

SV-TrustEval-C: Evaluating Structure and Semantic Reasoning in Large Language Models for Source Code Vulnerability Analysis

Yansong Li*, Paula Branco*, Alexander M. Hoole†, Manish Marwah†,
Hari Manassery Koduvely†, Guy-Vincent Jourdan*, Stephan Jou†

*University of Ottawa, †OpenText

Email: {yli627, pbranco, gjourdan}@uottawa.ca, {hoole, mmawah, hkoduvely, sjou}@opentext.com

Abstract—As Large Language Models (LLMs) evolve in understanding and generating code, accurately evaluating their reliability in analyzing source code vulnerabilities becomes increasingly vital. While studies have examined LLM capabilities in tasks like vulnerability detection and repair, they often overlook the importance of both structure and semantic reasoning crucial for trustworthy vulnerability analysis. To address this gap, we introduce SV-TRUSTEVAL-C, a benchmark designed to evaluate LLMs’ abilities for vulnerability analysis of code written in the C programming language through two key dimensions: structure reasoning—assessing how models identify relationships between code elements under varying data and control flow complexities; and semantic reasoning—examining their logical consistency in scenarios where code is structurally and semantically perturbed. Our results show that current LLMs are far from satisfactory in understanding complex code relationships and that their vulnerability analyses rely more on pattern matching than on robust logical reasoning. These findings underscore the effectiveness of the SV-TRUSTEVAL-C benchmark and highlight critical areas for enhancing the reasoning capabilities and trustworthiness of LLMs in real-world vulnerability analysis tasks. Our initial benchmark dataset is available at <https://huggingface.co/datasets/LLMs4CodeSecurity/SV-TrustEval-C-1.0>

1. Introduction

With the continued development of Large Language Models (LLMs) [1], [2], [3], specialized versions tailored for the code domain [4], [5], [6], have demonstrated promising capabilities in code understanding and generation. Notable applications such as Amazon CodeWhisperer and GitHub Copilot have significantly boosted developer productivity and been widely adopted by millions of developers [7]. Building on these advances, numerous studies have broadened the applications of LLMs to include cybersecurity tasks, particularly in code vulnerability analysis [8], [9], [10]. Unlike traditional tools that depend on post-hoc verification through static [11] or dynamic analysis [12], LLMs offer real-time security feedback during the coding process [13] and can generate secure code while following detailed instructions [14], significantly reducing the risk of software vulnerabilities.

Despite their potential, studies have shown that LLMs do not guarantee code security and are prone to generating syntactically correct but semantically insecure code [15], [16], [17]. This raises concerns about their utility in code vulnerability analysis and underscores the urgent need to assess their capabilities in the security domain [18], [19]. However, existing benchmarks for code vulnerability, such as CVEFixes [20], BigVul [21], and DiverseVul [22], primarily focus on straightforward tasks like identifying or repairing publicly known vulnerable code snippets within pre-defined scenarios. These benchmarks fall short as they fail to discern whether success in resolving vulnerabilities stems from leveraging pre-trained parameterized knowledge [23], [24], or from a logical reasoning between code and vulnerabilities. This distinction is vital for verifying the reliability of LLMs, as reliance exclusively on pre-trained patterns often leads to poor generalization [25] and inconsistent interoperability [26], especially when addressing semantic and structural variations in code within real-world scenarios. To overcome these limitations, we formally define the core capabilities required for vulnerability analysis and introduce SV-TRUSTEVAL-C, a new question-answering benchmark designed for the trustworthy evaluation of LLMs in source code vulnerability analysis. This benchmark involves:

Structure Reasoning: Unlike unstructured natural language, code inherently possesses rigorous structural information. For natural language-driven LLMs, the ability to comprehend this code structure is essential for effective code understanding and reasoning [27], [28]. In vulnerability analysis, understanding the relationships between code elements is critical for assessing the scope and impact of vulnerabilities and for implementing accurate repair strategies. This benchmark evaluates the ability of LLMs to identify and predict the effects of modifications within code elements on each other.

Semantic Reasoning: Leveraging domain knowledge to adapt to changes in code is crucial for maintaining robustness in dynamic environments. This benchmark component evaluates LLMs through scenarios that simulate real-world conditions, including: **Counterfactual scenarios**, where altered code semantics challenge LLMs to apply logical reasoning beyond learned patterns; **Goal-driven scenarios**, where LLMs are tasked with completing code while ensuring functionality and security, testing their ability to handle

complex modifications without introducing vulnerabilities; and **Predictive scenarios**, where LLMs must accurately identify and differentiate between various types of vulnerabilities, including scenarios where vulnerabilities will not be triggered at runtime, thereby assessing their application of security concepts.

To develop SV-TRUSTEVAL-C, we created a Structure-Oriented Variants Generator capable of extracting structural information, perturbing code semantics based on the provided base code and classifications (*Safe*, *Unsafe* code)¹, and increasing code complexity in alignment with data flow and control flow graphs [29]. We conducted extensive experiments using the SV-TRUSTEVAL-C benchmark with zero-shot and in-context learning inference to evaluate eleven popular LLMs across various parameter scales. The results reveal that LLMs struggle to recognize relationships between code elements and predominantly rely on pattern matching rather than logical reasoning in vulnerability analysis scenarios. This highlights the need for specialized improvements in LLMs to enhance their practical utility in real-world code vulnerability analysis. Our contributions are the following:

- 1) **Reasoning-based Benchmark for Vulnerability Analysis:** We present SV-TRUSTEVAL-C, the first benchmark designed to assess LLMs’ ability to analyze source code vulnerabilities through logical reasoning and structural understanding, moving beyond mere pattern recognition.
- 2) **Structure-Oriented Variants Generator:** We created a generator that systematically extracts structural information, alters code semantics, and increases complexity based on data and control flow graphs using *Safe* and *Unsafe* code pairs.
- 3) **Identifying Gaps in LLM Capabilities:** Evaluating eleven LLMs revealed their reliance on pattern matching over logical reasoning in vulnerability analysis, highlighting the need for enhancing their reasoning capabilities in security applications.

2. Background and Related Work

Historically, identifying and mitigating source code vulnerabilities involved manual code reviews, which were both time-consuming and error-prone [38]. Automated tools and methodologies, such as static analysis (e.g., FindBugs [39], PMD [40], and Checkstyle [41]) and dynamic analysis techniques like fuzz testing [42], have evolved to increase detection accuracy and efficiency [43], [44]. Advances in machine learning and deep learning have significantly enhanced vulnerability detection capabilities [43], [45], [46], [47], [48], [49]. More recently, LLMs [1] have been explored for their potential to understand complex code patterns, predict vulnerabilities, and even automate vulnerability repairs [4], [50], [51], [52], [53], [54]. Despite these advances, questions remain regarding the reliability of LLMs in detecting and repairing specific vulnerabilities under diverse conditions,

highlighting the need for rigorous benchmarking to establish their efficacy and trustworthiness in software security applications.

2.1. Code Vulnerability Benchmarks

Benchmarks in the code vulnerability domain are designed to evaluate various stages and aspects of how LLMs and other automated tools handle vulnerabilities in code. Broadly, these benchmarks can be categorized according to five key tasks: **Identification**, **Repair**, **Safe-Generation**, **Question-Answering (QA)** and **Reasoning**. Each category addresses a distinct component of understanding and managing code vulnerabilities, providing specific datasets and metrics to assess model performance. Table 1 highlights these benchmarks, detailing their focus and labels which of the five tasks they each satisfy. To clearly define the tasks and distinguish our work from existing benchmarks, we formally characterize each of the five vulnerability domain tasks using mathematical formulations in the following sections.

2.1.1. Identification Tasks. These tasks involve detecting and classifying vulnerable code snippets while evaluating whether each snippet contains vulnerabilities and pinpointing the specific vulnerable segments when necessary. Benchmarks such as *Devign* [30] and *VulPatchPairs* [31] aim to evaluate a model’s capability to accurately identify whether a piece of code is vulnerable. Additionally, benchmarks like *BigVul* [21], *CrossVul* [20], *CVEFixes* [20], and *DiverseVul* [22] not only classify code snippets as vulnerable or safe but also provide insights into different types of vulnerabilities gleaned from real-world open-source repositories. Formally, given a code snippet X , the objective is usually to predict a binary label y indicating vulnerability:

$$P(y \mid X), \quad \text{where } y \in \{0, 1\}, \quad (1)$$




Some benchmarks extend this task to identify specific lines or elements of code that are vulnerable, which can be formulated as:















$$P(Y \mid X) = \prod_{t=1}^{|X|} P(y_t \mid X, y_{<t}), \quad (2)$$

where $y_{<t}$ denotes the vulnerability predictions of all preceding lines, and $Y = \{y_1, \dots, y_{|X|}\}$ is the set of predicted labels for each line in X . These tasks challenge models to combine global contextual understanding with fine-grained local analysis, reflecting the complex nature of real-world vulnerability detection.

2.1.2. Repair Tasks. Repair tasks generate secure versions of vulnerable code snippets by accurately identifying and fixing the underlying issues. Benchmarks that combine identification and repair tasks provide pairs of vulnerable code snippets (X_v) and their corresponding fixed versions (X_r). Examples include *BigVul*, *CrossVul*, *CVEFixes*, and *DiverseVul*, which gather data from commit records on open-source platforms like GitHub. This data includes the original

1. *Safe* and *Unsafe*: defined by whether the code triggers a vulnerability.

TABLE 1: Comparison of recent code vulnerability evaluation benchmarks against our proposed SV-TRUSTEVAL-C. “CWE Scope” indicates the number of CWEs in a benchmark; “Num. of Func.” shows the total functions evaluated. Icons indicate dataset origin— for manually labeled real-world data,  for automatically labeled real-world data, and  for synthetic data. Support levels are shown as \checkmark (full), \bigcirc (partial), and \times (none). Key features of our benchmark—CWE coverage, function count, and reasoning capability—are **highlighted** relative to SECLLMHOLMES. Evaluation tasks: **Identification** (Eq. 1 and Eq. 2) measures vulnerability detection and localization; **Repair** (Eq. 3) assesses conversion of vulnerable code into secure versions; **Generation** (Eq. 4) evaluates producing safe code without introducing new vulnerabilities; **QA** (Eq. 5) tests domain knowledge via targeted question-answering; and **Reasoning** (Eq. 6) evaluates the underlying reasoning process.

Benchmark	CWE Scope	Num. of Func.	Source	Identification	Repair	Generation	QA	Reasoning
SVEN [14]	9	1,606		\times	\times	\checkmark	\times	\times
Devign [30]	N/A	26,037		\checkmark	\checkmark	\times	\times	\times
VulPatchPairs [31]	N/A	26,000		\checkmark	\checkmark	\times	\times	\times
BigVul [21]	91	264,919		\checkmark	\checkmark	\times	\times	\times
CrossVul [20]	168	134,126		\checkmark	\checkmark	\times	\times	\times
CVEFixes [20]	180	168,089		\checkmark	\checkmark	\times	\times	\times
DiverseVul [22]	150	330,492		\checkmark	\checkmark	\times	\times	\times
Juliet [32]	118	61,387		\checkmark	\checkmark	\times	\times	\times
ROBUSTAPI [33]	N/A	1,208		\times	\times	\checkmark	\checkmark	\times
CyberMetric [34]	N/A	N/A		\times	\times	\times	\checkmark	\times
CyberBench [35]	N/A	N/A		\checkmark	\times	\times	\checkmark	\times
CYBERSECEVAL2 [36]	N/A	N/A		\times	\times	\checkmark	\times	\times
SecLLMHolmes [37]	8	228		\checkmark	\times	\times	\times	\bigcirc
SV-TRUSTEVAL-C (<i>Ours</i>)	82	3,337		\checkmark	\times	\times	\times	\checkmark

vulnerabilities and their subsequent fixes, serving as ground truth for evaluating a model’s repair capabilities. The formal objective for vulnerability repair tasks is:

$$P(X_r | X_v) = \prod_{t=1}^{|X_r|} P(x_t | x_{<t}, X_v), \quad (3)$$

where X_v is the vulnerable code and X_r is the repaired code. This formulation assesses a model’s proficiency in transforming vulnerable code into a secure version, reflecting real-world scenarios where developers or automated tools must identify and rectify security flaws efficiently.

2.1.3. Safe-Generation Tasks. These tasks focus on generating secure code snippets while evaluating a model’s ability to either produce only safe code when prompted or deny requests that could lead to insecure code, ensuring that no new vulnerabilities are introduced. Benchmarks such as *SVEN* [14] emphasize controlled code generation, where the LLM generates code only if the input prompt (c) is considered “safe,” and outputs a denial response if it’s not. This behavior is expressed as:

$$P(X | c) = \begin{cases} \prod_{t=1}^{|X|} P(x_t | x_{<t}, c), & \text{if } c \text{ is safe;} \\ \delta(X = D), & \text{if } c \text{ is not safe,} \end{cases} \quad (4)$$

where the Dirac delta function $\delta(X = D)$ represents a probability distribution concentrated entirely on a denial response D . Benchmarks like *RobustAPI* [33] assess how frequently the code generated by an LLM contains vulnerabilities by measuring: $P(V(X) = 1 | c) = \mathbb{E}_{X \sim P(X|c)}[V(X)]$, where $V(X)$ is an indicator function that equals 1 if X

is vulnerable, and 0 otherwise. Other benchmarks, such as *CYBERSECEVAL 2* [36], investigate controlled code generation in the presence of prompt injection attacks [55], [56], measuring the resilience of LLMs to malicious instructions. Collectively, these safe-generation tasks and benchmarks not only evaluate the correctness and functionality of the generated code but also its security and the LLM’s robustness against introducing vulnerabilities.

2.1.4. Question-Answering Tasks. Question-Answering (QA) tasks tests the LLM model’s knowledge and understanding of vulnerability concepts by requiring it to accurately and thoroughly answer security related questions with contextually appropriate explanations. Benchmarks like *CyberBench* [35] and *CyberMetric* [34] present various types of vulnerability-related questions and require models to provide correct and contextually appropriate answers:

$$P(A | Q), \quad (5)$$

where Q is a question and A is the model’s answer. These tasks evaluate the depth and applicability of a model’s understanding of vulnerabilities, including its capacity to explain vulnerability concepts, implications, and potential mitigations or fixes.

