

CLEANVUL: Toward High-Quality Function-Level Vulnerability Datasets via LLM-Based Noise Reduction

YIKUN LI, Singapore Management University, Singapore
 TING ZHANG, Singapore Management University, Singapore
 RATNADIRA WIDYASARI, Singapore Management University, Singapore
 YAN NAING TUN, Singapore Management University, Singapore
 HUU HUNG NGUYEN, Singapore Management University, Singapore
 TAN BUI, Singapore Management University, Singapore
 IVANA CLAIRINE IRSAN, Singapore Management University, Singapore
 YIRAN CHENG, Singapore Management University, Singapore
 XIANG LAN, North Carolina State University, United States
 HAN WEI ANG, GovTech, Singapore
 FRANK LIAUW, GovTech, Singapore
 MARTIN WEYSSOW, Singapore Management University, Singapore
 HONG JIN KANG, Singapore Management University, Singapore
 ENG LIEH OUH, Singapore Management University, Singapore
 LWIN KHIN SHAR, Singapore Management University, Singapore
 DAVID LO, Singapore Management University, Singapore

In the dynamic field of cybersecurity, the accurate identification and mitigation of software vulnerabilities are paramount for maintaining system integrity. Vulnerability datasets, often derived from the National Vulnerability Database (NVD) or directly from GitHub, are essential for training machine learning models to detect and address these security flaws. However, these datasets frequently suffer from significant noise, typically 40% to 75% [4, 5], due primarily to the automatic and indiscriminate labeling of all modifications in vulnerability-fixing commits (VFCs) as vulnerability-related. This misclassification occurs because not all changes in a commit aimed at fixing vulnerabilities pertain to security threats; many are routine updates like bug fixes, test improvements, or unrelated code refactoring.

To address these challenges, this paper introduces the **first methodology** that leverages the Large Language Model (LLM) with a heuristic enhancement to automatically identify vulnerability-fixing changes from VFCs, achieving an F1-score of 0.82. VULSIFTER was applied to a large-scale study, where we conducted a

Authors' Contact Information: Yikun Li, yikunli@smu.edu.sg, Singapore Management University, Singapore, Singapore; Ting Zhang, Singapore Management University, Singapore, Singapore; Ratnadira Widyasari, Singapore Management University, Singapore, Singapore; Yan Naing Tun, Singapore Management University, Singapore, Singapore; Huu Hung Nguyen, Singapore Management University, Singapore, Singapore; Tan Bui, Singapore Management University, Singapore, Singapore; Ivana Clairine Irsan, Singapore Management University, Singapore, Singapore; Yiran Cheng, Singapore Management University, Singapore, Singapore; Xiang Lan, North Carolina State University, Raleigh, North Carolina, United States; Han Wei Ang, GovTech, Singapore, Singapore; Frank Liauw, GovTech, Singapore, Singapore; Martin Weyssow, Singapore Management University, Singapore, Singapore; Hong Jin Kang, Singapore Management University, Singapore, Singapore; Eng Lieh Ouh, elouh@smu.edu.sg, Singapore Management University, Singapore, Singapore; Lwin Khin Shar, lkshar@smu.edu.sg, Singapore Management University, Singapore, Singapore; David Lo, davidlo@smu.edu.sg, Singapore Management University, Singapore, Singapore.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM 1557-735X/2018/8-ART111

<https://doi.org/XXXXXXX.XXXXXXX>

comprehensive crawl of 127,063 repositories on GitHub, resulting in the acquisition of 5,352,105 commits. The LLM with heuristic enhancement approach involves utilizing an LLM to comprehend code semantics and contextual information, while applying heuristics to filter out unrelated changes. We developed CLEANVUL, a high-quality dataset comprising 8,203 functions using our LLM heuristic enhancement approach, demonstrating *Correctness* (90.6%) comparable to established datasets such as SVEN (94.0%) [10] and PrimeVul (86.0%) [5]. In this context, *Correctness* is defined as the percentage of genuine vulnerable functions in the vulnerability dataset, which helps assess the effectiveness of VULSIFTER in identifying and filtering vulnerability-fixing changes. To evaluate the effectiveness of our CLEANVUL dataset, we conducted experiments focusing on fine-tuning various LLMs on the CLEANVUL dataset and other high-quality datasets. Our evaluation was conducted on multiple popular programming languages, including Java, Python, JavaScript, C#, C, and C++. Evaluation results reveal that LLMs fine-tuned on CLEANVUL not only exhibit enhanced accuracy but also superior generalization capabilities compared to those trained on uncleaned datasets. Specifically, models trained on CLEANVUL and tested on PrimeVul achieve accuracy higher than those trained and tested exclusively on PrimeVul, validating the effectiveness of our methodology.

CCS Concepts: • **Security and privacy** → **Software security engineering**.

Additional Key Words and Phrases: Empirical Study, Large Language Models, Vulnerability Detection

ACM Reference Format:

Yikun Li, Ting Zhang, Ratnadira Widyasari, Yan Naing Tun, Huu Hung Nguyen, Tan Bui, Ivana Clairine Irsan, Yiran Cheng, Xiang Lan, Han Wei Ang, Frank Liauw, Martin Weyssow, Hong Jin Kang, Eng Lieh Ouh, Lwin Khin Shar, and David Lo. 2018. CLEANVUL: Toward High-Quality Function-Level Vulnerability Datasets via LLM-Based Noise Reduction. *J. ACM* 37, 4, Article 111 (August 2018), 25 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

In the rapidly evolving field of cybersecurity, accurately detecting and mitigating software vulnerabilities is crucial for safeguarding digital assets and infrastructure. These vulnerabilities pose significant security risks, underscoring the need for robust tools to identify and rectify them. Vulnerability datasets serve as crucial resources for training machine learning (ML) models to identify and address these security flaws [3, 4, 8, 23]. These datasets are typically derived from known vulnerabilities cataloged in databases like the National Vulnerability Database (NVD) or mined from software repositories such as GitHub. By leveraging these datasets, ML models can learn patterns and indicators of security vulnerabilities, thereby enhancing automated vulnerability detection systems.

However, existing vulnerability datasets [1, 4, 6, 17, 18] often contain significant noise, ranging from 40% to 75%, which substantially reduces the training effectiveness of ML models [4, 5]. This issue arises because many datasets automatically label all modifications in vulnerability-fixing commits (VFCs) as related to vulnerabilities, failing to recognize that not all changes pertain to security threats. Frequently, a commit intended to address a specific vulnerability might also include unrelated code adjustments. This leads to the erroneous classification of such benign modifications as security threats. A detailed example illustrating this challenge is provided in Section 2.

To address these shortcomings, several attempts have been made to enhance dataset accuracy by correlating function names from commit logs with the descriptions found in the NVD [5]. Specifically, a function is labeled as vulnerable if it is explicitly mentioned in the NVD description or if it is the only function changed in the file mentioned by the NVD. However, this method falls short when applied to VFCs that lack corresponding NVD entries, a common scenario for most of VFCs identified on GitHub. Consequently, there is a critical need to develop robust, automated techniques capable of discerning genuine vulnerability-fixing changes across the entirety of VFCs, regardless of their NVD linkage.

Our Solution This paper introduces the first methodology to address the gap in accurately identifying vulnerability-related changes within VFCs. We first examine the reasons why many changes cataloged as vulnerability fixes do not pertain to actual security vulnerabilities. Our empirical study reveals that approximately 80% of non-vulnerability changes consist of test-related modifications and general bug fixes, with additional changes spanning support updates, code refactoring, and documentation updates. Building on this analysis, we developed **VULSIFTER**, a novel approach that combines LLM with heuristic filtering to automatically analyze VFCs and eliminate noise, resulting in a refined vulnerability dataset called **CLEANVUL**. By applying VULSIFTER, the *Correctness* of genuine vulnerability fixes in the dataset improved from 28.7% to 90.6%. *Correctness* is defined as the percentage of genuine vulnerable functions in the vulnerability dataset, and it helps assess the effectiveness of VULSIFTER in identifying and filtering vulnerability-fixing changes. CLEANVUL comprises 8,203 functions, both vulnerable and benign, and achieves a level of *Correctness* (90.6%) comparable to established datasets such as SVEN [10] and PrimeVul [5]. For comparison, the manually curated SVEN dataset contains 803 functions with 94.0% *Correctness*, while PrimeVul, which identifies vulnerabilities by matching function names with NVD descriptions, contains 6,968 NVD-sourced functions with 86.0% *Correctness*. As the first automated approach to improve dataset *Correctness* without requiring NVD linkage or other constraints, CLEANVUL represents a significant advancement in function-level vulnerability detection. It complements PrimeVul’s NVD-based approach by automatically filtering noise from GitHub VFCs to obtain high-quality vulnerability datasets.

Effectiveness of CLEANVUL We evaluated our dataset CLEANVUL by comparing it with two established high-quality vulnerability detection datasets: SVEN and PrimeVul. Since our dataset consists of balanced vulnerability pairs (vulnerable/benign), we use accuracy as our primary evaluation metric rather than F1-score, as accuracy better reflects model performance when classes are balanced - a random baseline would achieve 50% accuracy. Through experiments with various LLMs, we found that models fine-tuned on CLEANVUL showed strong generalization capabilities. When testing on PrimeVul, our CLEANVUL-trained models achieved 58.09% accuracy (using CodeBERT), exceeding PrimeVul’s own intra-dataset performance of 56.61%. When testing on SVEN, CLEANVUL-trained models reached 64.87% accuracy, outperforming models fine-tuned on PrimeVul which only achieved 55.75% accuracy on SVEN. Additionally, CLEANVUL demonstrated robust intra-dataset performance with 68.96% accuracy when trained and tested on itself. These results suggest that CLEANVUL captures a more diverse and representative range of vulnerabilities compared to existing datasets, making it particularly effective for training vulnerability detection models.

