

Leveraging Mutation Analysis for LLM-based Repair of Quantum Programs

Chihiro Yoshida*, Yuta Ishimoto[†], Olivier Nourry*,
Masanari Kondo[†], Makoto Matsushita*, Yasutaka Kamei[†], Yoshiki Higo*

*The University of Osaka, Osaka, Japan

c-yosida@ist.osaka-u.ac.jp, nourry@ist.osaka-u.ac.jp, matusita@ist.osaka-u.ac.jp, higo@ist.osaka-u.ac.jp

[†]Kyushu University, Fukuoka, Japan

ishimoto@posl.ait.kyushu-u.ac.jp, kondo@ait.kyushu-u.ac.jp, kamei@ait.kyushu-u.ac.jp

Abstract—In recent years, Automated Program Repair (APR) techniques specifically designed for quantum programs have been proposed. However, existing approaches often suffer from low repair success rates or poor understandability of the generated patches. In this study, we construct a framework in which a large language model (LLM) generates code repairs along with a natural language explanation of the applied repairs. To investigate how the contextual information included in prompts influences APR performance for quantum programs, we design four prompt configurations with different combinations of static information, dynamic information, and mutation analysis results. Mutation analysis evaluates how small changes to specific parts of a program affect its execution results and provides more detailed dynamic information than simple execution outputs such as stack traces. Our experimental results show that mutation analysis can provide valuable contextual information for LLM-based APR of quantum programs, improving repair success rates (achieving 94.4% in our experiment) and in some cases also improving the quality of generated explanations. Our findings point toward new directions for developing APR techniques for quantum programs that enhance both reliability and explainability.

Index Terms—Quantum Programs, Automated Program Repair, Mutation Analysis, Large Language Models, Explanation

I. INTRODUCTION

Quantum computing has drawn significant attention for its potential to outperform classical machines on specific computational tasks [1]. To harness the computational power of quantum computers, developers must write *quantum programs* that specify sequences of operations on quantum bits (*qubits*), where each operation is called a *quantum gate*.

Similar to classical (non-quantum) programs, quantum programs also need to be debugged to ensure their correctness [2], [3]. However, debugging quantum programs poses distinctive challenges, as it involves quantum-specific concepts such as superposition and entanglement [4], [5], [6]. Indeed, these quantum-specific concepts lead to unique software engineering challenges, such as quantum-specific technical debt [7], [8] and code smells [9].

To assist developers in debugging quantum programs, researchers have proposed *automated program repair (APR)* techniques specifically designed for quantum software [10], [11], [12]. The goal of APR is to generate patches that modify buggy programs to pass all test cases in a given test suite [13].

Guo et al. [11] investigated the capability of ChatGPT [14] to repair quantum programs. Their results showed that ChatGPT could fix only 17% of complex bugs that involve implementing quantum algorithms. Tan et al. [12] and Li et al. [10] proposed APR techniques that generate quantum gates as repair patches and insert them into quantum programs. The current state-of-the-art approach, HornBro [12], has been shown to fix more bugs than ChatGPT-based APR methods. However, this improvement comes at the cost of increased program complexity due to the insertion of additional gates. In one instance, HornBro added up to 249 additional gates while fixing a buggy program, negatively impacting the understandability and maintainability of the repaired code [9]. In summary, existing APR techniques for quantum programs suffer from either (1) low repair success rates or (2) poor understandability and maintainability of the repaired programs.

In this study, we aim to improve APR for quantum programs by incorporating insights from *mutation analysis* into a large language model (LLM)-assisted repair process. Mutation analysis evaluates how small changes to specific parts of a program affect its execution results. For quantum programs, mutation analysis combines classical mutation operators applicable to classical programs (e.g., replacing arithmetic operators) with quantum mutation operations (e.g., adding and deleting quantum gates) [15]. This approach has demonstrated its effectiveness for testing [15], [16], [17], [18] and fault localization [19]. In our experiments, each program is first analyzed using mutation analysis. The LLM is then given the buggy quantum program, its stack trace, and the mutation analysis results as input to generate *fixed code* along with a *natural language explanation* of the applied repair. Generating a natural language explanation alongside the fixed code helps developers understand the intent and rationale behind the repair, which cannot be easily inferred from the fixed code alone. We hypothesize that mutation analysis results can improve the repair success rate and help the model produce richer explanations, as they provide more detailed dynamic information than simple execution outputs, such as stack traces. The research questions (RQs) are as follows:

RQ1 (Repair success rate) Can mutation analysis results improve the repair success rate of buggy quantum pro-

grams?

