Benchmarking Explainable Vulnerability Detection with Calling Contexts

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly applied to vulnerability detection in real-world software systems, where the task goes beyond binary classification and requires calling context exploration over large code repositories. However, existing benchmarks primarily formulate the vulnerability detection as a classification problem without calling contexts and provide no formalized explanations, making them inadequate for evaluating LLM-driven vulnerability detectors upon repositories. To fill the gap, we introduce ManBo, a benchmark designed to assess the capability of LLM-driven vulnerability detectors in the real-world scenarios. Specifically, ManBo provides human-verified vulnerability explanations together with calling contexts, enabling principled evaluation of the ability to both identify vulnerabilities and explain root causes. It contains 205 vulnerable instances and 164 safe instances collected from 74 high-profile C/C++ projects, covering three critical types of memory corruption. We evaluate a broad spectrum of LLM-driven detectors on ManBo, and find that effective calling context retrieval is essential for repository-level vulnerability detection, exposing a fundamental gap between the capabilities of existing LLM-driven detectors and real-world requirements.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities and are now widely applied across many domains (Achiam et al., 2023; Touvron et al., 2023; Liu et al., 2024). In software engineering, developers increasingly rely on LLMs to support tasks such as code generation and vulnerability detection (Xu et al., 2024; Chen, 2021). Among these tasks, vulnerability detection is the cornerstone of reliable software development in the AI era, attracting substantial research interest and industry attention (Cursor, 2025; CodeRabbit, 2025; Wang et al., 2024a; Guo et al., 2025b). With expanded context windows and improved reasoning abilities, LLMs are increasingly being deployed beyond controlled experimental settings and into real-world software projects, where they are used to uncover previously unknown vulnerabilities (Guo et al., 2025b,a; Zheng et al., 2025).

To evaluate the capability of LLM-driven vulnerability detection in real-world settings, researchers have proposed various benchmarks over the past few years, including Devign (Zhou et al., 2019), BigVul (Fan et al., 2020), PrimeVul (Ding et al., 2024), ReposVul (Wang et al., 2024b), and CrossVul (Nikitopoulos et al., 2021). These benchmarks collect vulnerable code and corresponding fixed versions from real-world projects and ask detectors to determine whether a given code snippet contains a vulnerability. While effective for evaluating classification accuracy, these benchmarks suffer from two fundamental limitations.

Limitation I: Lack of formalized explanations. Existing vulnerability detection benchmarks predominantly formulate vulnerability detection as a classification problem (Zhou et al., 2019; Steenhoek et al., 2024). As code snippets in these benchmarks are annotated only with binary labels, namely *Vulnerable* or *Safe*, they do not provide formalized explanations of how vulnerabilities arise within the repository. Such simple formulation is insufficient for real-world vulnerability detection, where detailed explanations, such as how vulnerabilities originate, evolve, and are ultimately triggered across the codebase, are essential for effective diagnosis and remediation (Do and Bodden, 2020). As illustrated in Figure 1 (a), the Use-After-Free (UAF) vulnerability spans three functions and involves non-trivial inter-procedural data dependencies. Without a formalized explanation, human auditors must invest significant effort to reconstruct the vulnerable path and identify its root cause. Moreover, the binary classification formulation provides weak evaluation criteria: binary labels alone cannot determine

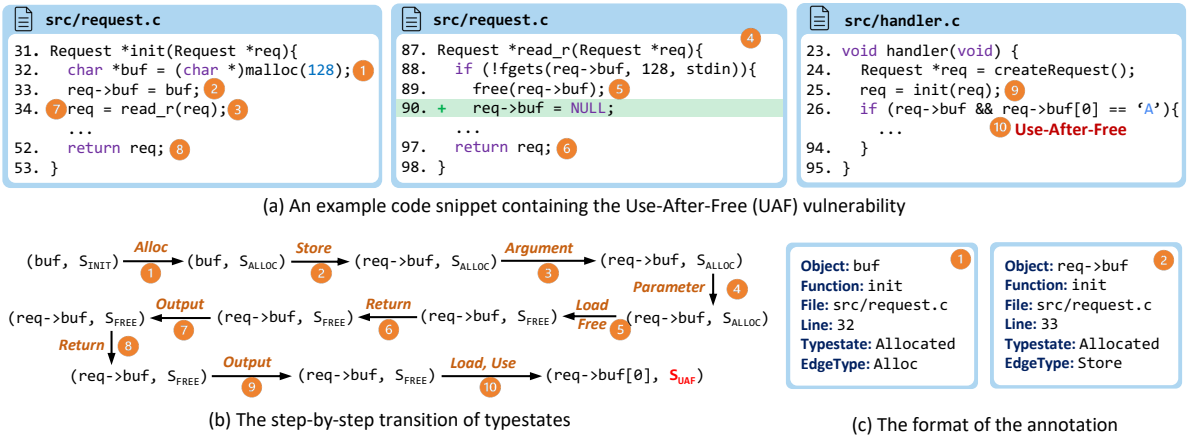


Figure 1: An example of Use-After-Free (UAF). The assignment req->buf = NULL is added in the fixing patch.

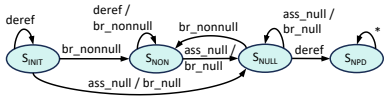
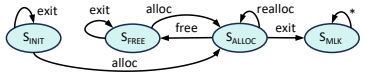
whether a detector reaches a correct prediction via valid reasoning paths (see Appendix J), which can consequently lead to false positives in evaluation.

Limitation II: Insufficient calling contexts. Existing benchmarks mainly define calling contexts based on patch commits (Fan et al., 2020; Ding et al., 2024; Wang et al., 2024b). During dataset construction, they locate a patch commit and treat the functions modified in the commit as related to the vulnerability, labeling the pre-patch version as vulnerable and the post-patch version as safe. Although the method of benchmarking is scalable, it cannot guarantee that the collected calling contexts are sufficient for reliable detection, as patch commits often modify only a subset of the functions involved in the complete vulnerable data flow. As shown in Figure 1 (a), the patch commit for the Use-After-Free (UAF) vulnerability only adds `req->buf = NULL` in the function `read_r`. Under the existing benchmarking strategy, the resulting calling context only includes the function `read_r`, while omitting other functions such as `init`, which allocates the memory, and `handler`, which later uses the freed memory. Obviously, evaluating LLM-driven detectors within such incomplete calling contexts leads to an ill-posed problem setting and can produce misleading conclusions (Risse et al., 2025).

Our Benchmark. To address these limitations, we introduce ManBo, a real-world vulnerability detection benchmark that provides both formalized explanations and sufficient calling contexts for evaluating LLM-driven vulnerability detectors. For rigorous formulation, we focus on *typestate vulnerabilities*, which cover a broad class of vulnerabilities involving resource and state mismanagement (Fink et al., 2008), and can be flexibly specified by users through customized typestate definitions. Specifically, we consider three representative defect types, namely Null Pointer Dereference (NPD), Memory Leak (MLK), and Use-After-Free (UAF), which are common memory corruptions in real-world software systems (MITRE, 2025). Conceptually, we introduce the notion of *typestate slice* to uniformly depict the formalized explanations of various typestate vulnerabilities (see Section 3.2). Since existing program analysis techniques cannot reliably recover the program statements relevant to the target vulnerability upon the source-code level (See Appendix A), we design a semi-automated pipeline for benchmark construction, consisting of issue collection, semantic filtering, and manual annotation, which ensures the high quality of both the calling contexts and explanations. In total, ManBo consists of 205 vulnerable instances and 158 safe instances collected from 74 real-world C/C++ projects, which have an average of 14.8K GitHub stars and 340K lines of code.

We further propose two novel evaluation tasks, *Explainable Vulnerability Detection* and *Calling Context Retrieval*, to assess the capabilities of state-of-the-art LLM-driven detectors upon our contructed benchmark ManBo. Our evaluation shows that LLMs exhibit notable limitations in generating accurate vulnerability explanations, achieving a coverage score of only 0.48. In particular, models may introduce hallucinations in the generated typestate slices, even when the predicted vulnerability label is correct. Moreover, our results demonstrate that the availability of sufficient calling context is a critical factor influencing vulnerability detection performance. However, existing LLM-based agents struggle to precisely retrieve all relevant calling contexts, recovering at best only 55.04% of the groundtruth relevant functions, which substantially limits their effectiveness in end-to-end explainable vulnerability detection.

Table 1: The FSMs of Null Pointer Dereference (NPD), Memory Leak (MLK), and Use-After-Free (UAF).

$\mathbf{FSM_{NPD}}$ $\langle \Sigma, \mathbb{S}, S_{\text{INIT}}, \delta, S_{\text{NPD}} \rangle$	$\mathbf{FSM_{MLK}}$ $\langle \Sigma, \mathbb{S}, S_{\text{INIT}}, \delta, S_{\text{MLK}} \rangle$	$\mathbf{FSM_{UAF}}$ $\langle \Sigma, \mathbb{S}, S_{\text{INIT}}, \delta, S_{\text{UAF}} \rangle$
$\mathbb{S} = \{S_{\text{INIT}}, S_{\text{NON}}, S_{N}, S_{\text{NPD}}\}$ S_{NON}: The variable is non-NULL. S_{NULL}: The variable is NULL.	$\mathbb{S} = \{S_{\text{INIT}}, S_{\text{ALLOC}}, S_{\text{FREE}}, S_{\text{MLK}}\}$ S_{ALLOC}: A heap object is allocated and not freed. S_{FREE}: A heap object is freed.	$\mathbb{S} = \{S_{\text{INIT}}, S_{\text{ALLOC}}, S_{\text{FREE}}, S_{\text{UAF}}\}$ S_{ALLOC}: A heap object is allocated. S_{FREE}: The heap object is freed.
$\Sigma = \{ass_null, br_null, br_nonnull, deref\}$ ***ass_null***: Assign NULL to a pointer. ***br_null***: Execute a branch where the pointer is NULL. ***br_nonnull***: Execute a branch where the pointer is non-NULL. ***deref***: Dereference a pointer.	$\Sigma = \{alloc, realloc, free, exit\}$ ***alloc***: Allocate a heap memory object. ***recalloc***: Reallocate a heap memory object. ***free***: Free a heap object. ***exit***: Return from main function.	$\Sigma = \{alloc, realloc, free, use, ass_null\}$ ***alloc***: Allocate a heap memory object. ***recalloc***: Reallocate a heap memory object. ***free***: Free a heap object. ***use***: use a heap object. ***ass_null***: Assign NULL to the memory pointer.
		