Contributions Overall, we make the following contributions:

- **Characterization of Code Change Categories in VFCs:** We conducted a manual analysis to categorize non-vulnerability-related changes within VFC datasets. We also developed a refined taxonomy capturing the nuances of function-level changes, revealing that the predominant non-vulnerability changes were test-related (41.2%) and bug fixes (38.2%).
- **VULSIFTER - LLM Heuristic for Identifying Vulnerability Fixes in VFCs:** We proposed and validated the first approach to automatically identify function-level vulnerability-fixing changes, namely VULSIFTER. We then demonstrated that LLM heuristic is particularly effective when analyzing function changes combined with commit messages and additional context, with GPT-4 achieving the highest F1-score of 0.82.
- **CLEANVUL - A Large-Scale, High-Quality Vulnerability Dataset:** We introduced CLEANVUL, a new high-quality dataset derived from the application of the heuristic approach,

containing 8,203 functions categorized as vulnerable and benign. We also showed that CLEANVUL maintains *Correctness* level (percentage of genuine vulnerable functions) comparable to existing high-quality datasets like SVEN and PrimeVul, enhancing the scale and reliability of data available for vulnerability detection research.

- **Evaluation and Comparison of Dataset Effectiveness:** We assessed the performance of LLMs trained on CLEANVUL and compared it to performances on an uncleaned dataset and other established datasets like SVEN and PrimeVul. We then confirmed that training on CLEANVUL improves model accuracy in comparison to the uncleaned dataset with a large margin, and highlighted the superior generalization capabilities of models trained on CLEANVUL when tested on external datasets.

In the spirit of open science, we make our source code and dataset publicly available¹.

Paper Structure The remainder of the paper is organized as follows. Section 2 presents motivating examples for our work. Section 3 describes the VulSifter approach, while Section 4 explains the CLEANVUL dataset curation process. Sections 5 and 6 present our evaluation settings and experiments respectively. Section 7 provides additional analyses, including sensitivity analysis and ablation study, followed by threats to validity in Section 8. Section 9 covers related work, and Section 10 concludes our paper.

2 Motivating Example

Consider the following real-world example² from the ThingsBoard project, which illustrates the complexity of identifying vulnerability fixes within commits. The commit message is presented below, targeting XSS vulnerabilities. This commit exemplifies what researchers call *tangled commits* [12] - commits that address multiple concerns or include changes serving different purposes simultaneously. While the primary purpose is introducing security measures such as NoXss validation to prevent XSS attacks, the commit also encompasses unrelated modifications including enhancing the codebase by removing unnecessary imports and updating license documentation, as well as improving test coverage to confirm the efficacy of the new validations.

```
Fixed xss vulnerabilities in attributes and telemetry (#8238)
* added noxss validation on kventries
* added ConstraintValidator usages for validation
* fixed licence
* added test
* removed redundant imports
```

This example underscores the challenge in analyzing commits categorized as VFCs. Assuming all modifications in such commits are responses to vulnerabilities would inaccurately label updates like documentation revisions, code cleanup, and test enhancements as security measures. This misclassification could distort the perceived security posture of the codebase and affect the effectiveness of training data for machine learning models. Therefore, it is essential to develop automated tools capable of discerning which changes within a VFC directly address security issues and which do not. By accurately segregating these changes, we can refine datasets to include only genuine vulnerability fixes, thereby enhancing the accuracy and reliability of vulnerability detection and analysis tools.

Empirical Study Building on this motivating example, we conducted an empirical study to better understand the types of changes commonly found in VFCs. Previous research constructing vulnerability datasets often identifies VFCs from the NVD dataset and treats all code changes within

¹<https://github.com/yikun-li/CleanVul>

²<https://github.com/thingsboard/thingsboard-edge/commit/1a3ee8512d58625940b25e46bc6488a3539fdc5e>

a commit as vulnerability-fixing changes [5]. However, as found by PrimeVul [5], this approach is inaccurate because many commits are complex and include changes unrelated to vulnerability fixes; some commits are monolithic and encompass various types of changes. Since no prior study has deeply analyzed the changes within VFCs, we aim to characterize these changes into different categories. This characterization will: (1) help us better understand why some changes are not related to vulnerability fixes; and (2) allow us to develop heuristic-based approaches to accurately identify vulnerability-fixing changes.

We conducted a manual analysis of 136 instances where non-vulnerability-related changes occurred within the VFC datasets. Our initial classification categories were derived from previous research [13]. We employed an open card sorting process to systematically analyze and categorize the instances. Throughout our analysis, we adapted and expanded these categories as necessary, ultimately developing a refined taxonomy that captures the nuances of function-level changes not directly related to vulnerability fixes. In the following sections, we present several reasons for these changes, accompanied by specific examples to illustrate our findings.

Example I: Test-Related Changes In some cases, developers introduce changes to the testing code when addressing vulnerabilities. These changes, while related to the vulnerability, do not directly alter the vulnerability’s resolution but instead aim to verify the fix’s effectiveness. For instance, in the Elasticsearch project (commit da3428), developers encountered a potential infinite loop issue when the span setting was too close to the length of the context part and the context ended in a word that tokenized to more than one token. To address this, they not only modified the core algorithm to prevent the infinite loop but also added a test case to ensure the issue was resolved:

```
+ public void testDetectInfiniteLoop() {
+     // These settings are known to produce an infinite loop.
+     // question and context are longer than max sequence length
+     // so the input must be spanned. With a span setting of 4
+     // there is only 1 more token that can go into the context part:
+     ...
+     var e = expectThrows(IllegalStateException.class, () -> tokenizer.tokenize(question,
+     context, Tokenization.Truncate.NONE, span, 0));
+     assertThat(e.getMessage(), containsString("Tokenization cannot be satisfied with the
+     current span setting"));
+ }
```

Example II: Bug Fixes Our analysis reveals a significant number of changes in VFCs from GitHub that were initially classified as vulnerability fixes but were actually bug fixes or feature enhancements. This discrepancy is primarily due to the use of automated keyword matching techniques [2] that are not always accurate and can generate false positives. Moreover, the inherent nature of commit messages themselves often limits the effectiveness of these detection methods. Many commit messages are vague and lack sufficient information to conclusively determine whether a change is related to vulnerability fixes. This ambiguity necessitates further examination, such as reviewing the actual code changes or consulting issue tracking descriptions, to accurately classify a commit. For instance, the commit message below was flagged as a VFC, but further investigation revealed it was related to gameplay features in a computer game, not a security vulnerability:

“Removed the attack peaceful towns validation check WHEN the peaceful feature is off” - [SiegeWar-1fbba5]

The presence of *attack* as a keyword in VFC identification algorithms led to its incorrect classification. The code change associated with this commit further illustrates this point:

```
- if (defendingTown.isNeutral())
+ if (SiegeWarSettings.getWarCommonPeacefulTownsEnabled() && defendingTown.isNeutral())
```

```
throw new TownyException(Translation.of("msg_war_siege_err_cannot_attack_peaceful_town")
);
```

Example III: Support Changes In certain instances, developers not only fix vulnerabilities but also modify other sections of code to support these fixes, which may involve adjusting to updated dependencies or configurations. This can be seen in the following commit message, where dependencies were updated to address specific CVEs, necessitating changes in the use of affected libraries or functions:

"Update of direct dependencies: kubernetes java-client to 19.0.0, docker-java-bom to 3.3.4. To address CVES: CVE-2023-3635 in okio, CVE-2023-33201 in bcjava" - [Druid-3c7dec]

This commit indicates that the modifications were targeted at fixing vulnerabilities CVE-2023-3635 and CVE-2023-33201 through updates to two key dependencies. The updates to these packages necessitated adjustments in the code where these dependencies are used, to ensure compatibility with new versions and continued secure operation. The following code snippet illustrates the changes made to the function calls, adapting to the updated interface of the dependencies:

```
Watch.createWatch(
    realK8sClient,
    coreV1Api.listNamespacedPodCall(namespace, null, true, null, null,
-   labelSelector, null, lastKnownResourceVersion, null, 0, true, null
+   labelSelector, null, lastKnownResourceVersion, null, null, 0, true, null
),
    new TypeReference<Watch.Response<V1Pod>>()
    {
```

Example IV: Code Refactoring During the process of addressing vulnerabilities, developers sometimes refactor the code to enhance its readability, maintainability, or extensibility. For example, in the Keycloak project (commit 15a21b), in addition to patching a vulnerability that could allow unauthorized access to user data, developers also refactored part of the code to improve its maintainability:

```
if (validRedirect.startsWith("/")) {
    validRedirect = relativeToAbsoluteURI(session, rootUrl, validRedirect);
    logger.debugv("replacing relative valid redirect with: {0}", validRedirect);
-   resolveValidRedirects.add(validRedirect);
- } else {
-   resolveValidRedirects.add(validRedirect);
- }
+ resolveValidRedirects.add(validRedirect);
```

Example V: Documentation Updates While addressing vulnerabilities, developers also take the opportunity to update documentation, enhancing the clarity and security of the code. For example, during enhancements to XML processing security to mitigate XXE attacks, javadocs were also updated:

```
/**
 * Configures a {@link DocumentBuilderFactory} to protect it against XML
 * External Entity attacks.
+ *
+ * @param factory the factory
 * @see <a href="https://www.owasp.org/index.php/XML_External_Entity_%
```

The results are summarized in Table 1. The data show that *Test-Related Changes* and *Bug Fixes* are the most prevalent categories of non-vulnerability changes, representing 41.2% and 38.2% of the cases, respectively. These two categories together account for nearly 80% of all non-vulnerability changes in the analyzed VFCs, indicating a significant focus on functionality enhancement and reliability testing in software updates. Additionally, *Support Changes* and *Code Refactoring* are

Table 1. Reasons for Non-Vulnerability Changes in Identified VFCs

Type	Number (#)	Percentage (%)
Test-Related Changes	56	41.2
Bug Fixes	52	38.2
Support Changes	20	14.7
Code Refactoring	7	5.1
Documentation Updates	1	0.7

observed to constitute 14.7% and 5.1% of the changes, respectively, reflecting a lesser but noteworthy commitment to adapting existing systems and improving code quality. The smallest category, *Documentation Updates*, makes up 0.7% of the changes, suggesting minimal alterations in documentation alongside other code changes.