RQ2 (Quality of explanation) Can mutation analysis results improve the quality of explanations generated by LLMs?

Our experimental setup includes 18 real-world buggy quantum programs from the Bugs4Q dataset [20]. All of these programs are written in Python using Qiskit [21], a widely used library for quantum programming. Throughout the experiment, we use GPT-5 [22], a state-of-the-art LLM, as the base model for generating the program repairs. For mutation analysis, we employ QMutPy [15], a mutation testing framework designed for quantum programs. We compared four prompt configurations to analyze how different contextual information affects repair performance. Our main contributions are as follows:

- We present the first empirical evidence that mutation analysis can serve as a valuable signal for improving LLM-based APR for quantum programs.
- We demonstrate that incorporating both dynamic runtime information and mutation analysis results into the prompt can yield a higher repair success rate.
- We show that incorporating mutation analysis results can improve the quality of LLM-generated explanations by making positional descriptions more accurate, complete, and concise.

II. RELATED WORK

A. Automated Program Repair for Classical Programs

Among the APR techniques, learning-based repairs [23] have recently gained attention, driven by rapid advances in machine learning, particularly LLMs. Numerous studies for LLM-based repairs highlight the importance of the contextual information included in the prompts [24], [25], [26], [27]. For example, InferFix [24] enhances prompts with static analysis results and bug location, enabling the LLM to generate repairs with explicit awareness of the bug type and its position in the code. Ehsani et al. [27] enhance prompts by hierarchically injecting bug-, repository-, and project-level knowledge to provide broader and more relevant repair context.

Our work follows a similar line of research in investigating the impact of prompts on the performance of LLM-based APR. The novelty of our approach lies in augmenting prompts with mutation analysis results that are effective for testing [15], [16], [17], [18] and fault localization [19] for quantum programs. To the best of our knowledge, no prior work has leveraged mutation analysis to enhance LLM-based APR of either quantum or classical programs.

B. Automated Program Repair for Quantum Programs

Fixing bugs in quantum programs is challenging because it requires domain-specific knowledge of quantum computing [4], [5], [6]. In a previous study, Guo et al. [11] used ChatGPT [14] to explore APR for quantum programs. Their results showed that while ChatGPT successfully repaired 97% of classical (non-quantum) bugs in their dataset, it only managed to repair 17% of quantum bugs. Li et al. [10] proposed UnitAR, which repairs quantum programs by automatically generating

unitary operations. Tan et al. [12] proposed HornBro, which removes faulty gates responsible for the bug and synthesizes new ones to produce the correct behavior. Both UnitAR and HornBro can be regarded as *synthesis-based repair* techniques because they generate and insert new gates to fix buggy programs. We chose an LLM-based repair approach for the following reasons:

- 1) Unlike synthesis-based approaches, which are limited to gate-related bugs, LLM-based approaches are more flexible and can handle a wider range of bugs such as API-related bugs, which are also commonly found in real-world quantum programs [5].
- 2) Synthesis-based approaches increase the number of gates in a quantum program, thereby increasing its complexity. Avoiding excessive gate counts is recommended as a best practice by Google Quantum AI [28] and has also been identified as a type of quantum code smell [9].

Our approach not only instructs the LLM to generate minimal repairs but also to provide natural language explanations for them. Unlike other APR techniques for quantum programs, this allows developers to examine and validate the rationale and intent behind each change.