2 Related Work

Recent studies have applied LLMs to real-world vulnerability detection tasks (Guo et al., 2025b; Li et al., 2025; Zheng et al., 2025; Guo et al., 2025a). As these techniques are increasingly evaluated on real-world codebases, the quality of benchmarks becomes critical for assessing their effectiveness and reliability. However, existing benchmarks for real-world vulnerability detection often lack sufficient context information and formalized explanations, which limits their ability to support such evaluations (Risse et al., 2025). Several widely used datasets, such as BigVul (Fan et al., 2020) and Devign (Zhou et al., 2019), construct vulnerability instances from patch commits. Although this approach enables large-scale dataset construction, the resulting instances typically provide incomplete calling contexts and only binary *Vulnerable* / *Safe* labels, without explicit explanations of root causes. Subsequent benchmarks, including PrimeVul (Ding et al., 2024), CrossVul (Nikitopoulos et al., 2021), and DiverseVul (Chen et al., 2023), build upon BigVul and improve data quality through additional filtering and cleaning, but they largely inherit the same fundamental limitations. Defects4J (Just et al., 2014) adopts a different construction methodology by incorporating human verification, ensuring that each vulnerability is executable and can be triggered by at least one test case. Nevertheless, Defects4J still does not provide formalized explanations that characterize the root cause or propagation of vulnerabilities, making it difficult to systematically evaluate a model's semantic reasoning ability.

To better evaluate LLM-driven vulnerability detectors, we focus on *typestate vulnerabilities*, a broad class that subsumes many common real-world vulnerability types (Fink et al., 2008; Li et al., 2022) and enables uniform, formalized explanations with necessary calling contexts. Additional preliminaries are introduced in Section 3. A detailed review of prior work on typestate verification and vulnerability detection is provided in Appendix B.

3 Preliminaries

This section introduces the notion of *typestate vulnerabilities* and the formal model used to characterize them, followed by the definition of *typestate slices* as their formalized explanations.

3.1 Typestate Vulnerabilities

A *typestate* describes the abstract state of a program entity and how this state evolves as program operations are applied to it (Strom and Yemini, 1986). A *typestate vulnerability* occurs when an operation is executed in a state where it is not semantically valid, thereby violating the intended usage protocol of the entity. Typestate vulnerabilities constitute a broad class of real-world defects, encompassing various memory corruptions (e.g., Null Pointer Dereference (NPD), Memory Leak (MLK), and Use-after-Free (UAF)) as well as various resource misuse issues (e.g., file descriptor leaks).

To detect typestate vulnerabilities, prior work models typestate evolution using finite-state machines (FSMs) (Fink et al., 2008; Li et al., 2022). An FSM formalizes relevant typestates and depicts how they are updated as the program executes. Specifically, an FSM abstracts program execution as transitions over a finite set of states. Formally, a typestate vulnerability is defined by an FSM

$$\text{FSM}_{\text{type}} = \langle \Sigma, \mathbb{S}, S_{\text{INIT}}, \delta, S_{\text{type}} \rangle.$$

Here, Σ is the set of operations that may affect the typestate, namely *typestate-transition operations*. $\mathbb{S}$ is the set of typestates and S_{INIT} is the initial state. The transition function $\delta : \mathbb{S} \times \Sigma \rightarrow \mathbb{S}$ defines how typestates are updated in response to typestate-transition operations. S_{type} denotes an error typestate whose reachability indicates a potential vulnerability.

Table 1 illustrates the FSMs formulating NPD, MLK, and UAF. The states S_{NPD}, S_{MLK}, and S_{UAF} denote the error states for the three vulnerability types, respectively. If program execution reaches one of these error states, the corresponding typestate vulnerability is reported. The error state has a self-loop and behaves as a trap state.

3.2 Typestate Slices

Detecting typestate vulnerabilities requires reasoning about both pointer manipulations and inter-procedural value flows. Specifically, pointer load and store operations enable flexible propagation of program values through memory, while passing values as function arguments and return values allows their typestates to evolve across different functions. To formalize how a typestate vulnerability arises in a program, we introduce the notion of a *typestate slice*, which captures the essential reasoning steps leading to the vulnerability. Formally, it is defined as an ordered sequence of program statements

$$\mathcal{S} = \langle s_1, \ldots, s_n \rangle$$

where each statement s_i belongs to one of the following three categories: (1) *typestate-transition operations* defined in the corresponding FSM, which trigger typestate transitions; (2) *pointer load and store operations*, which propagate vulnerability-related values through memory; and (3) *inter-procedural value-passing operations*, including function argument passing and return value propagation across call boundaries. Particularly, the first statement s_1 and the last statement s_n are the typestate-transition operations that change the typestate from the initial state S_{INIT} to a new typestate and update the typestate into an error typestate S_{type}, respectively.

Figure 1 (b) shows an example of typestate slice along with the updated typestates. To simplify the formulation, we only maintain the relevant statements and do not explicitly store typestates in the typestate slice. The FSM formulating the UAF is shown in the last column in Table 1. In the function `init`, a memory object `buf` is *allocated* via `malloc` at line 32 and subsequently *stored* into `req->buf` with the typestate S_{ALLOC}. The pointer `req` is then passed as the first *argument* to `read_r`. Later, the object is *loaded* and *freed* at line 89 in `read_r` and *used* at line 26 in `handler`, resulting in a UAF vulnerability.

Notably, typestate slices precisely capture how the typestate vulnerabilities arise by modeling the sequences of operations upon the concerned program values, making them suitable for evaluating the reasoning chains of the LLM-driven vulnerability detectors. With the inter-procedural propagation information, a typestate slice maintain the functions relevant to a given vulnerability as the calling context. Based on this essential concept, we will construct the benchmark and annotate the formalized explanation for a typestate vulnerability.

4 Our Benchmark: ManBo

In this section, we introduce a high-quality real-world vulnerability detection benchmark that provides both formalized vulnerability explanations and sufficient calling context. Unlike prior benchmarks that primarily evaluate the binary classification performance of LLM-driven vulnerability detectors, our benchmark assesses a broader set of capabilities essential for real-world vulnerability detection, including calling context retrieval and explainable vulnerability detection.

4.1 Benchmark Construction

We first introduce the construction of our benchmark. Since existing program analysis techniques cannot reliably recover complete typestate slices (see Appendix A), we develop a semi-automated pipeline that consists of three stages: issue collection, semantic filtering, and manual annotation.

4.1.1 Issue Collection

We begin by crawling real-world issues from GitHub repositories and selecting projects using two criteria: (1) large scale, with the project size exceeding 100K LoC, and (2) well-maintained, with more than 500 stars in GitHub. In total, we collect 75 projects with an average of 340K lines of code and 14K stars.

For each issue, we parse the description and apply keyword matching to identify those related to the three target vulnerability types. For each potentially relevant issue, we further retrieve supplementary information, including comments, vulnerable commit (i.e., the commit that contains the vulnerability), and referenced commits. For the vulnerable commit, we first attempt to extract the commit identifier directly from the issue description. When this information is unavailable, we approximate the referenced version by selecting the commit immediately preceding the issue creation time. This heuristic increases the likelihood that the extracted code matches the version discussed in the report. Referenced commits include all commits

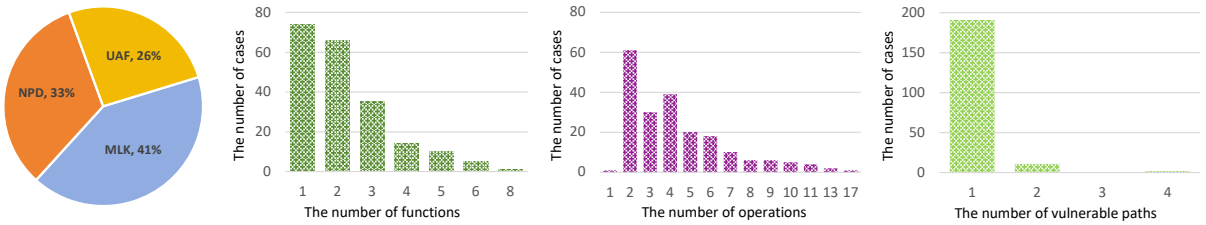


Figure 2: The detailed statistics of vulnerable cases in our benchmark ManBo.

that mention or link to the issue, many of which contain potential patches.

4.1.2 Semantic Filtering

Next, we employ a lightweight LLM agent to analyze each issue together with its supplementary information. The agent performs semantic filtering by retaining only issues that satisfy three constraints: (1) the issue has been confirmed or fixed by the developers; (2) it explicitly describes the vulnerable execution path; and (3) it contains function names or file paths that correspond to the vulnerable path. An issue is retained only when all constraints are satisfied. It is important to note that the goal of this semantic filtering is not to guarantee soundness, but rather to reduce manual effort by prioritizing issues that are more likely to be high quality.