3 VULSIFTER: Approach

The goal of this study is to analyze source code and commit messages from VFCs for the purpose of automatically identifying function-level vulnerability-fixing changes. The overview of our approach is demonstrated in Figure 1. The overview figure illustrates the methodological framework divided into two primary stages: *LLM Analysis* and *Heuristics*.

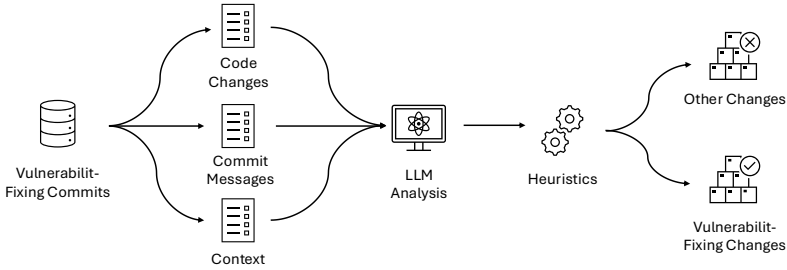


Fig. 1. Overview of the Methods and Experiments Conducted in This Paper

LLM Analysis We developed a prompt for LLMs that analyzes function changes along with their commit messages and context (other changed functions in the same commit). The complete prompt details are provided in Figure 2. The prompt produces a score from 0 to 4, representing the confidence in predicting vulnerability-fixing changes. To provide a clearer understanding of the scoring system, we offer the following explanations for each score:

- Score 0: No vulnerability detected
- Score 1: Low likelihood of vulnerability
- Score 2: Moderate likelihood of vulnerability
- Score 3: High likelihood of vulnerability
- Score 4: Very high likelihood of vulnerability

The motivation for using the prompt outputting a range from 0 to 4 instead of common binary output stems from how this output scale allows for fine-tuned control over dataset cleanliness. In particular, an exceptionally clean dataset might exclusively comprise items scored as 4, representing the utmost confidence in vulnerability-fixing changes. In contrast, for a more comprehensive

dataset, items scored 1 or higher could be retained, encompassing a wider array of potential vulnerability-fixing changes.

The Prompt Produces a Score from 0 to 4:
 As a cybersecurity expert, analyze the provided "Original" and "Revised" code snippets from a commit along with the commit message and other functions in the same commit, where the "Original" code represents the state prior to the changes, and the "Revised" code represents the state after the changes. **Evaluate the changes made in terms of vulnerability fixing on a scale of 0-4.** The length of the code snippet should not influence your assessment; focus on evaluating the logic line by line.

- A score of 0 indicates that the changes made from the "Original" code to the "Revised" code **are not related to fixing vulnerabilities.**
- A score of 4 indicates that the changes made from the "Original" code to the "Revised" code **are clearly focused on fixing vulnerabilities.**

Commit Message:
 {commit}

Original code snippet (code before changes):
 {original}

Revised code snippet (code after changes):
 {revised}

Here are the other functions in the same commit:
 {context}

Fig. 2. The Prompt Produces a Score from 0 to 4, Representing the Confidence in Predicting Vulnerability-Fixing Changes.

Heuristics To enhance the capability of LLMs in identifying vulnerability-fixing changes, we applied a heuristic approach. As identified in Section 2, *Test-Related Changes* significantly contribute to non-vulnerability fixes. Based on this insight, we devised heuristic rules to exclude test-related changes prior to processing the data with LLMs.

These rules were designed to identify and eliminate test-related code modifications by scrutinizing the naming conventions of the files and functions involved in the changes. We consulted widely-accepted testing frameworks and their naming conventions, such as Pytest for Python, to develop these rules. The comprehensive rules can be found in Figure 3.

We offer an example below to demonstrate the procedure. *Pytest Naming Conventions*: The file name should commence with the word "test" followed by the name of the file. An underscore symbol "_" is utilized to separate the terms "test" and the file name for visualization purposes. In accordance with this naming convention, we classify code modifications as test-related code changes if the function names contain the word *test*. By implementing these heuristic rules, we effectively filter out test-related changes, enabling the LLMs to concentrate on vulnerability-fixing changes and potentially enhancing their performance in identifying such changes.

4 CLEANVUL: Dataset Curation

To create a high-quality vulnerability dataset using VULSIFTER, we conducted a comprehensive crawl of 127,063 repositories, resulting in the acquisition of 5,352,105 commits. Using the keyword-based approach proposed by Bui et al. [2], we identified 43,029 function changes from Vulnerability-Fix Commits (VFCs) from these repositories. We analyzed these VFCs using VULSIFTER, which processed code changes across multiple programming languages: 26,423 Java, 6,591 Python, 5,578 C, 4,000 JavaScript, 312 C#, and 125 C++ changes.

Table 2. Number of Vulnerable Functions in Curated Datasets

Dataset Name	Threshold	Heuristic	Vulnerable Function (#)
CLEANVUL	1	✓	26,549
		✗	29,841
	2	✓	16,287
		✗	18,465
	3	✓	8,203
		✗	9,031
	4	✓	6,371
		✗	7,023

VULSIFTER assigns scores from 0 to 4 to each code change, where 0 indicates no relation to vulnerability fixes and 4 signifies a strong focus on fixing vulnerabilities. Table 2 presents the distribution of vulnerable functions across these threshold levels in our dataset CLEANVUL, evaluated both with (✓) and without (✗) heuristics. At threshold level 1, we identified 26,549 vulnerable functions with heuristics and 29,841 without. The number of detected functions decreases as the threshold increases: level 2 found 18,465 (✓) and 16,287 (✗) functions; level 3 identified 8,203 (✓) and 9,031 (✗) functions; and level 4 contained 6,371 (✓) and 7,023 (✗) functions.

By varying these threshold values, we created datasets of different quality levels. The dataset with threshold 4 represents the highest confidence in vulnerability-fixing changes, while the dataset with threshold 1 provides broader coverage by including potential vulnerability fixes.

5 Evaluation Settings

This evaluation examines the following research questions (RQs):

RQ1: (*Efficacy*) *What is the correctness improvement of CLEANVUL compared to the uncleaned dataset?*

Rationale: After applying VULSIFTER to clean the noise in the VFC dataset, it is important to evaluate the cleaned dataset (CLEANVUL) to understand if the percentage of vulnerable functions (*Correctness*) in the dataset has increased and the extent of the improvement.

RQ2: (*Effectiveness*) *How does the performance of LLMs fine-tuned on CLEANVUL compare to their performance on other established vulnerability datasets?*

Rationale: According to recent work [5], most vulnerability datasets include 40% to 75% of noisy data. In this study, we select the most refined high-quality datasets, including PrimeVul and SVEN, to train machine learning models and perform cross-dataset validation to evaluate the predictive performance of models trained on different datasets. This sheds light on the overall performance of models trained on different datasets, indicating the quality of these datasets.

RQ3: (*Effectiveness*) *How does the performance of LLMs fine-tuned on CLEANVUL compare to their performance on the uncleaned dataset?*

Rationale: Apart from cross-comparison with other datasets, we also compare the performance of models trained on our CLEANVUL dataset and uncleaned data to assess the improvement in performance on the original dataset. This helps us understand how much performance is improved after adopting VULSIFTER to remove non-vulnerability-fixing changes in VFCs.

5.1 Comparison with Established Datasets

Previous studies [4, 5] identified only two vulnerability datasets with *Correctness* above 70%: SVEN [10] and PrimeVul [5]. These datasets are comparable to our CLEANVUL, which achieves over 80% *Correctness* at thresholds 3 and 4. SVEN, developed through manual analysis, contains 803 functions and achieves 94.0% *Correctness*. PrimeVul identifies vulnerabilities by matching function names with NVD descriptions and contains 6,968 functions with 86.0% *Correctness*. However, PrimeVul is limited to NVD-linked vulnerabilities and cannot clean VFCs without NVD entries. We compare the generalizability of models fine-tuned on these established datasets versus our CLEANVUL.

5.2 Model Choices

Our experiments encompass various state-of-the-art LLMs. We have selected three widely-used encoder-only models, namely RoBERTa [15], CodeBERT [7], and GraphCodeBERT [9], along with three decoder-only models, which consist of two smaller models, GPT-2 [19] and CodeGPT [17], and one larger decoder-only model, CodeLlama [21]. These models have demonstrated exceptional performance in code-related tasks, making them suitable candidates for our study [25].

