Multilingual Trade-offs in Fine-tuned Language Models

This project investigates how fine tuning a language model on one language may affect its ability to handle other languages. This research follows XLM-RoBERTa, fine tuned on English and Hindi using the XNLI dataset. The XNLI Dataset is a benchmark dataset designed to evaluate language models across different languages. It explores the trade off that occurs when the language model fine tuned on English is run against Hindi, and vice versa. The goal is to understand if improving the performance in one language comes at the cost of another, what changes happen inside the model, and what kind of mistakes the model can make. The approach was to fine tune the same base model on hindi and english data separately, run experiments and analyze how and where it affects the model.

Experiment 1: Behavioral Analysis

The finetuned models were evaluated on English and Hindi XNLI testing samples, to measure performance tradeoffs.

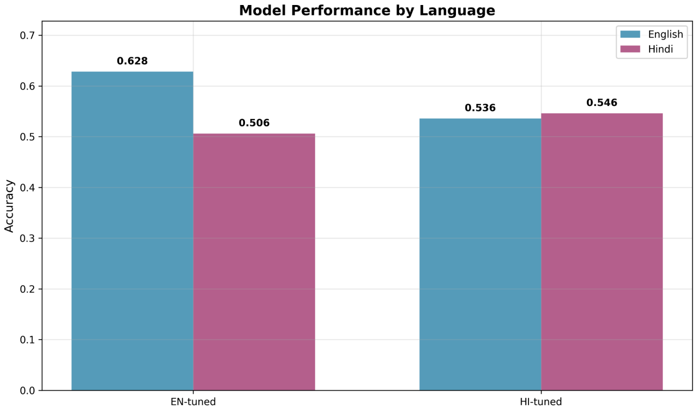


Figure 1: Calculates how each model accurately performs on Hindi versus English tasks. The English fine tuned model is better at English (62.8%) than Hindi (50.6%), performing 12.2% in the favour of English.

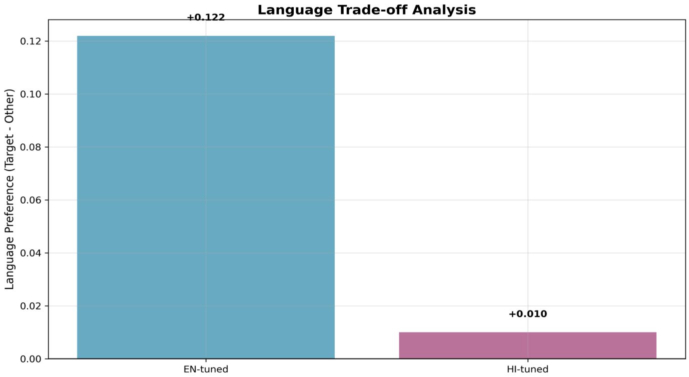


Figure 2: Calculates the bias toward the target language by finding the difference between target and non-target language performance within the same model. The english finetuned model shows a

+0.122 preference for English, while the hindi finetuned model shows only +0.01 preference for Hindi. This indicates that English fine tuning creates a stronger bias than Hindi fine tuning.

Experiment 2: Activation Analysis

Performed to measure the similarity between english tuned and hindi tune models across all 12 layers of the transformer.

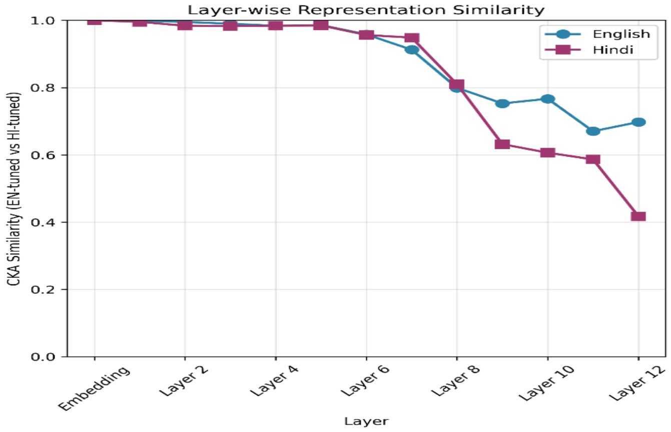


Figure 3: The line graph shows how similar the internal representations are between english tuned and hindi tuned models across all layers. Layers 0-6 show similar representations, changes begin around Layer 7 by layer 12, Hindi model shows a 42% similarity while the English model depicts 70%, indicating that the models develop completely different paths when nearing the final layer.

