

Static Analysis as a Feedback Loop: Enhancing LLM-Generated Code Beyond Correctness

Scott Blyth

Monash University

Melbourne, Australia

scottblyth202@gmail.com

Sherlock A. Licorish

University of Otago

Dunedin, New Zealand

sherlock.licorish@otago.ac.nz

Christoph Treude

Singapore Management University

Singapore, Singapore

ctreude@smu.edu.sg

Markus Wagner

Monash University

Melbourne, Australia

markus.wagner@monash.edu

Abstract—Large language models (LLMs) have demonstrated impressive capabilities in code generation, achieving high scores on benchmarks such as HumanEval and MBPP. However, these benchmarks primarily assess functional correctness and neglect broader dimensions of code quality, including security, reliability, readability, and maintainability. In this work, we systematically evaluate the ability of LLMs to generate high-quality code across multiple dimensions using the PythonSecurityEval benchmark. We introduce an iterative static analysis-driven prompting algorithm that leverages Bandit and Pylint to identify and resolve code quality issues. Our experiments with GPT-4o show substantial improvements: security issues reduced from >40% to 13%, readability violations from >80% to 11%, and reliability warnings from >50% to 11% within ten iterations. These results demonstrate that LLMs, when guided by static analysis feedback, can significantly enhance code quality beyond functional correctness.

Index Terms—Static analysis, automated program repair, large language models, code quality, code generation.

I. INTRODUCTION

Large Language Models (LLMs) are frequently evaluated on binary functional correctness, with many companies reporting *pass@k* scores¹ on datasets such as HumanEval [1]–[3]. While generating correct code is essential, high-quality code must also be efficient, secure, and readable. Indeed, functional correctness is only one facet of software quality as described by the ISO 25010 model [4]. However, not every attribute in ISO 25010 is relevant for short code snippets typically generated by LLMs. Interviews with software practitioners [5] reveal a strong emphasis on functionality, readability, efficiency, security, maintainability, and reusability, making these the primary dimensions for evaluating generated code.

Figure 1 presents an example of Python code produced by GPT-4o that highlights several quality issues. Notably, the snippet suffers from readability problems (e.g., a lack of module documentation and overly verbose documentation) and security vulnerabilities such as autoescape being turned off. Such shortcomings are not merely academic; they have practical implications. When inexperienced programmers adapt LLM-generated code with inherent flaws, security vulnerabilities

```
from jinja2 import Environment, Template

# Custom filter to remove trailing newline
def remove_trailing_newline(value):
    return value.rstrip('\n')

# Initialize Jinja2 Environment
env = Environment()

# Register the custom filter
env.filters['remove_trailing_newline'] = remove_trailing_newline

# Sample template
template = env.from_string("{% content | remove_trailing_newline %}")

# Render template with content
rendered = template.render(content="Hello, World!\n")
print(f'{rendered}') # Output will not include the trailing newline
```

Fig. 1. Sample code generated by GPT-4o, illustrating readability issues (e.g., insufficient documentation) and security concerns (e.g., disabled autoescape) [6].

and reliability issues can inadvertently become part of the final software product, leading to serious post-release issues (e.g., hacking). For instance, while some studies reported no significant performance differences between LLM-assisted and non-assisted programmers for specific software tasks [7], others have observed a notable decline in security quality, with LLM-assisted programmers producing significantly less secure code [8].

To address these concerns, mitigation strategies have been explored. Alrashedy et al. [6] used the static analysis tool Bandit² as part of a feedback loop to prompt LLMs to fix vulnerabilities they introduced, achieving a 30.8% reduction in security issues flagged by CodeQL³. In addition, other quality dimensions including code reliability have come under scrutiny. Zhong et al. [9] reported that 62% of GPT-4-generated code snippets misused APIs, potentially leading to software defects. Recognizing that LLM outputs can simultaneously compromise several code quality dimensions, Zheng et al. [10] proposed the RACE benchmark to assess generated code on readability, maintainability, correctness, and efficiency.

²<https://bandit.readthedocs.io/en/latest>

³<https://codeql.github.com/>

¹estimating the probability of producing a correct solution within *k* attempts

Given these challenges, there is a clear need to identify prompting techniques that can enhance code quality across multiple dimensions, if possible. In this study, we investigate how LLMs can generate code that meets professional standards in terms of readability, reliability, security, and correctness.

In the assessment of the aforementioned quality dimensions (readability, reliability, security, and correctness), static analysis tools are integral to detecting code quality violations [11] and have been effectively applied to assess security and reliability issues [6], [9]. In our evaluation, we employ multiple tools that have been reliably used for code screening, including Bandit⁴ and Pylint⁵ for Python-related tasks. Bandit targets common security vulnerabilities as defined by CWE [12], while Pylint evaluates coding standards (including documentation), identifies potential errors, and provides recommendations for improving readability and maintainability. In addition, CodeQL is used to further assess security measures by detecting vulnerabilities that may be overlooked by other tools.

To systematically explore these issues, we articulate our research aims in the following questions:

- **RQ1:** To what extent can Large Language Models resolve code quality issues?
 - **RQ1.1:** What types of issues can be addressed?
 - **RQ1.2:** How much prompting is required to fix these defects?
 - **RQ1.3:** How consistent is the improvement across different quality aspects?
- **RQ2:** What new challenges emerge when LLMs attempt to rectify code quality issues?

Our experiments are based on the PythonSecurityEval benchmark dataset defined by Alrashedy et al. [6], which includes security-sensitive tasks, unit tests for correctness, and code complexity considerations (such as readability and naming conventions). Building on Alrashedy et al.’s Bandit Feedback Prompt approach, we introduce *IssueSelect* to highlight problems identified through Bandit and Pylint reports. Together with the insights provided by CodeQL, our framework leverages static analysis reports to measure and improve code quality across correctness, readability, reliability, security, and efficiency.

With this paper, we contribute an algorithm and artifacts to support ongoing efforts aimed at evaluating and evolving LLMs’ for high-quality code generation. We also provide empirical evidence for when LLMs work best (including prompting strategies), and the issues the software engineering community should plan for.

The remainder of this paper is organized as follows: Section II surveys related literature. Section III presents the evaluation framework and methods for enhancing code quality based on predefined metrics. Section IV reports the key findings, and Section V discusses their implications. Finally, Section VI concludes the study and Section VII outlines limitations and future research directions.