5.3 Evaluation Metrics

Four popular metrics are adopted in our experiments to assess the performance of the LLMs on tasks such as identifying vulnerability-fixing changes, comparing models trained on different datasets, and more:

- **Accuracy:** This metric measures the overall correctness of the model across all classes. It is defined as the ratio of correctly predicted observations (both true positives and true negatives) to the total number of observations. Our dataset consists of paired samples (vulnerable/benign function pairs), making the classes inherently balanced. Therefore, accuracy is particularly useful for evaluating the performance of models trained on our dataset.
- **Precision:** Precision is the ratio of correctly predicted positive observations to the total predicted positives. This metric is crucial when the cost of a false positive is high. In the context of this work, high precision means that most of the changes identified by the model as vulnerability-fixing are indeed correct.
- **Recall:** Recall is the ratio of correctly predicted positive observations to all actual positives. It measures the model's ability to find all relevant cases within a dataset. High recall is important in scenarios where missing an actual positive (failing to identify a true vulnerability-fixing change) can have serious implications.
- **F1-Score:** The F1-score is the harmonic mean of precision and recall. It is a way of combining both precision and recall into a single measure that captures both properties. This metric is particularly useful when you need to balance precision and recall, which often have an inverse relationship.

Each of these metrics offers distinct insights into the performance characteristics of the models used in our experiments, allowing us to tailor our model selection and tuning to the specific needs of the application at hand. In addition to these metrics, we also employ the *Correctness* metric to evaluate the quality of the vulnerability dataset.

- **Correctness:** Correctness is defined as the percentage of genuine vulnerable functions in the vulnerability dataset. This metric is particularly relevant in our study, as it helps assess the effectiveness of our approach in identifying and filtering vulnerability-fixing changes.

5.4 Implementation Details

For inference with commercial LLMs, we employ the LangChain framework. In order to fine-tune models such as RoBERTa, CodeBERT, GraphCodeBERT, GPT-2, CodeGPT, and CodeLlama, we utilize the Huggingface Transformers library. We set the maximum number of input tokens to 512 to ensure efficient processing. In terms of training, we configure the number of epochs to 50 and store the optimal checkpoints for later use. To facilitate a seamless transition from pretraining to fine-tuning, we maintain a consistent learning rate that aligns with the LLM's pretraining learning rate. Our experiments were executed on NVIDIA H100 GPUs, utilizing a server equipped with an Intel(R) Xeon(R) Platinum 8480C CPU and running Ubuntu 22.04.2 LTS as the operating system.

6 Evaluation Experiments

In this section, we report our experimental results and answer the three research questions.

6.1 RQ1: What is the correctness improvement of CLEANVUL compared to the uncleaned dataset?

To evaluate *Correctness* of CLEANVUL, we randomly select a sample of 487 function-level code changes from the collected GitHub VFC dataset and manually analyze them to create a testing dataset. This sample size provides a confidence level of 95% with a margin of error of $\pm 4.4\%$ for the full dataset, allowing us to make statistically significant observations. We carefully examine each code change and mark them as either vulnerability-fixing changes or other code changes, establishing a ground truth for our evaluation. This evaluation method aligns with previous work [5]. As mentioned in Section 3, VULSIFTER's output score ranges from 0 to 4, with 0 indicating no relation to vulnerability fixes and 4 signifying a strong focus on fixing vulnerabilities. We construct four datasets by setting different threshold values. A very clean dataset includes only items scored as 4, representing the highest confidence in vulnerability-fixing changes. In contrast, a more extensive dataset retains items scored 1 or higher, encompassing potential vulnerability-fixing changes. The results are presented in Table 3.

Upon applying VULSIFTER, the *Correctness* (defined in Section 5.3 as the percentage of genuine vulnerable functions in the vulnerability dataset) improves from 28.7% to a range of 37.5% to 97.3% on the test sample, with improvements spanning from 30.6% to 239.0%. Notably, the heuristic approach enhances *Correctness* across different thresholds. When the threshold is set to 1, the *Correctness* increases from 28.7% to 43.1% with the heuristic, resulting in a 50.1% improvement. Without the heuristic, the *Correctness* reaches 37.5%, yielding an 30.6% improvement. For thresholds 2, 3, and 4, the heuristic approach yields *Correctness* improvements of 101.0%, 215.6%, and 239.0%, respectively. Thresholds 3 and 4 demonstrate particularly high *Correctness* levels.

Comparison with Other Datasets Most existing datasets exhibit low *Correctness* levels since they treat all changes in VFCs as vulnerability-fixing changes. Only SVEN [10] and PrimeVul [5] verify this property and achieve good *Correctness*. However, SVEN is limited in size, containing only 803 vulnerable functions due to its reliance on manual analysis. PrimeVul, while larger, requires NVD links, making it unsuitable for VFCs without associated NVD entries. The enhanced *Correctness* using VULSIFTER is competitive with these established datasets. At a threshold of 3, VULSIFTER achieves a *Correctness* rate of 90.6%, comparable to SVEN [10] (94.0%) and PrimeVul [5] (86.0%). This achievement is particularly noteworthy given that our CLEANVUL dataset is derived from GitHub and operates at a larger scale, in contrast to others that primarily rely on NVD data. As such, our CLEANVUL dataset serves as a crucial complement, broadening the scope and application of vulnerability datasets in real-world scenarios.

Table 3. Comparison of Accuracy Across Existing Vulnerability Datasets With Our Cleaned Dataset

Dataset	Threshold	Heuristic	Vulnerable Func.	Ori. Corr.	Corr.	Improvement
CLEANVUL	1	✓	26,549	28.7%	43.1%	+50.1%
		✗	29,841	28.7%	37.5%	+30.6%
	2	✓	16,287	28.7%	49.4%	+72.1%
		✗	18,465	28.7%	57.7%	+101.0%
	3	✓	8,203	28.7%	90.6%	+215.6%
		✗	9,031	28.7%	76.5%	+166.5%
	4	✓	6,371	28.7%	97.3%	+239.0%
		✗	7,023	28.7%	78.0%	+171.7%
SVEN [10]	-	-	803	-	94.0%*	-
PrimeVul [5]	-	-	6,968	-	86.0%*	-
DiverseVul [4]	-	-	18,945	-	60.0%†	-
CVEFixes [1]	-	-	5,495	-	51.7%†	-
CrossVul [18]	-	-	5,877	-	47.8%†	-
VulnPatchPairs [20]	-	-	13,100	-	36.0%*	-
BigVul [6]	-	-	11,823	-	25.0%†	-
CodeXGLUE [17, 26]	-	-	23,355	-	24.0%*	-

* Refers to results in Ding et al. [5].

† Refers to results in Chen et al. [4].

With a threshold of 3, we obtain a dataset of **8,203** vulnerability-fixing changes, achieving a *Correctness* of **90.6%** on the testing sample. Increasing the threshold to 4 results in **6,371** vulnerability-fixing changes with a perfect *Correctness* of **97.3%**. These results match the *Correctness* levels of existing high-quality datasets SVEN [10] (94.0%) and PrimeVul [5] (86.0%), while providing more samples compared to SVEN (803) and PrimeVul (6,968).

6.2 RQ2: How does the performance of LLMs fine-tuned on CLEANVUL compare to their performance on other established vulnerability datasets?

To evaluate the effectiveness of our CLEANVUL dataset, we conducted two experiments: 1) we focused on fine-tuning various LLMs on all programming languages in the CLEANVUL dataset and other high-quality datasets; 2) since our CLEANVUL primarily contains vulnerable code in Java, we trained various LLMs on Java only in the CLEANVUL and compared them with models fine-tuned on other datasets to assess cross-language and cross-dataset performance.

Our CLEANVUL features a vulnerability score ranging from 0 to 4. To test the highest quality dataset we could obtain, we selected a threshold of 4. As mentioned in Section 6.1, this process resulted in a collection of 6,371 vulnerable and benign function pairs. To facilitate comparison with two other high-quality datasets, SVEN [10] and PrimeVul [5], which exhibit high *Correctness* levels exceeding 85%, we also employed a balanced dataset setting for training and testing to eliminate other factors. It is important to note that PrimeVul contains 6,968 pairs of code changes exclusively in C and C++ code, while SVEN consists of 803 pairs of code changes, with equal parts C and C++ code and Python code. We partitioned the dataset into training, testing, and validation sets using a 7:1.5:1.5 ratio and subsequently compared their cross-dataset performance. The evaluation metrics employed were accuracy, precision, recall, and F1-score, which are presented for different LLMs trained on one dataset and tested on others, as illustrated in Table 5.

Table 4. Comparison with Different LLMs Fine-Tuned on CLEANVUL (**Incorporating All Programming Languages**) and Other Datasets