After this stage, 30% of the collected issues are filtered out, primarily due to the lack of developer confirmation, without altering the overall dataset distribution. The remaining 384 real-world issues include project identifiers, vulnerability types, vulnerable and fix commits, but lack complete typestate slices and may contain ambiguous or low-quality cases, necessitating manual filtering and annotation.

4.1.3 Manual Annotation

In the annotation stage, we label the explanation of each vulnerability in the form of typestate slices. As stated in Section 3.2, a typestate slice begins with the initialization of the object and follows its forward propagation along the execution path. Whenever a typestate transition, an inter-procedural propagation, or a pointer load or store operation is encountered, we record the corresponding function name, file path, line number, and the resulting typestate of the object, as illustrated in Figure 1 (c).

In addition, for pointer load and store operations, we annotate the corresponding memory location (as shown by the second annotation in Figure 1(c)), and for inter-procedural propagation operations, we record the indices of the associated arguments and parameters. These annotations enable precise tracking of data flow across function boundaries. This typestate slice constitutes a minimal sequence of operations that fully characterizes the vulnerability and provides line-level explanation, thereby supporting precisely evaluating vulnerability detection.

Wrapper Functions. During the manual annotation, we only consider standard library functions when identifying typestate transition operations. However, many real-world projects encapsulate standard malloc or free primitives within customized helper or wrapper functions. Expanding all propagation steps inside these wrappers would significantly increase the depth of calling contexts. Following common practice in program analysis (Lyu et al., 2022), when encountering a wrapper function, we treat the invocations of wrapper functions as explicit typestate-transition operations instead of inlining its internal propagations. This strategy yields a minimal and concise typestate slice without sacrificing semantic soundness.

All issues are independently annotated by two authors, with conflicts resolved by a senior third annotator. Detailed inter-annotator agreement statistics are reported in Appendix C. To avoid ambiguity, we annotate all vulnerable execution paths that lead to the same vulnerability at the same location, treating each path as a separate typestate slice.

4.2 Dataset Details

By applying the benchmarking methodology stated in Section 4.1, we obtain a total of 205 vulnerable cases, comprising 67 NPD cases, 85 MLK cases, and 53 UAF cases, accounting for 32.7%, 41.5%, and 25.9% of the vulnerable cases, respectively. Each case is annotated with human-verified typestate slices as the formalized explanation. As illustrated in Figure 2, 36.1% (74/205) of the cases are intra-procedural, involving only a single relevant function, while the remaining cases are inter-procedural, spanning up to 8 functions and 17 operations along the annotated typestate slice. Among the 205 vulnerable cases, 164 have associated fix

commits and can be paired with their corresponding fixed versions as safe cases, resulting in a final benchmark of 369 instances with a well-balanced composition of vulnerable and safe cases.

4.3 Task Design

Our benchmark is designed to evaluate the performance of LLM-driven detectors in realistic vulnerability detection scenarios. Specifically, we introduce two complementary tasks, namely *explainable vulnerability detection* and *calling context retrieval*, which are described in detail below.

4.3.1 Explainable Vulnerability Detection

This task evaluates the ability of LLMs to detect vulnerabilities. Specifically, we supply the LLMs with the final operation in the typestate slice and a *calling context* that contains all relevant functions involved in the slice. Note that the relevant functions also include wrapped functions. An ablation study examining the impact of wrapper functions is presented in Appendix D. To ensure that the LLMs follow the prescribed format of the typestate slice shown in Section 3.2, we explicitly specify the output format rules in the prompt and provide three few-shot examples for each vulnerability type.

For evaluation, we assess both the performance of the vulnerability identification and the quality of the generated typestate slice with the groundtruth. For the vulnerability identification, we consider both vulnerable and safe cases and evaluate whether the predicted label (i.e., Vulnerable and Safe) matches the ground-truth label, which yield the statistics of the precision, recall, and F1 score. To quantify the quality of the generated typestate slices, we only consider vulnerable instances and evaluate the consistency between the generated typestate slices and the groundtruth. Specifically, we compute the *typestate-slice coverage* that measures the fraction of the operations in the groundtruth that are covered by the generated typestate slice. More details are provided in Appendix E.

4.3.2 Calling Context Retrieval

In Section 4.3.1, we provide the calling contexts of a vulnerability as input to the explainable vulnerability detection task. However, it remains an open problem that how to automatically collect such calling contexts in a way that effectively facilitates vulnerability detection. Hence, we introduce the task of calling context retrieval, in which we evaluate state-of-the-art LLM-based coding agents on their ability to retrieve relevant functions that serve as calling contexts. In this task, an agent is provided with only a line of code and the corresponding function. It is asked to retrieve all relevant functions needed to determine whether a specific type of bug can arise at the specified program location or not.

Basically, the goal of calling context retrieval is to retrieve all relevant functions while minimizing the inclusion of irrelevant ones. Hence, we measure the precision and recall of the calling context retrieval by comparing the retrieved functions with the ones appearing in the labeled typestate slices. To quantify the efficiency, we measure the financial costs and execution time of each agent. We further conduct an ablation study to evaluate the impact of multiple vulnerable execution paths on retrieval performance (see Appendix F).

5 Evaluation

We organize our evaluation around four research questions.

- **RQ1:** How do LLMs perform in the explainable vulnerability detection with calling contexts?
- **RQ2:** How do LLMs perform in the explainable vulnerability detection without calling contexts?
- **RQ3:** How do existing agents perform in the calling context retrieval?
- **RQ4:** How do existing agents perform in the end-to-end explainable vulnerability detection?

Specifically, RQ1 and RQ2 target the explainable vulnerability detection task within a given scope of code, which is a primitive task that does not require interaction with the program environment. RQ3 and RQ4 target and require the calling context retrieval, respectively, which involves the interaction with the codebase to retrieve relevant functions. Hence, we use LLMs to investigate RQ1 and RQ2, and an LLM-based agent to investigate RQ3 and RQ4. For the primary metrics (F1 score and coverage score), we report 95% confidence intervals computed via bootstrap resampling over the evaluation instances using 10,000 samples to assess statistical stability.

5.1 RQ1: Detection with Calling Contexts

We evaluate eight LLMs on the explainable vulnerability detection task. The details of LLM settings are provided in Appendix G. For each task instance, we provide the LLMs with a specific program location that may trigger a vulnerability, together with all relevant functions in the calling context, and ask the LLMs to determine whether the vulnerability

Table 2: The performance of the LLMs in explainable vulnerability detection. **P**, **R**, and **F1** indicate precision, recall, and F1 score, respectively. **Cov** indicates typestate-slice coverage. **Time** and **Cost** indicate the total execution time and financial cost, respectively.

Model	P (%)	R (%)	F1	Cov	Time (s)	Cost ($)
GPT 5	74.36	71.12	0.73	0.50	13,647	1.55
Gemini 2.5 Pro	69.11	81.37	**0.75**	**0.57**	18,740	2.96
Claude Sonnet 4.5	63.02	73.21	0.68	0.49	16,845	5.22
GPT 5 mini	67.08	64.84	0.66	0.49	11,746	0.40
Gemini 2.5 Flash	62.41	71.02	0.66	0.52	20,139	0.72
Claude Haiku 4.5	74.05	75.63	**0.75**	0.51	19,814	1.74
Qwen3 Coder 30B	73.11	47.02	0.57	0.30	7,842	N/A
GPT OSS 120B	75.24	62.88	0.69	0.45	7,015	N/A

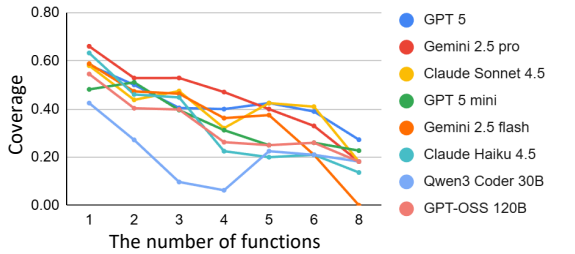


Figure 3: The typestate-slice coverage under increasing inter-procedural complexity.

exists or not and also generate a typestate slice as the formalized explanation of the vulnerability. The detailed task design and evaluation metrics are described in Section 4.3.1.

As shown in Table 2, all evaluated models achieve better performance on vulnerability identification, with an average F1 score of 0.69. However, the average coverage of the generated typestate slices is only 0.48. This discrepancy indicates that, although models can often predict the correct label, they frequently fail to recover the correct typestate slice, which represents the actual root cause and is critical in real-world vulnerability detection scenarios. Without a correct explanation, developers cannot reliably locate and fix the vulnerability, even when the model produces a correct prediction. Moreover, models may generate typestate slices that correspond to safe execution paths while still predicting the correct vulnerability label, highlighting a gap between existing evaluation metrics and the requirements of realistic vulnerability detection. The detailed case study is provided in Appendix J.

Figure 3 further illustrates typestate slices coverage under increasing inter-procedural complexity. As the number of functions involved in a typestate slice increases, the coverage scores of all models decline, highlighting that reasoning over inter-procedural data dependencies remains a major challenge for existing models.

Table 3: Comparison of explainable vulnerability detection performance with and without calling context.

