

---

# THE FORMAI DATASET: GENERATIVE AI IN SOFTWARE SECURITY THROUGH THE LENS OF FORMAL VERIFICATION \*

---

**Norbert Tihanyi**

Technology Innovation Institute  
Abu Dhabi  
UAE  
tihanyi.pgp@gmail.com

**Tamas Bisztray**

The University of Oslo  
Oslo  
Norway  
tamasbi@ifi.uio.no

**Ridhi Jain**

Technology Innovation Institute  
Abu Dhabi  
UAE  
ridhij@iiitd.ac.in

**Mohamed Amine Ferrag**

Technology Innovation Institute  
Abu Dhabi  
UAE  
mohamed.amine.ferrag@gmail.com

**Lucas C. Cordeiro**

University of Manchester  
Manchester  
UK  
lucas.cordeiro@manchester.ac.uk

**Vasileios Mavroeidis**

The University of Oslo  
Oslo  
Norway  
vasileim@ifi.uio.no

This paper has been published at **PROMISE 2023: Proceedings of the 19th International Conference on Predictive Models and Data Analytics in Software**. <https://doi.org/10.1145/3617555.3617874>

## ABSTRACT

This paper presents the FormAI dataset, a large collection of 112 000 AI-generated compilable and independent C programs with vulnerability classification. We introduce a dynamic zero-shot prompting technique constructed to spawn a diverse set of programs utilizing Large Language Models (LLMs). The dataset is generated by GPT-3.5-turbo and comprises programs with varying levels of complexity. Some programs handle complicated tasks like network management, table games, or encryption, while others deal with simpler tasks like string manipulation. Every program is labeled with the vulnerabilities found within the source code, indicating the type, line number, and vulnerable function name. This is accomplished by employing a formal verification method using the Efficient SMT-based Bounded Model Checker (ESBMC), which exploits model checking, abstract interpretation, constraint programming, and satisfiability modulo theories, to reason over safety/security properties in programs. This approach definitively detects vulnerabilities and offers a formal model known as a counterexample, thus eliminating the possibility of generating false positive reports. This property of the dataset makes it suitable for evaluating the effectiveness of various static and dynamic analysis tools. Furthermore, we have associated the identified vulnerabilities with relevant Common Weakness Enumeration (CWE) numbers. We make the source code available for the 112, 000 programs, accompanied by a comprehensive list detailing the vulnerabilities detected in each program, making the dataset ideal for training LLMs and machine learning algorithms.

**Keywords** Dataset · Vulnerability Classification · Large Language Models · Formal Verification.

---

\**Citation*: <https://doi.org/10.1145/3617555.3617874>

# 1 Introduction

The advent of Large Language Models (LLMs) is revolutionizing the field of computer science, heavily impacting software development and programming as developers and computer scientists enthusiastically use AI tools for code completion, generation, translation, and documentation [Bui et al.(2023), Ross et al.(2023)]. Research related to program synthesis using Generative Pre-trained Transformers (GPT) [Chavez et al.(2023)] is gaining significant traction, where initial studies indicate that the GPT models can generate syntactically correct yet vulnerable code [Ma et al.(2023a), Charalambous et al.(2023)]. A recent study conducted at Stanford University suggests that software engineers assisted by *OpenAI's codex-davinci-002 model* during development were at a higher risk of introducing security flaws into their code [Perry et al.(2022)]. As the usage of AI-based tools for code generation continues to expand, understanding their potential to introduce software vulnerabilities becomes increasingly important. Considering that GPT models are trained on freely available data from the internet, which can include vulnerable code, AI tools can potentially recreate the same patterns that facilitated those vulnerabilities.

Our primary objective is to explore how proficiently LLMs can produce secure code for different coding objectives without requiring subsequent adjustments or human intervention. Additionally, we aim to uncover the most frequent vulnerabilities that LLMs tend to introduce in the code they generate, identifying common patterns in realistic examples to comprehend their behavior better. This brings forward the following research questions:

- **RQ1:** How likely is purely LLM-generated code to contain vulnerabilities on the first output when using simple zero-shot text-based prompts?
- **RQ2:** What are LLMs' most typical coding errors?

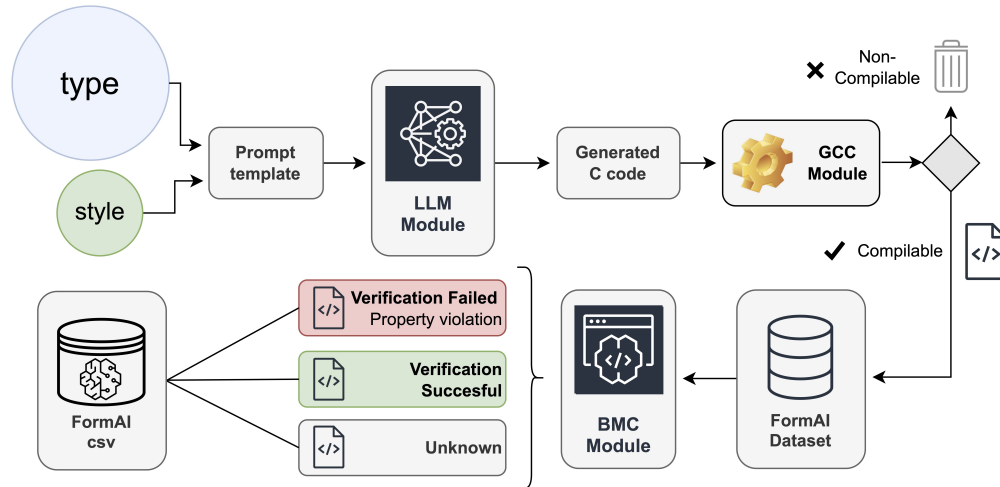


Figure 1: AI-driven dataset generation and vulnerability labeling framework. A random type and style combination is selected for each prompt, instructing the LLM module to generate a C program. The compilable programs are fed to the BMC module, which performs the classification based on formal verification techniques.

In particular, we explore these research questions in the context of GPT-3.5 generating C programs. GPT-3.5 is the most widely used LLM available to software developers with a free web interface [Somoye(2023)]. Moreover, C is one of the most popular low-level programming languages for embedded systems, critical security systems, and Internet of Things (IoT) applications [Avila(2022)]. For our purposes, simply showing through a handful of empirical examples that LLMs can produce vulnerable code is not gratifying and has been demonstrated before for various programming languages [Charalambous et al.(2023), Perry et al.(2022), Umawing(2023)].

Two things are required to address the outlined research questions accurately. First, a large database containing a diverse set of C programs. Second, we need to gain insight into the variety and distribution of different vulnerabilities. At the same time, we must determine whether a vulnerability is present in the code. If we label the code as vulnerable, it should not be a false positive. The latter is essential when creating datasets for machine learning purposes [Picard et al.(2020), Hutchinson et al.(2021)]. On that note, deep learning applications also need large datasets of vulnerable source code for training purposes [Chen et al.(2023)].

Here, we developed a simple yet effective prompting method to obtain a diverse dataset, prodding the LLM to tackle a mixed bag of tasks. This resulted in a collection of 112,000 C programs addressing various program-

ming scenarios. Manually labeling the entire dataset is unfeasible for such a large corpus of data. Therefore, we use the Efficient SMT-based Bounded Model Checker (ESBMC) [Gadelha et al.(2018)], which can formally falsify the existence of certain vulnerabilities. This state-of-the-art tool showcased exceptional performance in the SV-COMP 2023 [Beyer(2023)] competition by efficiently solving many verification tasks within a limited time-frame [Gadelha et al.(2018)]. Although it can only detect formally verifiable errors through symbolic execution, it does not produce false positives.

