Scaling Effects on Fine-Tuned Multilingual Models in Arabic-SQuAD Benchmarking

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does scaling the model size (e.g., increasing the number of parameters) of fine-tuned multilingual models affect their performance on Arabic-SQuAD compared to monolingual models in terms of exact. 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. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. 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 scaling the model size (e.g., increasing the number of parameters) of fine-tuned multilingual models affect their performance on Arabic-SQuAD compared to monolingual models in terms of exact match (EM) and F1 scores?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.3/10.

3 Results

13 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 3.3/10.

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 framework employs Accuracy, Precision, Recall, F1, and mRP metrics for structured prediction tasks.	×	0.05
Text generation quality is evaluated using ROUGE, BLEU, METEOR, and Cosine Similarity metrics.	×	0.05
Robustness and reliability are assessed using variance measures, consistency, entropy, Gini Index, and confidence margin	×	0.03
LLM-as-judge scores are used to capture quality judgments beyond surface similarity.	×	0.04
The LEXam–MC task achieved an Accuracy of 0.74.	×	0.04
The LEXam–Open task received an LLM Score of 0.51.	×	0.05
F1 scores for the model evaluations range between 0.18 and 0.23.	×	0.07
Precision values in the evaluation range from 0.14 to 0.22.	×	0.04
Recall values in the evaluation range from 0.15 to 0.21.	×	0.04
The mean R-Precision (mRP) values range between 0.15 and 0.24.	×	0.02

References

- http://arxiv.org/abs/2012.15613v2
- http://arxiv.org/abs/2509.22472v1
- http://arxiv.org/abs/2508.11281v3