Multi-Needle Retrieval F1 Variability on LongBench for Python Contexts: CAKE Scaling from 7B to 13B Parameters

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 9 peer-reviewed papers addressing the following research question: What is the variation in Multi-needle retrieval F1 scores on LongBench for Python contexts when scaling CAKE from 7B to 13B parameter models versus static baselines. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.8/10. 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: What is the variation in Multi-needle retrieval F1 scores on LongBench for Python contexts when scaling CAKE from 7B to 13B parameter models versus static baselines?.

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.8/10.

3 Results

9 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 5.8/10.

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/2505.21439v1
- http://arxiv.org/abs/2106.16020v1
- http://arxiv.org/abs/1811.08772v1