

Do Neutral Prompts Produce Insecure Code? FormAI-v2 Dataset: Labelling Vulnerabilities in Code Generated by Large Language Models

Norbert Tihanyi^{1*}, Tamas Bisztray², Mohamed Amine Ferrag¹,
Ridhi Jain¹, Lucas C. Cordeiro³

¹ Technology Innovation Institute (TII), Abu Dhabi, UAE.

² University of Oslo, Oslo, Norway.

³ University of Manchester, Manchester, UK.

*Corresponding author(s). E-mail(s): norbert.tihanyi@tii.ae;

Abstract

This study provides a comparative analysis of state-of-the-art large language models (LLMs), analyzing how likely they generate vulnerabilities when writing simple C programs using a neutral zero-shot prompt. We address a significant gap in the literature concerning the security properties of code produced by these models without specific directives. N. Tihanyi et al. introduced the FormAI dataset at PROMISE '23, containing 112,000 GPT-3.5-generated C programs, with over 51.24% identified as vulnerable. We expand that work by introducing the FormAI-v2 dataset comprising 265,000 compilable C programs generated using various LLMs, including robust models such as Google's GEMINI-pro, OpenAI's GPT-4, and TII's 180 billion-parameter Falcon, to Meta's specialized 13 billion-parameter CodeLLama2 and various other compact models. Each program in the dataset is labelled based on the vulnerabilities detected in its source code through formal verification using the Efficient SMT-based Context-Bounded Model Checker (ESBMC). This technique eliminates false positives by delivering a counterexample and ensures the exclusion of false negatives by completing the verification process. Our study reveals that at least 63.47% of the generated programs are vulnerable. The differences between the models are minor, as they all display similar coding errors with slight variations. Our research highlights that while LLMs offer promising capabilities for code generation, deploying their output in a production environment requires risk assessment and validation.

Keywords: Large Language Models, Vulnerability Classification, Formal Verification, Software Security, Artificial Intelligence, Dataset

1 Introduction

The introduction of Large Language Models (LLMs) is transforming software development and programming [1–3]. Every day, developers and computer scientists utilize various code creation and completion models to tackle different tasks [4, 5]. Research related to program synthesis using Generative Pre-trained Transformers (GPT) [6] is gaining significant traction, where initial studies indicate that the GPT models can generate syntactically correct yet vulnerable code [7]. A recent study conducted at Stanford University suggests that software engineers assisted by *OpenAI’s codex-davinci-002 model* were at a higher risk of introducing security flaws into their code [8]. As the usage of AI-based tools in coding continues to expand, understanding their potential to introduce software vulnerabilities becomes increasingly important. Given that LLMs are trained on data freely available on the internet, including potentially vulnerable code, there is a high risk that AI tools could replicate the same patterns. This leads to the question: *Is it safe to employ these models in real software projects?*

As a first step towards answering this question, N. Tihanyi et al. published the FormAI dataset [9] at the 19th International Conference on Predictive Models and Data Analytics in Software Engineering (PROMISE’23). This dataset stands out as the first and largest collection of AI-generated compilable C programs with vulnerability classification, featuring 112,000 samples. To guarantee the diversity of generated C codes, the authors developed a framework designed to produce a variety of programs that cover multiple coding scenarios, efficiently facilitating real-world bugs. The study employed Bounded Model Checking (BMC), a technique within Formal Verification (FV), to evaluate the security properties of the dataset. It revealed that 51.24% of the C programs generated by GPT-3.5 were vulnerable. The original research discussed in [9] had three primary limitations that this comprehensive study aims to address:

- a. Firstly, it exclusively focused on OpenAI’s GPT-3.5 without evaluating other models. To bridge this gap, our study has been extended to encompass eight advanced LLMs, such as Google’s *Gemini-pro* [10], TII’s *Falcon-180B* [11], or Meta’s *CodeLLama* [12]. The models can be seen in Figure 1.
- b. Secondly, the initial findings (51.24%) may have been under-reported due to the inherent limitations of BMC, indicating that the actual percentage of vulnerabilities could be higher. To address this issue, we transitioned our verification approach from bounded to *unbounded* verification, thereby enhancing the depth and accuracy of our security assessments [13–16].
- c. Lastly, the complexity of the programs generated by GPT-3.5 was not thoroughly analyzed. To address these gaps, we have calculated the *cyclomatic complexity* CC [17] of each program generated by these LLMs to understand their intrinsic complexity better.

This study answers the following research questions:

- **RQ1:** Does the security of LLM-generated code vary significantly between different models?
- **RQ2:** What are the most typical vulnerabilities introduced during C code generation by different LLMs using neutral zero-shot prompts?

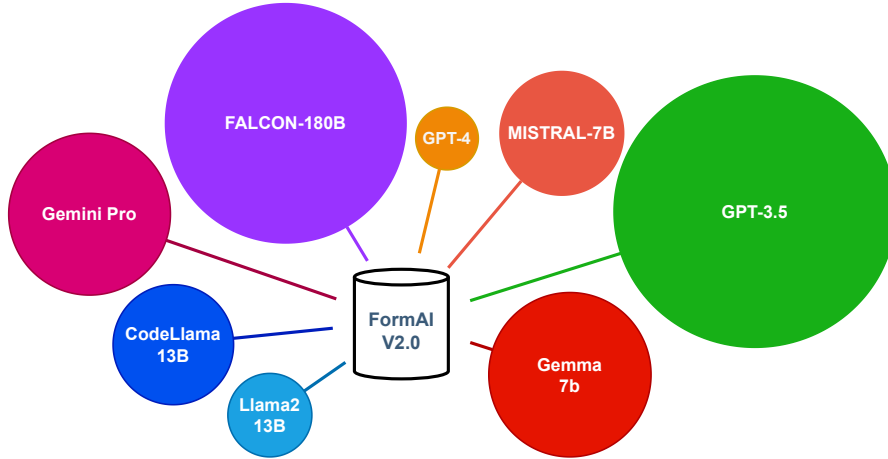


Fig. 1: The eight LLM models used in this study for code security comparison.

This is the first comprehensive study to compare vulnerabilities in code generated by various state-of-the-art LLMs. To summarize, this paper holds the following original contributions:

- We present the **FormAI-v2** dataset, consisting of 265,000 compilable C program samples generated by eight different LLM models. Each C sample has been systematically labeled based on vulnerabilities identified through FV methods, particularly using the Efficient SMT-based Bounded Model Checker (ESBMC) [14–16] tool with an *unbounded* setting;
- A detailed study to determine which LLMs produce code with the highest and the lowest number of vulnerabilities;
- We offer a thorough analysis of the generated programs, including distributions of C keywords and the cyclomatic complexity. This analysis enables the comparison of LLMs in terms of coding complexity;
- We made the **FormAI-v2** dataset available to the research community, including all generated C samples and classification results. The dataset can be accessed on our project website at <https://github.com/FormAI-Dataset>.

The remaining sections of this paper are organized as follows: Section 2 provides an in-depth discussion on the motivation behind our research, outlining the primary questions that drove our study. Section 3 presents a comprehensive overview of the literature related, highlighting significant previous studies and their findings. Section 4 introduces the concepts of formal verification, focusing on the ESBMC module. Section 5 details the methodology we adopted to develop and label our dataset. Section 6 thoroughly evaluates our research findings and discusses their implications. Section 7 explores the limitations and threats to the validity of our study and proposes potential future research directions that could expand upon our results. Finally, Section 8 concludes the paper by summarising our contributions and addressing the research questions posed in this study.

2 Motivation

LLMs are typically employed for simpler tasks, such as crafting a prime number generation function or composing a basic program to sort an array, rather than tackling extensive projects involving thousands of lines of code [8]. The latest generation of LLMs can easily solve these simple tasks without facing any challenges. So far, the main area of interest in LLM-based code generation has been correctness. Datasets such as HumanEval [18] provide programming challenges to assess the performance of models in correctly solving various problems. For example, GPT-4 achieves a 67% success rate in solving tasks compared to 48.1% for GPT-3.5 [19]. On the contrary, even basic programs can pose challenges for the state-of-the-art LLM models regarding secure coding. To illustrate the issue, imagine a situation where a programmer asks GPT-4 the following: “Create a C program that prompts the user to input two numbers and then calculate their sum”. The code generated by GPT-4 is presented on the left in Figure 2, while the output from the formal verification tool ESBMC 7.5 is shown on the right.

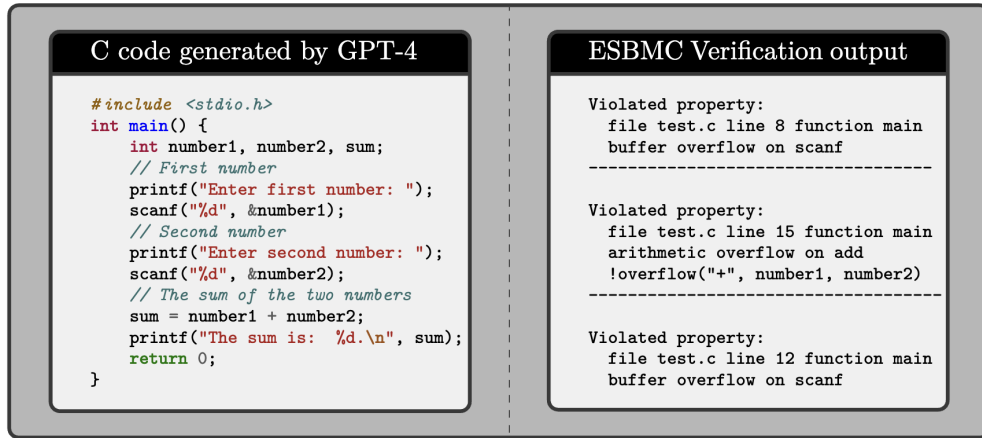


Fig. 2: Motivation example: GPT-4 produced code with security vulnerabilities, demonstrated through formal verification results

This simple code contains three potential security vulnerabilities. It exhibits an integer overflow during the addition of the variables *number1* and *number2*, as well as two buffer overflows through the `scanf()` functions that retrieve input from the user. 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.

When GPT-4 is requested to write a secure version of this code using the following prompt: “Create a C program that prompts the user to input two numbers and then calculates their sum. Be careful and avoid security vulnerabilities.”, it only attempts to fix entering non-integer inputs by adding the following code snippet (Fig. 3):

GPT-4 generated code snippet attempting a secure version

```
printf("Enter first number: ");
scanResult = scanf("%d", &number1);
if (scanResult != 1) {
    printf("Invalid input. Please enter an integer.\n");
    return 1;}
```

Fig. 3: Response to a generic prompt requesting a secure version.

Even after requesting a “secure” zero-shot prompt, all three original issues remain unresolved. Despite the significant advancements from GPT-3.5—which exhibited the same issue [9]—to GPT-4, our motivational example indicates that GPT-4 continues to produce code with vulnerabilities. Even if specifically requested in the prompt to avoid integer overflow in the program, the issue persists (Fig. 4):

Code snippet aiming to fix integer overflow (failing to do so)

```
if ((number1 > 0 && number2 > INT_MAX - number1) ||
    (number1 < 0 && number2 < INT_MIN - number1)) {
    printf("Error: Integer overflow detected.\n");
    return 1;}
```

Fig. 4: Zero-shot prompt requesting a fix for integer overflow

We want to emphasize that simply requesting a secure version is not an effective approach towards achieving a secure code for the following reason: Code completion tools such as GitHub Copilot ¹ or Amazon Code Whisperer ² suggest code snippets based on contextual analysis and training data, which has also been shown to produce vulnerable code [20]. In such scenarios, the ability to prompt is limited (it can be attempted through comments in the code). In addition, GitHub Copilot is powered by a variant of the GPT (Generative Pre-trained Transformer) model called Codex, which OpenAI developed. The underlying issue will remain if these models are not trained to produce secure code. Based on this observation, we aim to conduct a comprehensive study involving various state-of-the-art models to address our research questions.

¹<https://github.com/features/copilot/>

²<https://aws.amazon.com/codewhisperer/>

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 LLMs in Software Engineering

In software engineering (SE), it is essential to guarantee three main aspects of the code: correctness, safety, and security of the programs created. Functionally correct code should yield the expected outcomes for each input it processes. Safety in code means constructing fail-safe systems, protecting against accidental or unexpected inputs that might produce logically correct but undesirable results. Software security involves fortifying the software against external threats and deliberate attacks [21]. In a comprehensive study, Anwar et al. [22] highlight important safety issues related to LLMs beyond SE, from the disruptive socioeconomic impacts and cybersecurity risks to ethical issues. Vassilka et al. [23] discuss the need for SE education to adapt to AI advancements and prepare future software engineers to effectively and ethically utilize these technologies in their careers.

To assess correctness, datasets such as HumanEval [18] serve as a benchmark to measure the problem-solving abilities of AI models for problems related to language comprehension, reasoning, algorithms, simple mathematics, coding, and logical thinking. There are several other similar datasets, such as MBPP [24] to assess code synthesis capabilities on elementary Python challenges, or CodeXGLUE [25], to test code completion, translation, and understanding of different LLMs.

Frameworks and techniques for turning prompts into executable code for SE are rapidly emerging, but the main focus is often functional correctness, omitting important security aspects [26–28]. There has been an arms race between researchers to excel in benchmarks using zero or few-shot frameworks [29, 30], multi-agent frameworks [31], fine-tuned models [32], and various other methods. As AI models evolve, their problem-solving capabilities improve significantly. However, whether these advancements also enhance the safety and security properties of the code they generate remains largely unclear and under-researched.

In [33], Lin et al. assessed different software process models to evaluate how these models affect code correctness (Pass@1³). They also assessed the code quality of the AI-generated code by running static code checkers to uncover code smells⁴. This work had an interesting finding: the proposed software process models improved the quality of the generated code by significantly reducing code smells compared to what GPT-3.5 outputs by itself. Code smells or bad coding practices will not outright introduce vulnerabilities. However, several small-scale studies point to the fact that LLMs negatively impact software development from a security perspective. In [34], the authors generated 21 small programs in five different languages: C, C++, Python, HTML,

³This metric highlights the model’s ability to produce correct and functional code on its first try without any revisions or corrections.

⁴Code smells are patterns in code that hint at potential problems, making maintenance harder but not necessarily causing immediate errors. They suggest areas where the code may need to be refactored for better quality and reliability.

and Java. Combining manual verification with GPT-based vulnerability detection, the study found that only 5 of the 21 generated programs were initially secure.

In [35], Pearce et al. conclude that the control group utilized GitHub’s Copilot to solve arithmetic operations accurately. This work 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 scenarios, which is one of the primary goals the FormAI dataset strives to achieve. These studies indicate that AI tools, and in particular ChatGPT, can produce code containing vulnerabilities as of today.

Ma et al. [36] assessed the capabilities and limitations of ChatGPT for SE and provided initial insights into why the programs generated by language models are syntactically correct but potentially vulnerable. A study by Microsoft [37] found that GPT models encounter difficulties when accurately solving arithmetic operations. This aligns with our findings in Section 2.