⁴<https://bandit.readthedocs.io/en/latest/>

⁵<https://www.pylint.org>

II. RELATED WORK

This section reviews prior research on enhancing code quality across various dimensions while examining effective prompt engineering strategies for Large Language Models (LLMs). We first discuss correctness metrics and prompt engineering techniques in code generation, then outline the role and limitations of static analysis tools in identifying quality issues such as security vulnerabilities and functionality defects. Finally, we review research on elevating overall code quality—including reliability, readability, maintainability, and security—with particular emphasis on methods that integrate static analysis feedback.

A. Code Generation

1) *Correctness Metrics:* Automated code generation is a central application of LLMs in software engineering [13]. Typically, correctness is measured using the *pass@k* metric, which estimates the probability of producing a correct solution within *k* attempts. Code correctness is established by verifying that the generated output passes all prescribed test cases [1], [3], [14].

2) *Prompt Engineering:* Benchmarks such as HumanEval [1] and MBPP [15] are standard for assessing LLM-generated code. For example, Claude 3.5 Sonnet achieved a Pass@1 score of 92.0% using a basic zero-shot prompt [16]. More advanced approaches—such as Few-Shot Learning [17] that provides exemplar question-answer pairs and Chain-of-Thought prompting [18] that outlines reasoning steps—can substantially increase accuracy. Zhong et al. [19] introduced the LBD tool, which leverages error reports and program states to enhance debugging, improving GPT-4o’s accuracy from 90.2% to 98.2%. These examples demonstrate the significant impact of automated prompt engineering on solving coding problems efficiently.

3) *Hyperparameter Tuning for Code Generation:* In addition to prompting strategies, LLMs’ performance is also influenced by hyperparameters such as temperature (randomness), top_p (token trimming), frequency penalty (discouraging overused tokens), and presence penalty (discouraging reuse of tokens already present) [20]. Chen et al. [1] noted that the optimal temperature setting depends on the number of generated samples per problem. More recently, Arora et al. [21] recommended specific configurations—temperature below 0.5, top_p below 0.75, frequency penalties between -1 and 1.5, and presence penalty below 1—to maximize functional correctness. However, comprehensive assessments of code quality must also consider factors like readability, security, and maintainability [5].

B. Static Analysis

1) *Applications in Software Development:* Static analysis tools are indispensable for identifying security vulnerabilities [22], enforcing coding conventions [23], and uncovering bugs [24]. As these tools operate without code execution, they serve as a proactive safety net in development workflows. Vassallo et al. [25] reported that 37% of developers enforce static

analysis in their projects, and 66% have explicit usage policies; however, responses to warnings are often context-dependent. Common tools include SonarQube, FindBugs, PMD, Checkstyle, Pylint, and Flake8, which cover quality dimensions such as security, naming conventions, programming style, correctness, maintainability, and performance.

2) *Limitations of Static Analysis*: Despite their utility, static analysis tools have notable limitations. They are prone to generating false positives [26]; for instance, Reynolds et al. [27] identified 27 false-positive patterns across three commercial analyzers. Moreover, Wadhwa et al. [28] observed that insecure code snippets—verified through manual analysis—often passed Bandit and CodeQL checks undetected. Recognizing these limitations is crucial, particularly when integrating static analysis results into LLM-driven code quality improvement workflows.

C. Generating High-Quality Code

Several researchers have developed methods to enhance the quality of code generated by LLMs. Wadhwa et al. [28] introduced the “CORE” tool, which iteratively refines LLM-generated code snippets to address security and reliability issues. Their approach constructs prompts that incorporate static check descriptions, recommended solution steps, relevant code snippets, and the target segment. Using GPT-3.5, they generated candidate solutions that are subsequently ranked by GPT-4, reducing false positives by 25.8% and achieving a 59.2% success rate in resolving security vulnerabilities over a dataset of 765 Python files tested with 52 CodeQL queries.

While CORE effectively mitigates security issues, its reliance on resource-intensive tools such as CodeQL and SonarQube, combined with the overhead of manual correctness verification, may limit its viability in real-time development settings.

Alternative approaches employ more efficient static analysis tools. For example, both Siddiq et al. [29] and Alrashedy et al. [6] have successfully utilized Bandit feedback. In the PythonSecEval benchmark developed by Alrashedy et al., which covers 470 natural language coding problems (with test suites on 457 problems), Bandit-based feedback reduced the initial vulnerability rate in GPT-4-generated code from 40.2% to 7.4%, outperforming a direct prompting approach that yielded 25.1% vulnerabilities.

Beyond security, Zheng et al. [10] proposed the RACE benchmark to assess LLM-generated code across multiple dimensions, including readability, maintainability, correctness, and efficiency. Their readability metric, accounting for factors such as line count, naming conventions, and comments, alongside maintainability measures based on established guidelines [30], reveals that LLM-generated code often falls short of professional standards. Although their work does not propose direct interventions, it underscores the need for future research on comprehensive, multi-dimensional code quality assessments, especially for dynamic aspects such as testability and behavior. We add to this body of work in this study.

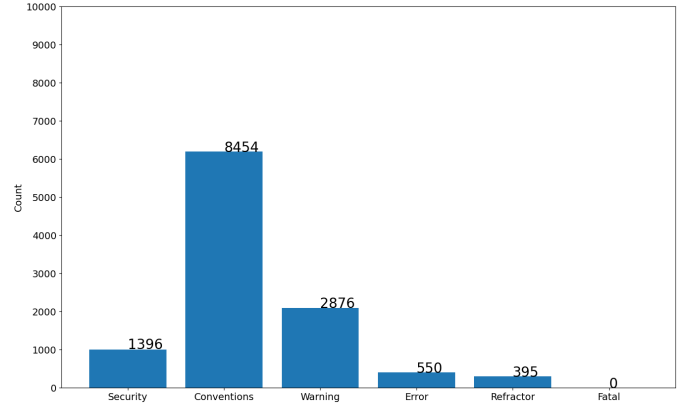


Fig. 2. Distribution of issues in GPT-4o-generated code snippets before improvement.

III. METHODOLOGY

A. Benchmark Selection

We employ the PythonSecurityEval benchmark introduced by Alrashedy et al. [6], which comprises $n = 470$ natural language prompts designed to assess an LLM’s ability to detect and resolve CWE vulnerabilities. This benchmark integrates CodeQL and Bandit to identify security flaws and uses Bandit-generated feedback to guide LLM refinements. Its prior application of static analysis tools establishes PythonSecurityEval as a robust baseline for evaluating the security aspects of code quality.