III. STUDY DESIGN

This section describes the design of our study, including the benchmark, prompt configurations, and LLMs. Figure 1 provides an overview of our experimental process. Our goal is to investigate how incorporating mutation analysis results into LLM-based repair affects the repair success rate and the quality of generated explanations. At a high level, for each buggy program, we prepare four different prompt configurations. The LLM is instructed to generate both the fixed code and an explanation of the fix regardless of the prompting configuration used.

A. Bug Benchmark

In this study, we use Bugs4Q [20], a benchmark of real-world bugs in quantum programs implemented using Qiskit [21]. It contains 42 buggy programs and their fixed versions, collected from GitHub, Stack Overflow, and Stack Exchange. We used Bugs4Q benchmark because it satisfied the following criteria:

- 1) It contains bugs found in real-world quantum programs written by human developers.
- 2) It includes test scripts to verify the behavior of both the buggy and fixed code.

The first criterion is necessary to compare the LLM’s explanations of the generated patches with the developers’ true intents behind the repairs. The second criterion is essential to perform mutation analysis and to verify whether the generated patches are correct.

Among the 42 programs in Bugs4Q, we selected 18 programs as our experimental targets. First, we cloned the replication package of Bugs4Q¹ and confirmed whether each

¹<https://github.com/Z-928/Bugs4Q-Framework>

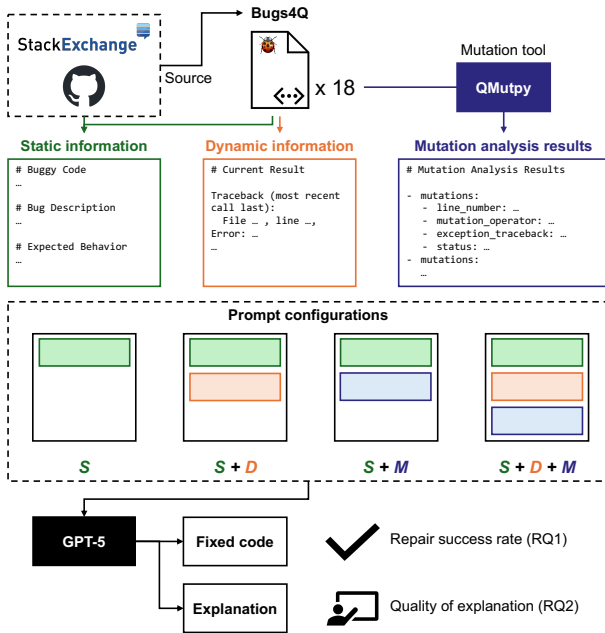


Fig. 1: Overview of the experiment.

bug could be reproduced. Bugs4Q bugs exhibit two types of symptoms: *Throw Exception (TE)* (e.g., an exception raised by the Qiskit library) and *Wrong Output (WO)* (e.g., an AssertionError triggered when the measured quantum state differs from the expected state). 19 programs were excluded because their bugs could not be reproduced. We also excluded 5 programs that did not satisfy either of the following criteria necessary to construct the prompts used in our experiment:

- 1) At least one mutant must be generated when performing mutation analysis.
- 2) The URL of the source repository must be accessible.

B. Prompt Construction

To investigate how the information included in prompts influences APR performance for quantum programs, we design the following four prompt configurations (see Figure 1):

- *S*: Static information only.
- *S+D*: Static and dynamic information.
- *S+M*: Static information and mutation analysis results.
- *S+D+M*: Static information, dynamic information, and mutation analysis results.

We describe how each type of information is obtained below. **Static Information.** We collected three types of static information: (1) *buggy code* obtained from Bugs4Q; (2) *bug descriptions* obtained from the source URLs; and (3) *expected behavior* taken from the source URLs, describing the intended correct program behavior. (2) and (3) were manually collected by two authors, who inspected the source URLs of each bug and extracted the necessary information. Any disagreements were resolved through discussion to reach a consensus.

Dynamic Information. Dynamic information consists of the execution results of buggy programs, including the error message and the accompanying stack trace.