In a comprehensive literature review, Hou et al. [38] examined LLMs’ application, effects, and possible limitations on SE. This study reveals that LLMs are extensively employed across software development, appearing in 229 papers for 24 distinct SE tasks, predominantly in code generation and program repair. It also identifies over 70 LLMs, classifying them into three architectural types: decoder-only, encoder-decoder, and encoder-only. Each architecture serves specific functions—encoder-only for in-depth understanding, encoder-decoder for combined understanding and generation, and decoder-only primarily for generation tasks. This work highlights an interesting gap: there are dozens of research papers aiming to perform vulnerability detection in source code using machine learning (ML) and LLMs [39–51], however, assessing software safety and security properties of LLM-generated code on a large-scale has not yet been performed apart from our original work [9] for C, and recently by [52] for PHP code. Both studies evaluated a single model in a zero-shot code generation scenario, while our current work also conducts a comparison of the performance of different models.

In [53] Shumailov et al. highlighted a phenomenon known as “*model collapse*”. 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 ML algorithms 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. This raises the question of whether the FormAI dataset is suitable for fine-tuning and ML purposes. It is important to note that the AI-generated code is just one part of the dataset. Most importantly, the vulnerability labeling was not done by AI but by the ESBMC formal verification tool. This way, models trained on this dataset can essentially learn the skills of a formal verification tool (or at least try to achieve the best optimal outcomes).

The programs are generated through a dynamic zero-shot prompting method, and the generated programs are not modified by any AI system afterward. While the primary goal of our paper is to investigate and compare the secure coding abilities of

different LLM models, these conditions make the **FormAI-v2** dataset suitable for ML purposes. On the other hand, AI models 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 [54].

3.2 Existing Databases for Vulnerable C Code

We show how the **FormAI-v2** dataset compares to seven widely studied datasets containing vulnerable code and the previous version of the dataset published in [9]. The examined datasets are: Big-Vul [55], Draper [56, 57], SARD [58], Juliet [59], Devign [60], REVEAL [61], DiverseVul [54], and FormAI-v1 [62]. 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 [54, 63].

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

Datasets Specs	Big-Vul	Draper	SARD	Juliet	Devign	REVEAL	Diverse Vul	FormAI	FormAI v2
Language	C/C++	C/C++	Multi	Multi	C	C/C++	C/C++	C	C
Source	RW	Syn + RW	Syn + RW	Syn	RW	RW	RW	AI	AI
Dataset size	189k	1,274k	101k	106k	28k	23k	379k	112k	150k
Vul. Snippets	100%	5.62%	100%	100%	46.05%	9.85%	7.02%	51.24%	61%
Multi. Vulns.	✗	✓	✗	✗	✗	✗	✗	✓	✓
Compilable	✗	✗	✓	✓	✗	✗	✗	✓	✓
Granularity	Func	Func	Prog	Prog	Func	Func	Func	Prog	Prog
Class. Type	CVE CWE	CWE	CWE	CWE	CVE	CVE	CWE	CWE	CWE
Avg. LOC.	30	29	114	125	112	32	44	79	82
Labelling Method	P	S	B/S/M	B	M	P	P	F	F

Legend:

Multi: Multi-Language Dataset, **RW:** Real World, **Syn:** Synthetic, **AI:** AI-generated,
Func: Function level granularity, **Prog:** Program level granularity,
CVE: Common Vulnerabilities and Exposures, **CWE:** Common Weakness Enumeration,
P: GitHub Commits Patching a Vulnerability, **S:** Static Analyzer,
B: By Design Vulnerable, **F:** Formal Verification with ESBMC, **M:** Manual Labeling

Big-Vul, Draper, Devign, REVEAL, and DiverseVul comprise vulnerable real-world functions from open-source applications. These five datasets do not include all the samples’ dependencies; 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, using various vulnerability labeling methodologies such as P, where functions are considered vulnerable before

receiving GitHub commits that address detected vulnerabilities; **M**, which involves manual labeling; **S**, which uses static analyzers; and **B**, designated as by design vulnerable without the use of a vulnerability verification tool. It’s important to note that the size of these datasets can be misleading, as many include samples from languages other than the one primarily studied. For example, SARD includes not only C and C++ but also Java, PHP, and C#. Moreover, newly released sets often incorporate previous datasets or scrape the same GitHub repositories, making them redundant.

For example, Draper 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: Cppcheck and Flawfinder [56]. However, the paper does not mention if vulnerabilities were manually verified or if any confirmation has been performed to root out false positives. In [54], 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 crucial for ensuring software’s safety and security properties. It employs a variety of techniques, each with its strengths and limitations. These techniques include manual verification, static analysis, dynamic analysis, formal verification, and increasingly, machine learning-based approaches such as those involving LLMs [36, 64–67].

Manual verification involves human-driven processes such as code reviews and manual testing. While these methods effectively catch complex errors that automated tools might miss, they are labor-intensive and not scalable to large codebases or frequent updates. Static analysis evaluates source code without executing it, using static symbolic execution, data flow analysis, and control flow analysis. Style checking enforces coding standards for better readability and maintainability. These methods collectively enhance software integrity. The drawbacks are that this method can miss vulnerabilities that manifest only during runtime interactions and often introduce false positive results. Dynamic analysis tests the software’s behavior during execution [68]. It includes fuzzing, automated testing, run-time verification, and profiling. This technique requires executable code and often significant setup to simulate different environments and may not cover all execution paths.

Formal Verification (FV) uses mathematical proofs to verify the correctness of algorithms against their specifications. It is the most rigorous form of software verification and is used in applications where reliability is critical, such as aerospace and medical devices. However, FV can be time-consuming and requires specialized knowledge, limiting its widespread adoption [21]. Recent advancements include machine learning techniques, particularly LLMs, in various aspects of software verification [69]. LLMs can assist in automated code review by suggesting improvements, detecting vulnerabilities, generating test cases, fixing bugs, and creating documentation. Despite their potential [70–76], LLMs, on their own face limitations such as a lack of understanding of code semantics and difficulty in handling highly domain-specific knowledge [77],

and they depend heavily on the quality and variety of the training data. Using LLMs as part of a framework to complement other techniques is, however, a promising area of research [7, 9, 78, 79].

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 further highlighted challenges in the tool’s ability to generate functionally correct code [80]. While LLMs struggle with detection by themselves, in [7], the authors demonstrated that GPT-3.5 could efficiently fix errors if the output of the ESBMC verifier is provided. Program repair is another emerging area where the application of LLMs is showing real promise, where in addition to fine-tuning strategies, the combination of LLMs with other tools appears to be an effective method [41, 81–97]. In [98], the authors call for innovations to enhance automated vulnerability repair, particularly for developing more extensive training datasets to optimize LLMs.

3.4 Overview on Cyclomatic Complexity

Diversity is vital in our dataset to ensure a comprehensive representation of real-life vulnerabilities, facilitating a fair comparison across different LLMs. We add a new metric in **FormAI-v2** to measure code complexity: cyclomatic complexity (CC). This metric, introduced by McCabe [17], quantifies the maintainability and complexity of programs by measuring the number of linearly independent paths within a program’s control-flow graph (CFG) [99]. It’s calculated as $V(G)$ for a graph G with n vertices, e edges, and p connected components.

$$V(G) = e - n + 2p.$$

This metric, derived from the CFG – a directed graph of basic program blocks linked by control paths – is a numerical indicator of a program’s structural complexity. While complexity in functions increases due to multiple decision points, such as *if*-statements and loops imply higher testing and maintenance demands, critics note that CC might not fully capture complexities in modern software that extensively uses polymorphism and multi-threading. As will be discussed in Section 6, our findings suggest that cyclomatic complexity (CC) can indicate whether a model is capable of generating well-maintained, well-structured code or if it tends to produce poorly designed programs with high cyclomatic complexity, where higher numbers imply worse outcomes and lower numbers are preferable. Cyclomatic complexity can provide insights into the potential difficulty of testing, maintaining, or troubleshooting a piece of code and an indication of the likelihood of the code producing errors. Additionally, incorporating CC as a new feature in the dataset could enhance classification accuracy when used during the machine learning training phase.

The next section will offer necessary insights into formal verification to clarify the methodology employed in developing the dataset.

4 Formal Verification (FV) and Bounded Model Checking (BMC)

As manually labelling the entire dataset is unfeasible for such a large corpus of data, we employ a Formal Verification (FV) method, called Bounded Model Checking (BMC), to accurately detect vulnerabilities. In contrast to traditional static analysis tools, which often yield a high incidence of false positives due to reliance on pattern recognition without mathematical grounding [100], BMC offers rigorous validation. The Efficient SMT-based Bounded Model Checker (ESBMC) [14] effectively minimizes false positives and negatives. This approach ensures a fair comparison between programs produced by different LLMs.

4.1 State Transition System

A state transition system $M = (S, T, S_0)$ represents an abstract machine consisting of a collection of states S , where $S_0 \subseteq S$ indicates the initial states, and $T \subseteq S \times S$ specifies the transition relation, illustrating the potential state transitions within the system. Every state $s \in S$ is characterized by the value of the program counter (pc) and the values of all program variables. The initial state s_0 sets the program's starting location. Transitions between states denoted as $T = (s_i, s_{i+1}) \in T$, between any two states s_i and s_{i+1} , are associated with a logical formula $T(s_i, s_{i+1})$ that describes the constraints on the program counter and program variables relevant to that transition.

4.2 Bounded Model Checking

BMC is employed in FV to ascertain the correctness of a system up to a finite number of steps. This approach models the system as a finite state transition system and methodically examines its state space to a predefined depth. Recent BMC modules are capable of processing a variety of programming languages such as C, C++, JAVA, or Kotlin [100–106]. The process starts with the program code from which a CFG is derived. In this CFG, nodes represent deterministic or nondeterministic assignments or conditional statements, while edges indicate potential changes in the program's control flow. Essentially, each node is a block that encapsulates a set of instructions with a unique entry and exit point, and edges indicate potential transitions to other blocks. The CFG is then converted into Static Single Assignment (SSA) form and further into a State Transition System (STS), which a Satisfiability Modulo Theories (SMT) solver can interpret. The SMT solver checks if a given formula, representing the program's correctness within a bound k , is satisfiable, indicating the existence of a counterexample to the properties being verified. If no errors are found and the formula is unsatisfiable within the bound k , it suggests the program has no vulnerabilities within that limit. Thus, if the solver concludes satisfiability within a bound $\leq k$, it confirms the presence of a vulnerability through a counterexample.

Consider a program $\mathcal{P}$ under verification modeled as a finite STS, denoted by the triplet $\mathcal{ST} = (S, R, I)$, where S represents the set of states, $R \subseteq S \times S$ represents the set of transitions, and $I \subseteq S$, including elements such as $s_n, \dots, s_m$, marks the initial state set. 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, s_0 , specifies the initial program location on the CFG. Every transition $T = (s_i, s_{i+1}) \in R$, connecting two states s_i and s_{i+1} , correlates with a logical expression $T(s_i, s_{i+1})$ that constrains the program counter (pc) and variable values pertinent to the transition.

In the context of BMC, the properties under examination are defined as follows: $\phi(s)$ represents a logical formula reflecting states that fulfill a safety or security criterion, whereas $\psi(s)$ encodes a logical statement representing states that meet the completeness threshold, synonymous with program termination. Notably, $\psi(s)$ incorporates loop unwinding to avoid surpassing the program’s maximum loop iterations. Termination and error conditions are mutually exclusive, rendering $\phi(s) \wedge \psi(s)$ inherently unsatisfiable. If $T(s_i, s_{i+1}) \vee \phi(s)$ is unsatisfiable, state s is considered a deadlock state.

4.2.1 The Bounded Model Checking Problem

Based on this information, we can define the bounded model checking problem as BMC_Φ , which involves creating a logical statement. The truth of this statement determines if the program $\mathcal{P}$ has a counterexample with a maximum length of k . The formula can only be satisfied if a counterexample fitting within the predetermined length restriction is present, i.e.:

$$BMC_\Phi(k) = I(s_0) \wedge \bigwedge_{i=1}^{k-1} T(s_i, s_{i+1}) \wedge \bigvee_{i=1}^k \neg\phi(s_i). \quad (1)$$

Herein, I denotes the initial state set of $\mathcal{ST}$, and $T(s_i, s_{i+1})$ embodies the transition relation within $\mathcal{ST}$ between consecutive time steps i and $i + 1$. Thus, the logical expression $I(s_0) \wedge \bigwedge_{i=1}^{k-1} T(s_i, s_{i+1})$ depicts the execution pathways of $\mathcal{ST}$ spanning a length 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 ϕ is violated. If $BMC_\Phi(k)$ is satisfied, it implies a violation of ϕ , permitting an SMT solver to deduce a satisfying assignment from which the program variables’ values can be derived to assemble a counterexample. By definition, a counterexample, or trace, for a violated property ϕ , is defined as a finite sequence of states $s_0, \dots, s_k$, where $s_0, \dots, s_k \in S$ and $T(s_i, s_{i+1})$ holds for $0 \leq i < k$. These counterexamples hold significant importance for us, as we explicitly seek out these violations to compare and determine which code generated by LLMs is “more secure”. Fewer violated properties indicate that the LLM can produce more secure code. In this context, it is important to note the influence of cyclomatic complexity. Merely experiencing fewer errors with an LLM does not necessarily indicate superiority; it could simply be generating less intricate problems. Therefore, assessing both property violations and cyclomatic complexity concurrently offers a more comprehensive understanding of an LLM’s proficiency in secure coding.

If the Equation (1) is unsatisfiable, it implicates no error state as reachable within k steps or fewer, hence no software vulnerability exists within the 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 method detects security concerns such as buffer overflows and

division by zero or null dereference failures. It is worth noting that if a program is classified as vulnerable, this assessment relies on counterexamples, effectively eliminating the possibility of false positives. Conversely, in situations where no counterexample exists, we can confidently assert that the program is free from vulnerabilities up to the bound k , ensuring the absence of false negatives. By adopting this strategy, we aim to classify each program by detecting violated properties in the generated code. While this method offers greater precision and improved outcomes compared to standard static analysis tools, it is also more time-consuming, demands significant resources, and is limited to identifying only a predefined set of vulnerabilities.

4.3 Efficient SMT-based Context-Bounded Model Checker

Annually, the International Competition on Software Verification, known as SV-COMP, challenges various programs to detect bugs and ensure software safety. ESBMC excels in this competition by solving the highest number of verification tasks within a strict 10-30 second time limit per program, as evidenced in SV-COMP 2023 [107]. Given its performance, ESBMC was selected as our primary BMC tool. As a robust, open-source model checker for C/C++, Kotlin, and Solidity programs, ESBMC addresses a wide range of safety properties and program assertions, including out-of-bounds array access, illegal pointer dereference, integer overflows, and memory leaks. It employs advanced verification techniques such as incremental BMC and k -induction, supported by state-of-the-art SMT and Constraint Programming (CP) solvers. ESBMC's effectiveness in falsification and bug-finding is underscored by its multiple accolades in SV-COMP, including 6 gold, 4 silver, and 10 bronze medals.

4.3.1 Unbounded model checking using ESBMC

In the original work of [9], the bounded model checking (BMC) problem was addressed with a bound set to $k = 1$. Although we have observed that most vulnerabilities can be detected with these settings empirically, the nature of BMC can sometimes lead to erroneous conclusions. For example, if a property violation occurs at level $k = 2$, then $BMC_{\Phi}(1)$ will fail to detect the bug and report the verification as successful, giving a false impression. As a result, in the **FormAI-v1** dataset, numerous samples were previously classified as "*NON-VULNERABLE up to bound k* ". We have transitioned from bounded to *unbounded* model checking to capture more vulnerabilities or prove their absence for each sample. This means that the program is incrementally unwound until a bug is found or until the completeness threshold is reached, i.e., states corresponding to the program terminating. This incremental BMC approach ensures that smaller problems are solved sequentially instead of guessing an upper bound for the verification. Applying these settings, we have successfully identified more vulnerabilities or prove that no vulnerability exists. Consequently, if the verification process is completed successfully, we can conclude that the program has no violated properties. While this approach requires significantly more computational power, it has proven effective in revealing a greater number of vulnerabilities or proving their absence, as we will demonstrate in Section 6.