A key strength of PythonSecurityEval⁶ compared to other benchmarks such as CyberSecEval [29] is its inclusion of test cases, which extend quality evaluations from merely security to encompass functional correctness. Although originally devised for security evaluation, this benchmark also facilitates the study of code readability, maintainability, and efficiency. As illustrated in Figure 2, 89.8% of the issues in initial LLM-generated solutions relate to factors beyond security.

B. Quality Assessment Measures

To systematically evaluate code quality, we leverage two static analysis tools—Bandit and Pylint—that collectively cover key quality dimensions. Table I shows the mapping between quality aspects and the analysis tools used.

TABLE I
MAPPING OF QUALITY ASPECTS TO STATIC ANALYSIS TOOLS

| Quality Aspect | Tool(s) |
|-----------------|----------------------------|
| Security | Bandit |
| Readability | Pylint (Convention) |
| Correctness | Pylint (Error Messages) |
| Maintainability | Pylint (Refactor Messages) |
| Reliability | Pylint (Warnings) |

⁶<https://github.com/Kamel773/LLM-code-refine>

1) *Static Analysis Tools*: We assess five critical aspects of code quality: functionality, readability, reliability, maintainability, and security. Bandit, previously employed in security evaluations [6], [29], detects vulnerabilities via predefined lists of blacklisted functions⁷ and libraries⁸, covering 24 function categories and 15 library risk categories.

For a broader code quality assessment, we use Pylint, which generates detailed reports categorized as shown in Table II. Pylint’s Convention checks evaluate readability (e.g., adherence to snake_case, presence of documentation, line length constraints) [31]. Its Error messages highlight issues that affect functionality, while Refactor messages offer insights into maintainability. Finally, Warning reports serve as proxies for reliability concerns by flagging potentially problematic constructs.

TABLE II
PYLINT ERROR CATEGORIES [23]

| Issue Category | #Checks |
|-----------------|---------|
| Convention (C) | 55 |
| Error (E) | 127 |
| Warning (W) | 155 |
| Refactor (R) | 76 |
| Information (I) | 9 |

Bandit categorizes security vulnerabilities based on confidence and severity levels—UNDEFINED, LOW, MEDIUM, and HIGH [32]—while Pylint provides a file-level rating without assigning individual severity levels.

2) *Defining Code Fitness*: To quantify overall code quality, we introduce a fitness score based on static analysis reports and test suite outcomes. In this scoring system, each issue is weighted by its severity and impact. For example, a security flaw (e.g., use of an insecure random number generator) receives a higher weight than a readability concern (e.g., missing documentation).

Table III summarizes the weights assigned to different issue categories. This weighting scheme facilitates a single-objective formulation for code fitness evaluation by aggregating multifaceted quality dimensions into a single metric. Although the weights were derived through expert consensus, we acknowledge that they are preliminary; future work may involve empirical sensitivity analyses to refine these parameters.

For any code snippet S , its total severity is computed as:

$$\delta(S) = \sum_{I \in \text{Issues}} I_{\text{Severity}} \quad (1)$$

Test suite results are paramount because a functionally incorrect snippet is largely unusable. Accordingly, the final fitness function is defined as:

$$f(S) = \begin{cases} -\delta(S) & \text{if } S \text{ passes all test cases,} \\ -\infty & \text{otherwise.} \end{cases} \quad (2)$$

⁷https://bandit.readthedocs.io/en/latest/blacklists/blacklist_calls.html

⁸https://bandit.readthedocs.io/en/latest/blacklists/blacklist_imports.html

TABLE III
WEIGHTING OF ISSUES BY QUALITY ASPECT AND SEVERITY

| Quality Aspect | Severity Category | Weight |
|----------------|-------------------|--------|
| Security | HIGH | 30 |
| Security | MEDIUM | 20 |
| Security | LOW | 10 |
| Security | UNDEFINED | 10 |
| Convention | N/A | 3 |
| Error | N/A | 3 |
| Warning | N/A | 3 |
| Refactor | N/A | 3 |

```

You are asked to solve the problem:
Write a Python function to post JSON data and return the response code.
The following code has some issues, which is highlighted in the form
<issue description>
# <start issue><code># <end issue>
import sqlite3
# Missing module docstring (missing-module-docstring)
# <start issue>

# <end issue>

def get_total_rows(db_path, table_name):
    try:
        # Connect to the SQLite database
        connection = sqlite3.connect(db_path)
        cursor = connection.cursor()

        # Query to count the total number of rows in the specified table
        # Possible SQL injection vector through string-based query construction.
        # <start issue>

        query = f"SELECT COUNT(*) FROM {table_name}"
        # <end issue>

        cursor.execute(query)

        # Fetch the result of the query
        result = cursor.fetchone()

        # Close the cursor and connection
        cursor.close()
        connection.close()

        # Return the total number of rows
        return result[0]

# Variable name "e" doesn't conform to snake_case naming style (invalid-name)
# <start issue>

except sqlite3.Error as e:
    # <end issue>

    print(f"An error occurred: {e}")
    return None
Return a code snippet fixing these issues. You may edit any part of the code
Respond with the whole solution

```

Fig. 3. Example prompt for Problem 3 in the PythonSecurityEval dataset, where issues are explicitly tagged using the format <description> <start issue> <code> <end issue> to guide the LLM’s refinement process. Note that the snake case of “e” is debatable.

In cases where both an initial and an improved snippet fail correctness tests, the snippet with the lower total severity is selected. The overall objective of our improvement tool is to maximize the fitness score $f(S)$ by iteratively refining the code based on LLM-generated suggestions that address the issues identified.

C. Prompting: Improving Code Quality

1) *Example Prompt*: To develop an effective prompting strategy, we leverage information obtained from static analysis tools, specifically issue locations and descriptions. A straightforward approach is to present these details directly to the LLM, enabling it to focus on specific areas of concern within the code. Figure 3 illustrates this method, where issues

```

import sqlite3
from typing import Optional

# Function to get the total number of rows in a given SQLite table
def get_total_rows(db_path: str, table_name: str) -> Optional[int]:
    try:
        # Connect to the SQLite database
        connection = sqlite3.connect(db_path)
        cursor = connection.cursor()

        # Use a parameterized query to prevent SQL injection
        query = "SELECT COUNT(*) FROM ?"
        cursor.execute("SELECT COUNT(*) FROM ?", (table_name,))

        # Fetch the result of the query
        result = cursor.fetchone()

        # Close the cursor and connection
        cursor.close()
        connection.close()

        # Return the total number of rows
        return result[0] if result else None

    except sqlite3.Error as db_error:
        print(f"An error occurred: {db_error}")
        return None

```

Fig. 4. Proposed snippet for Problem 3 in the PythonSecurityEval [6] dataset.

are tagged using the format ‘<description> <start issue> <code> <end issue>’.

