Microsoft

MARI Seminar

28 June 2023

MEGA: Multilingual Evaluation of Generative AI

Presenters:

Kabir Ahuja – Microsoft Research India (MSRI)

Millicent Ochieng – Microsoft Africa Research Institute (MARI)

Project Team:

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, Sunayana Sitaram (Lead)

MEGA Paper is under review at the EMNLP

Microsoft

Research Overview

Introduction

Generative AI models (LLMs) are AI systems that leverage large-scale training data to generate human-like text.

Recently, LLMs (GPT*) have demonstrated remarkable proficiency across various Natural Language Processing (NLP) tasks, including language comprehension, logical reasoning, and text generation.

This is now transforming a wide range of NLP applications.

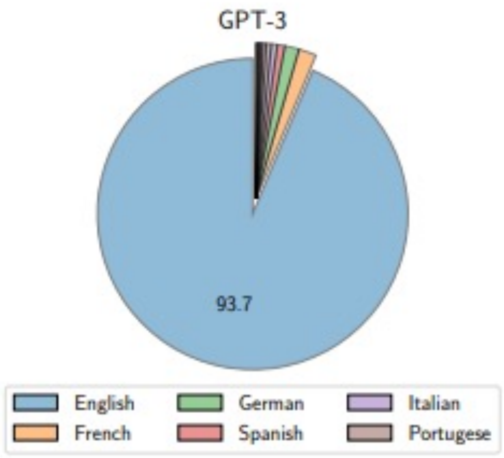
But **how well do GPT* models perform on Languages of the world?**

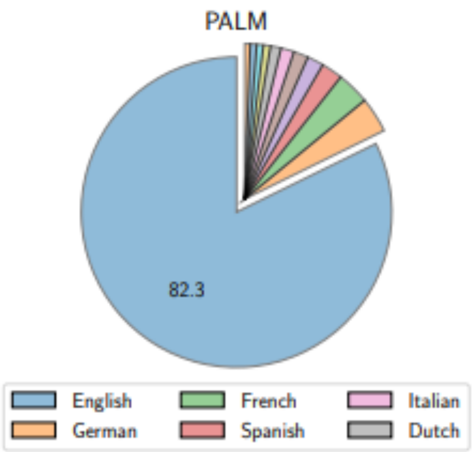
LLMs Training Data

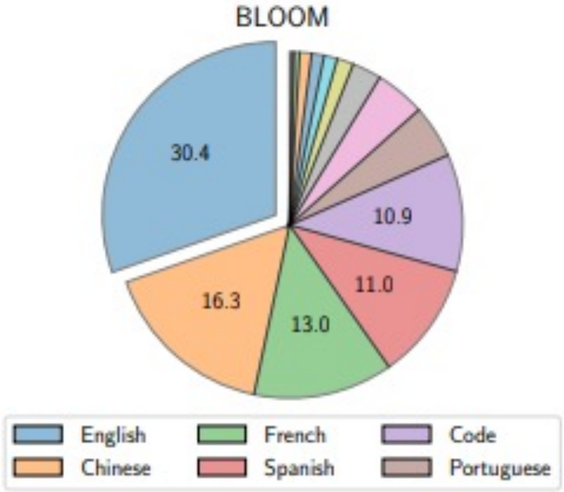
LLMs training data is **primarily** English content created in Global North. But **~ 6 billion people do not speak English, as their first language**.

This raises questions about the proficiency of LLMs in understanding and generating text in **other languages**, and what might this mean for non-English-speaking regions worldwide.

It is **crucial to evaluate multilingual capabilities** of these models as performance gains in high-resource languages may not generalize to all languages.







Advancing Multilingual Evaluation of Generative AI: MEGA

Microsoft

MEGA benchmark: Introduce a comprehensive evaluation of generative LLMs on 70 typologically diverse languages, covering 16 tasks and 4 LLMs i.e., *GPT-3.5 models (text-davinci-003 and gpt-3.5-turbo), GPT-4 and BLOOMZ*.

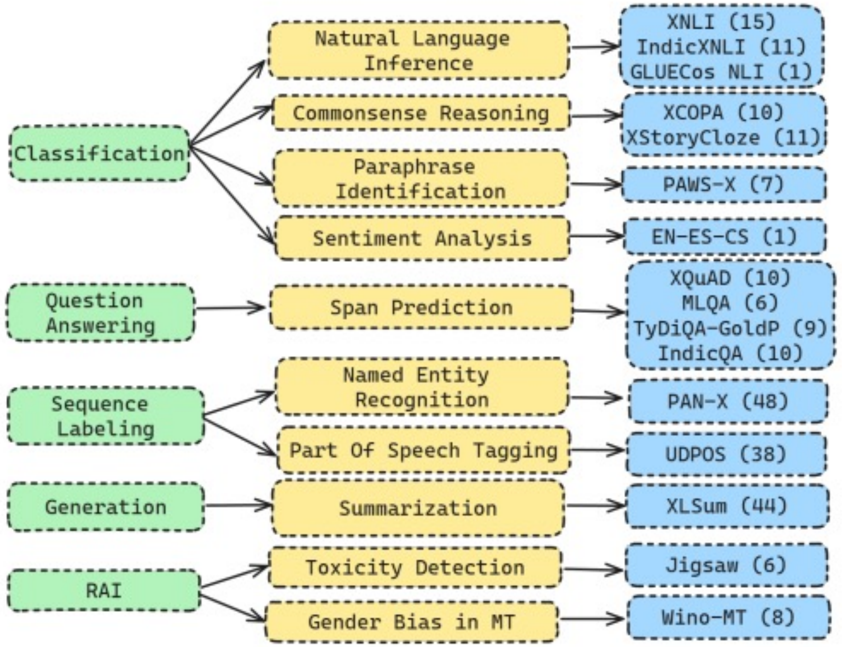
Performance comparison: Compare generative LLMs with state-of-the-art non-autoregressive models such as TULRv6, MuRIL to assess their effectiveness.

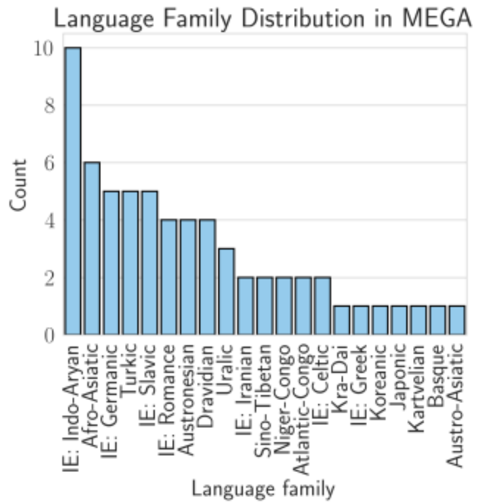
Optimal prompting strategies for non-English languages: Recommend effective strategies for using generative LLMs in diverse linguistic contexts, enhancing performance.