Mutation Analysis Results. We applied QMutPy [15] to each buggy program to obtain the mutation analysis results. QMutPy implements 20 classical and 5 quantum mutation operators, including insertion, deletion, and replacement of quantum gates, as well as insertion and deletion of quantum measurements. Each mutation operation is applied to a single location at a time. The mutation operators are applied repeatedly, producing multiple mutated programs (i.e., mutants) for each quantum program. The original program’s test suite is executed on all mutants to determine their outcomes. We refer to the resulting collection of data as the mutation analysis results, which includes the following information for each applied mutation operation:

- *line_number*: The line number where the mutation operation was applied.
- *mutation_operator*: The name of the mutation operator applied to the program.
- *exception_traceback*: The traceback output when a test execution for the mutant fails.
- *status*: The result of the mutation test, categorized into one of the following: (1) *killed*, at least one test produces a different outcome; (2) *survived*, no change in test outcome; (3) *incompetent*, non-viable mutant (e.g., compile error, crash); (4) *time_out*, execution exceeded the time limit (e.g., infinite loop).

The system prompt is shared across all four prompt configurations. It provides detailed instructions on how to interpret and use each piece of information included in the prompt, as well as the strict output format to follow. All prompts used in our study are available in our replication package [29].

C. LLM

In this study, we used the state-of-the-art *GPT-5* via the OpenAI API with default settings. Because LLMs are stochastic and may produce different outputs for the same input, we generated outputs five times for each prompt configuration. Consequently, we obtained a total of 18 programs \times 4 prompt configurations \times 5 generations = 360 generated repairs and their corresponding explanations.

D. Evaluation Metrics

(RQ1) Repair Success Rate. We count a repair as successful if the generated code passes all the tests provided in Bugs4Q. The repair success rate is defined as the proportion of the 18 programs for which at least one of the five repair attempts was successful.

(RQ2) Quality of Explanation. The LLM-generated explanations may describe the cause of the bug, the location of the repair, and the rationale behind the repair. To evaluate these explanations, we adopt three core elements of patch explanations [30]: (1) *Position* — where the bug occurs; (2) *Cause* — why the bug occurs; (3) *Change* — how the bug is repaired. The quality of explanations is evaluated using three criteria proposed in previous work [31], [32]:

- *Correctness*: Is the explanation accurate with respect to the ground-truth fixed code?

- *Completeness*: Does the explanation fully describe all changes in the patch?
- *Complexity*: Does the explanation contain unnecessary complexity?

For each of the three explanation elements (Position, Cause, Change), we evaluate whether it satisfies each of the three criteria (Correctness, Completeness, Complexity) through a binary judgment (yes/no). For example, given an LLM-generated explanation, we evaluate the *Position* element by answering the following three questions: 1) does the explanation *correctly* explain the position of the bug?, 2) is the explanation *complete* and mentions all locations that were modified for the repair?, and 3) does the explanation include unnecessary *complexity* to explain the location of the bug?

Evaluating the three elements of patch explanation (Position, Cause, Change) therefore resulted in 9 (3 elements \times 3 criteria) total binary evaluations per explanation generated by the LLM. We only considered the first generated repair patch among the five produced for each prompt, resulting in a total of 72 explanations. Two authors independently assessed each explanation, resolving any disagreements through discussion.

IV. RESULTS

We ran all experiments on simulators on classical computers due to the limited availability of quantum hardware and the innate instability of quantum computing posing a challenge for the reproducibility of our experiment.

A. RQ1: Can mutation analysis results improve the repair success rate of buggy quantum programs?

Table I shows the repair success rates for each prompt configuration and bug type. The “Total” column reports the repair success rate across all 18 programs, while the “TE” and “WO” columns show those for each bug type (eight and ten programs, respectively). Out of 18 programs, 17 of them could be repaired by at least one prompt configuration.

Mutation analysis results are most effective when used in combination with dynamic information. Among all prompt configurations, $S+D+M$ achieved the highest repair success rate at 94.4% in total. When considering only WO bugs, $S+D+M$ successfully repaired all programs. Interestingly, $S+M$ achieved a lower total repair success rate than $S+D$. Because mutation analysis requires running tests, the availability of M implies that D is also available. Thus, the lower success rate of $S+M$ is not concerning, as $S+D+M$ can easily be constructed in such cases. In contrast, for

TABLE I: Repair success rates by prompt configuration and bug type. WO: Wrong Output, TE: Throw Exception. The maximum value in each column is highlighted in bold.

