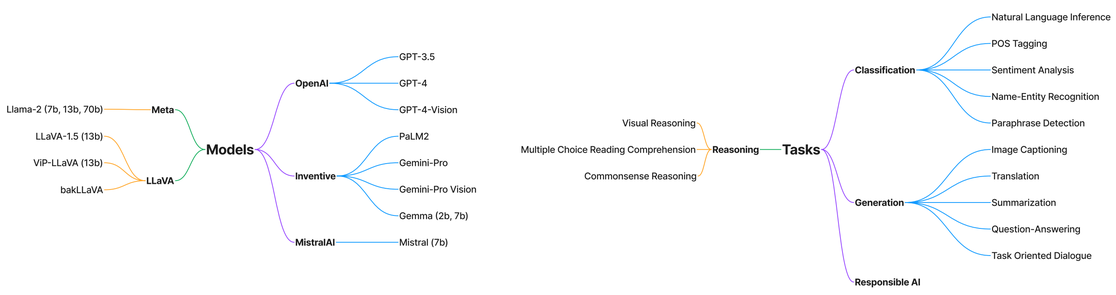

Figure 1. Hierarchy of Models and Tasks spread across MEGAVERSE

Benchmarking Results

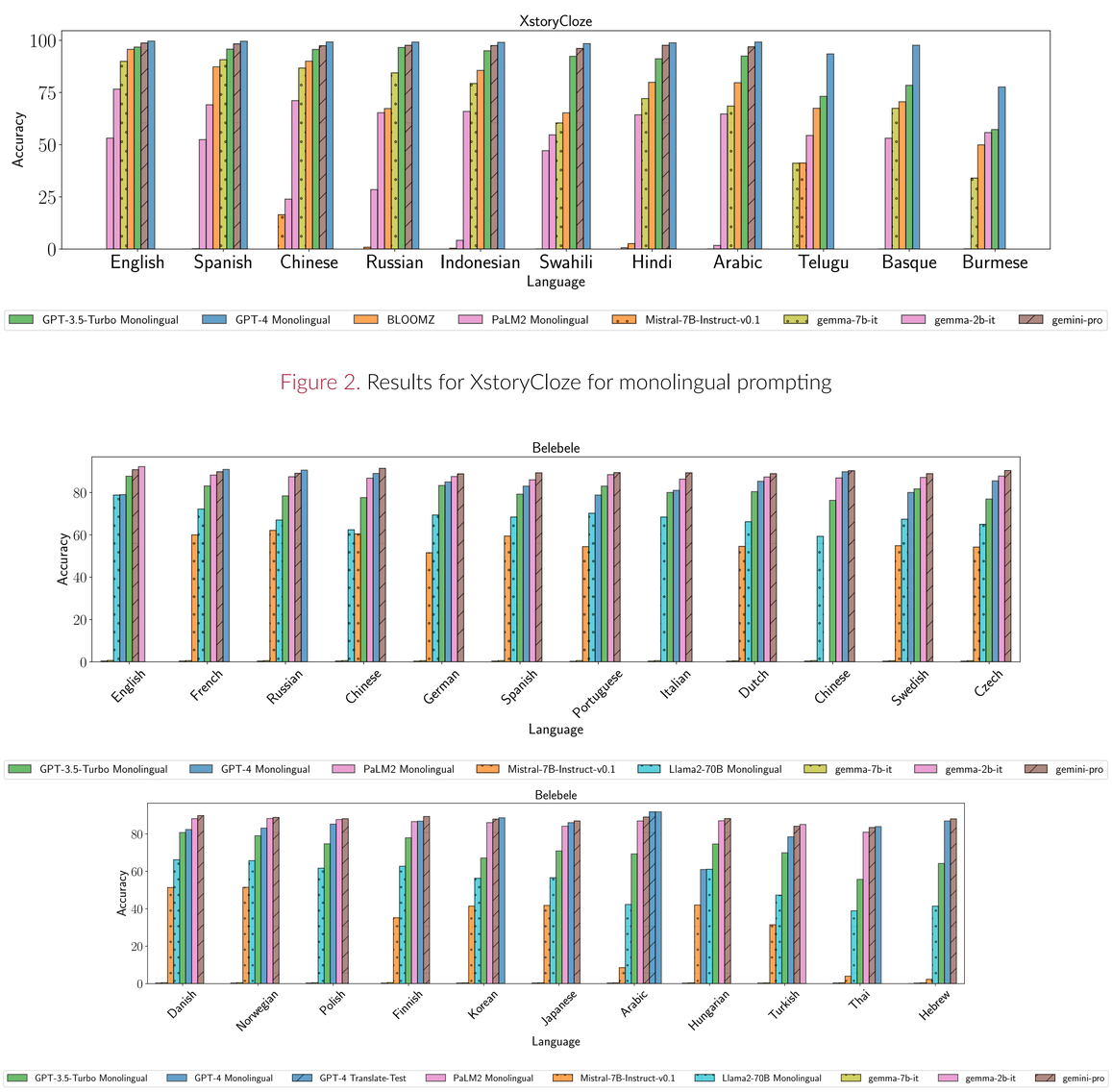

Figure 2. Results for XstoryCloze for monolingual prompting

Figure 3. Results for Belebele across all languages and models for monolingual prompting

