Quantized LoRA Model Performance in Legal RAG with Context Window Variations

Assignee Research

June 12, 2026

Abstract

The evaluation bottleneck in recommendation systems has become particularly acute with the rise of Generative AI, where traditional metrics fall short of capturing nuanced quality dimensions that matter in specialized domains like legal research. Can we trust Large Language Models to serve as reliable judges of their own kind? This paper investigates LLM-as-a-Judge as a principled approach to evaluating Retrieval-Augmented Generation systems in legal contexts, where the stakes of recommendation quality are exceptionally high. We tackle two fundamental questions that determine practical viabi

1 Introduction

This paper examines: LLM-as-a-Judge: Rapid Evaluation of Legal Document Recommendation for Retrieval-Augmented Generation. Research question: What is the impact of context window size on the retrieval-augmented generation performance of quantized LoRA-adapted models when evaluating unfair terms in specialized legal domains?.

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 8.2/10.

3 Results

15 papers retrieved. 13 claims extracted; 12 independently verified. Quality review score: 8.2/10.

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 evaluated inter-rater reliability metrics including Cohen's Kappa, Krippendorff's Alpha, Spearman's rank corre	✓	0.28
The study did not evaluate Fleiss's Kappa or the Brennan-Prediger coefficient.	✓	0.22
The evaluation was conducted on two legal RAG systems operating over a comprehensive legal corpus at Bloomberg Law.	✓	0.22
Bloomberg Law evaluates thousands of legal query-document pairs monthly across multiple products.	✓	0.19
Each evaluated RAG system consists of a retrieval component and an answer generation component.	×	0.15
System A utilizes traditional BM25 retrieval combined with an open-source LLM summarizer applied to the top 5 retrieved	✓	0.17
System B incorporates improvements in the retrieval system and employs the proprietary GPT-4 model by OpenAI as the summ	✓	0.20
The evaluation framework targeted retrieval effectiveness through search relevancy assessment and generation quality thr	✓	0.24
Retrieval Augmented Generation (RAG) enhances LLM capabilities by integrating external knowledge sources.	✓	0.16
Traditional automated evaluation metrics such as ROUGE and BLEU depend on reference responses.	✓	0.21
Human evaluations are considered the gold standard for accuracy but are impractical at large scales due to time and expe	✓	0.19
GPT-4 has shown promise in achieving human-level agreement on certain tasks.	✓	0.24
Recent studies have identified challenges in using LLMs as judges, including cognitive biases, self-preference, and syst	✓	0.26

References

- http://arxiv.org/abs/2402.12317v2
- http://arxiv.org/abs/2510.06999v1
- http://arxiv.org/abs/2509.12382v1