Multilingual Knowledge: an Informal, Initial Study

Ishika Agarwal, Nimet Beyza Bozdag, Dilek Hakkani-Tür
Department of Computer Science
UIUC
{ishikaa2, nbozdag2, dilek}@illinois.edu

Abstract

Code-switching is a common phenomenon of alternating between different languages in the same utterance, thought, or conversation. We posit that humans code-switch because they feel more comfortable talking about certain topics and domains in one language than another. With the rise of knowledge-intensive language models, we ask ourselves the next, natural question: *do language models have more knowledge in certain languages than others?* We run an experiment to test this hypothesis and find that language models can perform better when performing chain-of-thought reasoning in different languages. We find that language models do indeed know more about certain topics in certain languages than others. **This report serves as an initial, preliminary study of this hypothesis.**

1 Introduction

Humans have been known to code-switch, and it was formally documented in 1982 [Joshi, 1982]. Hinglish and Spanglish are common examples of hybrid languages between Hindi and English, and Spanish and English, respectively. There are many examples in the media of language models code-switching, potentially supporting our hypothesis that language models may perform better, or exhibit greater comfort in certain languages when addressing specific topics. For example, DeepSeek-r1 has been shown to suddenly "think" in Chinese when given an English query. While this is inherently not a problem, the final output tends to be in Chinese, as well.

We find that certain information is stored in certain languages. As a toy example, we ask GPT-4 turbo about the concept of dowry in Hindi and Arabic (see Figure 1). In Hindi, because it is linked to Indian culture, the language model talks about the concept of Dahej (where the bride's family gives money to the groom's family). In Arabic, because it is linked to Islamic culture, the language model talks about the concept of Mahr (where the groom's family gives money to the bride).

Speaker	Arabic	Arabic → English	Hindi	Hindi → English
User	ماذا يسمى تبادل الأموال من عائلة إلى أخرى من أجل الزواج؟	What is the tradition of one family giving another family gifts during marriage called?	शादी के लिए एक परिवार से दूसरे परिवार को धन का आदान-प्रदान क्या कहलाता है?	What is the tradition of one family giving another family gifts during marriage called?
GPT	يسمى المهر إذا كانت الأموال تُدفع من العريس أو عائلته إلى العروس أو عائلتها.	It is called "Mahr" if the money is paid by the groom or his family to the bride or her family.	जब वधू का परिवार वर के परिवार को धन देता है, तो उसे दहेज कहते हैं।	When the bride's family gives wealth to the groom's family for marriage, it is called dowry.

Figure 1: The user asks: "What is the tradition of one family giving another family money during marriage called?" in Hindi and Arabic. Because both languages have different cultures associated, they talk about different concepts of dowry (Dahej: bride to groom; Mahr: groom to bride).

We believe that humans code-switch because they know more about certain topics and domains in one language than another. We study this effect in language models as well. We coin the term

Preprint. Under review.

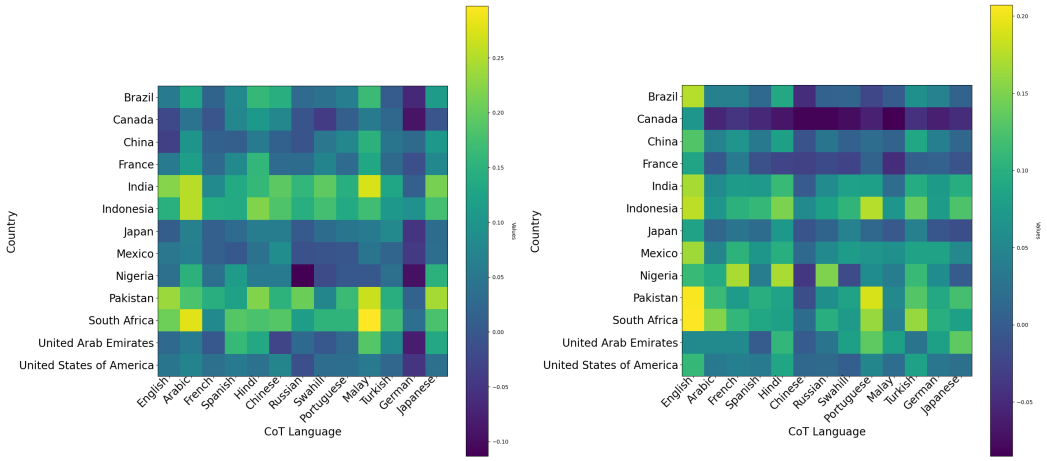

Figure 2: Multilingual Knowledge in Qwen2.5 (0.5B) and Phi-4 (3.4B)

"multilingual knowledge" to represent knowledge that is present in one or more languages, but not all. We design an experiment to showcase this.

2 Experimentation

To begin, we formulate experimentation by studying the effect of **multilingual chain-of-thought (CoT) reasoning** on the performance of language model reasoning. Essentially, we identify a set of languages $\mathcal{L}$. For each language ℓ in $\mathcal{L}$, we prompt the model to do CoT reasoning in ℓ on the entire dataset. Denote this as $LLM(D|\ell)$ which is the overall language model performance (LLM) on the dataset D with CoT reasoning using language ℓ. We also calculate the performance of the language model without reasoning, and denote it simply as $LLM(D)$. While the CoT language is ℓ, the final answer of the model is always English.