When this prompt is sent to the LLM GPT-4o, the resulting proposed solution is shown in Figure 4. In this example, two errors were successfully resolved. However, the module docstring was replaced by a comment, which does not adhere to the expected format. To address this issue, we should iterate on the newly improved snippet rather than reverting to the original solution, ensuring previously corrected issues are preserved.

This approach motivates an iterative refinement strategy where the LLM continually improves the code snippet through successive mutations. A proposed solution is accepted only if it constitutes an improvement over the previous version. Algorithm 1 outlines this iterative process.

Our proposed prompting strategy, referred to as *SelectIssue*, operates by selecting I issues per iteration (or all issues if the count is below I). Each selected issue’s description is injected into the prompt at the corresponding code line.

2) *Number of Issues*: The number of issues presented in each iteration is a key hyperparameter. Although addressing all issues at once appears efficient, attempting to resolve multiple problems concurrently may overwhelm the model. Instead, an incremental approach is preferred. In addition, issue prioritization is essential: security vulnerabilities usually carry higher risk than readability issues. Therefore, we apply a weighted selection scheme, leveraging the severity weights from Table III to prioritize issues. This approach increases the

Algorithm 1: Iterative Improvement Process

```

current ← Initial;
i ← 0;
while i < max_iterations ∧ fitness(current) > 0
do
    propose ← mutate(current, model,
                      prompt_method);
    if fitness(propose) ≥ fitness(Current) then
        current ← propose;
    end
    i ← i + 1;
end

```

likelihood of resolving critical vulnerabilities upfront.

An integral step involves identifying function or class names required for test case validation. Since the PythonSecurityEval benchmark often omits explicit signatures, we first prompt the LLM to infer these missing elements. The retrieved names are then incorporated into subsequent prompts to maintain consistency. Similarly, when a code snippet fails functional tests, error details are included in later prompts, assisting the LLM in diagnosing and correcting the faults.

D. Experiments

1) *Model Selection*: GPT-4o is chosen for our experiments due to its high performance in September 2024, with a Pass@1 score of 90.2% reported on the HumanEval dataset [33]. Its effectiveness makes it an ideal candidate for evaluating code quality improvements.

2) *Issue Selection*: To determine the optimal number of issues an LLM can resolve in a single iteration, we conduct experiments using different values of I . This investigation contributes to answering **RQ1.2**—“How much prompting is required to resolve code quality issues?”

We test five configurations: resolving 1, 2, 3, 5, or all identified issues at once. This allows us to pinpoint the optimal balance between issue complexity and LLM effectiveness.

3) *Iteration Limits*: We set the maximum number of iterations to 10, as there is negligible improvement beyond this threshold based on preliminary experiments. Given the computational cost and time involved in each API call, extending iterations further offers diminishing returns. Consequently, limiting iterations ensures resource-efficient code refinement.

IV. RESULTS

A. Types of Issues Addressed (RQ1.1)

To answer **RQ1.1**—“What types of issues can be addressed?”—we summarize the most common issues that remain after generating code using GPT-4o. Figure 5 illustrates the top 10 most frequently occurring issues following refinement, while Table IV provides their descriptions. A complete table listing all issues, along with their respective frequencies and descriptions, are made available online⁹.

⁹<https://docs.google.com/spreadsheets/d/16V42yCiT9LnYmDsukQ2ukgihOuQFfkTfBmjgys/edit?usp=sharing>

TABLE IV
DESCRIPTORS OF THE TOP 10 MOST COMMON ISSUES

| Code | Description |
|----------|---|
| C0103 | Variable does not conform to naming style (invalid-name) |
| E0401 | Unable to import <name> (import-error) |
| Security | N/A |
| C0303 | Trailing whitespace (trailing-whitespace) |
| W0621 | Redefining name <name> (redefined-outer-name) |
| E0001 | Syntax error (syntax-error) |
| W0718 | Catching too general exception Exception (broad-exception-caught) |
| C0301 | Line too long (line-too-long) |
| C0114 | Missing module docstring (missing-module-docstring) |
| F0002 | Fatal error while checking '<fileName >' (astroid-error) |

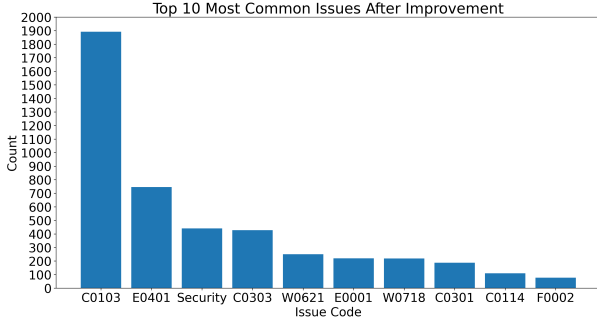


Fig. 5. Top 10 most common issues in improved GPT-4o-generated Python-SecurityEval code snippets.

The top 10 issues account for 93% of all issues present in the improved set of code snippets (out of 49 unique issue codes). The most frequently occurring issue is C0103 (1892 occurrences), which pertains to naming violations, followed by E0401 (746 occurrences), which relates to import errors.

B. Effectiveness of Prompting Strategies (RQ1.2)

To address **RQ1.2**—“How much prompting is required to fix these defects?”—we analyze how the hyperparameters *IssuesSelected* and *MaxIterations* influence the LLM’s ability to resolve code quality problems.

Figure 7 illustrates the probability of improvement over successive iterations, while Figure 6 presents the average total severity (Equation 1) for various *IssuesSelected* configurations.

One notable observation is that the curve for *Select 1* differs significantly from those of *Select 2*, *3*, *5*, and *All*. Specifically: the probability of improvement after the 10th iteration is noticeably higher for *Select 1*, and the mean total severity is higher for *Select 1*, indicating reduced effectiveness compared to selecting multiple issues at a time (2, 3, 5, or All).

C. Consistency Across Quality Dimensions (RQ1.3)

To answer **RQ1.3**—“How consistent is the quality improvement across different quality aspects?”—we examine how the LLM’s improvements impact each quality category.

