Polyglot or Not? Measuring Multilingual Encyclopedic Knowledge in Foundation Models

Tim Schott, Daniel Furman, and Shreshta Bhat

University of California, Berkeley, School of Information

Abstract

In this work, we assess the ability of foundation models to recall encyclopedic knowledge across a wide range of linguistic contexts. To support this, we: 1) produce a 20-language dataset that contains 303k factual associations paired with counterfactuals, 2) evaluate 5 models in a multilingual test, and 3) benchmark a diverse set of 24 models in an English-only test. Meta's LLaMA achieves the highest scores in both multilingual and English-only evaluations. Yet, an analysis of LLaMA's errors reveals significant limitations in its ability to recall facts in languages other than English, plus difficulties related to the location and gender of fact subjects. Overall, our findings suggest that today's foundation models are far from polyglots. Supporting code and data are openly released.

Introduction

Can foundation models be used as multilingual knowledge bases? Foundation models typify an emerging paradigm that warrants further study; all-purpose Large Language Models (LLMs) that are trained on internet-scale corpora excel in generalization to some new tasks. Their widespread adoption and ostensible credibility come with risks, though. For instance, foundation models inherit inaccuracies from training corpora , which are in turn propagated downstream to the models that are fine-tuned from them. Additionally, foundation models spend the majority of their training phase absorbing information in English; for example, LLaMA [4] devotes two-thirds of its training dataset to an English-only subset of the CommonCrawl. Thus, foundation models are potentially deficient on non-English tasks.

Task

We use cloze statements for the **Polyglot or Not?** test; given some context (the "stem"), a language model is prompted to predict the next token in the sentence (the "object"). The vocabulary-wide inference probabilities are subsequently probed to contrast the likelihood of the factual completion to the counterfactual(s), as formalized below.

Task Formalization

Factual associations are represented as a triplet $\langle s, r, o \rangle$ where s and o denote the subject and object entity and r is a linking relation. For example, the fact "Paris is the capital of France" is represented by the triplet ⟨Paris, capital of, France⟩ where "Paris" maps to s, "capital of" maps to r, and "France" maps to o. To conduct the test, a language model M is prompted with the sentence stem, where o is masked out. Erroneous "counterfactuals" $\langle s, r, o' \rangle$ like ⟨Paris, capital of, Italy⟩ are then used to assess M's understanding of $\langle s, r, o \rangle$. The test measures whether M correctly knows a fact $\langle s, r, o \rangle$ by calculating:

$$\text{CKA}_{\text{M}}(s, r, o) = \frac{P_M(o \mid s, r)}{\mathbb{E}_{o'}[\, P_M(o' \mid s, r)]}$$

When $\text{CKA}_{\text{M}}(s, r, o) > 1$, the model is said to have successfully recalled the association. To carry out the test, we solicit cloze completions across each of the languages in batch. The percentage of fact-completions that M recalls correctly per batch is calculated by tallying up the number of completions where $\text{CKA}_{\text{M}}(s, r, o) > 1$ and dividing by the total number of completions.

Dataset

The **Polyglot or Not?** dataset includes 303k unique knowledge statements spanning 20 languages. To curate this album, we first concatenated two datasets of English-only knowledge statements from [1] and [3]. After de-duplicating $\langle s, r, o \rangle$ triplets, we were left with 26.3k unique factual statements. We then used the Google Translate API to translate the data into 19 target languages (see Figure 1, y-axis). Our translation approach mirrors prior multilingual studies, such as [2], that show minimal practical differences in machine versus manually translated cloze statements. Additionally, we enforce that fact objects retain the right-most position in the sentence, thus supporting both masked and causal language modeling tasks. Each language contains different amounts of statements due to the varying syntactic capacities to support this requirement. On average, a given fact appears in 12 of the 20 languages tested.

stem	factual	counterfactual	relation
"Hungary, which has the capital"	"Budapest"	"Vienna"	P1376
"Sundar Pichai trabaja para"	"Google"	"Apple"	P108
"La Prius est produite par"	"Toyota"	"Honda"	P176

Table 1. A sample of 3 knowledge statements from the **Polyglot or Not?** dataset.