Model	With Context		Without Context	
	F1	Cov	F1	Cov
GPT-5	0.73 ± 0.04	0.50 ± 0.04	0.49 ± 0.04	0.27 ± 0.03
Gemini 2.5 Pro	0.75 ± 0.04	0.57 ± 0.05	0.62 ± 0.04	0.32 ± 0.04
Claude Sonnet 4.5	0.68 ± 0.04	0.49 ± 0.05	0.50 ± 0.05	0.29 ± 0.03
GPT-5 Mini	0.66 ± 0.05	0.49 ± 0.05	0.39 ± 0.04	0.23 ± 0.03
Gemini 2.5 Flash	0.66 ± 0.04	0.52 ± 0.04	0.50 ± 0.03	0.26 ± 0.04
Claude Haiku 4.5	0.75 ± 0.03	0.51 ± 0.04	0.63 ± 0.04	0.28 ± 0.03
Qwen3 Coder 30B	0.57 ± 0.04	0.30 ± 0.04	0.48 ± 0.05	0.13 ± 0.03
GPT-OSS 120B	0.69 ± 0.03	0.45 ± 0.04	0.55 ± 0.05	0.16 ± 0.04
Average	**0.69 ± 0.04**	**0.48 ± 0.05**	**0.52 ± 0.04**	**0.24 ± 0.03**

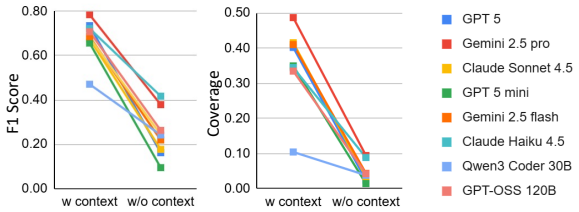


Figure 4: The effect of calling context on explainable vulnerability detection in inter-procedural cases.

> **Finding 1:** LLMs often fail to offer high-quality vulnerability explanations.

5.2 RQ2: Detection without Calling Contexts

To assess the effect of calling context on vulnerability detection, we introduce a controlled experiment in which the model is provided only with the vulnerability triggering program location without calling contexts. The evaluation metrics follow RQ1.

As shown in Table 3, removing calling contexts leads to substantial degradation in both F1 score and typestate-slice coverage. Specifically, the average F1 score drops from 0.69 to 0.52, while the average coverage decreases from 0.48 to 0.24. The performance gap becomes more pronounced in inter-procedural cases. As illustrated in Figure 4, the average F1 score decreases sharply from 0.68 to 0.24, while the average coverage score drops from 0.36 to 0.05. This performance degradation aligns with the intuition: without access to sufficient calling context, it is difficult to draw a precise conclusion about whether a vulnerability exists. Although models may attempt to infer missing context based on prior knowledge, such inferences are highly uncertain and prone to hallucination.

> **Finding 2:** Calling context is indispensable for effective vulnerability detection.

5.3 RQ3: Calling Context Retrieval

To evaluate the calling context retrieval capability of LLM-based coding agents, we select two widely

Table 4: Performance of different agents in the calling context retrieval task. **Time** and **Cost** indicate the total execution time and financial cost, respectively.

Agent	Model	P (%)	R (%)	Time (s)	Cost ($)
OpenHands	GPT 5 mini	32.58	**49.51**	26,181	4.95
	Gemini 2.5 Flash	12.58	26.65	13,189	3.79
	GPT OSS 120B	12.63	16.42	22,162	N/A
	Claude Sonnet 4.5	**33.42**	44.21	28,795	43.12
RepoAudit	GPT 5 mini	27.80	47.47	16,625	0.98
	Gemini 2.5 Flash	25.54	40.16	15,413	1.01
	GPT OSS 120B	17.25	29.26	8,507	N/A
	Claude Sonnet 4.5	**28.52**	**48.31**	19,041	10.25
Claude Code	Claude Sonnet 4.5	**81.13**	**55.04**	18,488	39.61

Table 5: Vulnerability detection and explanation performance of different agents and backbone models.

Agent	Model	P (%)	R (%)	F1	Cov
OpenHands	GPT 5 mini	70.42	61.41	**0.66 ± 0.06**	**0.33 ± 0.05**
	Gemini 2.5 Flash	71.34	57.43	0.64 ± 0.05	0.30 ± 0.04
	GPT OSS 120B	59.36	51.53	0.55 ± 0.06	0.19 ± 0.04
	Claude Sonnet 4.5	69.48	55.59	0.62 ± 0.05	0.31 ± 0.03
RepoAudit	GPT 5 mini	70.61	56.34	**0.63 ± 0.04**	0.38 ± 0.05
	Gemini 2.5 Flash	72.51	52.50	0.61 ± 0.05	**0.40 ± 0.05**
	GPT OSS 120B	74.42	45.24	0.56 ± 0.05	0.29 ± 0.04
	Claude Sonnet 4.5	71.53	56.32	**0.63 ± 0.04**	0.38 ± 0.04
Claude Code	Claude Sonnet 4.5	72.35	67.05	**0.70 ± 0.04**	**0.44 ± 0.04**

used code agents, Claude Code (Anthropic) and OpenHands (Wang et al., 2024c), and a state-of-the-art vulnerability detection agent RepoAudit (Guo et al., 2025b), as baselines. Detailed descriptions of these agents are provided in Appendix H. Each agent is provided with a vulnerability triggering code and the corresponding function, and is tasked with retrieving the calling context required for vulnerability detection. The detailed task design and evaluation metrics are described in Section 4.3.2.

As shown in Table 4, Claude Code with Sonnet 4.5 achieves the best performance, with a precision of 81.13% and a recall of 55.04%. In comparison, OpenHands and RepoAudit achieve comparable recall but substantially lower precision, indicating less effective filtering of irrelevant functions during calling-context retrieval. In terms of efficiency, OpenHands performs the worst. RepoAudit achieves the lowest financial cost. This difference arises because both OpenHands and Claude Code rely on LLM-generated grep-style searches and must process a large volume of retrieved results, whereas RepoAudit retrieves calling contexts based on a preconstructed call graph, thereby reducing the amount of context passed to the LLM.

Finding 3: Precise and efficient calling context retrieval remains a significant challenge for LLM-based coding agents.

5.4 RQ4: End-to-End Detection

We evaluate the end-to-end vulnerability detection capability of existing LLM-based coding agents. The input setting remains the same as that of RQ3. Each agent is required to first retrieve relevant calling context from the codebase and then perform explainable vulnerability detection.

The results are summarized in Table 5. Claude Code achieves the best overall performance, with an F1 score of 0.70 and a typestate-slice coverage of 0.43. Our analysis indicates that the low typestate coverage mainly stems from early termination due to wrong assumptions (44.09%) and incorrect reasoning (41.73%). Among the three vulnerability types, UAF exhibits the highest inter-procedural and aliasing complexity, leading to the worst performance with an average coverage of 0.38. This suggests that current LLM-based agents still struggle with complex data-flow reasoning involving deep inter-procedural propagation and heap aliasing. Detailed analysis is provided in Appendix I.

Compared with the results in RQ2, all agents outperform LLMs operating without calling context in explainable vulnerability detection. However, due to the limitation in calling context retrieval capability, their coverage scores remain lower than those observed in RQ1 (0.49 for Claude Sonnet 4.5 and 0.53 for Gemini 2.5 Flash), where LLMs are provided with sufficient calling context.

Finding 4: Calling context retrieval is essential for effective vulnerability detection by LLM-based coding agents.

6 Conclusion

We present ManBo, a benchmark designed to evaluate vulnerability detection under realistic, repository-scale settings. We formulate two core tasks, namely calling context retrieval and explainable vulnerability detection, and evaluates existing LLMs and LLM-driven coding agents on both tasks. Our evaluation reveals that deficiencies in calling context retrieval constitute a major bottleneck, substantially hindering the effectiveness of end-to-end explainable vulnerability detection. We expect ManBo to serve as a foundation for future research on code analysis in real-world codebases, extending beyond vulnerability detection to broader program reasoning tasks. We have uploaded the benchmark as supplementary data on OpenReview and will release it publicly upon acceptance.

7 Limitations

ManBo has several limitations that suggest directions for future extensions. First, the benchmark currently focuses on calling context for repository-level reasoning. Program values propagated through global variables or concurrent interactions across threads are not explicitly modeled, which limits coverage of certain real-world vulnerabilities in concurrent or stateful systems. Second, ManBo targets typestate-related vulnerabilities, with a focus on three representative memory corruption issues. Other vulnerability types, such as buffer overflows or numeric errors, are not yet included. Nevertheless, the data format introduced in Section 4.2 is general and can be naturally extended to support additional vulnerability categories. Finally, the benchmark construction relies primarily on manual annotation, which incurs non-trivial human effort and may introduce minor labeling errors despite multi-annotator cross-validation. A promising direction is to integrate static analysis tools to automatically generate candidate typestate paths, followed by human verification and refinement, to reduce annotation cost while maintaining data quality.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. Claude code: AI-powered coding assistant for developers.

Steven Arzt, Siegfried Rasthofer, Christian Fritz, Eric Bodden, Alexandre Bartel, Jacques Klein, Yves Le Traon, Damien Octeau, and Patrick McDaniel. 2014. Flowdroid: Precise context, flow, field, object-sensitive and lifecycle-aware taint analysis for android apps. *ACM sigplan notices*, 49(6):259–269.

Thomas Ball and Sriram K Rajamani. 2002. The slam project: debugging system software via static analysis. In *Proceedings of the 29th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, pages 1–3.

Thoms Ball. 1999. The concept of dynamic analysis. *ACM SIGSOFT Software Engineering Notes*, 24(6):216–234.

Max Brunsfeld. 2018. Tree-sitter-a new parsing system for programming tools. In *Strange Loop Conference,. Accessed–. URL: https://www. thestrangeloop. com//tree-sitter—a-new-parsing-system-for-programming-tools. html*.

Juan Caballero, Gustavo Grieco, Mark Marron, and Antonio Nappa. 2012. Undangle: early detection of dangling pointers in use-after-free and double-free vulnerabilities. In *Proceedings of the 2012 International Symposium on Software Testing and Analysis*, pages 133–143.