Microsoft

MEGA: Tasks, Datasets & Languages

Tasks & Datasets in MEGA





MEGA: LLMs

OpenAI models:

- GPT-3.5 text-davinci-003, supporting 4096 tokens,
- GPT-3.5-turbo, supporting 16k tokens,
- GPT-4 model, supporting 32k tokens.

Prompt-based Baselines:

- BLOOMZ

SOTA Fine-tuned Baselines:

- TULRv6
- XLM-R
- mT5
- MuRIL

Microsoft

Evaluation Methodology

MEGA Framework: **The Prompt Approach**

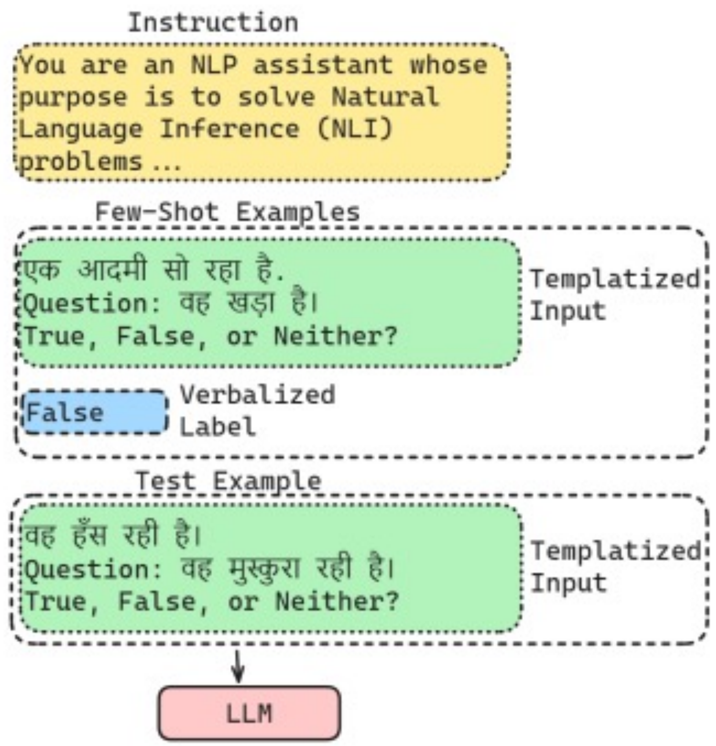
- We adopt the **prompt-based approach** to evaluate LLMs on multilingual benchmark.

- We use **Promptsource** for prompt tuning.

- Prompting Strategies
  - Monolingual
  - Zero-Shot Cross Lingual
  - Translate Test

Microsoft

MEGA Framework: **Prompting Structure**



Example of multilingual prompting

MEGA Framework: **Prompting Examples**

Microsoft

A.4.1 XNLI, IndicXNLI, GLUECoS NLI

Models : GPT-3.5-Turbo, GPT-4

Task Instruction $\mathcal{I}$: You are an NLP assistant whose purpose is to solve Natural Language Inference (NLI) problems. NLI is the task of determining the inference relation between two (short, ordered) texts: entailment, contradiction, or neutral. Answer as concisely as possible in the same format as the examples below:

Template f_{temp}:
`{premise}`
Question: `{hypothesis}`
True, False, or Neither?

Verbalizer f_{verb}:
Entailment : True,
Contradiction: False,
Neutral: Neither

Models : DV003

Template f_{temp}:
`{premise}` Based on previous passage is it true that `{hypothesis}` ? Yes, No, or Maybe?

Verbalizer f_{verb}:
Entailment : Yes,
Contradiction: No,
Neutral: Maybe

A.4.4 XQUAD, TyDiQA, MLQA

Models : GPT-3.5-Turbo, GPT-4

Task Instruction $\mathcal{I}$: You are an NLP assistant whose purpose is to solve reading comprehension problems. You will be provided questions on a set of passages and you will need to provide the answer as it appears in the passage. The answer should be in the same language as the question and the passage.

Template f_{temp}:
`{context}`
Q: `{question}`
Referring to the passage above, the correct answer to the given question is: `{answer}`

A.4.10 XLSum

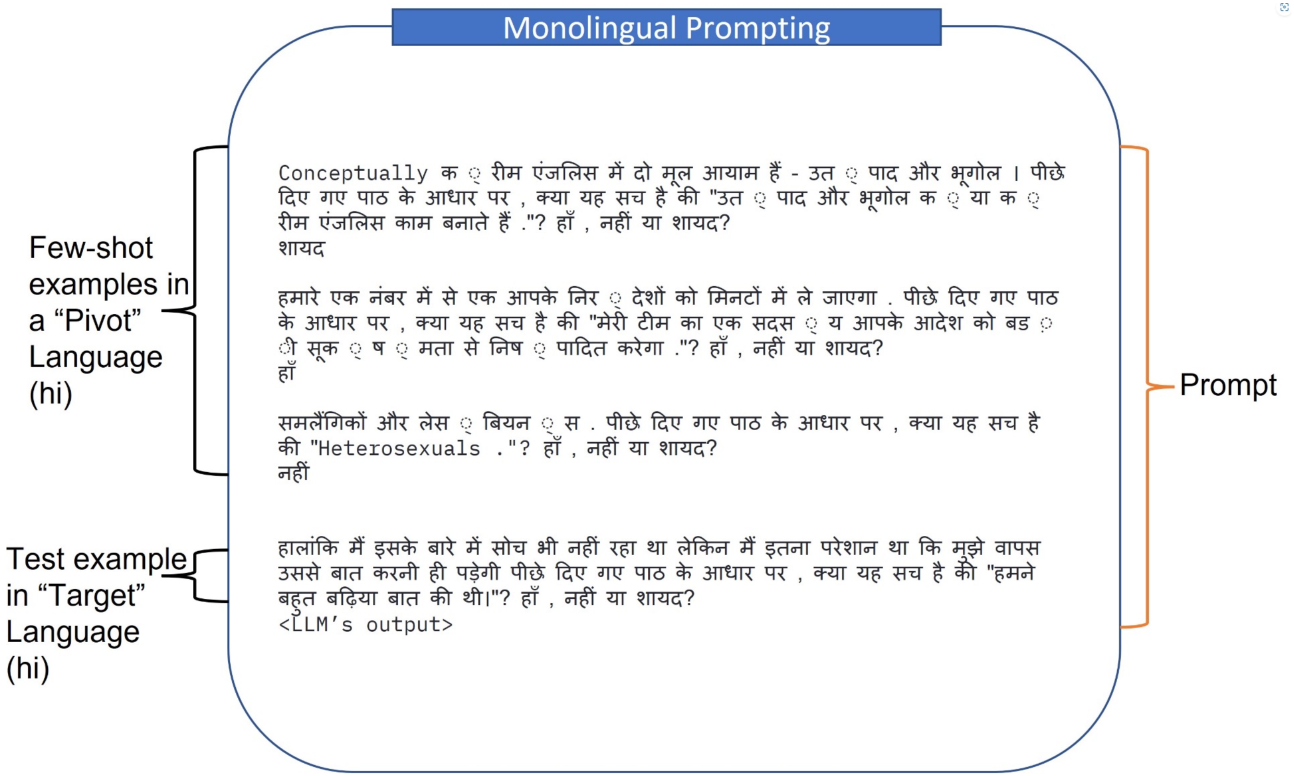
Models : GPT-3.5-Turbo, GPT-4

Task Instruction $\mathcal{I}$: You are an NLP assistant whose purpose is to summarize any given article. You should summarize all important information concisely in the same language in which you have been provided the document. Following the examples provided below:

