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

NUS National University of Singapore

AICET

Introduction

Introductory programming students typically struggle with errors, primarily due to inadequate real-time support during assignments. In this work, we investigate the potential of Large Language Models (LLMs) like **GPT-3.5T** and **GPT-4** for generating correct repair and valid feedback on incorrect submissions.

Specifically, we investigate:

a) Coverage: What is the repair coverage of GPTs, and can we improve it through multiple interactions with the model?

b) Reliability: How trustworthy is the feedback generated by GPTs?

Repair Coverage

On our dataset of **366** incorrect and **5928** correct student submissions across **69** high-school programming assignments, GPT-3.5T could repair **64.8%** incorrect submissions successfully while GPT-4 achieved **74.9%** repair coverage.

Our key insight is that despite the initial repair failure, a conversational interaction with the LLM, paired with an evaluation oracle that reveals failing testcases in each iteration, can significantly improve the repair coverage.

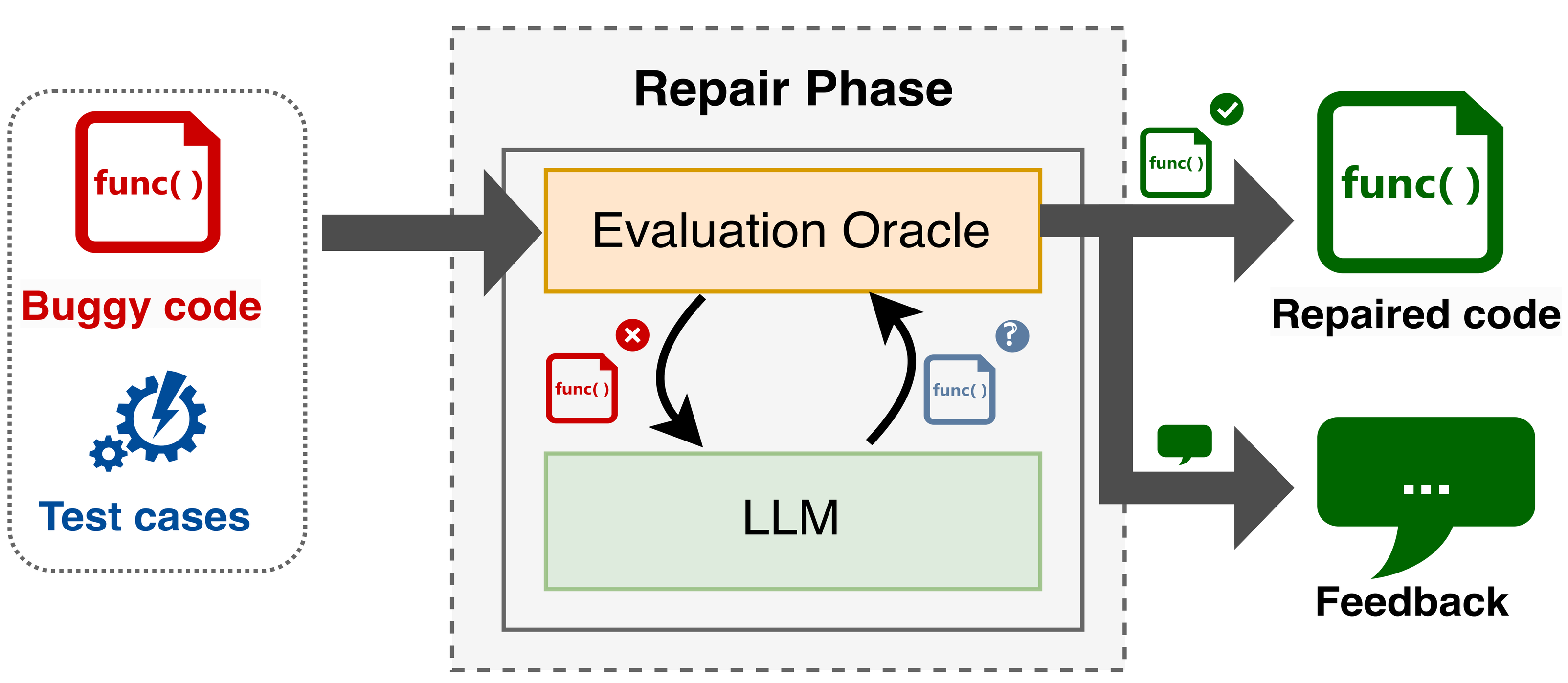

Figure 1: *Proposed architecture. LLM generated repair is validated by an evaluation oracle against testcases, prior to releasing feedback for students.*

