

Trace Gadgets: Minimizing Code Context for Machine Learning-Based Vulnerability Prediction

Felix Mächtle¹ , Nils Loose¹ , Tim Schulz² , Florian Sieck¹ , Jan-Niclas Serr¹ ,
Ralf Möller² , Thomas Eisenbarth¹ 

¹Institute for IT Security, University of Lübeck, Germany

²University of Hamburg, Institute for Humanities-Centered Artificial Intelligence, Germany

Email: {f.maechtle, n.loose, florian.sieck, j.serr, thomas.eisenbarth}@uni-luebeck.de,
{tim.schulz, ralf.moeller}@uni-hamburg.de

Abstract—As the number of web applications and API endpoints exposed to the Internet continues to grow, so does the number of exploitable vulnerabilities. Manually identifying such vulnerabilities is tedious. Meanwhile, static security scanners tend to produce many false positives. While machine learning-based approaches are promising, they typically perform well only in scenarios where training and test data are closely related. A key challenge for ML-based vulnerability detection is providing suitable and concise code context, as excessively long contexts negatively affect the code comprehension capabilities of machine learning models, particularly smaller ones.

This work introduces Trace Gadgets, a novel code representation that minimizes code context by removing non-related code. Trace Gadgets precisely capture the statements that cover the path to the vulnerability. As input for ML models, Trace Gadgets provide a minimal but complete context, thereby improving the detection performance. Moreover, we collect a large-scale dataset generated from real-world applications with manually curated labels to further improve the performance of ML-based vulnerability detectors. Our results show that state-of-the-art machine learning models perform best when using Trace Gadgets compared to previous code representations, surpassing the detection capabilities of industry-standard static scanners such as GitHub’s CodeQL by at least 4% on a fully unseen dataset. By applying our framework to real-world applications, we identify and report previously unknown vulnerabilities in widely deployed software.

I. INTRODUCTION

Software vulnerabilities pose a significant threat to organizations in all sectors. Systems with public interfaces are especially vulnerable to attacks. If these systems are hosted in critical environments such as hospitals, banks or transportation systems, their exploitation has a devastating impact [4], [16]. The rising number of web services and internet facing API endpoints leads to an immense attack surface where exposed API endpoints introduce a significant security risk [14].

Protecting against API abuse, especially in web applications, is more critical than ever. Yet, traditional security methods, such as penetration testing, require labor-intensive manual inspection by human experts, making penetration testing approaches costly and time-consuming. Considering the shortage of experts, the industry often turns to automated scanners [26], [53], [2] as a cost-effective alternative or addition. However, these scanners, which rely solely on rule-based systems without deep code understanding, fall short compared

Listing 1. Motivating Example: A Trace Gadget generated from the OWASP Benchmark test case *BenchmarkTest01314*. The original test case spans multiple classes, conditions and functions, but is distilled into a precise, single-function representation via code inlining and the removal of redundant computations, thereby isolating the sink-relevant functionality.

```
1 void TG(HttpServletRequest var0,  
2     HttpServletResponse var1){  
3     String var2 = var0.getParameter(  
4         "BenchmarkTest01314"); // Source  
5     if (var2 == null) var2 = "";  
6     var2 = "INSERT INTO users (username, password) VALUES  
7         ('foo', '" + var2 + "')";  
8     Statement var3 = DBHelper.getSqlStatement();  
9     var3.executeUpdate(var2); // Sink  
10 }
```

to human experts [60]. While they report high true positive (TP) rates, they are often accompanied by high false positive (FP) rates, leading to numerous false incidents and thus induce alert fatigue [30].

With recent successes in the area of natural language processing (NLP) [68], [83], the capabilities of machine learning (ML) models to reason about code have advanced. State-of-the-art models such as *UniXcoder* [29], *Traced* [20], *CodeT5+* [72] or *GPT-4o* [48] report promising results in code understanding [20], [72].

These advances have led to an active area of research focused on leveraging ML techniques to detect vulnerabilities in existing software [52], [47], [38], [15], [80], [55], [51], [43]. The goal is to be able to detect vulnerabilities in real-world datasets without the need to manually define vulnerability patterns. Moreover, ML techniques feature the ability to generalize the data seen during training, enabling the detection of yet unseen vulnerability instances.

However, machine learning techniques struggle with long inputs and excessive noise. Although modern commercial LLMs such as GPT-4o [48] or Claude Sonnet [7] can theoretically handle up to 128K+ tokens, their performance degrades as the input length increases (Section V-A). While most research has focused on modifying the input or exploring different models, little work has been devoted to minimizing the context length for vulnerability prediction [13]. Most approaches represent the input using function-level represen-

tations [24], [63], [84], [55], [80], [51], [84] or program slicing [15], [38]. These representations, however, include code that is irrelevant for the detection of vulnerabilities. Consequently, this paper introduces five requirements for an effective code representation tailored to injection vulnerabilities, a major class of web vulnerabilities [49] where a user-controlled value reaches a vulnerable statement without proper sanitization.

Furthermore, upon developing our vulnerability detection framework, we identified two additional key challenges: *Applicability* to closed and open source applications and transferability to *real-world data*.

Code representation. In machine learning, removing non-essential information from the input is imperative [54]. To this end we introduce a novel code representation that we refer to as *Trace Gadgets (TGs)*. An example is shown in Listing 1. Similar to *Code Gadgets* [38], the code is sliced with respect to a potentially vulnerable statement. However, before slicing, the program is statically traced to further reduce the number of statements for each input (Section III). We demonstrate that Trace Gadgets outperform previous representations, such as Code Gadgets or function-level granularity in the task of detecting injection vulnerabilities and show its superior performance by outperforming the best static scanners from industry [26], [53], [2], by at least 4%. Furthermore, we evaluate two representative methods from related work [56], [38] and show that replacing their original input representation with *TGs* reduces the False Positive Rate by 29-38%.

Broad Applicability. The source code for software applications is not always freely available and users might only have access to binaries or bytecode. Since Trace Gadgets are defined purely by control and data flow semantics, any representation that exposes them can be transformed into *TGs*. To demonstrate the idea, we prototype on JVM bytecode, since the JVM ecosystem (Java, Kotlin, Scala, ...) is among the most widely used [62], [6] and its well-specified bytecode facilitates static analysis. All experiments therefore run on JVM bytecode, covering closed proprietary web applications and open source software. However, porting the framework to other runtimes (e.g., .NET or native x86) mainly requires swapping out the backend analysis passes, the *TG* abstraction itself remains unchanged.

Dataset. Machine learning approaches require large labeled datasets, especially for training. In the context of vulnerabilities in JVM web services, available datasets are either large but synthetic [3], [5], raising concerns about their applicability to real-world programs [51], or small custom datasets from real-world applications [11], which are insufficient for robust training.

However, by focusing on bytecode instead of source code, we can access compiled, production-ready programs. To obtain a large variety of real-world production programs, we pull millions of docker images from Docker Hub and extract the included Jar files. From these Jar files we select those containing web applications and use our framework to extract Trace Gadgets. The *TGs* are labeled with static scanners and a

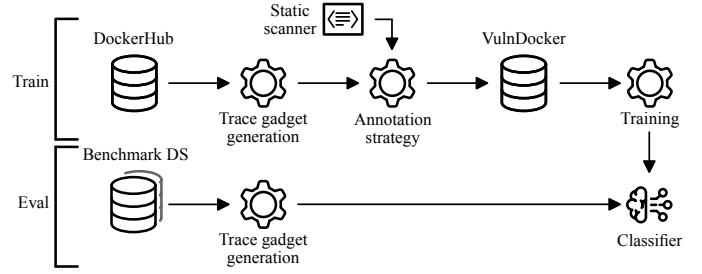


Fig. 1. Systematic overview of the training and evaluation phase.

partial manual verification. Based on this procedure, we create a large training dataset which we refer to as VulnDocker, containing real-world vulnerabilities and code.

New Vulnerability Detection Framework. To emphasize the practical effectiveness of Trace Gadgets and VulnDocker, we developed a prototype framework that we evaluated and successfully compared against related work. The components for training, evaluation, *TG* generation, and the creation of the novel VulnDocker dataset are depicted in Figure 1.

Findings and Responsible Disclosure. Using the proposed framework, we found and reported two new vulnerabilities in web applications we extracted from Docker containers, both of which have been acknowledged by the vendors. The first vulnerability was found in Atlassian’s build server Bamboo and the second one in the open-source project Geoserver. During the responsible disclosure process both vendors released a fix [10], [25].

Contributions.

To summarize, our contributions are:

- We demonstrate that modern Large Language Models, despite theoretically supporting large context sizes, exhibit reduced code comprehension performance for longer inputs.
- We introduce *Trace Gadgets*, a novel code representation that preserves the original program semantics while reducing code length by 28–34% compared to existing code representation. Using *TGs* as input reduces the False Positive Rate by 29–38% in related methods.
- We create a large-scale labeled dataset of Trace Gadgets based on real-world JVM applications retrieved from DockerHub, consisting of 32886 deduplicated samples with manually curated labels.

II. BACKGROUND

A. Program Slicing

Program slicing [73] is a technique used to reduce a program to only those statements necessary for a particular computation, known as the *slicing criterion*. This reduction is achieved by constructing a Program Dependence Graph (PDG), which integrates both the Control Flow Graph (CFG) and the Data Flow Graph (DFG) to represent the dependencies between program statements. By performing a reachability analysis from a given node in the PDG, either forward or backward,

it is possible to determine which statements influence or are influenced by the slicing criterion.

B. Code Gadgets

Li *et al.* introduced *Code Gadgets* as a representation for vulnerability detection [38]. In their approach, program slicing is performed with respect to a list of potentially vulnerable statements. The resulting slices, composed of lines of code, are then assembled into single code snippets suitable for machine learning. Since these code snippets contain all relevant statements from the PDG, they effectively represent all possible executions that could lead to the execution of a particular sink.

C. Program Traces

Program tracing is a technique used to monitor the execution of a program by recording the sequence of statements as they are executed [71]. A *program trace* describes a particular execution path through a program \mathcal{P} consisting of statements s_0, s_1, \dots, s_n , where each s_i belongs to a set of statements \mathcal{S} .

To model the entire program \mathcal{P} , it is necessary to consider the set of all possible execution traces, accounting for every potential path the program might take during execution. This set of traces \mathcal{T} collectively represents the program's behavior and can be expressed as:

$$\mathcal{P} = \bigcup_{t \in \mathcal{T}} t$$

where each trace $t \in \mathcal{T}$ is a sequence of statements from \mathcal{S} corresponding to a possible execution path. Listing 2 shows a simple example function whose traces would be $t_1 = [1, 2, 3, 4, 8, 9]$ and $t_2 = [1, 2, 3, 6, 8, 9]$.

