Quantized Influence Measures vs Attention-Based Retrieval in Multi-File Code Generation

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How does quantized influence measure compare to standard attention-based retrieval in improving code generation accuracy on multi-file dependency benchmarks. This study presents an innovative enhancement to retrieval-augmented generation (RAG) systems by seamlessly integrating fine-tuned large language models (LLMs) with vector databases. This integration capitalizes on the combined strengths of structured data retrieval and the. 9 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: A Fine-tuning Enhanced RAG System with Quantized Influence Measure as AI Judge. Research question: How does quantized influence measure compare to standard attention-based retrieval in improving code generation accuracy on multi-file dependency benchmarks?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.5/10.

3 Results

16 papers retrieved. 9 claims extracted; 0 independently verified. Quality review score: 3.5/10.

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 RAG (L) model achieved an average score of 0.934 in chatbot creation tasks.	×	0.05
The RAG (L) model outperformed the Davinci002 and Llama2 models in the experiments.	×	0.03
The RAG (L) model had the lowest standard deviation of 0.016 among the RAG models.	×	0.03
The RAG (L) model achieved peak scores as high as 0.960 in specific trials.	×	0.03
RAG (1E) and RAG (3E) had average scores of 0.925 and 0.921 respectively.	×	0.02
Foundation Model (FM) achieved a robust average of 0.848 in performance metrics.	×	0.04
Langchain + SerpAPI (L+S) approach had a lower average performance score of 0.495.	×	0.04
Fine-tuning pre-trained Large Language Models using customized data from internet scraping has been reported to improve	×	0.10
QLoRA is a fine-tuning method enabling a 65B model to be fine-tuned on a 48G GPU.	×	0.12

References

- http://arxiv.org/abs/2402.12317v2
- http://arxiv.org/abs/2402.17081v1
- http://arxiv.org/abs/2511.20417v2