Template f_{temp}:
`{document}`
===
Write a summary of the text above :

Microsoft

MEGA Framework: **Prompting Strategies**



The *k-shot* randomly selected examples for in-context supervision are of the same language as the test examples.

MEGA Framework: **Prompting Strategies**

Microsoft

Zero-Shot Cross-Lingual Prompting

Few-shot examples in a “Pivot” Language (en)

Conceptually cream skimming has two basic dimensions - product and geography . Based on the previous passage, is it true that "Product and geography are what make cream skimming work ."? Yes, no, or maybe?
Maybe

One of our number will carry out your instructions minutely . Based on the previous passage, is it true that "A member of my team will execute your orders with immense precision ."? Yes, no, or maybe?
Yes

Gays and lesbians . Based on the previous passage, is it true that "Heterosexuals ."? Yes, no, or maybe?
No

Test example in “Target” Language (hi)

हालांकि मैं इसके बारे में सोच भी नहीं रहा था लेकिन मैं इतना परेशान था कि मुझे वापस उससे बात करनी ही पड़ेगी पीछे दिए गए पाठ के आधार पर , क्या यह सच है की "हमने बहुत बढ़िया बात की थी।"? हाँ , नहीं या शायद?
<LLM’s output>

Prompt

The *k-shot* examples for in-context supervision are sampled from a pivot language which is different from the language of the test examples.

Microsoft

MEGA Framework: Prompting Strategies

Translate-Test

Few-shot examples in a "Pivot" Language (en)

Conceptually cream skimming has two basic dimensions - product and geography . Based on the previous passage, is it true that "Product and geography are what make cream skimming work ."? Yes, no, or maybe?
Maybe

One of our number will carry out your instructions minutely . Based on the previous passage, is it true that "A member of my team will execute your orders with immense precision ."? Yes, no, or maybe?
Yes

Gays and lesbians . Based on the previous passage, is it true that "Heterosexuals ."? Yes, no, or maybe?
No

Test example in "Target" Language (hi) translated to Pivot language

Well, I wasn't even thinking about that, but I was so frustrated, and, I ended up talking to him again. Based on the previous passage, is it true that "We had a great talk."? Yes, no, or maybe?
<LLM’s output>

Prompt

The *k-shot* examples are sampled from English data while the test examples are translated to English using Bing Translator.

Performance Analysis

MEGA Results: # Comparing Different Models

Microsoft

GPT-3.5 (DV003 and Turbo) performs worse than SOTA models. Best performance is with data point and context translated to English and back.

Gap between GPT4 and SOTA models is reduced (but significantly worse than English). GPT4 can be queried directly in target language for many high-resource and Latin script languages.

GPT4 is significantly better than GPT-3.5 (Turbo), showing how multilingual behavior is beginning to appear for some languages and tasks, where monolingual performance surpasses or comes close to translation*

For low-resource languages, **translating into English or other high-resource languages provides benefits.**

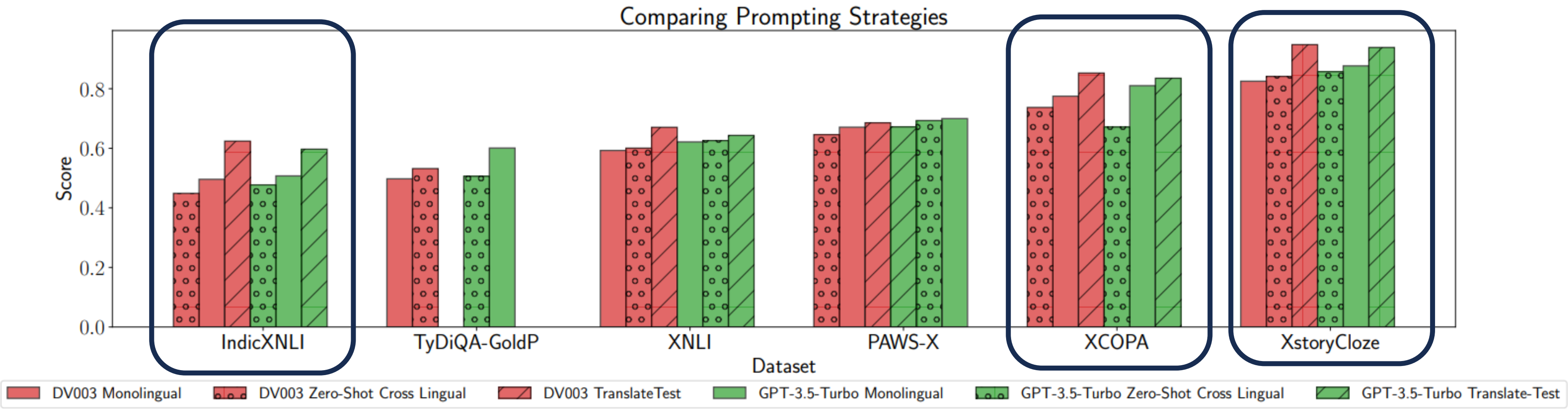
Model	Classification				Question Answering			Sequence Labelling		Summarization
	XNLI	PAWS-X	XCOPA	XStoryCloze	XQuAD	TyDiQA-GoldP	MLQA	UDPOS	PAN-X	XLSum
Metrics	Acc.	Acc.	Acc.	Acc.	F1 / EM	F1 / EM	F1 / EM	F1	F1	ROUGE-L
Fine-tuned Baselines										
mBERT	65.4	81.9	56.1	×	64.5 / 49.4	59.7 / 43.9	61.4 / 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.

**Caveat: it is unclear which evaluation datasets GPT4 has seen during training, working on creating new, harder multilingual evaluation benchmarks*

Microsoft

MEGA Results: Comparing different Prompting Strategies



We compare three prompting strategies: *monolingual, translate-test, and zero-shot cross-lingual*.