Prompt	Total [%]	WO [%]	TE [%]
S	77.8	70.0	87.5
$S+D$	88.9	90.0	87.5
$S+M$	83.3	80.0	87.5
$S+D+M$	94.4	100.0	87.5

TE bugs, the set of successfully repaired programs (seven out of eight) was identical across all prompt configurations. This indicates that dynamic information and mutation analysis results have no impact on buggy code in this category. For WO bugs, the program executes without crashing, making runtime information such as mutation analysis results helpful for repair. For TE bugs, the program terminates with an error, making the bug evident and static information alone sufficient.

Answer to RQ1: Combining dynamic information with mutation analysis results is most effective, yielding a total repair success rate of 94.4%. It also achieved a 100% repair success rate for the bug type of the Wrong Output.

B. RQ2: Can mutation analysis results improve the quality of explanations generated by LLMs?

Table II reports how many of the 18 programs satisfy each criterion for each prompt configuration. The two evaluators agreed on 79.2% of the evaluations (513 out of 648 = 72 explanations \times 9 evaluation items), and Cohen’s κ [33] was 0.48, indicating moderate agreement [34]. This value is comparable to the $\kappa = 0.55$ reported in prior work [35] evaluating LLM-generated explanations for fault localization.

TABLE II: Quality of explanation for each prompt configuration. \uparrow (\downarrow) indicates that a higher (lower) value is better. The best value in each row is highlighted in bold.

Criterion	Element	S	$S+D$	$S+M$	$S+D+M$
Correctness (\uparrow)	Position	12	12	13	14
	Cause	14	13	12	12
	Change	7	9	5	7
Completeness (\uparrow)	Position	13	15	12	15
	Cause	18	15	16	15
	Change	13	16	13	16
Complexity (\downarrow)	Position	2	2	1	1
	Cause	2	2	1	1
	Change	8	7	8	6

From the bolded values in Table II, $S+D+M$ achieved the best scores in six out of nine evaluation items, indicating that this prompt configuration produces better explanations compared to other prompts. In particular, for the position element, $S+D+M$ achieved the best scores across all criteria, indicating that this prompt configuration is especially effective at positional explanations, leveraging dynamic information and mutation analysis results. On the other hand, $S+D+M$ is less effective for the cause element. For this element, the best scores in correctness and completeness were achieved by the prompt configuration S . One possible reason is that most of the information relevant to the cause of the repair is already contained in the static information. Mutation analysis results are more effective for detailed positional explanations than for explaining the cause of the bug.

Answer to RQ2: Mutation analysis results combined with dynamic information achieved the best scores in six out of nine evaluation items by making positional descriptions more accurate, complete, and concise.

V. DISCUSSION

A. Unique Successful Repairs Across Prompt Configurations

Since RQ1 showed the effectiveness of dynamic information and mutation analysis results, we further investigate which programs particularly benefit from these components. Figure 2 shows the number of programs fixed by each prompt configuration. While all configurations repair 14 of the 18 buggy programs, the $S+D+M$ configuration is the only one that succeeds in repairing 17 programs. This indicates that dynamic information and mutation analysis results did not hinder repair effectiveness, as the $S+D+M$ configuration encompasses all successfully repaired programs. Our results demonstrate that mutation analysis can enhance the effectiveness of LLM-based APR for real-world buggy quantum programs, providing a concrete benefit.

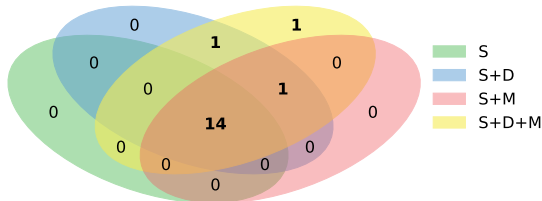


Fig. 2: Successful repairs by prompt configuration.

