Explaining GitHub Actions Failures with Large Language Models: Challenges, Insights, and Limitations – Summary

Pablo Valenzuela-Toledo [1,2], Chuyue Wu [1], Sandro Hernández [1], Alexander Boll [1], Roman Machacek [1], Sebastiano Panichella [1], and Timo Kehrer [1]

Abstract:

We report about recent research on the use of Large Language Models (LLMs) to support the diagnosis of failures in Continuous Integration and Continuous Deployment (CI/CD) workflows, originally published at the 33rd IEEE International Conference on Program Comprehension (ICPC 2025) [Va25].

GitHub Actions (GA) has become the *de facto* tool that developers use to automate software workflows, seamlessly building, testing, and deploying code. Yet when GA fails, it disrupts development, causing delays and driving up costs. Diagnosing failures becomes especially challenging because error logs are often long, complex and unstructured. Given these difficulties, this study explores the potential of LLMs to generate correct, clear, concise, and actionable contextual descriptions (or summaries) for GA failures, focusing on developers' perceptions of their feasibility and usefulness. Our results show that over 80% of developers rated LLM explanations positively in terms of correctness for simpler/small logs. Overall, our findings suggest that LLMs can feasibly assist developers in understanding common GA errors, thus, potentially reducing manual analysis. However, we also found that improved reasoning abilities are needed to support more complex CI/CD scenarios. For instance, less experienced developers tend to be more positive on the described context, while seasoned developers prefer concise summaries. Our work offers key insights for researchers enhancing LLM reasoning, particularly in adapting explanations to user expertise.

Keywords: CI/CD, GitHub Actions, Large Language Models, Run Failure Explanation

1 Summary

This paper investigates how Large Language Models (LLMs) can support developers in understanding and diagnosing failures occurring in GitHub Actions (GA) workflows. GA has become a central automation service for continuous integration and deployment, yet when workflows fail, developers must manually inspect long and unstructured logs. These

[1] Software Engineering Group, University of Bern, Switzerland,
pablo.valenzuela@unibe.ch, https://orcid.org/0000-0001-6349-4296;
chuyue.wu@students.unibe.ch, https://orcid.org/0009-0004-4952-817X;
sandro.hernandez@unibe.ch, https://orcid.org/0009-0006-5532-1700;
alexander.boll@unibe.ch, https://orcid.org/0000-0002-9881-9748;
roman.machacek@unibe.ch, https://orcid.org/0009-0007-9976-4420;
sebastiano.panichella@unibe.ch, https://orcid.org/0000-0003-4120-626X;
timo.kehrer@unibe.ch, https://orcid.org/0000-0002-2582-5557

[2] Universidad de La Frontera, Temuco, Chile,
pablo.valenzuela@unibe.ch, https://orcid.org/0000-0001-6349-4296

logs often contain cryptic error codes and fragmented contextual information, making diagnosis slow and error-prone. The study explores whether LLMs can generate concise and meaningful explanations that summarize the causes of failures and assist developers in identifying corrective actions.

To evaluate this idea, we conducted a mixed-methods feasibility study using LogExp, a custom web platform that presents GA failure logs alongside LLM-generated explanations. The study surveyed 31 developers experienced in maintaining GA workflows, who rated the explanations according to four key attributes: correctness, conciseness, clarity, and actionability. Ten Likert-scale statements and two open-ended questions captured both quantitative and qualitative insights about the usefulness of these explanations in real diagnostic scenarios.

The results reveal that developers perceive LLM-generated explanations as largely accurate and clear. More than eighty percent of participants agreed that the explanations correctly reflected the details and context of the run failures and were easy to understand. The analysis also shows that LLMs perform better on shorter and more structured logs, while their reasoning capacity decreases for complex or verbose cases that require deeper contextual understanding. Regarding actionability, five recurring attributes were identified as essential for effective explanations: clarity, actionable guidance, specificity, contextual relevance, and conciseness. Developers particularly valued explanations that provided precise steps directly applicable to resolving the observed errors.

In our paper, we contribute an empirical foundation for understanding how LLMs explain failures in CI/CD workflows and provide evidence that such explanations can support developers in diagnosing errors efficiently. We define a set of quality attributes that describe the usefulness of explanations, validate them through a mixed-methods study, and identify the conditions under which LLMs perform best. We present a web-based evaluation platform for controlled assessment of explanations, demonstrate the feasibility of using LLMs for failure comprehension in GitHub Actions, and outline directions for improving their reasoning and adaptability in complex diagnostic contexts.

2 Data Availability

The original publication is publicly accessible with the DOI 10.1109/ICPC66645.2025.00037. A preprint is also available online at https://arxiv.org/abs/2501.16495. Our artifact is available on Zenodo (10.5281/zenodo.14750197).

References

[Va25] Valenzuela-Toledo, P. et al.: Explaining GitHub Actions Failures with Large Language Models: Challenges, Insights, and Limitations. In: Proc. IEEE/ACM 33rd Int'l Conf. on Program Comprehension (ICPC). IEEE, Ottawa, Canada, pp. 286–297, 2025.