D. Injection vulnerabilities

Injection vulnerabilities are one of the most significant and prevalent security risks in web applications [49]. These occur when attacker-controlled input enters the program at a source statement ($s_k \in \mathcal{S}_{\text{source}}$), such as line 4 in Figure 2, and propagates to a sink ($s_l \in \mathcal{S}_{\text{sink}}$) without proper sanitization [77].

E. Java Vulnerability Datasets

NIST Juliet Java 1.3. The National Institute of Standards and Technology (NIST) provides several vulnerability datasets, including the *Juliet Java 1.3* test suite [3]. The Juliet dataset contains various small vulnerability samples categorized by their Common Weakness Enumeration (CWE). Each vulnerability is documented, classified and labelled. However, the dataset's synthetic nature and small toy examples limit the suitability for training ML models for detecting vulnerabilities in real-world applications. Additionally, Partenza *et al.* [51] observed that the dataset is biased due to different lengths of benign and vulnerable examples. This bias results in artificial discriminating features for ML models.

OWASP Benchmark. The Open Web Application Security Project (OWASP) provides a vulnerability dataset [5], focusing on the top ten web application security risks [49]. Similar to the Juliet test suite, each test case in this dataset is an endpoint annotated with labels indicating its vulnerability status. In

Table I
COMPARISON OF DIFFERENT CODE REPRESENTATIONS

	\mathcal{R}_1	\mathcal{R}_2	\mathcal{R}_3	\mathcal{R}_4	\mathcal{R}_5
Function-level [24], [63], [84], [55], [80], [51], [84]	○	○	●	○	●
Slices (Code Gadgets) [38], [15], [18], [45]	◐	●	○	○	●
Execution Traces [71], [70]	○	●	●	○	◐
Trace Gadgets	●	●	●	●	●

contrast to Juliet, the OWASP Benchmark presents a greater number of execution paths per endpoint, thus offering a more challenging environment for testing security tools.

III. CODE REPRESENTATION

In ML-based vulnerability detection, the code representation plays a vital role. This representation is the basis for the subsequent ML processes. Therefore, the quality and effectiveness of the code representation has a significant impact on the overall performance of the vulnerability detection system. A key challenge is to determine the optimal code representation [82]. This involves carefully considering what information should be included. The representation must be precise, capture relevant code patterns and logic necessary for identifying vulnerabilities while avoiding redundant information that could dilute the model's focus and efficiency. To identify a suitable representation \mathcal{R} , we propose five criteria that should be met:

- \mathcal{R}_1 *Conciseness* - \mathcal{R} should be compact, mimicking the human ability to spot vulnerabilities in shorter code snippets compared to longer ones (Section V-A).
- \mathcal{R}_2 *Completeness* - It is crucial that \mathcal{R} encompasses all vital statements that contribute to pinpointing vulnerabilities. This ensures that the model has all the information necessary for precise and accurate vulnerability detection [74], [45].
- \mathcal{R}_3 *Simplicity* - The representation should aim for simplicity, limiting its structural elements, such as the number of functions or classes, to as few elements as necessary for comprehensibility [34].
- \mathcal{R}_4 *Stylistic Consistency* - Maintaining a consistent programming style in \mathcal{R} is crucial [66], [54], [67]. This uniformity helps the model to focus on the substance rather than the form of the code, thereby reducing the cognitive load on the model and potentially enhancing its effectiveness.
- \mathcal{R}_5 *Computational Efficiency* - While ensuring the quality of the representation remains uncompromised, it is imperative that generating \mathcal{R} is computationally efficient.

A categorization of existing granularities is shown in Table I and explained in the following paragraphs:

Function-level. Often, a function-level granularity is chosen as granularity [80], [55], [47], [51], where each single function is individually analyzed without further context. This granularity is computationally efficient, thereby fulfilling \mathcal{R}_5 - *Computational Efficiency*, and simple due to its limited

Listing 2. Source Program

```

1 void doGet(HttpServletRequest request rq) {
2   String p;
3   if (rq.getParameter("A")!=null) {
4     p = rq.getParameter("A");
5   } else {
6     p = "DEFAULT";
7   }
8   Log.debug("Database update");
9   DB.executeUpdate(p); // Sink
10 }

```

Listing 3. Trace Gadget 1

```

1 void doGet(HttpServletRequest request v1) {
2   String v2;
3   if (v1.getParameter("A")!=null) {
4     v2 = v1.getParameter("A");
5   }
6
7   DB.executeUpdate(v2); // Sink
8 }

```

Listing 4. Trace Gadget 2

```

1 void doGet(HttpServletRequest request v1) {
2   String v2;
3
4   if (v1.getParameter("A")==null) {
5     v2 = "DEFAULT";
6   }
7
8   DB.executeUpdate(v2); // Sink
9 }

```

Fig. 2. Trace Gadget Generation: The logging statement is removed in both gadgets as it does not influence the value that flows into the sink. The IF-statement is split into then and else branches, each represented by one of the two Trace Gadgets.

scope, addressing \mathcal{R}_3 - *Simplicity*. However, it falls short in several areas. By analyzing an entire function, it includes extraneous statements that are irrelevant to the vulnerability being targeted, thus violating \mathcal{R}_1 - *Conciseness*. Additionally, function-level granularity fails to capture relevant statements if a vulnerability spans multiple functions or classes, which compromises \mathcal{R}_2 - *Completeness*. Lastly, without further preprocessing, the code representation still retains the original programming style and variable naming conventions, leading to a violation of \mathcal{R}_4 - *Stylistic Consistency*.

Slices (Code Gadgets). When additional information is provided, such as a *sink* statement s_{sink} , program slicing [73] can be used to reduce the number of statements to those that affect the sink. This technique ensures that only the relevant statements are retained, thereby supporting \mathcal{R}_1 - *Conciseness*. Nevertheless, the approach does not entirely eliminate all execution paths within a single trace, which means it only partially addresses this requirement. Nonetheless, program slicing remains computationally efficient to perform, thereby fulfilling \mathcal{R}_5 - *Computational Efficiency*. When combined with an inter-procedural analysis, all relevant statements are included, thus meeting \mathcal{R}_2 - *Completeness*. However, this broader scope can introduce multiple functions or classes, which compromises simplicity, thereby violating \mathcal{R}_3 - *Simplicity*. Additionally, similar to function-level granularity, without further preprocessing, Code Gadgets still reflect the original programming style, thereby failing to satisfy \mathcal{R}_4 - *Stylistic Consistency*.

Execution Traces. Capturing the execution of a program captures every line of code that is executed, ensuring that all relevant statements are included. This satisfies \mathcal{R}_2 - *Completeness*. In addition, since the trace is derived from a single execution run, \mathcal{R}_3 - *Simplicity* is inherently fulfilled. However, execution traces may contain numerous auxiliary computations that are unrelated to the vulnerability in question, thus failing to meet \mathcal{R}_1 - *Conciseness*. Furthermore, because the recorded source lines are aggregated without any enforced uniformity, does not fulfill \mathcal{R}_4 - *Stylistic Consistency*. Finally, obtaining such traces requires actual program execution, and often requires steering execution toward specific sinks using resource-intensive techniques such as symbolic execution, thereby compromising \mathcal{R}_5 - *Computational Efficiency*.

A. Trace Gadgets

Our novel code representation is developed, aligned with the previously established criteria, by building upon the concept of Code Gadgets, i.e., inter-procedural program slicing [73] used for source to sink vulnerabilities. Therefore, static program tracing is combined with Code Gadgets to generate *Trace Gadgets (TGs)*. By slicing a trace, the number of statements is further reduced while retaining all relevant ones. Furthermore, the code from multiple classes and functions is merged into a single function to simplify further analysis. An example of such a transformation is shown in Figure 2. Instead of a single Code Gadget, this approach results in multiple *TGs* (two in the referenced example). Each *TG* only contains the statements relevant to a particular sink (slicing), in a single function.

Formalization. Formally, the concept of tracing, as introduced in Section II-C, is used as a foundation. However, using the traditional concept of tracing, the number of traces is exponential in the number of all branches as every trace only contains a single control flow path. However, experimental evaluation on web applications have shown that many conditional statements contain only a single branch while the other branch is effectively empty (\emptyset). Such conditions are often used to enforce a boundary. Hence, to avoid exponential behavior for one-sided branches, we include elements that represent one-sided conditional statements, into the analytical model of traces. Formally, we denote an *if-then-else* construct as $(c_j, s_{j+1}^T, s_{j+1}^F)$, where c_j is the condition, s_{j+1}^T is the first instruction of the *then* branch, and s_{j+1}^F is the first instruction of the *else* branch. Thus, we define the extended model S' as:

$$S' = S \cup \{(c_j, s_{j+1}^T, \emptyset)\}$$

As explained in Section II-D, all source to sink vulnerabilities, such as injections, contain a source statement ($s'_k \in S'_{source} \subset S'$) later followed by a sink statement ($s'_l \in S'_{sink} \subset S'$). Hence, only traces containing these two in the correct order are of interest. Notably, we adapt the use of the 'in' operator (\in) to check for the presence of a statement s in a trace t which is a sequence of statements:

$$\exists s'_k, s'_l \in t_i : s'_k \in S'_{source} \wedge s'_l \in S'_{sink} \wedge k < l$$

Additionally, slicing is used to further remove statements within a trace that are irrelevant to the intended sink s'_l . Combining the previously mentioned modification, Trace Gadgets,

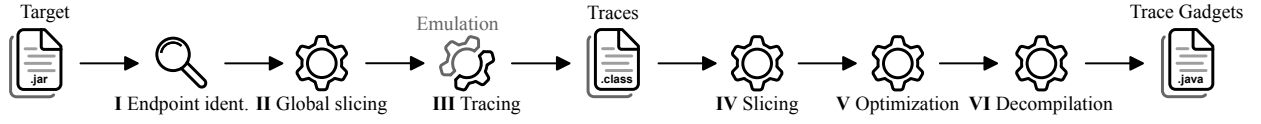


Fig. 3. Systematic overview of the code preparation steps. The Roman numerals denote the paragraphs in Section III-B.

by design, no longer model the entire program \mathcal{P} . Let \mathcal{TG} be the set of all Trace Gadgets. We define:

$$t \in \mathcal{TG} \iff s'_k \in t \wedge s'_l \in t \wedge k < l \\ \wedge \forall s'_i \in t : s'_i \text{ is relevant for } s'_l$$

While this work focuses on JVM-based analysis of injection vulnerabilities in the next sections, the representation can be applied to any programming language and vulnerabilities with a source to sink behavior.

B. Implementation

This section introduces the engine developed to generate Trace Gadgets for JVM-based web applications. Figure 3 shows an overview of our pipeline. A target Java archive (Jar) file is first analyzed to identify all possible endpoints (I). For each endpoint, Global Slicing identifies all functions that occur between a single source and a single sink (II). Each endpoint is then used as an entry point, along with its associated functions from II, to generate traces using a custom static emulation engine (III). Each trace is sliced (IV) before being optimized (V). Finally, the optimized TG is decompiled.