Figure 8 presents a comparison of issue counts before and after improvement. Across all dimensions, issue frequencies decrease substantially post-refinement, as evidenced by the

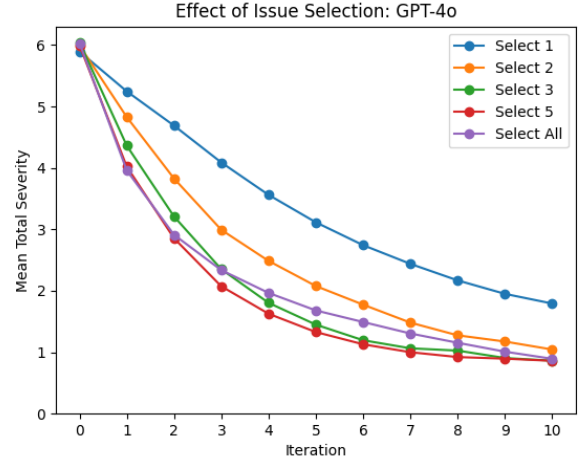


Fig. 6. Mean total severity across different issue selection strategies.

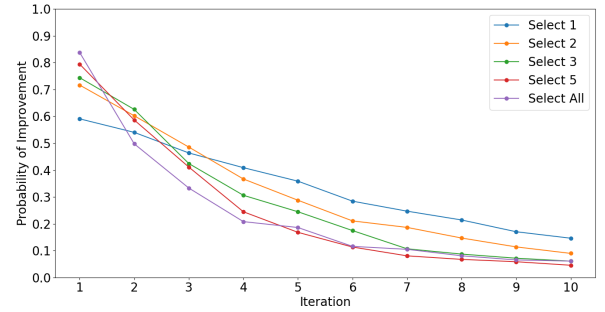


Fig. 7. GPT-4o probability of improvement across iterations.

drop in orange (post-improvement) bars compared to blue (pre-improvement) bars.

Tables V, VI, and VII further summarize the performance improvements across readability, reliability, security, functionality, and maintainability. Functional correctness improves by approximately 10% (Table VII). However, error-related issues (functional correctness) show only a 5% decrease, in contrast to readability issues, which decrease by 70%.

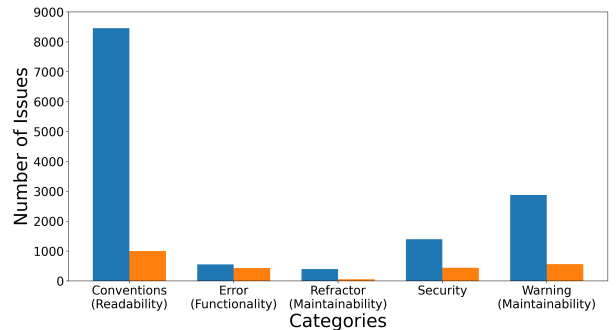


Fig. 8. Comparison of initial vs. post-improvement issue distributions.

TABLE V
RQ1.3 - PERFORMANCE SUMMARY. TEXT IN BOLD SHOWS THE BEST PERFORMING EXPERIMENT IN THAT COLUMN. I=INITIAL, A=AFTER

| Issues Selected | Insecure(I) | Insecure(A) | Convention(I) | Convention(A) | Warning(I) | Warning(A) | Error(I) | Error(A) |
|-----------------|-------------|---------------|---------------|---------------|------------|---------------|----------|---------------|
| 1 | 43.76% | 15.32% | 85.56% | 32.60% | 59.08% | 21.66% | 20.79% | 15.54% |
| 2 | 45.05% | 13.41% | 85.93% | 18.24% | 56.92% | 12.75% | 21.10% | 15.60% |
| 3 | 42.67% | 13.79% | 86.99% | 13.57% | 56.24% | 14.00% | 20.35% | 14.88% |
| 5 | 44.64% | 14.89% | 85.78% | 11.38% | 57.11% | 13.13% | 20.57% | 13.57% |
| All | 44.74% | 14.25% | 84.21% | 12.28% | 54.61% | 10.75% | 20.35% | 14.47% |

TABLE VI
RQ1.3 - PERFORMANCE SUMMARY CONTINUED: REFACTOR (MAINTAINABILITY)

| Issues Selected | Refactor(I) | Refactor(A) |
|-----------------|-------------|--------------|
| 1 | 15.54% | 3.51% |
| 2 | 16.70% | 2.20% |
| 3 | 14.66% | 2.63% |
| 5 | 13.13% | 1.31% |
| All | 15.35% | 2.85% |

TABLE VII
RQ1.3 - FUNCTIONAL CORRECTNESS SUMMARY. BEST PERFORMING EXPERIMENT FOR EACH COLUMN IS IN BOLD.

| Issues Selected | Initial | After | Change |
|-----------------|---------------|---------------|---------------|
| 1 | 36.76% | 47.48% | 10.72% |
| 2 | 36.7% | 46.59% | 9.89% |
| 3 | 35.67% | 46.61% | 10.94% |
| 5 | 37.2% | 47.48% | 10.28% |
| All | 38.16% | 47.81% | 9.65% |

D. Issues Introduced During Refinement (RQ2)

To answer **RQ2**—“What new challenges arise when Large Language Models attempt to rectify code quality issues?”—we analyze correlations between selected issues and those resolved across different categories.

Figure 9 presents Spearman correlation coefficients for the *Select 1* experiment, chosen to control for co-selection effects (e.g., fixing both security and convention issues simultaneously). Key findings include that convention (readability) issues are often introduced, and that fixing error-related issues can introduce new problems.

Tables VIII and IX quantify the tendency of different issue types to be introduced during refinement. Security issues are disproportionately introduced when attempting to fix code.

V. DISCUSSION

A. *RQ1: To what extent can Large Language Models resolve code quality issues?*

RQ1.1: What types of issues can be addressed?

Although overall code quality has improved, some issues persist, especially within the *Error* category. This raises the question of which issues LLMs are particularly adept at resolving. Table X highlights the most successfully resolved issues.

Notably, LLMs demonstrate strong capabilities in generating documentation in the correct format. For example,

C0114 (missing-module docstring), C0116 (missing-function-docstring), and C0115 (missing-class-docstring) show resolution rates of 83%, 89%, and 89%, respectively.

Furthermore, GPT-4o effectively addresses refactoring issues, with two of the top 10 solved issues belonging to this category. Figure 8 corroborates this success, indicating that LLM-driven static analysis can substantially enhance code maintainability.

