Alignment Tuning in Llama3 and DeepSeek-R1 for Adversarial Code Repair Robustness

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 9 peer-reviewed papers addressing the following research question: To what extent does alignment tuning in Llama3 and Deepseek R1 mitigate helpfulness degradation across diverse adversarial taxonomies in automated code repair tasks. 8 claims were extracted from source literature; 7 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.1/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Go Gentle Into the Final Dance: A Benchmark for Evaluating LLMs in Telecom. Research question: To what extent does alignment tuning in Llama3 and Deepseek R1 mitigate helpfulness degradation across diverse adversarial taxonomies in automated code repair tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.1/10.

3 Results

9 papers retrieved. 8 claims extracted; 7 independently verified. Quality review score: 8.1/10.

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 paper introduces Last Dance of Telecommunications (LDOT) as a comprehensive benchmark suite for telecom-related task	✓	0.25
LDOT encompasses problem categories including conceptual telecom questions, mathematical and logical reasoning problems,	✓	0.33
The study assesses state-of-the-art (SOTA) LLMs, including both closed-source and open-source models, using the LDOT ben	✓	0.19
General-purpose LLMs exhibit strong performance on basic telecom knowledge questions within the LDOT benchmark.	✓	0.26
General-purpose LLMs struggle with reasoning-intensive wireless problems in the LDOT benchmark.	✓	0.23
Certain multi-step optimization and planning tasks in LDOT remain unsolved by even the best evaluated models.	✓	0.25
Existing saturated benchmarks do not reveal the performance gaps exposed by LDOT.	×	0.14
The paper provides a failure analysis to distinguish whether model limitations arise from insufficient telecom-specific	✓	0.23

References

- https://doi.org/10.3390/bdcc9120320
- https://doi.org/10.1109/jsac.2025.3641905
- https://openalex.org/W7160458426