Improving the Coverage of GPT for Automated Feedback on High School Programming Assignments

Shubham Sahai, Umair Z. Ahmed, and **Ben Leong**
National University of Singapore

Shubham Sahai

Publicly Released dataset from **NUS High School**

69 Assignments

366 Buggy Solutions

5928 Correct Solutions

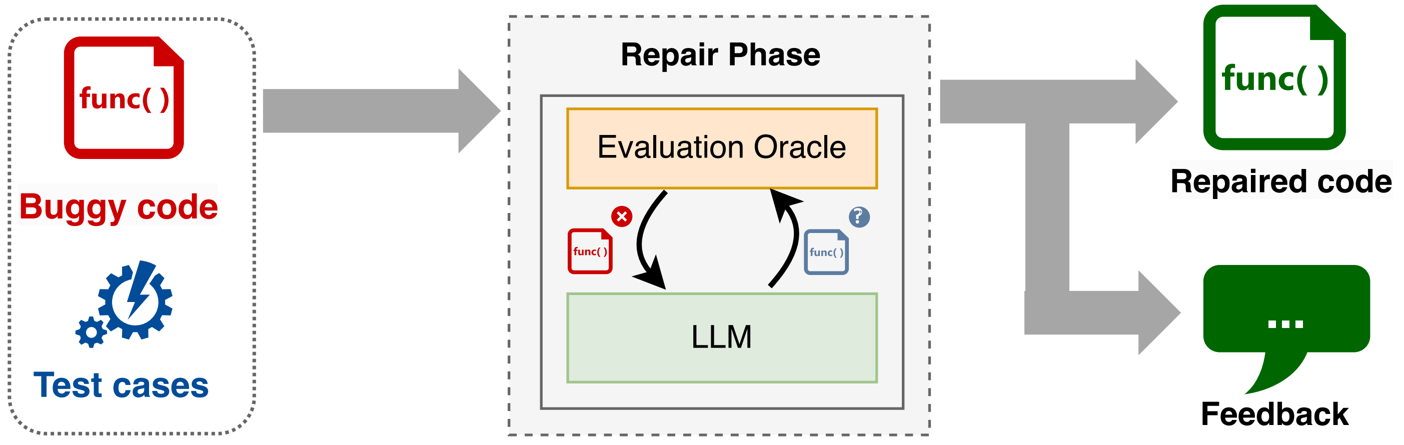

Figure 1: *Proposed architecture. LLM generates a repair and feedback which is validated by an evaluation oracle against testcases.*

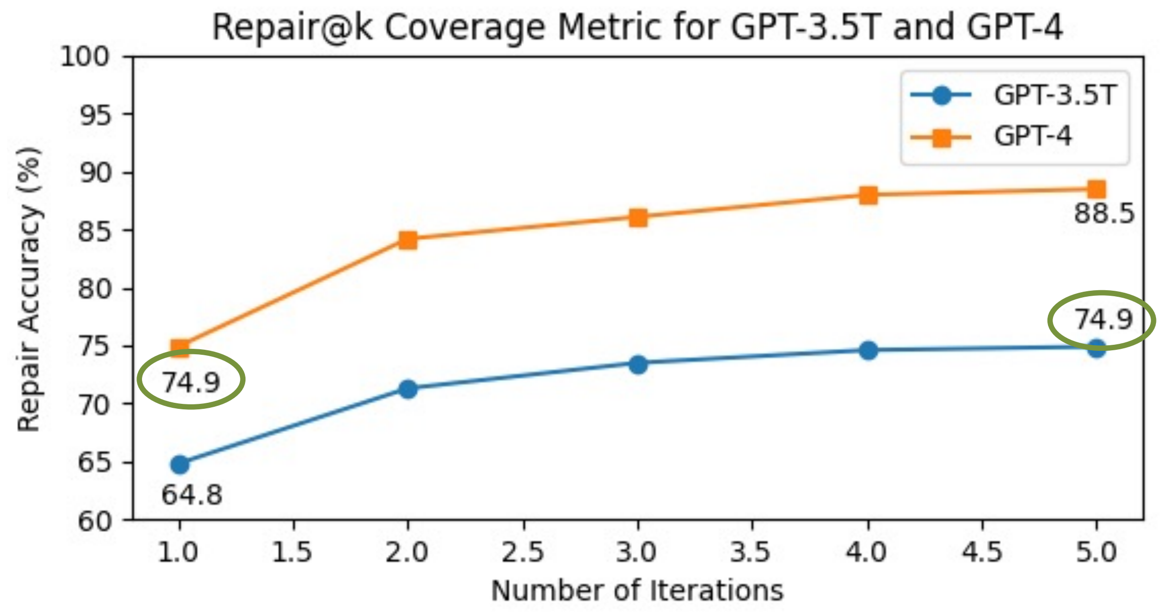

Figure 2: *Comparing repair accuracy of GPT-3.5T and GPT-4 after k interactive iterations*

To assess the reliability, we manually categorized GPT generated feedback into following 5 categories:

Category	Definition
True Positive	Valid feedback is generated
False Negative	Failed to detect the error and generate feedback
False Positive (Extra)	Unnecessary feedback, e.g., Optimization
False Positive (Invalid)	Incorrect feedback generated
False Positive (Hallucination)	Fabricated feedback (unrelated to the code) is generated.

	Precision	**Recall**	**False Positives**	
	Reliability	Coverage	Invalid	Hallucination
GPT 3.5T	51.2%	52.7%	15.0%	18.0%
GPT 4	72.0%	84.0%	9.0%	4.1%

Table 1: *Feedback quality of GPT-3.5T and GPT-4 LLMs, based on manual assessment by authors.*