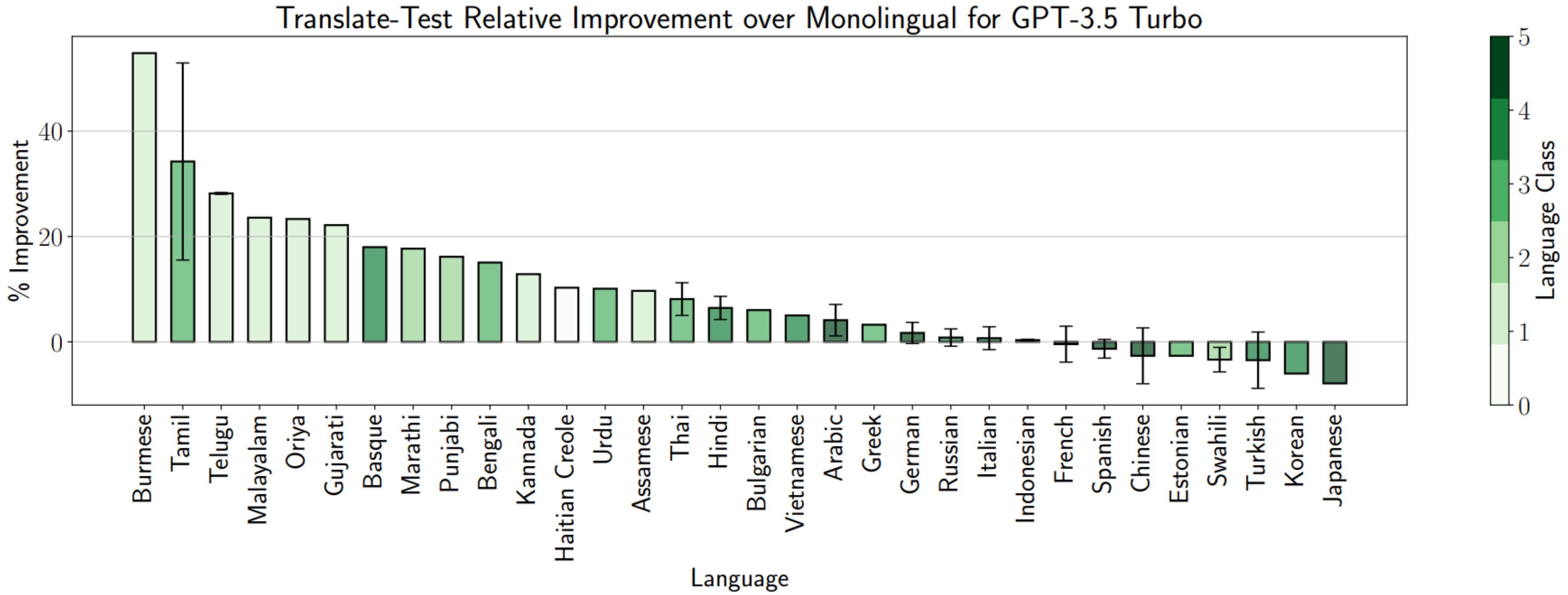
Zero-shot cross-lingual performs similarly to Monolingual for DV003 but shows a drop in performance for GPT-3.5-Turbo, especially for tasks involving extremely low-resource languages like Quechua and Haitian Creole.

Grounding the model through Monolingual prompting helps the model understand these languages better, resulting in improved predictions.

Translate-test generally improves performance, particularly for DV003. For datasets with low-resource and non-Latin script languages like IndicXNLI and XStoryCloze, the gains with translate-test are even more significant.

