Comparative Analysis of Domain-Adaptive Pre-training and Instruction Fine-tuning for Zero-Shot Cross-Lingual Legal Retrieval

Assignee Research

June 12, 2026

Abstract

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 integrated into legal workflows, their performance in multilingual, jurisdictionally diverse, and adversarial contexts remains insufficiently explored. This work evaluates LLaMA and Gemini on multilingual legal and non-legal benchmarks, and assesses their adversarial robustness in legal tasks through character and word-

1 Introduction

This paper examines: Evaluating the Limits of Large Language Models in Multilingual Legal Reasoning. Research question: How does the performance of domain-adaptive pre-training on LEMUR compare to instruction fine-tuning in zero-shot cross-lingual legal retrieval tasks when evaluated on adversarial perturbation benchmarks like MMLU or BIG-Bench?.

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 7.2/10.

3 Results

15 papers retrieved. 18 claims extracted; 13 independently verified. Quality review score: 7.2/10.

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 study employs classification metrics including Accuracy, Precision, Recall, F1, and mRP to assess correctness in str	✓	0.20
The study employs text generation metrics including ROUGE, BLEU, METEOR, and Cosine Similarity to evaluate the quality o	✓	0.22
The study employs robustness and reliability metrics including variance measures, consistency, entropy, Gini Index, and	✓	0.25
The study uses LLM-as-judge scores to capture quality judgments beyond surface similarity.	✓	0.16
Accuracy is defined as the proportion of correct predictions among all predictions, calculated as the sum of indicator f	✓	0.15
Precision is calculated as True Positives divided by the sum of True Positives and False Positives.	×	0.12
Recall is calculated as True Positives divided by the sum of True Positives and False Negatives.	×	0.12
F1 Score is the harmonic mean of Precision and Recall.	✓	0.24
Mean R-Precision (mRP) is the mean precision at rank k, where k equals the number of true labels.	✓	0.20
ROUGE-1 is an F1 score over unigram overlap that captures lexical similarity.	✓	0.22
ROUGE-2 is an F1 score over bigram overlap that reflects fluency and phrase structure.	✓	0.21
ROUGE-L is an F1 score based on the Longest Common Subsequence (LCS) over flat token sequences.	✓	0.25
ROUGE-L Sum is an F1 score based on LCS after sentence-level tokenization, optimized for evaluating multi-sentence summa	✓	0.43
BLEU measures the precision of n-gram overlaps between generated and reference texts.	✓	0.33
METEOR considers synonymy, stemming, and recall for n-gram overlaps.	✓	0.17
Cosine Similarity measures semantic similarity by calculating the cosine of the angle between vector representations of	×	0.12
In the LEXam–MC task, the Accuracy metric was recorded as 0.74.	×	0.02
In the LEXam–Open task, the LLM-Score was recorded as 0.51.	×	0.04

References

- http://arxiv.org/abs/2506.15415v1
- http://arxiv.org/abs/2509.22472v1
- http://arxiv.org/abs/2602.09570v1