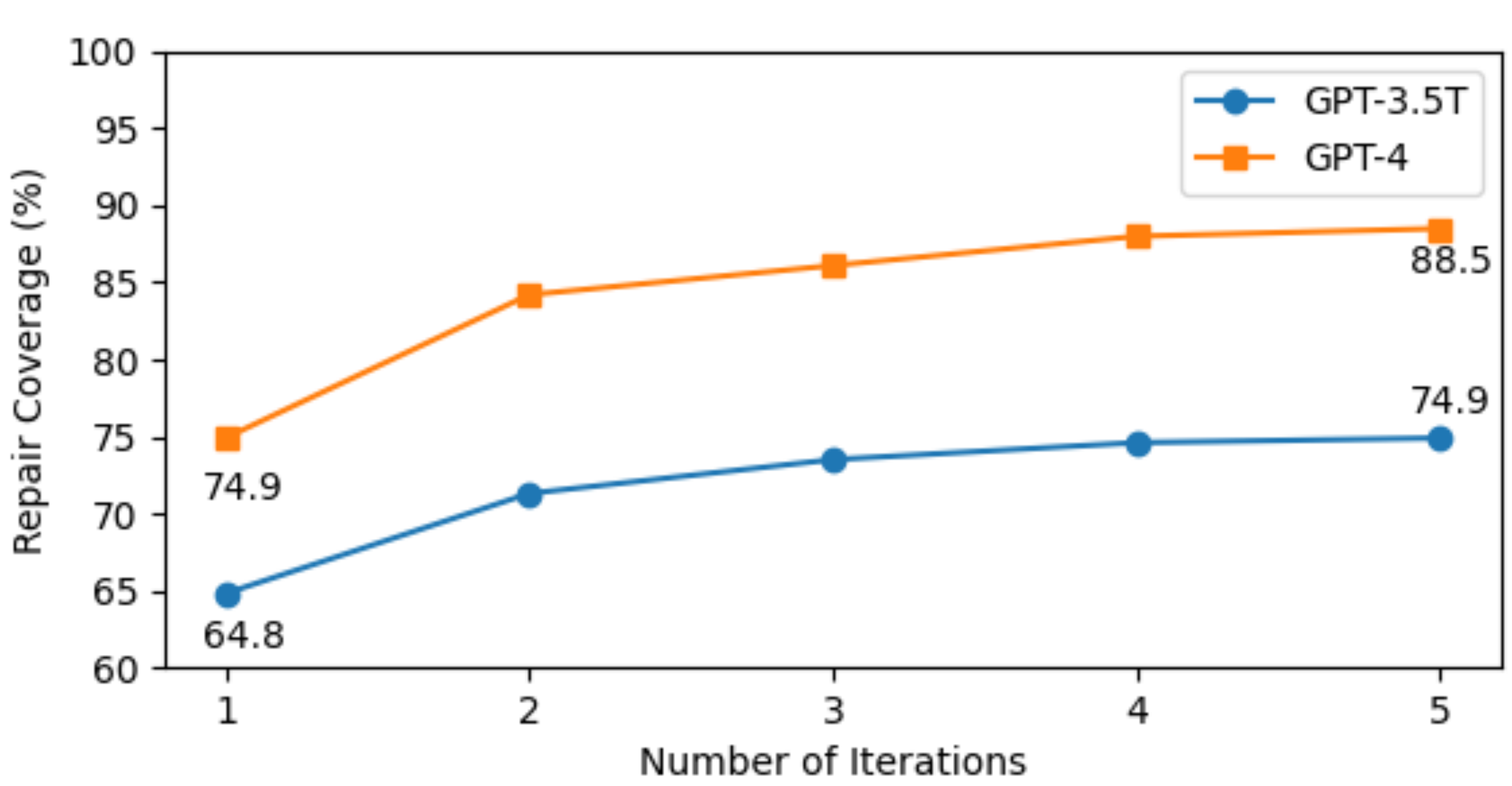

Figure 2: *Repair coverage of GPT-3.5T and GPT-4 after **k** iterations*

We found that within **5** iterations, repair coverage of

- GPT-3.5T improved from **64.8%** to **74.9%**
- GPT-4 saw an improvement from **74.9%** to **88.5%**.

In other words, the repair coverage of a weaker GPT-3.5T model could match that of state-of-art GPT-4 which is **20x** more expensive, with multiple conversational iterations.

Acknowledgement:
This research/project is supported by the National Research Foundation Singapore under the AI Singapore Programme (AISG Award No: AISG2-TC-2023-009-AICET).

Feedback Reliability

To the best of our knowledge, we are the first to manually evaluate the correctness of LLM-generated feedback on such a large dataset of more than 100 submissions.

We assigned the feedbacks into one of **5** categories, as seen in the following example:

src > is_prime.py > ...

```python
1 def is_prime(n):
2     for x in range(2, n):
3         if int(n / x) == n / x:
4             return False
5     return True
```

Figure 3: *Example incorrect code submitted by a student.*

Table 1: *Example categorization of GPT generated feedback for code in Figure 3.*

Line #	Feedback	Category
3	You should use the mod operator **(%)** to check if **n** is divisible by **x**, not division **(/)**.	True Positive **(Valid)**
2	You have missed a corner case to check for **n = 1**. Recall that **1** is not prime.	False Negative **(Missed)**
2	The range of your **loop** should be from **2** to the **square root of n**, not to n.	False Positive **(Extra)**
3	The condition in your **if** statement should be **n%2 == 0** to check if the number is even.	False Positive **(Invalid)**
5	The **return** statement should be indented to be inside the function.	False Positive **(Hallucination)**

Table 2: *Feedback quality of GPT-3.5T and GPT-4 on 366 student submissions.*

	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%

Our evaluation demonstrates the state-of-art GPT-4 model performs significantly better than the GPT-3.5T model. Specifically, GPT-3.5T suffers from serious hallucination issues in **18.0%** of the cases, as compared to the **4.1%** of cases by GPT-4. Nevertheless, the occurrence of hallucinations and invalid feedback in even the state-of-the-art models is a cause of concern.

Furthermore, while multiple conversational iterations with evaluation oracle significantly improved our repair coverage, they have a marginal improvement on feedback quality, sometimes even increasing the cases of hallucination.

Future Work

In this work, we focused on evaluating the correctness of repaired code and feedback generated by state-of-the-art LLMs.

In future, we plan to:

1. Conduct a large-scale user study to evaluate real-world efficacy.
2. Evaluate quality of feedback across more complex attributes, such as informativeness and comprehensibility.

NeurIPS'23 Workshop on Generative AI for Education (GAIED).