I Endpoint Identification. First, we identify entry point methods for our analysis. Thus, for web applications, we identify the request handling methods, i.e., endpoints, by using regular expressions to find functions with a `HttpRequest` parameter.

II Global Slicing. Given the previously identified endpoints, we run forward slicing on a function granularity basis. Initially, context-insensitive Control-Flow Analysis (0-CFA) [59] is used to quickly reduce the analysis scope. However, since this reduction may be too narrow and miss critical cases, we then use Class-Hierarchy Analysis (CHA) [19] to over-approximate and ensure correctness by covering all possible inheritance scenarios. Given all functions reachable per endpoint, we select those functions that contain a sink statement $s_l \in \mathcal{S}_{sink}$. To determine which statements are potentially vulnerable, we rely on a well-established list [23] by FindSecBugs [2]. For each function containing a sink, we create a backward slice. By taking the intersection of both slices, we retain only the functions on the path between an endpoint and a sink.

III Tracing. Given the previously identified pairs of endpoints and sinks along with their functions in between, i.e., the scope, we statically generate all reachable traces. We utilize the JVM emulation capabilities of the symbolic backend provided by SWAT [41], a dynamic symbolic execution engine for JVM bytecode. For our implementation, we developed a custom emulation engine that interacts with SWAT's execution tracking capabilities to enable accurate handling of

both control (our engine) and data flow (SWAT). This also holds for some dynamic features (i.e., lambda expressions or dynamic invocations) for which we provide wrappers. Pairing the JVM emulation with a static instruction execution module allows us to statically emulate methods without having access to any specific parameters or values determined at runtime. During the static emulation, each conditional branch with an *else* results in an independent duplication of the emulation. However, to ensure completeness and correctness we keep the branching statement and its condition in the TG. The static execution module records all observed instructions for each trace to assemble a single compiled method containing all statements observed during tracing. While this creates an exponential number of traces in the branching depth, it does so only within the narrow limits imposed by the Global Slicing.

IV Local Slicing. To generate our Trace Gadgets from the previously obtained traces, we utilize a self developed static slicing engine. Given the bytecode file of a trace, the engine slices the trace with respect to the potentially vulnerable sink.

V Optimization. To further optimize TGs, we utilize the Proguard Optimizer [28]. This further reduces the number of statements. As we perform all operations on bytecode level, we generate Java source code as model input with IntelliJ's Fernflower decompiler [32].

An example TG generated from an OWASP Benchmark sample is shown in Listing 1. Here, code from different classes and functions are merged into a single function, and unnecessary computations are stripped, leaving a single method that encapsulates the essential functionality relevant to the sink.

C. Limitations of the implementation

Currently, we support all JVM bytecode instructions, except the `throw` instruction.

Exception Handling. The engine executes try-catch blocks without considering potential exceptions, therefore effectively ignoring catch blocks. Moreover, we can not handle execution paths that contain the `throw` instruction. Upon experimental evaluation on real production programs (VulnDocker), this limitation is responsible for 4.5% of endpoints not being emulated, as detailed in Table II.

Internal Functionality. Currently, we only have limited wrapper functions for internal functionalities such as threading or reflection. The tool directly copies most of these functionalities without emulating their behavior. This approach succeeds for functionalities that do not alter program execution, such as string operations, but fails for, e.g., concurrency, as it does not replicate the execution order or resolve dynamic behaviors.

A detailed description of the consequences of these limitations is given in Appendix A.

IV. DATASETS

A diverse and high-quality dataset is essential for effective ML. If the training data misrepresents the target distribution or omits critical cases, the performance of applied algorithms suffers accordingly. For classifiers used beyond the training domain, the dataset must closely approximate real-world data, also implying the existence of both vulnerable and benign samples. Therefore, we created a new large-scale labeled dataset, *VulnDocker*, based on applications extracted from docker containers hosted on DockerHub [1]. To ensure transferability, we decouple the datasets used for training (*VulnDocker*), hyperparameter evaluation (Juliet) and evaluation (OWASP). Evaluating our models on OWASP also ensures comparability to other vulnerability detection frameworks, as this dataset is commonly used for evaluation [51], [26], [2], [58], [53].

A. Benchmark Datasets

As described before, we use the Juliet and OWASP datasets, for hyperparameter evaluation (Juliet) and the overall evaluation (OWASP). Therefore, we preprocess the datasets and filter non-injection vulnerability related samples.

For Juliet, many samples only differ marginally. Hence, after generating *TGs* we substantially reduced redundancy in the dataset. In particular, we found that most were duplicates, resulting in only 587 unique *TGs* from the original set of 8803 test cases. The final processed dataset contains 257 benign and 330 vulnerable *TGs*.

For OWASP, due to the complexity of the dataset, most test cases resulted in multiple *TGs* per endpoint. We obtain 5823 *TGs* from a total of 1572 endpoints. 819 of those endpoints are vulnerable. As this dataset is used for evaluation only, a label per endpoint, rather than per *TG*, is sufficient, as the ground truth of the OWASP dataset is endpoint-based. For evaluation purposes, we aggregate *TG* classifications: an endpoint is deemed vulnerable if the maximum prediction score among all its associated *TGs* surpasses a defined threshold. Only if all *TGs* for an endpoint fall below this threshold (i.e., are classified as benign) the endpoint itself is classified as benign. This strategy enables direct performance assessment against the established endpoint-level ground truth.

B. VulnDocker Dataset

During the training phase, it is essential to utilize a robust dataset designed for ML applications. To allow for arbitrary analysis of JVM programs, including closed-source security analysis, the input to our framework is JVM bytecode. To obtain program code used in production environments, we explored DockerHub [1] as a source of real-world JVM applications. While GitHub offers a wide variety of projects, they require compilation with the correct production configuration. Moreover, compiling arbitrary GitHub projects is a difficult challenge on its own [44]. We collected an extensive list of 7 459 528 unique containers and their respective description file. These description files provide a structured representation of the container's construction. We filter each based on its description for JVM applications, resulting in 9 807 555 Jar files.

Passing these files through the Trace Generation toolchain results in 182 346 unique traces and 32 886 unique *TGs*. The main problem that led from such a large number of Jar files to a comparatively small number of *TGs* was the presence of numerous duplicates.

To train ML models, we needed labels for the collected *TGs*. Since labeling code as vulnerable or benign is challenging without ground truth, we approximated labels using Find Security Bugs (FindSecBugs) [2], a state-of-the-art static scanner [50].

However, since static scanners tend to overpredict [50], i.e., have a high false positive rate, we manually reviewed all 10 610 supposedly vulnerable samples. To ensure accuracy, we followed the recommendation by Dorner *et al.* [21] and allocated effort to additional single reviews. After this extensive process, we ended up with 1 412 truly vulnerable samples. Relying on FindSecBugs labels for benign samples introduces the risk of false negatives, where vulnerable code is mislabeled as benign. Therefore, we additionally reviewed a subset of 500 random benign samples. Of these, only 5 samples were identified as false negatives. Thus, while we acknowledge the possibility that some vulnerable samples may be mislabeled as benign, such occurrences appear to be rare.

V. EXPERIMENTS AND RESULTS

In the previous sections, we introduced Trace Gadgets (*TGs*) and *VulnDocker* with the goal to improve the ability to automatically scan applications for software vulnerabilities. To evaluate the effectiveness of *TGs*, the collected dataset and our proposed framework in its entirety, we devised the following research questions:

- RQ1** How does the length of input context affect the code comprehension performance of modern LLMs? (Section V-A)
- RQ2** What is the efficiency and effectiveness of our Trace Gadget generation engine? (Section V-B)
- RQ3** Can state-of-the-art ML models be effectively combined with Trace Gadgets for vulnerability classification on unseen datasets? (Section V-D)
- RQ4** Can the minimal context of Trace Gadgets enhance the overall vulnerability detection performance of state-of-the-art? (Section V-E)
- RQ5** Do Trace Gadgets fulfill the proposed code representation requirements for vulnerability detection? (Section V-F)

A. LLM Performance and Code Context Size

Modern Large Language Models (LLMs) claim to support context lengths in excess of 128,000 tokens. However, recent studies on question-answering tasks indicate a performance degradation of LLMs as input sizes increase [39]. To the best of our knowledge, this issue has not been studied in the context of program code.

Experiment Design. To evaluate how effectively LLMs handle increasing code context lengths, we used the CRUXEval dataset [27]. CRUXEval tasks an LLM with reasoning

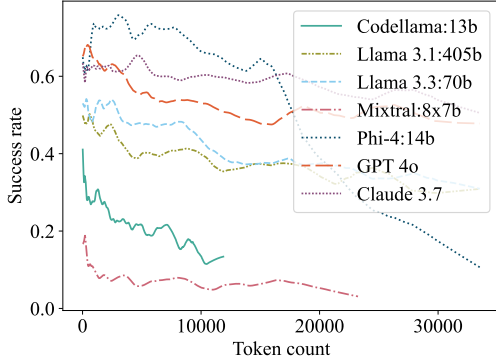


Fig. 4. LOWESS smoothed success rate of various LLMs at predicting the output of a function, depending on the context length of the input.

about program execution, by providing a single function and its input, while requiring the model to predict the function’s output. To increase the context length beyond a single function, we iteratively appended additional distractor functions from the dataset. In each iteration, the LLM was prompted with the original CRUXEval output-prediction prompt and a randomized sequence of distractor functions and the target function, along with the corresponding input. To avoid biases from harder or easier tasks, we repeated this experiment 100 times per model, each time selecting a different target function. Crucially, all models evaluated received identical target and distractor functions across corresponding experimental iterations, ensuring comparability.

Our experimental setup closely follows the methodology outlined by Liu *et al.* [39]. Specifically, they found that placing task-relevant instructions at the end of the prompt led to improved performance in the document retrieval setting. We adopted this configuration in our experiments. In addition, consistent with Liu *et al.*, we randomized function names across iterations to ensure that all functions were equally difficult to find.

To establish a conservative performance baseline, each trial contained exactly one target function with no interactions with other functions, providing an upper bound on performance by eliminating the complexity of reasoning about inter-function interactions.

We applied this experimental design to two state-of-the-art commercial closed-source models and five open-source models of varying size and specialization. In addition, we evaluated WizardCoder [42]. However, due to its consistent failure to capture the task instructions, often modifying or explaining the functions rather than predicting their outcomes, it achieved less than a 5% success rate without any distractor functions. Therefore, WizardCoder was excluded from further comparative analysis.

Results. Figure 4 plots the LOWESS-smoothed [17] average success rate of different LLMs on the CRUXEval output prediction task against the token count of the input prompts. At lower token counts, all evaluated LLMs achieve their highest

Table II
TRACE GADGET GENERATION RESULTS IN PROPORTION TO THE NUMBER OF ENDPOINTS. THE “THROW” AND “ERROR” COLUMNS REFER TO LIMITATIONS OF THE ENGINE.