2.1.5. Vulnerability Reasoning Tasks. Vulnerability Reasoning tasks evaluate a model’s ability to logically explain vulnerabilities, differentiating true analytical reasoning from mere pattern matching. While earlier tasks primarily assess an LLM’s effectiveness in detecting vulnerabilities, they typically do not clarify whether the model relies on superficial pattern recognition or engages in genuine logical reasoning.

This distinction is crucial, as vulnerability analysis demands deep code understanding to accurately interpret complex structures, dependencies, and contextual nuances beyond pattern matching capabilities. Logical reasoning empowers LLMs to identify novel and intricate security flaws, adapt effectively to emerging threats, and deliver accurate, context-sensitive evaluations essential for securing complex software systems. Addressing this critical gap, *SecLLMHolmes* [37] introduces an approach specifically designed to assess reasoning capabilities in code vulnerability analysis. Vulnerability Reasoning involves formulating both the reasoning process and the identification of vulnerabilities as:

$$P(O \mid X) = \prod_{t=1}^{|O|} P(o_t \mid o_{<t}, X), \quad (6)$$

where $O = R \parallel V$ represents the concatenation of reasoning steps R and vulnerability identification V . This task evaluates an LLM’s capacity to understand and explain the logic behind a vulnerability. However, *SecLLMHolmes* provides only a limited set of code scenarios (48 hand-crafted code samples) within MITRE’s Top 25 Common Weakness Enumeration (CWE)², and its evaluation relies primarily on sentence similarity, $\text{Similarity}(O_{\text{LLM}}, O_{\text{Human}})$, between LLM-generated reasoning and human-crafted reasoning. This method doesn’t conclusively show if LLMs rely on pattern matching or genuine reasoning due to limited tests and low interpretability.

In this work, we propose a systematic approach to evaluating LLMs’ reasoning abilities in vulnerability analysis scenarios. Our benchmark, built upon the synthetically generated Juliet Test Suite covering 82 distinct CWEs, ensures 100% label accuracy while introducing significantly scalable task complexity to comprehensively evaluate LLMs’ vulnerability reasoning capabilities. It is designed to overcome the limitations of existing benchmarks by offering extensive test scenarios and robust evaluation metrics that can better discern genuine logical reasoning from pattern recognition.

2.2. Data Sources and Quality

Existing benchmarks utilize data and labels from three primary sources. First, open-source repositories are used in an automated manner, where datasets like *BigVul* and *CrossVul* are created by scraping code and commit information from platforms such as GitHub, with vulnerability labels automatically assigned based on commit metadata and code diffs. While this method can produce extensive datasets, it often suffers from imprecise labeling and inconsistent data quality [9]. Second, some benchmarks, including *Devign* and *VulPatchPairs*, combine automated data collection with manual verification by human experts, ensuring greater label accuracy and more reliable ground truth information, although resulting in smaller datasets. Third, synthetic sources such as the *Juliet Test Suite* [32] and *Romeo* [57] systematically generate code with high

label accuracy, though risks of oversimplification remain. Our benchmark dataset is built on synthetic data but extends its complexity through our Structure-Oriented Variants Generator, ensuring both label accuracy and scalable complexity that approximates real-world scenarios.

2.3. Code Reasoning

Code reasoning involves analyzing and predicting a program’s behavior without direct execution, aiding tasks such as vulnerability detection, code generation, and program comprehension. Traditional approaches leverage code execution information—like inputs, outputs, and intermediate runtime states—to enhance model performance in these areas [58], [59], [60], [61]. For example, *Lever* [59] integrated a verifier that utilizes execution results to improve code generation performance, while Chen et al. [60] employed output-based feedback for guiding LLMs in self-debugging generated code. Meanwhile, some studies have incorporated dynamic features such as runtime program states to train language models capturing deeper code semantics [61], [62]. Recent benchmarks have started evaluating the code reasoning abilities of LLMs. CodeMind [28] tests how well models reason about input-output relations in code, whereas CRUXEval [27] goes further to measure how models infer runtime behavior, including code coverage and execution paths. These benchmarks demonstrate progress in assessing LLMs’ code understanding but primarily focus on general code correctness and runtime behavior inference.

Our work differs by explicitly emphasizing code vulnerability reasoning. Unlike general code reasoning tasks that concentrate on functionality and runtime behavior, vulnerability reasoning requires understanding complex code semantics and contextual security considerations. Existing benchmarks like CRUXEval and CodeMind do not focus on this specialized aspect. We introduce a benchmark tailored to evaluating how well LLMs handle vulnerability reasoning tasks, aiming to provide more targeted insights into model capabilities in analyzing code for security flaws and advancing our understanding of code reasoning in security contexts.

3. SV-TrustEval-C Benchmark

This section first outlines the capabilities necessary for reliable vulnerability analysis. Subsequently, we describe the key features of the SV-TRUSTEVAL-C Benchmark, which leverages our Structure-Oriented Variants Generator.

3.1. Core Capabilities

To thoroughly evaluate the reliability of LLMs in vulnerability analysis, we design two main categories of tasks: **Structure Reasoning** and **Semantic Reasoning**. These categories correspond to fundamental aspects of code vulnerability tasks, as illustrated in Figure 1.

2. <https://cwe.mitre.org/top25/>

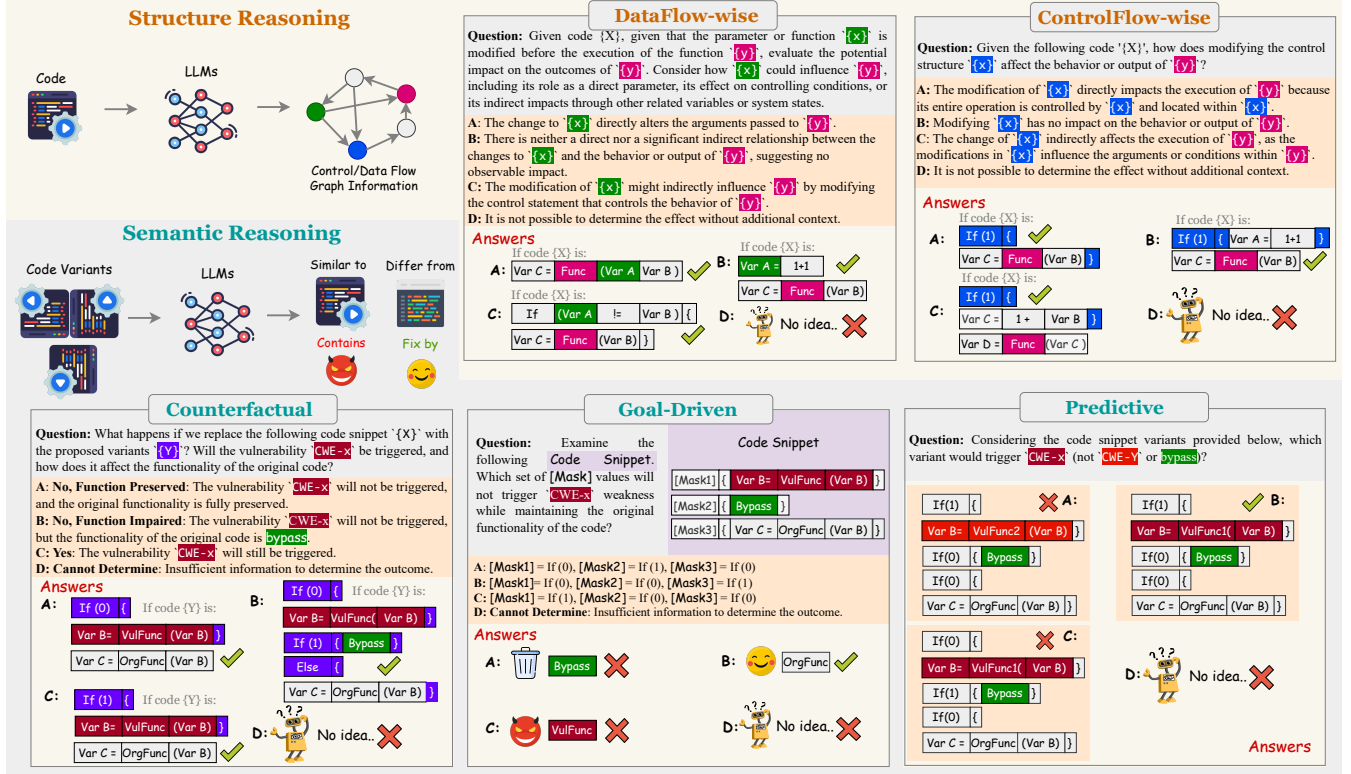


Figure 1: Overview of the SV-TRUSTEVAL-C Benchmark: This figure highlights question-answering templates for **Structure Reasoning**, divided into control-flow and data-flow segments, alongside **Semantic Reasoning**, encompassing Counterfactual, Goal-driven, and Predictive tasks. Each section illustrates various question types and potential answer scenarios, designed to assess LLMs' abilities to handle the diverse complexities of code semantics and structure.

3.1.1. Structure Reasoning. Structure Reasoning evaluates how accurately LLMs understand the relationships and interactions between code elements, focusing on both data flow and control flow, a critical capability for identifying and mitigating potential security threats. This category aligns with the task:

$$P(D | X) = \prod_{i=1}^M \prod_{j=1}^M P(D_{ij} | X), \quad (7)$$

where D_{ij} denotes the relationship between elements i and j (with $D_{ij} = 1$ indicating a connection, and 0 otherwise), M is the total number of elements, and X represents the code snippet. Specifically:

- **DataFlow-wise Reasoning:** Evaluates the model's understanding of how data moves through the code, essential for identifying vulnerabilities related to data handling.
- **ControlFlow-wise Reasoning:** Assesses the model's proficiency in analyzing the control flow of the program, vital for understanding how different parts of the code execute and interact.

These tasks examine the model's ability to discern how code elements correlate and how vulnerabilities can propagate through these interactions.

3.1.2. Semantic Reasoning. Semantic Reasoning evaluates an LLM's adaptability and understanding of changes in code semantics under various scenarios and transformations, ensuring security and functionality, and encompasses three sub-tasks: **Counterfactual** (predicting vulnerabilities when code is altered), **Goal-driven** (safely modifying code to meet specific aims), and **Predictive** (classifying code variants by their security impact). Formally, for the Counterfactual scenario, we consider:

$$P(V(f(X_s, t)) = 1 | f(X_s, t), X_s), \quad (8)$$

$t \in \{\text{safe, unsafe, impaired}\}$

where $V(X) = 1$ if X is vulnerable (otherwise 0), and $f(X_s, t)$ transforms X_s based on behavior t . We introduce three transformation types:

- **Safe Code Transformations:** Modifications that preserve functionality without adding vulnerabilities.
- **Unsafe Code Transformations:** Modifications that introduce new vulnerabilities into the code.
- **Impaired Code Transformations:** Alterations that keep the code vulnerability-free but diminish its original functionality.

These tasks assess the model's consistency and adaptability in analyzing vulnerabilities under semantic alterations, rein-

forcing the importance of understanding how code transformations affect security properties.

Goal-driven tasks evaluate an LLM’s ability to modify code to achieve a specific outcome without introducing vulnerabilities:

$$P(V(X) = 0, X \mid c), \quad (9)$$

where X is the modified code snippet to meet the goal, c is the given context or constraint, and $V(X) = 0$ means the modified code remains vulnerability-free. In these tasks:

- The model is prompted to insert or alter code to achieve specified goals (e.g., adding features or fixing bugs) while ensuring no new vulnerabilities are introduced.
- Code templates emphasize control statements that can either bypass, trigger, or prevent vulnerabilities.

By ensuring the resulting code remains secure and functional, these scenarios evaluate the model’s proficiency in context-aware vulnerability reasoning and code refinement. The joint probability formulation, $P(V(X) = 0, X \mid c)$, assesses both the safety of the modified code and its successful production under the given context c .

Predictive scenarios challenge the LLM to classify code variants based on whether they introduce, remove, or do not affect vulnerabilities and code functionality:

$$P(k \mid f(X)), \quad (10)$$

where k represents the code state or type of vulnerability, including non-vulnerable and potentially impaired states, and $f(X)$ is the code variants. These tasks:

- Provide code variants that may introduce new vulnerabilities, remove existing vulnerabilities, remain vulnerability-free but become functionally impaired, or have no effect on vulnerability.
- Require the LLM to accurately classify these code variants, distinguishing among various vulnerability states and the possibility of impaired functionality.

These Predictive tasks assess the model’s capability to distinguish among different vulnerabilities and its understanding of how code modifications influence security.

3.2. Structure-Oriented Variants Generator

This section describes the purpose of our Structure-Oriented Variants Generator, subsequently referred to as generator, and its role in generating our benchmark dataset. The generator contains a Flow Extractor and a Behaviour Simulator. Given a code snippet from an existing codebase, the generator strategically modifies the behavior of code variants into three categories: *safe* (preventing vulnerabilities), *unsafe* (triggering vulnerabilities), and *impaired* (bypassing the original function), while enhancing their structural complexity. It also scales the structural complexity to suit the benchmark questions, ensuring a diverse set of code variants.

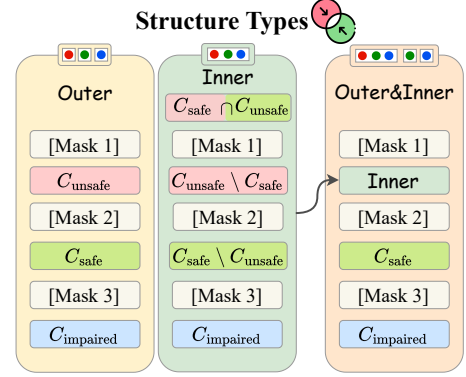


Figure 2: Overview of the Outer, Inner, and Outer&Inner structures. The *impaired* block prints or returns *None* and does not mimic the behavior of either C_{safe} or C_{unsafe} . Here, $C_{\text{safe}} \cap C_{\text{unsafe}}$ represents the common parts of both versions, while $C_{\text{safe}} \setminus C_{\text{unsafe}}$ denotes the unique parts of C_{safe} .

