The role of locally relevant content in evaluating multilingual LLMs

Nikolay Bogoychev
Kriz Chan
Dieuwke Hupkes

LOC. WORLD

#LocWorld54 Monterey

Agenda

- How to evaluate an LLM?
- What is MultiLoKo?
- Scope
- High level findings
- How can I use the benchmark

How to evaluate an LLM?

Benchmark, Benchmark, Benchmark

Hold on…What is a benchmark? What is it used for?

There can be many different types of benchmarks for various purposes:

- Reasoning and problem solving
- Safety
- Domain-specific knowledge, e.g. coding, maths, science, etc.
- Language understanding
- General knowledge

Our benchmark focuses on multilingual ***locally relevant*** knowledge.

Questions 'asked' to the LLM

Correct answer, present in the benchmark

Answer given by the LLM

Language	Question	Correct answer	LLM answer	
English	How old was Prince Phillip when he married Princess Elizabeth?	26	26	✅
Spanish	De todas las hermanas y medias hermanas de Frida Kahlo, ¿quién fue la mayor de todas?	María Luisa	Margarita	❌
Chinese	张无忌在哪座山修练猴腹之中的九阳神功？	昆仑山	光明顶	❌
....	....	....	....	❌

Whether the LLM answer is correct

Questions 'asked' to the LLM

Correct answer, present in the benchmark

Answer given by the LLM

Whether the LLM answer is correct

Language	Question	Correct answer	LLM answer	
English	How old was Prince Phillip when he married Princess Elizabeth?	26	26	✅
Spanish	De todas las hermanas y medias hermanas de Frida Kahlo, ¿quién fue la mayor de todas?	María Luisa	Margarita	❌
Chinese	张无忌在哪座山修练猴腹之中的九阳神功？	昆仑山	光明顶	❌
....	....	....	....	❌

How do we come up with the benchmark score?

We count how often the LLM is correct and incorrect, and use that to assign a score for the benchmark.

A score of 100 (usually) reflects the LLM answers everything perfectly,

0 means that it got everything wrong.

MultiLoKo? What?

"***Multi***lingual ***Lo***cal ***K***n***o***wledge benchmark"

- *Multilingual benchmark for LLMs*
- *spans over **31 languages***
- *500 questions per language, spread out over a validation and test set*
- ***Sourced independently** for each language, catering to locally relevant knowledge*

Why are we building MultiLoKo?

We need to assess how well language models retrieve and comprehend knowledge to generate responses that are culturally and linguistically relevant across various languages.

What is the problem we are solving?

1. Multilingual datasets are not very common and/or have scattered coverage
2. Most existing multilingual datasets are heavily skewed towards English language and Western knowledge
3. Even when there is non-English content included in those datasets, it's often translations rather than original language.

Why are we building MultiLoKo?

This leads to various issues:

- Blind spots in the model creating cultural bias
- Low quality for non-English portions of benchmarks
- Impact model performance in tasks involving non-English/non-Western languages

Example of failure in real world application of LLM

We asked an LLM about 中華電信MOD . This is a Taiwanese media company and the 9th most popular web page on the Traditional Chinese Wiki page. When we asked the question in English, we got the response we expected:

Example of failure in real world application of LLM

But when we asked about the same company in Spanish, the system choked and produced Traditional Chinese instead which is presumably the language the model got the information from (same happened when we tested multiple LLMs):

Imagine you are a Spanish person travelling in Taiwan, and was curious about some texts you came across, even though you entered the Chinese texts, you expected to see an explanation in your target language - Spanish.

This test shows that the model has not successfully transferred the knowledge across languages.

Benchmark How?

Sourcing

1. For each language, get the top 6k most visited wikipedia articles from the years 2016-2021
2. Sort them by most viewed and sample several paragraphs (up to 3k chars)
3. Send the paragraphs to human annotators ask to select 550 of them to produce locally relevant question answer pairs.
   - ~~*How old was Messi when he scored his first professional goal?*~~
   - *When did TV series Forensic Heroes 法證先鋒 first debut?*
4. Send to independent annotators (different vendor) for verification
5. Ask two new annotators to answer each question to ensure correctness
6. Annotators produce all possible answer variants (ie 1;1st;first;etc…)
7. Translate English data to non-English languages and vice versa (both human & Google Translate)

Prompt

We invested quite some time creating prompts separately for each language, using native speakers.