MEGA Results: **Comparing different Prompting Strategies**

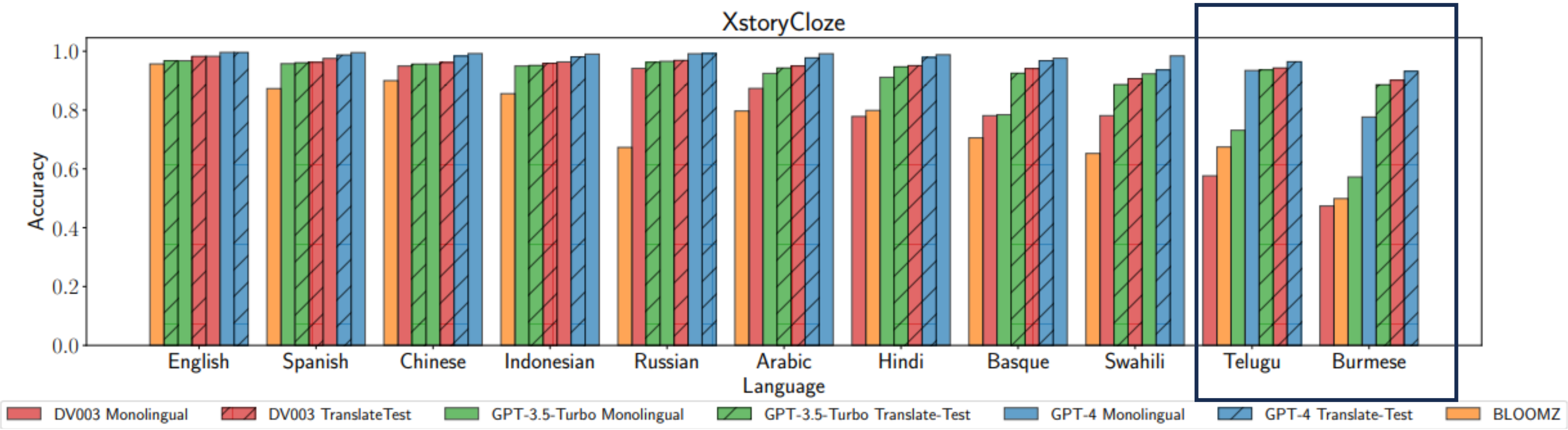
Microsoft

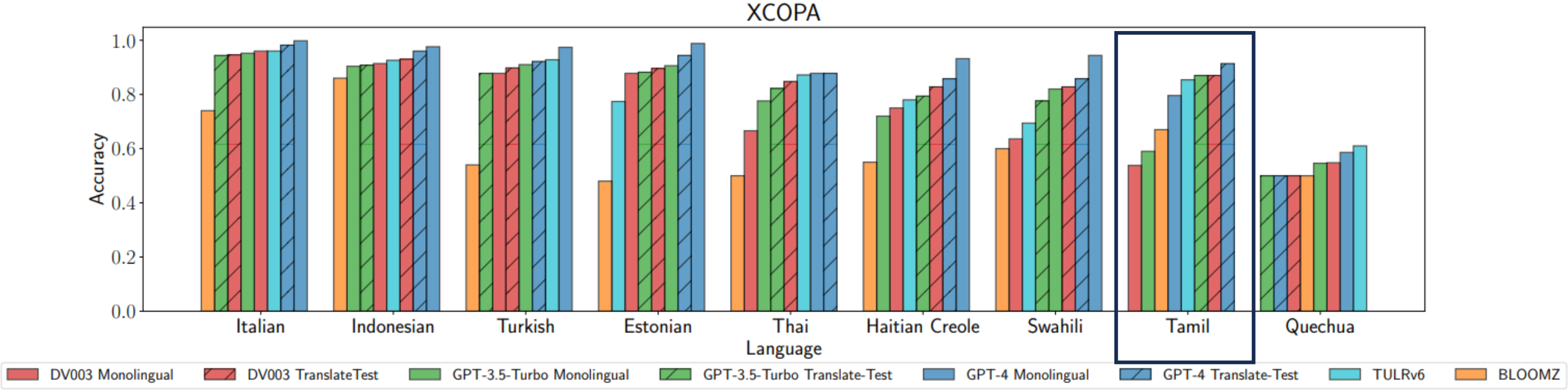


Translate-test: languages like Burmese, Tamil, and Telugu see upto > 30% relative improvement by Translate-Test over Monolingual, while for high-resource languages such as French and Spanish, the two perform similarly.

Microsoft

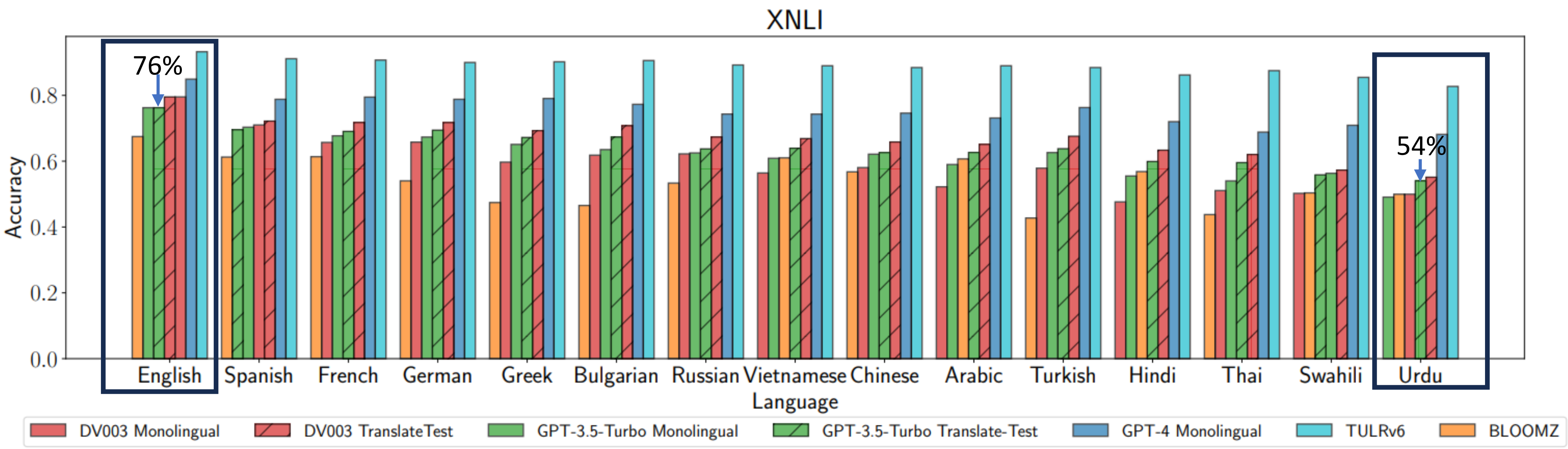
MEGA Results: Comparing different Prompting Strategies





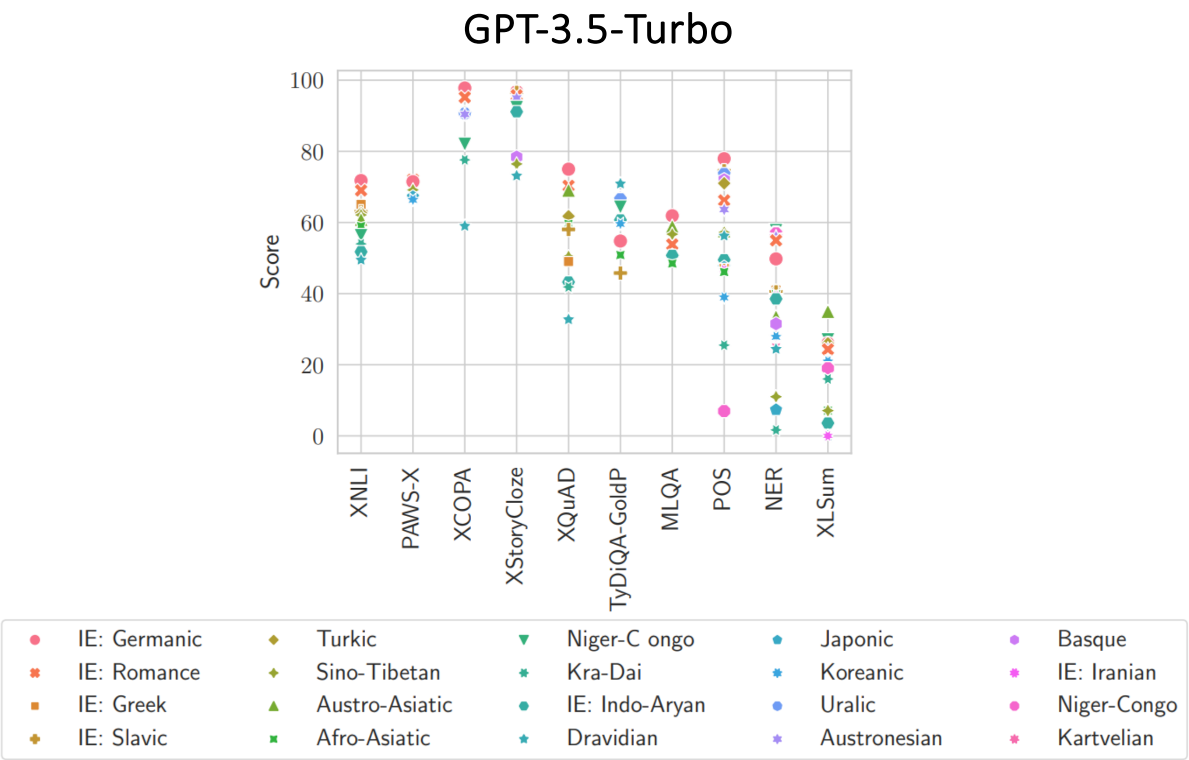
Monolingual and Translate Test are much more on par for GPT-4, but even there for low-resource languages like Burmese and Tamil, translate-test improves the performance by a significant margin

