Adversarial Robustness Degradation in Gemini Models from Multilingual Pretraining Imbalance

Assignee Research

June 13, 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: To what extent does multilingual pretraining data imbalance degrade the adversarial robustness of Gemini models when evaluating cross-jurisdictional legal constraints in non-English languages?.

2 Methodology

Systematic literature search across multiple databases yielded 14 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

14 papers retrieved. 18 claims extracted; 15 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 the indicat	✓	0.17
Precision is defined as the proportion of predicted positives that are actually correct, calculated as True Positives di	✓	0.18
Recall is defined as the proportion of actual positives that are correctly predicted, calculated as True Positives divid	✓	0.17
F1 Score is defined as the harmonic mean of Precision and Recall.	✓	0.21
Mean R-Precision (mRP) is defined as the mean precision at rank k, where k equals the number of true labels.	✓	0.19
ROUGE-1 is an F1 score over unigram overlap used to capture lexical similarity.	✓	0.15
ROUGE-2 is an F1 score over bigram overlap used to reflect fluency and phrase structure.	✓	0.17
ROUGE-L is an F1 score based on the Longest Common Subsequence (LCS) over flat token sequences to capture token-level or	✓	0.27
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.	✓	0.34
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 yielded a value of 0.74.	×	0.03
In the LEXam–Open task, the LLM Score metric yielded a value of 0.51.	×	0.04

References

- http://arxiv.org/abs/2605.29738v1
- http://arxiv.org/abs/2509.22472v1
- http://arxiv.org/abs/2310.00905v2