Sample prompts (English)

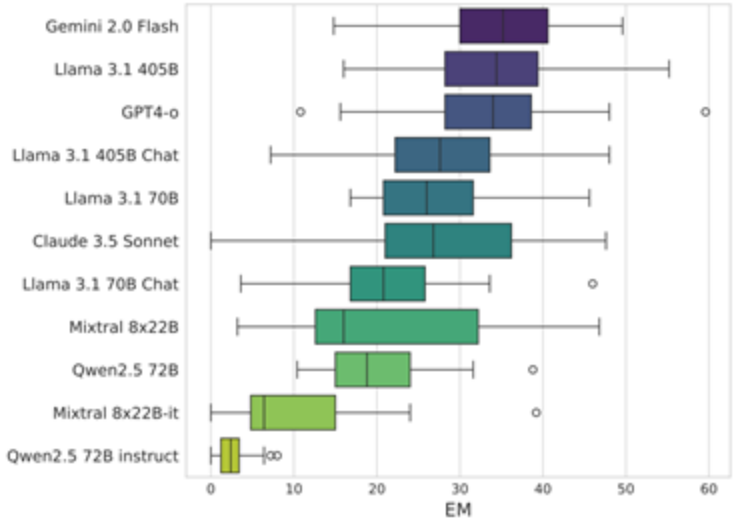
```
"""\
    Please answer the following question.
    Question: {{ question }}
    Your response should be as concise as possible and end with "The answer is [answer]", where [answer] is the response to the question.
    The answer should only be a {{ output_type }}.

    The answer is"""
```

- Also includes a comprehension version of the task where we provide the article source on which the question was based.
- All prompts were written by native speakers.

High Level Findings

Model performance (main partition)



(a) EM scores

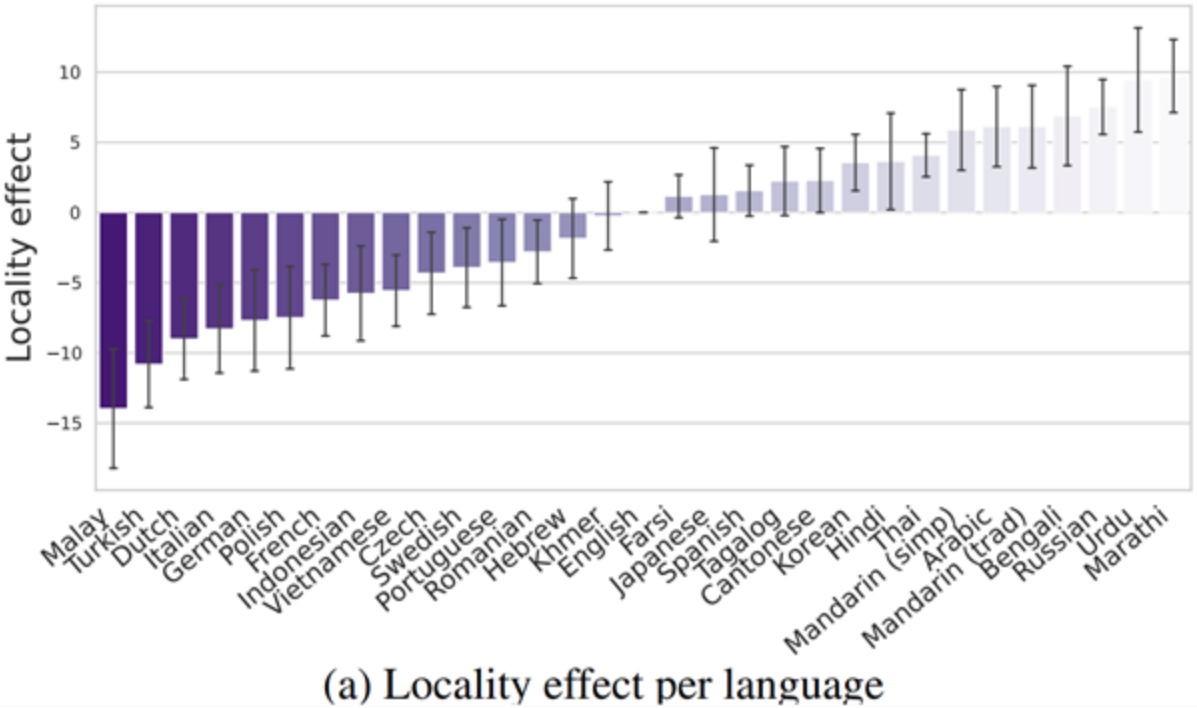
Take away message: There isn't a single model that scores highly on MultiLoKo across the board

Native or Translated data?

- Locality Effect -> Do models perform better when asked about English centric knowledge or local-specific knowledge?

