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

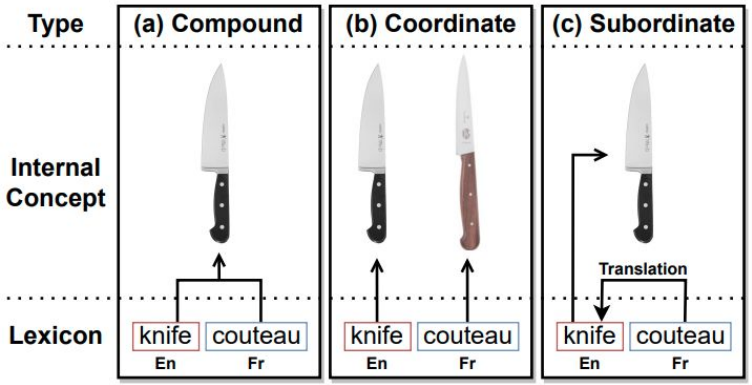
June 24, 2024

INTRODUCTION

Because of an imbalance of training data between languages, LLMs perform worse in low-resource languages across multiple tasks. This has big negative implications for downstream tasks. We explore the use of synthetic code-switched data via fine-tuning to improve performance in low-resource languages while preserving performance in high-resource languages.

TERMINOLOGY: Bi/Multi-lingualism

a. **Compound**: Learning two languages simultaneously in the same environment, resulting in a fused cognitive system with shared concepts. This promises consistent performance across multiple languages
b. **Coordinate**: Learning two languages in separate contexts, creating distinct cognitive systems for each language. This leads to inconsistencies with multi-lingual data input and output.
c. **Subordinate**: Learning a second language through the first language, leading to a translation-based understanding. This leads to poor performance in all low-resource languages



PROBLEM (Part 1)

Language Data Imbalance ⟶ Performance Disparity

Imbalance in Training Data:

- LLMs are trained with a significant amount of data in high-resource languages (e.g., English).
- Low-resource languages (e.g., Hindi, Yoruba) are underrepresented in training datasets.

Performance Disparity:

- LLMs perform much better for high-resource languages compared to low-resource languages.
- This creates a performance gap and inequity in language model outputs.

Meta AI Blog Post:

"To prepare for upcoming multilingual use cases, over 5% of the Llama 3 pre training dataset consists of high-quality non-English data that covers over 30 languages. However, we do not expect the same level of performance in these languages as in English."

PROBLEM (Part 2)

Impact and Implications

Impact on Language Equity:

- Speakers of low-resource languages receive consistently worse performance from LLMs.
- This goes against the principles of language equity (refers to the fair and equal treatment of all languages and their speakers)

Real-World Implications:

- In multilingual societies, equitable language support ensures access to services and opportunities. In bilingual classrooms, like Spanglish environments, poor LLM performance can hinder learning and widen educational disparities.

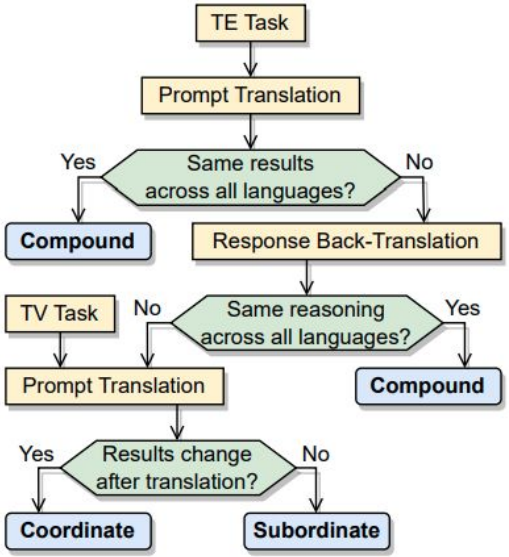
TERMINOLOGY: Tasks

Translation-Equivariant(TE) Tasks

- Reasoning (GSM8K and commonsenseQA)
- Knowledge Access (QA pairs)
- Articulation (prompt model to generate cover letter, followed by manual evaluation)