5 Methodology and Dataset Creation

The FormAI-v2 dataset consists of two main components: the AI-generated C programs and their vulnerability labeling. In the data generation phase, we have created 265,000 samples. This paper implements a similar methodology introduced in the original work [9]. The main difference is that we employ eight different LLMs rather than solely GPT-3.5. Moreover, as discussed in the previous section, we have adopted a more expansive *unbounded* verification model. This section covers the dataset generation methodology and the specific prompts for producing the C code samples. The overview of the generation and vulnerability labeling mechanism is depicted in Figure 5.

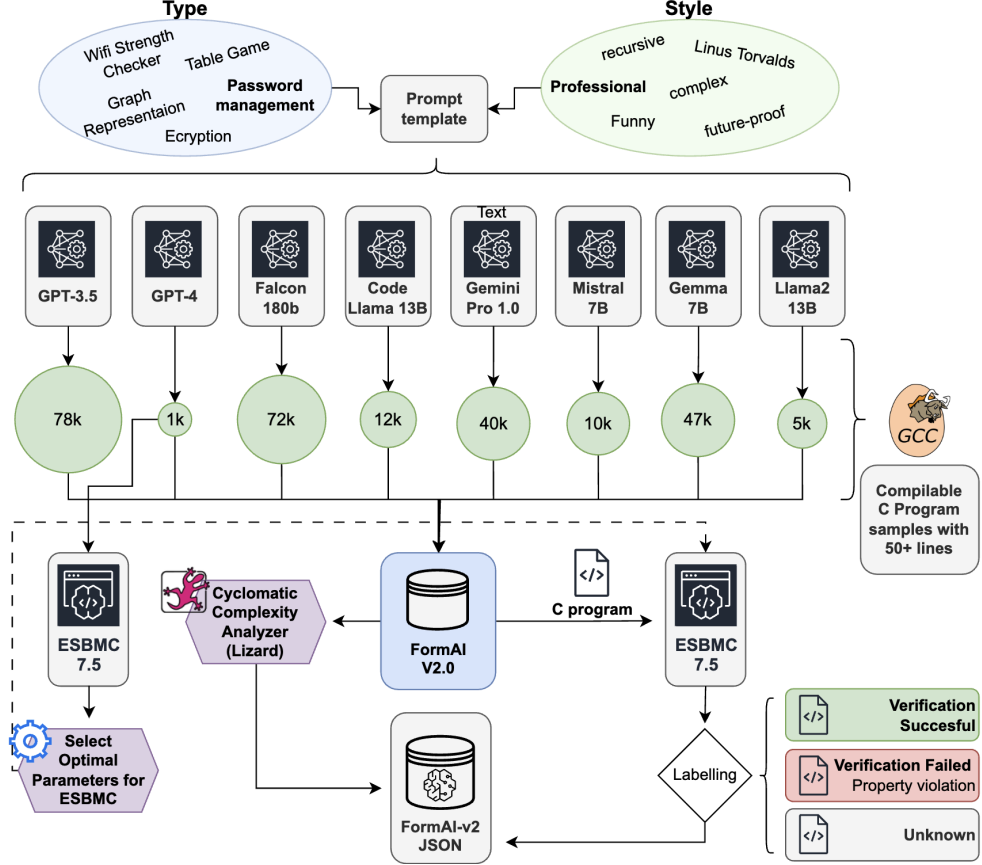


Fig. 5: FormAI-v2 dataset generation Framework using different LLMs.

5.1 Code Generation

During the creation process, special attention was given to ensure the diversity of the **FormAI-v2** dataset, which contains 265,000 compilable C samples. Using a basic prompt like *"generate a C program"* often results in similar outputs, such as programs that add two numbers or perform simple string manipulations, which do not meet our objectives. We need to systematically generate a comprehensive and varied collection of small programs. To meet this goal, we have developed a prompting method consisting of a dynamic 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 Figure 6.

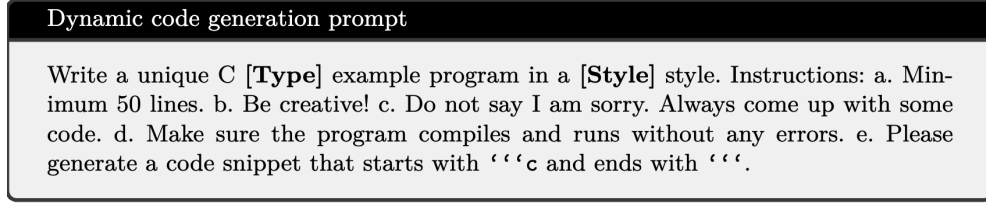


Fig. 6: Dynamic code generation prompt

The dynamic part of the prompt, highlighted as **[Type]** and **[Style]**, represent distinct categories within the prompt, each encompassing different elements. Every API call randomly selects a Type category from a set of 200 elements. This category contains topics such as Wi-Fi Signal Strength Analyzer, QR code reader, Image Steganography, Pixel Art Generator, Scientific Calculator Implementation, etc. Similarly, a coding Style is chosen from a set of 100 elements during each query. This helps minimize repetition, as coding styles such as "excited", "relaxed", or "mathematical" are randomly combined with a Type category. By employing this method, we can generate $200 \times 100 = 20,000$ distinct combinations. As demonstrated by insights from [80, 108], there's a need for a code base that supports diverse settings while ensuring tasks remain concise to fit within the token constraints of large language models (LLMs). With this consideration, we designed tasks in the Type category, selecting prompts that LLMs can efficiently process. For instance, complex prompts like "Create a CRUD application using React for the front-end, Node.js with Express for the back-end, and MongoDB for the database" must be broken down into smaller, manageable tasks. Furthermore, tasks with different styles, such as "File handling" with a "thriller" versus a "happy" style, lead to different representations in the vector space upon tokenization. Despite potential compatibility issues between certain Type-Style combinations, encouraging LLMs to code in varied styles has generally enhanced the diversity and distinctiveness of responses to identical tasks.

Decreasing the number of unsuccessful queries by refining the prompt is important from an efficiency perspective. We have established five instructions in each prompt to minimize the error within the generated code. These instructions, along with their corresponding explanations, are the following:

1. **Minimum 50 lines:** This encourages the LLM to avoid the generation of overly simplistic code with only a few lines (which occasionally still happens);
2. **Be creative!:** The purpose of this instruction is to generate a more diverse dataset;
3. **Do not say I am sorry:** This instruction aims to circumvent objections and responses such as “As an AI model, I cannot generate code”, and similar statements.
4. **Make sure the program compiles:** This instruction encourages the model to include header files and create a complete and compilable program.
5. **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 compiler⁵ is employed to verify whether the corresponding code is compilable. During the code generation process, we ensure that the **FormAI-v2** dataset exclusively consists of compilable code while excluding any other code that does not meet this criterion. Different models can generate varying percentages of compilable code depending on their parameter size. Models like GPT-4, GEMINI-pro, or FALCON-180B can achieve compilation rates higher than 90%, whereas smaller models with 7B parameters typically produce C code with a compilability rate between 55-70%. 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`. The generation of 265,000 code samples was completed within 48 hours. As the cost of generation associated with different models can significantly vary, we did not generate the same number of samples from each model. For example, GPT-4 API calls can be up to 60 times as expensive as GPT-3.5. Table 2 shows how many samples we acquired from each LLM.

Table 2: Content of the FormAI v2.0 dataset.

LLM Model	Company	Size	License	Sample Size
GPT-4	OpenAI	N/A	Proprietary	1000
Llama2-13B	Meta	13B	Open	5000
Mistral-7B	Mistral AI	7B	Apache 2.0	10000
CodeLlama-13B	Meta	13B	Proprietary	12000
Gemini 1.0 Pro	Google	N/A	Proprietary	40000
Gemma-7b	Google	7B	Gemma-terms-of-use	47000
Falcon-180B	TII	180B	Apache 2.0	72000
GPT-3.5	OpenAI	175B	Proprietary	78000

⁵<https://gcc.gnu.org>

5.2 Vulnerability Classification

Following the code generation, we executed ESBMC on each file to classify them. Let us denote the set of all the generated C samples by Σ , such that $\Sigma = \{c_1, c_2, \dots, c_{265,000}\}$, where each c_i represents an individual sample. As shown in Table 3, we can group all the samples into four distinct categories. These categories are mutually exclusive, meaning a single sample cannot belong to more than one category.

Table 3: Summary of Vulnerabilities with Detailed Descriptions.

Category	Description
$\mathcal{VS}$: Verification Success	The set of samples for which the verification process completed successfully with no vulnerabilities detected.
$\mathcal{VU}$: Verification Unknown (Timeout)	The set of samples for which the verification process was not completed within the provided time frame; no vulnerabilities found before timeout.
$\mathcal{VF}$: Verification Failed	The set of samples for which the verification status failed, vulnerabilities detected by ESBMC based on counterexamples.
$\mathcal{ER}$: Error	The set of samples for which the verification status resulted in an error.

A significant change has been made compared to the classification in [9]. The previous paper only included vulnerabilities if a program’s verification process was completely finished. Our new approach includes every instance where ESBMC found a counterexample. This decision was motivated by identifying as many vulnerabilities as possible and providing the most accurate ratio of vulnerable to non-vulnerable programs. Dismissing these intermediate detection results could lead to a misleading assessment, potentially underestimating the number of vulnerabilities produced by larger models. Thus, including all counterexamples ensures a more accurate representation of the true vulnerability landscape across different LLM outputs. For this reason, the category “*TIMEOUT*” ($\mathcal{TO}$) has been renamed to $\mathcal{VU}$: Verification Unknown. The category labeled as “*ERROR*” ($\mathcal{ER}$) encompasses all instances where the verification process faced errors or crashes in the core ESBMC module, GOTO converter, SMT solver, or the clang compiler module. These samples are omitted from the classification because they cannot be handled using the latest ESBMC module. “*Verification failed*” ($\mathcal{VF}$) represents the main focus of our interest. This set of samples, where the verification status was unsuccessful, had vulnerabilities identified by ESBMC through counterexamples. We divided $\mathcal{VF}$ into 6 main categories, and 19 subcategories. Note that in the JSON file that contains the vulnerability classification, the exact type of vulnerability is always indicated as provided by the ESBMC module, as shown in Appendix Fig. 1. The categories and subcategories, along with the precise distribution of vulnerabilities across the entire dataset, are detailed in Section 6, Table 7.

5.2.1 ESBMC Parameter Selection

Model-checking tools like ESBMC offer adjustable parameters that can optimize performance based on program complexity. Default parameters such as those used in competitions like SV-COMP, may not be suitable for all software types, potentially leading to fewer detected vulnerabilities. We conducted a detailed analysis to understand how various settings impact verification outcomes, particularly for programs generated by LLMs. The different switches are explained in Appendix Table 1.

Given our expectation that it generates more complex code, we have selected 1,000 samples generated by GPT-4, serving as the basis for selecting the ESBMC parameters for the entire dataset. For these samples, we tested multiple parameter configurations of ESBMC to determine which settings yielded the best results regarding runtime efficiency and vulnerability detection. We focus on two objectives. Firstly, to minimize verification unknown outcomes ($\mathcal{VU}$) through the t parameter and preferably completing the verification process; and secondly, to identify as many vulnerabilities as possible. Table 4 illustrates the verification outcomes of the 1,000 samples, demonstrating how various combinations of unwind (u) and time (t), alongside the utilization of k-induction, incremental-bmc, or falsification techniques, impact the results. Our analysis revealed that merely increasing the unwind parameter u while keeping a short timeout (e.g., 1 second) often leads to timeouts. For example, setting the unwind to 10 with a 1-second timeout resulted in most samples (684) falling into $\mathcal{VU}$. This limitation reflects the capability of the underlying architecture; in our case, the AMD Ryzen Threadripper PRO 3995WX with 32 CPU cores. A more powerful system could potentially handle more computations within the same timeframe.

A larger unwind parameter increases the detection of vulnerabilities in loops if adequate processing time is given. K-induction improves the detection rate for larger unwind settings, but the default k-bound is 1 unless adjusted. After extensive testing, we extended the timeout to 500 seconds and allowed unlimited k-steps, transitioning from bounded to *unbounded model* checking. This adjustment ensures that if the verification is completed within this time frame, we either identify a counterexample or confirm the absence of the examined vulnerabilities. This represents a significant change from our previous methodology in creating the FormAI-v1 dataset. We used the following ESBMC switches during our experiments, as depicted in Figure 7.

Note that `--overflow`, `--memory-leak-check`, and `--multi-property` are the same for the entire Table 4 experiments and were not changed. Using these parameters on our 1000 sample set, 416 files were deemed non-vulnerable, while 519 files

Sample verification command

```
esbmc <file.c> --overflow --memory-leak-check --k-induction
--timeout 500 --unlimited-k-steps --multi-property
--show-stacktrace --verbosity 6
```

Fig. 7: ESBMC command employed to verify each sample in the dataset.

Table 4: Results of classification for varying parameters based on a dataset of 1000 samples generated by GPT-4.

ESBMC Parameters					RESULTS					
u	time	k-ind	bmc	fls	Running time (m:s)	$ \phi $	$\mathcal{VS}$	$\mathcal{VF}$	$\mathcal{VU}$	$\mathcal{ER}$
X	300	✓	X	X	1698:53	1678	471	491	25	13
2	1000	X	X	X	1418:03	1638	505	407	70	18
3	1000	X	X	X	2100:36	1620	495	390	94	21
X	100	✓	X	X	653:05	1583	486	468	33	13
2	100	X	X	X	224:25	1580	496	393	96	15
1	1000	X	X	X	419:45	1529	538	428	21	13
X	30	✓	X	X	216:28	1513	494	448	45	13
X	30	X	✓	X	216:20	1511	494	448	45	13
X	30	X	X	✓	232:36	1511	494	448	45	13
2	30	X	X	X	99:05	1500	486	371	129	14
1	100	X	X	X	79:09	1465	536	421	30	13
X	10	✓	X	X	84:11	1432	500	430	57	13
3	100	X	X	X	344:01	1408	478	350	158	14
1	10	X	X	X	21:47	1351	527	399	61	13
2	10	X	X	X	47:48	1272	469	330	187	14
3	10	X	X	X	62:14	951	433	272	281	14
X	1	✓	X	X	13:23	941	474	336	177	13
X	1	X	X	✓	13:39	938	475	335	177	13
X	1	X	✓	X	13:34	936	476	334	177	13
1	1	X	X	X	7:41	911	487	323	177	13
2	1	X	X	X	10:30	559	404	205	377	14
10	1	X	X	X	14:14	158	224	79	684	13
X	10	X	X	X	152:41	69	75	25	887	13

Legend:

✓: Enabled; X: Not set; $|\phi|$: Number of Vulnerabilities; $\mathcal{VS}$: Verification Success;
 $\mathcal{VF}$: Verification Failed: vulnerabilities detected; $\mathcal{VU}$: Verification Unknown (Timeout)
 $\mathcal{ER}$: Error: Verification process resulted in an error; **k-ind**: k-induction;
bmc: incremental-bmc; **fls**: falsification technique; **u**: unwind

were vulnerable. Among these 519 files, a total of 2116 unique vulnerabilities were detected. Considering the classification of 256,000 programs, the worst-case scenario is that every program from FormAI-v2 would utilize its allocated time, resulting in 500 seconds dedicated to verifying each sample. Using 32 CPU threads, the entire verification process on our experimental setup would take approximately 46,3 days in this worst-case scenario, calculated as $256,000 \times 500 / 60 / 60 / 24 / 32$.

