From Scores to Evidence: A Framework for Epistemically Valid Evaluation of LLM Coding Agents

Anonymous Authors

September 18, 2026

Abstract

Large language models (LLMs) are increasingly evaluated as autonomous coding agents, yet existing benchmarks typically reduce performance to task-level success metrics that obscure why agents succeed, fail, or change behaviour across interaction stages.

We introduce EDASE-AIO, an evidence-based framework for analysing coding agents through claim-level trajectories rather than binary task outcomes. The framework models agent interactions as structured evidence chains spanning initial attempts, feedback-driven revisions, independent validation, and final outcomes.

We apply the framework to a controlled study of coding agents across 180 claim trajectories, 900 claim-level evaluation cells, and 66 replicated tasks spanning multiple programming languages and task categories. Through blinded dual-pass annotation, deterministic evidence freezing, and independent adjudication, we identify behavioural patterns including persistent degradation, stable partial alignment, soft improvement, and recovery trajectories.

Our framework treats evaluation evidence as a first-class object and enables more reliable conclusions about agent capabilities, limitations, and behavioural change under realistic software engineering workflows.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 Introduction

2 Related Work

3 Evidence-Based Evaluation Framework

4 Experimental Setup

5 Results

6 Discussion

7 Limitations

8 Conclusion

References