Translation-Variant(TV) Tasks

- Letter Writing
- Pun Understanding



Logical flowchart on how TE and TV Task performance indicates bilingualism (Zhang et al, 2023)

TERMINOLOGY: Task Examples

Example CommonSenseQA:

Q: Where on a river can you hold a cup upright to catch water on a sunny day?
A: 👍 waterfall, 👎 bridge, 👎 valley , 👎 pebble, 👎 mountain

Example KnowledgeAcessQA:

Q:Who invented the ford motor company?
A:Henry Ford, B: Michael Copon, D: Corbin Bleu, D: Joe Jonas, E: Jon Bernthal

Example Letter Writing Task:

You are Johnson Smith from University of Alberta with A GPA of 3.9. You like sapping. You want to join Huawei company. Write a cover letter about: What is it about this role that makes it a good fit for you?, What's something outside of your work that you're passionate about?, and What does your next ideal role look like?

Example Pun Understanding Task:

The bicycle can't stand on its own, since it's too-tired.

Task	**En**	Fr	De	Es	Ja	Zh
MR	**0.90**	0.80	0.78	0.80	0.82	0.78
CSR	**0.68**	0.58	0.52	0.54	0.48	0.52
KA	**0.96**	**0.96**	0.94	0.94	0.80	0.68

Table 1: Accuracy for TE tasks: math reasoning (MR), commonsense reasoning (CSR), and knowledge access (KA).

Results on TE Tasks (Zhang et al, 2023)

Chinese	**English Translation**	**Frequency**
诚挚地	Sincerely	54.0%
致意	Regards	38.4%
祝愿	Best Wishes	3.6%
此致敬礼	Salute (Proper Chinese Sign-off)	0.8%
No sign-off		3.2%

Table 2: The frequency of different sign-offs in 250 different Chinese cover letters generated by ChatGPT.

Results on Letter Writing Task (Zhang et al, 2023)

Language	**P-Acc**	**L-Acc**
Es	0.488	0.697
Es-En	0.507	0.714
Fr	0.500	0.886
Fr-En	0.513	0.813
En	0.506	0.965
En-Fr	0.500	0.646
En-De	0.519	-
En-Es	0.488	0.607
En-Ja	0.519	-
En-Zh	0.550	0.511

Table 3: Accuracy on pun detection (P-Acc) and location (L-Acc). X-Y means the puns were translated from language X to language Y before prompting.

Zhang et al, 2023

WHY THIS MATTERS

Promote Language Equity:

- Equal performance across all languages ensures fair access to technology

Supports multilingual education:

- Enhanced LLMs improve learning outcomes in multilingual classrooms by effectively handling multiple languages

Advance NLP Research:

- Tackling language biases advances the development of more inclusive and robust AI technologies. And of course, multiple wider implications in improving accessibility across applications

Gender Bias:

Teacher:
- शिक्षक (for male teacher)

- शिक्षिका (for female teacher)

Cultural Bias:

Date:
- डेट (for romantic meeting)

- तारीख (for calendar date)

TERMINOLOGY: Code-switching, CMI, SPI

Code-Switching:

- Practice of alternating between two or more languages or dialects within a single conversation, sentence, or even phrase.

