Math Speaks All Languages: Enhancing LLM Problem-Solving Across Multilingual Contexts

Maksim Kostritsya
Sber AI Lab

Kseniia Kuvshinova
Sber AI Lab

Rauf Parchiev
Sber AI Lab

Konstantin Polev
Sber AI Lab

With
Goodfire AI and Apart Research

Abstract

Large language models (LLMs) have shown significant adaptability in tackling various human issues; however, their efficacy in resolving mathematical problems remains inadequate. Recent research has identified steering vectors — hidden attributes that can guide the actions and outputs of LLMs. Nonetheless, the exploration of universal vectors that can consistently affect model responses across different languages is still limited. This project aims to confront two primary challenges in contemporary LLM research by utilizing the Goodfire API to examine whether common latent features can improve mathematical problem-solving capabilities, regardless of the language employed.
Keywords: AI Observability, Mechanistic Interpretability, Model Reprogramming, Math Solving, Steering vectors

1 Introduction

Large Language Models (LLMs) have exhibited exceptional performance across a variety of tasks in both few-shot and zero-shot scenarios [9, 3]. Progress in areas such as scaling laws [7, 1], model quantization [10, 5], and fine-tuning has further augmented their capabilities and efficiency. Nevertheless, despite these advancements, the resolution of mathematical problems continues to pose a considerable challenge for LLMs [8, 2]. They frequently struggle to generate accurate answers, revealing insufficiency in reasoning and cognitive skills – essential criteria for assessing LLM's intelligence and generalization. This project aims to tackle the ongoing limitations of LLMs in addressing mathematical problems, especially within multilingual frameworks. Instead of depending on specialized training or domain-specific modifications, we introduce another strategy utilizing the Goodfire API [6] to pinpoint steering vectors that reinforce mathematical reasoning. We hypothesize that mechanisms represented by such vectors should be universal across multiple languages, and so we aspire to identify shared characteristics across languages and establish a universal framework for directing LLM behavior in mathematical problem-solving. By emphasizing cross-linguistic universality, this strategy not only aims to elevate the quality of mathematical reasoning in multilingual LLMs but also prompts critical inquiries regarding their interpretability and adaptability across different languages—ultimately contributing to the development of more resilient and inclusive AI systems.

2 Overview

2.1 Hypotheses

Hypothesis: We hypothesize that there is a subset of SAE features that can enhance the LLM's mathematical abilities in case of problems that the original model (without steering) failed to solve. Moreover, we suspect that such features should be language-agnostic and capture universal patterns across different languages.

Methodology: We compare the performance of LLM with and without SAE feature steering on random 100 grade school math problems from the English GSM8K dataset [4]. The selected problems were checked to include different types of mathematics, including algebra, probability, and geometry, to ensure a comprehensive evaluation. These samples were then translated using Mistral and Google Translate to French and Russian languages. All our experiments were performed using Goodfire API's *meta-llama/Meta-Llama-3-8B-Instruct*; steering was used in nudge mode with "value" set to 0.5 (unless mentioned otherwise). Generations were done in 5-shot mode (prompts written entirely in each respective language). To reduce the effects of stochastic variability each generation was repeated 3 times; if the label (correct/incorrect) was not the same across all 3 samples, we performed additional generations and the final label was selected on a best of 5 basis.

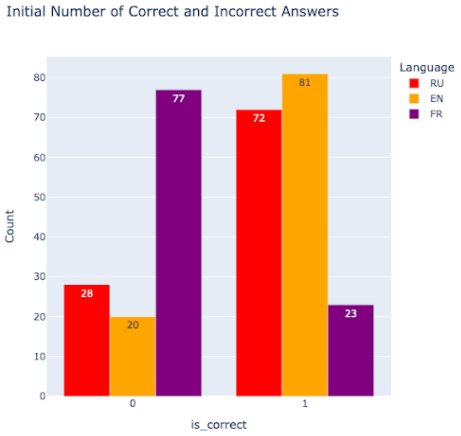


Figure 1: Models performance on cross-languages datasets. 0 stands for not correct answer, 1 for correct one.