It is important to notice that while the best-performing setting from Table 4 had a much lower time setting than the second best, more files utilized the maximum time frame, thus making the verification process longer.

6 Discussion

The following section summarizes our main results. First, we examine statistics on the entire dataset regarding overall verification results and vulnerability types. This is followed by evaluating each LLM and comparing their cyclomatic complexity, keyword

frequency and secure coding capabilities. In the original FormAI dataset, we only created 112,000 compilable C samples using GPT-3.5. Furthermore, the complexity of each program was not measured. This research closes this gap by comparing eight state-of-the-art LLMs and providing a vulnerability-labelled dataset to the research community, comprising 256,000 C programs.

6.1 Evaluation of the FormAI-v2 Dataset

We have examined over 21,994,613 lines of C code, with an average of 83.70 lines per sample. In total, we performed the verification process on 256,000 C program files, and our results for the entire dataset are shown in Table 5. The TOP 10 violations throughout the entire dataset are presented in Table 6.

Table 5: Summary of Verification Outcomes.

Category	Count	(%)
Total Files	265,000	100.00
Verification Success	23,036	8.69
Verification Unknown (Timeout)	66,899	25.24
Verification Failed	168,208	63.47
Error	6,857	2.59

Table 6: Total Violations Across All Categories.

Rank	Category	Violation Type	Count	Percentage
1	$\mathcal{DF}$	Dereference failure: NULL pointer	285,505	41.73%
2	$\mathcal{BO}$	Buffer overflow on <code>scanf</code>	181,122	26.47%
3	$\mathcal{DF}$	Dereference failure: invalid pointer	61,022	8.92%
4	$\mathcal{DF}$	Dereference failure: forgotten memory	21,222	3.10%
5	$\mathcal{DF}$	Dereference failure: array bounds violated	20,200	2.95%
6	$\mathcal{DF}$	Array bounds violated: upper bound	18,895	2.76%
7	$\mathcal{DF}$	Array bounds violated: lower bound	17,275	2.52%
8	$\mathcal{AO}$	Arithmetic overflow on sub	15,089	2.21%
9	$\mathcal{AO}$	Arithmetic overflow on add	12,552	1.83%
10	$\mathcal{AO}$	Arithmetic overflow on mul	10,088	1.47%

During the 500-second verification time-frame, ESBMC identified 168,208 unique programs with vulnerabilities. In contrast, only 23,036 programs, representing 8.69%, were verified as secure. Expanding computational resources may increase the number of programs uncovered from $\mathcal{VU}$, thereby potentially extending the $\mathcal{VF}$ category. These given results provide an even better lower bound compared to [9], on what percentage of LLM-generated code is vulnerable. The big picture is worse than simply saying 63.47%, as one file can contain more than one vulnerability. The total number of property violations detected by ESBMC is 684,227. A breakdown of the distribution of these different types of vulnerabilities is shown in Table 7.

Table 7: Detailed Categorisation of Vulnerabilities as Detected by ESBMC 7.5.

Category	Description	Count	Percentage
<i>DF</i>	Dereference failures:		
	- NULL pointer	285505	41.73%
	- Invalid pointer	61022	8.92%
	- Forgotten memory	21222	3.10%
	- Array bounds violated	20200	2.95%
	- Invalidated dynamic object	2904	0.42%
	- Access to object out of bounds	1898	0.28%
	- Accessed expired variable pointer	1091	0.16%
	- Write access to string constant	751	0.11%
	- Of non-dynamic memory	251	0.04%
	- Object accessed with incompatible base type	270	0.04%
	- Oversized field offset	141	0.02%
	- Data object accessed with code type	11	0.00%
<i>AO</i>	Arithmetic overflows:		
	- On sub	15089	2.21%
	- On add	12552	1.83%
	- On mul	10088	1.47%
	- IEEE mul	7545	1.10%
	- IEEE div	2755	0.40%
	- IEEE add	1729	0.25%
	- IEEE sub	1464	0.21%
	- On div	691	0.10%
	- On shl	669	0.10%
	- On modulus	220	0.03%
	- On neg	105	0.02%
<i>BO</i>	Buffer overflows:		
	- On <code>scanf</code>	181122	26.47%
	- On <code>fscanf</code>	6593	0.96%
	- On <code>sscanf</code>	2810	0.41%
<i>ABV</i>	Array bounds violations:		
	- Upper bound	18895	2.76%
	- Lower bound	17275	2.52%
	- VLA array size in bytes overflows address space size	4082	0.60%
<i>DBZ</i>	Division by zero	3592	0.52%
<i>MV</i>	Miscellaneous Vulnerabilities:		
	- The pointer to a file object must be a valid argument	1125	0.16%
	- Invalid Function argument issues	339	0.05%
	- Same object violation	137	0.02%
	- Operand of free must have zero pointer offset	84	0.01%

The most common type of vulnerability is related to “Dereference failures” accounting for 57.77% of the cases, predominantly due to NULL pointer issues. This category includes a variety of pointer-related issues such as invalid pointers, forgotten memory, and array bounds violations, among others. “Buffer overflows”, mainly triggered by the `scanf` function, comprise a significant 27.84% of the vulnerabilities. This highlights common issues in handling buffer sizes and input functions. “Arithmetic overflows” are also notable, covering various operations like subtraction, addition, multiplication, and division, indicating frequent issues in handling numeric calculations without adequate checks. The table further lists “Array bounds violations” and “Division by zero” as common issues, illustrating challenges in correctly managing

arrays and arithmetic operations. A smaller portion of the table covers "Miscellaneous Vulnerabilities," which includes a variety of less frequent but notable issues such as invalid file object pointers and operand violations in memory deallocation. Overall, the data emphasizes the need for robust handling of pointers, buffers, and numeric operations within the source code to mitigate the risk of vulnerabilities.

6.2 CWE Classification

Next, we link the vulnerabilities to Common Weakness Enumeration (CWE) identifiers by manually assigning the appropriate CWEs after reviewing source codes associated with each vulnerability group. The multifaceted nature of software flaws often results in a single vulnerability associated with multiple CWE identifiers. Table 9 shows the categorization of the most prevalent vulnerabilities and the associated 39 unique CWEs we identified across these categories. The "*Miscellaneous Vulnerabilities*" (*MV*) category includes Assertion failure, Same object violation, Operand of free must have zero pointer offset, function call: not enough arguments, and several types of dereference failure issues. For these categories we did not individually assign CWE identifiers, as in general they constitute a very small percentage of the dataset. The small sample size also prevents machine learning algorithms to efficiently recognise and learn these patterns. While in the .json file they are still marked as vulnerable and count for statistical purposes, we aggregated these samples under the other category so they can be conveniently excluded from the training data.

In total, 42 unique CWE were identified in the dataset. From MITRE's Top 25 Most Dangerous Software Weaknesses for 2023 list, six is present in our list as shown in Table 8. The remaining CWEs in the top 25 list are related to web vulnerabilities

Table 8: CWEs from 2023's MITRE Top 25.

Rank	CWE Description
1	CWE-787: Out-of-bounds Write
4	CWE-416: Use After Free
6	CWE-20: Improper Input Validation
7	CWE-125: Out-of-bounds Read
12	CWE-476: NULL Pointer Dereference
14	CWE-190: Integer Overflow or Wraparound

like SQL injection, XSS, and authentication, which are irrelevant to our C language samples. It is vital to emphasize that, in our 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. In the datasets ReVeal, BigVul, and Diversevul, a function is vulnerable if the vulnerability-fixing commit changes it, while in Juliet, a single vulnerability is introduced for each program. In FormAI, a single file often contains multiple vulnerabilities. As noted, a single vulnerability can be associated with multiple CWEs. Additionally, multiple CWEs can be required as a prerequisite for a vulnerability to manifest. As an example, "*CWE-120: Buffer Copy without Checking Size of Input (Classic Buffer*

Table 9: Detailed Categorisation of Vulnerabilities as Detected by ESBMC 7.5.

Description	CWE	Associated CWEs
Dereference failures:		
- NULL pointer	CWE-476	CWE-690, CWE-391
- Invalid pointer	CWE-822	CWE-119, CWE-787, CWE-822
- Forgotten memory	CWE-825	CWE-401, CWE-404, CWE-459
- Array bounds violated	CWE-125	CWE-119, CWE-787
- Invalidated dynamic object	CWE-824	CWE-416, CWE-415
- Access to object out of bounds	CWE-125	CWE-119, CWE-787
- Accessed expired variable pointer	CWE-416	CWE-825
- Write access to string constant	CWE-843	CWE-758
- Of non-dynamic memory	CWE-590	CWE-415, CWE-415, CWE-762
- Object accessed with incompatible base type	CWE-843	CWE-119
- Oversized field offset	CWE-787	CWE-119, CWE-125, CWE-823
- Data object accessed with code type	CWE-843	CWE-686, CWE-704
Arithmetic overflows:		
- On sub	CWE-191	CWE-20, CWE-190, CWE-192
- On add	CWE-190	CWE-20, CWE-191, CWE-192
- On mul	CWE-190	CWE-20, CWE-191, CWE-192
- Floating-point ieee.mul	CWE-190	CWE-681
- Floating-point ieee.div	CWE-682	CWE-369, CWE-681
- Floating-point ieee.add	CWE-190	CWE-681
- Floating-point ieee.sub	CWE-190	CWE-681
- On div	CWE-190	CWE-20, CWE-369
- On shl	CWE-190	CWE-192
- On modulus	CWE-190	CWE-20, CWE-191
- On neg	CWE-191	CWE-190, CWE-192
Buffer overflows:		
- On scanf	CWE-120	{CWE-20, CWE-121, CWE-122}
- On fscanf	CWE-120	CWE-129, CWE-131, CWE-628
- On sscanf	CWE-120	CWE-676, CWE-680, CWE-787}
Array bounds violations:		
- lower bound	CWE-129	{CWE-119, CWE-125, CWE-129}
- upper bound	CWE-788	CWE-131, CWE-193, CWE-787}
- VLA array size in bytes overflows	CWE-190	CWE-131, CWE-680
Division by zero	CWE-369	CWE-691
Miscellaneous Vulnerabilities:		
- The pointer to a file must be valid	CWE-476	CWE-690, CWE-459
- Same object violation	CWE-628	CWE-843, CWE-668
- Operand of free must have zero pointer offset	CWE-761	CWE-415, CWE-590

Overflow)”, can happen as a result of “*CWE-676: Use of Potentially Dangerous Function*”, which can be the *scanf* function. If this is combined with “*CWE-20: Improper Input Validation*”, it can result in “*CWE-787: Out-of-bounds Write*”.

Labelling the vulnerable function name, line number, and vulnerability type identified by the ESBMC module provides granular information that can benefit machine learning training. This level of detail can allow models to discern patterns and correlations with higher precision, thereby improving vulnerability prediction and detection capabilities. Since our programs exhibit numerous vulnerabilities, including multiple

occurrences of the same type, categorizing each solely into one CWE group, as seen with Juliet, would be sub-optimal for training purposes. This method fails to communicate crucial details about the vulnerabilities. For instance, both "Arithmetic overflow on add" and "Arithmetic overflow on div" are assigned the same primary CWE, they manifest differently in the source code. Therefore, merely labelling them with CWEs does not offer the same level of granularity and makes the dataset less suitable for ML. While other datasets focus more on CWEs related to vulnerabilities that could potentially be exploited, ESBMC also detects issues related to software safety.

6.3 LLMs comparison

6.3.1 Keyword Frequency

We employ a token-based keyword-counting mechanism to extract the cardinality of the 32 C keywords from each LLM-generated subset, as shown in Figure 8 and 9. The frequency is normalized to show the occurrence of keywords per million lines of code for each model. It is also important to note that while the distribution is similar, the

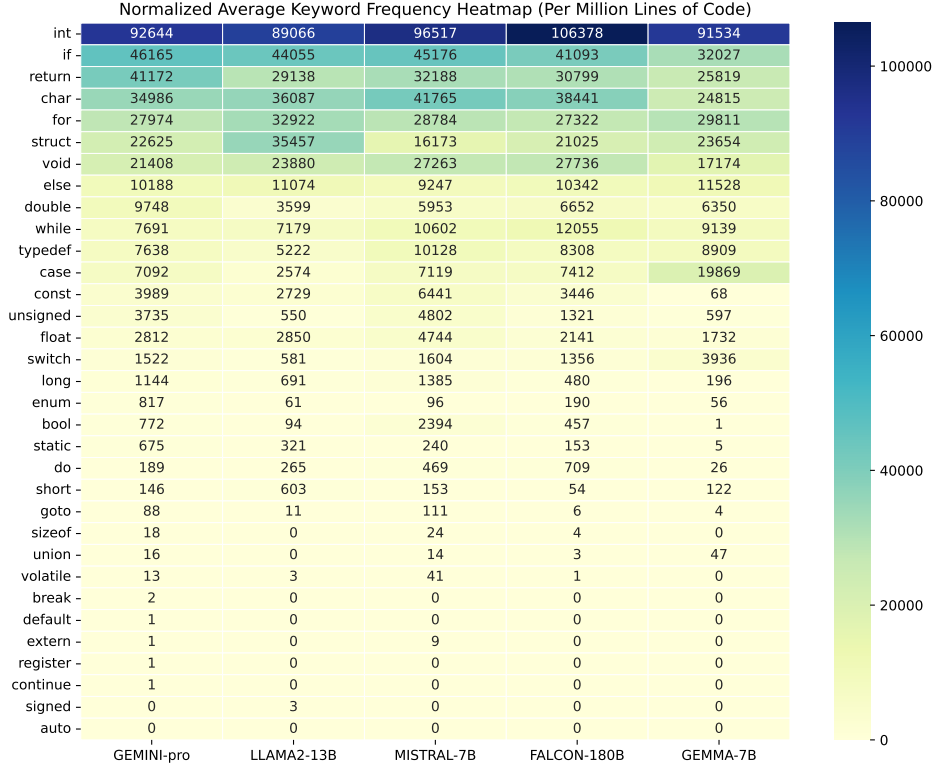


Fig. 8: 32 C keyword distribution - PART I.

slight fluctuations indicate that the different models handle statements, expressions, and variables differently. This observation is supported by examining the cyclomatic complexity statistics of each model.