3.2.1. Flow Extractor. Given a code snippet C , our Flow Extractor constructs the data flow graph $\mathcal{G}_d = (\mathcal{V}_d, \mathcal{E}_d)$ and the control flow graph $\mathcal{G}_c = (\mathcal{V}_c, \mathcal{E}_c)$. Initially, we parse C into an Abstract Syntax Tree (AST) using a parser generator tool³ to extract code syntactic information. We then apply Depth-First Search (DFS) to traverse the AST, identifying code elements $v_i \in \mathcal{V}_d$ and $v_i \in \mathcal{V}_c$ such as variables, literals, and expressions, assuming different contexts for data and control flows respectively. Concurrently, we establish edges $e_{i,j}^d \in \mathcal{E}_d$ and $e_{i,j}^c \in \mathcal{E}_c$ to represent data dependencies between v_i^d and v_j^d through assignments and function calls, as well as control flows between elements v_i^c and v_j^c .

3.2.2. Behaviour Simulator. Given the graph data extracted by the Flow Extractor and the required behaviors of the code (*safe*, *unsafe*, and *impaired*), our behavior simulator manages the final behavior of the code by modifying the safe code C_{safe} and the unsafe code C_{unsafe} within the original code C . Initially, the simulator incorporates control flow variants from Juliet [32], introducing additional control branches within \mathcal{G}_c . This step tests if LLMs can maintain consistency with code variants that have similar syntax but different structures. As illustrated in Figure 2, the simulator then introduces Outer, Inner, and Outer&Inner structures to manage the code’s behavior, infusing C_{safe} , C_{unsafe} and our C_{impaired} with masked control statements into various structural configurations. Next, we fill the masked control statements to create code variants with the specified behavior. Finally, the Behavior Simulator generates multiple variants for each base code C , categorized by structure and behavior, to test LLM adaptability and accuracy.

3.3. Question Generation

After creating code variants from existing codebases, we develop questions for structural and semantic reasoning

3. <https://tree-sitter.github.io/tree-sitter/>

scenarios. Sample questions are provided in Appendix A.

3.3.1. DataFlow-wise Questions. For DataFlow-wise structure reasoning, we use the Flow Extractor to construct the data flow graph \mathcal{G}_d . We then evaluate each source node $v_i^d \in \mathcal{V}_d$ together with function-only target nodes v_j^d to determine whether modifying v_i^d influences v_j^d 's behavior. Each generated question falls into one of the following categories (see Figure 3):

- **A:** Ground truth is derived directly from \mathcal{E}_d , which may involve direct or multi-hop connections within \mathcal{G}_d .
- **B:** No connection in either \mathcal{G}_d or \mathcal{G}_c .
- **C:** Some connections are not explicitly shown in \mathcal{G}_d but are implied through control flow \mathcal{G}_c . For instance, if v_i^d is within an `if` ($v_i^d == 1$) block, and v_i^d is involved in the condition, it has an indirect impact. Similarly, if v_i^d triggers a `break` that prevents execution from reaching v_j^d , it also constitutes an indirect impact.
- **A&C:** In some cases, both A and C apply—namely, when v_i^d is connected to v_j^d in \mathcal{G}_d and also influences a control statement governing v_j^d .

The difficulty of each question increases with the number of hops between v_i^d and v_j^d .

3.3.2. ControlFlow-wise. For ControlFlow-wise structure reasoning, we use \mathcal{G}_c , obtained from the Flow Extractor, to enumerate all control statements $v_i^c \in \mathcal{V}_c$ linked to function-only target nodes v_j^c , to assess whether modifying v_i^c would impact v_j^c . Each generated question belongs to one of the following classes/options.

- **A:** The impact of control statement v_i^c on function v_j^c can be directly obtained from \mathcal{G}_c .
- **B:** No perceivable connections within either \mathcal{G}_d or \mathcal{G}_c .
- **C:** The control statement v_i^c directly or indirectly affects the arguments of v_j^c as specified within \mathcal{G}_d .

The difficulty scale for ControlFlow-wise questions is scaled by the number of hops from v_i^c to v_j^c within \mathcal{G}_c .

3.3.3. Counterfactual. In the Counterfactual semantic reasoning task, for each code snippet C , we generate code variants \mathcal{C} using the Structure-Oriented Variants Generator to construct a set of options. Each generated question falls into one of the following classes.

- **A:** The *safe* code variants do not trigger the vulnerability while preserving C 's original functionality.
- **B:** The *impaired* code variants avoid triggering the vulnerability but fail to maintain C 's functionality.
- **C:** The *unsafe* code variants trigger the vulnerability.

The difficulty level for Counterfactual scenarios is scaled based on the number of control flow injections and the specific code structures introduced by the generator.

3.3.4. Goal-driven. The goal-driven semantic reasoning questions assess LLMs' ability to fix vulnerabilities without altering code functionality. We generate code variants using the Structure-Oriented Variants Generator with masked control statements as conditions. LLMs are then tasked with

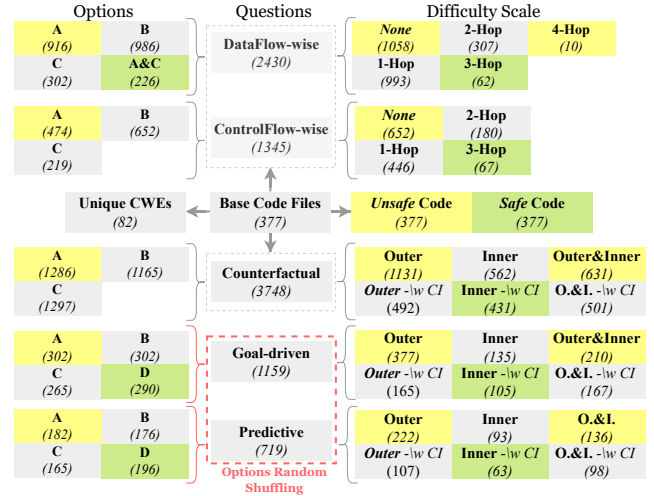


Figure 3: Overview of benchmark question distribution across two core capabilities, five question types, and multiple difficulty levels. The notation `-lw CI` indicates code variants with control flow injection. The term *None* indicates that there is no connection within the dataflow or control flow graph in the structure reasoning scenario. The question generation strategy is designed to ensure a uniform distribution across each base file, prioritizing high difficulty levels.

selecting the correct path to resolve the vulnerability while preserving the intended functionality. Answer options are generated using the Behavior Simulator and include: (1) **Bypass**, (2) **Resolve (Expected Behavior)**, and (3) **Trigger the Vulnerability**. The difficulty of each question scales with the complexity of the underlying code variant. Since the correct answer is always the same class (i.e., “Resolve”) in this task, we apply **random shuffling** of both the options and answer placements to prevent position bias.

3.3.5. Predictive. The Predictive semantic reasoning tasks also use code variants generated by the generator, challenging LLMs to identify which variants trigger specific vulnerabilities. This evaluation tests the models' ability to detect and accurately predict the presence of vulnerabilities, and to differentiate between various types of vulnerabilities. The answer options—(1) **Bypass**, (2) **Target CWE (Expected Behavior)**, and (3) **Different CWE**—follow the same **random shuffling** strategy as the Goal-driven questions. Difficulty scales with the complexity of each variant.

3.4. Benchmark Statistics

The SV-TRUSTEVAL-C benchmark, developed using the C programming language components from the Juliet Test Suite [32], includes 377 base files, each containing both a *safe* and an *unsafe* function, covering 82 distinct CWEs. Our generator produced a total of 1,297 *unsafe* and 1,286 *safe* compilable code variants. Utilizing these variants, we created 9,401 questions. The detailed questions' statistics are presented in Figure 3.

TABLE 2: Comparison of model performance based on accuracy. The **best performance** and **second-best performance** are highlighted to denote the top two scores, respectively. The \star symbol identifies models in the “instruct” version, which are specifically fine-tuned to enhance instruction-following capabilities. Additionally, we report the false positive rate (FPR) for misclassified “safe” instances (FPR^{safe}) in the baseline scenario. The \downarrow symbol indicates a decrease in model performance compared with Zero on the current task after introducing ICL, while the \uparrow symbol signifies an improvement in performance.

Models	Structure Reasoning						Semantic Reasoning						Base Scenario					
	DataFlow		ControlFlow		Average		Counterfactual		Goal-driven		Predictive		Average		Unsafe	Safe	Average	FPR ^{Safe}
	Zero	ICL	Zero	ICL	Zero	ICL	Zero	ICL	Zero	ICL	Zero	ICL	Zero	ICL				
>15B Param. Models																		
GPT-4-turbo	60.58	59.42↓	65.20	68.55↑	62.89	63.98↑	41.57	52.27↑	31.23	1.47↓	19.61	30.46↑	30.80	28.07↓	56.65	76.39	66.58	23.61
GPT-3.5-turbo	40.41	50.49↑	56.95	47.36↓	48.68	48.9↑	32.76	34.79↑	4.83	16.82↑	19.47	23.50↑	19.02	25.04↑	56.23	31.56	43.90	68.44
Llama3.1-405B*	63.17	61.40↓	73.98	76.88↑	68.58	69.14↑	45.17	62.57↑	2.67	29.08↑	28.93	43.95↑	25.59	45.20↑	87.00	24.14	55.57	75.86
<15B Param. Models																		
Llama3.1-8B*	45.80	54.57↑	54.94	52.64↓	50.37	53.60↑	36.87	38.42↑	1.12	23.04↑	24.90	27.54↑	20.96	29.67↑	55.17	66.31	60.74	33.69
Llama3-8B*	46.63	49.79↑	44.68	48.48↑	45.66	49.14↑	34.82	35.33↑	0.95	6.04↑	23.92	29.49↑	19.9	23.62↑	53.85	60.74	57.29	39.26
CodeLlama-7B*	38.40	37.70↓	34.65	29.22↓	36.52	33.46↓	34.26	34.61↑	0.00	33.74↑	24.48	29.07↑	19.58	32.47↑	100.00	0.00	50.00	100.00
CodeLlama-13B*	23.74	37.57↑	36.43	34.35↓	30.08	35.96↑	34.47	34.61↑	24.68	23.38↓	22.67	26.01↑	27.27	28.02↑	100.00	0.00	50.00	100.00
Gemma-7B*	36.17	33.87↓	27.88	29.44↑	32.02	31.66↓	33.11	33.62↑	28.90	41.59↑	24.48	26.70↑	28.83	33.97↑	99.73	0.53	50.13	99.47
CodeGemma-7B*	37.98	40.00↑	29.29	39.85↑	33.64	39.92↑	35.01	35.17↑	0.09	22.78↑	21.56	16.97↓	18.89	24.97↑	98.14	0.27	49.20	99.73
CodeQwen1.5-7B*	35.76	38.02↑	34.87	35.24↑	35.32	36.63↑	33.11	32.84↓	33.74	22.69↓	26.70	21.70↓	31.18	25.74↓	100.00	0.00	50.00	100.00
Mixtral-7B*	29.67	51.77↑	37.70	52.49↑	33.68	52.13↑	31.78	32.90↑	0.00	9.66↑	29.62	30.46↑	20.47	24.34↑	3.98	0.00	1.99	100.00

4. Experiments

4.1. Experimental Setups

We evaluate eleven of the most popular and representative large language models (LLMs), including GPT-4-turbo [2], GPT-3.5-turbo [1], Llama3.1 [63] in both 405B and 8B versions, Llama3 [64], CodeLlama [4] in 13B and 7B versions, Gemma [65], CodeGemma [66], CodeQwen [6], and Mixtral [67]. We excluded the latest LLM, GPT-o1 [68], due to budget constraints and prohibitively high time consumption. All models have undergone specialized pre-training in the code domain and are available in the “instruct” version, which is fine-tuned to follow prompted instructions, making them well-suited for our benchmark design. We conduct all model inferences at a temperature of **zero** to ensure more deterministic answers and set a maximum output length of 50 tokens, as only the selected option is required. All other inference hyperparameters are set to their default values for each LLM. Furthermore, we establish conversation-independent threads for each question in our experiment to eliminate potential information leakage during the question-answer (QA) process.

4.1.1. Inference Mode. We utilize both zero-shot inference (Zero) and in-context learning (ICL) approaches [69], [70], [71] to comprehensively evaluate LLMs. In the zero-shot setup, models generate responses based solely on the input question without additional examples, allowing us to assess their inherent understanding and reasoning abilities. In contrast, the in-context learning approach provides models with specific prompts that demonstrate question-and-answer patterns along with corresponding explanations for each answer. This contextual information guides their responses, enhancing their ability to learn from context and more accurately follow the question. The prompts used for in-context learning include a few sample question-response pairs following the template shown in Figure 4. All ques-

In-context Learning Prompt

Instruction: {Assign a code security expert persona with guidelines to answer the question.}

Context:

- **Question(s):** {Insert the demo question based on structural or semantic reasoning.}
- **Choices:** {Provide four possible answers based on the code scenario.}
- **Answer(s):** {Specify the correct answer, e.g., “B”.}
- **Explanation(s):** {Provide a detailed explanation justifying why the chosen answer is correct.}

Input Question: {Insert the selected question.}

Output: {LLM-generated answer.}

Figure 4: Prompt template for In-context learning pipeline

tions, answer choices, correct answers, and explanations are carefully crafted by human domain experts and subsequently refined by GPT-o1 to ensure accuracy and reduce instructional errors.

4.1.2. Label Masking. To avoid label leakage from the original Juliet test suite, we applied label masking to all code snippets, including: 1) removing annotations directly referencing vulnerabilities; 2) replacing vulnerability-specific function names with generic ones (e.g., “Sample_func()”); and 3) swapping variable names or tokens that suggest vulnerability (e.g., “good” or “bad”) with neutral terms such as “cat” or “apple.”