2.1 Set up

Datasets. There are at least three kinds of data that could exhibit the idea of multilingual knowledge: culture, history, and religion. In this study, we choose to study multilingual cultural knowledge of language models. Intuitively, a model should know more about a certain culture in the corresponding language. Hence, we choose the CultureAtlas dataset [Fung et al., 2024]. This dataset consists of correct and incorrect cultural norms, categorized by the region of the cultural norm. The goal is classification. Hence, we use accuracy ((True Positive + True Negative) / All Predictions) as a performance metric, and report the results on $LLM(D|\ell) - LLM(D)$. The maximum accuracy is 1.0 (100%) and minimum is 0.0 (0%), so the performance metric ranges from 1.0 to -1.0.

A nice feature of the CultureAtlas dataset is that it identifies the country of each cultural norm. So, we identify 15 countries and 13 languages to test our hypothesis:

- **Countries**: Brazil, Canada, China, France, India, Indonesia, Japan, Mexico, Nigeria, Pakistan, Russia, South Africa, Turkey, United Arab Emirates, United States of America
- **Languages**: Arabic, Chinese, English, French, German, Hindi, Japanese, Malay, Portuguese, Russian, Spanish, Swahili, Turkish

Models. We use four models that were trained on multilingual data: (1) `microsoft/Phi-4-mini-instruct`, (2) `ibm-granite/granite-3.1-8b-instruct`, (3) `Qwen/Qwen2.5-0.5B-Instruct`, and (4) `meta-llama/Llama-3.1-8B`. We choose these models because they are from different model families, differ in number of parameters, and are available on VLLM for fast inference.

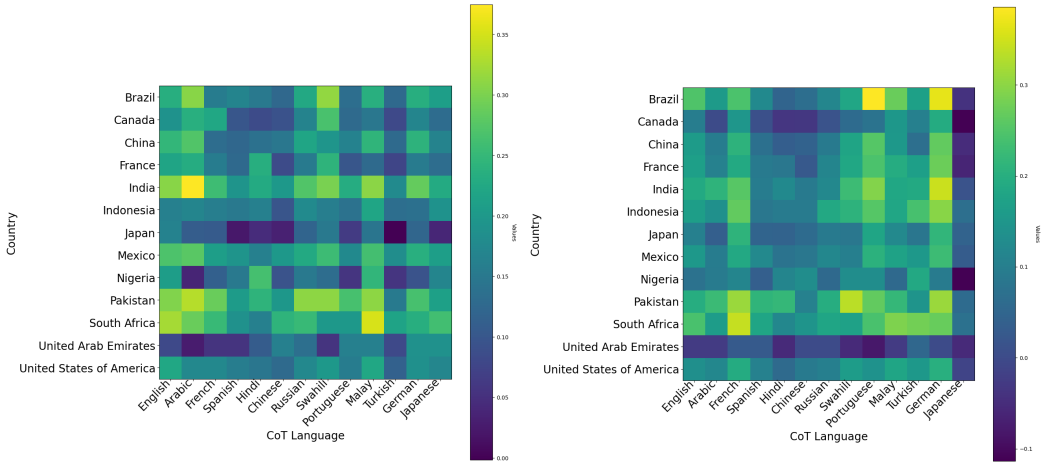

Figure 3: Multilingual Knowledge in Granite-3.1 (8B) and LLama-3.1 (8B).

2.2 Results

Figures 2 and 3 contain our results for this experiment. To begin, we see very different levels of multilingual knowledge in different models. We see some positive trends: Llama-3.1 correctly identifies Brazilian cultural norms when thinking in Portuguese, Granite-3.1 does well with Indian norms in Arabic, Phi-4 does much better in English than other languages, and Qwen-2.5 performs well for India, Indonesia, Pakistan and South Africa. Also, generally, there are more lighter colors than darker colors, indicating that the reasoning does help in most cases. The maximum difference in performance between $LLM(D|\ell) - LLM(D)$ is around 30% across all models – this is also a positive sign.

However, we see some negative trends as well: Qwen-2.5 does not perform well in German (despite it being a high resource language), Phi-4 does not correctly identify Canadian cultural norms, and Llama-3.1 does not perform well in Japanese, nor correctly identifies UAE cultural norms. We plan to further investigate the cause of these performance drops.

3 Future Work

We aim to use the findings here to extract multilingual knowledge from language models in a more deliberate manner. For example, Doddapaneni et al. [2024] formulates Cross-Lingual Prompting (CLP) which uses a similar prompting style to enable majority voting for answering questions. We want to smartly choose the languages we use for majority sampling, based on the context. Next, we also would like to train models to do multilingual chain-of-thought reasoning (CoT in multiple languages, not just one) using policy optimization and reinforcement learning.

References

S. Doddapaneni, M. S. U. R. Khan, D. Venkatesh, R. Dabre, A. Kunchukuttan, and M. M. Khapra. Cross-lingual auto evaluation for assessing multilingual llms, 2024. URL `https://arxiv.org/abs/2410.13394`.

Y. Fung, R. Zhao, J. Doo, C. Sun, and H. Ji. Massively multi-cultural knowledge acquisition & lm benchmarking, 2024. URL `https://arxiv.org/abs/2402.09369`.

A. K. Joshi. Processing of sentences with intra-sentential code-switching. In *Coling 1982: Proceedings of the Ninth International Conference on Computational Linguistics*, 1982. URL `https://aclanthology.org/C82-1023/`.