
Investigate arithmetic features in Multi-lingual LLMs

Akash Kundu
Independent

Ashish Rai
Independent

Suhas K R
Independent

With
Goodfire and Apart Research

Abstract

As the adoption of LLMs continue to grow worldwide, the usage of Multi-lingual models will increase, particularly in countries where English is not the primary language. Hence it is important to study the inner-workings of Multi-lingual LLM's and as a step in this direction we investigate the arithmetic related feature activations in Llama3.1 70b model across its 8 supported languages. We use arith-activation¹ strength to compare the 8 languages and unsurprisingly English has the highest strength and Hindi, Thai score the least.

We then studied the most important features related to the mathematical problems in the English examples. We created a variant of the base model by nudging these important features to see if there is an improvement in the performance in other languages. Surprisingly, the performance of the nudged model drops significantly. The responses became very verbose, the reasoning steps increased, but the model tends to hallucinate.

Keywords: Multi-lingual LLMs, AI Observability, Mechanistic Interpretability, Model Reprogramming

1. Introduction

In this project, we aim to study the feature space activation of multi-lingual LLM's in the case of arithmetic problems. LLama 3.1 70b model, supports 8 languages - 'English', 'French', 'Italian', 'Spanish', 'Portuguese', 'German', 'Hindi', and 'Thai'. We use math problems in these languages to study the feature activation space. We try to identify language dependent and language in-dependent features for solving mathematical problems.

We use the GSM8K dataset that comprises of grade school, math word problems. We used google-translate api to translate the first 100 examples in the test-set to the 8 languages.

2. Findings

Feature Activation:

We first use the examples to find the top feature activations in each language. We accumulate the features that contain the word “math” in description and call it activation-strength. We see that for samples in English, the activation is very high compared to other languages.

Language	Activation Strength
English	3.01
Spanish	1.44
Portuguese	1.23
French	1.14
German	1.04
Italian	0.91
Thai	0.54
Hindi	0.51

Language independent features:

Surprisingly, very few of the mathematical and arithmetic related features would fire-up across examples and languages.

We found the following two features consistently across languages -

- ("Question words asking about quantities in math problems")
- ("Mathematical word problems and calculation descriptions")

Feature Intervention:

Using the above two features, we created a variant of the base model. But unfortunately, the performance of the model dropped post this across all languages. The output became very verbose and also the reasoning steps were incorrect.

Language	Base LLama 70b	with Feature intervention
English	19	14
Spanish	17	8
Portuguese	12	7
French	17	9
German	17	9
Italian	17	8
Thai	16	5
Hindi	12	5

These scores are out of 20 examples. (could not run more, as the APIs were taking too long)

3. Discussion and Conclusion

Comparing feature activations across languages was very interesting. We uncovered a bunch of surprising facts. For example, most of the mathematical or arithmetic related features did not fire-up except for very few of them which would consistently fire-up across all examples.

Although we were not able to steer the model towards better performance, we would love to spend more time after the hackathon, to understand why feature-intervention caused a big degradation in performance.