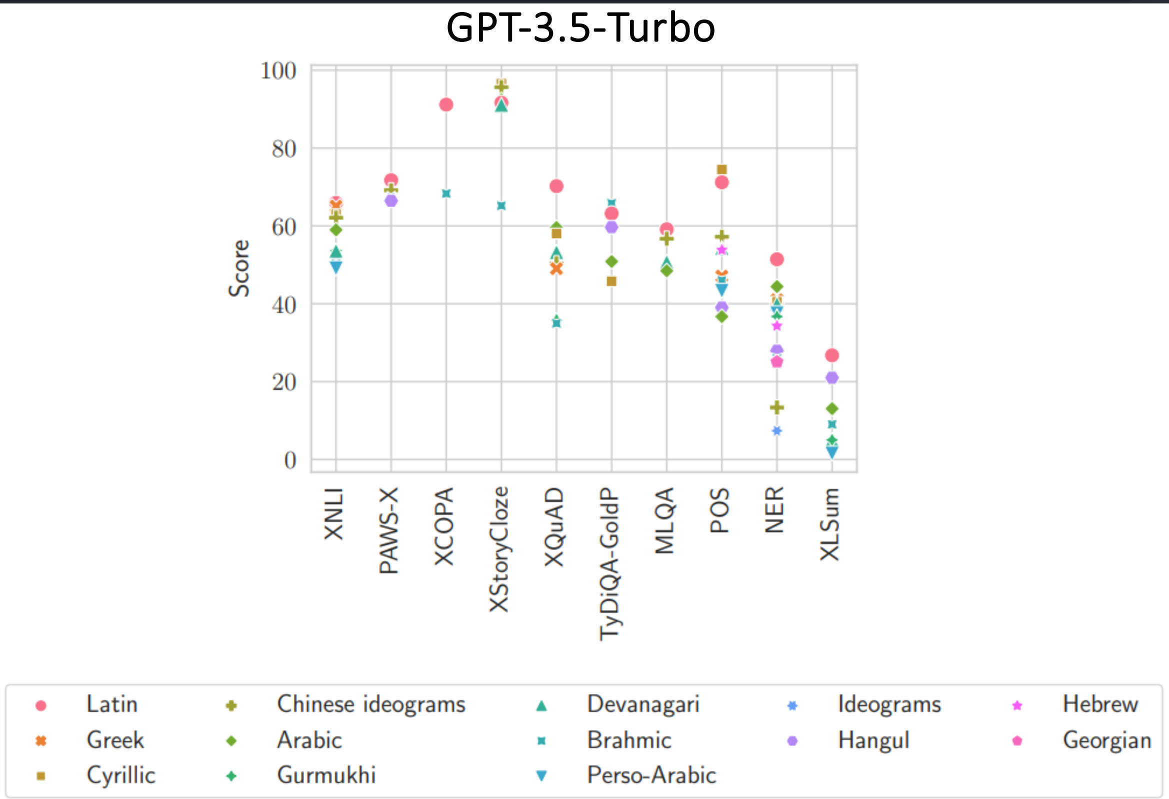
MEGA Results: **Comparing different Prompting Strategies. Does Translate-Test Solve the Problem?**



Well No! The gap between performance in English and performance obtained after translate-test for languages like Urdu can still be significantly high!

MEGA Results: **Linguistic Comparison**





LLMs tend to **work well on higher-resource languages families** (Indo-European: Germanic and Romance families) with **Latin Scripts**

Low-resource languages (Dravidian families) with limited training data and fewer available resources such as Tamil, Telugu **pose challenges for LLMs**.

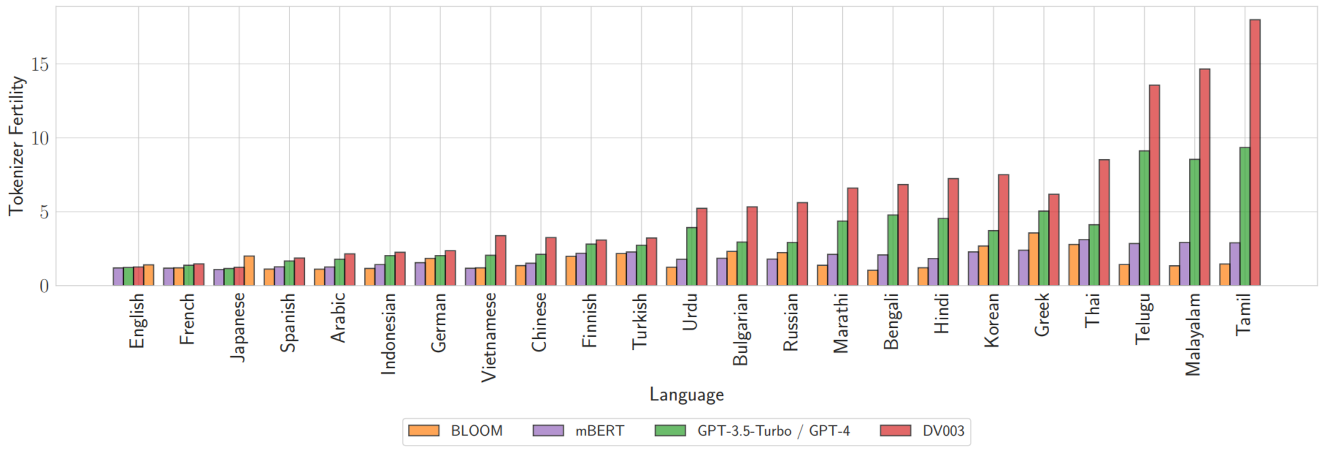
Factors Affecting Multilingual Performance in LLMs

Tokenization

Tokenization impact: Tokenization influences the performance of LLMs, as demonstrated by the disparity between Open AI models, mBERT and BLOOMZ tokenizers, and language-specific tokenizers.

Disparities in behavior: Differential behavior of tokenization across languages can explain the poor performance of generative models, especially in monolingual settings.