Dataset	Successful	Unsuccessful		
		Timeout	Throw	Error
VulnDocker	0.894	0.054	0.045	0.007
Juliet	0.989	0.011	0	0
OWASP	0.991	0.009	0	0

success rates, clearly demonstrating their ability to process short prompts and accurately predict the correct function output. Nonetheless, we can clearly see a performance gap between the closed-source models and most of the open-source models. As the token count grows, we observe a gradual decline in performance of around 10 percent success rate. Especially small models with fewer parameters are heavily affected by longer contexts. While Claude 3.7 and GPT 4o are almost able to maintain their abilities, Phi-4 drops from being the best model with over 70% success rate to being one of the worst models with around 10% success rate. Furthermore, Codellama and Mixtral faced practical limitations in running the full experimental setup due to their maximum context lengths of 16k and 32k tokens, respectively, as shown in Figure 4.

Discussion. These observations show that long contexts hurt code comprehension accuracy, particularly for smaller models. Although modern LLMs theoretically support extensive prompt lengths, their performance drops when dealing with excessively long prompts. This performance degradation is likely due to difficulties in maintaining relevant long-range dependencies or to constraints inherent in their internal architectures. Consequently, these findings underscore the importance of developing methods to minimize input size to maintain high model accuracy.

RQ1: Increasing context length has a negative impact on the code comprehension capabilities of LLMs, with smaller models exhibiting significantly greater performance degradation than larger ones.

B. Trace Gadget Generation Efficiency

To answer **RQ2**, we first evaluate the efficiency of our Trace Gadget generation in this section as well as the overall quality of the *TGs* in the next.

Experiment Design. To evaluate the efficiency of the trace generation, we consider two metrics: the proportion of endpoints for which *TGs* can be successfully generated and the average and median time required for trace generation per endpoint. We evaluate these metrics on three datasets: OWASP, Juliet and VulnDocker. For the Juliet and OWASP datasets, we consider all available endpoints. Due to the size of the VulnDocker dataset, we randomly sample 1000 Jar files from the dataset and generate *TGs* for one random endpoint of each

Jar file. The aggregated results for the success rate of Trace Gadget generation are shown in Table II. *TGs* could not be generated successfully if the processing time exceeds a timeout of 5 minutes or if an error occurs within the engine.

Results. The performance across both benchmark datasets is excellent, with *TGs* being successfully generated for around 99% of all endpoints. However, for the VulnDocker dataset the success rate drops to 89.4%. The 11.6% of endpoints for which no *TGs* could be generated can be attributed to 5.4% timeouts and 5.2% premature terminations. The 5.2% of premature terminations are attributable to technical constraints, primarily due to shortcomings in our engine’s handling of the JVM’s `throw` instruction (4.5%). The remaining 0.7% of errors in the engine are caused by other limitations.

To assess the time complexity, we analyzed the generation times of VulnDocker endpoints that produced at least one *TG*. We observed that on our machine equipped with a Neoverse-N1 CPU, the median time to generate a *TG* per endpoint is approximately 11.03 *s* whereas the mean is 63.07 *s*.

Discussion. For *TG* generation on our real-world VulnDocker dataset we observe only 5.4% of endpoints resulting in timeouts. Thus, the computational load of generating *TGs* remains controlled, despite concerns about exponential growth. This control is primarily achieved by applying global slicing (Step II of our toolchain) prior to trace generation, which limits growth to paths from a single source to a single sink. Without global slicing, i.e., when all target functions are in scope, only 55% of traces are successful.

Concerning the evaluation of the overall computation time for *TG* generation for a typical web application, we further evaluated the average number of endpoints per web application and the average number of *TGs* per endpoint. In the VulnDocker dataset, the number of endpoints per web application, are 4.9 in median and 57.54 on average. These endpoints result in a median of 4.0 and a mean of 24.41 *TGs*. Therefore, combining the numbers with the median or average generation time of a *TG*, our analysis for a typical (median) web application is rather fast. Some applications take longer to generate *TGs*, as indicated by the averages. However, even applications with twice the average number of endpoints complete in about an hour, which is significantly faster than a manual review by an expert.

C. Correctness of Trace Gadget Generation

Evaluating the correctness of the generated *TGs* is crucial for assessing the reliability and validity.

Experiment Design. To assess the correctness of *TG* generation, we use 1000 random test cases from a subset of the Juliet dataset. The subset consists of all test cases that execute deterministically, do not include resource exhaustion vulnerabilities and do not print memory addresses.

For each of these 1000 test cases, we compare the output of the original test case with the output of the corresponding generated *TGs*. If any *TG*’s output matches the actual execution, we consider the semantics to be correct.

Results. Our tracing engine successfully replicated the output in 937 instances, yielding a success rate of 93,7%. Upon manual inspection of the rest, we found that 38 cases were caused by known limitations by the engine as explained in Section III-C. A detailed analysis of the erroneous test cases is given in Appendix A.

RQ2: *TGs* for JVM-based web applications can be efficiently generated with our framework. All context required for proper functionality is maintained.

D. Machine learning-based Vulnerability Detection

To answer **RQ3**, we first select three suitable models that report state-of-the-art performance on code comprehension tasks. After model selection, we present our experiment design for the training phase and report the performance of each model on our dataset.

Model Selection. To evaluate the effectiveness of *TGs*, we fine-tune a pre-trained ML model from literature using our prepared datasets. While related work [38], [15], [80] presents freshly trained models or slightly improved architectures, we argue that the level of code-understanding achieved by pre-trained models cannot easily be achieved by training a model from scratch for the task at hand [75]. Consequently, we consider the design of a new architecture for the task of vulnerability detection out of scope and focus on fine-tuning models pre-trained with code understanding. Especially, given the very precise but also very small dataset crafted for vulnerability detection, it is required to utilize a model that has been trained for general code understanding [75]. Although, all selected models have undergone pre-training on a range of tasks related to code comprehension, it is noteworthy that none of these tasks addressed vulnerability detection.

Specifically, we selected three state-of-the-art models for our evaluation. We evaluate the performance of all three to determine the model best suited for our task. All chosen models are transformer-based [68], and have been trained on a number of code-related tasks in order to gain code understanding. Concretely, we picked *UniXcoder* [29], *CodeT5+* [72] and *Traced* [20], due to their performance results [72], [20], [46].

Experiment Design. Our training procedure consists of three training steps shown in Figure 5 and aligns with the three prepared datasets. This experiment focuses on the SQL-injection subsets of the evaluation datasets (i.e., OWASP and Juliet), as SQL injection attacks are the predominant vulnerability in the OWASP dataset. Moreover, Partenza *et al.* [51] also use the OWASP SQL injection subset. We begin by fine-tuning a pre-trained model with our VulnDocker dataset, enabling it to understand the underlying code structures and common patterns of injection attacks. To allow for hyperparameter selection we use a grid search with respect to the F1 score. We evaluate each set of hyperparameters on the subset of SQL injection examples of the unseen Juliet dataset. Applying an unseen dataset for hyperparameter selection enhances transferability. More information regarding

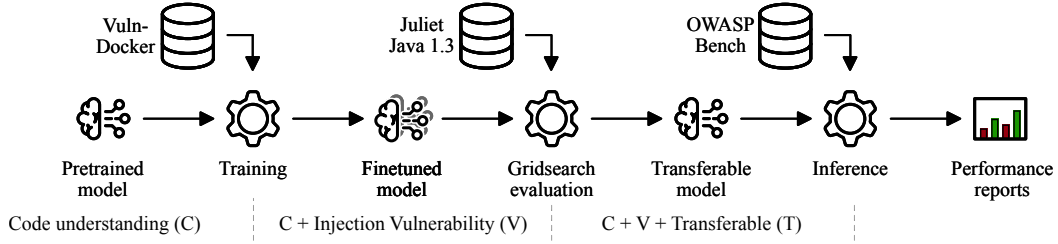


Fig. 5. Systematic overview of the training procedure (middle) alongside all utilized datasets (top) and the model capabilities (bottom).

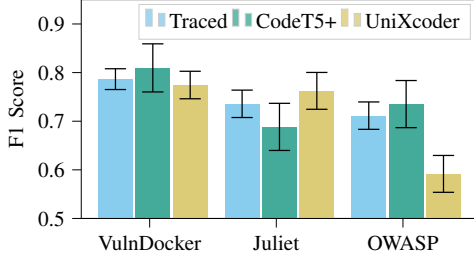


Fig. 6. The F1 scores for each model across all three datasets across 10 runs. The VulnDocker dataset was used as the training source and the hyperparameters of the models were tuned using the Juliet dataset. Consequently, the OWASP dataset represents an entirely unseen dataset in this evaluation.

the selection of hyperparameters can be found in Appendix C. Finally, we evaluate the fine-tuned models on the OWASP benchmark dataset due to its comparability with industry-grade scanners. To evaluate the performance on the OWASP dataset, we use the ground truth labels of the benchmark dataset for each endpoint. If any TG within an endpoint is identified as vulnerable, we classify the entire endpoint as vulnerable. This approach allows us to evaluate the models’ detection performance against the OWASP benchmark’s endpoint-based vulnerability labels.

Results. The previously conceptualized experiments were run against all three models. The F1 score of each model, including the standard deviation, is calculated over ten runs. All models were evaluated on an unused portion of the VulnDocker dataset, the SQL injection subset of the Juliet dataset, and the SQL injection subset of the OWASP dataset. The results are shown in Figure 6. As expected, all evaluated models achieve the best performance on the training dataset with an average F1 score between 0.77 and 0.81. The F1 scores reported for Juliet are expectantly a bit lower ranging from 0.69 to 0.74. When evaluating all models on the unseen OWASP dataset, the performance of Traced and CodeT5+ remains at a comparable level, with an average F1 score of 0.71 for Traced and 0.73 for CodeT5+. This highlights the generalizability of these two models under the given training procedure. With an F1 score of 0.59, the trained UniXcoder model shows a notably lower performance. As UniXcoder and Traced share the same architecture, this can potentially be explained with additional pre-training tasks in Traced [20].

RQ3: All three models show robust performance on datasets other than VulnDocker, with two models matching their performance between the hyperparameter selection dataset and the fully unseen dataset.

E. Comparison with State-of-the-Art

To evaluate the effectiveness of Trace Gadgets compared to other input representations presented in previous work, we selected two widely used static vulnerability detection approaches that exemplify different input representations. Although there are alternative approaches for vulnerability detection, the two major input representations are code slices or function-level input. Thus, we select two representative approaches for each input representation. The first approach is VulDeePecker [38], which pioneered the application of deep learning for vulnerability detection. The second approach by Shestov *et al.* [56] uses modern LLMs for the same task. To compare not only with academic work, we also evaluate four industry-leading static scanners in our experiment, i.e., ShiftLeft Scan (v2.1.1) [58], CodeQL (v2.19.0) [26], FindSecBugs (v1.12.0) [2] and Semgrep (v1.86.0) [53]. Complementing the evaluation with our own setup, we include our best model from the previous research question, i.e., CodeT5+ together with TGs.