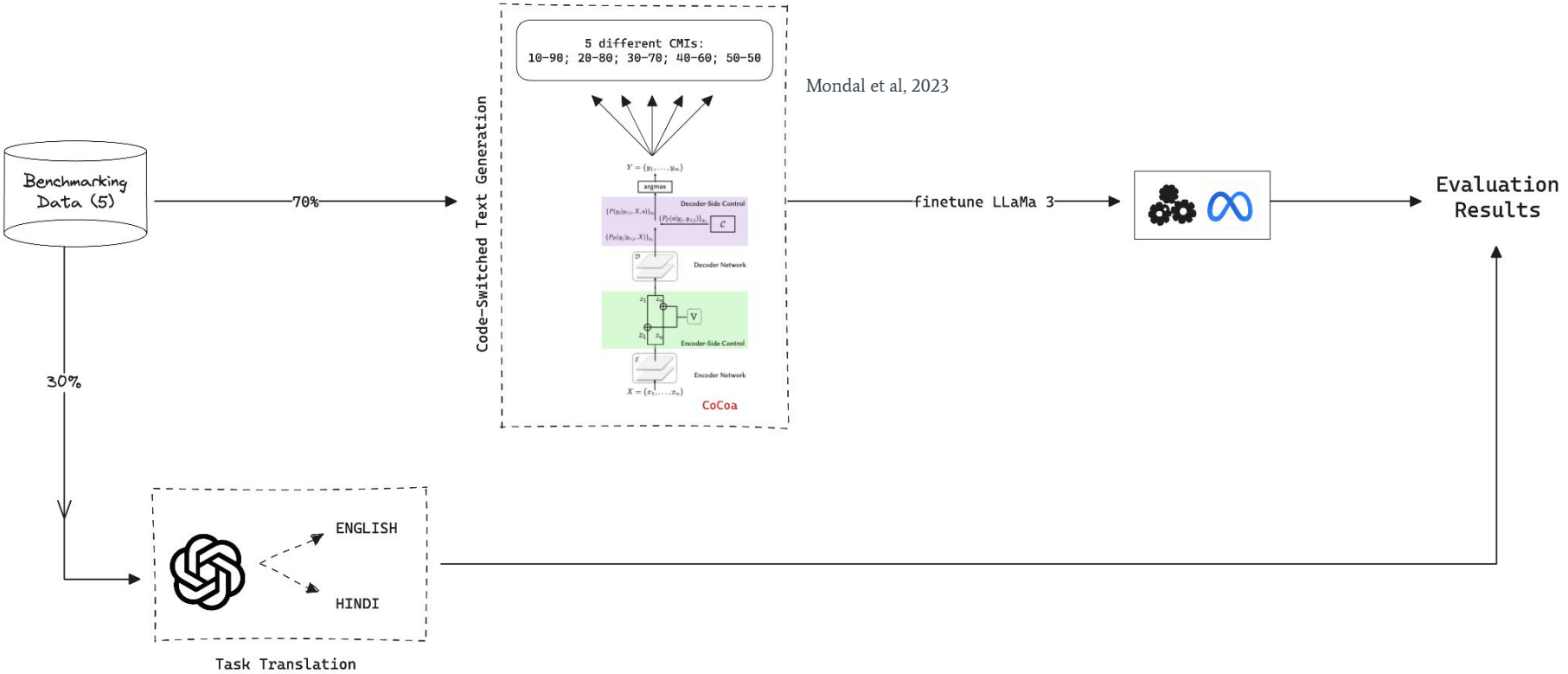
CMI: Code-Mixing Index

- Commonly used to quantify the ratio of L1 vs. L2

SPI: Switch-Point Index

- Commonly used to quantify the extent of burstiness

PIPELINE


Benchmarking Data (5)
70%
30%
Code-Switched Text Generation
5 different CMIs:
10-90; 20-80; 30-70; 40-60; 50-50
Mondal et al, 2023
argmax
Decoder-Side Control
Decoder Network
Encoder-Side Control
Encoder Network
CoCoa
finetune LLaMa 3
Evaluation Results
ENGLISH
HINDI
Task Translation

EVALUATION

Compare fine-tuned Llama performance with base Llama 3:

1. Based on Accuracy (number of tasks performed correctly):
   a. 5 different datasets; each task evaluated 5 times; then accuracy calculated

Task	En	Hi
CSR	85%	75%

Common Sense Reasoning

Our Contributions

Baseline Scores: We replicate Zhang et al's study with Hindi and new Llama 3 model

Innovative Data Generation: We create synthetic, code-switched datasets for fine tuning and evaluation + we can expand research to Spanglish too

Effects of fine-tuning with varying CMIs: We can pinpoint which code-switched text lends best results to transfer learning across languages

Challenges and Limitations

Generating Code-Switched Text:

Current Progress and Next Steps

Mostly just learning (fine-tuning techniques, literature reviews on past papers, basic famous models (Enc-Dec, LSTM, GRU, RNN,...), replicating code for TCS & GCM model...

Calculated baseline scores for Llama 3 for CSR...showed same difference across English and Hindi languages