Cristian Cadar, Daniel Dunbar, and Dawson R. Engler. 2008. KLEE: unassisted and automatic generation of high-coverage tests for complex systems programs. In *8th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2008, December 8-10, 2008, San Diego, California, USA, Proceedings*, pages 209–224. USENIX Association.

Mark Chen. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Yizheng Chen, Zhoujie Ding, Lamya Alowain, Xinyun Chen, and David Wagner. 2023. Diversevul: A new vulnerable source code dataset for deep learning based vulnerability detection. In *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses*, pages 654–668.

Clang. 2025. Clang: a C language family frontend for LLVM. https://clang.llvm.org/. [Online; accessed 29-May-2025].

CodeRabbit. 2025. Coderabbit: Ai-powered code review for github. https://www.coderabbit.ai/. Accessed: 2025-07-07.

Cursor. 2025. Bugbot documentation. https://docs.cursor.com/bugbot. Accessed: 2025-06-26.

Manuvir Das, Sorin Lerner, and Mark Seigle. 2002. Esp: Path-sensitive program verification in polynomial time. In *Proceedings of the ACM SIGPLAN 2002 Conference on Programming language design and implementation*, pages 57–68.

Yangruibo Ding, Yanjun Fu, Omniyyah Ibrahim, Chawin Sitawarin, Xinyun Chen, Basel Alomair, David Wagner, Baishakhi Ray, and Yizheng Chen. 2024. Vulnerability detection with code language models: How far are we? *arXiv preprint arXiv:2403.18624*.

Lisa Nguyen Quang Do and Eric Bodden. 2020. Explaining static analysis with rule graphs. *IEEE Transactions on Software Engineering*, 48(2):678–690.

Michael D Ernst, Jake Cockrell, William G Griswold, and David Notkin. 2001. Dynamically discovering likely program invariants to support program evolution. *IEEE transactions on software engineering*, 27(2):99–123.

Jiahao Fan, Yi Li, Shaohua Wang, and Tien N Nguyen. 2020. Ac/c++ code vulnerability dataset with code changes and cve summaries. In *Proceedings of the 17th international conference on mining software repositories*, pages 508–512.

Stephen J Fink, Eran Yahav, Nurit Dor, G Ramalingam, and Emmanuel Geay. 2008. Effective typestate verification in the presence of aliasing. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 17(2):1–34.

GitHub. 2025. CodeQL. https://codeql.github.com/. [Online; accessed 13-July-2025].

Jinyao Guo, Chengpeng Wang, Dominic Deluca, Jinjie Liu, Zhuo Zhang, and Xiangyu Zhang. 2025a. Bugscope: Learn to find bugs like human. *arXiv preprint arXiv:2507.15671*.

Jinyao Guo, Chengpeng Wang, Xiangzhe Xu, Zian Su, and Xiangyu Zhang. 2025b. Repoaudit: An autonomous llm-agent for repository-level code auditing. *arXiv preprint arXiv:2501.18160*.

Maria Jump and Kathryn S McKinley. 2007. Cork: dynamic memory leak detection for garbage-collected languages. In *Proceedings of the 34th annual ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, pages 31–38.

René Just, Darioush Jalali, and Michael D Ernst. 2014. Defects4j: A database of existing faults to enable controlled testing studies for java programs. In *Proceedings of the 2014 international symposium on software testing and analysis*, pages 437–440.

Chris Lattner and Vikram Adve. 2004. Llvm: A compilation framework for lifelong program analysis & transformation. In *International symposium on code generation and optimization, 2004. CGO 2004.*, pages 75–86. IEEE.

Tuo Li, Jia-Ju Bai, Yulei Sui, and Shi-Min Hu. 2022. Path-sensitive and alias-aware typestate analysis for detecting os bugs. In *Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 859–872.

Ziyang Li, Saikat Dutta, and Mayur Naik. 2025. Iris: Llm-assisted static analysis for detecting security vulnerabilities. In *The Thirteenth International Conference on Learning Representations*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Yunlong Lyu, Yi Fang, Yiwei Zhang, Qibin Sun, Siqi Ma, Elisa Bertino, Kangjie Lu, and Juanru Li. 2022. Goshawk: Hunting memory corruptions via structure-aware and object-centric memory operation synopsis. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 2096–2113. IEEE.

Meta. 2025. Infer Static Analyzer. https://fbinfer.com/. [Online; accessed 13-July-2025].

MITRE. 2025. Common Weakness Enumeration (CWE). https://cwe.mitre.org/index.html. Accessed: 2025-07-11.

Georgios Nikitopoulos, Konstantina Dritsa, Panos Louridas, and Dimitris Mitropoulos. 2021. Crossvul: a cross-language vulnerability dataset with commit data. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 1565–1569.

Jiayi Pan, Xingyao Wang, Graham Neubig, Navdeep Jaitly, Heng Ji, Alane Suhr, and Yizhe Zhang. 2024. Training software engineering agents and verifiers with swe-gym. *arXiv preprint arXiv:2412.21139*.

Abhinav Rastogi, Adam Yang, Albert Q Jiang, Alexander H Liu, Alexandre Sablayrolles, Amélie Héliou, Amélie Martin, Anmol Agarwal, Andy Ehrenberg, Andy Lo, and 1 others. 2025. Devstral: Fine-tuning language models for coding agent applications. *arXiv preprint arXiv:2509.25193*.

Henry Gordon Rice. 1953. Classes of recursively enumerable sets and their decision problems. *Transactions of the American Mathematical society*, 74(2):358–366.

Niklas Risse, Jing Liu, and Marcel Böhme. 2025. Top score on the wrong exam: On benchmarking in machine learning for vulnerability detection. *Proceedings of the ACM on Software Engineering*, 2(ISSTA):388–410.

Konstantin Serebryany, Derek Bruening, Alexander Potapenko, and Dmitriy Vyukov. 2012. {AddressSanitizer}: A fast address sanity checker. In *2012 USENIX annual technical conference (USENIX ATC 12)*, pages 309–318.

Ekaterina Shemetova, Ivan Smirnov, Anton Alekseev, Ilya Shenbin, Alexey Rukhovich, Sergey Nikolenko, Vadim Lomshakov, and Irina Piontkovskaya. 2025. Lamed: Llm-generated annotations for memory leak detection. In *Proceedings of the 29th International Conference on Evaluation and Assessment in Software Engineering*, pages 1024–1034.

Saurabh Sinha, Alessandro Orso, and Mary Jean Harrold. 2004. Automated support for development, maintenance, and testing in the presence of implicit flow control. In *Proceedings. 26th International Conference on Software Engineering*, pages 336–345. IEEE.

Yannis Smaragdakis, George Balatsouras, and 1 others. 2015. Pointer analysis. *Foundations and Trends® in Programming Languages*, 2(1):1–69.

Benjamin Steenhoek, Hongyang Gao, and Wei Le. 2024. Dataflow analysis-inspired deep learning for efficient vulnerability detection. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering, ICSE 2024, Lisbon, Portugal, April 14-20, 2024*, pages 16:1–16:13. ACM.

Robert E. Strom and Shaula Yemini. 1986. Typestate: A programming language concept for enhancing software reliability. *IEEE Transactions on Software Engineering*, SE-12(1):157–171.

Yulei Sui and Jingling Xue. 2016. SVF: interprocedural static value-flow analysis in LLVM. In *Proceedings of the 25th International Conference on Compiler Construction, CC 2016, Barcelona, Spain, March 12-18, 2016*, pages 265–266. ACM.

Hugo Touvron and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Raja Vallée-Rai, Phong Co, Etienne Gagnon, Laurie J. Hendren, Patrick Lam, and Vijay Sundaresan. 1999. Soot - a java bytecode optimization framework. In *Proceedings of the 1999 conference of the Centre for Advanced Studies on Collaborative Research, November 8-11, 1999, Mississauga, Ontario, Canada*, page 13. IBM.

Chengpeng Wang, Wuqi Zhang, Zian Su, Xiangzhe Xu, Xiaoheng Xie, and Xiangyu Zhang. 2024a. LLMDFA: Analyzing dataflow in code with large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Xinchen Wang, Ruida Hu, Cuiyun Gao, Xin-Cheng Wen, Yujia Chen, and Qing Liao. 2024b. Reposvul: A repository-level high-quality vulnerability dataset. In *Proceedings of the 2024 IEEE/ACM 46th International Conference on Software Engineering: Companion Proceedings*, pages 472–483.

Xingyao Wang, Boxuan Li, Yufan Song, Frank F Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, and 1 others. 2024c. Openhands: An open platform for ai software developers as generalist agents. *arXiv preprint arXiv:2407.16741*.

Xiangzhe Xu, Zian Su, Jinyao Guo, Kaiyuan Zhang, Zhenting Wang, and Xiangyu Zhang. 2024. Prosec: Fortifying code llms with proactive security alignment. *arXiv preprint arXiv:2411.12882*.

Chenyuan Yang, Zijie Zhao, Zichen Xie, Haoyu Li, and Lingming Zhang. 2025. Knighter: Transforming static analysis with llm-synthesized checkers. *Preprint*, arXiv:2503.09002.

Mingwei Zheng, Chengpeng Wang, Xuwei Liu, Jinyao Guo, Shiwei Feng, and Xiangyu Zhang. 2025. An llm agent for functional bug detection in network protocols. *Preprint*, arXiv:2506.00714.

Yaqin Zhou, Shangqing Liu, Jingkai Siow, Xiaoning Du, and Yang Liu. 2019. Devign: Effective vulnerability identification by learning comprehensive program semantics via graph neural networks. *Advances in neural information processing systems*, 32.