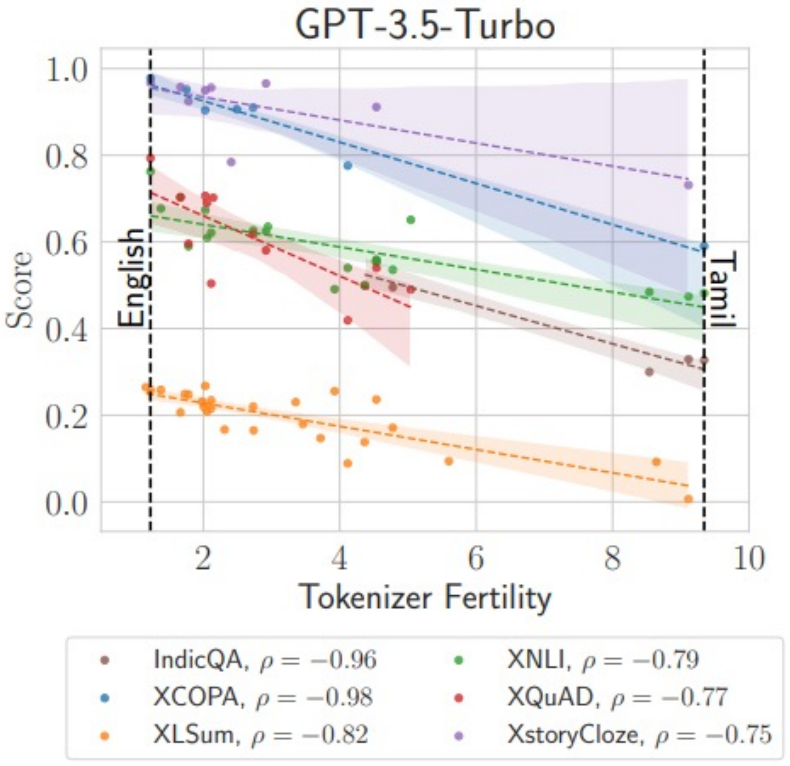
Limitations in lower-resource languages: Inadequate tokenization in lower-resource languages can restrict context encapsulation, resulting in issues such as poor context representation and performance on downstream tasks.



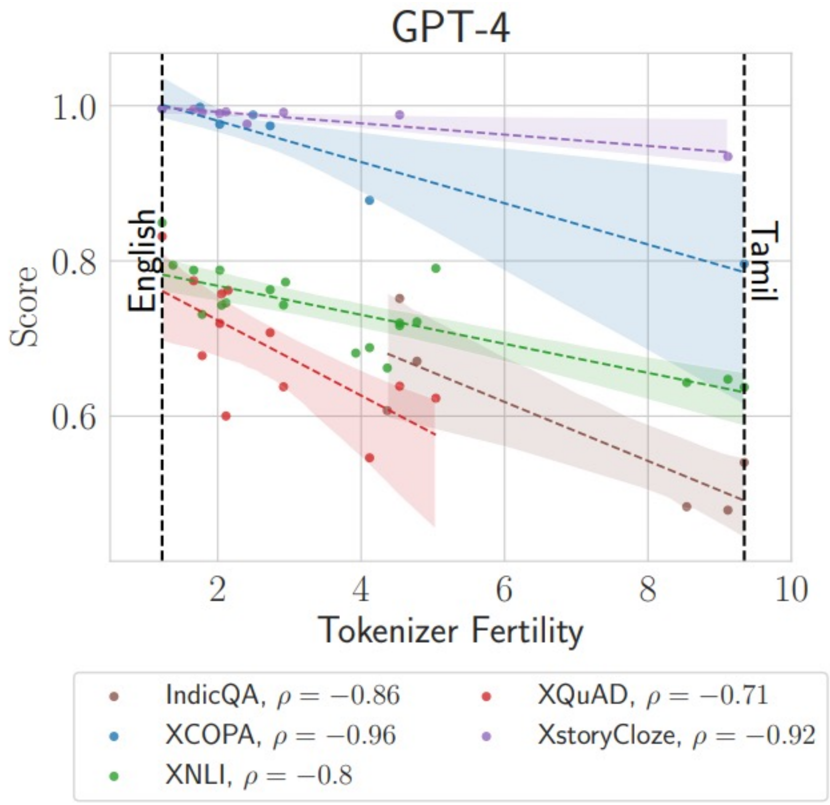
Tokenizer Fertility for GPT, BLOOMZ and mBERT for different languages

Tokenization

Tokenization impact: Strong correlations between tokenizer fertility and performance on many tasks!



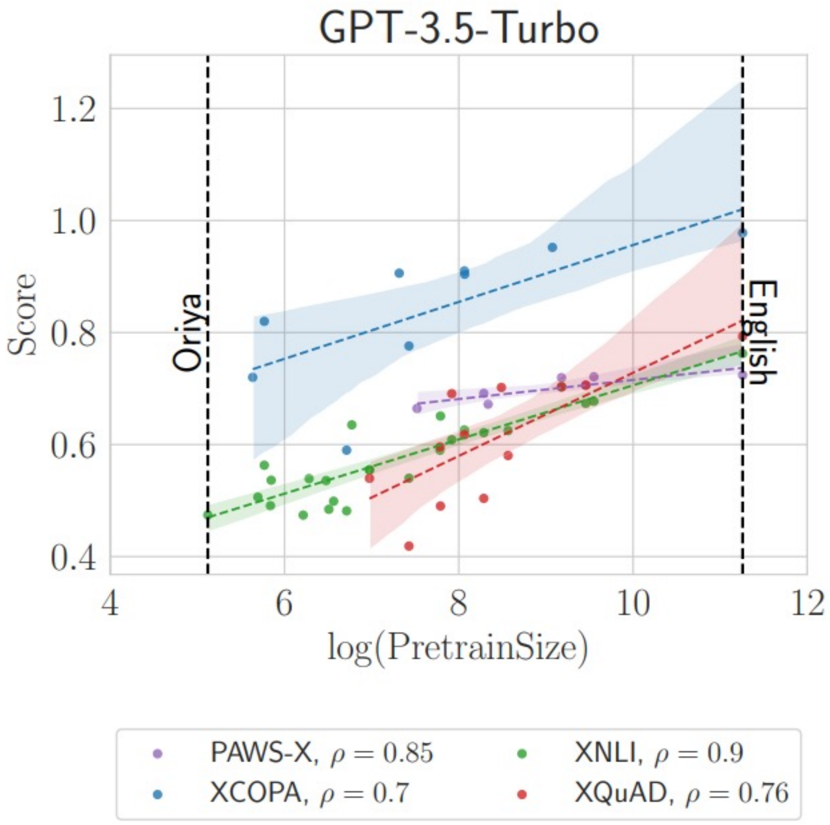
(a) Correlation between tokenizer fertility and performance for GPT-3.5-Turbo.