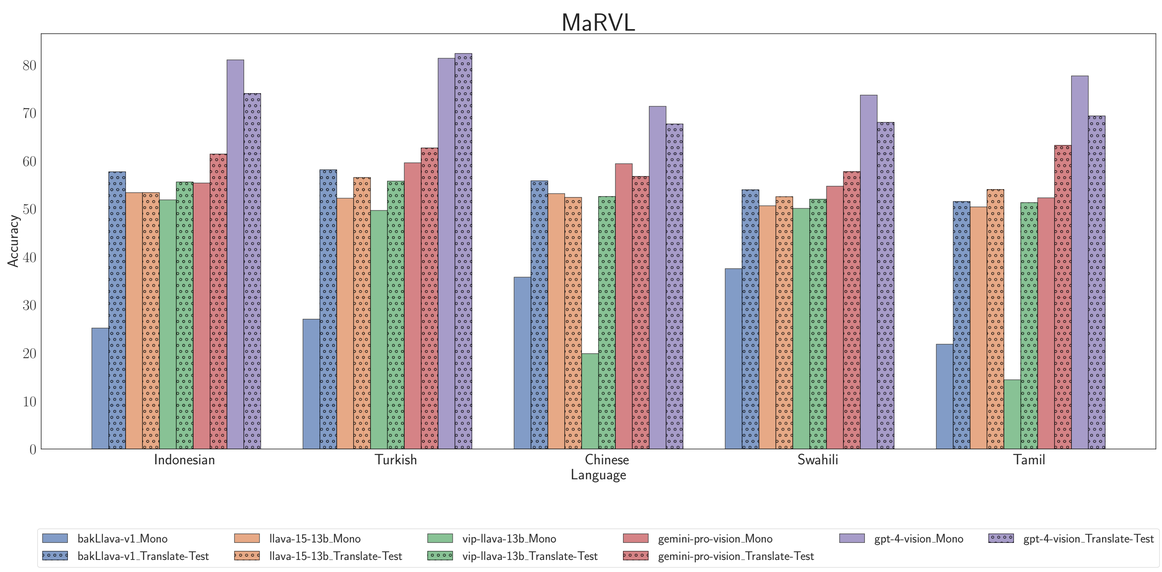

Figure 4. ChrF scores for the LLaVA models, GPT4-Vision, and Gemini-Pro-Vision on XM-3600 using monolingual prompting

The deviation of performance across language families and tasks

- Calculated the deviation of a given experiment i in each Language Family or Task j:

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

where $p\_score_{(i,j)}$ is the penalized score for the experiment i and is calculated by:

$$p\_score_{(i,j)} = (\frac{|X_j|}{\sum_i |X_j|}) * score_i$$

$score_i$ is the normalized score of given Task or a Language Family and $|X|$ stands for the number of samples

- The motivation to apply size penalization stems from the sparsity of the language,dataset and model combinations. We apply this penalization to limit the biase of outliers and combinations with limited support.

