NeurIPS 2023, Generative AI in Education workshop

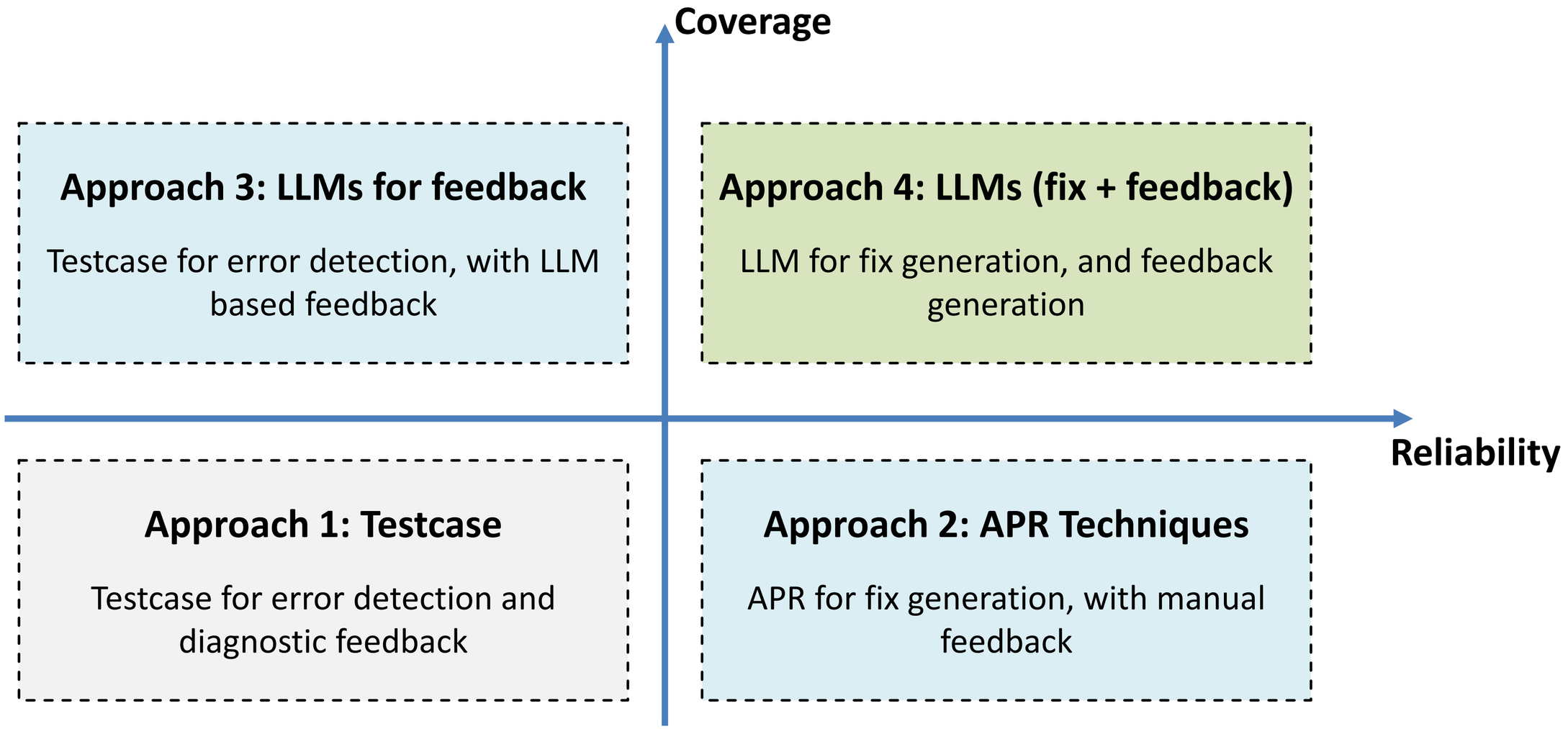
Improving the Coverage of GPT for Automated Feedback on High School Programming Assignments