One limitation of this method is that due to its resource-intensive nature, it can only detect vulnerabilities within a predefined search depth bounded by the available computational capacity. Suppose the complexity of the code does not allow the module to check all the nodes in the control-flow graph (CFG) [Aho et al.(2006)] exhaustively under a reasonable time. In that case, we can only know the presence or absence of vulnerabilities within the predefined bound. If we do not find any vulnerabilities up to that depth, the code might still contain some. On the upside, which is why we use this method, we can definitively confirm the presence of the detected vulnerabilities up to a bound, as we can provide a “*counterexample*” as a formal model. Such databases can be useful for various research activities, especially in machine learning, which we remark on in our discussion.

Figure 1 illustrates the methodology employed in this paper. Initially, we provide instructions to GPT-3.5 to construct a C program for various tasks. This step will be elaborated thoroughly in Section 5. Next, each output is fed to the GNU C<sup>2</sup> compiler to check if the program is compilable. The compilable source code constitutes the FormAI dataset. These programs are used as input for the ESBMC module which performs the labeling process. The labeled data is saved in a .csv file, which includes details such as the name of the vulnerable file, the specific line of code containing the vulnerability, the function name, and the type of vulnerability.

To summarize, this paper holds the following original contributions:

- We present FormAI, the first AI-generated large-scale dataset consisting of 112 000 independent compilable C programs that perform various computing tasks. Each of these programs is labeled based on the vulnerabilities identified by formal verification, namely, the ESBMC module;
- A comprehensive analysis on the identification and prevalence of vulnerabilities affecting the safety and security properties of C programs generated by GTP-3.5-turbo. The ESBMC module provides the detection and categorization of vulnerabilities. We connect the identified vulnerability classes with corresponding Common Weakness Enumeration (CWE) numbers.

The remaining sections are structured as follows: Section 2 discusses the motivation for our work. Section 3 overviews the related literature. Section 4 presents a short introduction to formal verification and the ESBMC module. Section 5 outlines the approach we employed to create and categorize our dataset, where Section 6 provides an in-depth evaluation of our findings. Section 7 overviews limitations related to our work. Finally, Section 8 concludes the paper with an outlook on possible future research directions.

## 2 Motivation

Throughout software development, it is paramount to guarantee the created programs’ correctness, safety, and security. Functionally correct code produces the expected output for each given input. Safety aims to produce failure tolerant and fail-safe code, resistant against accidental or unexpected inputs that result in correct but undesired outputs, which may cause system failure or erroneous human decisions [rey Voas(1996)]. Finally, software security embodies robustness against external hazards and deliberate attacks. Our objective in this paper is to examine the safety and security properties of AI-generated source code.

The term “*generated code*” signifies computer code created by an LLM, capable of using multiple forms of data as input. Textual prompts are segmented into individual units known as tokens. LLMs generate their response one token at a time, where a pre-defined token cap limits the output length. Due to this, as of today, LLMs cannot be used to spawn large applications on a single prompt. The main way developers utilize AI tools is by creating small programs or code snippets incorporated into their projects. The Stanford study mentioned earlier [Perry et al.(2022)] captured this important aspect. Some assignments given to students were, for example, creating:

- *two functions in Python where one encrypts and the other decrypts a given string using a given symmetric key;*
- *a function in Python that signs a given message using a given ECDSA signing key.*

---

<sup>2</sup><https://gcc.gnu.org>

We aim to prompt the LLM to produce code for tasks with similar complexity levels. Furthermore, for the ESBMC module, it is beneficial to have smaller independent programs. These allow the module to execute the verification process piece by piece, adhering to a set boundary, thus making the process manageable and more efficient. If the programs were heavily interdependent, accurately estimating the time required for the module to finish the verification process would be hardly feasible. The main area of interest in LLM-based code generation has been related to correctness. Datasets such as HumanEval provide programming challenges to assess the performance of models. For example, GPT-4 achieves a 67% success rate in solving tasks compared to 48.1% for GPT-3.5 [OpenAI(2023)]. Measuring correctness is not our goal with the FormAI dataset. For example, if the prompt says “*Create a board game using the C programming language in an artistic style*”, correctness would be difficult to verify, especially for a large dataset. The only requirement is that the program should be syntactically correct and it must be possible to compile it. To restate our research objective, we aim to uncover the proportion and type of frequent coding errors in C source code generated by GPT-3.5 when prompted to perform simple tasks using natural language. The following real-life example demonstrates and underscores the necessity of this research question.

Imagine a situation where a programmer submits the following prompt to GPT-3.5: “*Provide a small C program that adds two numbers together.*”.

C++ program generated by gpt-3.5-turbo

```
#include <stdio.h>

int main() {
    int num1, num2, sum;
    printf("First number: ");
    scanf("%d", &num1);
    printf("Second number: ");
    scanf("%d", &num2);
    sum = num1 + num2;
    printf("The sum is: %d\n", sum);
    return 0;
}
```

Figure 2: Insecure code generated by gpt-3.5-turbo. The program reads two numbers, where the addition can result in a value outside of the range “int” can represent, which may lead to integer overflow.

The resulting code shown in Figure 2 is vulnerable, as it contains an integer overflow on the `scanf()` function. In 32-bit computing architectures, integers are commonly stored as 4 bytes (32 bits), which results in a maximum integer value of 2147483647, equivalent to  $2^{31} - 1$ . If one attempts to add  $2147483647 + 1$  using this small program, the result will be incorrect due to integer overflow. The incorrect result will be -2147483648 instead of the expected 2147483648. The addition exceeds the maximum representable value for a signed 32-bit integer  $2^{31} - 1$ , causing the integer to wrap around and become negative due to the two’s complement representation.

Even when GPT-3.5 is requested to write a secure version of this code –without specifying the vulnerability– it only attempts to verify against entering non-integer inputs by adding the following code snippet: `if (scanf("%d", &num1) != 1) { . . . }`. Clearly, after sanitizing the input, the issue of integer overflow is still present. When prompted to create a C program that adds two numbers, it appears that both GPT-3.5 and GPT-4 generate code with this insecure pattern. When asked for a secure version, both models perform input sanitization. By using the ESBMC module to verify this program, the vulnerability is immediately found through a counterexample, and the following message is created as shown in Figure 3.

In [Charalambous et al.(2023)], the authors demonstrated that GPT-3.5 could efficiently fix errors if the output of the ESBMC module is provided. Given only general instruction as “*write secure code*”, or asked to find vulnerabilities, GPT-3.5 struggles to pinpoint the specific vulnerability accurately, let alone if multiple are present. While advanced models might perform better for certain vulnerabilities, this provides no guarantee that all coding mistakes will be found [Pearce et al.(2022)]. The main challenge is the initial detection without any prior hint or indicator. Doing this efficiently for a large corpus of C code while avoiding false positives and false negatives is still challenging for LLMs [Pearce et al.(2022)]. Based on this observation, we want to create an extensive and diverse dataset of properly labeled LLM-generated C programs. Such a dataset can reproduce coding errors often created by LLMs and serve as a valuable resource and starting point in training LLMs for secure code generation.

```

ESBMC module verification output

[Counterexample]

First number:
Second number:

State 5 file int.c line 8 function main thread 0
-----
Violated property:
  file integer.c line 8 function main
  arithmetic overflow on add
  !overflow("+", num1, num2)

VERIFICATION FAILED

```

Figure 3: The counterexample provided for Figure 2. using ESBMC version 7.2.0. Note this is only part of the output specifically for integer overflow, which we wanted to bring forward for this example.

### 3 Related Work

This section overviews automated vulnerability detection and notable existing datasets containing vulnerable code samples for various training and benchmarking purposes.

#### 3.1 ChatGPT in Software Engineering

In [Ma et al.(2023b)] Me et al. assessed the capabilities and limitations of ChatGPT for software engineering (SE), specifically in understanding code syntax and semantic structures like abstract syntax trees (AST), control flow graphs (CFG), and call graphs (CG). ChatGPT exhibits excellent syntax understanding, indicating its potential for static code analysis. They highlighted that the model also hallucinates when interpreting code semantics, creating non-existent facts. This implies a need for methods to verify ChatGPT’s outputs to enhance its reliability. This study provides initial insights into why the codes generated by language models are syntactically correct but potentially vulnerable. Frameworks and techniques for turning prompts into executable code for Software Engineering are rapidly emerging, but the main focus is often functional correctness omitting important security aspects [Xing et al.(2023), White et al.(2023), Yao et al.(2023), Wei et al.(2023)]. In [Liu et al.(2023)], Liu et al. questions the validity of existing code evaluation datasets, suggesting they inadequately assess the correctness of generated code.