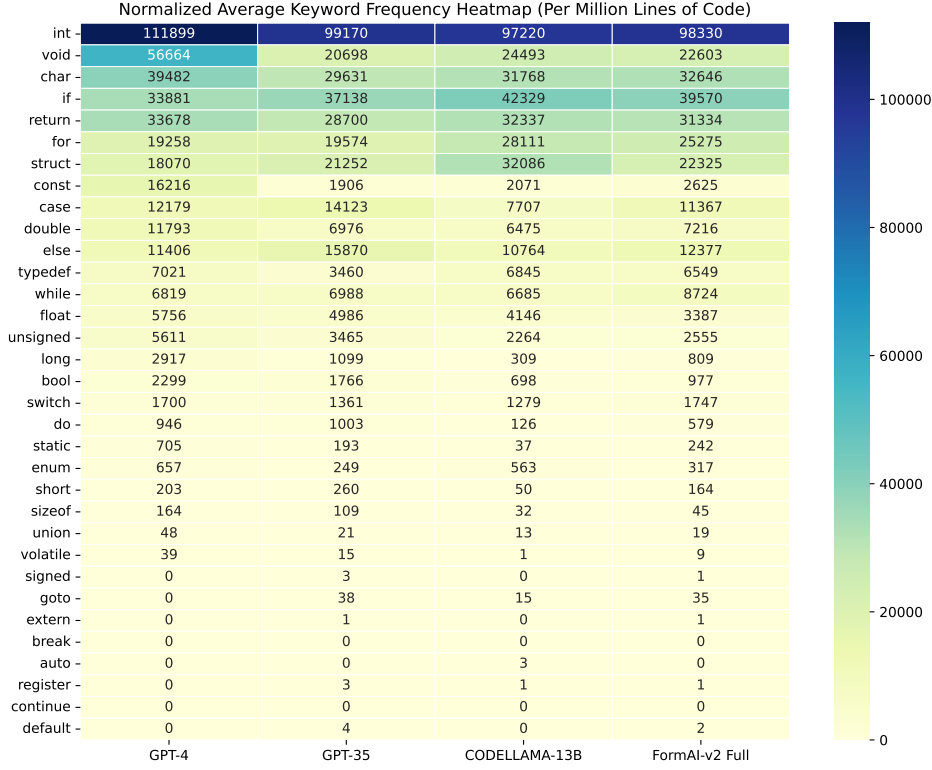


Fig. 9: 32 C keyword distribution - PART II.

6.4 Cyclomatic Complexity and General Statistics

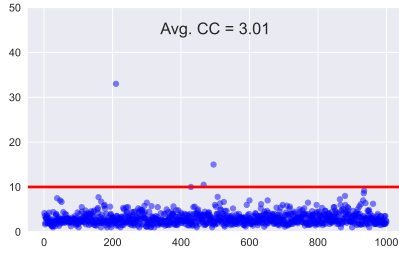
To comprehend the meaning and limitations of cyclomatic complexity, it's essential to delve into the industry-standard guidelines established by Microsoft and NIST [109, 110]. Simply put, *“higher numbers are bad and lower numbers are good”*. Cyclomatic complexity serves as a metric for assessing the difficulty involved in testing, maintaining, or troubleshooting code, as well as predicting the likelihood of errors. It is calculated by counting decision points in the source code. NIST defines it as *“the amount of decision logic in a source code function”* and sets the recommended maximum as 10 [110]. They note that a complexity threshold above 10 is only acceptable for bigger projects where experienced engineers can manage the extra testing

needed. Since the programs are small standalone units, a cyclomatic complexity number surpassing 10 is deemed unacceptable, and even the suggested threshold may be considered excessive.

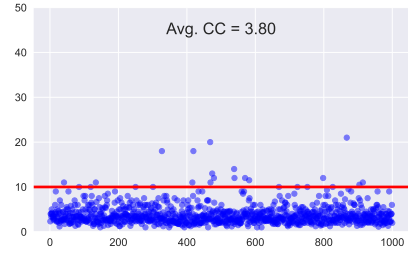
Table 10: Statistics of Cyclomatic Complexity, ordered by Average CC.

Model	Total Samples	Avg Lines	Min CC	Max CC	Avg CC	Median CC	Std. Dev. CC
GPT-4	1000	104.54	1	33	3.01	2.71	1.60
Mistral-7B	10000	75.02	1	33.0	3.85	3.33	2.22
Falcon-180B	72000	72.06	1	46	4.34	3.33	3.10
Llama2-13B	5000	73.35	1	34.0	4.29	3.50	2.77
Gemini-Pro	40000	98.25	0	154	4.52	3.25	4.05
Codellama-13B	12000	83.30	1	94	4.63	3.20	4.64
Gemma-7B	47000	66.53	1	109	5.03	3.33	5.23
GPT-3.5	78000	96.52	0	240.0	6.10	4.43	5.68

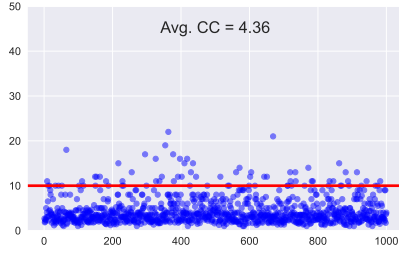
We compute each model’s cyclomatic complexity (CC) number by randomly selecting 1000 samples, as illustrated in Figure 10. The results unveil both expected and unexpected findings. While the exact parameter size of GPT-4 remains unknown, this model consistently generates the highest-quality code in terms of CC. Moreover, GPT-4 demonstrates remarkable performance on various benchmarks, such as HumanEval, indicating its proficiency in efficiently solving tasks without generating overly complex code. However, contrary to the expectations, the analysis of Table 10 reveals that GPT-4 does not necessarily produce shorter or simpler code. It tends to generate the longest C programs and exhibits the highest verification unknown score, implying that the verification process for ESBMC takes longer in the case of GPT-4 samples. Despite this, it manages to keep the CC number low. Among the examined LLMs, GPT-3.5 and Gemma7B demonstrate the poorest performance regarding cyclomatic complexity (CC). Interestingly, despite being a smaller model, Mistral7B closely follows GPT-4 in code quality. GPT-4, Mistral-7B, Falcon-180B, and Llama2-13B consistently exhibit low CC numbers across all samples, positioning them as the top performers regarding average CC. Conversely, other models display high maximum CC values, with corresponding high standard deviations, indicating less consistent performance. It is important to note, that due to the different total sample size, certain metrics like Max CC should be treated with caution. The more samples are generated the higher the chance for programs to appear with high CC numbers. Still, these metrics can serve as a good indicator, as for example Falcon-180B never managed to produce a program with a higher CC number than 46, despite the sample size for this model was the second largest. We aim to expand the dataset and equalise the sample size across the various models, to be able to make an even comparison. Despite this, such metrics are a useful indicator of performance.



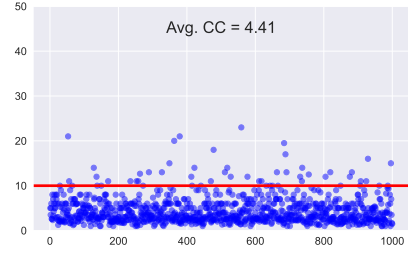
(a) GPT4.



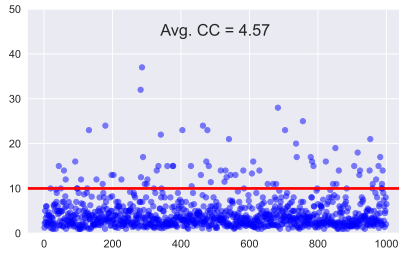
(b) MISTRAL-7B.



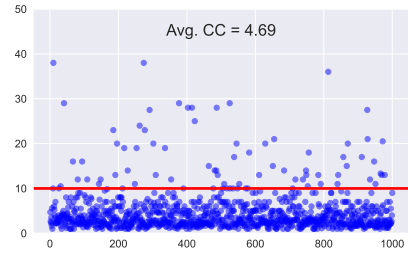
(c) FALCON-180B.



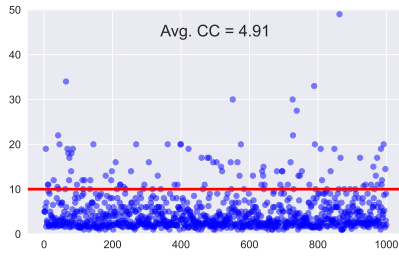
(d) LLAMA2-13B.



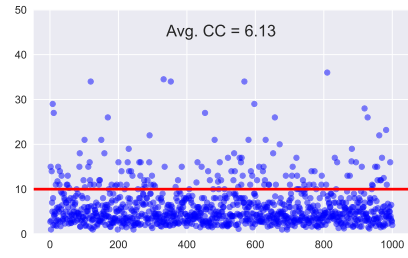
(e) GEMINI-pro.



(f) CODELLAMA-13B.



(g) GEMMA-7B.



(h) GPT-3.5.

Fig. 10: Cyclomatic complexity of different LLMs on 1000 randomly selected samples.

6.5 Vulnerability Ranking

The Tables 11, 12, 13, and 14 provide an overview of the top 10 vulnerabilities produced by each model. Important to note that since the number of samples created by each model is different, the individual vulnerability counts do not serve as a basis for comparison.

Buffer Overflow on `scanf` ranks among the top three vulnerabilities identified across every LLM model. The functions `scanf`, `fscanf`, and `sscanf` do not limit the input size to the size of their respective buffers. This oversight leads to a risk of buffer overflow, which could enable an attacker to execute arbitrary code or cause a crash. As earlier noted, this issue can be associated with several CWEs, including the CWE-676, CWE-20, and CWE-787. While buffer overflow is a type of out-of-bounds write, CWE-787 encompasses a wider array of vulnerabilities, where CWE-120 is focused specifically on the classic scenario of buffer overflow resulting from not checking input sizes during buffer copy operations. Note that this analysis does not provide an accurate count of every identified CWE, but rather on the vulnerabilities verifiably identified by ESBMC.

The results indicate that LLMs consistently produce these errors in a zero-shot setting. While more complex issues such as arithmetic overflows and array bounds violations require a deeper understanding of the programming scenario, issues related to `scanf` would be comparatively simpler to avoid. Nonetheless, every model tested consistently exhibits buffer overflow errors on `scanf`, although the frequency varies.

Table 11: Top 10 Vulnerabilities in GPT-3.5 and FALCON-180B.

Rank	Category	Violation Type	Count	Percentage
GPT-3.5				
1	<i>BO</i>	Buffer overflow on <code>scanf</code>	84,363	38.05%
2	<i>DF</i>	Dereference failure: NULL pointer	58,889	26.56%
3	<i>DF</i>	Dereference failure: invalid pointer	20,887	9.42%
4	<i>DF</i>	Dereference failure: forgotten memory	4,788	2.16%
5	<i>DF</i>	Array bounds violated: lower bound	7,836	3.53%
6	<i>DF</i>	Array bounds violated: upper bound	7,773	3.51%
7	<i>AO</i>	Arithmetic overflow on sub	6,422	2.90%
8	<i>DF</i>	Dereference failure: array bounds violated	6,701	3.02%
9	<i>AO</i>	Arithmetic overflow on add	5,169	2.33%
10	<i>AO</i>	Arithmetic overflow on mul	4,266	1.92%
FALCON-180B				
1	<i>BO</i>	Buffer overflow on <code>scanf</code>	52,811	34.99%
2	<i>DF</i>	Dereference failure: NULL pointer	44,957	29.79%
3	<i>DF</i>	Dereference failure: invalid pointer	15,822	10.48%
4	<i>DF</i>	Dereference failure: forgotten memory	5,942	3.94%
5	<i>DF</i>	Dereference failure: array bounds violated	4,675	3.10%
6	<i>DF</i>	Array bounds violated: upper bound	4,266	2.83%
7	<i>AO</i>	Arithmetic overflow on sub	3,697	2.45%
8	<i>DF</i>	Array bounds violated: lower bound	3,566	2.36%
9	<i>AO</i>	Arithmetic overflow on add	2,880	1.91%
10	<i>BO</i>	Buffer overflow on <code>fscanf</code>	2,555	1.69%

Table 12: Top 10 Vulnerabilities in LLAMA2-13B, and GEMMA-7B.

Rank	Category	Violation Type	Count	Percentage
llama2-13B				
1	$\mathcal{DF}$	Dereference failure: NULL pointer	4,649	54.12%
2	$\mathcal{BO}$	Buffer overflow on <code>scanf</code>	814	9.48%
3	$\mathcal{DF}$	Dereference failure: invalid pointer	811	9.44%
4	$\mathcal{DF}$	Dereference failure: array bounds violated	435	5.06%
5	$\mathcal{DF}$	Dereference failure: forgotten memory	389	4.53%
6	$\mathcal{AO}$	Arithmetic overflow on add	236	2.75%
7	$\mathcal{AO}$	Arithmetic overflow on mul	179	2.08%
8	$\mathcal{DF}$	Array bounds violated: upper bound	176	2.05%
9	$\mathcal{AO}$	Arithmetic overflow on sub	152	1.77%
10	$\mathcal{BO}$	Division by zero	125	1.46%
gemma7b				
1	$\mathcal{DF}$	Dereference failure: NULL pointer	89,888	65.76%
2	$\mathcal{BO}$	Buffer overflow on <code>scanf</code>	19,097	13.97%
3	$\mathcal{DF}$	Dereference failure: forgotten memory	5,491	4.02%
4	$\mathcal{DF}$	Dereference failure: invalid pointer	3,929	2.87%
5	$\mathcal{DF}$	Array bounds violated: upper bound	3,408	2.49%
6	$\mathcal{DF}$	Array bounds violated: lower bound	2,828	2.07%
7	$\mathcal{AO}$	Arithmetic overflow on sub	2,094	1.53%
8	$\mathcal{DF}$	Dereference failure: array bounds violated	2,060	1.51%
9	$\mathcal{AO}$	Arithmetic overflow on floating-point <code>ieee_mul</code>	1,458	1.07%
10	$\mathcal{AO}$	Arithmetic overflow on add	1,415	1.04%

Dereference failure-related issues were collectively the most numerous category for every LLM, and NULL pointer dereference consistently ranks first or second in every model. The generated code often includes pointers, but these models may not always correctly manage pointer operations or accurately simulate real-world applications’ complex memory management behaviors, leading to frequent dereference issues. However, LLMs, which operate primarily on pattern recognition without understanding underlying principles, often generate code with flawed pointer handling.

In addition, the datasets they were used for the training of these models often include a wide variety of pointer usage examples, which may not always adhere to safe programming standards. This, coupled with the complexities of dynamic memory management in languages like C, poses challenges for LLMs in consistently generating secure code. The recurrent issues with pointer dereferencing underscore the risks associated with deploying LLM-generated code in critical systems where security and reliability are paramount.

To mitigate these risks, it is crucial to develop enhanced training methodologies focused on robust memory handling. It is also paramount to implement advanced code analysis tools and frameworks. These tools will help detect and rectify vulnerabilities before deployment, thereby improving the security and reliability of LLM-generated code for real-world applications.

While dereference failures and buffer overflows pose risks, their severity and frequency vary significantly. For instance, GEMMA-7B shows a disproportionately high incidence of NULL pointer dereference failures at 65.76%, suggesting specific

Table 13: Top 10 Vulnerabilities in CODELLAMA-13B, and GEMINI-pro 1.0.