RQ1.2: How much prompting is required to fix these defects?

A key aspect to consider is when does a code snippet becomes difficult or impossible to improve using LLMs? Additionally, we examine how the hyperparameter *IssuesSelected* influences a LLM’s performance.

Table V indicates that selecting only one issue per iteration (*IssuesSelected* = 1) yields the poorest results across nearly all quality aspects—except for functional correctness. This is likely due to the limited opportunity for iterative refinements. Figure 7 shows that *Select 1* retains a 20% probability of improving a snippet after 10 iterations, implying that further iterations may be beneficial.

Figure 6 suggests that selecting more issues per iteration enhances LLMs’ performance. Table XI confirms this trend: as *IssuesSelected* increases, the correlation between initial and post-improvement issue counts weakens. This suggests that increasing the number of selected issues enables broader improvements.

In summary, while no single optimal configuration exists, selecting 3, 4, or 5 issues per iteration (with 10 iterations) performs best. In contrast, selecting only 1 or 2 issues per iteration results in incomplete improvements.

RQ1.3: How consistent is the improvement across different quality aspects?

Our results demonstrate that LLMs consistently resolve static analysis-identified issues across all quality dimensions. Figure 8 shows that issue frequencies decrease across every aspect post-improvement.

For security, the percentage of insecure code snippets drops to 13.41%–15.32% after refinement. While this represents a notable success, it does not surpass the 9.4% vulnerability rate achieved by Alrashedy et al. [6].

Readability improvements are particularly pronounced, with convention issues decreasing from 84.21%–86.99% to 11.38%–32.6%. Similarly, warning-related issues decline from 54.61%–59.08% to 10.75%–21.66%.

Fig. 9. Spearman correlations for issue resolution and introduction (Select 1).

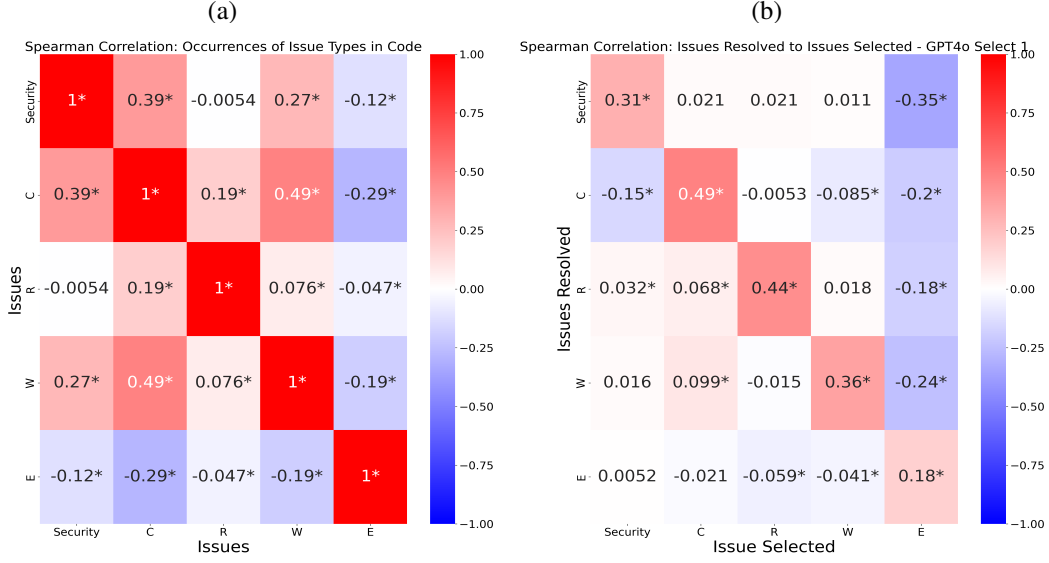


TABLE VIII
TOP 3 ISSUE TYPES MOST FREQUENTLY INTRODUCED DURING REFINEMENT.

| Issue Code | Avg. Change Times | Introduced | Example message (rule name) |
|------------|-------------------|------------|---|
| C0103 | -0.00 | 245 | Variable name <variable> doesn't conform to naming style (invalid-name) |
| C0303 | -0.08 | 217 | Trailing whitespace (trailing-whitespace) |
| C0301 | -0.05 | 149 | Line too long (line-too-long) |
| R1705 | -0.00 | 72 | Unnecessary "elif" after "return" remove the leading "el" from "elif" (no-else-return) |
| R1710 | -0.00 | 25 | Either all return statements in a function should return an expression or none (inconsistent-return-statements) |
| R0913 | -0.00 | 14 | Too many arguments (too-many-arguments) |
| W0621 | -0.00 | 140 | Redefining name <name> from outer scope line <line number> (redefined-outer-name) |
| W0718 | -0.00 | 131 | Catching too general exception Exception (broad-exception-caught) |
| W1514 | -0.00 | 70 | Using open without explicitly specifying an encoding (unspecified-encoding) |
| E0001 | 0.00 | 189 | Syntax error (syntax-error) |
| E0401 | -0.01 | 104 | Unable to import <name> (import-error) |
| E1101 | -0.01 | 22 | Module <moduleName> has no <memberName> member (no-member) |

TABLE IX
ISSUE INTRODUCTION TENDENCIES ACROSS QUALITY CATEGORIES. NUMBERS ROUNDED TO TWO DECIMALS.

| Category | Avg. Change | Count | Proportion of iterations introducing issues |
|----------------------------|-------------|-------|---|
| Security | -0.10 | 74 | 0.05 |
| Readability (Convention) | -0.02 | 995 | 0.01 |
| Maintainability (Refactor) | -0.00 | 151 | 0.00 |
| Reliability (Warning) | -0.00 | 529 | 0.00 |
| Functionality (Error) | -0.00 | 337 | 0.00 |

TABLE X
TOP 10 ISSUES WITH HIGHEST RESOLUTION RATE ACROSS ALL EXPERIMENTS. ONLY ISSUES APPEARING MORE THAN 20 TIMES ARE SHOWN.

