Multimodal Context Scaling and DeepSeek-R1 Robustness in Iterative Code Repair

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does the scaling of multimodal context (varying the ratio of text to diagram information) affect the robustness of DeepSeek-R1's iterative code repair performance across different programming. Code repair is a fundamental task in software development, facilitating efficient bug resolution and software maintenance. Although large language models (LLMs) have demonstrated considerable potential in automated code repair, their ability to comprehend and leverage diverse. 13 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: FeedbackEval: A Benchmark for Evaluating Large Language Models in Feedback-Driven Code Repair Tasks. Research question: How does the scaling of multimodal context (varying the ratio of text to diagram information) affect the robustness of DeepSeek-R1's iterative code repair performance across different programming languages in FeedbackEval?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.8/10.

3 Results

15 papers retrieved. 13 claims extracted; 2 independently verified. Quality review score: 4.8/10.

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
FeedbackEval is a new benchmark for evaluating large language models in feedback-driven code repair tasks.	✓	0.27
Each benchmark task in FeedbackEval consists of two main components: a core program repair instance and a full set of fe	×	0.06
The core components of each task include erroneous code, docstring, context, reference implementation, and test cases.	×	0.04
FeedbackEval provides multiple feedback modalities for each task, including compiler feedback, test feedback, minimal fe	×	0.13
Compiler feedback is structured and indicates syntax errors, style violations, and potential bugs.	×	0.05
Test feedback is structured and derived from the task's test cases, identifying failing tests and expected outcomes.	×	0.05
Minimal feedback is a fixed, concise message providing no detailed guidance.	×	0.03
LLM-skilled feedback is unstructured, natural-language suggestions resembling advice from a competent but non-expert dev	×	0.08
The benchmark includes a table comparing feedback types, their sources, forms, acquisition difficulty, and intended leve	×	0.04
The benchmark includes a table showing the performance of different models (GPT-4o, Claude-3.5, Deepseek-R1, GLM-4, Qwen	✓	0.19
The benchmark includes a table showing the performance of different models on different benchmarks (HumanEval, CoderEval	×	0.11
The benchmark includes a table showing the performance of different feedback types on different benchmarks (HumanEval, C	×	0.12
The benchmark includes a table showing the performance of different models (GPT, Claude, DeepSeek, GLM, Qwen) on differe	×	0.07

References

- http://arxiv.org/abs/2506.00404v1
- http://arxiv.org/abs/2505.21514v1
- http://arxiv.org/abs/2504.06939v2