Experiment Design. We evaluate each approach using both its original input representation and TGs to compare their performance. Additionally, we contrast the median number of tokens for all representations.

The authors of VulDeePecker did not publish their model nor their data preprocessing implementation. To compare VulDeePecker to our vulnerability detection approach, we use an open-source implementation of VulDeePecker [33]. To replicate the results presented by Shestov *et al.* [56], who use function-level granularity with a fine-tuned LLM (WizardCoder [42]), we refrained from fine-tuning due to computational resource constraints and the substantial complexity associated with fine-tuning LLMs. Instead, we used GPT-4o, which, as shown in Section V-A, achieves significantly better performance compared to WizardCoder. Details about this setup can be found in Appendix B. Additionally, in order to evaluate all code representations consistently, GPT-4o was tested with VulDeePecker’s CodeGadgets.

Code Gadgets were extracted with WALA [69] and functions with Tree-Sitter [65]. Consistent with our approach in the

previous section, all approaches requiring training were trained on the Juliet dataset. All evaluations in this section were performed on the SQL injections part of the OWASP Benchmark. VulnDocker could not be used for training as it consists only of Trace Gadgets. Extracting other representations would have required rerunning the entire pipeline for a large amount of Jar files, including the manual verification of all labels. However, we argue that using the same dataset for training allows for a fair comparison between the input representations.

The results for all static scanners were retrieved using the scripts included with the OWASP benchmark.

The metrics for all evaluations are the True Positive Rate (TPR) and False Positive Rate (FPR). These are the suggested metrics for the OWASP dataset [5]. The results are shown in Figure 7.

Performance. When the original VulDeePecker model is trained with CodeGadgets, it achieves a moderate performance, i.e. a True Positive Rate (TPR) of 0.96 and a False Positive Rate (FPR) of 0.93. However, when CodeGadgets are replaced with *TGs*, VulDeePecker’s performance improves significantly. Although the TPR drops only slightly to 0.84, the FPR drops dramatically to 0.68, a 29% decrease, making the model much more practical for vulnerability detection.

Comparing the results of the approach by Shestov *et al.*, we find that combining Code Gadgets as well as function-level granularity as input representation together with GPT-4o, both result in a TPR and FPR of 1, which is effectively random guessing. However, replacing the representation with *TGs* significantly improves performance, achieving a TPR of 0.95 and an FPR of 0.62, thereby reducing the FPR by 38%, rendering *TGs* together with GPT-4o the best combination.

The static scanners ShiftLeft Scan, CodeQL, and FindSecBugs have excellent TPRs of 1.00. However, this high sensitivity comes at the cost of elevated false positive rates: ShiftLeft Scan reports an FPR of 0.81, CodeQL yields 0.89, and FindSecBugs 0.90. For readability reasons, only the best-performing scanner, i.e., Semgrep, is shown in Figure 7. In comparison to ShiftLeft Scan, Semgrep achieves a slightly better result, by having a lower TPR of 0.86 but also a lower FPR of 0.60.

Upon comparing the best static scanner, i.e., Semgrep, and the best ML based scanner, i.e., *TGs* combined with GPT-4o, with our approach, i.e., CodeT5+ trained on VulnDocker, we see that Semgrep achieves the worst F1 score of 0.74, whereas CodeT5+ with *TGs* has a higher FPR of 0.73, but also a substantially higher TPR of 1, resulting in a F1 score of 0.76. The combination of *TGs* with GPT-4o outperforms all other scanners with an F1 score of 0.77, beating Semgrep’s F1 score by 4% and slightly exceeding CodeT5+’s F1 score by 1%.

Token count. The use of *TGs* not only improves detection accuracy but also results in the lowest token count after preprocessing. When comparing the preprocessed token count of the Juliet dataset, we observe that function-level granularity, despite lacking complete information, produces the highest number of tokens, with a median of 178. Code slices, such as

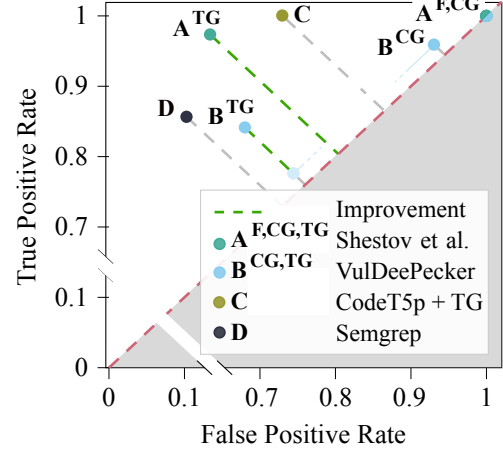


Fig. 7. True Positive Rate (TPR) and False Positive Rate (FPR) on the OWASP benchmark with clipped axis. The letters A, B, D refer to the approaches from related work and C to our approach from RQ3. Superscripts indicate the code representation, i.e. functions (F), Code Gadgets (CG), or our Trace Gadgets (TG). Green dashed lines show how switching to *TGs* improves performance.

CodeGadgets utilized by VulDeePecker, have a median token count of 165. On the other hand, *TGs*, have the lowest median token count of 118. Thus, *TGs* reduce the token count by 28–34% compared to existing approaches.

Discussion. Although GPT-4o may have encountered the OWASP dataset during training, our results show that using unmodified function-level code, which is closer to GPT-4o’s training data, results in worse detection performance compared to highly restructured *TGs*. A plausible explanation is that the OWASP dataset contains many apparent vulnerabilities that are unreachable due to broken flows. Such broken flows may not be adequately detected by GPT-4o. On the other hand, *TGs* implicitly capture those broken flows that render many vulnerabilities in the OWASP dataset inaccessible, thus substantially reducing the FPR.

Furthermore, it is notable that the superior performance of *TGs* with GPT-4o is not caused by the commercial model, but by the code representation: Upon comparing the different input representations (A, B, E) in Figure 7, it can be seen that GPT-4o with functions and slices results in a worse detection performance than with *TGs*.

RQ4: *TGs* improve the detection performance of ML-based vulnerability detection, compared to Code Gadgets or function-level granularity. Moreover, *TGs* contain the fewest statements while being complete.

F. Trace Gadgets and the Code Representation Requirements

In Section III we identified five requirements for code representations that have to be met for vulnerability detection. To assess whether *TGs* fulfill all five requirements, we revisit each and check whether they are fulfilled:

\mathcal{R}_1 - **Conciseness.** By including only statements necessary for a single execution path, *TGs* are highly concise, thus satisfying the conciseness requirement. This is underlined by the results in Section V-E, where *TGs* contain the fewest statements compared to other code representations.

\mathcal{R}_2 - **Completeness.** Despite their conciseness, *TGs* are complete because they contain all necessary statements, as shown in Section V-C, by successfully replicating most program outputs.

\mathcal{R}_3 - **Simplicity.** By design, *TGs* only consist of a single function representing the execution trace. Hence, they also fulfill the simplicity requirement.

\mathcal{R}_4 - **Stylistic Consistency.** The use of an optimizer and a decompiler in the generation of *TGs* standardizes the naming and coding style. This uniformity eliminates stylistic variations, satisfying the requirement for stylistic consistency.

\mathcal{R}_5 - **Computational Efficiency.** While the approach theoretically allows for exponential growth, in practice the Trace Gadget generation process remains computationally efficient, with a median generation time of 11 s per trace, as demonstrated in Section V-B.

RQ5 Trace Gadgets meet all requirements.

G. Evaluation on Real-World Targets

To evaluate our proposed setup beyond laboratory conditions, we applied it in real-world scenarios.

Experiment Design. From the list of Docker containers as described in Section IV-B we select those containers with over one million pulls and extract their Jar files. For those Jar files, we applied our described pipeline from Figure 3 together with a CodeT5+ model trained on VulnDocker.

Results. Our investigation revealed two previously unknown vulnerabilities in popular Java-based web services. The nature of our framework enables us to scan not only open-source applications but also closed-source Jar files.

First, we identified a cross-site scripting (XSS) vulnerability in Atlassian Bamboo, a widely deployed build server, with over 1.5 million pulls on DockerHub. Specifically we discovered a stored XSS vulnerability resulting in a privilege elevation. By exploiting this vulnerability, we were able to create a new user account with administrative privileges. After we reported this vulnerability, it was acknowledged and fixed [10] by the vendors.

Our second discovery was a server-side request forgery (SSRF) in Geoserver, an open-source project for sharing and editing geospatial data. This vulnerability has been concurrently reported by us and an independent security researcher. The vendor acknowledged both reports and released a fix [25].

VI. RELATED WORK

Vulnerability detection has a long history in the security community. Various code analysis methods from different areas have been applied to this challenge, ranging from rule-based approaches such as taint analysis [9], to formal methods

such as symbolic execution [35], random testing through fuzzing [79], and function similarity [76]. In the last decade, most of these methods have seen success through their joint application with machine learning techniques [38], [71], [70], [45]. In particular, the use of artificial code understanding can bridge the gap to accurate vulnerability classification based on representations extracted by classical analysis methods. As a result, two orthogonal areas of improvement have emerged: one focused on improving code representations through targeted refinements to code analysis methods, and the other focused on enhancing the capabilities of machine learning models by optimizing their architecture, size, and training procedures.

Program Slicing. Li *et al.* introduce deep learning-based vulnerability detection [38]. As an input representation, they use Code Gadgets, which are static program slices extracted with respect to potentially vulnerable API calls. Code Gadgets capture the data-flow dependencies leading to a vulnerable statement, but include multiple possible execution paths and thus extraneous statements, introducing noise into the learning process and thus violating \mathcal{R}_1 - *Conciseness*. Several works such as SySeVR [37] or Snopy [12] extended upon Code-Gadgets by computing both forward and backward slices. Although such extensions ensure that both causes and consequences of vulnerabilities are included, they may be redundant, especially for injection vulnerabilities where forward slices (beyond the sink) add little value because the injection has already occurred. Other approaches explored custom slicing strategies tailored to the vulnerability type, such as combining multiple sinks with forward and/or backward slicing [74].

Unlike traditional slicing approaches that capture multiple paths, *TGs* only contain a single execution path, thereby eliminating irrelevant code. In addition, *TGs* produce a single, inlined function rather than multiple functions or classes, a property that traditional slices do not offer, thus failing \mathcal{R}_3 - *Simplicity*. Additionally, all slicing approaches don't use further preprocessing, thus they still reflect parts of the original programming style, thereby violating \mathcal{R}_4 - *Stylistic Consistency*.

