DeepSeek R1 Performance on MultiMedQA Under Controlled Training Set Contamination

Assignee Research

June 2, 2026

Abstract

This report synthesises findings from 1 peer-reviewed paper addressing the following research question: How does the performance of Deepseek R1 on MultiMedQA vary when fine-tuned on datasets with controlled levels of training set contamination across Bloom's Taxonomy levels. Public health reasoning requires population level inference grounded in scientific evidence, expert consensus, and safety constraints. However, it remains underexplored as a structured machine learning problem with limited supervised signals and benchmarks. 4 claims were extracted from source literature; 4 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: From Knowledge to Inference: Formalizing Specialized Public Health Reasoning on GlobalHealthAtlas. Research question: How does the performance of Deepseek R1 on MultiMedQA vary when fine-tuned on datasets with controlled levels of training set contamination across Bloom's Taxonomy levels?.

2 Methodology

Systematic literature search across multiple databases yielded 1 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.5/10.

3 Results

1 papers retrieved. 4 claims extracted; 4 independently verified. Quality review score: 8.5/10.

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
GlobalHealthAtlas is a large scale multilingual dataset of 280,210 instances spanning 15 public health domains and 17 la	✓	0.36
The dataset construction and quality control pipeline includes retrieval, deduplication, evidence grounding checks, and	✓	0.23
A domain aligned evaluator is distilled from high confidence judgments of diverse LLMs to assess outputs along six dimen	✓	0.37
The project codebase, evaluator, and model are publicly released at https://github.com/Jan8217/GlobalHealthAtlas, https:	✓	0.33

References

- https://openalex.org/W7127542513