Scaling Retriever Portfolios and Robustness of RAG Systems Under Adversarial Queries

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the effect of scaling the number of retrievers in a portfolio on the robustness of RAG systems against adversarial query perturbations across different domains (e.g., News, Science,. 16 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Towards More Robust Retrieval-Augmented Generation: Evaluating RAG Under Adversarial Poisoning Attacks. Research question: What is the effect of scaling the number of retrievers in a portfolio on the robustness of RAG systems against adversarial query perturbations across different domains (e.g., News, Science, Literature) on the AmbiEval benchmark?.

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 2.7/10.

3 Results

14 papers retrieved. 16 claims extracted; 0 independently verified. Quality review score: 2.7/10.

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 primary evaluation measure is F1 score, which balances precision and recall to provide a comprehensive measure of pe	×	0.04
Additional measures such as Precision, Recall, and Abstention Rate are included in Appendix E.1 for reference.	×	0.02
The retrieval component includes adversarial context, untouched context, guiding context, and a Non-RAG setting.	×	0.09
The generation component experiments with two types of prompts: Neutral and Skeptical.	×	0.05
Neutral prompts do not explicitly steer the model's behavior.	×	0.03
Skeptical prompts explicitly steer the model's behavior to critically assess retrieved content.	×	0.03
The number of episodes in Chicago Fire Season 4 is 24.	×	0.02
GPT-3.5 Neutral has an F1 score of 50.00.	×	0.02
GPT-3.5 Skeptical has an F1 score of 37.00.	×	0.02
GPT-3.5 Faithful has an F1 score of 32.98.	×	0.02
GPT-4 Neutral has an F1 score of 38.37.	×	0.02
GPT-4 Skeptical has an F1 score of 79.00.	×	0.02
GPT-4 Faithful has an F1 score of 5.00.	×	0.02
GPT-4o Neutral has an F1 score of 61.70.	×	0.02
GPT-4o Skeptical has an F1 score of 86.00.	×	0.02
GPT-4o Faithful has an F1 score of 26.04.	×	0.02

References

- http://arxiv.org/abs/2412.16708v2
- http://arxiv.org/abs/2605.31176v1

- http://arxiv.org/abs/2510.25518v1