Tree of Reviews vs. Chain-Based Retrieval Robustness in Llama-3-8B-128K on Adversarial SQuAD

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 9 peer-reviewed papers addressing the following research question: How does the robustness of Tree of Reviews retrieval compare to chain-based retrieval for Llama-3-8B-128K when evaluated on adversarial or noisy versions of SQuAD using different embedding models. Dense retrieval is becoming one of the standard approaches for document and passage ranking. The dual-encoder architecture is widely adopted for scoring question-passage pairs due to its efficiency and high performance. 8 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Analysing the Robustness of Dual Encoders for Dense Retrieval Against Misspellings. Research question: How does the robustness of Tree of Reviews retrieval compare to chain-based retrieval for Llama-3-8B-128K when evaluated on adversarial or noisy versions of SQuAD using different embedding models (e.g., Sentence-T5 vs. MPNet)?.

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 3.8/10.

3 Results

9 papers retrieved. 8 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
On clean questions, data augmentation as well as the proposed approaches (contrastive learning and data augmentation com	×	0.11
All approaches for robustifying DR perform significantly better compared to the original DR on questions with typos.	×	0.07
The data augmentation combined with contrastive learning approach holds the best performance on questions with typos.	×	0.09
The effectiveness of the methods varies across the three settings: random typos, typos in non-stopwords, and typos in di	×	0.03
Robustness deteriorates when typos do not appear randomly, with the most significant losses occurring when typos appear	×	0.06
The data augmentation combined with contrastive learning approach remains the best performing one across all settings.	×	0.06
There is a strong connection between the frequency of the typoed words and the retrieval performance.	×	0.05
As the frequency of the typoed words decreases, the performance drops significantly.	×	0.03

References

- http://arxiv.org/abs/2205.02303v1
- http://arxiv.org/abs/2404.14464v1
- http://arxiv.org/abs/1811.08772v1