4.1.3. Question Design. To develop an effective question-answering template that ensures clear comprehension by LLMs, we manually crafted seven distinct prompts for each question. We then employed GPT-4 as an automatic evaluator [72] to select the most suitable format for each QA.

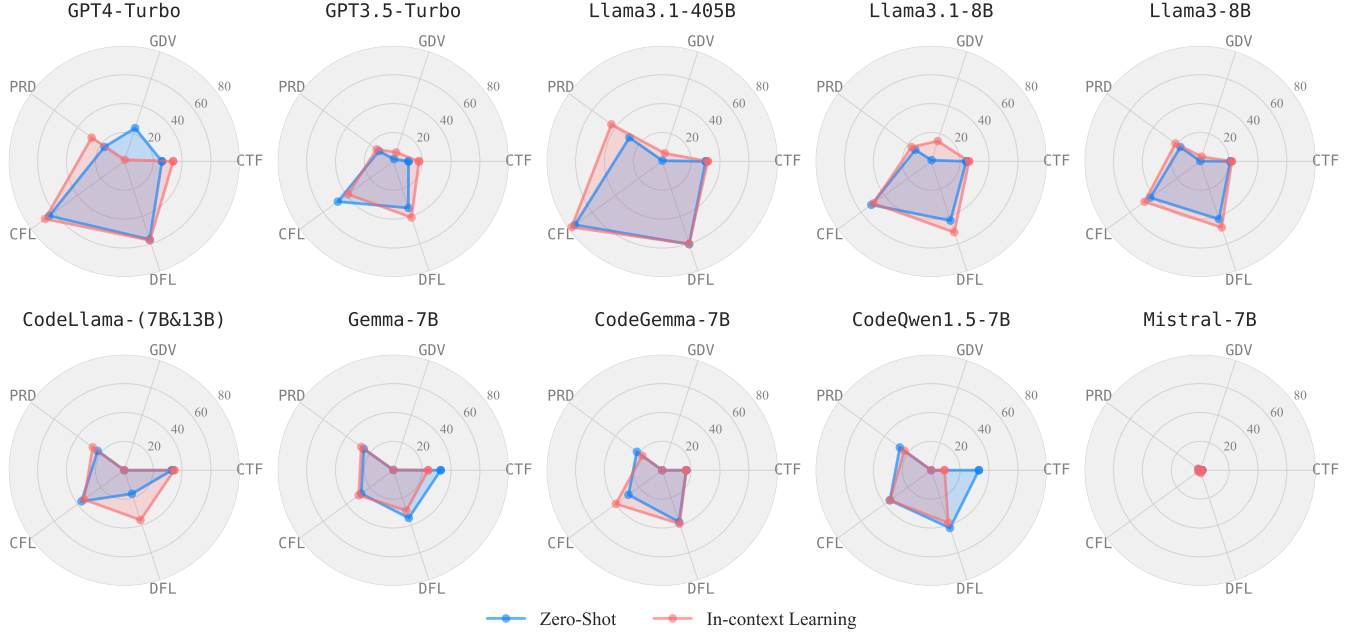


Figure 5: Overall consistency scores of LLMs on SV-TRUSTEVAL-C. For CodeLlama, the average consistency score across the 7B and 13B versions is reported due to their high similarity in performance. Abbreviations: **DFL** (DataFlow-wise questions), **CFL** (ControlFlow-wise questions), **CTF** (Counterfactual questions), **GDV** (Goal-driven questions), **PRD** (Predictive questions).

Additionally, to verify the absence of misleading syntax or semantics, we manually reviewed the intermediate explanations for 50 randomly selected QAs from each question type, ensuring that the LLMs’ responses were well-aligned with the QAs.

4.2. Main Results

As shown in Table 2, while LLMs generally require further improvement in structure reasoning, models with over 15 billion parameters significantly outperform their smaller counterparts. Notably, the recently released Llama3.1B-405B performs better in both data flow and control flow dimensions, achieving average scores of 68.58% in zero-shot and 69.14% in in-context learning, compared to GPT-4.

However, in semantic reasoning scenarios, most LLMs perform poorly, with average scores falling below 32% for zero-shot and 46% for in-context learning modes—particularly in Goal-driven tasks that require targeted vulnerability fixes and Predictive scenarios demanding extensive domain knowledge. We also conducted a baseline assessment by classifying code from the original Juliet Test Suite as *safe* or *unsafe* using a QA format with the options: A) Vulnerable, B) Non-Vulnerable, C) Do Not Know. The results presented in Table 2 indicate that most models perform unsatisfactorily, especially given that the code originates from synthetically generated datasets like Juliet. This suggests that the models are inadequately pre-trained in code vulnerability detection.

Additionally, several models—including CodeLlama, CodeQwen, and Gemma—exhibit high false positive rates⁴, erroneously classifying *safe* code as *unsafe* in nearly 100% of cases, as illustrated in Table 2. This raises concerns about the efficacy of LLMs in code analysis.

Finally, we observe that in-context learning can enhance models’ performance in most cases. Specialized LLMs such as CodeLlama and CodeGemma, which are less effective at understanding complex natural language scenarios, benefit significantly from this approach. Conversely, general-purpose LLMs like Mistral, which have less pre-training in the code domain, often struggle with code contexts; here, in-context learning serves as a bridge, helping these models adapt to intersecting scenarios. Similar improvements are also observed in models like Llama3.1B-405B and GPT-3.5. However, GPT-4 exhibits a dramatic decline in performance in Goal-driven scenarios with in-context learning, indicating that different models may respond variably to this technique and may require specific customization for vulnerability analysis.

4.2.1. Consistency Analysis. Our consistency analysis assesses how models maintain reliable vulnerability evaluations across different scenarios by focusing on two distinct reasoning tasks: **structure reasoning** and **semantic reasoning**. In the structure reasoning scenario, we transition

4. For FPR^{safe} , false positives are instances where *safe* code is incorrectly classified as *unsafe* or as an unknown category; true negatives are cases where code is correctly classified as *safe*.

from base code vulnerability analysis to tasks that require understanding the code’s structure, specifically DataFlow-wise (DFL) and ControlFlow-wise (CFL) analyses. Grasping the relationships among code components is essential for accurate vulnerability detection. If a model can identify vulnerabilities but fails to differentiate variable relationships within the source code, it indicates that its capabilities are driven by pre-trained patterns rather than genuine logical code analysis [27], [28]. Conversely, in the semantic reasoning scenario, we investigate whether LLMs can maintain consistent vulnerability assessments across diverse contexts by addressing the logical equivalence of statements such as $A = B$ and $B = A$ [73]. Using the base scenario as a benchmark, we evaluate if consistency is preserved in corresponding Counterfactual (CTF), Goal-driven (GDV), and Predictive (PRD) scenarios. This means that if LLMs accurately assess vulnerabilities in the base scenario, they should provide reliable analyses for any code variants within the same semantic context. For each question derived from the base scenario, we define the consistency scores as:

$$\text{Cons}_{\text{DFL}} = \frac{\sum_{i=1}^{N_{\text{DFL}}} \mathbb{I}(C_{\text{base}}^i = 1 \wedge C_{\text{DFL}}^i = 1)}{N_{\text{DFL}}} \quad (11)$$

$$\text{Cons}_{\text{CFL}} = \frac{\sum_{i=1}^{N_{\text{CFL}}} \mathbb{I}(C_{\text{base}}^i = 1 \wedge C_{\text{CFL}}^i = 1)}{N_{\text{CFL}}} \quad (12)$$

$$\text{Cons}_{\text{CTF}} = \frac{\sum_{i=1}^{N_{\text{CTF}}} \mathbb{I}(C_{\text{CTF}}^i = 1 \wedge C_{\text{base}}^i = 1)}{N_{\text{CTF}}} \quad (13)$$

$$\text{Cons}_{\text{GDV}} = \frac{\sum_{i=1}^{N_{\text{GDV}}} \mathbb{I}(C_{\text{GDV}}^i = 1 \wedge C_{\text{safe}}^i = 1)}{N_{\text{GDV}}} \quad (14)$$

$$\text{Cons}_{\text{PRD}} = \frac{\sum_{i=1}^{N_{\text{PRD}}} \mathbb{I}(C_{\text{PRD}}^i = 1 \wedge C_{\text{unsafe}}^i = 1)}{N_{\text{PRD}}} \quad (15)$$

where:

- $\mathbb{I}(\cdot)$ is the indicator function, which equals 1 if the condition inside is true, and 0 otherwise.
- C_{safe}^i and C_{unsafe}^i denote the correctness indicators for *safe* and *unsafe* classifications in the base scenario for the i -th case.
- $C_{\text{base}}^i = \mathbb{I}(C_{\text{safe}}^i = 1 \wedge C_{\text{unsafe}}^i = 1)$ is the correctness indicator for the base scenario for the i -th case.
- C_{DFL}^i and C_{CFL}^i are the correctness indicators for DataFlow-wise and ControlFlow-wise scenarios for the i -th case, respectively.
- C_{CTF}^i is the correctness indicator for the Counterfactual scenario of the i -th case.
- C_{GDV}^i and C_{PRD}^i are the correctness indicators for Goal-driven and Predictive scenarios for the i -th case.
- N_{DFL} , N_{CFL} , N_{CTF} , N_{GDV} , and N_{PRD} represent the total number of cases in the DataFlow-wise, ControlFlow-wise, Counterfactual, Goal-driven, and Predictive scenarios, respectively.

Observations: As shown in Figure 5, Mistral-7B exhibits an almost negligible consistency score in all scenarios, largely due to its weak performance in the Base scenarios. By contrast, GPT-4 and Llama3.1-405B substantially

outperform other models in both Cons_{CFL} and Cons_{DFL} , reflecting a stronger ability to track dependencies between control statements, variables, and functions within the code. In comparison, the remaining LLMs achieve under 60% consistency in structure reasoning scenarios, indicating that effectiveness in vulnerability analysis alone does not necessarily translate into a robust understanding of these code elements. In the semantic reasoning scenario, the consistency scores for all LLMs are notably low, averaging below 50%. This suggests that, despite some LLMs being pre-trained in the code domain, they struggle to maintain consistent analysis across variant scenarios, even when these variants incorporate the original code snippet. Particularly in Goal-driven scenarios, all LLMs exhibit minimal consistency. For instance, CodeQwen1.5 and Gemma, despite achieving the highest scores of 33.74% and 41.59% respectively in zero-shot and in-context learning modes for Goal-driven scenarios, record a 0% Cons_{GDV} . This indicates that their success does not reflect a genuine understanding of code vulnerabilities. Similarly, models like Llama3 and Llama3.1 have nearly zero consistency scores, highlighting a failure to effectively apply their knowledge of vulnerabilities in practical scenarios. While introducing in-context learning generally improves Cons_{GDV} and Cons_{PRD} , the enhancements are insufficient, and further efforts are required to bolster LLMs’ consistency in semantic reasoning tasks. This underscores the necessity for more specialized training and fine-tuning to enable LLMs to better understand and analyze code vulnerabilities across diverse and complex scenarios.

5. Analysis

5.1. Effects of Difficulty

As detailed in Table 3, we evaluated the performance of various LLMs across different difficulty levels within each scenario. The analysis encompasses three primary scenarios: Counterfactual, Goal-driven, and Predictive.

5.1.1. Structure Reasoning. In the structure reasoning tasks for DataFlow and ControlFlow scenarios, the LLMs’ performance generally falls into two categories:

High Performance at the “None” Level, but Struggles with Complexity: Models such as Llama3.1 and GPT-4 perform well at detecting when no connection exists between two code elements (the “None” level), but they face challenges with more complex relationships. However, their performance declines significantly in more complex scenarios that involve connections of three or more hops. This indicates a robust foundational understanding that diminishes as the reasoning required becomes more intricate.

Low Performance at the “None” Level, but Improved with Increased Complexity: Conversely, models like CodeQwen, CodeLlama, and CodeGemma exhibit average performance below 30% at the “None” level. Surprisingly, these models perform better in scenarios requiring connections of ≥ 3 hops. This counterintuitive outcome arises because these LLMs tend to assume connectivity

TABLE 3: LLMs’ performance across difficulty levels by question type. The symbol \odot represents unknown parameter scales; terms **D-Flow** and **C-Flow** refer to DataFlow-wise and ControlFlow-wise questions, respectively; **O.&I.** signifies the Outer&Inner structure, **#** denotes the number of questions, and **CI** indicates code variants with control flow injection. “None” indicates that there are no direct or indirect connections between the target code elements within the code graphs. The \downarrow symbol denotes a decrease in model performance on the current task compared to Zero after introducing ICL, while \uparrow signifies an improvement, and \uparrow indicates no change in performance.