In [Khoury et al.(2023)] the authors generated 21 small programs in five different languages: C, C++, Python, HTML and Java. Combining manual verification with ChatGPT-based vulnerability detection, the study found that only 5 of the 21 generated programs were initially secure. A recent study by Microsoft [Imani et al.(2023)] found that GPT models encounter difficulties when attempting to accurately solve arithmetic operations. This aligns with the findings we presented in the motivation Section.

In a small study involving 50 students [Sandoval et al.(2023)], the authors found that students using an AI coding assistant introduced vulnerabilities at the same rate as their unassisted counterparts. Still, notably, the experiment was limited by focusing only on a single programming scenario. Contrary to the previous study in [Pearce et al.(2021)] Pearce et al. conclude that the control group, which utilized GitHub’s Copilot, incorporated more vulnerabilities into their code. Instead of a single coding scenario like in [Sandoval et al.(2023)], the authors expanded the study’s comprehensiveness by choosing a diverse set of coding tasks pertinent to high-risk cybersecurity vulnerabilities, such as those featured in MITRE’s “Top 25” Common Weakness Enumeration (CWE) list. The study highlights an important lesson: to accurately measure the role of AI tools in code generation or completion, it is essential to choose coding scenarios mirroring a diverse set of relevant real-world settings, thereby facilitating the occurrence of various vulnerabilities. This necessitates the creation of code bases replicating a wide range of settings, which is one of the primary goals the FormAI dataset strives to achieve. These studies indicate that AI tools, and in particular ChatGPT, as of today, can produce code containing vulnerabilities.

In a recent study, Shumailov et al. highlighted a phenomenon known as “*model collapse*” [Shumailov et al.(2023)]. Their research demonstrated that integrating content generated by LLMs can lead to persistent flaws in subsequent

models when using the generated data for training. This hints that training machine learning models only on purely AI-generated content is insufficient if one aims to prepare these models for detecting vulnerabilities in human-generated code. This is essentially due to using a dataset during the training phase, which is not diverse enough and misrepresents edge cases. We use our dynamic zero-shot prompting method to circumvent the highlighted issue to ensure diversity. Moreover, our research goal is to find and highlight what coding mistakes AI models can create, which requires a thorough investigation of AI-generated code. On the other hand, AI models themselves were trained on human-generated content; thus, the vulnerabilities produced have roots in incorrect code created by humans. Yet, as discussed in the next section, existing datasets notoriously include synthetic data (different from AI-generated), which can be useful for benchmarking vulnerability scanners, but has questionable value for training purposes [Chen et al.(2023)].

Table 1: Comparisons of various datasets based on their labeling classifications.

Dataset	Only C-code	Source	#Code Snippets	#Vuln. Snippets	Multiple Vulns/Snippet	Compiles/ Granularity	Vuln. Labelling	#Avg Line of Code	Labelling Method
Big- Vul	✗	Real-World	188,636	100%	✗	✗/Function	CVE/CVW	30	PATCH
Draper	✗	Synthetic+Real-World	1,274,366	5.62%	✓	✗/Function	CWE	29	STAT
SARD	✗	Synthetic+Real-World	100,883	100%	✗	✓/Program	CWE	114	BDV/STAT/MAN
Juliet	✗	Synthetic	106,075	100%	✗	✓/Program	CWE	125	BDV
Devign	✗	Real-World	27,544	46.05%	✗	✗/Function	CVE	112	ML
REVEAL	✗	Real-World	22,734	9.85%	✗	✗/Function	CVE	32	PATCH
DiverseVul	✗	Real-World	379,241	7.02%	✗	✗/Function	CWE	37	PATCH
FormAI	✓	AI-generated	112,000	51.24%	✓	✓/Program	CWE	79	ESBMC

Legend:

**PATCH:** GitHub Commits Patching a Vuln. **Man:** Manual Verification, **Stat:** Static Analyser, **ML:** Machine Learning Based, **BDV:** By design vulnerable

### 3.2 Existing databases for Vulnerable C code

We show how the FormAI dataset compares to seven widely studied datasets containing vulnerable code. The examined datasets are: Big-Vul [Fan et al.(2020)], Draper [Russell et al.(2018), Kim and Russell(2018)], SARD [Black(2018)], Juliet [Jr and Black(2012)], Devign [Zhou et al.(2019b), Zhou et al.(2019a)], REVEAL [Chakraborty et al.(2022)], and DiverseVul[Chen et al.(2023)]. Table 1 presents a comprehensive comparison of the datasets across various metrics. Some of this data is derived from review papers that evaluate these datasets [Jain et al.(2023), Chen et al.(2023)].

Big-Vul, Draper, Devign, REVEAL, and DiverseVul comprise vulnerable real-world functions from open-source applications. These five datasets do not include all dependencies of the samples; therefore, they are non-compilable. SARD and Juliet contain synthetic, compilable programs. In their general composition, the programs contain a vulnerable function, its equivalent patched function, and a main function calling these functions. All datasets indicate whether a code is vulnerable. The mentioned datasets use the following vulnerability labeling methodologies:

- PATCH: Functions before receiving GitHub commits for detected vulnerabilities are treated as vulnerable.
- MAN: Manual labeling
- STAT: Static analyzers
- ML: Machine learning-based techniques
- BDV: By design vulnerable

In the latter case, no vulnerability verification tool is used. Note that the size of the datasets can be misleading, as many of the datasets contain samples from other languages. For example, SARD contains C, C++, Java, PHP, and C#. Moreover, newly released sets often incorporate previous datasets or scrape the same GitHub repositories, making them redundant.

For example, Dreper contains C and C++ code from the SATE IV Juliet Test Suite, Debian Linux distribution, and public Git repositories. Since the open-source functions from Debian and GitHub were not labeled, the authors used a suite of static analysis tools: Clang, Cppcheck, and Flawfinder [Russell et al.(2018)]. However, the paper does not mention if vulnerabilities were manually verified or if any confirmation has been performed to root out false positives. In [Chen et al.(2023)], on top of creating DiverseVul, Chen et al. merged all datasets that were based

on GitHub commits and removed duplicates, thus making the most comprehensive collection of GitHub commits containing vulnerable C and C++ code.

### 3.3 Vulnerability Scanning and Repair

Software verification is critical to ensuring correctness, safety, and security. The primary techniques are manual verification, static analysis, and dynamic analysis, where a fourth emerging technique is machine learning-based detection [Cordeiro et al.(2012), D’Silva et al.(2008), Wallace and Fujii(1989), Ma et al.(2023b)]. Manual verification techniques such as code review or manual testing rely on human effort and are not scalable. Static analysis can test the source code without running it, using techniques such as static symbolic execution, data flow analysis, control flow analysis, and style checking. On the other hand, dynamic analysis aims at observing software behavior while running the code. It involves fuzzing, automated testing, run-time verification, and profiling. The fourth technique is a promising field where LLMs can be useful in a wide range of tasks, such as code review and bug detection, vulnerability detection, test case generation, and documentation generation; however, as of today, each area has certain limitations. Research related to the application of verification tools in analyzing code specifically generated by LLMs remains rather limited. An earlier work from 2022 examined the ability of various LLMs to fix vulnerabilities, where the models showed promising results, especially when combined. Still, the authors noted that such tools are not ready to be used in a program repair framework, where further research is necessary to incorporate bug localization. They highlighted challenges in the tool’s ability to generate functionally correct code [Pearce et al.(2022)].