B. Future Research Plan

(1) Exploring Additional Contextual Information. One limitation of our study is that it applies only to quantum programs for which mutation analysis can be performed. For three programs, we were unable to generate any mutants even though the mutation tool execution was successful. To address this limitation, we plan to incorporate additional types of contextual information (e.g., relevant sections of the Qiskit documentation) alongside mutation analysis results.

(2) Identifying Domain Knowledge Gaps. As another direction for future work, we plan to explore how LLM-generated explanations can help developers identify gaps in their quantum-specific knowledge. A prior survey shows that contributors to quantum open-source software (OSS) often lack sufficient understanding of physics and quantum computing [4]. Our workflow could help developers recognize such gaps through explanations generated during the repair. To validate the practical utility of this approach, we plan to conduct surveys with OSS developers. While LLM-based explanations of quantum algorithms have shown promise [36], explanations derived from program repair represent an unexplored avenue.

VI. THREATS TO VALIDITY

Construct Validity. In RQ1, we defined a successful repair as one that passes all tests, but this does not guarantee that the repair fully matches the developer’s intended modification. This may affect the validity of the reported repair success rate.

Internal Validity. Due to the stochastic nature of GPT-5, the same prompt does not necessarily yield identical results. Although we generated five outputs for each prompt configuration, the inherent variability of LLM outputs may still influence the findings. In RQ2, two authors manually evaluated the quality of the explanations to mitigate subjectivity. However, some degree of subjective judgment is unavoidable, and misclassifications may have occurred due to limited domain knowledge in quantum programming.

External Validity. This study was conducted using quantum programs on the Qiskit simulator, and it is unclear whether similar results would be obtained on actual quantum computers. Since our investigation is limited to a single benchmark (Bugs4Q) and a single quantum framework (Qiskit), a full validation using a wider range of experimental subjects (i.e., different bug benchmarks, other frameworks, or LLMs other than GPT-5) is needed.

VII. CONCLUSION

In this study, we demonstrated the effectiveness of incorporating mutation analysis results into prompts for LLM-based APR of quantum programs using real-world bug benchmarks. From our experiment, we find that the $S+D+M$ prompt configuration, which includes static information, dynamic information, and mutation analysis results, achieves the highest repair success rate of 94.4%. We also find that explanations generated using $S+D+M$ exhibit the lowest complexity while providing accurate, complete, and concise descriptions of bug positions. These results indicate that combining mutation analysis results with dynamic information can improve both the repair success rate and the quality of LLM-generated explanations. This work presents the first empirical evidence that mutation analysis can serve as valuable context information for improving LLM-based APR of quantum programs. These results point toward new directions for developing APR techniques for quantum programs that enhance both reliability and explainability.

VIII. DATA AVAILABILITY

All data, benchmarks, scripts, and prompts used in this study are publicly available in our replication package [29].

ACKNOWLEDGMENT

We gratefully acknowledge the financial support of: (1) Japan Society for the Promotion of Science, Grant-in-Aid for Scientific Research (B), Grant Number JP25K03102, (2) Japan Society for the Promotion of Science, Grant-in-Aid for Scientific Research (A), Grant Number JP24H00692, (3) Japan Society for the Promotion of Science, Grant-in-Aid for Scientific Research (B), Grant Number JP23K24823, (4) Japan Science and Technology Agency (JST) as part of Adopting Sustainable Partnerships for Innovative Research Ecosystem

(ASPIRE), Grant Number JPMJAP2415, and (5) the Inamori Research Institute for Science for supporting Yasutaka Kamei via the InaRIS Fellowship.

