

LLM-Generated CI/CD YAML Workflows Evaluation

Experiment Design & Pipeline Overview

Dataset Creation

MSR'24 Workflow Histories Dataset

A dataset of GitHub Actions workflow histories

*repositories.csv
workflows.csv
workflows.tar*

Collated Dataframe

Full dataframe merging all repositories, workflows and available metadata from MSR'24

Filter

★ **Top-500 Starred
Repositories**



Workflow Dataset

Output: 500 repos, 3,887 workflows

*actionlint v1.7.9 Gold
Pass/Fail + Diagnostics*

RQ1 - Gold Baseline Validation

One-Shot Generation Evaluation

Reverse Prompting (Prompt Synthesis)

Input

Gold YAML (+ repo metadata)

Gemini 2.5

Flash/Pro (Thinking)

Generated Prompt

*NL Prompt Per Workflow
(1:1)*

Workflow Generation

Gemini 2.5

Flash/Pro (Thinking)

Output

*Generated YAML
(Pro/Flash)*

Validate Generated Workflows

Actionlint Diagnostics

*Run actionlint on Pro/Flash
outputs → diagnostics*

RQ2 - Generated Workflow Evaluation

Iterative Repair Loop

Repair Loop (K=5)

Input

*Failing generated YAML + NL prompt
+ actionlint diagnostics*

Gemini 2.5

Flash/Pro (Thinking)

Actionlint Diagnostics

*Run actionlint on Pro/Flash
outputs → diagnostics*

Output

*Repaired YAML candidates
per iteration*

*Error
Feedback Loop*

*Break on
success*

RQ3 - Iterations-to-Fix Analysis