- Mother Tongue Effect -> Can models answer questions about Bulgarian history better if they are asked in Bulgarian or English

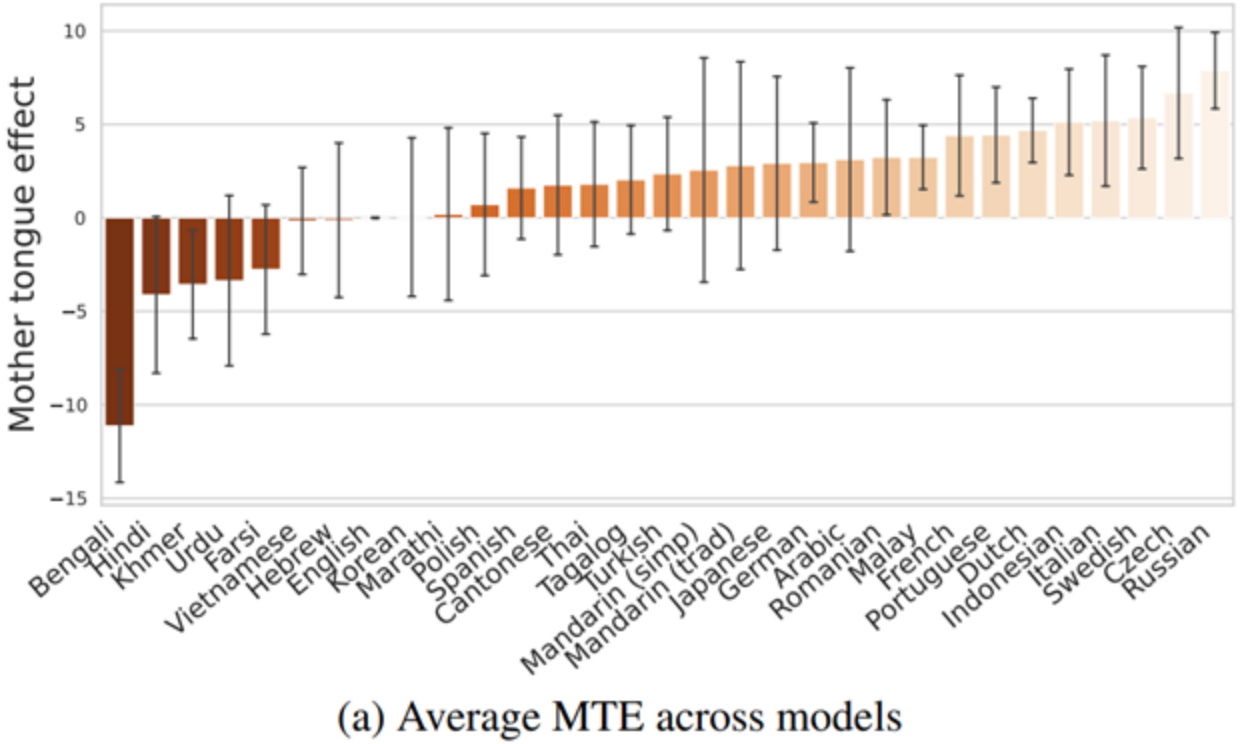
Locality effect



(a) Locality effect per language

Negative locality effect: English questions Translated into X are easier:
Positive locality effect: English questions Translated into X are more difficult.