Model	Train	Test	Acc (%)	Pre (%)	Rec (%)	F1 (%)
RoBERTa 125M Encoder-Only	CLEANVUL	CLEANVUL	65.83	64.55	71.17	67.56
		PrimeVul	56.26	55.79	60.80	58.18
		SVEN	60.97	57.91	80.20	67.23
	PrimeVul	CLEANVUL	52.53	52.01	65.88	57.85
		PrimeVul	51.63	51.77	52.17	51.15
		SVEN	49.72	49.75	75.17	59.34
	SVEN	CLEANVUL	50.98	51.28	51.33	50.10
		PrimeVul	52.32	52.72	52.83	48.98
		SVEN	75.00	77.90	70.91	73.77
CodeBERT 125M Encoder-Only	CLEANVUL	CLEANVUL	68.10	66.98	71.52	69.06
		PrimeVul	58.09	58.06	61.85	59.39
		SVEN	64.87	60.99	85.14	70.87
	PrimeVul	CLEANVUL	54.97	53.92	68.29	60.25
		PrimeVul	56.61	56.29	58.44	57.03
		SVEN	54.94	53.42	77.48	63.23
	SVEN	CLEANVUL	52.93	51.98	79.24	62.67
		PrimeVul	53.57	52.81	74.80	61.33
		SVEN	81.32	77.25	88.99	82.64
GraphCodeBERT 125M Encoder-Only	CLEANVUL	CLEANVUL	68.96	66.09	78.11	71.57
		PrimeVul	54.65	53.75	66.87	59.59
		SVEN	62.07	58.03	87.91	69.88
	PrimeVul	CLEANVUL	54.19	52.89	77.08	62.72
		PrimeVul	57.19	55.50	72.39	62.81
		SVEN	55.75	54.43	70.86	61.57
	SVEN	CLEANVUL	54.74	53.22	79.36	63.69
		PrimeVul	52.50	52.01	64.94	57.76
		SVEN	81.86	79.60	85.71	82.54
GPT-2 124M Decoder-Only	CLEANVUL	CLEANVUL	61.03	61.63	58.47	59.94
		PrimeVul	53.13	53.01	60.43	56.03
		SVEN	54.37	53.61	66.12	59.17
	PrimeVul	CLEANVUL	51.75	51.75	51.12	51.32
		PrimeVul	51.26	51.42	45.28	48.14
		SVEN	55.74	56.54	49.18	52.58
	SVEN	CLEANVUL	52.36	51.92	67.41	58.38
		PrimeVul	51.50	51.44	60.97	54.67
		SVEN	78.69	78.34	80.33	78.95
CodeGPT 124M Decoder-Only	CLEANVUL	CLEANVUL	66.81	67.32	65.57	66.38
		PrimeVul	56.25	57.40	49.69	53.11
		SVEN	56.83	54.97	75.96	63.76
	PrimeVul	CLEANVUL	52.62	52.08	63.79	57.14
		PrimeVul	53.02	54.20	39.16	45.44
		SVEN	51.91	51.57	60.11	53.87
	SVEN	CLEANVUL	52.91	52.71	56.49	54.42
		PrimeVul	55.20	56.75	43.12	48.11
		SVEN	79.78	83.38	74.86	78.63
CodeLlama 7B Decoder-Only	CLEANVUL	CLEANVUL	60.98	60.51	64.67	62.46
		PrimeVul	53.46	53.74	45.92	49.05
		SVEN	53.14	52.63	64.27	57.83
	PrimeVul	CLEANVUL	51.59	48.73	44.17	40.06
		PrimeVul	52.82	51.90	38.88	39.04
		SVEN	52.56	68.36	48.16	41.37
	SVEN	CLEANVUL	50.53	51.21	49.36	45.60
		PrimeVul	50.27	49.43	49.86	42.19
		SVEN	51.66	52.19	53.65	48.40

Intra-Dataset Performance Fine-Tuned on All Languages Considering that the datasets used for training and testing are balanced in this study, we primarily compare and report accuracy. The results indicate that models trained on CLEANVUL generally exhibit good performance on the same dataset. For example, GraphCodeBERT achieved the best intra-dataset performance for all three datasets. Specifically, GraphCodeBERT fine-tuned and tested on CLEANVUL achieved an accuracy of 68.96%, 57.19% on PrimeVul, and 81.86% on SVEN. The accuracy is higher than PrimeVul (68.96% vs 57.19%), but lower than SVEN (81.86%). This discrepancy might be due to the low diversity of the SVEN dataset, which contains only around 800 vulnerability functions across 10 CWE vulnerabilities. Another reason could be the diversity in programming languages, as CLEANVUL includes six different languages: Java, Python, C, C++, C#, and JavaScript, while SVEN contains only Python, C, and C++. Interestingly, larger models such as CodeLlama-7B achieved the lowest accuracy of 60.98% compared to the smaller models, which could be attributed to a form of underfitting, possibly due to insufficient training data or inadequate model architecture for capturing the nuances of the diverse programming languages and vulnerabilities.

Inter-Dataset Generalization Fine-Tuned on All Languages Our dataset CLEANVUL demonstrated excellent performance during cross-dataset generalization experiments. Specifically, regarding generalization to PrimeVul, when fine-tuned on our dataset and tested on PrimeVul, the best accuracy is 58.09% with CodeBERT, which is even higher than solely fine-tuning and testing on PrimeVul with the best accuracy of 57.19%. The best model that fine-tuned on SVEN and tested on PrimeVul is CodeGPT with an accuracy of 55.20%, which is lower than fine-tuning and testing on PrimeVul with the best accuracy of 57.19%. This shows that models fine-tuned on our dataset achieved the best performance on PrimeVul compared to models fine-tuned on PrimeVul or SVEN.

Regarding generalization to SVEN, when fine-tuned on CLEANVUL and tested on SVEN, the best accuracy is 64.87% with CodeBERT, which is higher than the models fine-tuned on PrimeVul and tested on SVEN with an accuracy of 55.75% with GraphCodeBERT. Although both are lower than only training and testing on SVEN, with an accuracy of 81.86% with GraphCodeBERT, our dataset demonstrated impressive performance in generalizing knowledge to an unknown dataset.

In terms of training on other two datasets and testing on CLEANVUL, we can notice that the best accuracy for SVEN is 54.74% with GraphCodeBERT, and the best accuracy for PrimeVul is 54.97% with CodeBERT. Both accuracies are lower than the accuracy trained and tested on our dataset (68.96%). This shows that our dataset might be more diverse than PrimeVul and SVEN.

Intra-Dataset Performance Fine-Tuned Exclusively on Java Similar to the models fine-tuned on all programming languages on CLEANVUL, the results fine-tuned exclusively on Java on CLEANVUL also demonstrate very good intra-dataset performance. For instance, CodeGPT achieved an accuracy of 73.36%, outperforming the intra-dataset performance of PrimeVul (54.49%) but still lower than the accuracy of SVEN (82.79%). Comparing different LLMs, we observed that GraphCodeBERT is the best-performing encoder-only model with an accuracy of 74.78%, while CodeGPT is the top-performing decoder-only model with an accuracy of 73.36%. Similarly, larger models such as CodeLlama-7B achieved a lower accuracy of 70.21% compared to these smaller models. Moreover, the intra-dataset performance fine-tuned exclusively on Java is improved in comparison with fine-tuning on all programming languages on CLEANVUL, as the best accuracy increased from 68.96% to 74.78%.

Inter-Dataset Generalization Fine-Tuned Exclusively on Java In cross-dataset testing, when training GPT-2 on CLEANVUL on Java and testing on SVEN, the accuracy reached 59.02%, which is comparable to the best accuracy achieved by training GraphCodeBERT on PrimeVul and testing on SVEN (59.84%). However, this is still lower than the 85.25% accuracy obtained when training

Table 5. Comparison with Different LLMs Fine-Tuned on CLEANVUL (**Java Only**) and Other Datasets

Model	Train	Test	Acc (%)	Pre (%)	Rec (%)	F1 (%)
RoBERTa 125M Encoder-Only	CLEANVUL	CLEANVUL	66.96	65.60	71.30	68.33
		PrimeVul	50.66	50.47	70.61	58.87
		SVEN	51.64	51.28	65.57	57.55
	PrimeVul	CLEANVUL	52.61	51.79	75.65	61.48
		PrimeVul	53.51	53.23	57.89	55.46
		SVEN	51.64	51.00	83.61	63.35
	SVEN	CLEANVUL	53.48	53.17	58.26	55.60
		PrimeVul	53.73	53.36	59.21	56.13
		SVEN	76.23	75.00	78.69	76.80
CodeBERT 125M Encoder-Only	CLEANVUL	CLEANVUL	73.04	68.53	85.22	75.97
		PrimeVul	52.41	51.80	69.30	59.29
		SVEN	50.82	50.44	93.44	65.52
	PrimeVul	CLEANVUL	53.91	52.98	69.57	60.15
		PrimeVul	54.17	54.15	54.39	54.27
		SVEN	57.38	54.95	81.97	65.79
	SVEN	CLEANVUL	54.35	54.24	55.65	54.94
		PrimeVul	53.51	52.58	71.49	60.59
		SVEN	85.25	85.25	85.25	85.25
GraphCodeBERT 125M Encoder-Only	CLEANVUL	CLEANVUL	74.78	71.11	83.48	76.80
		PrimeVul	52.19	51.48	76.32	61.48
		SVEN	54.10	52.48	86.89	65.43
	PrimeVul	CLEANVUL	50.43	50.30	73.04	59.57
		PrimeVul	55.26	54.92	58.77	56.78
		SVEN	59.84	56.98	80.33	66.67
	SVEN	CLEANVUL	51.30	51.22	54.78	52.94
		PrimeVul	52.63	52.13	64.47	57.65
		SVEN	83.61	85.96	80.33	83.05
GPT-2 124M Decoder-Only	CLEANVUL	CLEANVUL	71.30	71.30	71.30	71.30
		PrimeVul	53.29	52.49	69.30	59.74
		SVEN	59.02	56.79	75.41	64.79
	PrimeVul	CLEANVUL	50.00	50.00	50.43	50.22
		PrimeVul	52.63	53.23	43.42	47.83
		SVEN	56.56	56.06	60.66	58.27
	SVEN	CLEANVUL	54.78	52.97	85.22	65.33
		PrimeVul	53.73	54.03	50.00	51.94
		SVEN	80.33	79.37	81.97	80.65
CodeGPT 124M Decoder-Only	CLEANVUL	CLEANVUL	73.36	73.17	73.77	73.47
		PrimeVul	55.13	54.51	61.97	58.00
		SVEN	52.46	51.58	80.33	62.82
	PrimeVul	CLEANVUL	53.28	52.11	81.15	63.46
		PrimeVul	54.49	57.34	35.04	43.50
		SVEN	50.00	50.00	86.89	63.47
	SVEN	CLEANVUL	50.00	50.00	76.23	60.39
		PrimeVul	52.99	52.82	55.98	54.36
		SVEN	82.79	83.33	81.97	82.64
CodeLlama 7B Decoder-Only	CLEANVUL	CLEANVUL	70.21	70.92	68.49	69.69
		PrimeVul	52.50	53.02	43.85	48.00
		SVEN	51.52	51.52	51.52	51.52
	PrimeVul	CLEANVUL	50.68	62.50	3.42	6.49
		PrimeVul	49.81	47.37	3.46	6.45
		SVEN	50.76	100.00	1.52	2.99
	SVEN	CLEANVUL	49.32	49.02	34.25	40.32
		PrimeVul	50.00	50.00	69.62	58.20
		SVEN	50.76	50.65	59.09	54.55

and testing CodeBERT on SVEN. This finding indicates that CLEANVUL (Java only) and PrimeVul exhibit similar performance when generalizing to SVEN. However, considering that PrimeVul contains only C++ and C code and SVEN contains almost half of C++ and C vulnerable functions, and that models are fine-tuned on CLEANVUL only on Java code, the generalizability of CLEANVUL (Java only) shows comparable performance across languages rather than not cross languages from PrimeVul to SVEN. This demonstrates impressive cross-language performance of CLEANVUL.