Graph-based Models. Other works, e.g., DeepWukong [15] or GLICE [18], use one of the presented slicing methods but change the underlying model architecture to Graph-based models, to better capture the graph structure of code. Thus, graph-based node classification can be used to determine the vulnerable node, which can then be mapped to a single vulnerable line of code [45]. Another possibility to retrieve the vulnerable line is proposed by Li *et al.* [36], who use GNNExplainer [78] to get an explanation, in the form of a small subgraph, which causes the graph model to predict the graph as vulnerable. This work is extended by Steenhoek *et al.* [63], who propose DeepDFA, a framework that employs an abstract dataflow embedding using a custom one-hot encoding. Another extension introduced by Zhang *et al.* [81] is the integration of both code graphs and raw source code textual sequences. This is extended by Wu *et al.* [75], who use a single model for the graph and the textual sequences. They

also discovered, that pretraining is beneficial for the task at hand. Hence, our decision to utilize pretrained models for our evaluation. The interested reader is referred to [82] for a comparison between different model architectures for the task of vulnerability prediction.

However, the change of model architecture is orthogonal to our approach, as Trace Gadgets could serve as an intermediate preprocessing step for graph-based models, effectively reducing its graph size, aiding \mathcal{R}_1 - *Conciseness*.

Execution Traces. LiGer [71] adopts a hybrid approach by using both symbolic execution traces and actual execution traces. By combining these two sources, the authors achieve state-of-the-art results in name prediction and semantic classification. Similar to LiGer, Concoction [70] uses dynamic symbolic traces in conjunction with function-level code, effectively combining static analysis with dynamic execution. Using a custom driver, their approach directs symbolic execution towards functions of interest within the target program. The resulting function source code and dynamic symbolic trace are then used as input for classification.

In contrast, our method statically derives a single new function that represents the compressed execution, without noise and redundancy introduced by capturing all executed statements. However, *TGs* could also complement Concoction. For example, replacing function-level source code in Concoction’s workflow with *TGs* could further streamline the representation, thereby satisfying \mathcal{R}_1 - *Conciseness* and \mathcal{R}_4 - *Stylistic Consistency* requirements that their current approach does not satisfy. In addition, execution traces inherently rely on dynamic execution of the problematic function, which is fundamentally different from our purely static approach. This methodological difference is the reason we did not include execution traces in our evaluation in Section V-E.

Function-Level Representations. Many approaches [24], [63], [84], [55], [80], [51], [84] rely on using entire functions as input representation. While function-level granularity is computationally efficient, it often includes many irrelevant statements that introduce noise into the learning process, thereby violating \mathcal{R}_1 - *Conciseness*. Additionally, they are incomplete and might lack important details [74], [45], thus they don’t fulfill \mathcal{R}_2 - *Completeness*. Moreover, all named approaches don’t use further preprocessing, thus violating \mathcal{R}_4 - *Stylistic Consistency*. *TGs*, on the other hand, don’t suffer from these limitations.

A. LLM-based Vulnerability Detection

The rise of Large Language Models (LLMs) has opened new avenues for vulnerability detection. Recent works by Khare *et al.* [34] and Shestov *et al.* [56] use LLMs, either off-the-shelf or fine-tuned, to predict vulnerabilities. While these approaches are promising, our experiments indicate that even modern LLMs work better with shorter input sizes (Section V-A). However, even with short inputs, we showed that LLMs are sensitive to the quality of the input representation. When provided with traditional function-level code or Code

Gadgets, even powerful models such as GPT-4o may exhibit high false positive rates.

B. Java Vulnerability Detection

With Data Snooping. When focusing on JVM-based languages, Achilles [55] is one of the first deep learning systems for Java. It predicts vulnerabilities, with a function-level input, similar to our experiment in Section V-E, but with a different model, i.e., a Long Short-Term Memory network (LSTM) [31]. ISVSF [80] works similarly to Achilles, but instead of passing the code of a method in vectorized form directly to an LSTM, they first convert each basic block into a vector. The concatenation of those vectors is then used during classification.

All previously described systems rely on a single dataset for training and evaluation, thus risking Data Snooping [8]. We depart from this approach, ensuring robust results and transferability.

Without Data Snooping. In contrast, Partenza *et al.* [51] did not train and evaluate on the same dataset. Their approach uses a sequential version of the program’s Abstract Syntax Tree (AST) together with the corresponding source code. After training on Juliet, they evaluated this model using the SQL injection subset of the OWASP Benchmark dataset. While having an excellent F1 score of 0.93 in recognizing SQL injections in the Juliet dataset, they only gain a F1 score of 0.57 on the SQL injection subset. Hence, our F1 score on the unseen OWASP Benchmark of 0.76 / 0.77 (Section V-E) reflects a performance improvement of 33 – 35%. Similarly, Mamede *et al.* [43], used a pre-trained version of CodeBERT [22], to identify vulnerabilities using the function’s source code and categorize the vulnerabilities according to their Common Weakness Enumeration (CWE). While their approach showed impressive performance achieving F1 scores as high as 0.94 on the Juliet dataset, they reported a significant performance drop of 50-70% in F1 scores when evaluated on a custom, real-world dataset. Unlike all the aforementioned approaches, our approach does not require the source code of the target because it works on bytecode.

VII. THREATS TO VALIDITY

In this section, we discuss potential threats to the validity of our approach and outline measures taken to mitigate them.

A. Cross-Language Applicability

Our implementation and the resulting VulnDocker dataset are currently focused on JVM-based languages, as the prototype works directly on JVM bytecode. While languages such as C/C++ are not included, our approach is generally applicable to any programming language with source-to-sink vulnerabilities and analyzable control and data flow. Adapting the approach to other languages would primarily require a reimplementing of core analysis components, a task whose feasibility is supported by the availability of such analysis equivalent tools [57].

While the application of machine learning in the security domain promises great potential, the validity and interpretability of the results highly depend on a thoughtful design of the workflow and avoiding common pitfalls [8].

Sampling Bias: While the underlying true distribution of injection vulnerabilities is unknown, we identified biases in existing benchmark datasets commonly used for training and evaluation. To avoid these biases, we created VulnDocker, a large-scale dataset based on real applications to minimize the sampling bias.

Label Inaccuracy: With no available ground truth labels for our collected samples, we need to rely on an approximation for labeling. To ensure the label inaccuracy does not lead to unrealistic results, we strictly evaluate our approach on the benchmark dataset where the ground truth labels are known. However, possible inaccuracies introduced during labeling could lead to a reduction on performance. To avoid this, we manually inspected all samples flagged as vulnerable by the static scanner.

Data snooping: To prevent data snooping, we ensure the OWASP dataset remains completely unseen until the final evaluation. This approach guarantees that our model’s performance is not influenced by prior exposure to the test data.

Spurious Correlations and Biased Parameter Selection: Our model’s susceptibility to spurious correlations is minimized by training on diverse datasets and testing on a completely unseen dataset (OWASP). This practice not only reduces the risk of spurious correlations but also avoids biased parameter selection, as no parameters are tuned based on the unseen dataset.

Inappropriate Baseline and Performance Measures: We establish baselines using both industrial static scanners and academic literature, ensuring a comprehensive comparison with state-of-the-art methods. For performance evaluation, we adopt the metrics from the Benchmark Framework, emphasizing true positive and false positive rates (Figure 7).

Base Rate Fallacy: We report all our metrics using the F1 score, which is sensitive to class imbalances, thus avoiding the base rate fallacy.

Lab-Only Evaluation: Beyond theoretical assessments, we demonstrate our model’s real-world applicability, using a case study of real-world application leading to the identification of previously unknown vulnerabilities in widely used software.

Inappropriate Threat Model: As our approach does not guard against adversarial attacks targeting its ML component, we need to emphasize the potential for such threats. So while Trace Gadgets are well suited for scenarios where the expected vulnerabilities are caused by oversights, in their current form they do not protect against intentionally placed vulnerabilities in scenarios such as supply chain attacks. Several works have demonstrated successful attacks against ML-based code detection engines [40], [61], [64]. Securing our model against malicious actors to enable its utilization in adverse conditions is left as an interesting research topic.

In summary, this research introduces *Trace Gadgets*, a novel minimal code representation specifically designed for identifying injection vulnerabilities. Through comprehensive experimentation, we observed that increasing the length of the code context negatively affects the comprehension capabilities of machine learning models, especially smaller ones. This finding underscores the importance of concise code representations such as Trace Gadgets. Utilizing Trace Gadgets, we compiled a unique large-scale dataset, *VulnDocker*, featuring real-world applications. To create labels for this dataset, we employed a leading industry scanner, namely FindSecBugs [2], and manually checked 10640 potentially vulnerable samples. Using the VulnDocker dataset, we fine-tuned three advanced pre-trained machine learning models [20], [29], [72]. Using those models and a commercial LLM, namely GPT-4o [48], we conducted a comparative analysis against both traditional static scanners and other ML-based frameworks and demonstrated a performance improvement over existing methods by 4–35%. In comparison with other code representations, namely slices (Code Gadgets) and functions, we show that Trace Gadgets achieve superior performance while requiring 28–34% fewer tokens. Ultimately, Trace Gadgets and VulnDocker represent a significant advancement towards the practical application of ML for web-based vulnerability detection.

REFERENCES

- [1] Dockerhub. <https://hub.docker.com/>. Accessed: 11/2023.
- [2] Findsecbugs. <https://find-sec-bugs.github.io/>. Accessed: 11/2023.
- [3] Juliet java 1.3. <https://samate.nist.gov/SARD/test-suites>. Accessed: 11/2024.
- [4] An nsa-derived ransomware worm is shutting down computers worldwide. <https://arstechnica.com/information-technology/2017/05/an-nsa-derived-ransomware-worm-is-shutting-down-computers-worldwide/>. Accessed: 10/2023.
- [5] Owasp java benchmark. <https://owasp.org/www-project-benchmark/>. Accessed: 11/2023.
- [6] Tiobe index. <https://www.tiobe.com/tiobe-index/>. Accessed: 11/2023.
- [7] Anthropic. Claude 3.7 sonnet. <https://www.anthropic.com/claude/sonnet>, Accessed: 04/2025.
- [8] Daniel Arp, Erwin Quiring, Feargus Pendlebury, Alexander Warnecke, Fabio Pierazzi, Christian Wressnegger, Lorenzo Cavallaro, and Konrad Rieck. Dos and don’ts of machine learning in computer security. In *31st USENIX Security Symposium, USENIX Security 2022, Boston, MA, USA, August 10-12, 2022*, pages 3971–3988. USENIX Association, 2022.
- [9] Steven Arzt, Siegfried Rasthofer, Christian Fritz, Eric Bodden, Alexandre Bartel, Jacques Klein, Yves Le Traon, Damien Ocheau, and Patrick D. McDaniel. Flowdroid: precise context, flow, field, object-sensitive and lifecycle-aware taint analysis for android apps. In *ACM SIGPLAN Conference on Programming Language Design and Implementation, PLDI ’14, Edinburgh, United Kingdom - June 09 - 11, 2014*, pages 259–269. ACM, 2014.
- [10] Bugcrowd. Xss: Unescaped output of directory name. <https://bugcrowd.com/disclosures/29e6d45a-8829-42c9-a3b4-683d71318e37/xss-unescaped-output-of-directory-name>, Accessed: 09/2024.
- [11] Quang-Cuong Bui, Riccardo Scandariato, and Nicolás E. Díaz Ferreyra. Vul4j: A dataset of reproducible java vulnerabilities geared towards the study of program repair techniques. In *19th IEEE/ACM International Conference on Mining Software Repositories, MSR 2022, Pittsburgh, PA, USA, May 23-24, 2022*, pages 464–468. ACM, 2022.

