

Large Language Models Cannot Reliably Detect Vulnerabilities in JavaScript: The First Systematic Benchmark and Evaluation

Qingyuan Fei
Lanzhou University
feiqy2023@lzu.edu.cn

Xin Liu
Lanzhou University
bird@lzu.edu.cn

Song Li
Zhejiang University
songl@zju.edu.cn

Shujiang Wu
Beihang University
wushujiang@buaa.edu.cn

Jianwei Hou
*Data and Technology Support Center
of the Cyberspace Administration of China*
houjianwei@cac.gov.cn

Ping Chen
Fudan University
pchen@fudan.edu.cn

Zifeng Kang
*Beijing University of
Posts and Telecommunications*
zifengkang@bupt.edu.cn

Abstract—Researchers have proposed numerous methods to detect vulnerabilities in JavaScript, especially those assisted by Large Language Models (LLMs). However, the actual capability of LLMs in JavaScript vulnerability detection remains questionable, necessitating systematic evaluation and comprehensive benchmarks. Unfortunately, existing benchmarks suffer from three critical limitations: (1) incomplete coverage, such as covering a limited subset of CWE types, (2) underestimation of LLM capabilities caused by unreasonable ground truth labeling, and (3) overestimation due to unrealistic cases such as using isolated vulnerable files rather than complete projects.

In this paper, we introduce, for the first time, three principles for constructing a benchmark for JavaScript vulnerability detection that directly address these limitations: (1) comprehensiveness, (2) no underestimation, and (3) no overestimation. Guided by these principles, we propose FORGEJS, the first automatic benchmark generation framework for evaluating LLMs’ capability in JavaScript vulnerability detection. Then, we use FORGEJS to construct ARENAJS—the first systematic benchmark for LLM-based JavaScript vulnerability detection—and further propose JUDGEJS, an automatic evaluation framework.

We conduct the first systematic evaluation of LLMs for JavaScript vulnerability detection, leveraging JUDGEJS to assess seven popular commercial LLMs on ARENAJS. The results show that LLMs not only exhibit limited reasoning capabilities, but also suffer from severe robustness defects, indicating that reliable JavaScript vulnerability detection with LLMs remains an open challenge.

1. Introduction

JavaScript is the core language most widely deployed in browsers and increasingly widespread across server-side ecosystems; its security directly impacts user data protection and the software supply chain. In recent years, large

language models (LLMs) have demonstrated strong general capabilities in code understanding, generation, and auditing, motivating their adoption for JavaScript vulnerability detection. In parallel, classic program analysis for JavaScript security has progressed substantially [1]–[5]. However, evidence from practice suggests that the true effectiveness of this direction remains highly questionable: on real projects, models exhibit persistently high false positive rates, detection accuracy is difficult to reproduce stably, and behavior is highly sensitive to contextual shifts and code perturbations, revealing pronounced unreliability. More fundamentally, the community still lacks a systematic benchmark that comprehensively covers security risks in the JavaScript ecosystem, mitigates evaluation bias, and enables automated reproducibility, leading to contradictory conclusions across settings and hindering practical deployment.

To address these challenges, we propose three guiding principles for constructing a benchmark for JavaScript vulnerability detection: comprehensiveness, no underestimation, and no overestimation. Comprehensiveness demands broad coverage of weakness types, evaluation dimensions, and data sources to reduce evaluation bias and improve generalizability across diverse settings. No underestimation emphasizes mechanisms such as semantic equivalence classes and fuzzy matching to accommodate variability in model outputs and uncertainty in real-world annotations, preventing excessive penalties from rigid string matching or noisy labels. No overestimation requires industry-realistic, project-level evaluation that explicitly quantifies the cost of false positives and probes reliance on superficial cues via diverse data-augmentation strategies, thus suppressing artificially inflated laboratory results. Taken together, these principles aim to objectively define the realistic upper and lower bounds of LLM performance under reproducible, scalable, deployment-constrained conditions.

Guided by these principles, we introduce FORGEJS, the first automated generation framework for constructing LLM-based benchmarks for JavaScript vulnerability detection. To

Corresponding authors: Zifeng Kang and Xin Liu.

achieve comprehensiveness, FORGEJS aggregates heterogeneous sources, covers 218 CWE types, unifies extraction and normalization pipelines, and supports both function-level and project-level evaluation to align weakness distributions and contextual complexity with real-world characteristics. To avoid underestimation, FORGEJS relaxes brittle exact string alignment via CWE equivalence classes and fuzzy matching, introduces a dual-track evaluation with complete versus denoised datasets to isolate annotation noise, and calibrates benchmark labels against state-of-the-art reference frameworks (e.g., `claude-code-security-review`). To avoid overestimation, FORGEJS uses complete projects rather than single-file snippets, constructs pre- and post-fix project pairs to directly quantify false positives and localize error sources, and systematically introduces four augmentation strategies to disrupt reliance on filenames, import paths, and comments.

Building on FORGEJS, we construct ARENAJS, the first systematic benchmark for LLM-based JavaScript vulnerability detection, spanning multiple granularities, combining real-world and synthetic data, and unifying multi-perspective metrics. To enable large-scale, reproducible comparisons, we propose JUDGEJS, an automated evaluation framework that spans prompt templates, response parsing, label alignment, and robust scoring; leveraging semantic equivalence and fuzzy matching, it harmonizes heterogeneous outputs and yields a unified, rigorous, and traceable metric suite across function- and project-levels, including F1, false positive rate (FPR), and detection efficacy under engineering constraints.