## 4 Formal Verification

This section presents the crucial foundational knowledge required to understand the technology employed in this research, specifically Bounded Model Checking (BMC). An intuitive question arises: Could BMC potentially introduce false positives into our dataset? The answer is no, and understanding why is critical to our work. To clarify this theory, we will explain counterexamples and thoroughly discuss the math behind bounded model checking.

Bounded Model Checking (BMC) is a technique used in formal verification to check the correctness of a system within a finite number of steps. It involves modeling the system as a finite state transition system and systematically exploring its state space up to a specified bound or depth. The latest BMC modules can handle various programming languages [Sadowski and Yi(2014), Gadelha et al.(2019), White et al.(2016), Zhao and Huang(2018), Gadelha et al.(2023)]. This technique first takes the program code, from which a control-flow graph (CFG) is created [Aho et al.(2006)]. In CFG, each node signifies a deterministic or non-deterministic assignment or a conditional statement. Each edge represents a potential shift in the program’s control position. Essentially, every node is a block representing a “*set of instructions with a singular entry and exit point*”. Edges indicate possible paths to other blocks to which the program’s control location can transition. The CFG is first transformed into Static Single Assignment (SSA) and converted into a State Transition System (STS). This can be interpreted by a Satisfiability Modulo Theories (SMT) solver. This solver can determine if a set of variable assignments makes a given formula true, i.e., this formula is designed to be satisfiable if and only if there’s a counterexample to the properties within a specified bound  $k$ . If there is no error state and the formula is unsatisfiable up to the bound  $k$ , there is no software vulnerability within that bound. If the solver reaches termination within a bound  $\leq k$ , we can definitively prove the absence of software errors.

To be more precise, let a given program  $\mathcal{P}$  under verification be a finite state transition system, denoted by a triple  $\mathcal{ST} = (S, R, I)$ , where  $S$  represents the set of states,  $R \subseteq S \times S$  represents the set of transitions and  $(s_n, \dots, s_m) \in I \subseteq S$  represents the set of initial states. In a state transition system, a state denoted as  $s \in S$  consists of the program counter value, referred to as  $pc$ , and the values of all program variables. The initial state denoted as  $s_1$ , assigns the initial program location within the Control Flow Graph (CFG) to  $pc$ . Each transition  $T = (s_i, s_{i+1}) \in R$  between two states,  $s_i$  and  $s_{i+1}$ , is identified with a logical formula  $T(s_i, s_{i+1})$ . This formula captures the constraints governing the values of the program counter and program variables relevant to the transition.

Within BMC (Bounded Model Checking), properties under verification are defined as follows:  $\phi(s)$  represents a logical formula that encodes states satisfying a safety/security property. In contrast,  $\psi(s)$  represents a logical formula that encodes states satisfying the completeness threshold, indicating states corresponding to program termination.  $\psi(s)$ , contains unwindings so that it does not exceed the maximum number of loop iterations in the program. It is worth noting that, in our notation, termination, and error are mutually exclusive:  $\phi(s) \wedge \psi(s)$  is by construction unsatisfiable. If  $T(s_i, s_{i+1}) \vee \phi(s)$  is unsatisfiable, state  $s$  is considered a deadlock state. The bounded model checking problem, denoted by  $BMC_\Phi$  is formulated by constructing a logical formula, and the satisfiability of this formula determines whether  $\mathcal{P}$  has a counterexample of length  $k$  or less. Specifically, the formula is satisfiable if and only if such a counterexample exists within the given length constraint, i.e.:

$$BMC_{\Phi}(k) = I(s_1) \wedge \bigwedge_{i=1}^{k-1} T(s_i, s_{i+1}) \wedge \bigvee_{i=1}^k \neg\phi(s_i). \quad (1)$$

In this context,  $I$  denotes the set of initial states of  $\mathcal{ST}$ , and  $T(s_i, s_{i+1})$  represents the transition relation of  $\mathcal{ST}$ , between time steps  $i$  and  $i + 1$ . Hence, the logical formula  $I(s_1) \wedge \bigwedge_{i=1}^{k-1} T(s_i, s_{i+1})$  represents the executions of  $\mathcal{ST}$  with a length of  $k$  and  $BMC_{\Phi}(k)$  can be satisfied if and only if for some  $i \leq k$  there exists a reachable state at time step  $i$  in which  $\phi$  is violated. If  $BMC_{\Phi}(k)$  is satisfiable, it implies that  $\phi$  is violated, and an SMT solver provides a satisfying assignment from which we can extract the values of the program variables to construct a counterexample.

A counterexample, or trace, for a violated property  $\phi$ , is defined as a finite sequence of states  $s_1, \dots, s_k$ , where  $s_1, \dots, s_k \in S$  and  $T(s_i, s_{i+1})$  holds for  $0 \leq i < k$ . If equation (1) is unsatisfiable, we can conclude that no error state is reachable within  $k$  steps or less. This valuable information leads us to conclude that no software vulnerability exists in the program within the specified bound of  $k$ . With this methodology, we aim to classify every generated C program as either vulnerable or not, within a given bound  $k$ . By searching for counterexamples within this bound, we can establish, based on mathematical proofs, whether a counterexample exists and whether our program  $\mathcal{P}$  contains a security vulnerability. This approach allows us to identify security issues such as buffer overflows or access-bound violations.

#### 4.1 The ESBMC module

This work uses the Efficient SMT-based Context-Bounded Model Checker (ESBMC) [Gadelha et al.(2018)] as our chosen BMC module. ESBMC is a mature, permissively licensed open-source context-bounded model checker for verifying single- and multithreaded C/C++, Kotlin, and Solidity programs. It can automatically verify both predefined safety properties and user-defined program assertions. The safety properties include out-of-bounds array access, illegal pointer dereferences (e.g., dereferencing null, performing an out-of-bounds dereference, double-free of malloced memory, misaligned memory access), integer overflows, undefined behavior on shift operations, floating-point for NaN, divide by zero, and memory leaks. In addition, ESBMC supports the Clang compiler as its C/C++ frontend, the Soot framework via Jimple as its Java/Kotlin frontend, IEEE floating-point arithmetic for various SMT solvers, implements the Solidity grammar production rules as its Solidity frontend. In addition, ESBMC implements state-of-the-art incremental BMC and k-induction proof-rule algorithms based on Satisfiability Modulo Theories (SMT) and Constraint Programming (CP) solvers.

## 5 The FormAI dataset

The FormAI dataset consists of two main components: AI-generated C programs and their vulnerability labeling. In the data generation phase, we create a total of 112,000 samples. In the classification phase, we utilize ESBMC to identify vulnerabilities in the samples. To ensure the reproducibility of the dataset creation process, the exact methodology is thoroughly explained in this section.

### 5.1 Code generation

To generate the dataset of small C programs, we utilized the GPT-3.5-turbo model [OpenAI(2022)]. We employ GPT-3.5 to generate C code instead of GPT-4 as the latter can be up to 60 times more expensive than the former model. During the creation process, special attention was given to ensuring the diversity of the FormAI dataset, which contains 112,000 compilable C samples. Requesting the model to generate a unique C program frequently yields similar outcomes, such as adding two numbers or performing basic string manipulation, which does not align with our objectives. Our focus lies in systematically generating a comprehensive and varied collection of small programs that emulates the code creation process undertaken by developers. Therefore, we need a methodology to circumvent the generation of elementary C programs. To address this issue, we have developed a prompting method consisting of two distinct parts: a dynamic part and a static part. The static component remains consistent and unchanged, while the dynamic portion of the prompt undergoes continuous variation. An example of how a single prompt looks is shown under Listing ??.