REFERENCES

- [1] M. A. Nielsen and I. L. Chuang, *Quantum computation and quantum information*. Cambridge university press, 2010.
- [2] J. M. Murillo, J. Garcia-Alonso, E. Moguel, J. Barzen, F. Leymann, S. Ali, T. Yue, P. Arcaini, R. Pérez-Castillo, I. García-Rodríguez de Guzmán *et al.*, “Quantum software engineering: Roadmap and challenges ahead,” *ACM Transactions on Software Engineering and Methodology*, vol. 34, no. 5, pp. 1–48, 2025.
- [3] N. C. Leite Ramalho, H. Amario de Souza, and M. Lordello Chaim, “Testing and debugging quantum programs: The road to 2030,” *ACM Transactions on Software Engineering and Methodology*, vol. 34, no. 5, pp. 1–46, 2025.
- [4] R. Shaydulin, C. Thomas, and P. Rodeghero, “Making quantum computing open: Lessons from open source projects,” in *Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering Workshops*, 2020, p. 451–455.
- [5] J. Luo, P. Zhao, Z. Miao, S. Lan, and J. Zhao, “A comprehensive study of bug fixes in quantum programs,” in *Proceedings of the 2022 IEEE International Conference on Software Analysis, Evolution and Reengineering*. IEEE, 2022, pp. 1239–1246.
- [6] J. Zappin, T. Stalnakar, O. Chaparro, and D. Poshyvanyk, “When quantum meets classical: Characterizing hybrid quantum-classical issues discussed in developer forums,” in *Proceedings of the IEEE/ACM 47th International Conference on Software Engineering*, 2025, pp. 2931–2943.
- [7] M. Openja, M. M. Morovati, L. An, F. Khomh, and M. Abidi, “Technical debts and faults in open-source quantum software systems: An empirical study,” *Journal of Systems and Software*, vol. 193, p. 111458, 2022.
- [8] Y. Ishimoto, Y. Nakamura, R. Katsube, N. Sato, H. Ogawa, M. Kondo, Y. Kamei, and N. Ubayashi, “An empirical study on self-admitted technical debt in quantum software,” in *Proceedings of the 31st Asia-Pacific Software Engineering Conference*, 2024, pp. 41–50.
- [9] Q. Chen, R. Câmara, J. Campos, A. Souto, and I. Ahmed, “The smelly eight: An empirical study on the prevalence of code smells in quantum computing,” in *Proceedings of the IEEE/ACM 45th International Conference on Software Engineering*, 2023, pp. 358–370.
- [10] Y. Li, H. Pei, L. Huang, B. Yin, and K.-Y. Cai, “Automatic repair of quantum programs via unitary operation,” *ACM Transactions on Software Engineering and Methodology*, vol. 33, no. 6, pp. 1–43, 2024.
- [11] X. Guo, J. Zhao, and P. Zhao, “On repairing quantum programs using chatgpt,” in *Proceedings of the 5th ACM/IEEE International Workshop on Quantum Software Engineering*, 2024, pp. 9–16.
- [12] S. Tan, L. Lu, D. Xiang, T. Chu, C. Lang, J. Chen, X. Hu, and J. Yin, “Hornbro: Homotopy-like method for automated quantum program repair,” *Proceedings of the ACM on Software Engineering*, vol. 2, no. FSE, pp. 734–756, 2025.
- [13] C. Le Goues, M. Pradel, and A. Roychoudhury, “Automated program repair,” *Communications of the ACM*, vol. 62, no. 12, pp. 56–65, 2019.
- [14] OpenAI, “ChatGPT,” <https://chat.openai.com/>, 2022, accessed: 2025-11-04.
- [15] D. Fortunato, J. Campos, and R. Abreu, “Mutation testing of quantum programs: A case study with qiskit,” *IEEE Transactions on Quantum Engineering*, vol. 3, pp. 1–17, 2022.
- [16] S. Ali, P. Arcaini, X. Wang, and T. Yue, “Assessing the effectiveness of input and output coverage criteria for testing quantum programs,” in *Proceedings of the 14th IEEE Conference on Software Testing, Verification and Validation*, 2021, pp. 13–23.
- [17] E. Mendiluze, S. Ali, P. Arcaini, and T. Yue, “Muskit: A mutation analysis tool for quantum software testing,” in *Proceedings of the 36th IEEE/ACM International Conference on Automated Software Engineering*, 2021, pp. 1266–1270.