- [12] Sicong Cao, Xiaobing Sun, Xiaoxue Wu, David Lo, Lili Bo, Bin Li, Xiaolei Liu, Xingwei Lin, and Wei Liu. Snopy: Bridging sample denoising with causal graph learning for effective vulnerability detection. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering, ASE 2024, Sacramento, CA, USA, October 27 - November 1, 2024*, pages 606–618. ACM, 2024.
- [13] Beatrice Casey, Joanna C. S. Santos, and George Perry. A survey of source code representations for machine learning-based cybersecurity tasks. *CoRR*, abs/2403.10646, 2024.
- [14] Chien-An Chen. With great abstraction comes great responsibility: Sealing the microservices attack surface. In *2019 IEEE Cybersecurity Development, SecDev 2019, Tysons Corner, VA, USA, September 23-25, 2019*, page 144. IEEE, 2019.
- [15] Xiao Cheng, Haoyu Wang, Jiayi Hua, Guoai Xu, and Yulei Sui. Deepwukong: Statically detecting software vulnerabilities using deep graph neural network. *ACM Trans. Softw. Eng. Methodol.*, 30(3):38:1–38:33, 2021.
- [16] CISA. Ransomware activity targeting the healthcare and public health sector. <https://www.cisa.gov/news-events/cybersecurity-advisories/aa20-302a>. Accessed: 11/2023.
- [17] William S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.
- [18] Wesley de Kraker, Harald Vranken, and Arjen Hommersom. GLICE: combining graph neural networks and program slicing to improve software vulnerability detection. In *IEEE European Symposium on Security and Privacy, EuroS&P 2023 - Workshops, Delft, Netherlands, July 3-7, 2023*, pages 34–41. IEEE, 2023.
- [19] Jeffrey Dean, David Grove, and Craig Chambers. Optimization of object-oriented programs using static class hierarchy analysis. In Walter G. Olthoff, editor, *ECOOP’95 - Object-Oriented Programming, 9th European Conference, Århus, Denmark, August 7-11, 1995, Proceedings*, volume 952 of *Lecture Notes in Computer Science*, pages 77–101. Springer, 1995.
- [20] Yangruibo Ding, Benjamin Steenhoeck, Kexin Pei, Gail E. Kaiser, Wei Le, and Baishakhi Ray. TRACED: execution-aware pre-training for source code. *CoRR*, abs/2306.07487, 2023.
- [21] Florian E. Dörner and Moritz Hardt. Don’t label twice: Quantity beats quality when comparing binary classifiers on a budget. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.
- [22] Zhangyin Feng, Daya Guo, Duyu Tang, Nan Duan, Xiaocheng Feng, Ming Gong, Linjun Shou, Bing Qin, Ting Liu, Daxin Jiang, and Ming Zhou. Codebert: A pre-trained model for programming and natural languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1536–1547. Association for Computational Linguistics, 2020.
- [23] FindSecBugs. Injection sinks. <https://github.com/find-sec-bugs/tree/master/findsecbugs-plugin/src/main/resources/injection-sinks>. Accessed: 11/2023.
- [24] Michael Fu and Chakkrit Tantithamthavorn. Linevul: A transformer-based line-level vulnerability prediction. In *19th IEEE/ACM International Conference on Mining Software Repositories, MSR 2022, Pittsburgh, PA, USA, May 23-24, 2022*, pages 608–620. ACM, 2022.
- [25] Geoserver. [geos-11390] replace testwfspost with javascript demo page. <https://github.com/geoserver/geoserver/pull/7672>. Accessed: 11/2024.
- [26] Github. Codeql. <https://codeql.github.com/>. Accessed: 11/2023.
- [27] Alex Gu, Baptiste Rozière, Hugh James Leather, Armando Solar-Lezama, Gabriel Synnaeve, and Sida Wang. Cruxeval: A benchmark for code reasoning, understanding and execution. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.
- [28] Guardsquare. Proguard. <https://github.com/Guardsquare/Proguard>. Accessed: 11/2023.
- [29] Daya Guo, Shuai Lu, Nan Duan, Yanlin Wang, Ming Zhou, and Jian Yin. Unixcoder: Unified cross-modal pre-training for code representation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7212–7225. Association for Computational Linguistics, 2022.
- [30] Wajih Ul Hassan, Shengjian Guo, Ding Li, Zhengzhang Chen, Kangkook Jee, Zhichun Li, and Adam Bates. Nodoe: Combatting threat alert fatigue with automated provenance triage. In *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*. The Internet Society, 2019.
- [31] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.
- [32] JetBrains. Fernflower decompiler. <https://github.com/JetBrains/intellij-community/tree/master/plugins/java-decompiler/engine>. Accessed: 11/2023.
- [33] Johnb110. Vdpython. <https://github.com/johnb110/VDPython>. Accessed: 04/2025.
- [34] Avishree Khare, Saikat Dutta, Ziyang Li, Alaia Solko-Breslin, Rajeev Alur, and Mayur Naik. Understanding the effectiveness of large language models in detecting security vulnerabilities. *CoRR*, abs/2311.16169, 2023.
- [35] James C. King. Symbolic execution and program testing. 1976.
- [36] Yi Li, Shaohua Wang, and Tien N. Nguyen. Vulnerability detection with fine-grained interpretations. In *ESEC/FSE ’21: 29th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Athens, Greece, August 23-28, 2021*, pages 292–303. ACM, 2021.
- [37] Zhen Li, Deqing Zou, Shouhuai Xu, Hai Jin, Yawei Zhu, Zhaoxuan Chen, Sujuan Wang, and Jialai Wang. Sysevr: A framework for using deep learning to detect software vulnerabilities. <https://doi.org/10.21227/fhg0-1b35>, November 2018. Accessed on YYYY-MM-DD.
- [38] Zhen Li, Deqing Zou, Shouhuai Xu, Xinyu Ou, Hai Jin, Sujuan Wang, Zhijun Deng, and Yuyi Zhong. Vuldeepecker: A deep learning-based system for vulnerability detection. In *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018*. The Internet Society, 2018.
- [39] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *CoRR*, abs/2307.03172, 2023.
- [40] Nils Loose, Felix Mächtle, Claudius Pott, Volodymyr Bezsmertnyi, and Thomas Eisenbarth. Madvex: Instrumentation-based adversarial attacks on machine learning malware detection. In *Detection of Intrusions and Malware, and Vulnerability Assessment - 20th International Conference, DIMVA 2023, Hamburg, Germany, July 12-14, 2023, Proceedings*, volume 13959 of *Lecture Notes in Computer Science*, pages 69–88. Springer, 2023.
- [41] Nils Loose, Felix Mächtle, Florian Sieck, and Thomas Eisenbarth. SWAT: modular dynamic symbolic execution for java applications using dynamic instrumentation (competition contribution). In *Tools and Algorithms for the Construction and Analysis of Systems - 30th International Conference, TACAS 2024, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2024, Luxembourg City, Luxembourg, April 6-11, 2024, Proceedings, Part III*, volume 14572 of *Lecture Notes in Computer Science*, pages 399–405. Springer, 2024.
- [42] Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. Wizardcoder: Empowering code large language models with evol-instruct. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.
- [43] Cláudia Mamede, Eduard Pinconschi, Rui Abreu, and José Campos. Exploring transformers for multi-label classification of java vulnerabilities. In *22nd IEEE International Conference on Software Quality, Reliability and Security, QRS 2022, Guangzhou, China, December 5-9, 2022*, pages 43–52. IEEE, 2022.
- [44] Pedro Martins, Rohan Achar, and Cristina V. Lopes. 50k-c: a dataset of compilable, and compiled, java projects. In *Proceedings of the 15th International Conference on Mining Software Repositories, MSR 2018, Gothenburg, Sweden, May 28-29, 2018*, pages 1–5. ACM, 2018.
- [45] Yisroel Mirsky, George Macon, Michael D. Brown, Carter Yagemann, Matthew Pruett, Evan Downing, J. Sukarno Mertoguno, and Wenke Lee. Vulchecker: Graph-based vulnerability localization in source code. In *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, pages 6557–6574. USENIX Association, 2023.
- [46] Felix Mächtle, Jan-Niclas Serr, Nils Loose, Jonas Sander, and Thomas Eisenbarth. Ocean: Open-world contrastive authorship identification. *CoRR*, abs/2412.05049, 2024.
- [47] Van-Anh Nguyen, Dai Quoc Nguyen, Van Nguyen, Trung Le, Quan Hung Tran, and Dinh Q. Phung. Regvd: Revisiting graph neural networks for vulnerability detection. *CoRR*, abs/2110.07317, 2021.