LLMs Param. Mode	GPT-4 0		GPT-3.5 0		Llama3.1 405B		Llama3.1 8B		Llama3 8B		CodeLlama 13B		CodeLlama 7B		Gemma 7B		CodeGemma 7B		CodeQwen1.5 7B		Mixtral 7B		#	
	Zero	ICL	Zero	ICL	Zero	ICL	Zero	ICL	Zero	ICL	Zero	ICL	Zero	ICL	Zero	ICL	Zero	ICL	Zero	ICL	Zero	ICL		
D-Flow	Structure Reasoning																							
	None	81.66	82.04 \uparrow	8.03	51.23 \uparrow	81.76	80.91 \downarrow	86.48	77.32 \downarrow	51.89	87.81 \uparrow	2.08	0.00 \downarrow	4.63	26.47 \uparrow	1.70	2.74 \uparrow	0.00	1.89 \uparrow	14.56	1.23 \downarrow	5.58	75.99 \uparrow	1058
	1-Hop	52.92	54.83 \uparrow	72.71	62.74 \downarrow	57.85	59.57 \uparrow	23.01	47.73 \uparrow	50.91	30.92 \downarrow	70.69	74.02 \uparrow	31.42	53.32 \uparrow	68.08	57.60 \downarrow	73.72	74.92 \uparrow	69.13	73.82 \uparrow	40.38	38.32 \downarrow	993
	2-Hop	32.25	26.87 \downarrow	74.43	39.09 \downarrow	36.32	35.67 \downarrow	11.89	33.22 \uparrow	23.13	17.26 \downarrow	74.10	76.38 \uparrow	46.09	44.79 \downarrow	64.33	53.91 \downarrow	76.06	75.90 \downarrow	73.94	76.38 \uparrow	40.23	32.25 \downarrow	307
	3-Hop	51.61	50.00 \downarrow	79.03	67.74 \downarrow	53.23	45.16 \downarrow	34.68	49.19 \uparrow	58.06	33.87 \downarrow	70.16	79.84 \uparrow	42.74	70.97 \uparrow	79.84	67.74 \downarrow	79.84	78.23 \downarrow	79.03	79.84 \uparrow	47.58	54.84 \uparrow	62
4-Hop	30.00	30.00 \uparrow	100.00	80.00 \downarrow	0.00	10.00 \uparrow	40.00	70.00 \uparrow	30.00	0.00 \downarrow	90.00	100.00 \uparrow	0.00	100.00 \uparrow	100.00	100.00 \uparrow	100.00	100.00 \uparrow	100.00	100.00 \uparrow	0.00	30.00 \uparrow	10	
C-Flow	None	73.77	80.52 \uparrow	52.76	49.23 \downarrow	87.27	82.67 \downarrow	90.80	94.63 \uparrow	71.32	93.71 \uparrow	2.30	0.92 \downarrow	3.07	8.74 \uparrow	0.00	0.00 \downarrow	0.15	14.72 \uparrow	0.15	0.77 \uparrow	28.07	70.86 \uparrow	652
	1-Hop	81.39	81.84 \uparrow	83.18	61.21 \downarrow	87.67	88.12 \uparrow	29.37	18.83 \downarrow	19.73	5.38 \downarrow	95.96	79.37 \downarrow	97.31	80.72 \downarrow	30.04	42.38 \uparrow	49.33	46.86 \downarrow	100.00	98.88 \uparrow	40.13	43.05 \uparrow	446
	2-Hop	16.11	13.89 \downarrow	24.44	20.00 \downarrow	14.44	32.22 \uparrow	7.78	1.67 \downarrow	22.78	1.67 \downarrow	12.78	15.56 \uparrow	20.00	15.00 \downarrow	96.67	92.22 \downarrow	67.78	98.90 \uparrow	15.56	15.56 \uparrow	63.33	15.00 \downarrow	180
	3-Hop	5.97	10.45 \uparrow	10.45	10.45 \uparrow	13.43	65.67 \uparrow	2.99	5.97 \uparrow	10.45	20.90 \uparrow	0.00	7.46 \uparrow	0.00	26.87 \uparrow	100.00	71.64 \downarrow	76.12	79.10 \uparrow	0.00	0.00 \downarrow	46.27	37.31 \downarrow	67
Counterfactual	Semantic Reasoning																							
	Outer	43.59	58.00 \uparrow	35.46	34.84 \downarrow	51.37	68.52 \uparrow	37.22	38.11 \uparrow	32.71	34.31 \uparrow	33.24	33.33 \uparrow	33.33	33.33 \uparrow	32.54	35.37 \uparrow	33.42	34.66 \uparrow	32.54	31.56 \downarrow	33.51	36.07 \uparrow	1131
	-w CI	41.46	50.00 \uparrow	35.77	32.32 \downarrow	38.62	55.28 \uparrow	36.99	34.15 \downarrow	31.10	32.32 \downarrow	32.32	32.93 \uparrow	33.13	32.93 \downarrow	32.93	34.55 \uparrow	35.57	32.72 \downarrow	32.93	32.52 \downarrow	36.59	31.91 \downarrow	492
	Inner	45.73	53.38 \uparrow	38.79	40.04 \downarrow	44.48	62.81 \uparrow	40.04	39.32 \downarrow	37.19	37.90 \uparrow	37.54	37.54 \uparrow	37.54	37.54 \uparrow	32.38	31.67 \downarrow	38.43	37.72 \downarrow	32.38	36.65 \uparrow	37.72	37.19 \downarrow	562
	-w CI	42.69	48.72 \uparrow	37.59	39.68 \downarrow	46.87	61.95 \uparrow	38.28	42.92 \uparrow	39.21	39.21 \uparrow	39.21	39.21 \uparrow	38.75	39.21 \uparrow	37.12	32.02 \downarrow	36.89	38.52 \uparrow	37.12	33.41 \downarrow	37.35	34.11 \downarrow	431
O.&I.	36.77	50.87 \uparrow	35.82	32.65 \downarrow	43.58	62.92 \uparrow	34.07	38.19 \uparrow	36.61	35.18 \downarrow	33.91	33.44 \downarrow	33.28	33.44 \downarrow	32.49	33.28 \downarrow	32.96	32.17 \downarrow	32.49	32.33 \downarrow	35.66	28.05 \downarrow	631	
-w CI	37.52	45.11 \uparrow	32.93	32.53 \downarrow	38.92	56.06 \uparrow	34.73	38.72 \uparrow	34.53	34.53 \uparrow	32.93	33.33 \uparrow	32.73	33.33 \uparrow	32.73	32.73 \uparrow	35.13	36.73 \uparrow	32.73	31.94 \downarrow	35.73	28.14 \downarrow	501	
Goal-driven	Outer	66.31	2.39 \downarrow	1.33	19.39 \downarrow	2.39	27.59 \uparrow	0.00	24.67 \uparrow	0.00	7.43 \downarrow	0.00	34.22 \uparrow	4.51	20.69 \uparrow	5.31	35.54 \uparrow	0.00	23.34 \uparrow	34.22	22.55 \uparrow	0.00	8.49 \uparrow	377
	-w CI	0.61	0.00 \downarrow	21.21	19.39 \downarrow	7.88	37.58 \uparrow	3.64	22.42 \uparrow	0.00	4.24 \downarrow	0.00	29.09 \uparrow	54.55	24.85 \downarrow	46.06	42.42 \downarrow	0.00	27.27 \uparrow	36.36	27.88 \uparrow	0.00	16.36 \uparrow	165
	Inner	55.56	2.22 \downarrow	2.22	17.04 \downarrow	0.00	18.52 \uparrow	0.00	23.70 \uparrow	0.00	8.15 \downarrow	0.00	39.26 \uparrow	11.11	22.22 \uparrow	7.41	45.93 \uparrow	0.00	20.74 \uparrow	25.19	23.70 \uparrow	0.00	7.41 \uparrow	135
	-w CI	4.76	1.90 \downarrow	1.90	18.10 \downarrow	0.00	33.33 \uparrow	3.81	26.67 \uparrow	10.48	1.90 \downarrow	0.00	31.43 \uparrow	41.50	24.76 \downarrow	30.48	41.90 \uparrow	0.00	22.86 \uparrow	34.29	25.71 \uparrow	0.00	3.81 \uparrow	105
	O.&I.	13.81	1.90 \downarrow	1.43	18.57 \downarrow	1.43	30.00 \uparrow	0.00	18.57 \uparrow	0.00	4.76 \downarrow	0.00	34.29 \uparrow	21.43	23.81 \uparrow	33.81	41.62 \uparrow	0.00	21.90 \uparrow	32.38	15.71 \uparrow	0.00	7.62 \uparrow	210
-w CI	1.20	0.00 \downarrow	4.79	15.57 \uparrow	3.59	28.74 \uparrow	1.80	22.75 \uparrow	0.00	7.19 \downarrow	0.00	33.53 \uparrow	44.91	27.54 \downarrow	55.85	43.11 \downarrow	0.60	19.76 \uparrow	38.32	23.95 \uparrow	0.00	13.77 \uparrow	167	
Predictive	Outer	19.37	25.23 \uparrow	16.22	25.68 \uparrow	27.93	38.74 \uparrow	31.08	28.38 \uparrow	18.02	27.48 \uparrow	23.87	31.08 \uparrow	23.42	28.38 \uparrow	32.88	25.23 \downarrow	24.32	9.46 \downarrow	27.03	24.32 \downarrow	31.53	27.48 \uparrow	222
	-w CI	15.89	29.91 \uparrow	18.69	24.30 \uparrow	26.17	48.60 \uparrow	18.69	25.23 \uparrow	28.97	28.97 \uparrow	23.36	23.36 \uparrow	26.17	22.43 \downarrow	17.76	31.78 \uparrow	16.82	24.30 \uparrow	28.04	17.76 \uparrow	30.84	39.25 \uparrow	107
	Inner	26.88	34.41 \uparrow	23.66	15.05 \downarrow	32.26	47.31 \uparrow	19.35	25.81 \uparrow	22.58	24.73 \uparrow	31.18	29.03 \uparrow	22.58	32.26 \uparrow	18.28	27.96 \uparrow	26.88	17.20 \downarrow	24.73	25.81 \uparrow	24.73	35.48 \uparrow	93
	-w CI	25.40	34.92 \uparrow	22.22	22.22 \downarrow	30.16	42.86 \uparrow	23.81	26.98 \uparrow	26.98	34.92 \uparrow	25.40	34.92 \uparrow	30.16	25.40 \downarrow	34.92	30.16 \downarrow	23.81	20.63 \downarrow	36.51	17.46 \downarrow	33.33	30.16 \downarrow	63
	O.&I.	19.12	32.35 \uparrow	19.12	25.74 \uparrow	30.88	46.32 \uparrow	24.26	27.94 \uparrow	26.47	28.68 \uparrow	24.26	27.21 \uparrow	17.65	30.88 \uparrow	23.53	18.38 \downarrow	19.12	21.32 \uparrow	25.74	22.79 \uparrow	27.94	27.94 \uparrow	136
-w CI	14.29	33.67 \uparrow	22.45	24.49 \uparrow	27.55	44.90 \uparrow	24.49	29.59 \uparrow	27.55	37.76 \uparrow	21.43	29.59 \uparrow	20.41	19.39 \downarrow	19.39	37.76 \uparrow	18.37	19.39 \uparrow	22.45	26.53 \uparrow	32.65	37.76 \uparrow	98	

among all code elements within a context, thereby increasing the likelihood of selecting correct answers. Specifically, explicit and implicit connections account for 50% of the options (Options A and C), with some scenarios allowing both A and C to be simultaneously correct.

Despite these patterns, the ability to identify that two code elements are unconnected remains crucial, as it demonstrates the LLMs’ capacity for logical and critical analysis of relationships between code elements. Introducing in-context learning significantly improved the “None” level performance for structure reasoning tasks in most cases. For instance, Mixtral showed a dataflow-wise performance increase from 5.58% to 75.99%, and a controlflow-wise increase from 28.07% to 70.86%. However, this enhancement was not consistently observed across other difficulty levels for most LLMs. This inconsistency suggests that inherent code structure reasoning abilities require more sophisticated training methodologies and advanced model architectures to effectively handle varying levels of complexity.

5.1.2. Vulnerability Reasoning. Our evaluation across various scenarios reveals the following key insights:

- **Model Size and Complexity Handling:** Larger models (e.g., GPT-4, Llama3.1) typically perform better in Counterfactual and Goal-driven tasks, especially with complex code variants. However, they show comparatively weaker performance on vulnerability-resolving tasks, suggesting a need for more advanced reasoning capabilities.
- **Impact of In-Context Learning:** In-context learning often leads to performance gains for models such as CodeLlama, CodeGemma, and Llama3.1-serious,

particularly in Goal-driven scenarios. Conversely, models like CodeQwen1.5 exhibit consistent declines, highlighting varying sensitivities to in-context prompts and emphasizing the need for tailored prompt adaptation. These mixed outcomes underscore that model-specific strategies are crucial.

- **Impact of Control Flow Injection:** Most LLMs experience performance degradation after control flow injection (-w CI), indicating that even minimal syntactic perturbations, without altering semantics, can adversely affect performance in diverse scenarios.

Counterfactual Scenarios. Larger models such as GPT-4 and Llama3.1 consistently outperform smaller open-source counterparts (e.g., CodeQwen, Mistral), though their overall performance wanes as code complexity grows. In-context learning generally enhances the capabilities of GPT-4 and Llama3.1, whereas it negatively impacts CodeQwen and Mistral. This discrepancy suggests that architectural differences or training data variations impact how models respond to additional contextual information.

Goal-driven Scenarios. Performance in identifying correct fixes varies considerably with task difficulty: while GPT-4 performs well at simpler fixes in a zero-shot setting, CodeGemma outperforms it on more complex variants, reaching 75.45% accuracy. The introduction of in-context learning yields a dual effect—it significantly degrades performance for GPT-4 and CodeQwen, yet consistently enhances it for open-source models like Llama3.1-405B, CodeLlama-7B, CodeGemma, and Mistral—highlighting the need for tailored adaptations to optimize in-context learning.

Predictive Scenarios. The Llama-series models demon-



Figure 6: Distribution of choices made by all LLMs across all question scenarios, with the **shaded portions** of each bar showing how much the predicted option exceeds its corresponding ground truth.

strate robust capability in identifying target vulnerabilities, reflecting strong security-domain pre-training. In-context learning exerts mixed influences: models such as GPT-4 and Llama-series display consistent improvements, whereas CodeQwen1.5 and certain Gemma models face performance drops. These variations stress the significance of model-specific adaptation to capitalize on domain knowledge. Additionally, arbitrary prompts can degrade performance, mirroring the challenges noted in Goal-driven tasks. Hence, optimizing both prompt design and adaptation tech-

niques remains pivotal for maximizing the effectiveness of LLMs in vulnerability analysis.

