Results: Zero-shot Multilingual NLI Evaluation on XNLI

Experimental Setting

All results reported in this document correspond to *zero-shot* evaluation on the XNLI validation set. No additional task-specific fine-tuning or calibration was performed. The goal of these experiments is to analyze cross-lingual model behavior rather than optimize performance.

The evaluation considers English and Indic languages (Hindi and Urdu) to enable controlled cross-lingual comparison.

Quantitative Results

Language	**mBERT**	**XLM-R**
English	∼0.33	∼0.33
Hindi	∼0.33	∼0.34
Urdu	∼0.33	∼0.33

Table 1: Zero-shot XNLI accuracy across languages

Key Observations

- Both mBERT and XLM-R achieve near-chance accuracy across all evaluated languages.
- XLM-R does not significantly outperform the base mBERT model in this zero-shot setting.
- Performance differences across languages are small, indicating that uniformly low accuracy can mask deeper cross-lingual limitations.

Interpretation

These results highlight the difficulty of zero-shot multilingual Natural Language Inference and the sensitivity of such evaluations to task adaptation and preprocessing choices. While multilingual models promise broad language coverage, naive deployment without task-specific calibration can lead to uniformly poor performance across languages.