A Challenges for Automation

As discussed in Section 4, typestate slices correspond to fundamentally semantic program properties. According to Theorem (Rice, 1953), deciding whether a semantic program property holds or not is inherently undecidable. As a consequence, it is theoretically infeasible to design a fully automated algorithm that can precisely determine such properties and produce perfect ground truth annotations. We initially explored the possibility of leveraging existing static and dynamic analysis techniques to automatically recover the typestate slices. However, our investigation revealed that current techniques are insufficient to substantially reduce the annotation burden in practice.

Existing Static Analyzers. Existing static analysis tools, including SVF (Sui and Xue, 2016), FlowDroid (Arzt et al., 2014), CodeQL (GitHub, 2025), and KLEE (Cadar et al., 2008), predominantly operate on intermediate representations (IRs) rather than directly on source code. The IRs generated by compiler are well suited for foundational analyses such as pointer analysis (Smaragdakis et al., 2015). As a consequence, the outputs of these analyzers typically characterize relationships among IR-level program values instead of source-level variables and expressions. However, LLMs perform reasoning directly over source code. This mismatch introduces a semantic gap between traditional static analysis tools and the source-level vulnerability detection capability that our benchmark is designed to evaluate. As a result, existing static analyzers cannot be directly leveraged to generate source-level typestate slices required by our benchmark. Hence, automating data annotation using existing static analysis tools remains impractical at present.

Existing Dynamic Analyzers. Dynamic analysis (Ernst et al., 2001; Ball, 1999; Sinha et al., 2004) observes program behavior by executing the program on concrete inputs and analyzing properties of its runtime execution (e.g., via traces/profiles and observed variable values at selected program points). It can capture execution-specific behaviors and has been widely used for program understanding and vulnerability detection. However, dynamic analysis (especially when used as raw tracing) does not by itself guarantee a minimal, source-level explanation of a vulnerability, such as a compact typestate-transition sequence or a summarized interprocedural pointer-propagation narrative, without

additional abstraction or inference. In practice, execution traces can remain large and include operations irrelevant to the root cause, which makes post-mortem inspection difficult without further structuring/summarization. We also explored leveraging dynamic analysis to automatically generate vulnerability traces, but found that the generated traces could not precisely or consistently capture the typestate transitions and propagation semantics required for our benchmark.

B Related Work on Typestate Vulnerability Detection

Our benchmark focuses on typestate vulnerabilities and constructs the dataset covering three representative defect types: null pointer dereference (NPD), memory leak (MLK), and use-after-free (UAF). These defects have long been central to both academic research and industrial software engineering practice due to their prevalence and severe security implications. To better position our work, this section reviews prior research on typestate and memory vulnerability detection, including both traditional program-analysis–based approaches and recent LLM-based detection methods.

Domain-specific Detectors Over the past decades, researchers have developed a wide range of domain-specific detectors targeting particular classes of memory and typestate violations. AddressSanitizer (Serebryany et al., 2012) is a lightweight runtime memory error detector that tracks the state of allocated memory via compiler instrumentation and shadow memory, and is widely used to detect memory-related bugs such as memory leaks and use-after-free errors. Clang Static Analyzer (Clang, 2025) performs path-sensitive symbolic execution over LLVM intermediate representations (IRs), tracking symbolic program states along feasible execution paths to identify potential null pointer dereferences. Infer (Meta, 2025) employs compositional separation-logic–based analysis to infer heap summaries and propagate nullability constraints interprocedurally, enabling scalable detection of null pointer dereference bugs in large codebases. Cork (Jump and McKinley, 2007) detects memory leaks by identifying monotonic heap growth patterns and reporting diagnostic slices that explain the evolution of leaking data structures. Undangle (Caballero et al., 2012) introduces early detection of dangling pointers through runtime tracking to diagnose use-after-free and double-free vulnerabilities. Beyond vulnerability-specific detectors, classic typestate verification systems aim to enforce temporal safety properties and correct API usage protocols. ESP (Das et al., 2002) models program properties as finite-state machines and performs path-sensitive verification guided by property-relevant control flow. SLAM (Ball and Rajamani, 2002) checks API-usage rules using static analysis combined with counterexample-guided abstraction refinement (CEGAR), reducing false positives while maintaining scalability for system software. Overall, these detectors predominantly model program behavior through runtime instrumentation or symbolic reasoning over compiler-generated intermediate representations (Lattner and Adve, 2004; Vallée-Rai et al., 1999), and rely on deterministic rules or formally specified properties to identify violations.

LLM-based Detectors Recent progress in large language models (LLMs) has introduced a complementary paradigm for vulnerability detection, shifting from rule-driven analysis toward semantic reasoning over source code. Unlike traditional analyzers that depend on compiler infrastructures, LLM-based approaches leverage the semantic understanding capabilities of large language models to reason about program behavior directly from source code. In industrial settings, tools such as Bugbot (Cursor, 2025) and CodeRabbit (CodeRabbit, 2025) integrate LLMs into development workflows to automatically analyze code changes and flag potential issues during code submission. In academic research, IRIS (Li et al., 2025) employs LLMs to infer taint specifications, reducing reliance on manually crafted analysis rules. KNighter (Yang et al., 2025) synthesizes symbolic checkers from historical patches to enable customized vulnerability detection. LAMeD (Shemetova et al., 2025) generates function-level annotations using LLMs to guide static analyses for memory leak detection, alleviating manual specification effort and mitigating path explosion. LLMDFA (Wang et al., 2024a) proposes a neural-symbolic framework in which LLMs infer variable dependencies to support data-flow tracking. RepoAudit (Guo et al., 2025b) leverages LLMs to analyze intra-procedural data flows and employs an LLM-based validator to assess the inter-procedural feasibility of identified flows, demonstrating scalable vulnerability auditing on large real-world repositories.

Compared with traditional domain-specific de-

tectors, LLM-based approaches exhibit stronger generalization, enabling direct analysis of source code without compilation and thereby shortening auditing cycles. Furthermore, because LLMs are trained on multilingual code corpora, they naturally support cross-language analysis. However, LLM-based detectors also introduce new challenges. In particular, hallucination may reduce interpretability and reliability, and analyzing large-scale repositories often incurs substantial context-management and computational costs.

C Inter-Annotator Agreement

To ensure the quality of our dataset, we manually annotate the explanation of each vulnerability in the form of typestate slices. All issues are independently annotated by two authors, each with over five years of experience in program analysis. We then compare the annotated typestate slices to determine whether the two annotations are consistent. Overall, the disagreement ratio is 14.15% (29 out of 205 cases). The main sources of disagreement include differences in identifying wrapper functions and ambiguities caused by aliasing through `Load`/`Store` operations. When disagreements occur, a senior third annotator re-examines the case and discusses it with the two annotators to reach a consensus annotation.

D Ablation of Wrapper Functions

To evaluate the impact of wrapper functions on explainable vulnerability detection, we conduct an ablation study. For each task instance, the LLM is provided with a specific program location that may trigger a vulnerability, together with all relevant functions involved in the annotated typestate slice. The model is then asked to determine whether the given location can trigger a vulnerability and to generate a typestate slice as a formalized explanation of the detected issue. For the **Wrapper** setting, the model receives the complete set of relevant functions, including the bodies of wrapped functions inside customized wrapper APIs (e.g., `x_malloc`). In contrast, under the **No-Wrapper** setting, the bodies of wrapped functions are omitted from the provided context while all other relevant functions remain unchanged.

As shown in Table 6, the Wrapper setting achieves an average F1 score of 0.69 and a typestate slice coverage score of 0.48, while the No-Wrapper setting obtains an average F1 score of 0.68 and a coverage score of 0.47. These results indicate that removing wrapped function bodies does not substantially affect vulnerability detection performance. One possible explanation is that LLMs infer the functionality of wrapper functions from semantic cues such as function names and calling patterns, allowing them to approximate the behavior of omitted implementations. Although such inference is not always sound and may introduce false positives in certain cases, wrapper functions that encapsulate standard library calls typically follow predictable conventions. Consequently, the model's inferred semantics are often sufficient for accurate vulnerability detection, resulting in only limited performance degradation.

Table 6: Comparison of explainable vulnerability detection performance with and without wrapper functions.

Model	Wrapper		No-Wrapper	
	F1	Cov	F1	Cov
GPT 5	0.73 ± 0.04	0.50 ± 0.04	0.69 ± 0.04	0.51 ± 0.04
Gemini 2.5 Pro	0.75 ± 0.04	0.57 ± 0.05	0.76 ± 0.04	0.56 ± 0.05
Claude Sonnet 4.5	0.68 ± 0.04	0.49 ± 0.05	0.65 ± 0.05	0.48 ± 0.05
GPT 5 mini	0.66 ± 0.05	0.49 ± 0.06	0.67 ± 0.04	0.47 ± 0.05
Gemini 2.5 Flash	0.66 ± 0.04	0.52 ± 0.04	0.68 ± 0.03	0.52 ± 0.04
Claude Haiku 4.5	0.75 ± 0.03	0.51 ± 0.04	0.72 ± 0.04	0.48 ± 0.03
Qwen3 Coder 30B	0.57 ± 0.04	0.30 ± 0.04	0.58 ± 0.05	0.29 ± 0.04
GPT OSS 120B	0.69 ± 0.03	0.45 ± 0.04	0.70 ± 0.05	0.46 ± 0.04
Average	**0.69 ± 0.04**	**0.48 ± 0.05**	**0.68 ± 0.04**	**0.47 ± 0.04**

E Typestate-Slice Coverage

In this section, we describe the procedure used to compute the typestate slice coverage score. For each vulnerable instance, we treat the annotated typestate slice as the groundtruth representing the minimal sequence of operations required to precisely characterize the vulnerability. From this slice, we extract the set of operations, where each operation is represented as a tuple consisting of a function name and the corresponding source-code line number. We then extract the set of operations from the generated typestate slice and compare it against the groundtruth. Since a single operation may span multiple lines of code (up to four lines in our annotated typestate slices), we consider a ground-truth operation to be covered if there exists a generated operation with the same function name whose line number falls within a window of ±3 lines around the annotated line number. This approximation mitigates mismatches arising when the model correctly identifies the same operation but reports slightly different line numbers. Although this relaxed matching criterion may classify an incorrect operation as correct, such cases require the

generated operation to fall consistently within a narrow positional window around the ground truth, which is rare in practice. Importantly, this criterion reflects the granularity of typestate annotations rather than inflating performance, ensuring that semantically correct slices are not unfairly penalized due to line-level variance. Based on this criterion, the coverage score is defined as the proportion of groundtruth operations covered by the generated typestate slice

$$\text{Coverage} = \frac{|\text{Ops}_{\text{generated}} \cap \text{Ops}_{\text{groundtruth}}|}{|\text{Ops}_{\text{groundtruth}}|}.$$

F Performance of Calling Context Retrieval with multiple vulnerable paths.

