Impact of Syntactic Perturbation in Arabic Self-Invoking Code on Multilingual LLM Pass@k Metrics Relative to English Baselines

Assignee Research

June 11, 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 syntactic perturbation in Arabic self-invoking code tasks impact pass@k metrics of multilingual LLMs relative to English baselines on HumanEval Pro?.

2 Methodology

Systematic literature search across multiple databases yielded 8 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.6/10.

3 Results

8 papers retrieved. 20 claims extracted; 16 independently verified. Quality review score: 7.6/10.

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 evaluation employs a broad set of metrics spanning four categories: classification metrics, text generation metrics,	✓	0.19
Classification metrics include Accuracy, Precision, Recall, F1, and mRP.	✓	0.16
Text generation metrics include ROUGE, BLEU, METEOR, and Cosine Similarity.	✓	0.15
Robustness and reliability metrics include variance measures, consistency, entropy, Gini Index, and confidence margin.	✓	0.20
LLM-as-judge scores capture quality judgments beyond surface similarity.	✓	0.17
Accuracy is calculated as the proportion of correct predictions among all predictions.	×	0.12
Precision is calculated as the proportion of predicted positives that are actually correct.	✓	0.17
Recall is calculated as the proportion of actual positives that are correctly predicted.	✓	0.16
F1 Score is the harmonic mean of Precision and Recall.	✓	0.22
Mean R-Precision (mRP) is the mean precision at rank k, where k is the number of true labels.	✓	0.24
ROUGE-1 is the F1 score over unigram overlap, capturing lexical similarity.	✓	0.16
ROUGE-2 is the F1 score over bigram overlap, reflecting fluency and phrase structure.	✓	0.17
ROUGE-L is the F1 score based on the Longest Common Subsequence (LCS) over flat token sequences, capturing token-level o	✓	0.27
ROUGE-L Sum is the F1 score based on LCS after sentence-level tokenization, optimized for evaluating multi-sentence summ	✓	0.44
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 the semantic similarity between generated and reference texts using vector representations.	×	0.13
The benchmark table includes metrics such as Accuracy, Precision, Recall, F1 Score, and Mean R-Precision.	✓	0.17
The benchmark table includes 4values for LEXam–MC (Accuracy) and LEXam–Open (LLM Score).	×	0.04
The benchmark table includes F1, Precision, Recall, and mRP values with their respective standard deviations.	×	0.07

References

- http://arxiv.org/abs/2410.15308v2
- http://arxiv.org/abs/2412.21199v2
- http://arxiv.org/abs/2509.22472v1