NIASM Hybrid Approach Performance in Cross-Lingual Factual Consistency Benchmarks

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: How does the NIASM hybrid approach perform on cross-lingual factual consistency (F1 score) compared to monolingual fine-tuning in multilingual models like Bloom and Llama-2 on the XSUM and CNN/DM. In an era dominated by Large Language Models (LLMs), understanding their capabilities and limitations, especially in high-stakes fields like law, is crucial. While LLMs such as Meta's LLaMA, OpenAI's ChatGPT, Google's Gemini, DeepSeek, and other emerging models are increasingly. 7 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Evaluating the Limits of Large Language Models in Multilingual Legal Reasoning. Research question: How does the NIASM hybrid approach perform on cross-lingual factual consistency (F1 score) compared to monolingual fine-tuning in multilingual models like Bloom and Llama-2 on the XSUM and CNN/DM datasets?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.5/10.

3 Results

11 papers retrieved. 7 claims extracted; 0 independently verified. Quality review score: 3.5/10.

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
Accuracy is the proportion of correct predictions among all predictions.	×	0.04
Precision is the proportion of predicted positives that are actually correct.	×	0.00
Recall is the proportion of actual positives that are correctly predicted.	×	0.00
F1 Score is the harmonic mean of Precision and Recall.	×	0.02
Mean R-Precision (mRP) is the mean precision at rank k, where k equals the number of true labels.	×	0.03
ROUGE-1 captures lexical similarity through unigram overlap.	×	0.06
ROUGE-2 reflects fluency and phrase structure by measuring bigram overlap.	×	0.00

References

- http://arxiv.org/abs/2506.15415v1
- http://arxiv.org/abs/2212.10622v2
- http://arxiv.org/abs/2509.22472v1