5.2. Behavior Distribution

To further study the behavior of LLMs, we have detailed the distribution of choices made by LLMs during their evaluation, categorizing the options for each question into four types. In the DataFlow and ControlFlow-wise scenarios, the categories are: “Direct C.”, indicating direct connections between code elements in the graph; “Not C.”, denoting no

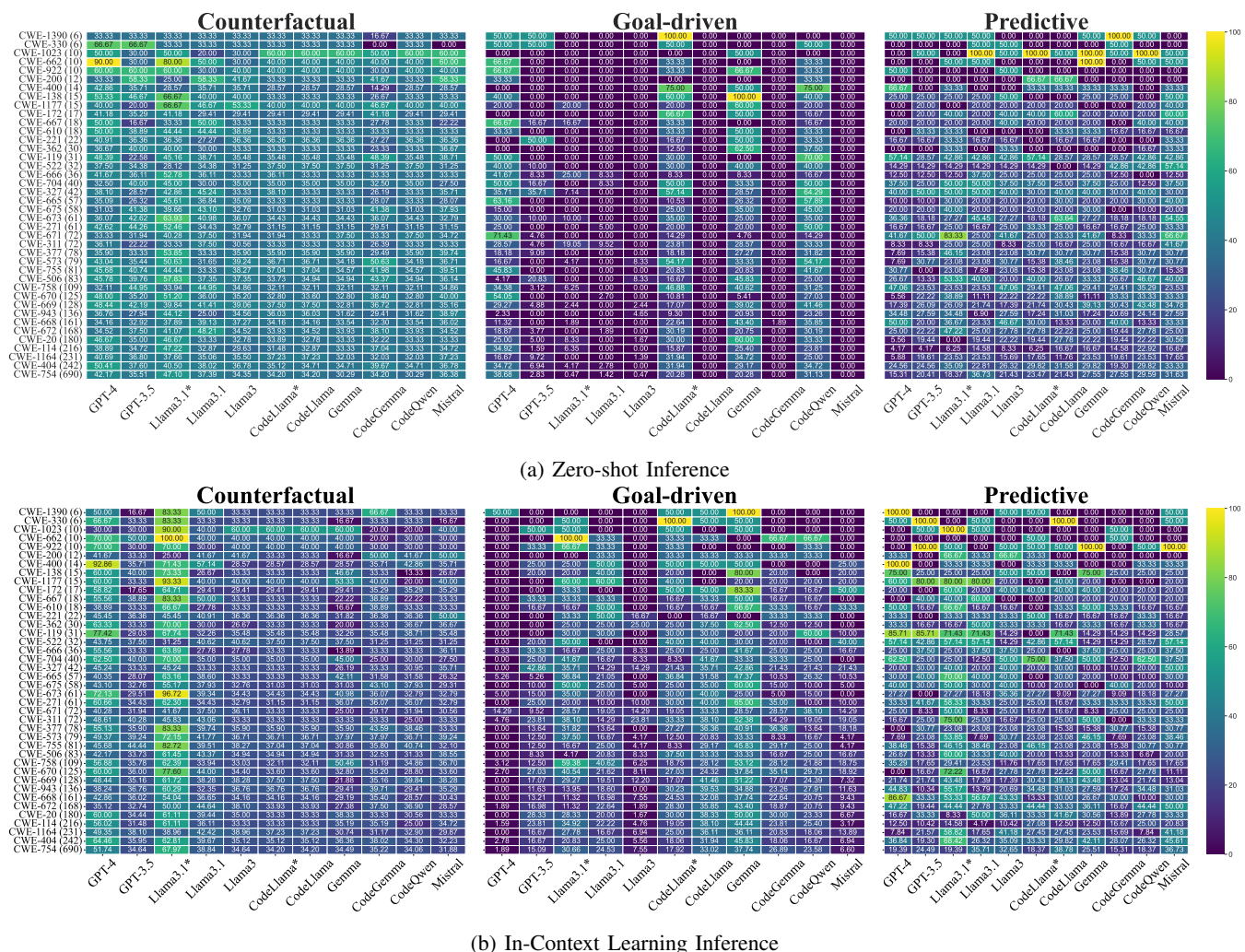


Figure 7: Performance of LLMs across various question types for **class-level** CWEs under different inference modes. Models marked with \star have larger parameter sizes (e.g., Llama3.1 \star denotes Llama3.1–405B). The number in brackets beside each CWE indicates the total number of associated questions.

significant connections; “Indirect C.”, for indirect connections as described in our method section; and “Unknown”, where the LLM is unable to respond. For the Counterfactual and Goal-driven scenarios, the responses are classified as “Safe”, indicating that the current code variants are secure and do not trigger vulnerabilities; “Bypass”, where the code neither triggers vulnerabilities nor executes the intended function; and “Unsafe”, where the code variants do activate vulnerabilities. Lastly, in the Predictive scenario, “Others” refers to CWEs specifically addressed in the question, and “Bypass” represents variants that circumvent the targeted CWEs. Our key findings are as follows:

- 1) **Pattern Matching Over Logic:** Our study reveals that LLMs in vulnerability analysis primarily rely on pattern matching based on pre-trained knowledge, rather than on logical analysis of the code.
- 2) **Need for Customized Approaches:** Effectively ad-

ressing vulnerabilities with LLMs requires more than general in-context learning; it demands advanced prompt engineering or scenario-specific fine-tuning.

- 3) **Need for Domain Adaption:** LLMs often struggle to differentiate between various CWEs, underscoring a critical need for domain adaptation to enhance their effectiveness in vulnerability analysis.

Specifically, as illustrated in Figure 6, in the Dataflow-wise and Controlflow-wise scenarios, we observe that LLMs like GPT-3.5, CodeLlama, CodeQwen and Gemma predominantly choose options classified as “Direct C.” and “Indirect C.” This suggests these LLMs assume that all code elements in a given piece of code are interconnected. These findings align with our analysis in Section 5.1, confirming that such LLMs lack rigorous logical analysis capabilities to understand contextual relationships in code. Instead, their vulnerability analysis tends to rely heavily on pattern matching, a conclusion supported by a substantial body of existing

TABLE 4: Ablation study on Llama3.1-8B with varying inference temperatures. **Temp.** denotes the temperature setting.

Temp.	Structure Reasoning						Semantic Reasoning								Base Scenario			
	DataFlow		ControlFlow		Average		Counterfactual		Goal-driven		Predictive		Average		Unsafe	Safe	Average	FPR ^{Safe}
	Zero	ICL	Zero	ICL	Zero	ICL	Zero	ICL	Zero	ICL	Zero	ICL	Zero	ICL				
0.0	45.80	54.57	54.94	52.64	50.37	53.60	36.87	38.42	1.12	23.04	24.90	27.54	20.96	29.67	55.17	66.31	60.74	33.69
0.2	48.19	53.37	55.91	51.75	52.05	52.56	35.57	35.78	12.94	25.02	25.73	27.54	24.75	29.45	61.54	53.58	57.56	46.42
0.4	47.33	51.77	57.10	51.08	52.22	51.42	34.28	35.30	14.41	25.19	25.59	28.93	24.76	29.81	62.60	47.21	54.91	52.79
0.6	44.65	46.34	56.88	47.66	50.76	47.00	32.98	34.87	16.74	24.16	26.56	28.51	25.43	29.18	60.48	53.58	57.03	46.42
0.8	43.00	44.94	52.71	47.58	47.86	46.26	33.06	34.90	20.45	25.63	27.40	28.09	26.97	29.54	57.82	49.07	53.45	50.93
1.0	39.96	40.91	51.67	44.16	45.82	42.54	29.99	34.42	22.86	27.26	27.68	27.12	26.84	29.60	55.97	53.05	54.51	46.95

research [74], [75], [76], [77]. Introducing in-context learning slightly mitigates this tendency, though a substantial gap remains. In the Counterfactual scenario, some LLMs cannot differentiate cases of “Bypass” code, struggling to correctly determine the code’s execution path. This so-called runtime reasoning ability [27] is crucial for vulnerability analysis, particularly for runtime-dependent vulnerabilities like Buffer Errors and Input Validation. Additionally, many LLMs show a distribution of responses heavily skewed toward “unsafe”, implying the code contains vulnerabilities. Notably, this happens even when our generator places vulnerable code along a dead path in the scenario. This phenomenon further supports the hypothesis that LLMs rely on pattern matching rather than genuine runtime reasoning.

In the more complex Goal-driven scenario, most LLMs tend to select “Unknown.” For instance, GPT-4 and LLaMA-series models, although they perform relatively well in other scenarios, frequently select “Unknown” option in this scenario. This indicates that current LLMs have difficulty applying parameterized knowledge to resolve vulnerabilities effectively. Unlike in other use cases, addressing vulnerabilities requires advanced prompt engineering or fine-tuning customized for specific scenarios. Lastly, in the Predictive scenario, most LLMs distribute their choices evenly across all available options. Incorrect classifications such as “Other,” “Bypass,” and “Unknown” dominate their responses. This distribution suggests that most LLMs: i) cannot reliably determine whether a vulnerability would be triggered in the current scenario. Simple pattern matching is insufficient for this level of judgment; and ii) struggle to distinguish fundamental differences between CWEs. This highlights the need for domain adaptation and specialized approaches. Without these, LLMs are not well-suited for direct application to vulnerability analysis.

5.3. Pairwise Performance Evaluation

Pairwise Performance Evaluation assesses pairs of samples that share the same functionality but differ in vulnerability. In our single Q&A setting, each question was presented in a single prompt per thread. By contrast, in the pairwise setting, each safe&unsafe pair was presented together in one prompt per thread. Both experiments were conducted independently. Inspired by PrimeVul [9], we use these pairs to examine how effectively LLMs handle syntactically similar yet semantically distinct scenarios within the same context. In the “Single” evaluation, each question Q is presented independently, producing $A = f(\{Q\})$, where A

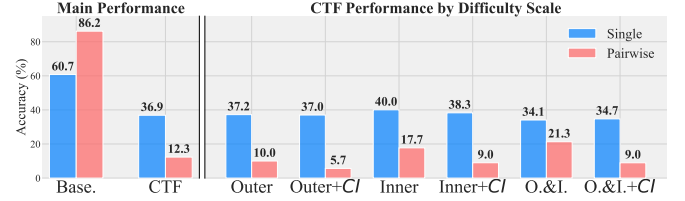


Figure 8: Pairwise performance analysis. **Base** and **CTF** indicate base and Counterfactual scenarios performance, respectively. **+CI** denotes code variants with control-flow injection. “Single” means one question per thread, while “Pairwise” means questions with similar context are combined for inference.

is the answer and f the LLM. In the “Pairwise” evaluation, a set of questions $\{Q_1, \dots, Q_N\}$ is presented simultaneously, yielding $\{A_1, \dots, A_N\} = f(\{Q_1, \dots, Q_N\})$. We measure performance in the “Single” setting as $\frac{\#(\text{Correct Answers})}{\#(\text{Total Questions})}$ and in the “Pairwise” setting as $\frac{\#(\text{Correct Pairs})}{\#(\text{Total Pairs})}$, where a pair is correct only if **all** its answers are correct.

We restricted pairwise evaluations to the *Base* and *Counterfactual* scenarios using Llama3.1-8B, excluding Goal-driven and Predictive scenarios because they do not produce semantically distinct safe&unsafe label pairs required for meaningful pairwise comparison. As shown in Figure 8, performance in the Base scenario increases from 60.7% to 86.2% under pairwise evaluation, but drops from 36.9% to 12.3% in the Counterfactual scenario, consistently across all difficulty levels. We also examined the consistency score in the pairwise setting, $\text{Cons}_{\text{CTF}}^{\text{Pairwise}}$, defined as:

$$\text{Cons}_{\text{CTF}}^{\text{Pairwise}} = \frac{\sum_{i=1}^{N_{\text{CTF}}^{\text{Pairwise}}} \mathbb{I}(C_{\text{CTF}^{\text{Pairwise}}}^i = 1 \wedge C_{\text{base}^{\text{Pairwise}}}^i = 1)}{N_{\text{CTF}}^{\text{Pairwise}}}.$$

This measure extends the consistency score introduced in Eq. 13 for the “single” setting. Under single evaluation, it is 24.3%, whereas under pairwise evaluation, it drops to 15.3%. A likely explanation involves the Juliet dataset, which is widely available and may have been part of the model’s pre-training data. Since Juliet often places safe and vulnerable code side by side, Llama3.1-8B effectively recognizes these patterns in the Base Scenario. However, the Counterfactual Scenario relies on our generator to generate semantically perturbed variants that deviate from the model’s training data, resulting in a marked performance decline. This outcome further indicates that LLMs tend

TABLE 5: Comparison of LLM performance on benchmark generated based on **PrimeVul** Dataset [9].

Models	DataFlow (601)		Structure Reasoning ControlFlow (419)		Average		Counterfactual (658)		Semantic Reasoning Goal-driven (189)		Predictive (205)		Average		Base Scenario			
	Zero	ICL	Zero	ICL	Zero	ICL	Zero	ICL	Zero	ICL	Zero	ICL	Zero	ICL	Unsafe (64)	Safe (64)	Average	FPR ^{Safe}
>15B Param. Models																		
GPT-4-turbo	51.91	55.75↑	73.75	76.27↑	62.83	66.01↑	44.83	48.98↑	13.23	21.40↑	14.15	17.17↑	24.07	29.18↑	17.19	81.25	49.22	18.75
GPT-3.5-turbo	35.27	37.11↑	45.35	46.15↑	40.31	42.63↑	45.74	49.35↑	28.57	29.10↑	21.95	22.08↑	32.09	33.51↑	32.81	62.50	47.66	37.50
Llama3.1-405B*	34.41	55.07↑	70.88	71.84↑	62.64	63.46↑	50.76	57.75↑	24.87	31.22↑	10.24	15.12↑	28.62	34.70↑	85.94	15.62	50.78	84.38
<15B Param. Models																		
Llama3.1-8B*	42.43	47.75↑	53.70	57.76↑	48.06	52.76↑	28.72	33.89↑	17.46	12.70↓	27.32	28.29↑	24.5	24.96↑	40.62	68.75	54.69	31.25
Llama3-8B*	41.60	44.76↑	49.16	55.13↑	45.38	49.94↑	35.87	36.47↑	4.76	11.64↑	12.20	25.37↑	17.61	24.49↑	21.88	81.25	51.56	18.75
Gemma-7B*	35.11	34.28↓	21.96	21.96↓	28.54	28.12↓	29.48	27.05↓	32.80	30.16↓	23.41	30.73↑	28.56	29.31↑	98.44	0.00	49.22	100.00
CodeQwen1.5-7B*	33.61	31.45↓	26.49	33.89↑	30.05	32.67↑	34.50	35.11↑	35.45	26.46↓	24.88	26.34↑	31.61	29.30↓	71.88	0.00	35.94	100.00
Mixtral-7B*	33.61	41.93↑	26.49	42.96↑	30.05	42.44↑	29.64	29.33↓	1.59	4.23↑	20.98	26.34↑	17.40	19.97↑	0.00	0.00	0.00	100.00

to rely more on pattern matching than on deeper logical reasoning.