Experiment 3: Attention Analysis

Tests whether the attention becomes more scattered (higher entropy) or becomes more focused (lower entropy) after fine tuning.

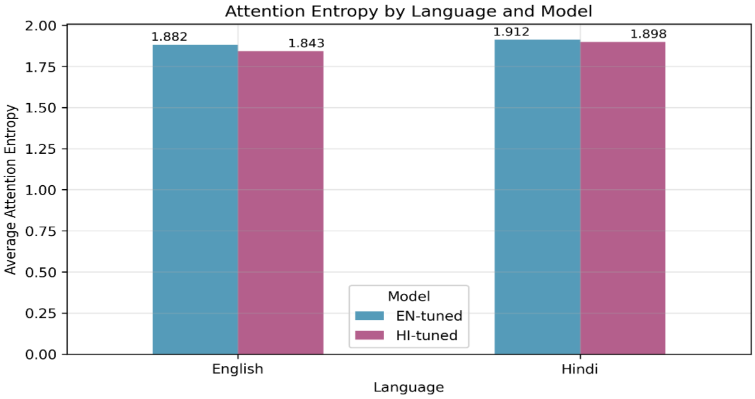


Figure 4: Both models show more scattered attention when processing Hindi inputs (1.9) compared to English inputs (1.85). The hindi tuned model shows more focused attention overall, with a clear difference on English inputs (1.843 vs 1.882).

Experiment 4: Confidence Analysis

Calculated prediction confidence and calibration with the help of softmax probabilities, taken from the model output samples.

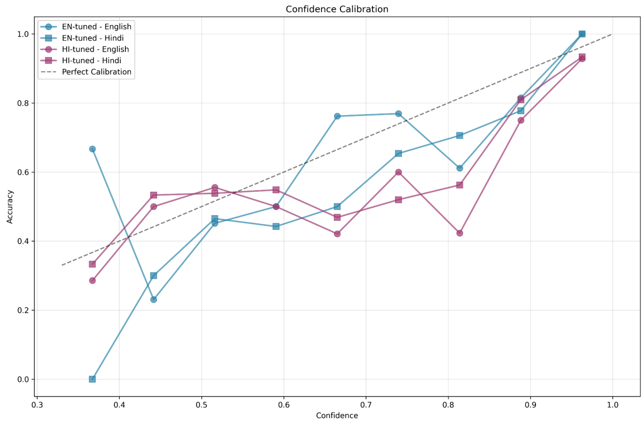


Figure 5: This scatter plot shows whether high confidence correlates with high accuracy. The English tuned model on English shows good calibration, as it is close to the diagonal. However, both models show poor calibration on non-target languages ,they point below the line at medium confidence levels. This shows systematic overconfidence on non-target languages.

Experiment 5: Error Analysis

Predicts errors using confusion matrices to identify systematic biases. Analyzed samples to determine which NLI categories (entailment / neutral / contradiction) suffered more from multilingual trade offs.

