Impact Of Domain-Specific Fine-Tuning (E.G., Legal Domain) On The Robustness Of Rag Models Against Adversarial Attacks

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What is the impact of domain-specific fine-tuning (e.g., legal domain) on the robustness of RAG models against adversarial attacks compared to general-domain fine-tuning, as measured by Recall@1000. Retrieval Augment Generation (RAG) is a recent advancement in Open-Domain Question Answering (ODQA). RAG has only been trained and explored with a Wikipedia-based external knowledge base and is not optimized for use in other specialized domains such as healthcare and news. 17 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering. Research question: What is the impact of domain-specific fine-tuning (e.g., legal domain) on the robustness of RAG models against adversarial attacks compared to general-domain fine-tuning, as measured by Recall@1000 and answer accuracy on JURIS-AQA?.

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 2.5/10.

3 Results

16 papers retrieved. 17 claims extracted; 0 independently verified. Quality review score: 2.5/10.

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
Most ODQA datasets like Natural Questions, TriviaQA, WebQuestions, and CuratedTrec are answered with Wikipedia-based kno	×	0.06
Neural retrievers like DPR are already trained with Wikipedia-based datasets.	×	0.04
Three domain-specific datasets were selected for the experiment: COVID-19 QA, News QA, and Conversation QA.	×	0.10
The COVID-19 QA domain knowledge base was created with 250,000 100-word passages extracted from 5,000 full-text scientif	×	0.09
RAG-end2end outperforms RAG-original even in other Wikipedia-based datasets.	×	0.14
RAG-end2end updates the context encoder and embeddings during the training process.	×	0.09
The retriever component is crucial in domain-specific question answering.	×	0.11
Future research directions include exploring RAG-end2end on tasks like Fact Checking, Summarisation, and conversational	×	0.07
Exploring generative capabilities with qualitative metrics could improve factual consistency and reduce hallucinations i	×	0.02
Updating the retriever and document embeddings during the training phase could improve factual consistency and reduce ha	×	0.05
The statement reconstruction signal acts as a good auxiliary signal for improving the overall performance of RAG models.	×	0.11
The most common signs and symptoms on admission included fever and cough.	×	0.01
32% of children had complaints of difficulty in respiration.	×	0.00
Other symptoms observed were myalgia, headache, and vomiting.	×	0.00
On examination, 66% of cases had crepitations and 42% had wheezing.	×	0.03
Hypoxemia was observed in 31% of cases at admission.	×	0.03
The Kiwi girl that Darren T Maloney spoke to on the phone commutes from a location that involves a border crossing with	×	0.01

References

- http://arxiv.org/abs/2210.02627v1
- http://arxiv.org/abs/2505.03970v1
- http://arxiv.org/abs/2307.02055v1