Breaking Language Barriers: Equitable Performance in Multilingual Language Models

Tanay Nagar*, Anna Sokol, Grigorii Khvatskii, Nitesh V. Chawla

Lucy Family Institute for Data & Society · University of Notre Dame · University of Wisconsin - Madison

INTRODUCTION

Language models (LMs) excel on high-resource languages but underperform on low-resource ones[1] due to imbalanced training data.

On the CommonsenseQA benchmark, accuracy falls from **78%** in English to **54%** in Hindi, despite identical task formats. This gap highlights a structural bias that limits the utility of LMs for hundreds of millions of non-English speakers.

Motivation:

- Data imbalance: Pre-training corpora are overwhelmingly dominated by English and other high-resource languages, leaving low-resource languages under-represented. LRL fine-tuning leads to reduced HRL performance too.[3]
- Equity implications: A **24 percentage-point** accuracy gap excludes over **600 million** Hindi speakers from reliable AI-powered reasoning tools in education, healthcare, and more.

Research Questions:

1. Can synthetic code-switched data narrow the performance gap on commonsense reasoning tasks for Hindi?
2. Will code-switched fine-tuning preserve—or even improve—English reasoning accuracy?

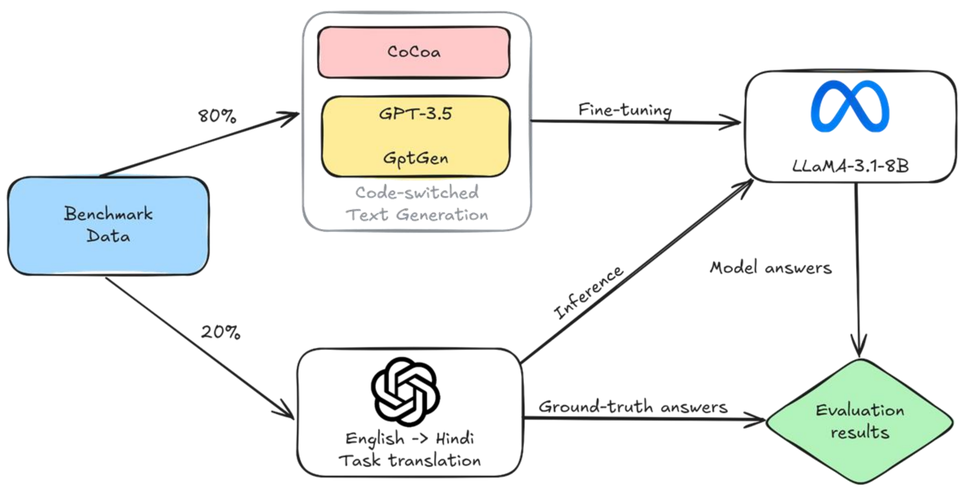
TERMINOLOGY

Code-Switching
Alternating between two or more languages within the same sentence.

CMI (Code-Mixing Index)
Percentage-based measure of language mixing (0 % = monolingual, 50 % = equal mix).

HRL / LRL
High-Resource Language / Low-Resource Language (data availability)

PIPELINE



METHODS

Data generation

- **GPT-3.5** prompting for free-mix Hinglish
- **CoCoa**[2] model to control CMI at low/medium/high levels

Fine-tuning setup

- Model: Meta LLaMA-3 8B-Instruct
- Method: QLoRA quantized training
- Hyperparams: 5 epochs, LR=3×10^{-5}, batch=32

Evaluation protocol

- **5-fold cross-validation** on CommonsenseQA in English & Hindi
- Each question run 5×; select majority vote → report mean ± std accuracy
- **Baseline**: un-fine-tuned LLaMA-3 scores (EN/HI) for direct comparison

DATASET

Source: CommonsenseQA (12 102 multiple-choice questions).

Conversion: Transformed questions to Hindi–English code-switched text; answers left in English.

Splits (1 200 questions each):

- GPTgen (free mixing via GPT-3.5)
- CMI-1 (0–16.7 %)
- CMI-2 (16.7–30 %)
- CMI-3 (30–50 %)

Verification: Manual review of 1/50 samples per split to ensure coherence.

CMI EXAMPLE

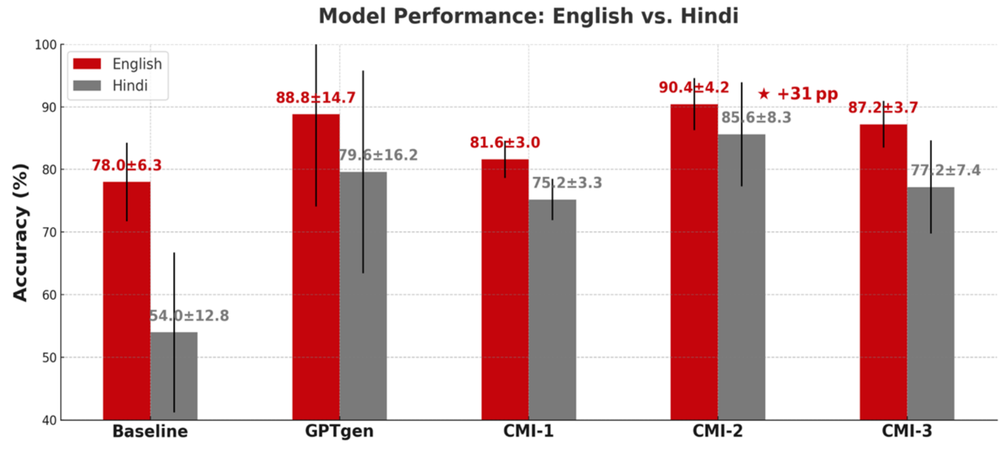
What is it called when you slowly cook using a grill? A) backyard B) restaurant C) crockpot D) neighbor's house E) barbeque	
CMI 1	जब आप grill का उपयोग करके slowly खाना पकाते हैं तो उसे क्या कहते हैं
CMI 2	जब आप grill का use करके slowly खाना पकाते हैं तो उसे क्या कहते हैं
CMI 3	जब आप grill का use करके slowly dinner पकाते हैं तो उसे क्या कहते हैं
GPTgen	इसे क्या कहते हैं जब आप धीरे-धीरे ग्रिल का उपयोग करके खाना पकाते हैं?

RESULTS

Our study demonstrates that **fine-tuning a multilingual language model on synthetic code-switched text can significantly improve performance on low-resource languages**. Specifically, we found that a medium level of code-mixing intensity yields optimal results, with accuracy improving to 85.6% on Hindi tasks and 90.4% on English tasks. This approach effectively closes the performance gap between high-resource and low-resource languages, reducing the delta from 24% to just 5%.

- CMI-2 (medium mixing) yields 90.4 % EN and 85.6 % HI, outranking all other variants.
- All code-switched fine-tuned models beat baseline by +12.4 pp (EN) and +31.6 pp (HI).

Model Performance: English vs. Hindi



CMI-2 fine-tuning closes 80 % of the Hindi–English reasoning gap without any English degradation

DISCUSSION

- Limitations: evaluation limited to a single language pair (Hindi–English) and reliance on synthetic code-switched data.
- Future work: expand to additional low-resource language pairs and benchmark against monolingual fine-tuning baselines.
- Implications: this code-switched fine-tuning approach bridges HRL–LRL performance gaps, advancing more accessible AI for underrepresented communities.

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