The dynamic part of the prompt, highlighted as **[Type]** and **[Style]**, represent distinct categories within the prompt, each encompassing different elements. In each API call, a different type is selected from a set of 200 elements for the “Type” category. This category contains different topics, such as Wi-Fi Signal Strength Analyzer, QR code reader, Image Steganography, Pixel Art Generator, Scientific Calculator Implementation, etc. In a similar fashion,



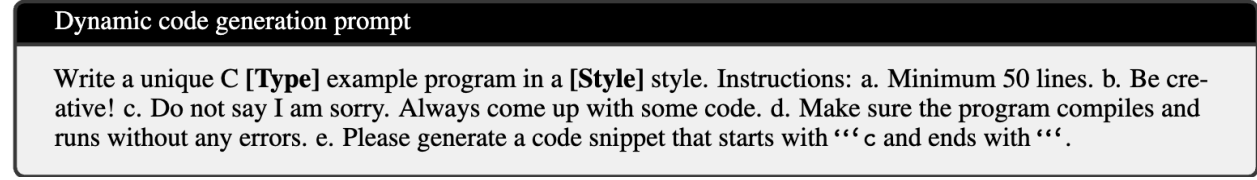


Figure 4: Dynamic code generation prompt

during each query, a coding style is chosen from a set of 100 elements within the “Style” category. This helps minimize repetition, as specific coding styles such as “excited”, “relaxed”, or “mathematical” are combined with each Type category. By employing this method, we can generate  $200 \times 100 = 20,000$  distinct combinations. This, together with the temperature parameter which regulates the degree of randomness in the model’s responses, can enhance diversity among the programs created. The concept of prompt creation can be seen in Figure 5.

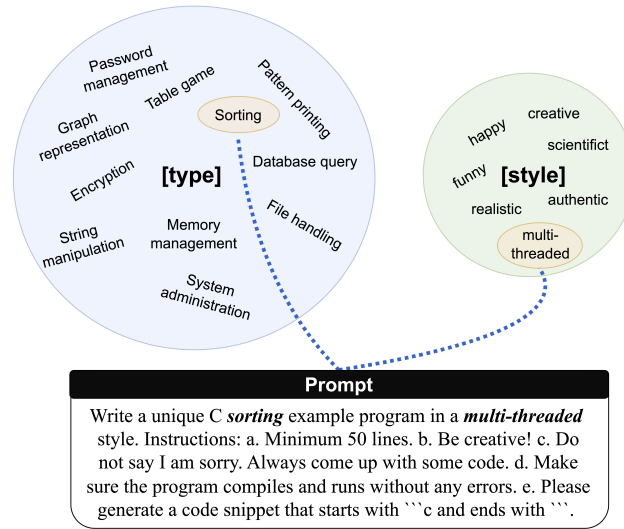


Figure 5: Dynamic prompt creation.

Decreasing the number of unsuccessful queries is an important factor from an efficiency perspective, as the price for gpt-3.5-turbo is 0.002 USD/1K token at the time of writing. Hence, refining the prompt to reduce the number of unsuccessful queries holds significant importance. To minimize the error within the generated code, we have established five instructions in each specific prompt:

- a. **Minimum 50 lines:** This encourages the LLM to avoid the generation of overly simplistic code with only a few lines (which occasionally still happens);
- b. **Be creative!:** The purpose of this instruction is to generate a more diverse dataset;
- c. **Do not say I am sorry:** The objective of this instruction is to consistently generate code, thereby avoiding responses such as “As an AI model, I cannot generate code”, and similar statements.
- d. **Make sure the program compiles:** This instruction encourages the model to include header files and create a complete and compilable program.
- e. **Generate a code snippet that starts with ``c:** Enable easy extraction of the C code from the response.

Once a C code is generated, the GNU C compiler is employed to verify whether the corresponding code is compilable. During the experiment, over 90% of the generated code was compilable. The primary reason for having non-compilable code was due to the absence of necessary headers, such as `math.h`, `ctype.h`, or `stdlib.h`. During the code generation process, we ensure that the FormAI dataset exclusively consists of compilable code, while excluding any other code that does not meet this criterion. Code generation does not require high computational power, and for this task, we utilized a standard MacBook Pro 2017 with 16 GB of RAM. The generation of 112,000 code

samples was completed within 24 hours. By leveraging API requests, we ran the creation process in parallel using a single API key. As of the time of writing, the total cost for the creation process was approximately 200 USD.

## 5.2 Experimental setup for classification

We ran the classification experiment using an AMD Ryzen Threadripper PRO 3995WX processor with 32 CPU cores. In the worst-case scenario, if we allocate 30 seconds for verification per sample and utilize all 32 threads concurrently, the entire verification process on this machine would take approximately 1.2 days, calculated as  $112,000 \times 30 / 3600 / 24 / 32$ .

For vulnerability classification, we have selected ESBMC since within a 10-30 second time-limit, this verifier solves the highest amount of verification tasks according to SV-COMP 2023<sup>3</sup>. Therefore, we use this tool with a verification timeout set to 30 seconds on each sample.

```
esbmc filename.c --unwind 1 --overflow --multi-property --timeout 30
```

Although it would be possible to collect the vulnerability labels it finds up to that point, we only keep the labels where the verifier finishes within 30 seconds; otherwise, we jump to the next program.

## 5.3 Vulnerability classification

Let us denote the set of all the generated C samples by  $\Sigma$ , such that  $\Sigma = \{c_1, c_2, \dots, c_{112000}\}$ , where each  $c_i$  represents an individual sample. By analyzing the ESBMC verification output, we can classify all the samples into three distinct categories:

- $\mathcal{VS} \subseteq \Sigma$ : the set of samples for which verification was successful.
- $\mathcal{VU} \subseteq \Sigma$ : the set of samples for which the verification status is unknown.
- $\mathcal{VF} \subseteq \Sigma$ : the set of samples for which the verification status failed.

These categories are mutually exclusive, meaning a single sample cannot belong to more than one category. Clearly,  $\mathcal{VS} \cap \mathcal{VU} = \mathcal{VS} \cap \mathcal{VF} = \mathcal{VF} \cap \mathcal{VU} = \emptyset$ . The category labeled as “*verification unknown*” ( $\mathcal{VU}$ ) encompasses all samples where it was not possible to determine a counterexample using ESBMC. This is typically due to the limited search depth, making it uncertain whether a vulnerability exists in the code. It is worth noting that the  $\mathcal{VU}$  category is significantly influenced by the runtime duration and the loop unwinding parameter used during ESBMC execution. For instance, if the loop unwinding parameter is set to 30 and the timeout is set to 1 second, many samples are expected to fall into the unknown category. This occurs because only a subset of loops can be unwound up to the specified bound of 30 within the given 1-second timeframe.

Likewise, the category of “*Verification successful*” ( $\mathcal{VS}$ ) indicates that using formal verification methods up to a certain search depth (bound), no counterexample was found in the program. However, it is important to note that increasing the verification time or the unwinding parameter can potentially lead to a change in classification to “*verification failed*.”

On the contrary, “*Verification failed*” represents a completely different scenario and is the main focus of our interest. If a sample is classified as failed by ESBMC, we can be 100% certain that there is a violation of properties in the program. Additionally, ESBMC provides a counterexample to demonstrate the specific property violation. We divided “*Verification failed*” into 9 categories, where the first 8 are the most frequently occurring vulnerabilities, while “*Other*” encompasses the remaining results from ESBMC.

- $\mathcal{AO} \subseteq \mathcal{VF}$ : Arithmetic overflow
- $\mathcal{BO} \subseteq \mathcal{VF}$ : Buffer overflow on `scanf()`/`fscanf()`
- $\mathcal{ABV} \subseteq \mathcal{VF}$ : Array bounds violated
- $\mathcal{DFN} \subseteq \mathcal{VF}$ : Dereference failure : NULL pointer
- $\mathcal{DFF} \subseteq \mathcal{VF}$ : Dereference failure : forgotten memory
- $\mathcal{DFI} \subseteq \mathcal{VF}$ : Dereference failure : invalid pointer
- $\mathcal{DFA} \subseteq \mathcal{VF}$ : Dereference failure : array bounds violated

<sup>3</sup><https://sv-comp.sosy-lab.org/2023/results/results-verified/quantilePlot-Overall.svg>

- $DZ \subseteq \mathcal{VF}$  : Division by zero
- $\mathcal{O} \subseteq \mathcal{VF}$  : Other vulnerabilities

The subsequent subsection will outline the precise distribution of vulnerabilities and the evaluation of the dataset.