5.4. Effects of CWEs

The CWE hierarchy comprises pillars, classes, bases, and variants [78]. Figure 7 shows our evaluation of various LLMs (zero-shot and in-context) mapped from Juliet’s CWEs to their **class-levels**. We observe notable performance variations tied to the LLMs’ pre-training domains. For example, Gemma and CodeLlama excel in Goal-driven scenarios like CWE-20 and CWE-1390, indicating a need for targeted enhancements in other CWEs. Comparing zero-shot and in-context heatmaps reveals that in-context learning typically improves understanding of specific CWE classes; however, models like GPT-4 may underperform in certain Goal-driven cases. Overall, these findings underscore the importance of tailoring prompt designs to each CWE for effective vulnerability analysis.

5.5. Effects of Temperature

We conducted an ablation study on inference temperature (ranging from 0.0 to 1.0 in 0.2 increments) using both zero-shot and in-context learning with Llama3.1-8B. As shown in Table 4, lower temperatures (e.g., 0.0) produce deterministic outputs that perform well in tasks requiring consistency, such as structural reasoning, base scenarios, and Counterfactual questions. Conversely, for more challenging tasks like Goal-driven and Predictive scenarios, higher temperatures yield better performance by fostering greater output diversity, which allows the model to explore varied reasoning paths and generate more creative answers.

5.6. Generalization of SV-TrustEval-C

We evaluated our generator and benchmark on the **PrimeVul** dataset, which provides higher label accuracy and broader CWE coverage than other sources. We manually validated samples from each CWE category and selected one verified *safe* vs. *unsafe* pair per category. Note that we skipped compilation validation, as PrimeVul lacks a direct compilation source—a step required to match the precision of our Juliet experiments. This process generated 64 *safe* and 64 *unsafe* base scenario samples, along with

601 DataFlow-based, 419 ControlFlow-based, 658 Counterfactual, 189 Goal-driven, and 205 Predictive questions, all of which are included in our GitHub repository⁵.

Table 5 shows trends similar to those observed with Juliet. Larger models (e.g., GPT-4, GPT3.5, Llama3.1-405B) perform well in structural tasks, while most models struggle with semantic tasks, often scoring below 50%. High false positive rates persist, with some models misclassifying safe code as unsafe. In-context learning generally boosts performance—especially for models with limited code pre-training (e.g., CodeQwen, CodeGemma, Mixtral)—although gains vary. Overall, these findings confirm that **SV-TRUSTEVAL-C** generalizes well to datasets with broader CWE coverage, though it is advisable to use datasets that support full quality verification. (i.e., label accuracy and compilation validation).

6. Conclusion

We present SV-TRUSTEVAL-C, our initial benchmark for evaluating LLMs’ vulnerability reasoning in C. Our study shows that most LLMs, including GPT-4, rely more on pattern matching than logical analysis, struggle with semantic reasoning, and perform inconsistently across code variants. They are also easily misled by bypass attempts, indicating limited structural understanding and domain-specific knowledge. These findings highlight the need to improve LLMs’ code analysis. Future work will expand SV-TRUSTEVAL-C to additional programming languages and more complex scenarios.

References

- [1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [2] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.
- [3] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.

5. <https://github.com/Jackline97/SV-TrustEval-C>

- [4] B. Roziere, J. Gehring, F. Gloeckle, S. Sootla, I. Gat, X. E. Tan, Y. Adi, J. Liu, T. Remez, J. Rapin *et al.*, “Code llama: Open foundation models for code,” *arXiv preprint arXiv:2308.12950*, 2023.
- [5] D. Y. Daya Guo, Qihao Zhu, “Deepseek-coder: When the large language model meets programming – the rise of code intelligence,” 2024. [Online]. Available: <https://arxiv.org/abs/2401.14196>
- [6] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang *et al.*, “Qwen technical report,” *arXiv preprint arXiv:2309.16609*, 2023.
- [7] GitHub, “Github copilot x: the ai-powered developer experience,” <https://github.blog/2023-03-22-github-copilot-x-the-ai-powered-developer-experience/>, 2023.
- [8] Q. Zhang, C. Fang, B. Yu, W. Sun, T. Zhang, and Z. Chen, “Pre-trained model-based automated software vulnerability repair: How far are we?” *IEEE Transactions on Dependable and Secure Computing*, 2023.
- [9] Y. Ding, Y. Fu, O. Ibrahim, C. Sitawarin, X. Chen, B. Alomair, D. Wagner, B. Ray, and Y. Chen, “Vulnerability detection with code language models: How far are we?” *arXiv preprint arXiv:2403.18624*, 2024.
- [10] M. Fu, C. Tantithamthavorn, T. Le, Y. Kume, V. Nguyen, D. Phung, and J. Grundy, “Aibughunter: A practical tool for predicting, classifying and repairing software vulnerabilities,” *Empirical Software Engineering*, vol. 29, no. 1, p. 4, 2024.
- [11] O. Tal, S. Knight, and T. Dean, “Syntax-based vulnerability testing of frame-based network protocols,” in *PST*, 2004, pp. 155–160.
- [12] C. Ko, G. Fink, and K. Levitt, “Automated detection of vulnerabilities in privileged programs by execution monitoring,” in *Tenth Annual Computer Security Applications Conference*. IEEE, 1994, pp. 134–144.
- [13] S. Barke, M. B. James, and N. Polikarpova, “Grounded copilot: How programmers interact with code-generating models,” *Proceedings of the ACM on Programming Languages*, vol. 7, no. OOPSLA1, pp. 85–111, 2023.
- [14] J. He and M. Vechev, “Large language models for code: Security hardening and adversarial testing,” in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023, pp. 1865–1879.
- [15] H. Pearce, B. Ahmad, B. Tan, B. Dolan-Gavitt, and R. Karri, “Asleep at the keyboard? assessing the security of github copilot’s code contributions,” in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 754–768.
- [16] A. Khare, S. Dutta, Z. Li, A. Solko-Breslin, R. Alur, and M. Naik, “Understanding the effectiveness of large language models in detecting security vulnerabilities,” *arXiv preprint arXiv:2311.16169*, 2023.
- [17] Snyk, “Copilot amplifies insecure codebases by replicating vulnerabilities,” <https://snyk.io/blog/copilot-amplifies-insecure-codebases-by-replicating-vulnerabilities/>, 2024, accessed: 2024-02-22.
- [18] O. Asare, “Security evaluations of github’s copilot,” Master’s thesis, University of Waterloo, 2023.
- [19] Y. Fu, P. Liang, A. Tahir, Z. Li, M. Shahin, and J. Yu, “Security weaknesses of copilot generated code in github,” *arXiv preprint arXiv:2310.02059*, 2023.
- [20] G. Nikitopoulos, K. Dritsa, P. Louridas, and D. Mitropoulos, “Crossvul: a cross-language vulnerability dataset with commit data,” in *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2021, pp. 1565–1569.
- [21] J. Fan, Y. Li, S. Wang, and T. N. Nguyen, “Ac/c++ code vulnerability dataset with code changes and cve summaries,” in *Proceedings of the 17th International Conference on Mining Software Repositories*, 2020, pp. 508–512.
- [22] Y. Chen, Z. Ding, L. Alowain, X. Chen, and D. Wagner, “Diversevul: A new vulnerable source code dataset for deep learning based vulnerability detection,” in *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses*, 2023, pp. 654–668.
- [23] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg *et al.*, “Sparks of artificial general intelligence: Early experiments with gpt-4,” *arXiv preprint arXiv:2303.12712*, 2023.
- [24] W. Zhong, R. Cui, Y. Guo, Y. Liang, S. Lu, Y. Wang, A. Saied, W. Chen, and N. Duan, “Agieval: A human-centric benchmark for evaluating foundation models,” *arXiv preprint arXiv:2304.06364*, 2023.
- [25] J. Huang and K. C.-C. Chang, “Towards reasoning in large language models: A survey,” in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 1049–1065. [Online]. Available: <https://aclanthology.org/2023.findings-acl.67>
- [26] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, “Survey of hallucination in natural language generation,” *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.
- [27] J. Chen, Z. Pan, X. Hu, Z. Li, G. Li, and X. Xia, “Reasoning runtime behavior of a program with llm: How far are we?” in *2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE)*. IEEE Computer Society, 2024, pp. 140–152.
- [28] C. Liu, S. D. Zhang, A. R. Ibrahimzada, and R. Jabbarvand, “Code-mind: A framework to challenge large language models for code reasoning,” *arXiv preprint arXiv:2402.09664*, 2024.
- [29] F. Yamaguchi, N. Golde, D. Arp, and K. Rieck, “Modeling and discovering vulnerabilities with code property graphs,” in *2014 IEEE symposium on security and privacy*. IEEE, 2014, pp. 590–604.
- [30] Y. Zhou, S. Liu, J. Siow, X. Du, and Y. Liu, “Devign: Effective vulnerability identification by learning comprehensive program semantics via graph neural networks,” *Advances in neural information processing systems*, vol. 32, 2019.
- [31] N. Risse and M. Böhme, “Limits of machine learning for automatic vulnerability detection,” *arXiv preprint arXiv:2306.17193*, 2023.
- [32] F. Meade, “Juliet test suite v1. 2 for c/c++ user guide,” no. December, 2012.
- [33] L. Zhong and Z. Wang, “Can llm replace stack overflow? a study on robustness and reliability of large language model code generation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 21 841–21 849.
- [34] N. Tihanyi, M. A. Ferrag, R. Jain, and M. Debbah, “Cybermetric: A benchmark dataset for evaluating large language models knowledge in cybersecurity,” *arXiv preprint arXiv:2402.07688*, 2024.
- [35] M. T. Alam, D. Bhusal, L. Nguyen, and N. Rastogi, “Ctiben: A benchmark for evaluating llms in cyber threat intelligence,” *arXiv preprint arXiv:2406.07599*, 2024.
- [36] M. Bhatt, S. Chennabasappa, Y. Li, C. Nikolaidis, D. Song, S. Wan, F. Ahmad, C. Aschermann, Y. Chen, D. Kapil *et al.*, “Cyberseceval 2: A wide-ranging cybersecurity evaluation suite for large language models,” *arXiv preprint arXiv:2404.13161*, 2024.
- [37] S. Ullah, M. Han, S. Pujar, H. Pearce, A. Coskun, and G. Stringhini, “Llms cannot reliably identify and reason about security vulnerabilities (yet?): A comprehensive evaluation, framework, and benchmarks,” in *IEEE Symposium on Security and Privacy*, 2024.
- [38] B. Liu, L. Shi, Z. Cai, and M. Li, “Software vulnerability discovery techniques: A survey,” in *2012 fourth international conference on multimedia information networking and security*. IEEE, 2012, pp. 152–156.
- [39] N. Ayewah, W. Pugh, D. Hovemeyer, J. D. Morgenthaler, and J. Penix, “Using static analysis to find bugs,” *IEEE software*, vol. 25, no. 5, pp. 22–29, 2008.