Regarding generalization to PrimeVul, training CodeGPT on CLEANVUL on Java and testing on PrimeVul resulted in an accuracy of 55.13%, which is very close to the best performance achieved by training and testing GraphCodeBERT on PrimeVul (55.26%). Conversely, when training on PrimeVul and testing on CLEANVUL on Java, the best performance was 53.91% for CodeBERT, which is considerably lower than the 74.78% accuracy achieved when training and testing on CLEANVUL (Java only). Additionally, the highest accuracy obtained when training on SVEN and testing on PrimeVul was 53.73% using GPT-2 or RoBERTa, which is lower than the results from training on CLEANVUL on Java and testing on PrimeVul. This evidence highlights the high quality of our dataset, suggesting that CLEANVUL (Java only) may contain more comprehensive or diverse examples of vulnerability-fixing changes than PrimeVul and SVEN, and that the knowledge from training on CLEANVUL can be generalized to PrimeVul but not vice versa.

In terms of training on SVEN and testing on CLEANVUL (Java only), the best accuracy achieved was 54.78% using GPT-2. This is similar to the results obtained when training on SVEN and testing on PrimeVul (53.73%), indicating that the diversity of CLEANVUL and PrimeVul could be much higher than that of SVEN.

Models fine-tuned on CLEANVUL demonstrate **much better generalization capabilities comparable to those fine-tuned on PrimeVul** when tested on SVEN (64.87% vs 55.75%). Additionally, these models exhibit **superior generalization abilities when tested on PrimeVul** compared to models trained on SVEN (58.09% vs 55.20%). Remarkably, the accuracy achieved by models trained on CLEANVUL and tested on PrimeVul is even better than those trained and tested solely on PrimeVul (**58.09% vs 57.19%**), highlighting the effectiveness and robustness of CLEANVUL in model fine-tuning.

6.3 RQ3: How does the performance of LLMs fine-tuned on CLEANVUL compare to their performance on the uncleaned dataset?

To further assess the effectiveness of VULSIFTER in cleaning noisy data, we train the same LLMs on both CLEANVUL and the uncleaned dataset, and test their performance on these datasets as well as two other high-quality datasets, PrimeVul and SVEN. Since GraphCodeBERT demonstrates the best generalization ability when trained on CLEANVUL in Section 6.2, we train GraphCodeBERT on the uncleaned dataset again and compare the performance.

The results are presented in Table 6. When training and testing on the same dataset, GraphCodeBERT trained on CLEANVUL achieves a higher accuracy compared to the uncleaned dataset, with accuracies of 68.96% and 55.23%, and F1-scores of 71.57% and 46.32%, respectively. Notably, when trained on CLEANVUL and tested on unclean data, the accuracy reaches 56.63% and the F1-score is 65.09%, which is higher than training and testing both on uncleaned data. Furthermore, when trained on uncleaned data and tested on CLEANVUL, the accuracy is 59.90% and the F1-score is 50.42%, which is considerably lower than the accuracy of 68.96% and F1-score of 71.57% when trained and tested on CLEANVUL.

Table 6. Comparison of CLEANVUL and Uncleaned Dataset Performance Across Other Datasets

Model	Train	Test	Acc (%)	Pre (%)	Rec (%)	F1 (%)
GraphCodeBERT 125M Encoder-Only	CLEANVUL	CLEANVUL	68.96	66.09	78.11	71.57
		PrimeVul	54.65	53.75	66.87	59.59
		SVEN	62.07	58.03	87.91	69.88
		Uncleaned Data	56.63	54.49	80.85	65.09
	Uncleaned Data	CLEANVUL	59.90	60.63	48.17	50.42
		PrimeVul	52.42	68.41	47.22	40.50
		SVEN	52.19	51.45	62.46	54.33
		Uncleaned Data	55.23	55.98	43.98	46.32

Upon examining the performance of training on CLEANVUL and uncleaned data and testing on other high-quality datasets, the differences become even more pronounced. CLEANVUL achieves 54.65% accuracy and 59.59% F1-score on PrimeVul, and 62.07% accuracy and 69.88% F1-score on SVEN, while the uncleaned dataset only attains accuracies of 52.42% and 52.19%, and F1-scores of 40.50% and 54.33% on PrimeVul and SVEN, respectively.

Training on the CLEANVUL dataset improves accuracy and F1-score when tested on the same dataset and other high-quality datasets (PrimeVul and SVEN) compared to using uncleaned data. The accuracy rates are 68.96% versus 55.23%, 54.65% versus 52.42%, and 62.07% versus 52.19%, respectively. Similarly, the F1-scores are 71.57% versus 46.32%, 59.59% versus 40.50%, and 69.88% versus 54.33%, respectively, demonstrating the benefits of training on cleaned data.

7 Additional Analyses

Beyond the primary evaluation discussed in Section 6, we conducted additional analyses to gain deeper insights into VULSIFTER's performance. Our additional investigations include: 1) an ablation study examining the impact of input combinations, the heuristic module, and the decision to use a 0-4 rating scale rather than binary output, and 2) a sensitivity analysis across different LLMs.

7.1 Sensitivity Analyses

Manual Analysis To evaluate the performance of different LLMs, as no evaluation dataset existed for this task, and no prior studies had proposed automatic approaches for this, we curated a test dataset manually. We conducted a detailed manual analysis involving multiple rounds:

- In the **first round**, we clarified our objectives and engaged five researchers to analyze a batch of previously studied VFCs [2]. Out of these five researchers, two were assigned to independently analyze each function change. We examined a total of 125 function changes in 50 VFCs, categorizing each change as a vulnerability-fixing change or not, while taking into account the function changes and commit messages. The independent analyses conducted by the two researchers from the group of five were later used to calculate Kappa's agreement.
- In the **second and third rounds**, the same five researchers analyzed another 152 and 84 function changes, respectively.
- In the **fourth and fifth rounds**, three researchers analyzed an additional 414 function changes in total.

In total, we analyzed 775 function changes, which exceeds the statistically significant sample size (385) with a confidence level of 95% and a margin of error of 5%. After three rounds of analysis, all function changes were analyzed by two independent researchers. The Kappa's agreement score was calculated based on the agreement between these two researchers for the entire set of 775 function changes. The obtained Kappa's agreement score of 0.681 indicates substantial agreement among the researchers.

LLM Performance Evaluation Following our manual analysis, we proceeded to assess the performance of several leading LLMs in identifying vulnerability-fixing changes. Our study incorporated some of the most widely recognized models, including GPT3.5, GPT4, GPT4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro, for their widespread popularity and cutting-edge performance. It is important to note that the specific versions of the closed-source models employed in our experiments. For GPT3.5, we use the chatgpt-35-0301 version, while for GPT4, we deploy the gpt-4-0314 version. In the case of GPT4o, we adopt the gpt-4o-2024-08-06 version.

Table 7. Comparison of F1-Scores (%) Across Different Models

Model	GPT3.5	GPT4	GPT4o	Claude	Gemini
F1-score	75.85	82.24	74.73	70.91	71.77

The results of our evaluation are presented in Table 7, which displays the F1-scores achieved by each model in identifying vulnerability-fixing changes. GPT4 demonstrated superior performance with the highest F1-score of 82.24%, followed by GPT3.5 at 75.85% and GPT4o at 74.73%. Both Claude 3.5 Sonnet and Gemini 1.5 Pro achieved F1-scores of 70.91% and 71.77%. These results suggest that while all models show competence in identifying vulnerability-fixing changes, GPT4 maintains a notable edge in this specific task.

7.2 Ablation Studies

We perform an ablation study by comparing different variants of VULSIFTER. Table 8 shows the comparison results across different inputs. The largest numbers are highlighted in bold. The variants considered in the study are as follows:

w/o Context Our tool without the contextual information of other changed functions in the commit. Without this information, the model may miss important relationships between interrelated changes, impacting the effectiveness of vulnerability detection. The F1-score drops from 82.24% to 77.64%, showing a decrease of 4.6%. These results confirm that contextual information improves detection accuracy.

Table 8. Ablation Study Results Comparing F1-Scores (%) Across Different Inputs

	F1-score (%)
VULSIFTER	82.24
w/o Context	77.64
w/o Commit Message	81.17
w/o Commit Message & Context	76.00

w/o Commit Message Our tool without utilizing commit messages in the analysis. Commit messages often contain valuable information about the nature and purpose of changes. Without

them, the F1-score decreases by 1.07% (from 82.24% to 81.17%). While the impact is smaller than removing context, this still demonstrates the value of commit message information in vulnerability detection.

w/o Commit Message & Context Our tool without both commit messages and contextual information. This represents the most basic configuration, relying solely on function changes. The F1-score drops to 76.00%, a decrease of 6.24% from the full configuration. This substantial reduction highlights the importance of having both commit messages and contextual information for effective vulnerability detection.