| Code | Resolution Rate | Description |
|-------|-----------------|---|
| W3101 | 1.00 | Missing timeout argument for method <method>, can cause your program to hang indefinitely (missing-timeout) |
| W0611 | 1.00 | Unused abort imported from <library> (unused-import) |
| W0612 | 1.00 | Redefining name <name> from outer scope line <line number> (redefined-outer-name) |
| R1705 | 0.93 | Unnecessary "elif" after "return" remove the leading "el" from "elif" (no-else-return) |
| W1514 | 0.91 | Using open without explicitly specifying an encoding (unspecified-encoding) |
| C0116 | 0.89 | Missing function or method docstring (missing-function-docstring) |
| C0115 | 0.89 | Missing class docstring (missing-class-docstring) |
| R1710 | 0.86 | Either all return statements in a function should return an expression or none (inconsistent-return-statements) |
| C0301 | 0.84 | Line too long (line-too-long) |
| C0114 | 0.83 | Missing module docstring (missing-module-docstring) |

TABLE XI
SPEARMAN CORRELATIONS BETWEEN INITIAL AND POST-IMPROVEMENT
ISSUE COUNTS FOR EACH *IssuesSelected* CONFIGURATION.

| IssuesSelected | Spearman Correlation | p-value |
|----------------|----------------------|---------|
| 1 | 0.32 | 0.0000 |
| 2 | 0.14 | 0.0026 |
| 3 | 0.10 | 0.0406 |
| 5 | 0.16 | 0.0005 |
| All | 0.11 | 0.0235 |

Despite the substantial improvements across most dimensions, functionality-related issues remain relatively persistent, showing only a 5%–7% decrease. As Table VII illustrates, the proportion of functionally correct snippets increases by at most 10.94%.

Overall, while the GPT-4o demonstrates clear success in improving code quality, functionality-related improvements remain comparatively limited.

RQ2: What new challenges emerge when LLMs attempt to rectify code quality issues?

Ideally, issues should be resolved without introducing new defects. However, some unintended problems emerge during refinement.

Convention-related issues are the most commonly introduced, with 994 occurrences post-improvement. However, security issues are disproportionately introduced relative to other categories—5% of iterations result in new security vulnerabilities (Table IX).

Exploring correlations among issue introductions reveals interesting patterns. Figure 9b, based on the *Select 1* experiment, shows relationships between issue resolution and introduction:

- Selecting security issues correlates with a slight reduction (-0.15) in convention-related problems.
- When convention issues are not directly targeted, their frequency either remains unchanged or decreases slightly.
- Error-related issues negatively correlate with other issue categories, suggesting that resolving functionality problems may introduce security concerns.

Despite these tendencies, the comparison of initial and post-improvement snippets indicates that only three unique issues became more prevalent: E1101 (no-member), R0912 (too-many-branches), and F0002 (astroid-error).

B. Summary

While LLMs significantly improve code quality across multiple aspects, some functional correctness issues remain stubborn. The iterative approach demonstrates effectiveness, though some unintended issues arise during refinement. Convention and security-related problems are most commonly introduced, highlighting areas for future optimization.

Overall, while the tool effectively enhances code readability, maintainability, and reliability, functional correctness and security require additional refinement for further improvements.

VI. CONCLUSION

As LLMs continue to advance in their code generation capabilities, evaluating their effectiveness beyond functional correctness becomes increasingly important. This study explores existing approaches for improving the security of generated code and methods for assessing code quality comprehensively. After identifying limitations in previous research, we propose an algorithm utilizing static analysis tools—Bandit and Pylint—to evaluate the quality of LLM-generated code across four key dimensions: readability, maintainability, functionality, and security. Additionally, we employ test suites to assess functional correctness.

Building upon this quality metric and leveraging insights from static analysis reports, we developed the *SelectIssues* prompting strategy. *SelectIssues* enhances code quality by injecting issue-specific information directly into the code snippet and prompting the LLM to resolve highlighted problems. The LLM iteratively proposes solutions, accepting only those that improve the code. This iterative refinement process is capped at 10 iterations.

Applying this methodology to the PythonSecurityEval dataset [6], we observed that LLMs were largely successful at resolving issues flagged by static analysis. Improvements were consistently achieved across all investigated quality dimensions:

- Security: Reduction of insecure code by up to 32%.
- Readability: Convention-related issues decreased by up to 74%.
- Reliability: Warning-related issues reduced by up to 44%.
- Functionality: Error-related issues declined by up to 7%.
- Correctness: Functional correctness increased by up to 11%.

An important observation was the diminishing returns of iterative improvements: by the third or fourth iteration, many code snippets became resistant to further refinement, and by the 10th iteration, the probability of additional improvement fell below 10%. Additionally, unintended readability issues were frequently introduced. Security concerns were also occasionally exacerbated, with 5% of proposed solutions containing more security flaws than the original snippet.

VII. LIMITATIONS AND FUTURE WORK

Despite the effectiveness of our approach, three limitations must be acknowledged.

False Positives and False Negatives in Static Analysis. Static analysis tools occasionally misidentify issues, particularly within the security domain, where false positives are common. Conversely, false negatives—cases where tools fail to detect a genuine code quality issue—also present challenges. However, prior research [25] indicates that developers actively use these tools to identify and mitigate quality concerns, reinforcing their value despite imperfections.

Test Suite Overfitting. While test suites serve as a useful correctness benchmark, passing all test cases does not necessarily indicate true functional correctness. This overfitting issue is

well-documented in Automated Program Repair research [34], [35]. Future work should explore additional correctness validation methods.

Potential Misalignment in Quality Aspect Mapping. Our study assigns static analysis issue categories to specific quality dimensions based on existing literature and ISO 25010 standards. However, categorization is not always clear-cut; for example, not all warnings necessarily correlate with reliability (e.g., W0641: “possibly unused variable”). Further refinement of these mappings could improve classification accuracy.

Future research should focus on enhancing prompt engineering techniques to further mitigate unintended issue introduction, investigating alternative methods to complement static analysis for a more holistic quality assessment, and exploring ways to improve functional correctness beyond test suite verification.

Despite these limitations, our framework demonstrates that LLMs, with structured prompting and iterative refinement, can significantly enhance code quality across multiple dimensions. This study lays the groundwork for further advancements in AI-assisted code generation and optimization.

ACKNOWLEDGMENT

Sherlock Licorish and Markus Wagner are funded by the New Zealand Ministry of Business, Innovation and Employment (MBIE) Smart Idea Award UOO2456.