- [48] OpenAI. Hello gpt-4o. <https://openai.com/index/hello-gpt-4o/>. Accessed: 09/2024.
- [49] OWASP. Owasp top ten. <https://owasp.org/www-project-top-ten/>. Accessed: 11/2023.
- [50] Tosin Daniel Oyetoan, Bisera Milosheska, Mari Grini, and Daniela Soares Cruzes. Myths and facts about static application security testing tools: An action research at telenor digital. In *Agile Processes in Software Engineering and Extreme Programming - 19th International Conference, XP 2018, Porto, Portugal, May 21-25, 2018, Proceedings*, volume 314 of *Lecture Notes in Business Information Processing*, pages 86–103. Springer, 2018.
- [51] Garrett Partenza, Trevor Amburgey, Lin Deng, Josh Dehlinger, and Suranjan Chakraborty. Automatic identification of vulnerable code: Investigations with an ast-based neural network. In *IEEE 45th Annual Computers, Software, and Applications Conference, COMPSAC 2021, Madrid, Spain, July 12-16, 2021*, pages 1475–1482. IEEE, 2021.
- [52] Dinglan Peng, Shuxin Zheng, Yatao Li, Guolin Ke, Di He, and Tie-Yan Liu. How could neural networks understand programs? In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8476–8486. PMLR, 2021.
- [53] r2c. Semgrep. <https://semgrep.dev/>. Accessed: 11/2023.
- [54] Niklas Risse and Marcel Böhme. Uncovering the limits of machine learning for automatic vulnerability detection. In *33rd USENIX Security Symposium, USENIX Security 2024, Philadelphia, PA, USA, August 14-16, 2024*. USENIX Association, 2024.
- [55] Nicholas Saccente, Josh Dehlinger, Lin Deng, Suranjan Chakraborty, and Yin Xiong. Project achilles: A prototype tool for static method-level vulnerability detection of java source code using a recurrent neural network. In *34th IEEE/ACM International Conference on Automated Software Engineering Workshops, ASE Workshops 2019, San Diego, CA, USA, November 11-15, 2019*, pages 114–121. IEEE, 2019.
- [56] Aleksei Shestov, Rodion Levichev, Ravil Mussabayev, Evgeny Maslov, Pavel Zadrozny, Anton Cheshkov, Rustam Mussabayev, Alymzhan Toleu, Gulmira Tolegen, and Alexander Krassovitskiy. Finetuning large language models for vulnerability detection. *IEEE Access*, 13:38889–38900, 2025.
- [57] ShiftLeft. Joern - the bug hunter's workbench. <https://joern.io/>. Accessed: 2022-08-7.
- [58] ShiftLeftSecurity. Shiftleft scan. <https://github.com/ShiftLeftSecurity/sast-scan>. Accessed: 12/2024.
- [59] Olin Shivers. Control-flow analysis in scheme. In Richard L. Wexelblat, editor, *Proceedings of the ACM SIGPLAN'88 Conference on Programming Language Design and Implementation (PLDI)*, Atlanta, Georgia, USA, June 22-24, 1988, pages 164–174. ACM, 1988.
- [60] Navneet Singh, Vishtasp Meherhomji, and B. R. Chandavarkar. Automated versus manual approach of web application penetration testing. In *11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020, Kharagpur, India, July 1-3, 2020*, pages 1–6. IEEE, 2020.
- [61] Wei Song, Xueziyang Li, Sadia Afroz, Deepali Garg, Dmitry Kuznetsov, and Heng Yin. Mab-malware: A reinforcement learning framework for blackbox generation of adversarial malware. In *ASIA CCS '22: ACM Asia Conference on Computer and Communications Security, Nagasaki, Japan, 30 May 2022 - 3 June 2022*, pages 990–1003. ACM, 2022.
- [62] Statista. Most used programming languages among developers worldwide as of 2024. <https://www.statista.com/statistics/793628/worldwide-developer-survey-most-used-languages/>, 2024.
- [63] Benjamin Steenhoek, Hongyang Gao, and Wei Le. Dataflow analysis-inspired deep learning for efficient vulnerability detection. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering, ICSE 2024, Lisbon, Portugal, April 14-20, 2024*, pages 16:1–16:13. ACM, 2024.
- [64] Octavian Suci, Scott E. Coull, and Jeffrey Johns. Exploring adversarial examples in malware detection. In *2019 IEEE Security and Privacy Workshops, SP Workshops 2019, San Francisco, CA, USA, May 19-23, 2019*, pages 8–14. IEEE, 2019.
- [65] Tree-sitter. <https://tree-sitter.github.io/tree-sitter/>, Accessed: 2022-11-23.
- [66] Saad Ullah, Mingji Han, Saurabh Pujar, Hammond Pearce, Ayse K. Coskun, and Gianluca Stringhini. Can large language models identify and reason about security vulnerabilities? not yet. *CoRR*, abs/2312.12575, 2023.
- [67] Saad Ullah, Mingji Han, Saurabh Pujar, Hammond Pearce, Ayse K. Coskun, and Gianluca Stringhini. Llms cannot reliably identify and reason about security vulnerabilities (yet?): A comprehensive evaluation, framework, and benchmarks. In *IEEE Symposium on Security and Privacy, SP 2024, San Francisco, CA, USA, May 19-23, 2024*, pages 862–880. IEEE, 2024.
- [68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.
- [69] WALA. The t. j. watson libraries for analysis. <https://wala.sourceforge.net/>. Accessed: 2022-11-29.
- [70] Huaning Wang, Zhanyong Tang, Shin Hwei Tan, Jie Wang, Yuzhe Liu, Hejun Fang, Chunwei Xia, and Zheng Wang. Combining structured static code information and dynamic symbolic traces for software vulnerability prediction. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering, ICSE 2024, Lisbon, Portugal, April 14-20, 2024*, pages 169:1–169:13. ACM, 2024.
- [71] Ke Wang and Zhendong Su. Blended, precise semantic program embeddings. In *Proceedings of the 41st ACM SIGPLAN International Conference on Programming Language Design and Implementation, PLDI 2020, London, UK, June 15-20, 2020*, pages 121–134. ACM, 2020.
- [72] Yue Wang, Hung Le, Akhilesh Deepak Gotmare, Nghi D. Q. Bui, Junnan Li, and Steven C. H. Hoi. Codet5+: Open code large language models for code understanding and generation. *CoRR*, abs/2305.07922, 2023.
- [73] Mark D. Weiser. Program slicing. In *Proceedings of the 5th International Conference on Software Engineering, San Diego, California, USA, March 9-12, 1981*, pages 439–449. IEEE Computer Society, 1981.
- [74] Bozhi Wu, Shangqing Liu, Yang Xiao, Zhiming Li, Jun Sun, and Shang-Wei Lin. Learning program semantics for vulnerability detection via vulnerability-specific inter-procedural slicing. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2023, San Francisco, CA, USA, December 3-9, 2023*, pages 1371–1383. ACM, 2023.
- [75] Hongjun Wu, Zhuo Zhang, Shangwen Wang, Yan Lei, Bo Lin, Yihao Qin, Haoyu Zhang, and Xiaoguang Mao. Peculiar: Smart contract vulnerability detection based on crucial data flow graph and pre-training techniques. In *32nd IEEE International Symposium on Software Reliability Engineering, ISSRE 2021, Wuhan, China, October 25-28, 2021*, pages 378–389. IEEE, 2021.
- [76] Fabian Yamaguchi, Nico Golde, Daniel Arp, and Konrad Rieck. Modeling and discovering vulnerabilities with code property graphs. In *2014 IEEE Symposium on Security and Privacy, SP 2014, Berkeley, CA, USA, May 18-21, 2014*, pages 590–604. IEEE Computer Society, 2014.
- [77] Fabian Yamaguchi, Alwin Maier, Hugo Gascon, and Konrad Rieck. Automatic inference of search patterns for taint-style vulnerabilities. In *2015 IEEE Symposium on Security and Privacy, SP 2015, San Jose, CA, USA, May 17-21, 2015*, pages 797–812. IEEE Computer Society, 2015.
- [78] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9240–9251, 2019.
- [79] Andreas Zeller, Rahul Gopinath, Marcel Böhme, Gordon Fraser, and Christian Holler. *The Fuzzing Book*. CISA Helmholtz Center for Information Security, 2024. Retrieved 2024-07-01 16:50:18+02:00.
- [80] Haibin Zhang, Yifei Bi, Hongzhi Guo, Wen Sun, and Jianpeng Li. ISVSF: intelligent vulnerability detection against java via sentence-level pattern exploring. *IEEE Syst. J.*, 16(1):1032–1043, 2022.
- [81] Yulin Zhang, Yong Hu, and Xiao Chen. Context and multi-features-based vulnerability detection: A vulnerability detection frame based on context slicing and multi-features. *Sensors*, 24(5):1351, 2024.
- [82] Yuting Zhang, Jiahao Zhu, Yixin Yang, Ming Wen, and Hai Jin. Comparing the performance of different code representations for learning-based vulnerability detection. In *Proceedings of the 14th Asia-Pacific Symposium on Internetware, Internetware 2023, Hangzhou, China, August 4-6, 2023*, pages 174–184. ACM, 2023.
- [83] Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, Hao Peng, Jianxin Li, Jia Wu, Ziwei Liu, Pengtao Xie, Caiming Xiong, Jian Pei, Philip S. Yu, and

- Lichao Sun. A comprehensive survey on pretrained foundation models: A history from BERT to chatgpt. *CoRR*, abs/2302.09419, 2023.
- [84] Yaqin Zhou, Shangqing Liu, Jing Kai Siow, Xiaoning Du, and Yang Liu. Devign: Effective vulnerability identification by learning comprehensive program semantics via graph neural networks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 10197–10207, 2019.

APPENDIX

A. Evaluating Trace Gadgets Semantics

In Section V-C, we discuss the observed discrepancies between the output of the generated Trace Gadgets and the original test cases. The following section provides a detailed analysis of all erroneous cases, highlighting two prevalent issues:

The primary problem, affecting 27 test cases, was the use of try-catch blocks in the original test cases. For example, the generated Trace Gadget crashed when attempting to load a configuration file that did not exist on our test system. In contrast, the original test case defaulted to a catch block. Our current engine does not encompass the handling of catch-blocks, leading to their absence in the Trace Gadgets generated. Upon manual inspection, it was confirmed that while the generated code was semantically accurate, it lacked the necessary catch-blocks. Another 17 test cases produced different outputs because of issues in fixing the random crypto seed. As a result, while they correctly encrypted a string, the final ciphertext was different. However, we confirmed that despite the different output, the generation process was the same. The remaining 19 test cases failed due to minor problems. Five were related to the loss of default initialization values in variables, resulting in incorrect output. Three cases involved threading, which our system currently fails to handle. The missing 11 are attributed to the inability to emulate Javas internal behavior as discussed in Section III-C. Given the rarity of this problem, we have decided not to modify our slicing engine at this time, although it is recognized as an area for future improvement.

B. LLM Prompt

The experiments in Section V-E involved the use of LLMs, specifically OpenAI’s GPT-4o, version gpt-4o-2024-05-13. For reproducibility purposes, the following prompt was used with a temperature of 0:

```
1 system:
2 You are an expert in sql injection detection

4 human:
5 As an expert in vulnerability detection, your task is to
  analyze
6 the provided code snippet for the presence of CWE-89
7 SQL Injection vulnerabilities.

9 Task:
10 - Focus exclusively on identifying CWE-89 SQL Injection
    issues.
11 - Do not address any other vulnerabilities or code issues
    .
12 - Provide a clear and concise assessment stating whether
    the
13 code contains CWE-89 SQL Injection vulnerabilities.
```

```
14 - If vulnerabilities are found, explain where they occur
    and why
15 they are vulnerabilities.

17 Code Snippet:
18 {code_snippet}
```

C. Hyper-parameter Grid and Search Protocol

To ensure a fair comparison, we applied an identical grid-search procedure to all models, i.e., UNIXCODER, TRACED, and CODET5+. The Cartesian product of the values below generates $3 \times 3 \times 3 \times 2 \times 2 = 108$ configurations per model. Each model was trained on *VulnDocker* and evaluated on the *Juliet* validation split. The setting with the highest F_1 score was finally deployed on the OWASP benchmark with a higher number of seeds.

Hyper-parameter	Values
Learning rate	$\{1 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}\}$
Drop-out rate	$\{0.1, 0.2, 0.3\}$
Warm-up fraction	$\{0, 0.1, 0.2\}$ (as proportion of total steps)
Frozen encoder layers [†]	$\{9, 6\}$
Random seed	$\{23, 42\}$

[†]Layer count starts at the embedding layer; a higher number freezes more lower layers during fine-tuning.