Rank	Category	Violation Type	Count	Percentage
CODELLAMA-13B				
1	$\mathcal{DF}$	Dereference failure: NULL pointer	11,741	45.22%
2	$\mathcal{BO}$	Buffer overflow on <code>scanf</code>	5,162	19.88%
3	$\mathcal{DF}$	Dereference failure: invalid pointer	3,528	13.59%
4	$\mathcal{DF}$	Dereference failure: array bounds violated	880	3.39%
5	$\mathcal{DF}$	Dereference failure: forgotten memory	711	2.74%
6	$\mathcal{DF}$	Array bounds violated: upper bound	659	2.54%
7	$\mathcal{AO}$	Arithmetic overflow on add	512	1.97%
8	$\mathcal{DF}$	Array bounds violated: lower bound	465	1.79%
9	$\mathcal{AO}$	Arithmetic overflow on mul	454	1.75%
10	$\mathcal{AO}$	Arithmetic overflow on sub	373	1.44%
GEMINI-pro 1.0				
1	$\mathcal{DF}$	Dereference failure: NULL pointer	68,570	57.27%
2	$\mathcal{DF}$	Dereference failure: invalid pointer	13,361	11.16%
3	$\mathcal{BO}$	Buffer overflow on <code>scanf</code>	13,106	10.95%
4	$\mathcal{DF}$	Dereference failure: array bounds violated	4,651	3.88%
5	$\mathcal{DF}$	Dereference failure: forgotten memory	3,433	2.87%
6	$\mathcal{AO}$	Arithmetic overflow on mul	2,450	2.05%
7	$\mathcal{DF}$	Array bounds violated: upper bound	2,154	1.80%
8	$\mathcal{AO}$	Arithmetic overflow on add	1,882	1.57%
9	$\mathcal{DF}$	Array bounds violated: lower bound	1,840	1.54%
10	$\mathcal{AO}$	Arithmetic overflow on sub	1,815	1.52%

Table 14: Top 10 Vulnerabilities in MISTRAL-7B and GPT-4.

Rank	Category	Violation Type	Count	Percentage
MISTRAL-7B				
1	$\mathcal{DF}$	Dereference failure: NULL pointer	6,318	33.19%
2	$\mathcal{BO}$	Buffer overflow on <code>scanf</code>	5,169	27.15%
3	$\mathcal{DF}$	Dereference failure: invalid pointer	2,474	13.00%
4	$\mathcal{DF}$	Dereference failure: array bounds violated	747	3.92%
5	$\mathcal{DF}$	Array bounds violated: lower bound	621	3.26%
6	$\mathcal{AO}$	Arithmetic overflow on sub	475	2.50%
7	$\mathcal{DF}$	Array bounds violated: upper bound	452	2.37%
8	$\mathcal{DF}$	Dereference failure: forgotten memory	413	2.17%
9	$\mathcal{AO}$	Arithmetic overflow on add	402	2.11%
10	$\mathcal{BO}$	Buffer overflow on <code>sscanf</code>	393	2.06%
GPT-4				
1	$\mathcal{BO}$	Buffer overflow on <code>scanf</code>	600	34.70%
2	$\mathcal{DF}$	Dereference failure: NULL pointer	493	28.51%
3	$\mathcal{DF}$	Dereference failure: invalid pointer	210	12.15%
4	$\mathcal{AO}$	Arithmetic overflow on sub	61	3.53%
5	$\mathcal{AO}$	Arithmetic overflow on floating-point <code>ieee_mul</code>	57	3.30%
6	$\mathcal{AO}$	Arithmetic overflow on add	56	3.24%
7	$\mathcal{DF}$	Dereference failure: forgotten memory	55	3.18%
8	$\mathcal{DF}$	Dereference failure: array bounds violated	51	2.95%
9	$\mathcal{DF}$	Array bounds violated: lower bound	34	1.97%
10	$\mathcal{AO}$	Arithmetic overflow on mul	25	1.45%

weaknesses in its memory management capabilities. Arithmetic overflows appear consistently for all models, mostly populating the bottom 5 ranks. They vary in terms of the specific operations (add, sub, mul) affected, indicating differing arithmetic handling across the models. Interestingly, lallama2-13B stands out as the only model below 10% on *scanf()* related violations, with Gemini-pro following suit around 11%. However, similar to GEMMA-7B, both models exhibit a disproportionately high incidence of NULL pointer dereference failures.

The consistent appearance of certain errors across different models emphasizes the need for comprehensive testing and validation frameworks to address these recurrent issues before deployment. While all models exhibit the same vulnerabilities, significant differences in the frequency and type of other vulnerabilities, such as arithmetic overflows, suggest that model-specific optimizations and enhancements are necessary.

6.6 LLM Ranking: Which model is the most secure

To compare which model is the “worst” or the “best” when it comes to secure coding—and to do this as fairly as possible—we will investigate several metrics, such as the ratio of verification results, average property violation per file, and average property violation per line of code.

Table 15: Verification Results Summary, Sorted by Average Property Violation per Line.

Category	Avg Prop. Viol. per Line	Rank	$\mathcal{VS}$	Rank	$\mathcal{VF}$	$\mathcal{VU}$ (Time out)	Avg Prop. Viol. per File
GPT4	0.0165	3	11.40%	2	50.80%	36.50%	3.40
Llama2-13B	0.0234	2	13.36%	1	47.50%	35.58%	3.62
Mistral-7B	0.0254	7	7.11%	4	62.00%	28.19%	3.07
Codellama-13B	0.0260	1	15.24%	3	52.34%	30.12%	4.13
Falcon180B	0.0291	8	6.49%	5	62.10%	28.61%	3.38
GPT35	0.0295	6	7.57%	7	64.25%	26.65%	4.42
Gemini_Pro	0.0305	5	9.56%	6	63.63%	24.39%	4.70
Gemma7B	0.0437	4	11.29%	8	69.29%	15.28%	4.20

Legend:

$\mathcal{VS}$: Verification Success; $\mathcal{VF}$: Verification Failed; $\mathcal{VU}$: Verification Unknown (Timeout)

The results indicate that there is no clear winner. Mistral-7B, despite having the fewest property violations per file, writes shorter code, reducing its likelihood of coding errors. However, this model also performs poorly in the $\mathcal{VS}$ metric, with only 7.11% of its samples categorized as being free of vulnerabilities. Codellama-13B achieved the highest $\mathcal{VS}$ rate, followed by Llama2-13B, and their $\mathcal{VF}$ ratio is ranking is third and second respectively, which is a good result for the LLama family. Still, it is best to remember that nearly half of their samples had vulnerabilities. Moreover, their $\mathcal{VU}$ is fairly high at 30% and 35%, which means that with further verification, there is still a chance that other models will take the lead.

GPT-4 outperforms GPT-3.5 while showing the highest $\mathcal{VU}$ percentage under current ESBMC settings, indicating its ability to produce more complex and longer outputs. It is important to note that this complexity is not reflected by the CC number as discussed earlier, which confirms the criticism towards Cyclomatic Complexity by practitioners. While GPT-4 ranks third in $\mathcal{VS}$ and second in $\mathcal{VF}$, it finishes first with an average property violation per line. This might be the fairest way to compare models, as the more lines, the more chances to have vulnerabilities, while this metric doesn't punish models producing shorter codes.

There is no definitive winner in this analysis; however, Gemma-7B, Gemini-Pro, and GPT-3.5—with the current verification settings—has the highest $\mathcal{VF}$ ratios and highest average property violation both per line and file. This unveils an interesting and unexpected finding: empirically, a slight correlation between having high cyclomatic complexity and vulnerable coding patterns can be observed. As Table 10 shows, these three models had the worst CC statistics. The correlation is, however, not clear, and there are certainly other factors. For example, CodeLlama-13B, with a maximum CC number of 94, did not show poor performance in the verification results.

It is important to underline that it might be tempting to speculate on a winner. However, having such a high verification failed ratio is unacceptable from a SE perspective for any model. All models surpassed the $\mathcal{VF}$ threshold of 47%, indicating that nearly half or more of the generated programs are vulnerable. The conclusions of this analysis must be clear: **Using code generated by the state-of-the-art Large Language Models, without any additional framework for validation and vulnerability analysis, carries severe risks.** While LLMs can be useful for automating simple tasks and scripting, directly including **such codes in production software without oversight from experienced software engineers is irresponsible and should be avoided.**

7 Limitations and Future Research

7.1 Future Research Directions

The dataset containing all 265,000 C program files, along with the .json and .csv files are published on GitHub⁶. The dataset is specifically prepared to support Machine Learning and fine-tuning. The absence of false positives makes the dataset suitable for benchmarking the effectiveness of various static and dynamic analysis tools and ML-based classifiers. Training on data that is void of false negatives is really important, so we indicate in the .json file for each file whether the verification process has finished, as such files do not contain the vulnerabilities detectable by ESBMC. The diverse structure of the C programs generated in the FormAI-v2 dataset made it excellent for an unexpected use case: fuzzing different applications. We uncovered and reported over thirteen bugs while executing ESBMC using the FormAI dataset. 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 formulas. Additionally, we identified bugs in the CBMC [111] and the Clang compiler, which failed to

⁶<https://github.com/FormAI-Dataset>

compile several programs, while GNU C had no issue. We promptly communicated these findings to the respective developers. The FormAI-v2 dataset aims to be a useful resource for training machine learning algorithms to possess the capabilities of the ESBMC module.

Our results give rise to several interesting research directions:

- It would be important to investigate why programs under “Verification Successful” are void of vulnerabilities. Is it because of better coding practices or simply because for example they don’t take user input, thereby avoiding buffer overflows?
- What is the right path towards LLMs producing secure code: Re-training models on better data, fine-tuning, or using current models in various few-shot frameworks with better prompting?
- Since several codes contain multiple vulnerabilities, this dataset is ideal for benchmarking and testing various vulnerability detection tools.
- As our motivation section showcased, GPT-4 did not excel at avoiding and fixing the vulnerability in the example. How do different LLMs compare in understanding, correctly fixing, and detecting coding errors?
- We aim to further grow the FormAI dataset including more state-of-the-art models, and also by increasing the number of samples for each LLM to have an overall larger dataset.
- How do different programming Tasks or Styles impact vulnerable coding patterns? Are there tasks that LLMs consistently mess up?

While we can partially address the last question, noting the use of insecure functions and poor input sanitization in handling user inputs, exploring this issue across various domains, such as networking or cryptography would be beneficial.

7.2 Limitations and Threats to Validity

With a larger timeout setting, ESBMC might find slightly more vulnerabilities in a given program. Whether the verifier can finish the process under a given timeout is up to the available computational capacity. The same parameter setting can yield a higher or lower detection rate on different architectures. To find all errors detectable by ESBMC, unwind must be set to infinite, and ESBMC must complete the verification process. 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. As the “Verification Unknown” category still contains samples for every model, the current results are strongly bound to the percentage of vulnerable files LLMs produce.

While ESBMC is a robust tool for detecting many types of errors in C, it is not currently suited to detect design flaws, semantic errors, or performance issues. As such, more vulnerabilities might be present besides the detected ones in the code. Thus we recommend that the training and fine-tuning applications to be restricted to the vulnerabilities detectable by ESBMC on this dataset.

8 Conclusions

In this research, we analyzed eight state-of-the-art Large Language Models to assess their likelihood of introducing vulnerabilities during neutral prompt based code generation. The models included in our analysis were Mistral-7B, Falcon-180B, GPT-4, Llama2-13B, Codellama-13B, Gemma-7B, GPT-3.5, and Gemini-Pro. We employed a zero-shot prompting method to encompass numerous programming scenarios for C code generation. These programs constitute the FormAI-v2 dataset, containing 256,000 independent compilable C programs.

We used the Efficient SMT-based Bounded Model Checker (ESBMC), a state-of-the-art formal verification tool, to identify vulnerabilities. Each program was given a verification period of 500 seconds with the unwinding parameter set to infinite, uncovering a total of 684,227 vulnerabilities. Overall 67% of the codes were vulnerable. Detailed labelling of each sample—including filename, type of vulnerability, function name, error type, and source code—is documented in a .json file, as detailed in Appendix Fig. 1, to facilitate the dataset’s use in machine learning applications.

Additionally, the FormAI-v2 dataset proved instrumental for fuzzing various applications, leading to the identification of multiple bugs in ESBMC, CBMC, and the Clang compiler. We have identified 42 distinct CWE identifiers, six of which are featured on MITRE’s Top 25 list for 2023. Notably, "CWE-787 Out-of-bounds Write," the top vulnerability on the list, was confirmed in at least 181,122 cases. These findings provide clear answers to our research questions:

- **RQ1:** Does the security of LLM-generated code vary significantly between different models?
 - **Answer:** Codellama-13B, Llama-13B, and GPT-4 performs slightly better, but all examined models notoriously introduce vulnerabilities into the C code they generate at unacceptable rates.
-
- **RQ2:** What are the most typical vulnerabilities introduced by different LLM models during code generation (focusing on C)?
 - **Answer:** Dereference failures and buffer overflow issues are the most prevalent vulnerabilities across all models, ranking arithmetic overflow as the third most common type. No model is completely free from any of the examined vulnerabilities; the variations lie in the frequency of occurrence.

While the literature reveals significant variations in the ability of these models to solve tasks, this is not mirrored in their susceptibility to produce vulnerabilities in source code. Our findings conclusively show that despite differences among the examined models, in terms of generating code, they all consistently introduce severe vulnerabilities when prompted with simple coding tasks. Our study indicates that despite the impressive capabilities of Large Language Models in code generation, employing their output in production requires meticulous risk assessment. Relying on these models without expert oversight in a production context is inadvisable.

Appendix

Switch	Description
-timeout <seconds>	This switch establishes a time limit in seconds for the analysis. If the verification process exceeds this duration, ESBMC will automatically halt the analysis.
-k-induction <number>	This enables k-induction proof, a formal verification technique to prove properties of loops.
-unlimited-k-steps	Allows the k-induction process to proceed without a pre-defined limit on the number of iterations, aiming for more thorough verification.
-unwind <number>	Sets the unwind limit for loops and recursion. This specifies how many times loops or recursive functions should be unwound during the analysis.
-no-unwinding-assertions	Disables assertions that check whether the loop or recursion unwinding was sufficient, useful when the unwind bound is considered adequate.
-falsification	ESBMC prioritizes finding violations of the properties, and the tool terminates as soon as it discovers a counterexample.
-overflow	Enables checking for arithmetic overflows within the program, helping detect when operations exceed the data type limits.
-memory-leak-check	Activates the detection of memory leaks, ensuring that all allocated memory is properly freed before program termination.
-multi-property	Enables the verification of multiple properties simultaneously, optimizing the analysis by checking all specified properties in a single run.
-show-stacktrace	When an error occurs, this option provides a stack trace, useful for debugging by showing the sequence of function calls leading to the error.
-verbosity 6	Sets the verbosity level of the output to detailed, offering a deeper insight into the tool's operations and the verification process.
-incremental-bmc	Enhances standard BMC by gradually increasing the bound depth, beginning at a low level and continuing until a bug is detected, a specified depth is reached, or resources are exhausted.

Table 1: Description of Investigated ESBMC Switches.

JSON Labels for a GPT-3.5 Generated Sample

```
{
  "category": "VULNERABLE",
  "file_name": "gpt35-58832.c",
  "verification_finished": "yes",
  "vulnerable_line": 42,
  "column": 9,
  "function": "sanitize_url",
  "violated_property": "file gpt35-58832.c line 42 column 9 function
    sanitize_url",
  "stack_trace": "c:@F@sanitize_url at file gpt35-58832.c line 15
    column 5 function main
    c:@F@main
    array bounds violated: array 'temp' lower bound
    (signed long int)(return_value_strlen3 - 1) >= 0",
  "error_type": "array bounds violated: lower bound",
  "code_snippet": "temp[j] = '\\0'; // null-terminate the temporary
    string
    // Remove any trailing slash at the end of the URL
    if (temp[strlen(temp) - 1] == '/')
    {
      temp[strlen(temp) - 1] = '\\0';
    }
    // Convert the URL to lowercase
    for (i = 0; i < strlen(temp); i++)
    {",
  "source_code": "code placeholder",
  "num_lines": 53,
  "cyclomatic_complexity": 3.0
}
```

Fig. 1: Example from the JSON file, demonstrating the labelling process. The source code object has been omitted.

