Training on Artificially Code-Switched Data for Robust Multilingual LLMs in Cross-Lingual NLI

Assignee Research

July 2, 2026

Abstract

Large language models (LLMs) are increasingly applied in multilingual contexts, yet their capacity for consistent, logically grounded alignment across languages remains underexplored. We present a controlled evaluation framework for multilingual natural language inference (NLI) that generates synthetic, logic-based premise-hypothesis pairs and translates them into a typologically diverse set of languages. This design enables precise control over semantic relations and allows testing in both monolingual and mixed-language (code-switched) conditions. Surprisingly, code-switching does not degrade

1 Introduction

This paper examines: Evaluating Multilingual and Code-Switched Alignment in LLMs via Synthetic Natural Language Inference. Research question: How does training on artificially code-switched data affect the robustness of multilingual LLMs against adversarial perturbations in cross-lingual natural language inference tasks compared to monolingual training?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.5/10.

3 Results

15 papers retrieved. 14 claims extracted; 11 independently verified. Quality review score: 7.5/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The selected languages—Arabic (ar), German (de), French (fr), Hindi (hi), and Swahili (sw)—cover both high- and low-reso	✓	0.33
Their scripts include Latin, Arabic, and Devanagari, introducing distinct orthographic and tokenization challenges.	✓	0.22
This selection also varies in morphological complexity, syntactic structure, and resource availability, providing a comp	✓	0.26
The resulting diversity helps surface weaknesses that might remain hidden in homogeneous and high-resource-only evaluati	✓	0.19
To further stress-test semantic alignment, a code-switching condition is introduced in which the premise and hypothesis	✓	0.24
For each ordered pair of languages L1 and L2, examples are constructed with the premise in L1 and the hypothesis in L2,	✓	0.28
This setup evaluates whether models can preserve semantic accuracy under mixed-lingual input—a common phenomenon in mult	✓	0.24
All experiments are executed using the Hugging Face Transformers library with a PyTorch backend.	✓	0.18
Inference is performed on A100 GPUs with device_map="auto" enabled for memory-efficient model parallelism.	✓	0.24
Generation uses greedy decoding with a maximum of 10 new tokens per prompt to produce concise outputs while limiting hal	✓	0.27
All models are evaluated in a zero-shot setting without task-specific fine-tuning.	✓	0.22
Six multilingual models are evaluated.	×	0.12
The benchmark tables show performance metrics for different language pairs in terms of entailment, contradiction, and ne	×	0.08
The performance metrics for French (fr), German (de), Swahili (sw), Hindi (hi), and Arabic (ar) are 0.912, 0.895, 0.841,	×	0.09

References

- http://arxiv.org/abs/2409.07054v2
- http://arxiv.org/abs/2102.12407v1
- http://arxiv.org/abs/2508.14735v1