Multilingual Efficacy

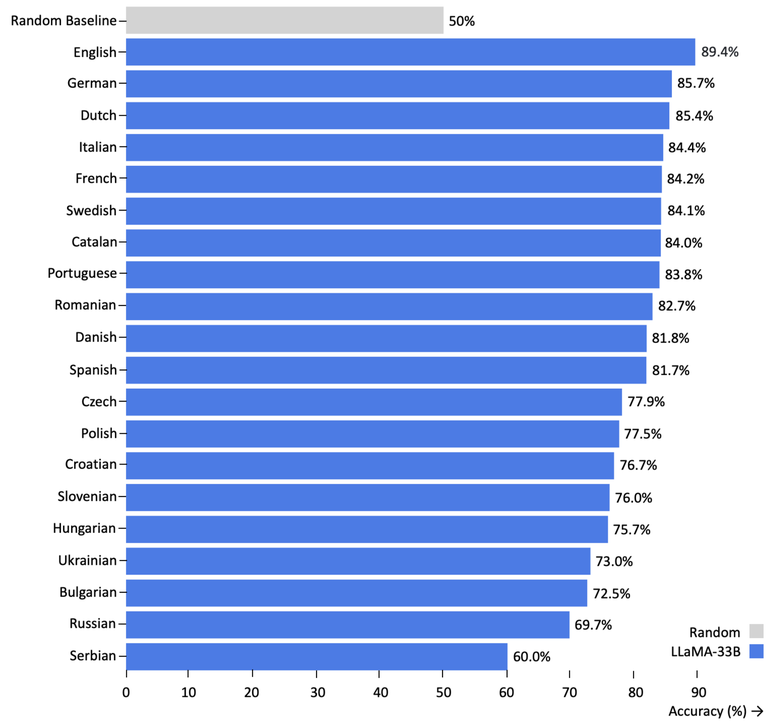

Figure 1. **LLaMA-33B's multilingual test.** LLaMA-33B scores higher on languages written in Latin script than those written in Cyrillic script (Ukrainian, Bulgarian, Russian, Serbian). A chi-squared test confirms that performance depends on language script ($\chi^2 = 3570, p < 0.001$).

Test Leaderboard

model	accuracy (%)
llama-33b	**79.31** (+/- 0.74)
m-bert	**62.00** (+/- 0.87)
bloom-7b1	**57.70** (+/- 0.88)
xlm-roberta	**56.03** (+/- 0.90)
mt5-xl	**52.51** (+/- 0.91)
Random Baseline	50

Table 2. **Multilingual test leaderboard**. Here, **accuracy** refers to the average performance of each model across 20 languages. The uncertainty estimates are averaged 95% confidence intervals computed from 10k bootstrap iterations per language. The results suggest tested models struggle to recall facts in a multilingual setting relative to English-only performance (see below).

model	accuracy (%)
llama-65b	**89.56** (+/- 0.37)
llama-33b	**89.40** (+/- 0.38)
falcon-40b	**87.01** (+/- 0.41)
llama-13b	**86.66** (+/- 0.42)
llama-7b	**85.53** (+/- 0.43)
redpajama-7b	**85.07** (+/- 0.44)
Random Baseline	50

Table 3. **English-only test leaderboard, top 6 models**. Here, **accuracy** refers to model performance on the English subset of the dataset. The uncertainty estimates are 95% confidence intervals computed from 10k bootstrap iterations. Consistent with the trends in the multilingual test, LLaMAs of varying sizes emerge as the front-runners.

Results

1. **Multilingual performance lags** behind English scores, particularly for Cyrillic script, suggesting an absence of robust cross-lingual knowledge transfer.
2. **An analysis of LLaMA's errors** reveals significant differences in its ability to recall facts by location and gender of knowledge statement subjects.
3. **Quality of training data** appears to outweigh the number of model parameters in boosting the efficacy of enyclopedic knowledge recall.

References

[1] Qingxiu Dong, Damai Dai, Yifan Song, Jingjing Xu, Zhifang Sui, and Lei Li. Calibrating factual knowledge in pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5937–5947, Abu Dhabi, United Arab Emirates, Dec 2022. Association for Computational Linguistics.

[2] Nora Kassner, Philipp Dufter, and Hinrich Schütze. Multilingual lama: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online, Apr 2021. Association for Computational Linguistics.

[3] Kevin Meng, David Bau, and Alex Andonian. Locating and editing factual associations in gpt. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, 2022.

[4] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. (arXiv:2302.13971), Feb 2023.