Evaluation: To assess performance, we initially chose to utilize GPT-4o. However, upon observing suboptimal performance on some samples, we switched to a manual human comparison between the generated answers and the ground truth.

Analysis: We compare the performance and the nature of the features extracted for each language. By examining the commonalities and differences in the feature representations, we aim to identify universal SAE features that can be leveraged to improve mathematical reasoning in multilingual LLMs. In a nutshell, we get top-n features for each dataset. These features are unique and annotated. Then we create a set of name strings for each dataset, and find a set intersection of these strings.

2.2 Experiments

We used Goodfire API to evaluate models answers on these datasets without steering and got such correct/incorrect rate (see Figure 1).

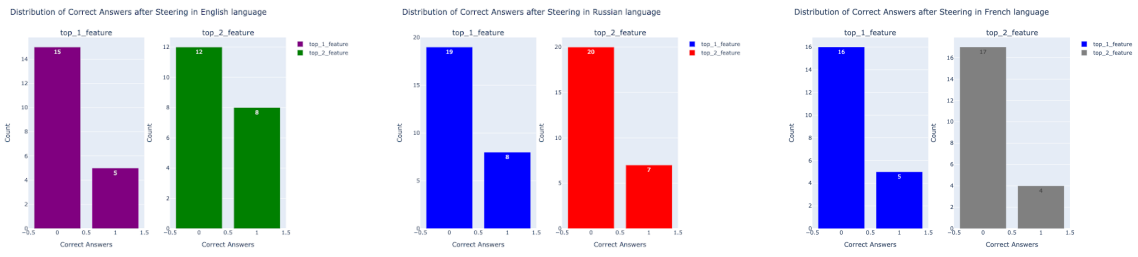


Figure 2: Steering generation results with top 2 features for each language.

The analysis revealed that the large language model (LLM) demonstrates inferior performance in French relative to its performance in Russian and English. This methodology also provided a solution to the limitations associated with the Goodfire API, where sampling parameters could not be explicitly defined or controlled.

After that we used contrastive search between incorrect and correct answers to find features leading the model to make errors in her reasoning. For each language we saved the top 2 most accurate features and generated them separately (see Figure 2).

2.3 Results

We improved models performance on samples where the base model fails to solve tasks! The next step is to find any intersections between these sets. We inspected Russian and English SAE Features have intersection - feature: "Arithmetic operations in word problems". After that our goal was to check if this feature transfers to the French language and improve generation (see Figure 3).

To further investigate the robustness of latent features across languages, we conducted an experiment focusing on the transferability of a feature derived from the French dataset to other languages. Specifically, we aimed to determine whether a feature extracted from contrasting correct and incorrect French answers could improve the model's performance not only in French but also in English and Russian.

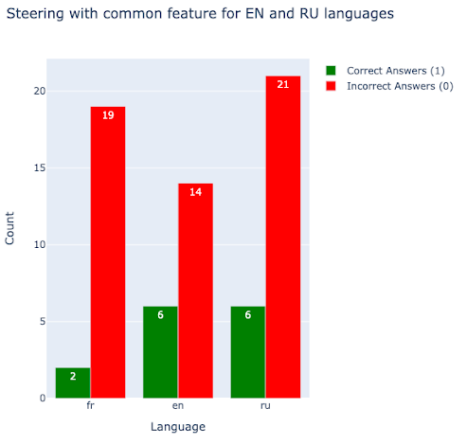


Figure 3: Steering generation results with common features for EN and RU languages.