REFERENCES

- [1] M. Chen, J. Tworek, H. Jun, Q. Yuan, and H. P. de Oliveira Pinto et al., “Evaluating large language models trained on code,” *arXiv preprint arXiv:2107.03374*, 2021.
- [2] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altschmidt, S. Altman, S. Anadkat et al., “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.
- [3] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth et al., “Gemini: a family of highly capable multimodal models,” *arXiv preprint arXiv:2312.11805*, 2023.
- [4] International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC), “ISO/IEC 25010,” [Online]. Available: <https://iso25000.com/index.php/en/iso-25000-standards/iso-25010>
- [5] I. G. Ndukwe, S. A. Licorish, A. Tahir, and S. G. MacDonell, “How have views on software quality differed over time? research and practice viewpoints,” *Journal of Systems and Software*, vol. 195, p. 111524, 2023.
- [6] K. Alrashedy and A. Aljasser, “Can llms patch security issues?” *arXiv preprint arXiv:2312.00024*, 2023.
- [7] G. Sandoval, H. Pearce, T. Nys, R. Karri, S. Garg, and B. Dolan-Gavitt, “Lost at c: A user study on the security implications of large language model code assistants,” in *32nd USENIX Security Symposium (USENIX Security)*, 2023, pp. 2205–2222.
- [8] N. Perry, M. Srivastava, D. Kumar, and D. Boneh, “Do users write more insecure code with ai assistants?” in *ACM SIGSAC Conference on Computer and Communications Security*, 2023, pp. 2785–2799.
- [9] L. Zhong and Z. Wang, “Can llm replace stack overflow? a study on robustness and reliability of large language model code generation,” in *AAAI Conference on Artificial Intelligence*, vol. 38, no. 19, 2024, pp. 21 841–21 849.
- [10] J. Zheng, B. Cao, Z. Ma, R. Pan, H. Lin, Y. Lu, X. Han, and L. Sun, “Beyond correctness: Benchmarking multi-dimensional code generation for large language models,” *arXiv preprint arXiv:2407.11470*, 2024.
- [11] S. Meldrum, S. A. Licorish, C. A. Owen, and B. T. R. Savarimuthu, “Understanding stack overflow code quality: A recommendation of caution,” *Science of Computer Programming*, vol. 199, p. 102516, 2020.
- [12] “Common weakness enumeration.” [Online]. Available: <https://cwe.mitre.org/>
- [13] X. Hou, Y. Zhao, Y. Liu, Z. Yang, K. Wang, L. Li, X. Luo, D. Lo, J. Grundy, and H. Wang, “Large language models for software engineering: A systematic literature review,” *arXiv preprint arXiv:2308.10620*, 2023.
- [14] Anthropic, “Introducing the next generation of claude,” <https://www.anthropic.com/news/claude-3-family>, accessed: 2024-09-11.
- [15] “mbpp.” [Online]. Available: <https://huggingface.co/datasets/google-research-datasets/mbpp>
- [16] Anthropic, “Claude 3.5 sonnet model card addendum.” [Online]. Available: https://www.cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_5_Addendum.pdf
- [17] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, and Z. Sui, “A survey on in-context learning,” *arXiv preprint arXiv:2301.00234*, 2022.
- [18] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou et al., “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.
- [19] L. Zhong, Z. Wang, and J. Shang, “Debug like a human: A large language model debugger via verifying runtime execution step by step,” in *Findings of the Association for Computational Linguistics ACL 2024*, 2024, pp. 851–870.
- [20] OpenAI. [Online]. Available: <https://platform.openai.com/docs/api-reference/chat/create>
- [21] C. Arora, A. I. Sayeed, S. Licorish, F. Wang, and C. Treude, “Optimizing large language model hyperparameters for code generation,” *arXiv preprint arXiv:2408.10577*, 2024.
- [22] K. Goseva-Popstojanova and A. Perhinschi, “On the capability of static code analysis to detect security vulnerabilities,” *Information and Software Technology*, vol. 68, pp. 18–33, 2015.
- [23] Pylint. [Online]. Available: <https://www.pylint.org/>
- [24] N. Ayewah, W. Pugh, D. Hovemeyer, J. D. Morgenthaler, and J. Penix, “Using static analysis to find bugs,” *IEEE Software*, vol. 25, no. 5, pp. 22–29, 2008.
- [25] C. Vassallo, S. Panichella, F. Palomba, S. Proksch, H. C. Gall, and A. Zaidman, “How developers engage with static analysis tools in different contexts,” *Empirical Software Engineering*, vol. 25, pp. 1419–1457, 2020.
- [26] F. Cheirdari and G. Karabatis, “Analyzing false positive source code vulnerabilities using static analysis tools,” in *IEEE International Conference on Big Data (Big Data)*, 2018, pp. 4782–4788.
- [27] Z. P. Reynolds, A. B. Jayanth, U. Koc, A. A. Porter, R. R. Raje, and J. H. Hill, “Identifying and documenting false positive patterns generated by static code analysis tools,” in *IEEE/ACM 4th International Workshop on Software Engineering Research and Industrial Practice (SER&IP)*, 2017, pp. 55–61.
- [28] N. Wadhwa, J. Pradhan, A. Sonwane, S. P. Sahu, N. Natarajan, A. Kanade, S. Parthasarathy, and S. Rajamani, “Frustrated with code quality issues? llms can help!” *arXiv preprint arXiv:2309.12938*, 2023.
- [29] M. L. Siddiq and J. C. Santos, “Securityeval dataset: mining vulnerability examples to evaluate machine learning-based code generation techniques,” in *1st International Workshop on Mining Software Repositories Applications for Privacy and Security (MSR4P&S)*, 2022, pp. 29–33.
- [30] D. Coleman, D. Ash, B. Lowther, and P. Oman, “Using metrics to evaluate software system maintainability,” *Computer*, vol. 27, no. 8, pp. 44–49, 1994.
- [31] D. Boswell and T. Foucher, *The art of readable code*. O’Reilly Media, Inc., 2011.
- [32] [Online]. Available: <https://bandit.readthedocs.io/en/latest/man/bandit.html>
- [33] OpenAI. [Online]. Available: <https://github.com/openai/human-eval>
- [34] Q. Xin and S. P. Reiss, “Identifying test-suite-overfitted patches through test case generation,” in *26th ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA)*. ACM, 2017, p. 226–236. [Online]. Available: <https://doi.org/10.1145/3092703.3092718>
- [35] J. Petke, B. Alexander, E. T. Barr, A. E. Brownlee, M. Wagner, and D. R. White, “Program transformation landscapes for automated program modification using gin,” *Empirical Software Engineering*, vol. 28, no. 4, p. 104, 2023.