Table 9. Ablation Study Results Comparing F1-Scores (%) Between Our 0-4 Output and Simple Binary Output

	F1-score (%)
VULSIFTER	82.24
w/o 0-4 Output	74.66

w/o 0-4 Output We then conducted an additional ablation study to evaluate the impact of our 0-4 scoring output design compared to a simple binary output approach, using our best-performing input combination of function changes plus commit message and context. The results are presented in Table 9. The analysis demonstrates that our proposed 0-4 output design improves the detection performance. When replacing the 0-4 output design with a simple binary output, the F1-score decreases from 82.24% to 74.66%, showing a reduction of 7.58%. This substantial difference highlights the effectiveness of our fine-grained vulnerability scoring approach compared to traditional binary classification.

Table 10. Ablation Study Results Comparing F1-Scores (%) With and Without Heuristic Filtering

	F1-score (%)
VULSIFTER	82.24
w/o Heuristics	78.97

w/o Heuristics VULSIFTER without the heuristic approach that filters out test-related changes before processing. Without these heuristics, the model may waste computational resources and potentially be misled by test code modifications that are not actual vulnerability fixes. The F1-score decreases from 82.24% to 78.97%, showing a reduction of 3.27%. This decline demonstrates that our heuristic approach effectively improves the tool's ability to identify genuine vulnerability-fixing changes by focusing the analysis on the most relevant code modifications.

8 Threats to Validity

Internal Validity Internal validity refers to the extent to which a study establishes a causal relationship between the independent and dependent variables, free from systematic errors and bias. One potential threat to internal validity arises from the manual labeling process used to create our evaluation dataset. To mitigate this threat, we employed multiple researchers for the labeling process and calculated Cohen's Kappa agreement score (0.681), indicating substantial inter-rater reliability. Additionally, we conducted multiple rounds of analysis and cross-validation among

researchers to ensure consistency. Another internal threat stems from the probabilistic nature of LLMs, which can lead to performance variations between runs. To address this, we conducted three runs for each experiment and reported the average results.

External Validity External validity refers to the extent to which research findings can be generalized to other contexts and settings. To ensure generalizability across programming languages, our dataset includes code from multiple popular languages (Java, Python, C, JavaScript, C#, C++). Additionally, we selected projects with diverse development practices and team sizes that mirror typical software development environments, enhancing the applicability of our findings to real-world scenarios.

Construct Validity Construct validity concerns the extent to which a study's measurements actually represent the intended theoretical constructs. To ensure robust evaluation, we employed a comprehensive set of standard metrics (F1-score, precision, recall, accuracy, and correctness) commonly used in vulnerability detection research.

9 Related Work

In this section, we review existing research related to vulnerability datasets, methods for improving dataset accuracy, and the application of machine learning models in vulnerability detection.

9.1 Vulnerability Datasets

The creation and curation of high-quality vulnerability datasets is critical for advancing automated vulnerability detection techniques. Several notable datasets have been developed in recent years: BigVul [6] and CodeXGLUE [17, 26] are large-scale datasets containing over 10K vulnerable functions. However, as noted by Chen et al. [4], these datasets suffer from low correctness rates of around 25%, meaning a significant portion of the labeled vulnerabilities may be inaccurate. More recent efforts have improved dataset quality. CrossVul [18] and CVEFixes [1] achieved correctness rates of 47.8% and 51.7% respectively [5]. DiverseVul [4] further improved on this with a 60% correctness rate across nearly 19,000 functions. The current state-of-the-art in terms of dataset quality is represented by SVEN [10] and PrimeVul [5], which report very high correctness rates of 94% and 86% respectively. However, the SVEN dataset is notably small, containing only 803 functions, primarily due to the constraints of manual analysis. PrimeVul leverages NVD descriptions to match function names, but this method limits its applicability to VFCs that do not correspond to NVD entries.

Our work aims to bridge the gap between dataset size and quality. CLEANVUL achieves a correctness rate of 90.6% (comparable to SVEN and PrimeVul) while maintaining a larger scale of 8,203 functions. Importantly, CLEANVUL is derived from GitHub data rather than relying solely on NVD entries, making it a valuable complement to existing high-quality datasets.

9.2 Automated Vulnerability Detection Techniques

Research in automated vulnerability detection has seen significant progress, evolving from traditional static analysis techniques to more sophisticated machine learning approaches: Early work focused on static analysis tools that use predefined rules to identify potential vulnerabilities [11, 16]. While effective for certain types of vulnerabilities, these approaches often suffer from high false positive rates and struggle with complex, context-dependent vulnerabilities. Machine learning techniques have emerged as a promising direction for improving vulnerability detection. Supervised learning approaches have been applied to classify code as vulnerable or benign based on features extracted from source code [22]. These methods have shown improved accuracy over traditional static analysis but are highly dependent on the quality of training data. Deep learning models

have recently gained traction in this domain. Wang et al. [24] proposed using deep belief networks for vulnerability detection, while Li et al. [14] introduced VulDeePecker, a neural network-based approach for detecting vulnerabilities in source code. More recent work has explored the use of graph neural networks [26] and transformer-based models [4] for vulnerability detection, showing promising results.

9.3 Dataset Cleaning and Noise Reduction

The challenge of noisy labels in security datasets has been recognized in several studies: Ding et al. [5] highlighted the issue of noise in vulnerability datasets, reporting that existing datasets often contain 40% to 75% noisy data. They proposed a method to improve dataset quality by correlating function names from commit logs with NVD descriptions. Chen et al. [4] addressed the dataset noise problem by developing a multi-stage filtering process to create a more diverse and accurate vulnerability dataset.

Our work contributes to this area by proposing the first automatic approach for identifying and filtering out non-vulnerability-related changes in commits, without requiring NVD entry links. We provide a detailed analysis of the types of changes affecting non-vulnerable functions commonly found in VFCs and demonstrate the effectiveness of VULSIFTER in reducing noise in the resulting dataset.

10 Conclusion and Future Work

In this paper, we addressed the critical challenge of accurately identifying vulnerability-fixing changes within vulnerability-fixing commits (VFCs), a task essential for improving the effectiveness of machine learning models in automated vulnerability detection. Our comprehensive study revealed that a significant portion of changes within VFCs are not directly related to fixing vulnerabilities, with *Test-Related Changes* and *Bug Fixes* accounting for 41.2% and 38.2% of non-vulnerability changes, respectively. This insight informed the development of our LLM heuristic approach - **VULSIFTER**, which demonstrated superior performance in identifying genuine vulnerability fixes in VFCs. Notably, GPT-4, enhanced with our heuristic method, achieved an F1-score of 0.82, making it the first approach to automatically identify genuine vulnerability fixes in VFCs, without requiring NVD entry links.

We created **CLEANVUL**, a new high-quality vulnerability dataset, by analyzing 5,352,105 commits from 127,063 GitHub repositories. Using VULSIFTER, we filtered out noise from vulnerability-fixing commits (VFCs) in this corpus. VULSIFTER has a configurable threshold that can be set based on data cleanliness requirements, allowing us to balance dataset size and quality. With a threshold of 3, CLEANVUL contains 8,203 vulnerable function pairs achieving 90.6% *Correctness*, while a stricter threshold of 4 yields 6,371 function pairs with 97.3% *Correctness* in our test sample. These results are comparable to established datasets such as SVEN (94.0% *Correctness*) and PrimeVul (86.0% *Correctness*), while overcoming their limitations. Unlike SVEN, which relies on manual analysis and contains only 803 samples, our approach is scalable. Additionally, while PrimeVul requires NVD entry links, our VULSIFTER can analyze all VFCs, including those without such links. Our evaluation of various LLMs fine-tuned on CLEANVUL revealed its superior generalization capabilities across different datasets. Notably, models fine-tuned on CLEANVUL significantly outperformed PrimeVul-trained models when tested on SVEN, achieving 64.87% accuracy compared to 55.75%. When evaluated on PrimeVul, our models demonstrated better performance than those trained on SVEN (58.09% vs 55.20%). Most remarkably, models trained on CLEANVUL and tested on PrimeVul achieved higher accuracy than models trained and tested on PrimeVul itself (58.09% vs 57.19%). These results underscore CLEANVUL's effectiveness and robustness for model fine-tuning, particularly in its ability to capture generalizable vulnerability patterns that transfer well across different contexts.

In the future, we plan to extend our approach in several key directions. First, we aim to develop techniques that can effectively process long-form software artifacts by exploring hierarchical analysis methods that can maintain model attention across extensive codebases. Given our observation that smaller specialized models sometimes outperform larger ones, we plan to investigate architectures specifically optimized for vulnerability detection tasks. We also intend to explore methods for intelligent context integration, combining commit messages with code diffs and project metadata while respecting model input limitations. Finally, we plan to work on semi-automated dataset curation approaches that leverage both LLM capabilities and expert validation to further improve the quality of vulnerability datasets, building upon the success of our current dataset.