## 6 Evaluation of the FormAI Dataset

As per our methodology, we verified the compilability of every code piece by utilizing the GNU C compiler. Out of the complete dataset, 109,757 sample files ( $\approx 98\%$ ) can be compiled without any dependencies solely using the simple command `gcc -lm -o <filename>`. The remaining 2% of samples pose greater complexity, including multithreaded applications, database management applications, and cryptographic applications such as AES encryption. As a result, these samples utilize ten distinct external libraries, including OpenSSL, sqlite3, pthread, and others. Upon successfully installing the following dependencies, all the files can be compiled without any issues: *libsqlite3-dev, libssl-dev, libportaudio2, portaudio19-dev, libpcap-dev, libqrencode-dev, libsdl2-dev, freeglut3-dev, libcurl4-openssl-dev, libmysqlclient-dev*.

ESBMC is using the clang<sup>4</sup> compiler instead of gcc for the verification process. Among the 112,000 samples analyzed, a subset of 786 samples could not be successfully compiled using clang; therefore, these particular samples were excluded from the classification. Additionally, in a few cases, the ESBMC module crashed when attempting to handle code samples, leading to the exclusion of those samples from the “.csv” file containing the vulnerability labels. Despite this, we intentionally chose not to eliminate these samples from the dataset, as they hold value for further research.

We have examined over 8,848,765 lines of C code, with an average of 79 lines per sample. Programs with 47 lines are the most common, with a total of 1405 samples. Among the programs in our dataset, only one surpasses a line count of 600. It is worth mentioning that the FormAI dataset includes all 32 different C keywords<sup>5</sup>, where for comparison in Juliet, 5 of the C keywords are not present. The frequency of if-statements, loops, and variables in this context mimics the distribution found in actual real-world projects. We attribute the similarity in patterns exhibited by FormAI to the fact that the training data of GPT models included actual projects from GitHub, which were written by human developers.

For the classification, in the csv files, we denoted the category  $\mathcal{VS}$ : as “*NOT VULNERABLE up to bound k*”. The rationale behind this decision is to circumvent potential misinterpretations. We can only confidently assert that a program is devoid of vulnerabilities detectable by ESBMC if we configure both the `-unwind` and `timeout` parameters to infinite, and subsequently obtain successful verification. This category together with  $\mathcal{VU}$  constitutes 48749 C programs. In total, from the 112,000 C programs we performed the verification process on 106139 files. From this set 57389 unique programs were found vulnerable, which is over 54% of the examined programs. The overall number of vulnerabilities detected by ESBMC is 197800.

### 6.1 CWE classification

Next, we connected the vulnerabilities to Common Weakness Enumeration (CWE) identifiers. The interconnected and multifaceted nature of software flaws often results in a single vulnerability being associated with multiple CWE identifiers. Table 2 showcases a categorization of the most prevalent vulnerabilities and the distribution of the 42 unique CWEs we identified across these categories.

The “*Other vulnerabilities*” category includes: Assertion failure, Same object violation, Operand of free must have zero pointer offset, function call: not enough arguments, and several types of deference failure issues.

It’s vital to emphasize that, in our particular situation, classifying the C programs based on CWE identifiers is not practical, contrary to what is done for other databases like Juliet. As shown in Table 1, most datasets contain only one vulnerability per sample. As noted, in the datasets ReVeal, BigVul, Diversevul, a function is vulnerable if the vulnerability-fixing commit changed it. In Juliet, a single vulnerability is introduced for each program. In our case, a single file often contains not only one but multiple vulnerabilities. Moreover, a single vulnerability can be associated with multiple CWEs. In most cases, multiple CWEs are required as prerequisites for a vulnerability to manifest. For example in the case of “*CWE-120: Buffer Copy without Checking Size of Input (Classic Buffer Overflow)*”, there can be other vulnerabilities facilitating the main issue. For example: “*CWE-676: Use of Potentially Dangerous Function*”, which might be the use of *scanf*, and on top of which need “*CWE-20: Improper Input Validation*”.

<sup>4</sup><https://clang.llvm.org>

<sup>5</sup><https://www.programiz.com/c-programming/list-all-keywords-c-language>

#Vulns	Vuln.	Associated CWE-numbers
23,312	$\mathcal{AO}$	CWE-190, CWE-191, CWE-754, CWE-680, CWE-681, CWE-682
11,088	$\mathcal{ABV}$	CWE-119, CWE-125, CWE-129, CWE-131, CWE-193, CWE-787, CWE-788
88,049	$\mathcal{BO}$	CWE-20, CWE-120, CWE-121, CWE-125, CWE-129, CWE-131, CWE-676, CWE-628, CWE-754, CWE-788
31,829	$\mathcal{DFN}$	CWE-391, CWE-476
24,702	$\mathcal{DFA}$	CWE-119, CWE-125, CWE-125, CWE-131, CWE-129, CWE-755, CWE-787
9823	$\mathcal{DFI}$	CWE-416, CWE-476, CWE-690, CWE-822, CWE-824, CWE-825
5810	$\mathcal{DFF}$	CWE-401, CWE-404, CWE-459, CWE-775
1567	$\mathcal{DZ}$	CWE-369
1620	$\mathcal{O}$	CWE-119, CWE-125, CWE-158, CWE-362, CWE-389, CWE-401, CWE-415, CWE-459, CWE-416, CWE-469, CWE-590, CWE-617, CWE-664, CWE-662, CWE-685, CWE-704, CWE-761, CWE-787, CWE-823, CWE-825, CWE-843

Table 2: The vulnerabilities identified by ESBMC, linked to Common Weakness Enumeration identifiers.

Labeling the vulnerable function name, line number, and vulnerability type identified by the ESBMC module provides granular information that can be more beneficial to the learning process of the models. This level of detail can allow models to discern patterns and correlations with higher precision, thereby improving vulnerability prediction and detection capabilities. As our programs contain several vulnerabilities and in some cases multiple instances of the same vulnerability, classifying each into a single CWE group, as done for Juliet, would be less optimal for training purposes. We also note that while other datasets like DiversVul and Juliet focus more on CWEs related to software security and vulnerabilities that could potentially be exploited, ESBMC detects issues related to software safety as well.

## 6.2 Emerging Research Directions

The dataset containing all 112,000 C program files, along with the two .csv files are published on GitHub under the following link: <https://github.com/FormAI-Dataset>.

The diverse structure of the C programs generated in the FormAI dataset made it excellent for an unexpected use case, namely: fuzzing different applications. While running ESBMC on the dataset, we discovered and reported seven bugs in the module. After validating these issues, ESBMC developers managed to resolve them. These included errors in the goto-cc conversion and the creation of invalid SMT solver equations. Additionally, we identified bugs in the CBMC model checker—which is another BMC module—and the Clang compiler which failed to compile several programs that GNU C had no issue with. We promptly communicated these findings to the respective developers.

## 7 Limitations and Threats to Validity

While ESBMC is a robust tool for detecting many types of errors in C, as of today it is not suited to detect design flaws, semantic errors, or performance issues. As such, more vulnerabilities might be present in the code besides the detected ones. Also, to find all errors detectable by ESBMC, the unwind limit and verification time must be set to infinite. As we provided the original C programs and the instructions on how to run ESBMC, researchers

who invest additional computational resources have the potential to enhance our findings. Our results were reached with: “*–unwind 1 –overflow –multi-property –timeout 30*”. Even with the same settings but using a weaker CPU, for example, ESBMC might not be able to complete the verification process for a several programs, resulting in “*ERROR: Timed out*”. We were able to run the verification process for 106,139 programs. The 5861 difference is because approximately 1% of the samples were not compilable by Clang while ESBMC encountered crashes on the rest; as such, 5% of samples were not classified.

