Instruction Tuning Datasets and Out-of-Distribution Accuracy in Long-Context Policy Learning

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 5 peer-reviewed papers addressing the following research question: How does instruction tuning with Vicuna-style conversational data versus Llama-3-style curated datasets impact out-of-distribution accuracy on the MLNeedle benchmark for long-context policy learning. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: FinGPT: Instruction Tuning Benchmark for Open-Source Large Language Models in Financial Datasets. Research question: How does instruction tuning with Vicuna-style conversational data versus Llama-3-style curated datasets impact out-of-distribution accuracy on the MLNeedle benchmark for long-context policy learning?.

2 Methodology

Systematic literature search across multiple databases yielded 5 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.3/10.

3 Results

5 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 3.3/10.

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
In the NER task, Llama2's F1-score increased from 0.637 in the Task-Specific phase to 0.678 in the Multi-Task phase.	×	0.02
In the NER task, MPT achieved a performance gain of 6.7% when moving from Task-Specific to Multi-Task tuning.	×	0.04
In the Headline Classification (HC) task, all tested models (Llama2, Falcon, MPT, BLOOM, ChatGLM2, Qwen) experienced a d	×	0.02
In the Relation Extraction (RE) task, MPT showed the highest percentage improvement (+35.8%) among all models when trans	×	0.03
BLOOM outperformed Falcon in the Relation Extraction (RE) Multi-Task phase with an F1-score of 0.697 compared to Falcon'	×	0.01
In zero-shot Sentiment Analysis on the FPB dataset, ChatGLM2 achieved the highest F1-score of 0.803 among the tested mod	×	0.06
The FinRED dataset contains 6,768 samples for Relation Extraction and 9,657 samples for Relation Extraction (CLS).	×	0.01
Llama2 achieved an average ranking of 2.0 across SA, NER, HC, and RE tasks, ranking first overall among the tested model	×	0.02
In Multi-Task Sentiment Analysis, BLOOM experienced the largest average performance decline (-4.7%) compared to its Task	×	0.03
Qwen was the only model to show an average performance gain (+1.1%) in Sentiment Analysis when moving from Task-Specific	×	0.05

References

- http://arxiv.org/abs/2310.04793v2
- http://arxiv.org/abs/2402.11690v1
- http://arxiv.org/abs/2312.10793v3