

Retrieval-Augmented Generation Effects on F1 and Hallucination Rates in Legal Domain Models

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does retrieval-augmented generation impact the F1-score and hallucination rates of 7B vs. 70B models on legal domain benchmarks like LegalQA or Lexical when evaluated in a zero-shot setting. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Anomaly Detection: How to Artificially Increase your F1-Score with a Biased Evaluation Protocol. Research question: How does retrieval-augmented generation impact the F1-score and hallucination rates of 7B vs. 70B models on legal domain benchmarks like LegalQA or Lexical when evaluated in a zero-shot setting?.

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. 0 claims extracted; 0 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.

References

- <http://arxiv.org/abs/2311.08890v1>
- <http://arxiv.org/abs/2106.16020v1>
- <http://arxiv.org/abs/2506.06962v3>