Reverse Operation Data Augmentation and Sample Efficiency in Fine-Tuned Language Models

Assignee Research

June 5, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the impact of reverse operation data augmentation on the sample efficiency of language models when fine-tuned on limited MMLU STEM subsets. 11 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: Linearization Explains Fine-Tuning in Large Language Models. Research question: What is the impact of reverse operation data augmentation on the sample efficiency of language models when fine-tuned on limited MMLU STEM subsets?.

2 Methodology

Systematic literature search across multiple databases yielded 13 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

13 papers retrieved. 11 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
LoRA was implemented on RoBERTa base and evaluated on the GLUE benchmark, IMDb, and Yelp datasets.	×	0.03
RoBERTa base has 125M parameters.	×	0.05
RoBERTa base has proven to be one of the most powerful models for various NLP tasks, including text classification, ques	×	0.09
The GLUE benchmark is a collection of diverse tasks that test a model's natural language understanding abilities.	×	0.05
The tasks included in the experiments are linguistic acceptability judgment (CoLA) and sentiment analysis (SST-2).	×	0.03
The IMDb dataset is a large dataset for binary sentiment classification, containing 50k highly popular movie reviews fro	×	0.02
The Yelp dataset contains customer reviews from Yelp, a popular platform for crowd-sourced reviews about businesses, pri	×	0.01
The Yelp dataset originally contains reviews with ratings from 1 to 5.	×	0.01
To convert the Yelp dataset into a binary classification task, reviews with ratings less than 3 are considered label 0 (	×	0.02
Table 9 in Appendix N shows specific hyperparameters for RoBERTa base across various benchmarks, including GLUE tasks (C	×	0.03
For all experiments, LoRA was used on the RoBERTa-base model from the Hugging Face transformers library.	×	0.04

References

- http://arxiv.org/abs/2602.08239v1
- http://arxiv.org/abs/2110.06500v2
- http://arxiv.org/abs/1910.03560v2