Exploring performance of LLMs fine-tuned on synthetic code-switched text

Tanay Nagar, Anna Sokol, Grigorii Khvatskii, Nitesh V. Chawla
Center for Computer-Assisted Synthesis, University of Notre Dame

Introduction

Language models (LMs) often underperform in low-resource languages due to imbalanced training data[1]. Our research aims to mitigate this bias by fine-tuning LMs on synthetic code-switched data, where multiple languages are mixed within sentences.

Motivation:
- **Language Imbalance**: High-resource languages dominate training data, disadvantaging low-resource languages.
- **Performance Gap**: LMs perform poorly in low-resource languages, creating inequities.

Research Questions:
1. Can synthetic code-switched data improve LLM performance in low-resource languages?
2. Does this fine-tuning affect high-resource language performance?

Methods

Data Generation:

Synthetic Code-Switched Text:
- **GPT-3.5**: Generated Hindi-English code-switched text using specific prompts.
- **mt5-Small[2]**: Controlled code-mixed text generation with language ratios in three buckets: [0, 0.167] (cmi 1), [0.167, 0.3](cmi 2), and [0.3, 0.5] (cmi 3).

Dataset Preparation:

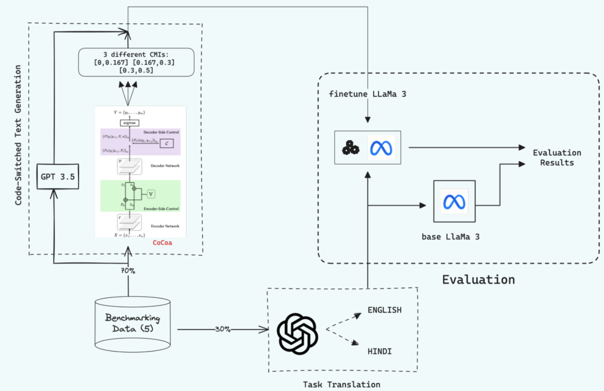
•**Common Sense Reasoning (CSR) Dataset**: Converted multiple-choice questions in English into code-switched text using the above methods, creating four datasets: GPT generated, CMI1, CMI2, and CMI3.

Fine-Tuning: LLaMa3 model fine-tuned for 2 epochs on the code-switched CSR dataset.

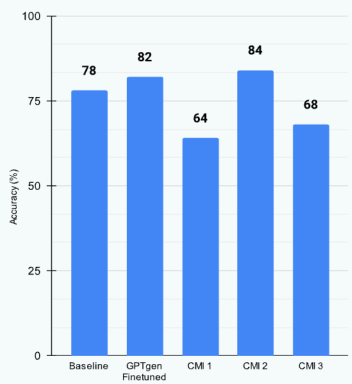
Evaluation:

•**Performance Metric**: Accuracy of correct answers over 5 iterations.
•**Baseline Comparison**: Calculated baseline scores for LLAMA3 without fine-tuning for comparison.

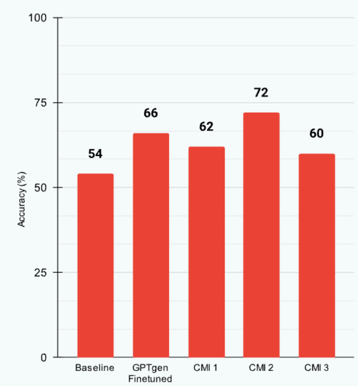
Methods Pipeline



Results



Accuracy for English CSR Questions



Accuracy for Hindi CSR Questions

Key Findings:
- **Improvement in Hindi**: All fine-tuned models demonstrated significant improvements in Hindi accuracy, showcasing the effectiveness of synthetic code-switched text in enhancing performance in low-resource languages.
- **Preservation of English Performance:** Two models, GPTgen and CMI2, not only improved performance in Hindi but also preserved or enhanced performance in English, indicating a balanced approach to multilingual enhancement.

Conclusion

Our research demonstrates partial success in improving low-resource language performance using synthetic code-switched text. Models finetuned on GPTgen and CMI2 showed significant improvements in Hindi while preserving or enhancing English performance.

Future Work:
Further experimentation with more specific Code-Mixed Indexes (CMIs) is needed to identify the optimal language ratios.

Impact:
- **Language Equity**: Ensures fair treatment of all languages.
- **Real-World Benefits**: Enhances multilingual support in diverse linguistic settings.

References

1. Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't Trust ChatGPT when your Question is not in English: A Study of Multilingual Abilities and Types of LLMs. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023. Association for Computational Linguistics, Singapore, 7915–7927. https://doi.org/10.18653/v1/2023.emnlp-main.491
2. Sneha Mondal, Ritika ., Shreya Pathak, Preethi Jyothi, and Aravindan Raghuveer. 2022. CoCoa: An Encoder-Decoder Model for Controllable Code-switched Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2466–2479, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
3. Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2024. Quantifying Multilingual Performance of Large Language Models Across Languages. Retrieved July 22, 2024 from http://arxiv.org/abs/2404.11553

Acknowledgements

This work sponsored by the NSF Center for Computer-Assisted Synthesis, an NSF Center for Chemical Innovation (CHE – 2202653).