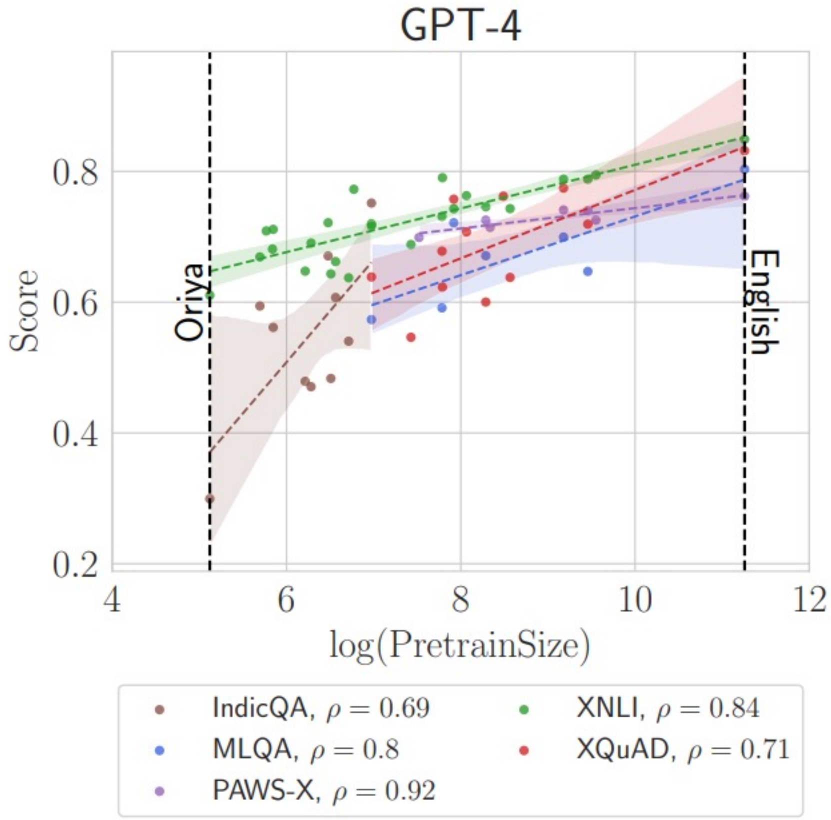
(b) Correlation between tokenizer fertility and performance for GPT-4

Amount of Pre-training Data

Similarly, we see strong correlations for a subset of tasks with amount of pre-training data and performance



(a) Correlation between pre-training size and performance for GPT-3.5-Turbo.



(b) Correlation between pre-training size and performance for GPT-4

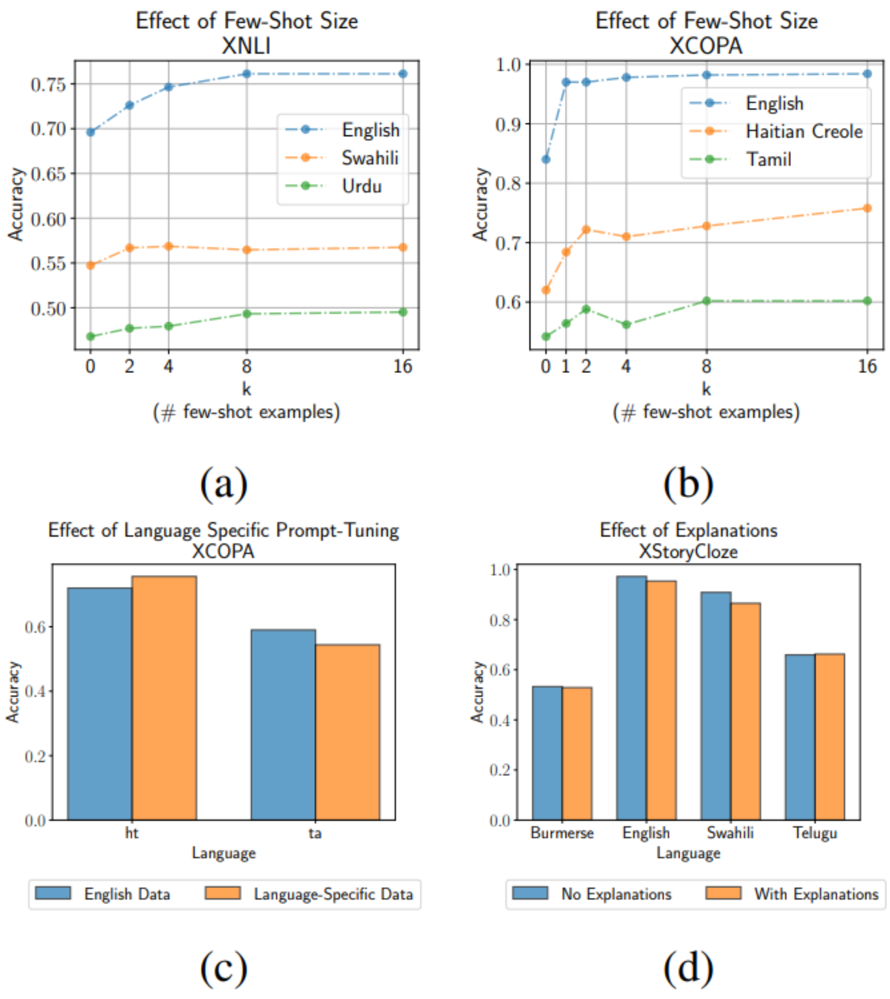
Microsoft

Challenges with Multilingual Benchmarking

Benchmarking Challenges: **Did we try out everything?**

A Kaleidoscope of Choices. So many decisions to be made while evaluation

- **Choice of Prompt**
- **Choice of Few-shot samples (size and type)**
- **Prompting Strategies (Explanations, CoT?)**
- **Choice of language of prompts**
- **Use of External Tools**
- **Decoding Hyper-parameters**



Benchmarking Challenges : **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
- We consider three factors to get some sense of dataset contamination: i) LLM's knowledge of the dataset, ii) availability of test datasets on the internet, and iii) dataset release date.
- 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).

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 NLI	None	No	June 2020
EN-ES-CS	None	No	May 2016

Table 3: 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 evidence.

MEGA Benchmark:

Summary

- There is a significant disparity between the performance of LLMs in English vs non-English languages, especially low-resource languages with non-Latin scripts
- Previous generation fine-tuned models fare much better for most tasks we evaluate
- It if often difficult to do better than translating target language inputs to English to solve the problem, and even that is vastly sub-optimal!
- Bad tokenization and poor representation in the pre-training data might explain the sub-par performance on low-resource languages

Looking Forward

Microsoft

Advancing Multilingual Evaluation of Generative LLMs:

Future Directions

Expand language coverage: Include more diverse and low-resource languages for comprehensive evaluation (Masakhane and AmericasNLP datasets).

Model coverage: Include PaLM and other models to expand comparison beyond OpenAI models, BLOOMZ, and SOTA models.

Explore additional evaluation dimensions: Incorporate calibration, bias, and disinformation to provide a holistic assessment beyond traditional metrics (Example: ROUGE-L limitations; Need for Human Evaluation).

Incorporate more NLP tasks and real-world datasets: Extend benchmark to cover a wider range of standard NLP tasks and real-world applications (MARI's LLMs evaluation on EPOCh data).

Microsoft

Questions

Get in touch

sunayana.sitaram@microsoft.com
t-kabirahuja@microsoft.com
mochieng@microsoft.com

Microsoft