Applying the French feature corrected only 2 out of 21 incorrect French samples but significantly improved performance in English and Russian, correcting 12 out of 20 and 9 out of 27 incorrect samples, respectively (see Figure 4). This unexpected outcome suggests that while the feature captures universal aspects of mathematical reasoning, the model's proficiency in English and Russian allows it to leverage the steering more effectively. These findings support our hypothesis of language-agnostic latent features within LLMs and demonstrate the potential of universal steering for mathematical problem-solving capabilities across different languages.

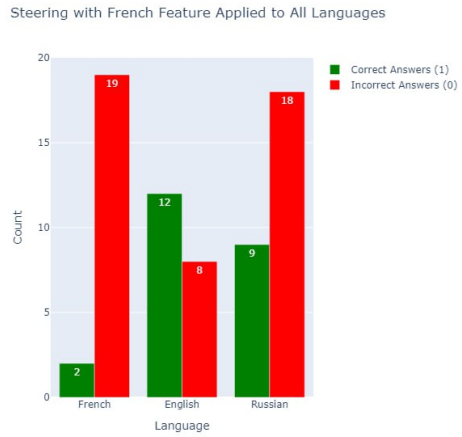


Figure 4: Cross-language feature transfer: Steering with French feature applied to all languages.

The next step was to find nearest neighbors features in decoder weights, provided as a method in Goodfire API. We took the top 2 features from each language and got a very interesting set (See in Appendix 1).

The next experiment was generated with these features, we used the first one in this set, due time limitations (see Figure 5(a)).

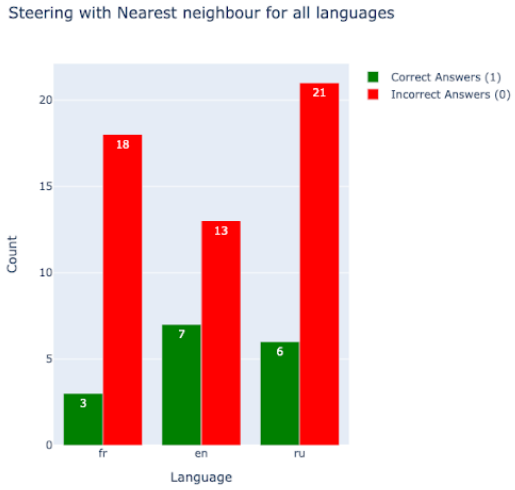
In previous experiments we steer model generation only with one SAE feature. Next step was to check out the generation with a group of features. This group consists of the top 5 features found using nearest neighbors method (see Appendix 2). Using this features set we got new results (see Figure 5(b)).

We achieve the best correct/incorrect ratio on incorrect cross-language samples!

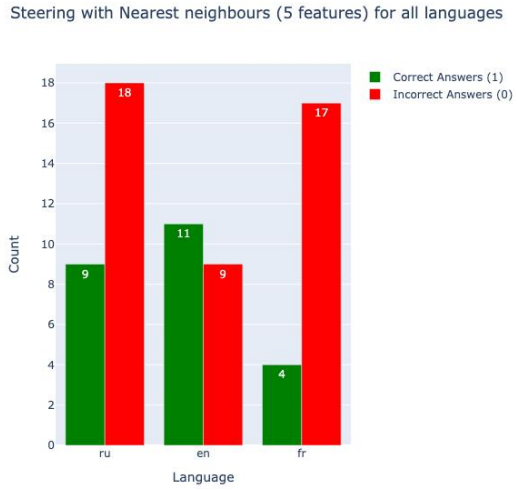
2.4 Future work

The findings from our experiments reveal some universal features that can boost mathematical problem-solving skills in different languages. By applying these features, we significantly improved the performance of LLMs on mathematical tasks in a multilingual setting, all without requiring extra training. These insights pave the way for future investigations into AI interpretability for mathematical tasks.

The expansion of this research presents various opportunities to improve the multilingual capabilities and interpretability of large language models (LLMs). One of these is comprehensive benchmarking for our hypotheses, including more both feature-steered and baseline model, as well as improved comparisons on a wider range of mathematical tasks. Another important



(a) Steering generation results with top feature from common nearest neighbor for all languages.