In this section, we analyze the impact of multiple vulnerable execution paths on calling-context retrieval performance. We select 14 cases containing multiple vulnerable execution paths, including 9 memory leak cases and 5 null pointer dereference (NPD) cases, with an average of 4.9 operations per case. As a control group, we additionally select 14 cases that contain only a single vulnerable execution path (9 memory leak cases and 5 NPD cases), with an average of 4.8 operations per trace. The experimental setting follows Section 5.4. LLM-based agents are provided with a program location that may trigger a bug and are tasked with retrieving all relevant functions and identifying all vulnerable execution paths that can lead to a bug at the given location.

As shown in Table 7, the average typestate coverage under both settings is 0.21, suggesting that the number of vulnerable execution paths has limited impact on the bug detection performance of LLM-based agents. One reason is that, in both settings, agents must first identify all relevant execution paths reaching the target program location, retrieve the corresponding calling contexts along these paths, and then determine which path leads to the typestate violation. Therefore, the context retrieval process remains essentially identical across the two settings, differing only in the number of vulnerable paths required in the final output. Additionally, different vulnerable execution paths often share substantial portions of their call chains, with higher overlap closer to the bug-triggering location. Since existing agents exhibit relatively low recall in the calling-context retrieval task, they frequently fail to retrieve the branching points that distinguish multiple vulnerable paths. As a result, the observable performance under the two settings becomes similar.

G Details of LLM Settings

We evaluate eight LLMs, including three flagship commercial models: GPT 5, Gemini 2.5 Pro, and Claude Sonnet 4.5; three cost-efficient commercial models: GPT 5 Mini, Gemini 2.5 Flash, and Claude Haiku 4.5; and two open-source models: Qwen3 Coder 30B A3B Instruct (denoted as Qwen3 Coder 30B) and GPT OSS 120B. The open-source models are hosted on our local servers equipped with eight NVIDIA RTX A6000 GPUs, while all commercial models are accessed via their respective platform APIs [1][2][3].

For Gemini and Claude models, we configure an output token limit of 4096 and a reasoning token budget of 2048. For GPT 5 and GPT 5 Mini, which do not expose explicit controls over token limits, we use the default configuration and set the reasoning effort to *medium*. For GPT OSS 120B, we set the output token limit to 4096 and the reasoning budget to 2048 tokens. For Qwen3 Coder 30B, which is not a reasoning model, we set the output token limit to 4096 and the decoding temperature to 0 to minimize randomness.

H Details of Evaluated Agents

To evaluate the calling context retrieval capability of LLM-based coding agents, we select OpenHands (Wang et al., 2024c), one of the most widely adopted coding agents in practice, and RepoAudit (Guo et al., 2025b), a state-of-the-art vulnerability detection agent, as complementary baselines.

OpenHands provides the underlying LLM with an execution environment that allows it to generate and execute search scripts, enabling grep-style context retrieval over the codebase. The execution results are returned to the LLM for further analysis. Completing a single task involves multiple iterations, during which the LLM performs planning, invokes execution tools, analyzes retrieved results, and determines whether the task has been completed. Following common practice in prior work (Rastogi et al., 2025; Pan et al., 2024), we set the maximum number of iterations to 50 during

[1] https://platform.openai.com/docs/models
[2] https://ai.google.dev/gemini-api/docs
[3] https://platform.claude.com/docs/en/about-claude/models/overview

Table 7: Calling-context retrieval performance under single-path and multiple-path settings. **P** and **R** denote precision and recall in the calling-context retrieval task, respectively. **Coverage** indicates typestate-slice coverage.

Agent	Model	Single			Multiple		
		P (%)	R (%)	Coverage	P (%)	R (%)	Coverage
OpenHands	GPT 5 mini	29.14	38.53	**0.29 ± 0.05**	42.34	48.65	**0.31 ± 0.05**
	Gemini 2.5 Flash	15.00	25.34	0.13 ± 0.04	23.26	25.63	0.11 ± 0.04
	GPT OSS 120B	14.49	12.72	0.10 ± 0.05	18.24	11.84	0.09 ± 0.05
	Claude Sonnet 4.5	25.75	34.62	0.25 ± 0.04	39.56	35.99	0.29 ± 0.04
RepoAudit	GPT 5 mini	32.74	36.46	0.24 ± 0.05	41.12	38.37	0.24 ± 0.05
	Gemini 2.5 Flash	45.54	32.65	0.18 ± 0.03	49.98	34.75	0.20 ± 0.04
	GPT OSS 120B	26.24	21.63	0.15 ± 0.04	24.87	26.53	0.14 ± 0.04
	Claude Sonnet 4.5	34.26	37.26	**0.25 ± 0.04**	32.54	39.62	**0.24 ± 0.04**
Claude Code	Claude Sonnet 4.5	71.24	43.31	**0.26 ± 0.03**	82.31	38.56	**0.26 ± 0.03**

evaluation.

While OpenHands is general and effective for broad context exploration, it does not explicitly model the structural properties of the program, such as the call graph. As a complementary baseline, we also evaluate RepoAudit, a vulnerability detection agent that performs calling-context retrieval in a structure-aware manner. RepoAudit first leverages static analysis tools (e.g, Tree-sitter (Brunsfeld, 2018)) to construct a call graph. It then selectively traverses call graph edges to retrieve the relevant calling context. Starting from a given program location, the LLM extracts the data-flow path associated with that location and triggers the retrieval of adjacent functions along the call graph whenever inter-procedural dependencies are involved. During evaluation, we set the maximum call-depth to 8, which matches the maximum number of functions in our dataset.

I Detailed Analysis on End-to-End Detection

In this section, we provide a detailed failure analysis of the end-to-end detection task. We focus on the results produced by Claude Code, which achieves the best overall performance, and analyze the relationship between detection performance and vulnerability types, as well as the factors that cause typestate coverage scores to be substantially lower than vulnerability identification performance.

Vulnerability Type Correlation As shown in Table 8, Claude Code performs worst on UAF vulnerabilities, both in terms of calling-context retrieval quality and final typestate slice coverage. One contributing factor is that UAF typestate slices contain the largest number of operations and functions on average, resulting in substantially higher **inter-procedural complexity**. As illustrated in Figure 3, the coverage of generated typestate slices decreases noticeably as inter-procedural complexity increases.

Another important factor lies in the higher **memory-indirection complexity** of UAF vulnerabilities. As shown in Table 8, 71.72% of UAF cases contain `Load`/`Store` operations in their typestate traces, compared to 61.47% for NPD and 38.74% for memory leaks. These operations introduce additional pointer aliasing and indirect data dependencies, requiring the model to reason over memory-based value propagation rather than explicit variable flows. Alias reasoning across heap objects significantly increases the difficulty of tracking correct typestate transitions, causing the model to miss critical intermediate steps in the ground-truth typestate slice.

Failure Mode Analysis Among the results produced by Claude Code, we identify 127 cases in which the predicted labels are correct but the typestate coverage is below 1, accounting for 61.95% of all evaluated instances. By comparing the generated typestate slices with the ground-truth slices, we categorize these failures into four major types:

1. **Totally Wrong Reasoning**: the model performs incorrect reasoning and hallucinates a different vulnerable path, resulting in false positives.
2. **Partially Wrong Reasoning**: part of the reasoning process is correct, but the model follows an incorrect propagation path during analysis.
3. **Early Stop**: the reasoning process is largely correct, but the model prematurely terminates after making unsupported assumptions (e.g., assuming a parameter value is `null`).

Table 8: Comparison across vulnerability types. **#Operation** and **#Function** denote the average number of operations and functions in typestate slices. **L/S (%)** indicates the proportion of cases containing `Load`/`Store` operations. **P (%)** and **R (%)** denote precision and recall of calling-context retrieval task. **Cov** denotes typestate-slice coverage.

Bug Type	#Operation	#Function	L/S (%)	P (%)	R (%)	F1	Cov
NPD	4.55	2.20	61.47	83.54	61.03	0.73	0.41
MLK	3.72	1.74	38.74	83.02	59.32	0.69	0.51
UAF	6.28	2.82	71.72	74.76	40.40	0.68	0.38

Table 9: Distribution of failure modes for cases with correct predictions but incomplete typestate coverage. **R** denotes the percentage of cases in each category, and **Cov** indicates typestate-slice coverage.