Shubham Sahai, Umair Z. Ahmed, and **Ben Leong**

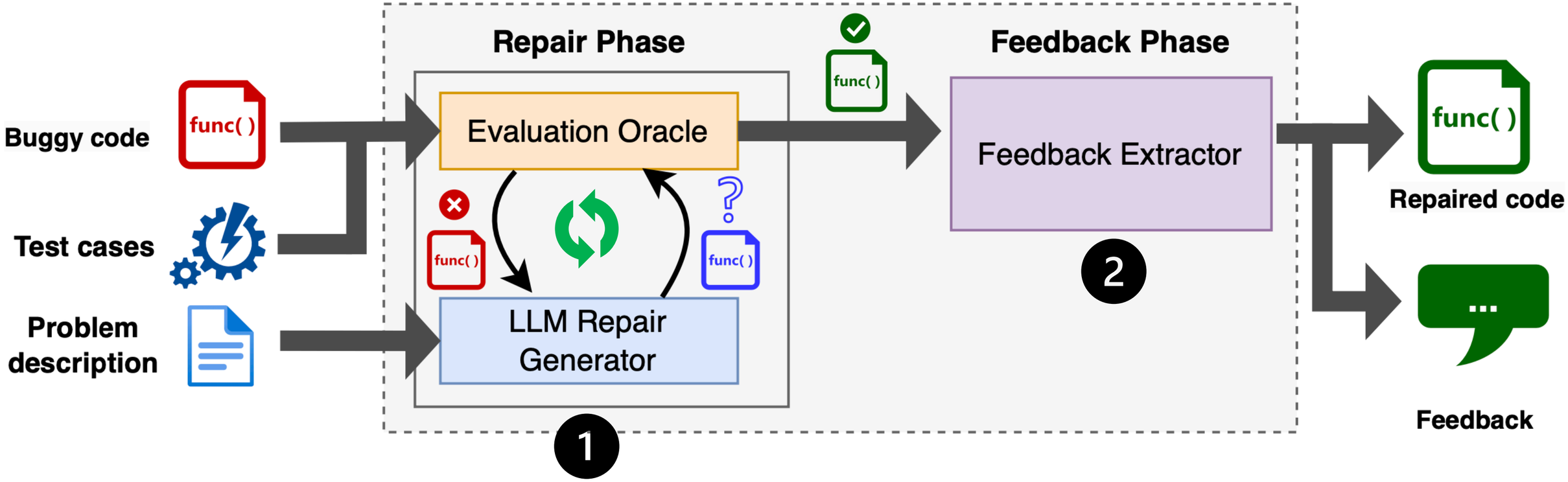
NUS National University of Singapore

AICET

Feedback Landscape



Architecture



Dataset

Public Releasable dataset from **NUS High**

69

Assignments

366

Buggy Solutions

5928

Correct Solutions

Results: Coverage

Model	Repair @ 1	Repair @ 5	
GPT 3.5T	64.8%	74.9%	↑ 10.1%
GPT 4	74.9%	88.5%	↑ 13.6%

CODAVERI

Confusion Matrix

		Model Output	
		Positive	**Negative**
Ground Truth	**Positive**		
	Negative		

Concrete Example

```
1  def convert16Bit(decimal):
2      s = '{0:b}'.format(decimal)
3      s = s.zfill(16)
4      print(s)
```

#	Category	Feedback
1	False Positive (Hallucination)	The function definition is missing a colon (:) at the end
4	True Positive	Instead of printing the result, you should return it.

(a) Feedback by GPT-3.5T

#	Category	Feedback
2	False Positive (Extra)	Use "0:016b" instead of "0:b". This will automatically pad the binary number with leading zeros to make it 16 bits.
4	True Positive	In Python, functions should return their result for use elsewhere in the program. Replace "print(s)" with "return s"

(b) Feedback by GPT-4

Results: Quality

Iteration	Model	Recall (TPR)	Miss (FNR)	Precision (PPV)	False Positive Rate (FPR) Extra	Invalid	Hallucination
Single	**GPT-3.5T**	52.7%	47.3%	51.2%	15.7%	15.0%	18.0%
Single	**GPT-4**	84.0%	16.0%	72.0%	14.8%	9.0%	4.1%
Multiple	**GPT-3.5T**	53.1%	46.9%	51.4%	15.2%	16.5%	16.9%
Multiple	**GPT-4**	87.2%	12.8%	72.4%	14.4%	7.7%	5.4%

Future Work

1. Large scale user study evaluate its real-world usability, in terms of pedagogical effectiveness on student’s learning outcomes and teacher’s grading process.