(b) Steering generation results with set of nearest neighbors (5 features).

direction of work is to check our approach not only on european languages, but Arabic, Japanese, or Chinese. This approach will yield more profound insights into the universality and constraints of steering features across languages with distinct syntactic and morphological structures.

3 Code

All our code and datasets are available at `https://github.com/MaksimKoster/goodfire_ai_hack`. We used Mistral API and Google translation to transfer datasets from English to Russian and French languages. For feature searching we used features.contrast and features.search methods from Goodfire API, in our experiments the first one shows better performance. All experiments were run on meta-llama/Meta-Llama-3-8B-Instruct.

4 Discussion and Conclusion

Identifying common features for mathematical tasks between different languages presents numerous opportunities for advancing AI research. It can boost the creation of AI agents that use language-agnostic mathematical reasoning to address challenges in multilingual environments without need of retraining or fine-tuning, which is especially important for small or low-resource languages. It also opens a path to understand universal representations that bridge the gap between text-based and mathematical reasoning. An interesting direction is to investigate whether models grounded on universal mathematical features are inherently more robust to adversarial attacks or domain/distribution shifts. We also believe our research will help to enhance the capacity of LLMs to work with mathematical proofs and symbolic computations.

References

[1] Armen Aghajanyan et al. “Scaling laws for generative mixed-modal language models”. In: *International Conference on Machine Learning*. PMLR. 2023, pp. 265–279.

[2] Janice Ahn et al. “Large language models for mathematical reasoning: Progresses and challenges”. In: *arXiv preprint arXiv:2402.00157* (2024).

[3] Jinze Bai et al. “Qwen technical report”. In: *arXiv preprint arXiv:2309.16609* (2023).

[4] Karl Cobbe et al. “Training Verifiers to Solve Math Word Problems”. In: *arXiv preprint arXiv:2110.14168* (2021).

[5] Amir Gholami et al. “A survey of quantization methods for efficient neural network inference”. In: *Low-Power Computer Vision*. Chapman and Hall/CRC, 2022, pp. 291–326.

[6] Goodfire. *Goodfire API*. `https://docs.goodfire.ai/`. Accessed: 2024-11-24. 2024.

[7] Berivan Isik et al. “Scaling laws for downstream task performance of large language models”. In: *arXiv preprint arXiv:2402.04177* (2024).

[8] KV Srivatsa and Ekaterina Kochmar. “What Makes Math Word Problems Challenging for LLMs?” In: *arXiv preprint arXiv:2403.11369* (2024).

[9] Hugo Touvron et al. “LLaMA: open and efficient foundation language models. arXiv”. In: *arXiv preprint arXiv:2302.13971* (2023).

[10] Zifei Xu et al. “Scaling laws for post-training quantized large language models”. In: *arXiv preprint arXiv:2410.12119* (2024).

5 Appendix

0	”Processing simple arithmetic word problems”
1	”Calculating time or money required to complete a task or reach a goal”
2	”Explanatory connectives in step-by-step reasoning”
3	”Mathematical reasoning in word problems”
4	”Numerical values in math problems or quantitative descriptions”
5	”Small numbers (especially 2) in arithmetic contexts”
6	”Animals in logical reasoning puzzles”
7	”Multiplication and division operators in mathematical expressions”
8	”The model is fact-checking a summary against a document”
9	”Mathematical problem-solving with costs and quantities”
10	”Polite and appreciative language in structured formal communication”
11	”The user has a question about task duration or time calculations”

Table 1: Set of common Math SAE features.

0	”Processing simple arithmetic word problems”
1	”Calculating time or money required to complete a task or reach a goal”
2	”Explanatory connectives in step-by-step reasoning”
3	”Mathematical reasoning in word problems”
4	”Numerical values in math problems or quantitative descriptions”

Table 2: Set of Math correctness SAE features.