Deviation Scores Visualized

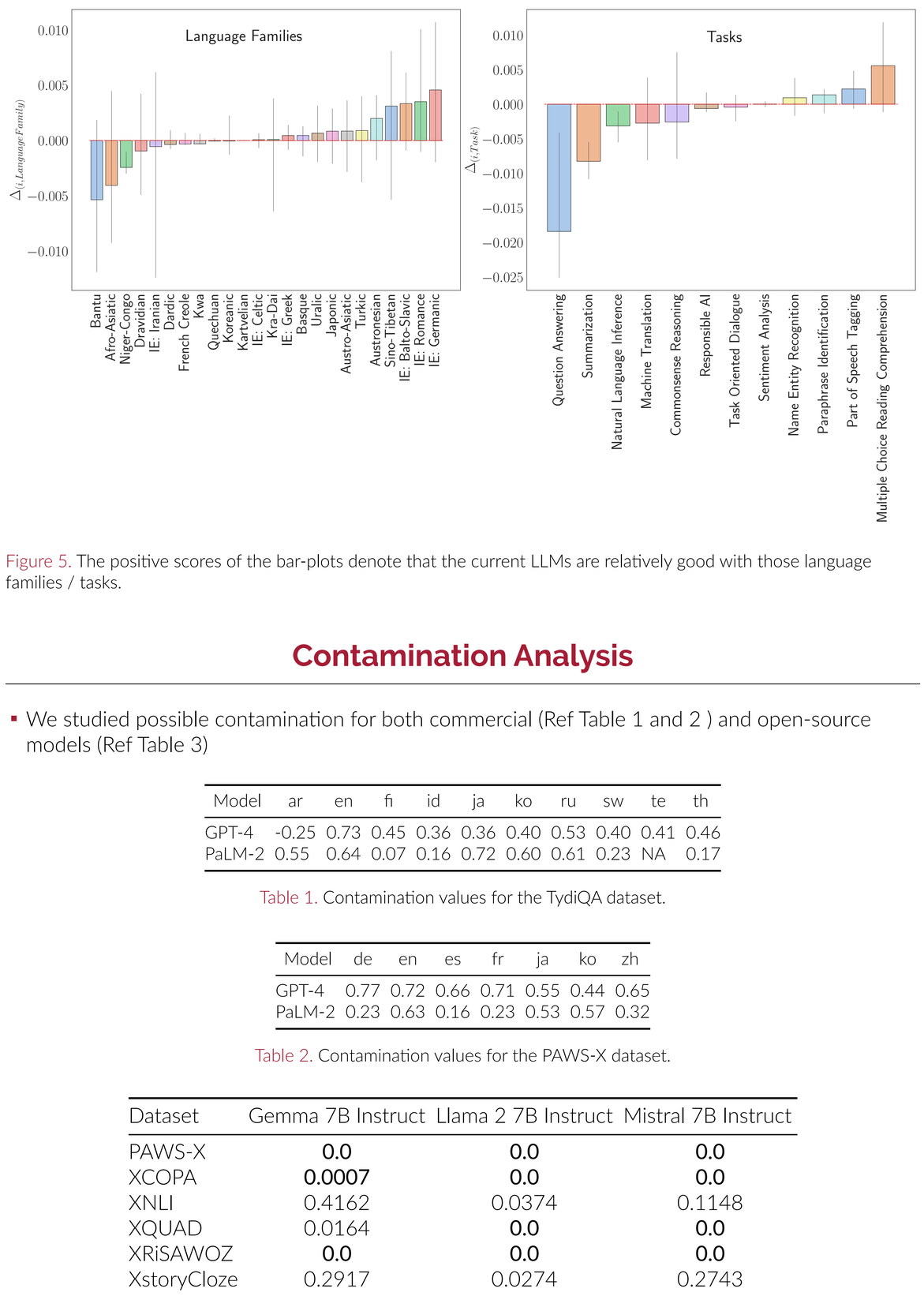

Figure 5. The positive scores of the bar-plots denote that the current LLMs are relatively good with those language families / tasks.

Contamination Analysis

- We studied possible contamination for both commercial (Ref Table 1 and 2) and open-source models (Ref Table 3)

Model	ar	en	fi	id	ja	ko	ru	sw	te	th
GPT-4	-0.25	0.73	0.45	0.36	0.36	0.40	0.53	0.40	0.41	0.46
PaLM-2	0.55	0.64	0.07	0.16	0.72	0.60	0.61	0.23	NA	0.17

Table 1. Contamination values for the TydiQA dataset.

Model	de	en	es	fr	ja	ko	zh
GPT-4	0.77	0.72	0.66	0.71	0.55	0.44	0.65
PaLM-2	0.23	0.63	0.16	0.23	0.53	0.57	0.32

Table 2. Contamination values for the PAWS-X dataset.

Dataset	Gemma 7B Instruct	Llama 2 7B Instruct	Mistral 7B Instruct
PAWS-X	**0.0**	**0.0**	**0.0**
XCOPA	**0.0007**	**0.0**	**0.0**
XNLI	0.4162	0.0374	0.1148
XQUAD	0.0164	**0.0**	**0.0**
XRiSAWOZ	**0.0**	**0.0**	**0.0**
XstoryCloze	0.2917	0.0274	0.2743

Table 3. The statistical test was performed on a total of 5000 test points equally divided amongst all the languages of a given dataset. Our significance value is 0.001 which is calculated using $1/(1+r)$, where r is the number of permutations per shard (for us it is 700). If a value is less than 0.001, then that test set is contaminated for the given model. The it suffix for the above model stands for Instruction-Tuned variant of that said model.

Takeaways

- **Comprehensive Benchmarking**
  - Evaluated 22 datasets spanning 83 languages across various models including GPT-4, PaLM2, Gemma, and Mistral
- **Performance Disparity**
  - Larger commercial models generally outperformed smaller open-source ones, especially in low-resource language scenarios
- **GPT-4 Dominance**
  - GPT-4 emerged as the top performer across most tasks in our study.
  - GPT-4-Vision surpassed Gemini-Vision and LLaVA on multimodal datasets.
- **Critical Contamination Insights**
  - Contamination analysis on both commercial and open-source models underscored its criticality.
  - Emphasizes the importance of safeguarding new evaluation datasets from inadvertent inclusion in training sets.

sanchitahuja205@gmail.com

NAACL 2024, Mexico City

sunayana.sitaram@microsoft.com