In addition to the reported findings, we found several instances of “*CWE-242: Use of Inherently Dangerous Function*”. Although ESBMC correctly reports several related functions as vulnerable, the reported line number of the vulnerability is often misleading. For instance, when the `gets()` function is invoked—which is declared in `io.c`—ESBMC marks a line number in `io.c` as the place of the vulnerability. This would be misleading for machine learning applications aiming to detect or fix vulnerabilities in the source code; therefore, we excluded such reports from the two csv files.

## 8 Conclusions

This paper shows that GPT-3.5-turbo notoriously introduces vulnerabilities when generating C code. The broad range of programming scenarios was instrumental in unveiling a variety of coding strategies and evaluating how GPT-3.5 manages specific tasks. This provided an opportunity to pinpoint situations where it might utilize questionable techniques that could potentially introduce a vulnerability. To facilitate a wide range of programming scenarios we created a zero-shot prompting technique and used GPT-3.5-turbo to generate C code. These programs constitute the FormAI dataset, containing 112,000 independent compilable C programs. We used the ESBMC bounded model checker to produce formally verifiable labels for bugs and vulnerabilities. In our experiment, we allocated a verification time of 30 seconds to each program, with the unwinding parameter set to 1. In total 197800 vulnerable functions were detected by ESBMC. Some programs contain multiple different errors. The labeling is provided in a .csv file which includes: Filename, Vulnerability type, Function name, Line number, and Error type. In addition, we provide an additional .csv file containing the C code as a separate column. Next, we connected the identified vulnerabilities to CWE identifiers. The FormAI dataset is a valuable resource for benchmarking vulnerability detection tools, or to train machine learning algorithms to possess the capabilities of the ESBMC module. The FormAI dataset proves to be a valuable instrument for fuzzing different applications, as we have demonstrated by identifying multiple bugs in the ESBMC and CBMC modules, as well as the Clang compiler.

## References

- [Aho et al.(2006)] Alfred V. Aho, Monica S. Lam, Ravi Sethi, and Jeffrey D. Ullman. 2006. *Compilers: Principles, Techniques, And Tools* (2nd ed.). Addison-Wesley Longman Publishing Co., Inc.
- [Avila(2022)] Risto Avila. 2022. *Embedded Software Programming Languages: Pros, Cons, and Comparisons of Popular Languages*. <https://www.qt.io/embedded-development-talk/embedded-software-programming-languages-pros-cons-and-comparisons-of-popular-languages>
- [Beyer(2023)] Dirk Beyer. 2023. Competition on Software Verification and Witness Validation: SV-COMP 2023. In *Tools and Algorithms for the Construction and Analysis of Systems*, Sriram Sankaranarayanan and Natasha Sharygina (Eds.). Springer Nature Switzerland, Cham, 495–522.
- [Black(2018)] Paul E. Black. 2018. A Software Assurance Reference Dataset: Thousands of Programs With Known Bugs. *Journal of Research of the National Institute of Standards and Technology* 123 (April 2018), 1–3. <https://doi.org/10.6028/jres.123.005>
- [Bui et al.(2023)] Nghi D. Q. Bui, Hung Le, Yue Wang, Junnan Li, Akhilesh Deepak Gotmare, and Steven C. H. Hoi. 2023. CodeTF: One-stop Transformer Library for State-of-the-art Code LLM. (May 2023). <http://arxiv.org/abs/2306.00029> arXiv:2306.00029 [cs].
- [Chakraborty et al.(2022)] Saikat Chakraborty, Rahul Krishna, Yangruibo Ding, and Baishakhi Ray. 2022. Deep Learning Based Vulnerability Detection: Are We There Yet? *IEEE Transactions on Software Engineering* 48, 9 (Sept. 2022), 3280–3296. <https://doi.org/10.1109/TSE.2021.3087402> Conference Name: IEEE Transactions on Software Engineering.
- [Charalambous et al.(2023)] Yiannis Charalambous, Norbert Tihanyi, Ridhi Jain, Youcheng Sun, Mohamed Amine Ferrag, and Lucas C. Cordeiro. 2023. A New Era in Software Security: Towards Self-Healing Software via Large Language Models and Formal Verification. (May 2023). <https://doi.org/10.48550/arXiv.2305.14752> arXiv:2305.14752 [cs].

- [Chavez et al.(2023)] Martin R. Chavez, Thomas S. Butler, Patricia Rekawek, Hye Heo, and Wendy L. Kinzler. 2023. Chat Generative Pre-trained Transformer: why we should embrace this technology. *American Journal of Obstetrics and Gynecology* 228, 6 (June 2023), 706–711. <https://doi.org/10.1016/j.ajog.2023.03.010>
- [Chen et al.(2023)] Yizheng Chen, Zhoujie Ding, Xinyun Chen, and David Wagner. 2023. DiverseVul: A New Vulnerable Source Code Dataset for Deep Learning Based Vulnerability Detection. (April 2023). <http://arxiv.org/abs/2304.00409> arXiv:2304.00409 [cs].
- [Cordeiro et al.(2012)] Lucas Cordeiro, Bernd Fischer, and Joao Marques-Silva. 2012. SMT-Based Bounded Model Checking for Embedded ANSI-C Software. *IEEE Transactions on Software Engineering* 38, 4 (July 2012), 957–974. <https://doi.org/10.1109/TSE.2011.59> Conference Name: IEEE Transactions on Software Engineering.
- [D’Silva et al.(2008)] Vijay D’Silva, Daniel Kroening, and Georg Weissenbacher. 2008. A Survey of Automated Techniques for Formal Software Verification. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 27, 7 (July 2008), 1165–1178. <https://doi.org/10.1109/TCAD.2008.923410> Conference Name: IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems.
- [Fan et al.(2020)] Jiahao Fan, Yi Li, Shaohua Wang, and Tien N. Nguyen. 2020. A C/C++ Code Vulnerability Dataset with Code Changes and CVE Summaries. In *Proceedings of the 17th International Conference on Mining Software Repositories (MSR ’20)*. Association for Computing Machinery, New York, NY, USA, 508–512. <https://doi.org/10.1145/3379597.3387501>
- [Gadelha et al.(2018)] Mikhail R Gadelha, Felipe R Monteiro, Jeremy Morse, Lucas C Cordeiro, Bernd Fischer, and Denis A Nicole. 2018. ESBMC 5.0: an industrial-strength C model checker. In *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*. 888–891.
- [Gadelha et al.(2023)] Mikhail R. Gadelha, Felipe R. Monteiro, Jeremy Morse, Lucas C. Cordeiro, Bernd Fischer, and Denis A. Nicole. 2023. *ESBMC: 5.0: An Industrial-Strength Model Checker*. <https://github.com/esbmc/esbmc> original-date: 2015-06-20T19:35:34Z.
- [Gadelha et al.(2019)] Mikhail Y. R. Gadelha, Enrico Steffnlongo, Lucas C. Cordeiro, Bernd Fischer, and Denis A. Nicole. 2019. SMT-based refutation of spurious bug reports in the clang static analyzer. In *Proceedings of the 41st International Conference on Software Engineering: Companion Proceedings, ICSE 2019, Montreal, QC, Canada, May 25-31, 2019*, Joanne M. Atlee, Tevfik Bultan, and Jon Whittle (Eds.). IEEE / ACM, 11–14. <https://doi.org/10.1109/ICSE-Companion.2019.00026>
- [Hutchinson et al.(2021)] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’21)*. Association for Computing Machinery, New York, NY, USA, 560–575. <https://doi.org/10.1145/3442188.3445918>
- [Imani et al.(2023)] Shima Imani, Liang Du, and Harsh Shrivastava. 2023. Mathprompter: Mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398* (2023).
- [Jain et al.(2023)] Ridhi Jain, Nicole Gervasoni, Mthandazo Ndhlovu, and Sanjay Rawat. 2023. A Code Centric Evaluation of C/C++ Vulnerability Datasets for Deep Learning Based Vulnerability Detection Techniques. In *Proceedings of the 16th Innovations in Software Engineering Conference*. 1–10.
- [Jr and Black(2012)] Frederick E. Boland Jr and Paul E. Black. 2012. The Juliet 1.1 C/C++ and Java Test Suite. *NIST* 45, 10 (Oct. 2012), 88–90. <https://www.nist.gov/publications/juliet-11-cc-and-java-test-suite> Last Modified: 2021-10-12T11:10-04:00 Publisher: Frederick E. Boland Jr., Paul E. Black.
- [Khoury et al.(2023)] Raphaël Khoury, Anderson R. Avila, Jacob Brunelle, and Baba Mamadou Camara. 2023. How Secure is Code Generated by ChatGPT? (April 2023). <http://arxiv.org/abs/2304.09655> arXiv:2304.09655 [cs].
- [Kim and Russell(2018)] Louis Kim and Rebecca Russell. 2018. Draper VDISC Dataset - Vulnerability Detection in Source Code. (Nov. 2018). <https://osf.io/d45bw/> Publisher: OSF.
- [Liu et al.(2023)] Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2023. Is Your Code Generated by ChatGPT Really Correct? Rigorous Evaluation of Large Language Models for Code Generation. (May 2023). <https://doi.org/10.48550/arXiv.2305.01210> arXiv:2305.01210 [cs].
- [Ma et al.(2023a)] Wei Ma, Shangqing Liu, Wenhan Wang, Qiang Hu, Ye Liu, Cen Zhang, Liming Nie, and Yang Liu. 2023a. The Scope of ChatGPT in Software Engineering: A Thorough Investigation. *arXiv preprint arXiv:2305.12138* (2023).