Data Availability Statements

In this study, a total of 265,000 C samples were generated and examined. The findings and all the generated C samples are available for access and download from the project's website at <https://github.com/FormAI-Dataset>.

Conflicts of interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

- [1] Wang, J., Huang, Y., Chen, C., Liu, Z., Wang, S., Wang, Q.: Software testing with large language models: Survey, landscape, and vision. *IEEE Transactions on Software Engineering* (2024)
- [2] Xu, F.F., Alon, U., Neubig, G., Hellendoorn, V.J.: A systematic evaluation of large language models of code. In: *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, pp. 1–10 (2022)
- [3] Jain, N., Vaidyanath, S., Iyer, A., Natarajan, N., Parthasarathy, S., Rajamani, S., Sharma, R.: Jigsaw: Large language models meet program synthesis. In: *Proceedings of the 44th International Conference on Software Engineering*, pp. 1219–1231 (2022)
- [4] Bui, N.D.Q., Le, H., Wang, Y., Li, J., Gotmare, A.D., Hoi, S.C.H.: CodeTF: One-stop Transformer Library for State-of-the-art Code LLM. *arXiv* (2023). <http://arxiv.org/abs/2306.00029> Accessed 2023-06-22
- [5] Ross, S.I., Martinez, F., Houde, S., Muller, M., Weisz, J.D.: 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*, pp. 491–514. Association for Computing Machinery, New York, NY, USA (2023). <https://doi.org/10.1145/3581641.3584037> . <https://dl.acm.org/doi/10.1145/3581641.3584037> Accessed 2023-06-22
- [6] Chavez, M.R., Butler, T.S., Rekawek, P., Heo, H., Kinzler, W.L.: Chat Generative Pre-trained Transformer: why we should embrace this technology. *American Journal of Obstetrics and Gynecology* **228**(6), 706–711 (2023) <https://doi.org/10.1016/j.ajog.2023.03.010> . Accessed 2023-06-22
- [7] Charalambous, Y., Tihanyi, N., Jain, R., Sun, Y., Ferrag, M.A., Cordeiro, L.C.: A New Era in Software Security: Towards Self-Healing Software via Large Language Models and Formal Verification. *arXiv* (2023). <https://doi.org/10.48550/arXiv.2305.14752> . <http://arxiv.org/abs/2305.14752> Accessed 2023-05-31
- [8] Perry, N., Srivastava, M., Kumar, D., Boneh, D.: Do users write more insecure code with ai assistants? In: *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. CCS ’23*, pp. 2785–2799. Association for Computing Machinery, New York, NY, USA (2023). <https://doi.org/10.1145/3576915.3623157> . <https://doi.org/10.1145/3576915.3623157>
- [9] Tihanyi, N., Bisztray, T., Jain, R., Ferrag, M.A., Cordeiro, L.C., Mavroeidis, V.: The formai dataset: Generative ai in software security through the lens of formal verification. In: *Proceedings of the 19th International Conference on Predictive Models and Data Analytics in Software Engineering. PROMISE 2023*, pp. 33–43. Association for Computing Machinery, New York, NY, USA (2023). <https://doi.org/10.1145/3576915.3623157>

- [10] Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
- [11] Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, É., Hesslow, D., Launay, J., Malartic, Q., et al.: The falcon series of open language models. arXiv preprint arXiv:2311.16867 (2023)
- [12] Roziere, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X.E., Adi, Y., Liu, J., Remez, T., Rapin, J., et al.: Code llama: Open foundation models for code. arXiv preprint arXiv:2308.12950 (2023)
- [13] Gadelha, M.Y.R., Ismail, H.I., Cordeiro, L.C.: Handling loops in bounded model checking of C programs via k-induction. *Int. J. Softw. Tools Technol. Transf.* **19**(1), 97–114 (2017) <https://doi.org/10.1007/s10009-015-0407-9>
- [14] Gadelha, M.R., Monteiro, F.R., Morse, J., Cordeiro, L.C., Fischer, B., Nicole, D.A.: ESBMC 5.0: an industrial-strength c model checker. In: *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*, pp. 888–891. ACM, Montpellier, France (2018)
- [15] Gadelha, M.Y.R., Monteiro, F.R., Cordeiro, L.C., Nicole, D.A.: ESBMC v6.0: Verifying C programs using k-induction and invariant inference - (competition contribution). In: Beyer, D., Huisman, M., Kordon, F., Steffen, B. (eds.) *Tools and Algorithms for the Construction and Analysis of Systems (TACAS)*. LNCS, vol. 11429, pp. 209–213 (2019). Springer
- [16] Menezes, R.S., Aldughaim, M., Farias, B., Li, X., Manino, E., Shmarov, F., Song, K., Brauße, F., Gadelha, M.R., Tihanyi, N., Korovin, K., Cordeiro, L.C.: ESBMC v7.4: Harnessing the power of intervals - (competition contribution). In: *Tools and Algorithms for the Construction and Analysis of Systems (TACAS)*. LNCS, vol. 14572, pp. 376–380 (2024). Springer
- [17] McCabe, T.J.: A complexity measure. *IEEE Transactions on Software Engineering* **SE-2**(4), 308–320 (1976) <https://doi.org/10.1109/TSE.1976.233837>
- [18] Chen, M., Tworek, J., Jun, H., Yuan, Q., Oliveira Pinto, H.P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F.P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W.H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A.N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., Zaremba,

- W.: Evaluating large language models trained on code (2021) [arXiv:2107.03374](#) [cs.LG]
- [19] OpenAI: GPT-4 Technical Report. arXiv (2023). <http://arxiv.org/abs/2303.08774> Accessed 2023-05-29
 - [20] Nehorai, N.: Analyzing Common Vulnerabilities Introduced by Code-Generative AI — HackerNoon (2024). <https://hackernoon.com/analyzing-common-vulnerabilities-introduced-by-code-generative-ai> Accessed 2024-02-28
 - [21] Cordeiro, L.C., Lima Filho, E.B., Bessa, I.V.: Survey on automated symbolic verification and its application for synthesising cyber-physical systems. *IET Cyber-Phys. Syst.: Theory & Appl.* **5**(1), 1–24 (2020) <https://doi.org/10.1049/IET-CPS.2018.5006>
 - [22] Anwar, U., Saparov, A., Rando, J., Paleka, D., Turpin, M., Hase, P., Lubana, E.S., Jenner, E., Casper, S., Sourbut, O., et al.: Foundational challenges in assuring alignment and safety of large language models. arXiv preprint arXiv:2404.09932 (2024)
 - [23] Kirova, V.D., Ku, C.S., Laracy, J.R., Marlowe, T.J.: Software engineering education must adapt and evolve for an llm environment. In: *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1. SIGCSE 2024*, pp. 666–672. Association for Computing Machinery, New York, NY, USA (2024). <https://doi.org/10.1145/3626252.3630927> . <https://doi.org/10.1145/3626252.3630927>
 - [24] Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C., Terry, M., Le, Q., et al.: Program synthesis with large language models (2021)
 - [25] Lu, S., Guo, D., Ren, S., Huang, J., Svyatkovskiy, A., Blanco, A., Clement, C., Drain, D., Jiang, D., Tang, D., et al.: Codexglue: A machine learning benchmark dataset for code understanding and generation. arXiv preprint arXiv:2102.04664 (2021)
 - [26] White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., Schmidt, D.C.: A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. arXiv (2023). <https://doi.org/10.48550/arXiv.2302.11382> . <http://arxiv.org/abs/2302.11382> Accessed 2023-06-24
 - [27] Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T.L., Cao, Y., Narasimhan, K.: Tree of Thoughts: Deliberate Problem Solving with Large Language Models. arXiv (2023). <http://arxiv.org/abs/2305.10601> Accessed 2023-05-29
 - [28] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E.,

- Le, Q., Zhou, D.: Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv (2023). <https://doi.org/10.48550/arXiv.2201.11903> . <http://arxiv.org/abs/2201.11903> Accessed 2023-06-24
- [29] Guo, D., Zhu, Q., Yang, D., Xie, Z., Dong, K., Zhang, W., Chen, G., Bi, X., Wu, Y., Li, Y., et al.: Deepseek-coder: When the large language model meets programming—the rise of code intelligence. arXiv preprint arXiv:2401.14196 (2024)
- [30] Wang, H., Liu, Z., Wang, S., Cui, G., Ding, N., Liu, Z., Yu, G.: Intervenor: Prompt the coding ability of large language models with the interactive chain of repairing. arXiv preprint arXiv:2311.09868 (2023)
- [31] Huang, D., Bu, Q., Zhang, J.M., Luck, M., Cui, H.: Agentcoder: Multi-agent-based code generation with iterative testing and optimisation. arXiv preprint arXiv:2312.13010 (2023)
- [32] Muennighoff, N., Liu, Q., Zebaze, A., Zheng, Q., Hui, B., Zhuo, T.Y., Singh, S., Tang, X., Von Werra, L., Longpre, S.: Octopack: Instruction tuning code large language models. arXiv preprint arXiv:2308.07124 (2023)
- [33] Lin, F., Kim, D.J., et al.: When llm-based code generation meets the software development process. arXiv preprint arXiv:2403.15852 (2024)
- [34] Khoury, R., Avila, A.R., Brunelle, J., Camara, B.M.: How Secure is Code Generated by ChatGPT? arXiv (2023). <http://arxiv.org/abs/2304.09655> Accessed 2023-05-30
- [35] Pearce, H., Ahmad, B., Tan, B., Dolan-Gavitt, B., Karri, R.: Asleep at the Keyboard? Assessing the Security of GitHub Copilot’s Code Contributions. arXiv (2021). <https://doi.org/10.48550/arXiv.2108.09293> . <http://arxiv.org/abs/2108.09293> Accessed 2023-06-10
- [36] Ma, W., Liu, S., Wang, W., Hu, Q., Liu, Y., Zhang, C., Nie, L., Liu, Y.: The Scope of ChatGPT in Software Engineering: A Thorough Investigation. arXiv (2023). <https://doi.org/10.48550/arXiv.2305.12138> . <http://arxiv.org/abs/2305.12138> Accessed 2023-06-10
- [37] Imani, S., Du, L., Shrivastava, H.: Mathprompter: Mathematical reasoning using large language models (2023). <https://doi.org/10.48550/arXiv.2303.05398>
- [38] Hou, X., Zhao, Y., Liu, Y., Yang, Z., Wang, K., Li, L., Luo, X., Lo, D., Grundy, J., Wang, H.: Large Language Models for Software Engineering: A Systematic Literature Review (2024)
- [39] Chan, A., Kharkar, A., Moghaddam, R.Z., Mohylevskyy, Y., Helyar, A., Kamal,

- E., Elkamhawy, M., Sundaresan, N.: Transformer-based vulnerability detection in code at edittime: Zero-shot, few-shot, or fine-tuning? arXiv preprint arXiv:2306.01754 (2023)
- [40] Nguyen, V., Yuan, X., Wu, T., Nepal, S., Grobler, M., Rudolph, C.: Deep learning-based out-of-distribution source code data identification: How far we have gone? arXiv preprint arXiv:2404.05964 (2024)
 - [41] Gao, Z., Wang, H., Zhou, Y., Zhu, W., Zhang, C.: How far have we gone in vulnerability detection using large language models. arXiv preprint arXiv:2311.12420 (2023)
 - [42] Gao, S., Mao, W., Gao, C., Li, L., Hu, X., Xia, X., Lyu, M.R.: Learning in the wild: Towards leveraging unlabeled data for effectively tuning pre-trained code models. arXiv preprint arXiv:2401.01060 (2024)
 - [43] Grishina, A., Hort, M., Moonen, L.: The earlybird catches the bug: On exploiting early layers of encoder models for more efficient code classification. In: Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, pp. 895–907 (2023)
 - [44] Khare, A., Dutta, S., Li, Z., Solko-Breslin, A., Alur, R., Naik, M.: Understanding the effectiveness of large language models in detecting security vulnerabilities. arXiv preprint arXiv:2311.16169 (2023)
 - [45] Noever, D.: Can large language models find and fix vulnerable software? arXiv preprint arXiv:2308.10345 (2023)
 - [46] Shestov, A., Cheshkov, A., Levichev, R., Mussabayev, R., Zadorozhny, P., Maslov, E., Vadim, C., Bulychev, E.: Finetuning large language models for vulnerability detection. arXiv preprint arXiv:2401.17010 (2024)
 - [47] Steenhoek, B., Gao, H., Le, W.: Dataflow analysis-inspired deep learning for efficient vulnerability detection. In: Proceedings of the IEEE/ACM 46th International Conference on Software Engineering. ICSE '24. Association for Computing Machinery, New York, NY, USA (2024). <https://doi.org/10.1145/3597503.3623345> . <https://doi.org/10.1145/3597503.3623345>
 - [48] Sun, Y., Wu, D., Xue, Y., Liu, H., Ma, W., Zhang, L., Shi, M., Liu, Y.: Llm4vuln: A unified evaluation framework for decoupling and enhancing llms' vulnerability reasoning. arXiv preprint arXiv:2401.16185 (2024)
 - [49] Tang, W., Tang, M., Ban, M., Zhao, Z., Feng, M.: Csgvd: A deep learning approach combining sequence and graph embedding for source code vulnerability detection. J. Syst. Softw. **199**(C) (2023) <https://doi.org/10.1016/j.jss.2023.111623>

- [50] Thapa, C., Jang, S.I., Ahmed, M.E., Camtepe, S., Pieprzyk, J., Nepal, S.: Transformer-based language models for software vulnerability detection. In: Proceedings of the 38th Annual Computer Security Applications Conference. ACSAC '22, pp. 481–496. Association for Computing Machinery, New York, NY, USA (2022). <https://doi.org/10.1145/3564625.3567985> . <https://doi.org/10.1145/3564625.3567985>
- [51] Zhang, C., Liu, H., Zeng, J., Yang, K., Li, Y., Li, H.: Prompt-enhanced software vulnerability detection using chatgpt. arXiv preprint arXiv:2308.12697 (2023)
- [52] Tóth, R., Bisztray, T., Erdodi, L.: LLMs in Web-Development: Evaluating LLM-Generated PHP code unveiling vulnerabilities and limitations (2024)
- [53] Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., Anderson, R.: The Curse of Recursion: Training on Generated Data Makes Models Forget. arXiv (2023). <http://arxiv.org/abs/2305.17493> Accessed 2023-06-27
- [54] Chen, Y., Ding, Z., Chen, X., Wagner, D.: DiverseVul: A New Vulnerable Source Code Dataset for Deep Learning Based Vulnerability Detection. arXiv (2023). <http://arxiv.org/abs/2304.00409> Accessed 2023-06-27
- [55] Fan, J., Li, Y., Wang, S., Nguyen, T.N.: 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, pp. 508–512. Association for Computing Machinery, New York, NY, USA (2020). <https://doi.org/10.1145/3379597.3387501> . <https://doi.org/10.1145/3379597.3387501> Accessed 2023-06-27
- [56] Russell, R.L., Kim, L.Y., Hamilton, L.H., Lazovich, T., Harer, J.A., Ozdemir, O., Ellingwood, P.M., McConley, M.W.: Automated Vulnerability Detection in Source Code Using Deep Representation Learning. In: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 757–762. IEEE, Orlando, FL, USA (2018). <https://doi.org/10.1109/ICMLA.2018.00120> . <https://api.semanticscholar.org/CorpusID:49670513>
- [57] Kim, L., Russell, R.: Draper VDISC Dataset - Vulnerability Detection in Source Code. Publisher: OSF (2018). <https://osf.io/d45bw/> Accessed 2023-06-27
- [58] Black, P.E.: A Software Assurance Reference Dataset: Thousands of Programs With Known Bugs. Journal of Research of the National Institute of Standards and Technology **123**, 1–3 (2018) <https://doi.org/10.6028/jres.123.005> . Accessed 2023-06-27
- [59] Jr, F.E.B., Black, P.E.: The Juliet 1.1 C/C++ and Java Test Suite. NIST **45**(10), 88–90 (2012). Last Modified: 2021-10-12T11:10-04:00 Publisher: Frederick E. Boland Jr., Paul E. Black. Accessed 2023-05-28