2. Evaluate on qualitative attributes such as informativeness and comprehensibility.

3. Exploring the effectiveness of our techniques for college level CS1 course.

CODAVERI

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

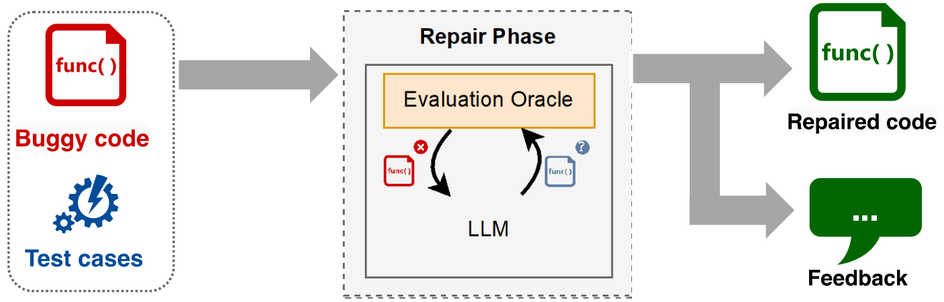


Figure 1: *Proposed architecture. LLM generates repair and feedback which is validated by an evaluation oracle against testcases.*

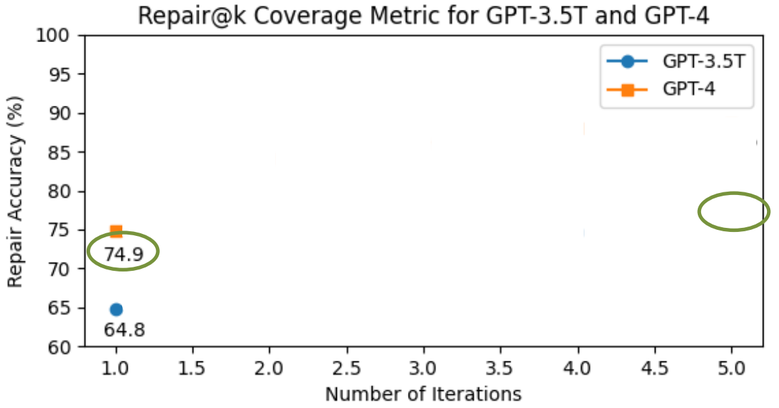


Figure 2: *Comparing repair accuracy of GPT-3.5T and GPT-4 after k interactive iterations*

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 Reliability	**Recall** Coverage	**False Positives** Invalid	**False Positives** 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.*

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

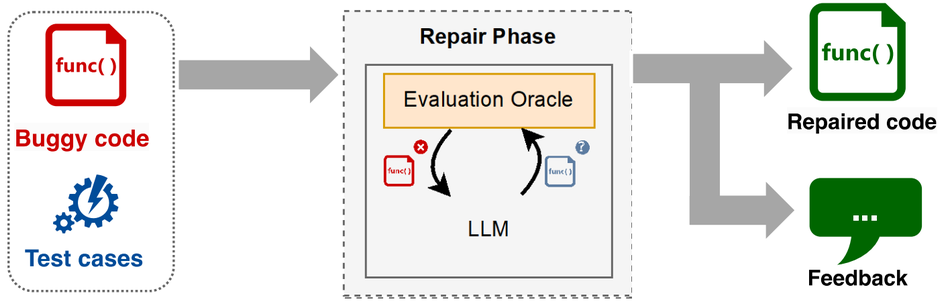


Figure 1: *Proposed architecture. LLM generates a repair and feedback which is validated by an evaluation oracle against testcases.*

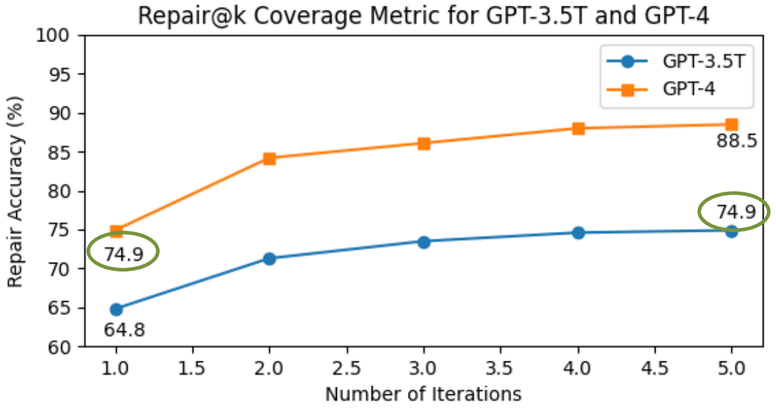


Figure 2: *Comparing repair accuracy of GPT-3.5T and GPT-4 after k interactive iterations*

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 Reliability	**Recall** Coverage	**False Positives** Invalid	**False Positives** 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.*