References

- [1] Guru Bhandari, Amara Naseer, and Leon Moonen. 2021. CVEfixes: automated collection of vulnerabilities and their fixes from open-source software. In *Proceedings of the 17th International Conference on Predictive Models and Data Analytics in Software Engineering*. 30–39.
- [2] Tan Bui, Yan Naing Tun, Yiran Cheng, Ivana Clairine Irsan, Ting Zhang, and Hong Jin Kang. 2024. JavaVFC: Java Vulnerability Fixing Commits from Open-source Software. *arXiv preprint arXiv:2409.05576* (2024).
- [3] Saikat Chakraborty, Rahul Krishna, Yangruibo Ding, and Baishakhi Ray. 2021. Deep learning based vulnerability detection: Are we there yet? *IEEE Transactions on Software Engineering* 48, 9 (2021), 3280–3296.
- [4] Yizheng Chen, Zhoujie Ding, Lamy Alowain, Xinyun Chen, and David Wagner. 2023. Diversevul: A new vulnerable source code dataset for deep learning based vulnerability detection. In *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses*. 654–668.
- [5] Yangruibo Ding, Yanjun Fu, Omniyyah Ibrahim, Chawin Sitawarin, Xinyun Chen, Basel Alomair, David Wagner, Baishakhi Ray, and Yizheng Chen. 2024. Vulnerability detection with code language models: How far are we? *arXiv preprint arXiv:2403.18624* (2024).
- [6] Jiahao Fan, Yi Li, Shaohua Wang, and Tien N Nguyen. 2020. AC/C++ code vulnerability dataset with code changes and CVE summaries. In *Proceedings of the 17th International Conference on Mining Software Repositories*. 508–512.
- [7] Zhangyin Feng, Daya Guo, Duyu Tang, Nan Duan, Xiaocheng Feng, Ming Gong, Linjun Shou, Bing Qin, Ting Liu, Daxin Jiang, et al. 2020. Codebert: A pre-trained model for programming and natural languages. *arXiv preprint arXiv:2002.08155* (2020).
- [8] Michael Fu and Chakkrit Tantithamthavorn. 2022. Linevul: A transformer-based line-level vulnerability prediction. In *Proceedings of the 19th International Conference on Mining Software Repositories*. 608–620.
- [9] Daya Guo, Shuo Ren, Shuai Lu, Zhangyin Feng, Duyu Tang, Shujie Liu, Long Zhou, Nan Duan, Alexey Svyatkovskiy, Shengyu Fu, et al. 2020. Graphcodebert: Pre-training code representations with data flow. *arXiv preprint arXiv:2009.08366* (2020).
- [10] Jingxuan He and Martin Vechev. 2023. Large language models for code: Security hardening and adversarial testing. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*. 1865–1879.
- [11] Nenad Jovanovic, Christopher Kruegel, and Engin Kirda. 2006. Pixy: A static analysis tool for detecting web application vulnerabilities. In *2006 IEEE Symposium on Security and Privacy (S&P'06)*. IEEE, 6–pp.
- [12] Hiroyuki Kirinuki, Yoshiki Higo, Keisuke Hotta, and Shinji Kusumoto. 2014. Hey! are you committing tangled changes?. In *Proceedings of the 22nd International Conference on Program Comprehension*. 262–265.
- [13] Stanislav Levin and Amiram Yehudai. 2017. Boosting automatic commit classification into maintenance activities by utilizing source code changes. In *Proceedings of the 13th International Conference on Predictive Models and Data Analytics in Software Engineering*. 97–106.
- [14] Zhen Li, Deqing Zou, Shouhuai Xu, Xinyu Ou, Hai Jin, Sujuan Wang, Zhijun Deng, and Yuyi Zhong. 2018. Vuldeepecker: A deep learning-based system for vulnerability detection. *arXiv preprint arXiv:1801.01681* (2018).
- [15] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
- [16] Benjamin Livshits and Thomas Zimmermann. 2005. Dynamine: finding common error patterns by mining software revision histories. *ACM SIGSOFT Software Engineering Notes* 30, 5 (2005), 296–305.
- [17] Shuai Lu, Daya Guo, Shuo Ren, Junjie Huang, Alexey Svyatkovskiy, Ambrosio Blanco, Colin Clement, Dawn Drain, Daxin Jiang, Duyu Tang, et al. 2021. Codexglue: A machine learning benchmark dataset for code understanding and generation. *arXiv preprint arXiv:2102.04664* (2021).
- [18] Georgios Nikitopoulos, Konstantina Dritsa, Panos Louridas, and Dimitris Mitropoulos. 2021. CrossVul: a cross-language vulnerability dataset with commit data. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering*

- Conference and Symposium on the Foundations of Software Engineering*. 1565–1569.
- [19] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
 - [20] Niklas Risse and Marcel Böhme. 2023. Limits of machine learning for automatic vulnerability detection. *arXiv preprint arXiv:2306.17193* (2023).
 - [21] Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950* (2023).
 - [22] Riccardo Scandariato, James Walden, Aram Hovsepyan, and Wouter Joosen. 2014. Predicting vulnerable software components via text mining. *IEEE Transactions on Software Engineering* 40, 10 (2014), 993–1006.
 - [23] Benjamin Steenhoek, Md Mahbubur Rahman, Richard Jiles, and Wei Le. 2023. An empirical study of deep learning models for vulnerability detection. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 2237–2248.
 - [24] Song Wang, Taiyue Liu, and Lin Tan. 2016. Automatically learning semantic features for defect prediction. In *Proceedings of the 38th international conference on software engineering*. 297–308.
 - [25] Frank F Xu, Uri Alon, Graham Neubig, and Vincent Josua Hellendoorn. 2022. A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*. 1–10.
 - [26] Yaqin Zhou, Shangqing Liu, Jingkai Siow, Xiaoning Du, and Yang Liu. 2019. Devign: Effective vulnerability identification by learning comprehensive program semantics via graph neural networks. *Advances in neural information processing systems* 32 (2019).

A Appendix

```

self.function_patterns = {
    # Java test method patterns
    'java': [
        r'@Test\s+.*?(?:public\s+)?void\s+(\w+)\s*\([^)]*\)',
        r'@Before\s+.*?(?:public\s+)?void\s+(\w+)\s*\([^)]*\)',
        r'@After\s+.*?(?:public\s+)?void\s+(\w+)\s*\([^)]*\)',
        r'@BeforeEach\s+.*?(?:public\s+)?void\s+(\w+)\s*\([^)]*\)',
        r'@AfterEach\s+.*?(?:public\s+)?void\s+(\w+)\s*\([^)]*\)'
    ],
    # C/C++ test function patterns
    'cpp': [
        r'TEST\s*\(\s*(\w+)\s*,\s*(\w+)\s*\)',
        r'TEST_F\s*\(\s*(\w+)\s*,\s*(\w+)\s*\)',
        r'TEST_P\s*\(\s*(\w+)\s*,\s*(\w+)\s*\)'
    ],
    # C# test method patterns
    'csharp': [
        r'\[Test(?:Case)?\]\s+.*?(?:public\s+)?void\s+(\w+)\s*\([^)]*\)',
        r'\[TestMethod\]\s+.*?(?:public\s+)?void\s+(\w+)\s*\([^)]*\)',
        r'\[Fact\]\s+.*?(?:public\s+)?void\s+(\w+)\s*\([^)]*\)',
        r'\[Theory\]\s+.*?(?:public\s+)?void\s+(\w+)\s*\([^)]*\)'
    ],
    # JavaScript test function patterns
    'javascript': [
        r'test\s*\(\s*["'].*?["']\s*,\s*(?:function|[\(\)]*)\s*=>',
        r'it\s*\(\s*["'].*?["']\s*,\s*(?:function|[\(\)]*)\s*=>',
        r'describe\s*\(\s*["'].*?["']\s*,\s*(?:function|[\(\)]*)\s*=>',
        r'beforeEach\s*\(\s*(?:function|[\(\)]*)\s*=>',
        r'afterEach\s*\(\s*(?:function|[\(\)]*)\s*=>'
    ],
    # Python test function patterns
    'python': [
        r'@pytest.mark\..*?\s*def\s+(\w+)\s*\([^)]*\):',
        r'@unittest\..*?\s*def\s+(\w+)\s*\([^)]*\):',
        r'def\s+(test_\w+)\s*\([^)]*\):', # unittest style test_* functions
        r'@pytest.fixture\s*.*?\s*def\s+(\w+)\s*\([^)]*\):',
        r'@pytest\.(?:mark\.)?parametrize\s*.*?\s*def\s+(\w+)\s*\([^)]*\):'
    ]
}

# Test indicators specifically for matching file names
self.test_indicators = [
    r'^test', # File starts with 'test'
    r'test$', # File ends with 'test'
    r'Test', # File contains 'Test'
    r'_test$', # File ends with '_test'
    r'^test_', # File starts with 'test_'
    r'_Test$', # File ends with '_Test'
    r'^Test' # File starts with 'Test'
]

```

Fig. 3. Regular expression patterns for identifying test functions and test files across multiple programming languages (Java, C++, C#, JavaScript, and Python). The patterns capture both function declarations and test file naming conventions commonly used in various testing frameworks.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

The Prompt Produces a Binary Output:

As a cybersecurity expert, analyze the provided "Original" and "Revised" code snippets from a commit, along with the commit message and other functions in the same commit. The "Original" code represents the state before the changes, while the "Revised" code represents the state after the changes. **Determine if the changes are focused on fixing vulnerabilities; if so, output 1, otherwise output 0.** The length of the code snippet should not influence your assessment; concentrate on evaluating the logic line by line.

- A score of 0 indicates that the changes made from the "Original" code to the "Revised" code **do not address vulnerability fixes.**
- A score of 1 indicates that the changes made from the "Original" code to the "Revised" code **are aimed at fixing vulnerabilities.**

Commit Message:

{commit}

Original code snippet (code before changes):

{original}

Revised code snippet (code after changes):

{revised}

Here are the other functions in the same commit:

{context}

Fig. 4. The Prompt Produces a Binary Output, Indicating Whether Changes are Vulnerability-Fixing or Not.