- [60] Zhou, Y., Liu, S., Siow, J., Du, X., Liu, Y.: Devign: Effective Vulnerability Identification by Learning Comprehensive Program Semantics via Graph Neural Networks, pp. 10197–10207. Curran Associates Inc., Red Hook, NY, USA (2019)
- [61] Chakraborty, S., Krishna, R., Ding, Y., Ray, B.: Deep Learning Based Vulnerability Detection: Are We There Yet? *IEEE Transactions on Software Engineering* **48**(9), 3280–3296 (2022) <https://doi.org/10.1109/TSE.2021.3087402>
- [62] Tihanyi, N., Bisztray, T., Jain, R., Amine Ferrag, M., C. Cordeiro, L., Mavroeidis, V.: FormAI Dataset: A Large Collection of AI-Generated C Programs and Their Vulnerability Classifications. *IEEE Dataport* (2023). <https://doi.org/10.21227/vp9n-wv96> . <https://dx.doi.org/10.21227/vp9n-wv96>
- [63] Jain, R., Gervasoni, N., Ndhlovu, M., Rawat, S.: 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*, pp. 1–10. ACM, Prayagraj, India (2023)
- [64] Cordeiro, L., Fischer, B., Marques-Silva, J.: SMT-Based Bounded Model Checking for Embedded ANSI-C Software. *IEEE Transactions on Software Engineering* **38**(4), 957–974 (2012) <https://doi.org/10.1109/TSE.2011.59>
- [65] D’Silva, V., Kroening, D., Weissenbacher, G.: A Survey of Automated Techniques for Formal Software Verification. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **27**(7), 1165–1178 (2008) <https://doi.org/10.1109/TCAD.2008.923410>
- [66] Morse, J., Cordeiro, L.C., Nicole, D.A., Fischer, B.: Context-bounded model checking of LTL properties for ANSI-C software. In: Barthe, G., Pardo, A., Schneider, G. (eds.) *Software Engineering and Formal Methods - 9th International Conference, SEFM 2011, Montevideo, Uruguay, November 14-18, 2011. Proceedings. Lecture Notes in Computer Science*, vol. 7041, pp. 302–317 (2011). Springer
- [67] Wallace, D.R., Fujii, R.U.: Software verification and validation: an overview. *IEEE Software* **6**(3), 10–17 (1989) <https://doi.org/10.1109/52.28119> . Accessed 2023-06-22
- [68] Alshmrany, K.M., Aldughaim, M., Bhayat, A., Cordeiro, L.C.: Fusebmc: An energy-efficient test generator for finding security vulnerabilities in C programs. In: Loulergue, F., Wotawa, F. (eds.) *Tests and Proofs - 15th International Conference, TAP 2021, Held as Part of STAF 2021, Virtual Event, June 21-22, 2021. Proceedings. Lecture Notes in Computer Science*, vol. 12740, pp. 85–105 (2021). Springer
- [69] Braberman, V.A., Bonomo-Braberman, F., Charalambous, Y., Colonna, J.G., Cordeiro, L.C., Freitas, R.: Tasks People Prompt: A Taxonomy of LLM

- [70] Hao, Y., Chen, W., Zhou, Z., Cui, W.: E&v: Prompting large language models to perform static analysis by pseudo-code execution and verification. arXiv preprint arXiv:2312.08477 (2023)
- [71] Yang, A.Z., Le Goues, C., Martins, R., Hellendoorn, V.: Large language models for test-free fault localization. In: Proceedings of the 46th IEEE/ACM International Conference on Software Engineering, pp. 1–12 (2024)
- [72] Quan, V.L.A., Phat, C.T., Van Nguyen, K., Duy, P.T., Pham, V.-H.: Xgv-bert: Leveraging contextualized language model and graph neural network for efficient software vulnerability detection. arXiv preprint arXiv:2309.14677 (2023)
- [73] Sun, T., Allix, K., Kim, K., Zhou, X., Kim, D., Lo, D., Bissyandé, T.F., Klein, J.: Dexbert: Effective, task-agnostic and fine-grained representation learning of android bytecode. *IEEE Transactions on Software Engineering* **49**(10), 4691–4706 (2023) <https://doi.org/10.1109/TSE.2023.3310874>
- [74] Tian, H., Liu, K., Li, Y., Kaboré, A.K., Koyuncu, A., Habib, A., Li, L., Wen, J., Klein, J., Bissyandé, T.F.: The best of both worlds: Combining learned embeddings with engineered features for accurate prediction of correct patches. *ACM Trans. Softw. Eng. Methodol.* **32**(4) (2023) <https://doi.org/10.1145/3576039>
- [75] Wang, W., Wang, Y., Joty, S., Hoi, S.C.H.: Rap-gen: Retrieval-augmented patch generation with codet5 for automatic program repair. In: Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2023, pp. 146–158. Association for Computing Machinery, New York, NY, USA (2023). <https://doi.org/10.1145/3611643.3616256> . <https://doi.org/10.1145/3611643.3616256>
- [76] Zhang, Y., Jin, Z., Xing, Y., Li, G.: Steam: simulating the interactive behavior of programmers for automatic bug fixing. arXiv preprint arXiv:2308.14460 (2023)
- [77] Wu, Y., Li, Z., Zhang, J.M., Papadakis, M., Harman, M., Liu, Y.: Large language models in fault localisation. arXiv preprint arXiv:2308.15276 (2023)
- [78] Mohajer, M.M., Aleithan, R., Harzevili, N.S., Wei, M., Belle, A.B., Pham, H.V., Wang, S.: Skipanalyzer: An embodied agent for code analysis with large language models. arXiv preprint arXiv:2310.18532 (2023)
- [79] Li, T.-O., Zong, W., Wang, Y., Tian, H., Wang, Y., Cheung, S.-C.: Finding Failure-Inducing Test Cases with ChatGPT (2023)
- [80] Pearce, H., Tan, B., Ahmad, B., Karri, R., Dolan-Gavitt, B.: Examining Zero-Shot Vulnerability Repair with Large Language Models. arXiv (2022). <http://arxiv.org/abs/2112.02125> Accessed 2023-06-10

- [81] Cao, J., Li, M., Wen, M., Cheung, S.-c.: A study on prompt design, advantages and limitations of chatgpt for deep learning program repair. arXiv preprint arXiv:2304.08191 (2023)
- [82] Deligiannis, P., Lal, A., Mehrotra, N., Rastogi, A.: Fixing rust compilation errors using llms. arXiv preprint arXiv:2308.05177 (2023)
- [83] Fan, Z., Gao, X., Mirchev, M., Roychoudhury, A., Tan, S.H.: Automated repair of programs from large language models. In: 2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE), pp. 1469–1481 (2023). IEEE
- [84] Huang, Q., Zhu, J., Xing, Z., Jin, H., Wang, C., Xu, X.: A chain of ai-based solutions for resolving fqns and fixing syntax errors in partial code. arXiv preprint arXiv:2306.11981 (2023)
- [85] Islam, N.T., Najafirad, P.: Code security vulnerability repair using reinforcement learning with large language models. arXiv preprint arXiv:2401.07031 (2024)
- [86] Jin, M., Shahriar, S., Tufano, M., Shi, X., Lu, S., Sundaresan, N., Svyatkovskiy, A.: InferFix: End-to-End Program Repair with LLMs (2023)
- [87] Lajkó, M., Csuvik, V., Vidács, L.: Towards javascript program repair with generative pre-trained transformer (gpt-2). In: 2022 IEEE/ACM International Workshop on Automated Program Repair (APR), pp. 61–68 (2022). <https://doi.org/10.1145/3524459.3527350>
- [88] Paul, R., Mohib Hossain, M., Hasan, M., Iqbal, A.: Automated program repair based on code review: How do pre-trained transformer models perform? arXiv e-prints, 2304 (2023)
- [89] Peng, Y., Gao, S., Gao, C., Huo, Y., Lyu, M.: Domain knowledge matters: Improving prompts with fix templates for repairing python type errors. In: Proceedings of the IEEE/ACM 46th International Conference on Software Engineering. ICSE '24. Association for Computing Machinery, New York, NY, USA (2024). <https://doi.org/10.1145/3597503.3608132> . <https://doi.org/10.1145/3597503.3608132>
- [90] Tian, H., Liu, K., Kaboré, A.K., Koyuncu, A., Li, L., Klein, J., Bissyandé, T.F.: Evaluating representation learning of code changes for predicting patch correctness in program repair. In: Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering. ASE '20, pp. 981–992. Association for Computing Machinery, New York, NY, USA (2021). <https://doi.org/10.1145/3324884.3416532> . <https://doi.org/10.1145/3324884.3416532>
- [91] Wei, Y., Xia, C.S., Zhang, L.: Copiloting the copilots: Fusing large language models with completion engines for automated program repair. In: Proceedings

- of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering. ESEC/FSE 2023, pp. 172–184. Association for Computing Machinery, New York, NY, USA (2023). <https://doi.org/10.1145/3611643.3616271> . <https://doi.org/10.1145/3611643.3616271>
- [92] Widjojo, P., Treude, C.: Addressing compiler errors: Stack overflow or large language models? arXiv preprint arXiv:2307.10793 (2023)
 - [93] Xia, C.S., Wei, Y., Zhang, L.: Practical program repair in the era of large pre-trained language models. arXiv preprint arXiv:2210.14179 (2022)
 - [94] Xia, C.S., Zhang, L.: Keep the conversation going: Fixing 162 out of 337 bugs for \$0.42 each using chatgpt. arXiv preprint arXiv:2304.00385 (2023)
 - [95] Zhang, Q., Fang, C., Sun, W., Liu, Y., He, T., Hao, X., Chen, Z.: Appt: Boosting automated patch correctness prediction via fine-tuning pre-trained models. *IEEE Transactions on Software Engineering* **50**(3), 474–494 (2024) <https://doi.org/10.1109/TSE.2024.3354969>
 - [96] Zhang, Q., Fang, C., Zhang, T., Yu, B., Sun, W., Chen, Z.: Gamma: Revisiting template-based automated program repair via mask prediction. In: 2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE), pp. 535–547. IEEE Computer Society, Los Alamitos, CA, USA (2023). <https://doi.org/10.1109/ASE56229.2023.00063> . <https://doi.org/10.1109/ASE56229.2023.00063>
 - [97] Zhang, Y., Li, G., Jin, Z., Xing, Y.: Neural program repair with program dependence analysis and effective filter mechanism. arXiv preprint arXiv:2305.09315 (2023)
 - [98] Wu, Y., Jiang, N., Pham, H.V., Lutellier, T., Davis, J., Tan, L., Babkin, P., Shah, S.: How effective are neural networks for fixing security vulnerabilities. In: Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis, pp. 1282–1294 (2023)
 - [99] Aho, A.V., Lam, M.S., Sethi, R., Ullman, J.D.: *Compilers: Principles, Techniques, And Tools*, 2nd edn. Addison-Wesley Longman Publishing Co., Inc., Boston, MA (2006)
 - [100] Gadelha, M.Y.R., Steffinlongo, E., Cordeiro, L.C., Fischer, B., Nicole, D.A.: Smt-based refutation of spurious bug reports in the clang static analyzer. In: Atlee, J.M., Bultan, T., Whittle, J. (eds.) Proceedings of the 41st International Conference on Software Engineering, pp. 11–14. IEEE / ACM, Montreal, QC, Canada (2019). <https://doi.org/10.1109/ICSE-Companion.2019.00026>
 - [101] Sadowski, C., Yi, J.: How developers use data race detection tools. In: Proceedings of the 5th Workshop on Evaluation and Usability of Programming

- Languages and Tools, pp. 43–51. ACM, Portland, USA (2014)
- [102] White, M., Tufano, M., Vendome, C., Poshyvanyk, D.: Deep learning code fragments for code clone detection. In: Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering, pp. 87–98. Association for Computing Machinery, New York, USA (2016)
 - [103] Zhao, G., Huang, J.: 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, pp. 141–151. ACM, Lake Buena Vista, USA (2018)
 - [104] Cordeiro, L.C., Kroening, D., Schrammel, P.: JBMC: bounded model checking for java bytecode - (competition contribution). In: Tools and Algorithms for the Construction and Analysis of Systems (TACAS). LNCS, vol. 11429, pp. 219–223 (2019). Springer
 - [105] Menezes, R., Moura, D., Cavalcante, H., Freitas, R., Cordeiro, L.C.: Esbmc-jimple: verifying kotlin programs via jimple intermediate representation. In: Ryu, S., Smaragdakis, Y. (eds.) ISSTA '22: 31st ACM SIGSOFT International Symposium on Software Testing and Analysis, Virtual Event, South Korea, July 18 - 22, 2022, pp. 777–780 (2022). ACM
 - [106] Gadelha, M.R., Monteiro, F.R., Morse, J., Cordeiro, L.C., Fischer, B., Nicole, D.A.: Esbmc 5.0: an industrial-strength c model checker. In: Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering. ASE '18, pp. 888–891. Association for Computing Machinery, New York, NY, USA (2018). <https://doi.org/10.1145/3238147.3240481> . <https://doi.org/10.1145/3238147.3240481>
 - [107] Beyer, D.: Competition on software verification and witness validation: Sv-comp 2023. In: Sankaranarayanan, S., Sharygina, N. (eds.) Tools and Algorithms for the Construction and Analysis of Systems, pp. 495–522. Springer, Cham (2023)
 - [108] Sandoval, G., Pearce, H., Nys, T., Karri, R., Garg, S., Dolan-Gavitt, B.: Lost at C: A User Study on the Security Implications of Large Language Model Code Assistants. arXiv (2023)
 - [109] Wallace, D.R., Watson, A.H., McCabe, T.J.: Structured testing : a testing methodology using the cyclomatic complexity metric. Technical Report NIST SP 500-235, National Institute of Standards and Technology, Gaithersburg, MD (1996). <https://doi.org/10.6028/NIST.SP.500-235> . Edition: 0. <https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication500-235.pdf> Accessed 2024-04-19
 - [110] Mikejo5000: Code metrics - Cyclomatic complexity - Visual Studio (Windows) (2024). <https://learn.microsoft.com/en-us/visualstudio/code-quality/>

[code-metrics-cyclomatic-complexity?view=vs-2022](#) Accessed 2024-04-18

- [111] Kroening, D., Tautschnig, M.: Cbmc-c bounded model checker: (competition contribution). In: Tools and Algorithms for the Construction and Analysis of Systems: TACAS 2014, pp. 389–391. Springer, Grenoble, France (2014)