- [18] E. Mendiluze Usandizaga, S. Ali, T. Yue, and P. Arcaini, “Quantum circuit mutants: Empirical analysis and recommendations,” *Empirical Software Engineering*, vol. 30, no. 3, p. 100, 2025.
- [19] Y. Ishimoto, M. Kondo, N. Ubayashi, Y. Kamei, R. Katsube, N. Sato, and H. Ogawa, “Evaluating mutation-based fault localization for quantum programs,” *arXiv preprint arXiv:2505.09059*, 2025.
- [20] P. Zhao, Z. Miao, S. Lan, and J. Zhao, “Bugs4q: A benchmark of existing bugs to enable controlled testing and debugging studies for quantum programs,” *Journal of Systems and Software*, vol. 205, p. 111805, 2023.
- [21] A. Javadi-Abhari, M. Treinish, K. Krsulich, C. J. Wood, J. Lishman, J. Gacon, S. Martiel, P. D. Nation, L. S. Bishop, A. W. Cross, B. R. Johnson, and J. M. Gambetta, “Quantum computing with Qiskit,” 2024.
- [22] OpenAI, “Introducing gpt-5,” <https://openai.com/index/introducing-gpt-5/>, 2025, accessed: 2025-11-12.
- [23] Q. Zhang, C. Fang, Y. Ma, W. Sun, and Z. Chen, “A survey of learning-based automated program repair,” *ACM Transactions on Software Engineering and Methodology*, vol. 33, no. 2, pp. 1–69, 2023.
- [24] M. Jin, S. Shahriar, M. Tufano, X. Shi, S. Lu, N. Sundaresan, and A. Svyatkovskiy, “Inferfix: End-to-end program repair with llms,” in *Proceedings of the 31st ACM joint european software engineering conference and symposium on the foundations of software engineering*, 2023, pp. 1646–1656.
- [25] C. S. Xia and L. Zhang, “Automated program repair via conversation: Fixing 162 out of 337 bugs for \$0.42 each using chatgpt,” in *Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis*, 2024, pp. 819–831.
- [26] I. Bouzenia, P. Devanbu, and M. Pradel, “Repairagent: An autonomous, llm-based agent for program repair,” in *Proceedings of the IEEE/ACM 47th International Conference on Software Engineering*, 2025, pp. 694–694.
- [27] R. Ehsani, E. Parra, S. Haiduc, and P. Chatterjee, “Hierarchical knowledge injection for improving llm-based program repair,” 2025.
- [28] Google, “Google best practice,” 2025, accessed: 2025-11-13. [Online]. Available: https://quantumai.google/cirq/google/best_practices
- [29] Anonymous, “Leveraging mutation analysis for llm-based automated quantum program repair,” 2025. [Online]. Available: <https://doi.org/10.5281/zenodo.17626083>
- [30] J. Liang, Y. Hou, S. Zhou, J. Chen, Y. Xiong, and G. Huang, “How to explain a patch: The empirical study of patch explanations in open source projects,” in *Proceedings of the 2019 IEEE 30th International Symposium on Software Reliability Engineering*, 2019, pp. 58–69.
- [31] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, and C. Seifert, “From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai,” *ACM Comput. Surv.*, vol. 55, no. 13s, p. 42, 2023.
- [32] D. Sobania, A. Geiger, J. Callan, A. Brownlee, C. Hanna, R. Moussa, M. Z. López, J. Petke, and F. Sarro, “Evaluating explanations for software patches generated by large language models,” in *Proceedings of the International Symposium on Search Based Software Engineering*. Springer, 2023, pp. 147–152.
- [33] J. Cohen, “A coefficient of agreement for nominal scales,” *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.
- [34] J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data,” *biometrics*, pp. 159–174, 1977.
- [35] S. Kang, G. An, and S. Yoo, “A quantitative and qualitative evaluation of llm-based explainable fault localization,” *Proceedings of the ACM on Software Engineering*, vol. 1, no. FSE, pp. 1424–1446, 2024.
- [36] G. d’Aloisio, S. Fortz, C. Hanna, D. Fortunato, A. Bensoussan, E. Mendiluze Usandizaga, and F. Sarro, “Exploring llm-driven explanations for quantum algorithms,” in *Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, 2024, pp. 475–481.