- [Ma et al.(2023b)] Wei Ma, Shangqing Liu, Wenhan Wang, Qiang Hu, Ye Liu, Cen Zhang, Liming Nie, and Yang Liu. 2023b. The Scope of ChatGPT in Software Engineering: A Thorough Investigation. (May 2023). <https://doi.org/10.48550/arXiv.2305.12138> arXiv:2305.12138 [cs].
- [OpenAI(2022)] OpenAI. 2022. GPT-3.5. Accessed May 17, 2023. <https://platform.openai.com/docs/models/gpt-3-5>.
- [OpenAI(2023)] OpenAI. 2023. *GPT-4 Technical Report*. Technical Report. <http://arxiv.org/abs/2303.08774> arXiv:2303.08774 [cs].
- [Pearce et al.(2021)] Hammond Pearce, Baleegh Ahmad, Benjamin Tan, Brendan Dolan-Gavitt, and Ramesh Karri. 2021. Asleep at the Keyboard? Assessing the Security of GitHub Copilot’s Code Contributions. (Dec. 2021). <https://doi.org/10.48550/arXiv.2108.09293> arXiv:2108.09293 [cs].
- [Pearce et al.(2022)] Hammond Pearce, Benjamin Tan, Baleegh Ahmad, Ramesh Karri, and Brendan Dolan-Gavitt. 2022. Examining Zero-Shot Vulnerability Repair with Large Language Models. (Aug. 2022). <http://arxiv.org/abs/2112.02125> arXiv:2112.02125 [cs].
- [Perry et al.(2022)] Neil Perry, Megha Srivastava, Deepak Kumar, and Dan Boneh. 2022. Do Users Write More Insecure Code with AI Assistants? (Dec. 2022). <http://arxiv.org/abs/2211.03622> arXiv:2211.03622 [cs].
- [Picard et al.(2020)] S. Picard, C. Chapdelaine, C. Cappi, L. Gardes, E. Jenn, B. Lefevre, and T. Soumarmon. 2020. Ensuring Dataset Quality for Machine Learning Certification. In *2020 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*. 275–282. <https://doi.org/10.1109/ISSREW51248.2020.00085>
- [rey Voas(1996)] Je rey Voas. 1996. Testing software for characteristics other than correctness: Safety, failure tolerance, and security. (1996).
- [Ross et al.(2023)] Steven I. Ross, Fernando Martinez, Stephanie Houde, Michael Muller, and Justin D. Weisz. 2023. The Programmer’s Assistant: Conversational Interaction with a Large Language Model for Software Development. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI ’23)*. Association for Computing Machinery, New York, NY, USA, 491–514. <https://doi.org/10.1145/3581641.3584037>
- [Russell et al.(2018)] Rebecca Russell, Louis Kim, Lei Hamilton, Tomo Lazovich, Jacob Harer, Onur Ozdemir, Paul Ellingwood, and Marc McConley. 2018. Automated Vulnerability Detection in Source Code Using Deep Representation Learning. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 757–762. <https://doi.org/10.1109/ICMLA.2018.00120>
- [Sadowski and Yi(2014)] Caitlin Sadowski and Jaeheon Yi. 2014. How developers use data race detection tools. In *Proceedings of the 5th Workshop on Evaluation and Usability of Programming Languages and Tools*. 43–51.
- [Sandoval et al.(2023)] Gustavo Sandoval, Hammond Pearce, Teo Nys, Ramesh Karri, Siddharth Garg, and Brendan Dolan-Gavitt. 2023. Lost at C: A User Study on the Security Implications of Large Language Model Code Assistants. (Feb. 2023). <http://arxiv.org/abs/2208.09727> arXiv:2208.09727 [cs].
- [Shumailov et al.(2023)] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The Curse of Recursion: Training on Generated Data Makes Models Forget. (May 2023). <http://arxiv.org/abs/2305.17493> arXiv:2305.17493 [cs].
- [Somoye(2023)] Funmi Looi Somoye. 2023. *Is ChatGPT free and unlimited? In short - yes.* <https://www.pcguides.com/apps/chat-gpt-free/>
- [Umawing(2023)] Jovi Umawing. 2023. *ChatGPT writes insecure code.* <https://www.malwarebytes.com/blog/news/2023/04/chatgpt-creates-not-so-secure-code-study-finds>
- [Wallace and Fujii(1989)] D.R. Wallace and R.U. Fujii. 1989. Software verification and validation: an overview. *IEEE Software* 6, 3 (May 1989), 10–17. <https://doi.org/10.1109/52.28119>
- [Wei et al.(2023)] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. (Jan. 2023). <https://doi.org/10.48550/arXiv.2201.11903> arXiv:2201.11903 [cs].
- [White et al.(2023)] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. (Feb. 2023). <https://doi.org/10.48550/arXiv.2302.11382> arXiv:2302.11382 [cs].
- [White et al.(2016)] Martin White, Michele Tufano, Christopher Vendome, and Denys Poshyvanyk. 2016. Deep learning code fragments for code clone detection. In *Proceedings of the 31st IEEE/ACM international conference on automated software engineering*. 87–98.

- [Xing et al.(2023)] Zhenchang Xing, Qing Huang, Yu Cheng, Liming Zhu, Qinghua Lu, and Xiwei Xu. 2023. Prompt Sapper: LLM-Empowered Software Engineering Infrastructure for AI-Native Services. (June 2023). <https://doi.org/10.48550/arXiv.2306.02230> arXiv:2306.02230 [cs].
- [Yao et al.(2023)] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. (May 2023). <http://arxiv.org/abs/2305.10601> arXiv:2305.10601 [cs].
- [Zhao and Huang(2018)] Gang Zhao and Jeff Huang. 2018. Deepsim: deep learning code functional similarity. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 141–151.
- [Zhou et al.(2019a)] Yaqin Zhou, Shangqing Liu, Jingkai Siow, Xiaoning Du, and Yang Liu. 2019a. Devign. <https://sites.google.com/view/devign>
- [Zhou et al.(2019b)] Yaqin Zhou, Shangqing Liu, Jingkai Siow, Xiaoning Du, and Yang Liu. 2019b. Devign: Effective Vulnerability Identification by Learning Comprehensive Program Semantics via Graph Neural Networks. (Sept. 2019). <http://arxiv.org/abs/1909.03496> arXiv:1909.03496 [cs, stat].