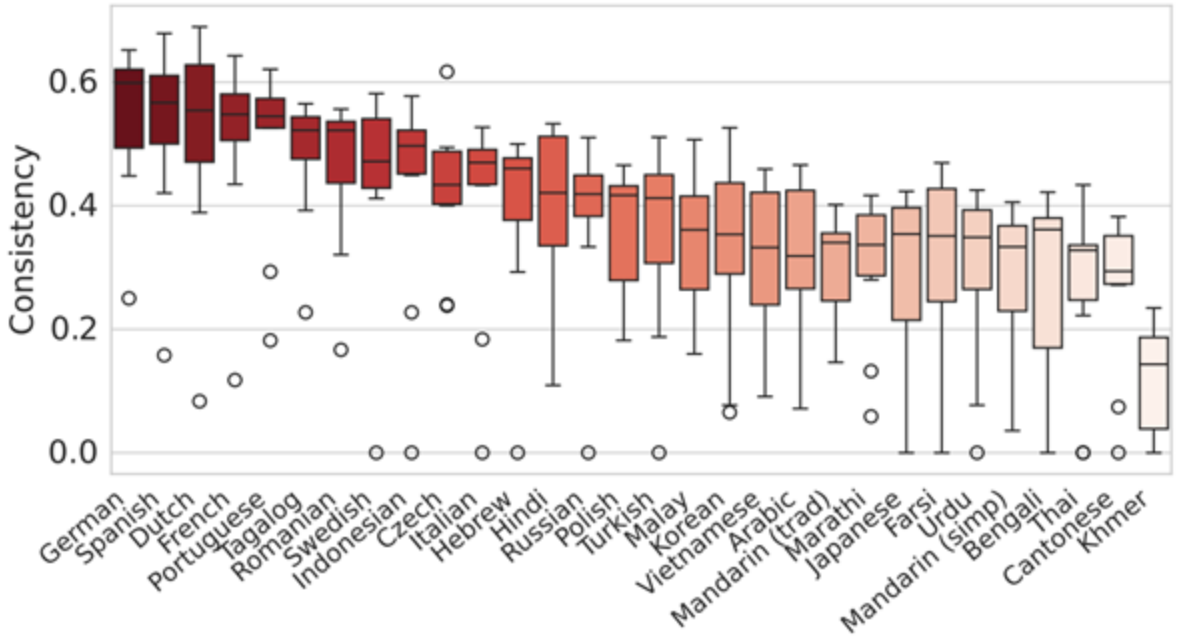
Mother tongue effect



(a) Average MTE across models

Languages where the model proficiency is bad, benefit from being asked in English, but otherwise it is better for models to be asked locally relevant questions in the local language.

Consistency

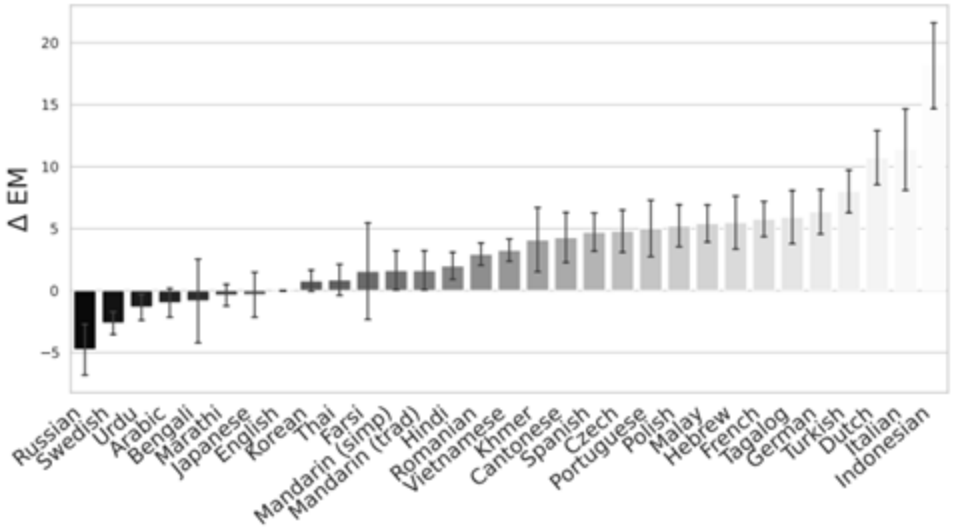


Models find it very hard to answer questions consistently, when the same question is asked in English and non-English.

Translation Effect

Model	R	min Δ	max Δ	avg Δ
Gemini 2.0 Flash	0.80	-10.00	21.60	4.35
Llama 3.1 405B	0.83	-4.40	18.80	5.82
GPT4-o	0.85	-6.00	21.60	4.46
Llama 3.1 405B Chat	0.80	-10.40	22.40	3.08
Llama 3.1 70B	0.77	-7.60	22.00	4.59
Claude 3.5 Sonnet	0.90	-9.60	20.80	2.84
Llama 3.1 70B Chat	0.87	-6.00	20.00	3.12
Mixtral 8x22B	0.91	-3.20	20.00	4.13
Qwen2.5 72B	0.83	-4.00	16.80	3.47
Mixtral 8x22B-it	0.92	-4.80	12.40	2.41
Qwen2.5 72B instruct	0.80	-0.80	3.20	0.36

(a) Language difficulty stats across human- and machine translations



(b) MT vs human translations

Using machine translation results in overall lower performance.

Can I use MultiLoKo?

1. Get the dataset from kaggle/huggingface/github:
   https://www.kaggle.com/datasets/metaresearch/multiloko
   https://github.com/facebookresearch/multiloko/
   https://huggingface.co/datasets/facebook/multiloko

2. 250 questions per language available publicly, 250 secret

Questions?

Dieuwke Hupkes
Research Scientist
AI Research

Nikolay Bogoychev
Research Scientist
AI Research

Van Phung
Project Manager, PDO

Dunant Hin
Project Manager, PDO

Milena Hoffman
Admin, PDO

Kriz Chan
Localization Program Manager

Antonio Gai
Localization Program Manager