- [40] T. Copeland, *PMD applied*. Centennial Books San Francisco, 2005, vol. 10.
- [41] O. Burn, "Checkstyle," <http://checkstyle.sourceforge.net>, 2003.
- [42] G. Klees, A. Ruef, B. Cooper, S. Wei, and M. Hicks, "Evaluating fuzz testing," in *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, 2018, pp. 2123–2138.
- [43] J. A. Harer, L. Y. Kim, R. L. Russell, O. Ozdemir, L. R. Kosta, A. Rangamani, L. H. Hamilton, G. I. Centeno, J. R. Key, P. M. Ellingwood *et al.*, "Automated software vulnerability detection with machine learning," *arXiv preprint arXiv:1803.04497*, 2018.
- [44] T. Ball, "The concept of dynamic analysis," *ACM SIGSOFT Software Engineering Notes*, vol. 24, no. 6, pp. 216–234, 1999.
- [45] Z. Bilgin, M. A. Ersoy, E. U. Soykan, E. Tomur, P. Çomak, and L. Karaçay, "Vulnerability prediction from source code using machine learning," *IEEE Access*, vol. 8, pp. 150 672–150 684, 2020.
- [46] Z. Li, D. Zou, S. Xu, X. Ou, H. Jin, S. Wang, Z. Deng, and Y. Zhong, "Vuldeepecker: A deep learning-based system for vulnerability detection," *arXiv preprint arXiv:1801.01681*, 2018.
- [47] Z. Li, D. Zou, S. Xu, H. Jin, Y. Zhu, and Z. Chen, "Sysevr: A framework for using deep learning to detect software vulnerabilities," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 4, pp. 2244–2258, 2021.
- [48] S. Chakraborty, R. Krishna, Y. Ding, and B. Ray, "Deep learning based vulnerability detection: Are we there yet?" *IEEE Transactions on Software Engineering*, vol. 48, no. 9, pp. 3280–3296, 2021.
- [49] L. Wartschinski, Y. Noller, T. Vogel, T. Kehrner, and L. Grunske, "Vudenc: vulnerability detection with deep learning on a natural codebase for python," *Information and Software Technology*, vol. 144, p. 106809, 2022.
- [50] Z. Feng, D. Guo, D. Tang, N. Duan, X. Feng, M. Gong, L. Shou, B. Qin, T. Liu, D. Jiang *et al.*, "Codebert: A pre-trained model for programming and natural languages," *arXiv preprint arXiv:2002.08155*, 2020.
- [51] D. Guo, Q. Zhu, D. Yang, Z. Xie, K. Dong, W. Zhang, G. Chen, X. Bi, Y. Wu, Y. Li *et al.*, "Deepseek-coder: When the large language model meets programming—the rise of code intelligence," *arXiv preprint arXiv:2401.14196*, 2024.
- [52] Y. Nong, R. Fang, G. Yi, K. Zhao, X. Luo, F. Chen, and H. Cai, "Vgx: Large-scale sample generation for boosting learning-based software vulnerability analyses," in *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, 2024, pp. 1–13.
- [53] D. de Fitero-Dominguez, E. Garcia-Lopez, A. Garcia-Cabot, and J.-J. Martinez-Herrera, "Enhanced automated code vulnerability repair using large language models," *arXiv preprint arXiv:2401.03741*, 2024.
- [54] M. A. Ferrag, F. Alwahedi, A. Battah, B. Cherif, A. Mechri, and N. Tihanyi, "Generative ai and large language models for cyber security: All insights you need," *arXiv preprint arXiv:2405.12750*, 2024.
- [55] Y. Liu, G. Deng, Y. Li, K. Wang, Z. Wang, X. Wang, T. Zhang, Y. Liu, H. Wang, Y. Zheng *et al.*, "Prompt injection attack against llm-integrated applications," *arXiv preprint arXiv:2306.05499*, 2023.
- [56] J. Piet, M. Alrashed, C. Sitawarin, S. Chen, Z. Wei, E. Sun, B. Alo-mair, and D. Wagner, "Jatmo: Prompt injection defense by task-specific finetuning," in *European Symposium on Research in Computer Security*. Springer, 2024, pp. 105–124.
- [57] C.-A. Brust, T. Sonnekalb, and B. Gruner, "Romeo: A binary vulnerability detection dataset for exploring juliet through the lens of assembly language," *Computers & Security*, vol. 128, p. 103165, 2023.
- [58] J. Henkel, S. K. Lahiri, B. Liblit, and T. Reps, "Code vectors: Understanding programs through embedded abstracted symbolic traces," in *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2018, pp. 163–174.
- [59] A. Ni, S. Iyer, D. Radev, V. Stoyanov, W.-t. Yih, S. Wang, and X. V. Lin, "Lever: Learning to verify language-to-code generation with execution," in *International Conference on Machine Learning*. PMLR, 2023, pp. 26 106–26 128.
- [60] X. Chen, M. Lin, N. Schärli, and D. Zhou, "Teaching large language models to self-debug," *arXiv preprint arXiv:2304.05128*, 2023.
- [61] C. Liu, S. Lu, W. Chen, D. Jiang, A. Svyatkovskiy, S. Fu, N. Sundaresan, and N. Duan, "Code execution with pre-trained language models," *arXiv preprint arXiv:2305.05383*, 2023.
- [62] Y. Ding, B. Steenhoeck, K. Pei, G. Kaiser, W. Le, and B. Ray, "Traced: Execution-aware pre-training for source code," in *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*, 2024, pp. 1–12.
- [63] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.
- [64] AI@Meta, "Llama 3 model card," *Github*, 2024. [Online]. Available: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md
- [65] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Riviere, M. S. Kale, J. Love *et al.*, "Gemma: Open models based on gemini research and technology," *arXiv preprint arXiv:2403.08295*, 2024.
- [66] C. Team, "Codegemma: Open code models based on gemma," *arXiv preprint arXiv:2406.11409*, 2024.
- [67] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. I. Casas, E. B. Hanna, F. Bressand *et al.*, "Mixtral of experts," *arXiv preprint arXiv:2401.04088*, 2024.
- [68] OpenAI, "Openai o1 system card," <https://openai.com/index/openai-o1-system-card/>, Sep. 2024.
- [69] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, and W. Chen, "What makes good in-context examples for gpt-3?" *arXiv preprint arXiv:2101.06804*, 2021.
- [70] Y. Lu, M. Bartolo, A. Moore, S. Riedel, and P. Stenetorp, "Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity," *arXiv preprint arXiv:2104.08786*, 2021.
- [71] Z. Wu, Y. Wang, J. Ye, and L. Kong, "Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering," *arXiv preprint arXiv:2212.10375*, 2022.
- [72] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [73] L. Berglund, M. Tong, M. Kaufmann, M. Balesni, A. C. Stickland, T. Korbak, and O. Evans, "The reversal curse: LLMs trained on 'a is b' fail to learn 'b is a'," *arXiv preprint arXiv:2309.12288*, 2023.
- [74] S. Mirchandani, F. Xia, P. Florence, B. Ichter, D. Driess, M. G. Arenas, K. Rao, D. Sadigh, and A. Zeng, "Large language models as general pattern machines," *arXiv preprint arXiv:2307.04721*, 2023.
- [75] I. Weber, "Large language models are pattern matchers: Editing semi-structured and structured documents with chatgpt," *arXiv preprint arXiv:2409.07732*, 2024.
- [76] I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio, and M. Farajtabar, "Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models," *arXiv preprint arXiv:2410.05229*, 2024.
- [77] C. Xie, Y. Huang, C. Zhang, D. Yu, X. Chen, B. Y. Lin, B. Li, B. Ghazi, and R. Kumar, "On memorization of large language models in logical reasoning," *arXiv preprint arXiv:2410.23123*, 2024.
- [78] R. A. Martin and S. Barnum, "Common weakness enumeration (cwe) status update," *ACM SIGAda Ada Letters*, vol. 28, no. 1, pp. 88–91, 2008.

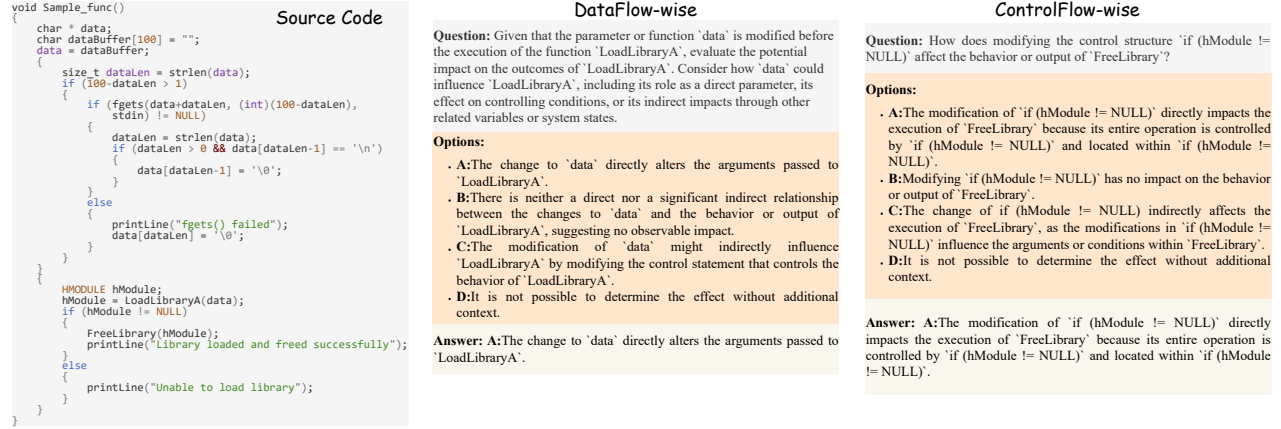


Figure 9: Examples of Structure Reasoning Questions.

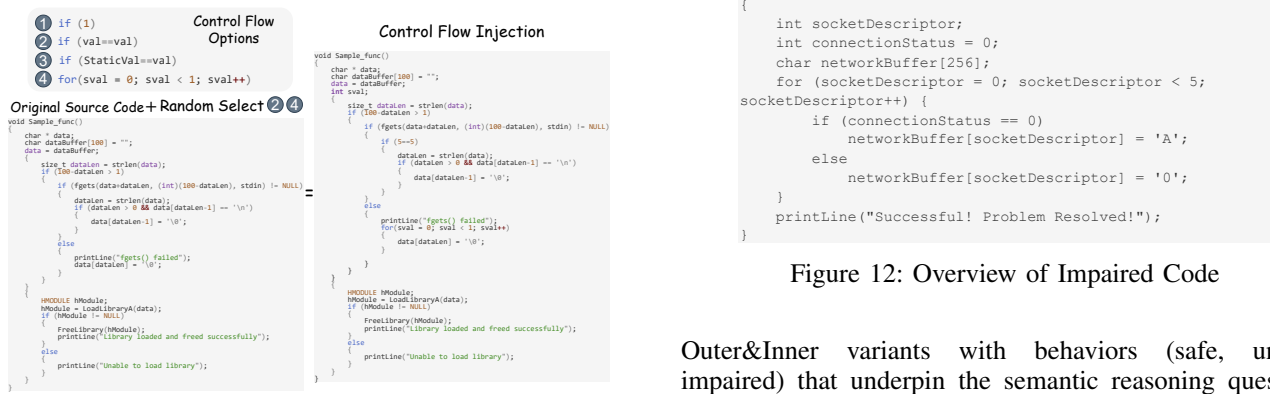


Figure 12: Overview of Impaired Code

Outer&Inner variants with behaviors (safe, unsafe, impaired) that underpin the semantic reasoning questions in Figure 11.

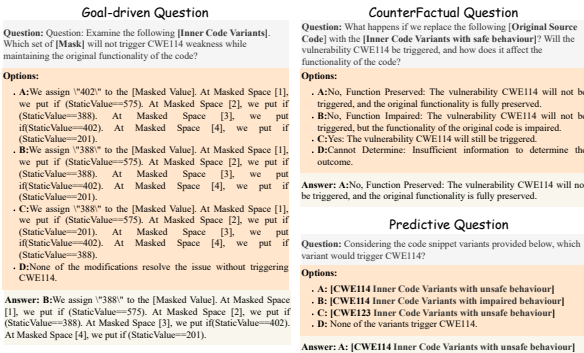


Figure 11: Examples of Semantic Reasoning Questions.

Appendix A. Question Examples

Our SV-TrustEval-C Benchmark is demonstrated through example questions. Figure 9 shows how structural reasoning questions are derived from the original code, while Figure 10 presents our Variants Generator, which adds control flow branches to create Outer, Inner, and

A.1. Impaired Code Sample

We introduced the snippet in Figure 12 as impaired code *C_{impaired}*, simulating false success to test if LLMs grasp the original code's vulnerabilities.

Appendix B. Error Case Analysis

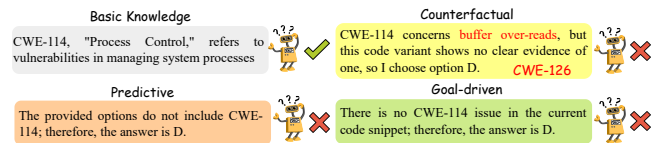


Figure 13: Error cases analysis of GPT-4

Our analysis shows that 33% of zero-shot and 21% of in-context semantic errors result in "Do not know" responses due to inconsistent vulnerability interpretation. For example, as Figure 13 illustrates, GPT-4 initially identifies CWE-114 correctly but later misclassifies it or responds "I don't know," indicating an overreliance on pattern matching over deep semantic reasoning.

Appendix C. Meta-Review

The following meta-review was prepared by the program committee for the 2025 IEEE Symposium on Security and Privacy (S&P) as part of the review process as detailed in the call for papers.

C.1. Summary

This paper introduces SV-TrustEval-C, a novel benchmark designed to evaluate the reasoning capabilities of Large Language Models (LLMs) in detecting source code vulnerabilities. The benchmark focuses on two key dimensions: structural reasoning, which assesses how well models identify relationships between code elements under varying data and control flow complexities, and semantic reasoning, which evaluates their logical consistency when code is altered both structurally and semantically. The authors evaluate 11 LLMs on this benchmark, revealing that current models rely more on pattern matching than on robust logical reasoning for vulnerability analysis.

C.2. Scientific Contributions

- **Creates a New Tool to Enable Future Science:** The SV-TrustEval-C benchmark provides a systematic way to evaluate LLMs' reasoning capabilities in vulnerability analysis, enabling future research in this area.
- **Provides a Valuable Step Forward in an Established Field:** The paper advances the field of LLM-based vulnerability detection by introducing a novel approach that leverages multiple-choice questions to assess reasoning, moving beyond traditional prompt engineering methods.

C.3. Reasons for Acceptance

- 1) The SV-TrustEval-C benchmark is a significant contribution to the field, as it allows for the systematic evaluation of LLMs' reasoning capabilities in vulnerability analysis. The benchmark is designed to be extensible, enabling future researchers to apply it to other datasets and codebases.
- 2) The paper addresses a critical gap in the evaluation of LLMs for vulnerability detection by focusing on both structural and semantic reasoning. The findings highlight the limitations of current models, providing valuable insights for future improvements in LLM-based vulnerability analysis.
- 3) The paper includes a detailed evaluation of 11 LLMs, offering a broad perspective on the current state of LLM performance in vulnerability detection. The results underscore the need for further research into enhancing the logical reasoning capabilities of LLMs.

C.4. Noteworthy Concerns

- 1) **Bias in the Benchmark:** While the authors have addressed concerns about bias by removing keywords like "good" and "bad" from the Juliet Test Suite, there is still a potential for bias due to the reliance on a single dataset. The authors have taken steps to mitigate this by extending their evaluation to the PrimeVul dataset, but further validation on real-world codebases would strengthen the benchmark's generalizability.
- 2) **Generalization Beyond Juliet Dataset:** The benchmark's reliance on the Juliet Test Suite raises questions about its applicability to real-world vulnerabilities. The authors have demonstrated extensibility by applying their approach to the PrimeVul dataset, but further validation on open-source vulnerability datasets (e.g., CVE-based datasets) would enhance the benchmark's credibility.

Appendix D. Response to the Meta-Review

We appreciate the reviewers' insightful feedback and the shepherd's guidance in improving our paper. While we acknowledge the value of extending our benchmark to additional real-world codebases, we must also maintain its quality in terms of label accuracy and code compilation validation. Our use of the Juliet and PrimeVul datasets is informed by these considerations: Juliet includes a compilable check environment for our generator and its artificially generated nature ensures label accuracy, whereas PrimeVul incorporates real-world codebases with rigorous label validation. Although further exploration of additional real-world datasets would likely reinforce our conclusions, the extensive effort required for label verification and code-compilation checks goes beyond our current scope. We plan to address this in future work.