LongRAG Fine-Tuning of Llama-3-8B Enhances Cross-Domain Long-Context QA Generalization

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: Does fine-tuning Llama-3-8B with LongRAG objectives improve generalization scores on cross-domain long-context QA tasks relative to domain-specific fine-tuning alone. Large Language Models (LLMs) have been widely applied in various professional fields. By fine-tuning the models using domain specific question and answer datasets, the professional domain knowledge and Q\&A abilities of these models have significantly improved, for example. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Fine-Tuning Medical Language Models for Enhanced Long-Contextual Understanding and Domain Expertise. Research question: Does fine-tuning Llama-3-8B with LongRAG objectives improve generalization scores on cross-domain long-context QA tasks relative to domain-specific fine-tuning alone?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.0/10.

3 Results

12 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 4.0/10.

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 context ability and instruction following ability of the general LLMs are both good, with the majority of testing ac	×	0.06
Models trained with multiple rounds of dialogue or long contextual dialogue can achieve excellent results in the test.	×	0.07
Medical LLMs from the Medical Benchmark in Chinese (CMB) achieved excellent results in the CMB professional medical prof	×	0.06
The results of medical LLM (especially models with better medical capabilities) were relatively unsatisfactory compared	×	0.14
HuatuoGPT-II has excellent medical abilities among these tested models, but the average accuracy rate of the open book e	×	0.08
PULSE often relies on multiple rounds of dialogue to obtain more information for more accurate diagnosis.	×	0.03
Data in the medical field is usually more professional, covering relatively narrow content and form, while general quest	×	0.09
IvyGPT proposed a new improvement on fine-tuning training data by mixing real question answering data with generated dat	×	0.09
HuatuoGPT trained a reward model brought by the two types of data, following the approach of RLAIF (reinforcement learni	×	0.07
HuatuoGPT-II proposed a unified domain adaptation protocol that combines continuous pre-training and fine-tuning stages	×	0.07
LLMs can be used to develop intelligent Q&A systems and chatbots to help patients answer common health questions, book o	×	0.04
An evaluation method was designed to test the model's contextual capabilities and instruction following capabilities.	×	0.05
Chinese medical exams, including physician exams, nursing exams, pharmacist exams, medical technology exams, professiona	×	0.10

References

- http://arxiv.org/abs/2410.18050v2
- http://arxiv.org/abs/2110.06500v2
- http://arxiv.org/abs/2407.11536v1