Category	#Cases	R (%)	Cov
Totally Wrong Reasoning	24	18.90	0.208
Partially Wrong Reasoning	29	22.83	0.443
Early Stop	56	44.09	0.486
Missing Propagation Path	18	14.17	0.597

4. **Missing Propagation Path**: the model correctly tracks typestate transitions but fails to capture intermediate propagation steps.

The distribution of failure categories is summarized in Table 9. Among all categories, *Early Stop* is the most prevalent, accounting for 44.09% of failure cases. In these instances, the reasoning trajectory is largely correct; however, the model tends to speculate about unknown variable values (e.g. function parameters) or function behaviors, hallucinating supporting evidence and terminating analysis prematurely. Although we attempt to mitigate this behavior through prompt constraints that discourage unsupported assumptions, the phenomenon persists. Notably, such assumptions not only cause premature termination and reduce explanation quality, but may also lead to incorrect program understanding. As shown in the false-positive case shown in Appendix J.2, the model incorrectly assumes the behavior of the function `JS_CallConstructorInternal`, resulting in an wrong vulnerability report.

Across the remaining categories, totally and partially wrong reasoning together account for 41.73% of failures, while *Missing Propagation Path* contributes 14.17%. In the latter case, although the model correctly identifies typestate-changing operations on the target object, it fails to capture intermediate propagation steps such as load/store operations and inter-procedural data propagation, leading to incomplete typestate slices and low-quality explanations.

J Case Study

In this section, we present representative false positive cases caused by hallucinations in LLM reasoning. Although the models correctly predict the decision as *Vulnerable*, the generated typestate slices do not match actual vulnerabilities.

J.1 False Positive 1

Listing 1 shows a code fragment containing a memory leak vulnerability. The wrapper function `LZ4F_calloc` allocate a memory object at line 3 inside function `LZ4F_createDecompressionContext_advanced`. The allocated memory is then returned to the caller function `LZ4F_createDecompressionContext` and subsequently propagated to `LZ4F_readOpen`. Within `LZ4F_readOpen`, the memory object is stored in the pointer `ret` at line 17. When `LZ4F_readOpen` returns along an error-handling branch at line 27, the allocated memory is not freed, resulting in a memory leak.

In the Explainable Vulnerability Detection task, the LLM correctly predicts the final decision as *Vulnerable*; however, the generated typestate slice does not correspond to the actual vulnerability. Specifically, the model-generated path starts from a `calloc` operation at line 12 in `LZ4F_readOpen` and ends at the same `return` statement at line 27. In this inferred path, the memory is already freed at line 26 prior to the return, and therefore no memory leak occurs along this data flow. As a result, the model produces a false positive at the explanation level for the memory leak.

J.2 False Positive 2

Listing 2 presents a code fragment that contains a use-after-free vulnerability. In function `JS_NewObjectFromShape`, the wrapper function `js_malloc` allocates a memory object at line 6. This object is subsequently returned through `JS_NewObjectProtoClass` and `js_create_from_ctor` to the function

Listing 1: A memory leak example.

```
LZ4F_dctx* LZ4F_createDecompressionContext_advanced(LZ4F_CustomMem customMem, unsigned version)
{
    LZ4F_dctx* const dctx = (LZ4F_dctx*)LZ4F_calloc(sizeof(LZ4F_dctx), customMem);
    if (dctx == NULL) return NULL;

    dctx->cmem = customMem;
    dctx->version = version;
    return dctx;
}

LZ4F_errorCode_t LZ4F_createDecompressionContext(LZ4F_dctx** LZ4F_decompressionContextPtr, unsigned versionNumber)
{
    assert(LZ4F_decompressionContextPtr != NULL);
    RETURN_ERROR_IF(LZ4F_decompressionContextPtr == NULL, parameter_null);

    *LZ4F_decompressionContextPtr = LZ4F_createDecompressionContext_advanced(LZ4F_defaultCMem, versionNumber);
    if (*LZ4F_decompressionContextPtr == NULL) {
        RETURN_ERROR(allocation_failed);
    }
    return LZ4F_OK_NoError;
}

LZ4F_errorCode_t LZ4F_readOpen(LZ4_readFile_t** lz4fRead, FILE* fp)
{
  char buf[LZ4F_HEADER_SIZE_MAX];
  size_t consumedSize;
  LZ4F_errorCode_t ret;
  LZ4F_frameInfo_t info;

  if (fp == NULL || lz4fRead == NULL) {
    return -LZ4F_ERROR_GENERIC;
  }

  *lz4fRead = (LZ4_readFile_t*)calloc(1, sizeof(LZ4_readFile_t));
  if (*lz4fRead == NULL) {
    return -LZ4F_ERROR_allocation_failed;
  }

  ret = LZ4F_createDecompressionContext(&(*lz4fRead)->dctxPtr, LZ4F_getVersion());
  if (LZ4F_isError(ret)) {
    free(*lz4fRead);
    return ret;
  }

  (*lz4fRead)->fp = fp;
  consumedSize = fread(buf, 1, sizeof(buf), (*lz4fRead)->fp);
  if (consumedSize != sizeof(buf)) {
    free(*lz4fRead);
    return -LZ4F_ERROR_GENERIC;
  }

  ret = LZ4F_getFrameInfo((*lz4fRead)->dctxPtr, &info, buf, &consumedSize);
  if (LZ4F_isError(ret)) {
      LZ4F_freeDecompressionContext((*lz4fRead)->dctxPtr);
      free(*lz4fRead);
      return ret;
    }
  ...
}
```

`JS_CallConstructorInternal`. The memory object is then passed as the third argument to the callee function `JS_CallInternal`, where it is stored in the pointer `val` at line 23 and further propagated to the stack pointer `sp` at line 27. Later, the object is loaded from `sp` at line 29 and freed via the wrapper function `JS_FreeValue` at line 30. Subsequently, in the caller function `JS_CallConstructorInternal`, the freed memory is used at line 13, resulting in a use-after-free vulnerability.

When provided with these functions, the LLM correctly predicts the final decision as *Vulnerable* and successfully identifies the memory object returned by `js_create_from_ctor`. However, it incorrectly interprets the operation at line 10 in `JS_CallConstructorInternal` as a store operation that stores the memory object from pointer `obj` into `ret`. In the subsequently generated reasoning path, the model assumes that the memory is freed at line 13 and returned to `JS_CallInternal` at line 14. It then treats the statement `*sp++ = ret_val;` at line 14 in `JS_CallInternal` as the operation that triggers the vulnerability. Although the generated path appears plausible and correctly tracks the target memory object, it is ultimately a false positive caused by the model's hallucinated interpretation of the functionality of `JS_CallInternal`.

Listing 2: A use after free example.

```
1  static JSValue JS_NewObjectFromShape(JSContext *ctx, JSShape *sh, JSClassID class_id)
2  {
3      JSObject *p;
4
5      js_trigger_gc(ctx->rt, sizeof(JSObject));
6      p = js_malloc(ctx, sizeof(JSObject));
7      ...
8      return JS_MKPTR(JS_TAG_OBJECT, p);
9  }

1  JSValue JS_NewObjectProtoClass(JSContext *ctx, JSValueConst proto_val,
2                                 JSClassID class_id)
3  {
4      ...
5     return JS_NewObjectFromShape(ctx, sh, class_id);
6  }

1  static JSValue js_create_from_ctor(JSContext *ctx, JSValueConst ctor,
2                                     int class_id)
3  {
4      ...
5      obj = JS_NewObjectProtoClass(ctx, proto, class_id);
6      JS_FreeValue(ctx, proto);
7      return obj;
8  }

1  static JSValue JS_CallConstructorInternal(JSContext *ctx,
2                                            JSValueConst func_obj,
3                                            JSValueConst new_target,
4                                            int argc, JSValue *argv, int flags)
5  {
6      ...
7      obj = js_create_from_ctor(ctx, new_target, JS_CLASS_OBJECT);
8      if (JS_IsException(obj))
9          return JS_EXCEPTION;
10     ret = JS_CallInternal(ctx, func_obj, obj, new_target, argc, argv, flags);
11     if (JS_VALUE_GET_TAG(ret) == JS_TAG_OBJECT ||
12         JS_IsException(ret)) {
13         JS_FreeValue(ctx, obj);
14         return ret;
15     } else {
16         JS_FreeValue(ctx, ret);
17         return obj;
18     }
19 }

1  static JSValue JS_CallInternal(JSContext *caller_ctx, JSValueConst func_obj,
2                         JSValueConst this_obj, JSValueConst new_target,
3                         int argc, JSValue *argv, int flags)
4  {
5      ...
6      ret_val = JS_CallConstructorInternal(ctx, call_argv[-2],
7                                           call_argv[-1],
8                                           call_argc, call_argv, 0);
9      if (unlikely(JS_IsException(ret_val)))
10         goto exception;
11     for(i = -2; i < call_argc; i++)
12         JS_FreeValue(ctx, call_argv[i]);
13     sp -= call_argc + 2;
14     *sp++ = ret_val;
15     ...
16     JSValue val;
17     uint32_t tag = JS_VALUE_GET_TAG(this_obj);
18     if (likely(tag == JS_TAG_OBJECT))
19         goto normal_this;
20     if (tag == JS_TAG_NULL || tag == JS_TAG_UNDEFINED) {
21         val = JS_DupValue(ctx, ctx->global_obj);
22     } else {
23         val = JS_ToObject(ctx, this_obj);
24         if (JS_IsException(val))
25             goto exception;
26     }
27     *sp++ = val;
28     ...
29     for(pval = local_buf; pval < sp; pval++) {
30         JS_FreeValue(ctx, *pval);
31     }
32     ...
33 }
```