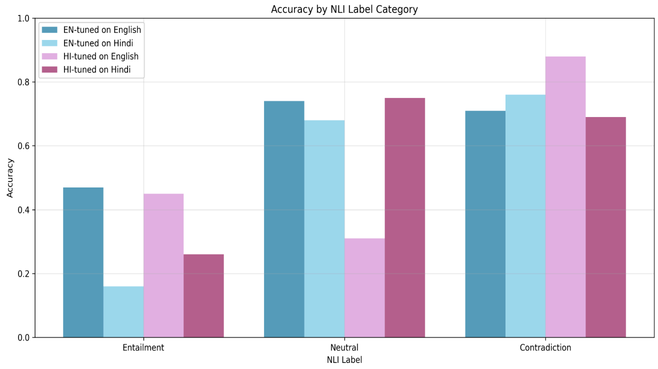


Figure 6: English tuned model drops from 47% to 16% on entailment accuracy when tested against Hindi. Contradiction remains relatively strong across all conditions i.e. 69-88%, and neutral shows moderate performance i.e. 31-75%. This signifies that entailment, the positive logical relationship is most vulnerable to cross language interference.

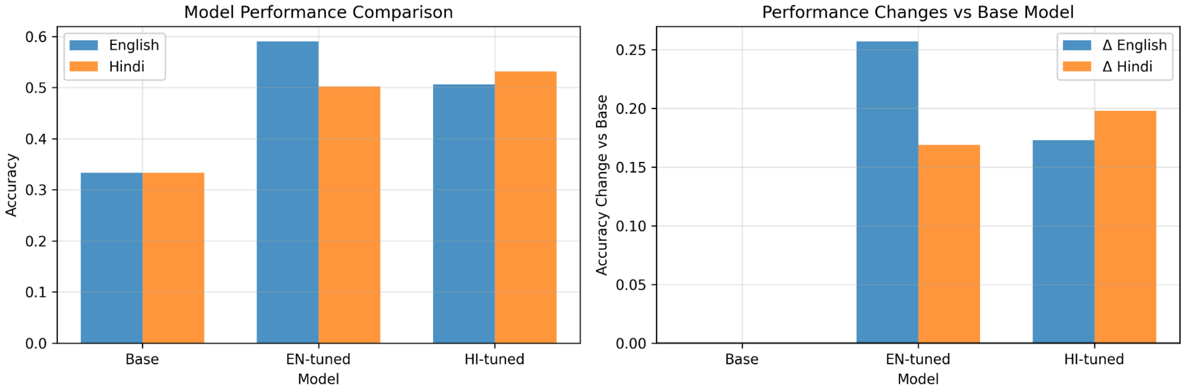
Dataset

- XNLI: Cross-lingual Natural Language Inference corpus
- Core Task: Natural Language Interface (NLI), the model determines the logical relationship between a premise and a hypothesis. The three possible relationships are entailment (the

hypothesis is true), contradiction (the hypothesis is false), and neutral (the relationship is unclear).

Limitations and Problems faced

The most significant setback was discovering that XLM-RoBERTa's classification head has randomly initialized weights, making base model comparison meaningless. Had to abandon a planned analysis component and reduced the interpretability of the model's improvements.



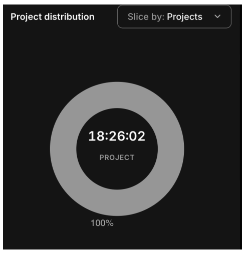
Core Findings

1. English finetuning creates a strong language bias, while Hindi finetuning has minimal specialization. This asymmetry suggests different languages respond differently to similar fine tuning methods.
2. The multilingual tradeoffs result from late layer specialization while preserving early layer language knowledge.
3. Models develop labelling bias towards non target language, as shown from the entailment drop. It reveals that logical relationships are disproportionately affected by language transfer.

Related Work

1. https://arxiv.org/pdf/2502.11223
   The asymmetric language effects they found in machine translation also happens in this research, through the NLI classification task. The English centric bias they observed is also present in this research, as fine tuning for English creates a much stronger model than that for the non target language.
2. https://aclanthology.org/2025.fever-1.5.pdf
   They identify general label bias and performance drop, this research provides a finer analysis, focus on logical relationships and pinpointing to the drop in entailment.

Total Project Time:


Project distribution
Slice by: Projects
18:26:02
PROJECT
100%