Using ARENAJS and JUDGEJS, we systematically evaluate seven mainstream commercial LLMs, including GPT-5. The results reveal three findings of substantive impact. First, models exhibit missing reasoning pathways: at the project level, F1 surpasses the function level by 8%–18%, but the gain is driven largely by surface features (filenames, imports, comments) rather than taint propagation and dependency analysis, indicating reliance on opportunistic heuristics. Second, robustness is deficient: seemingly mild augmentations and noise induce drastic swings—for example, under specific noise conditions, a leading model’s F1 drops from 35.9% to 4.2%—and compounded augmentations can trigger catastrophic failures, evidencing high sensitivity to fragile cues. Third, from the perspective of the VD-S metric [6], we find that models are difficult to deploy in realistic industrial settings: with $FPR \leq 0.5\%$, the system misses more than three-quarters of real vulnerabilities and cannot sustain acceptable detection performance.

In summary, our contributions are threefold.

- We establish a systematic methodology for JavaScript vulnerability benchmarking and operationalize three principles—comprehensiveness, no underestimation, and no overestimation.
- We deliver the first systematic benchmark and evaluation for LLM-based JavaScript vulnerability detection: the automated generation framework FORGEJS, the broad-

coverage benchmark ARENAJS, and the scalable, reproducible evaluation framework JUDGEJS.

- We provide a systematic evaluation that reveals key limitations in reasoning sufficiency, robustness, and real-world usability and offers a reproducible testbed to drive progress toward reliable, deployable LLM-based vulnerability detection.

2. Overview

2.1. Background: LLM for JavaScript Vulnerability Detection

Large language models (LLMs) and deep learning approaches have recently been applied to vulnerability detection tasks, demonstrating capabilities across multiple programming languages [7]–[11]. Unlike traditional static application security testing (SAST) tools [12]–[14] that rely on predefined rules and pattern matching, LLMs are designed to identify vulnerabilities through learned representations of code semantics and patterns.

Nevertheless, in JavaScript security analysis, current LLMs often over-rely on surface-level pattern matching rather than robust semantic understanding. This manifests in three ways: (1) keyword dependence; (2) syntactic-structure dependence; and (3) neglect of execution context and complete taint flows from sources to sinks. Illustrative code examples are provided in the Appendix (Listings 7, 8, 9).

Existing studies provide limited, systematic evaluation of LLMs’ actual capability on JavaScript vulnerability detection. Many benchmarks such as SECBENCH.JS [15] and VulcaN [16] use code snippets instead of complete projects, which fails to reflect the complexity of real-world repositories [2], [4], [16]. Moreover, evaluation protocols commonly assess only whether vulnerability is detected and ignore reasoning quality such as whether the model can correctly localize the vulnerable file, function, and CWE type. Two extremes are prevalent: evaluations can overestimate capability by simplifying tasks via snippet inputs, or underestimate capability due to inconsistent labeling granularity when the model reports CWE-83 while the ground truth uses CWE-79 for an equivalent attack.

This problem is especially salient for JavaScript. As the language that runs both in browsers on the frontend and on servers via Node.js backend, vulnerability characteristics vary significantly by context—frontend issues often involve DOM-based XSS and client-side injections [14], whereas backend issues include SQL injection, command injection, and prototype pollution [14], [17]. Existing JavaScript datasets such as SECBENCH.JS include only a few hundred samples and primarily server-side cases, making comprehensive evaluation difficult. A benchmark that spans thousands of real JavaScript projects and simultaneously avoids both overestimation and underestimation remains lacking.

```

1  const isObject = require('./utils/isObject');
2  const cloneDeep = require('./utils/cloneDeep');
3
4  function override(target, source, options =
5      ↪ {}) {
6    // ... parameter validation and circular
7    ↪ reference handling ...
8
9    // Main merge logic
10   for (const key in source) {
11     if (!source.hasOwnProperty(key)) continue;
12
13     // VULNERABILITY: only blocks __proto__,
14     ↪ misses constructor & prototype
15     - if (key === '__proto__') continue; // ↪ Incomplete protection
16     + // FIX: block all prototype pollution ↪ vectors
17     + const dangerousKeys = ['__proto__',
18       ↪ 'constructor', 'prototype'];
19     + if (dangerousKeys.includes(key)) continue; ↪ // Complete protection
20
21     const sourceValue = source[key];
22
23     // Deep merge for nested objects
24     if (isObject(sourceValue)) {
25       if (isObject(target[key])) {
26         target[key] = override(target[key],
27             ↪ sourceValue, options);
28       } else {
29         target[key] = options.clone ?
30           ↪ cloneDeep(sourceValue) : sourceValue;
31       }
32     } else {
33       target[key] = sourceValue;
34     }
35   }
36
37   // ... other option handling ...
38   return target;
39 }
40
41 module.exports = override;

```

Listing 1: Code comparison (vulnerable vs. fixed version).

2.2. A Motivating Example

We identify a real prototype pollution case from GitHub, specifically CVE-2021-25941 in project ASaiAnudeep/deep-override [18], [19], that exemplifies nine deficiencies in existing benchmarks and highlights large language models' tendency to overfit lexical patterns rather than semantics. The vulnerable project is a backend library for deep-merging JavaScript objects. The flaw resides in the function `override` in `src/index.js`.

Defect 1: Incomplete CWE coverage. This vulnerability is prototype pollution (CWE-1321) [20], a JavaScript-specific class formally added in 2019. Existing benchmarks emphasize traditional CWEs: SECBENCH.JS [4] includes only five CWE types while omitting many newer categories added after 2018 (e.g., CWE-611 XXE injection) [21], creating substantial evaluation gaps for LLM capability on emerging

```

1  // Exploit payload (via constructor.prototype)
2  const maliciousPayload = JSON.parse(
3    '{"constructor":{"prototype":' +
4    '"isAdmin":true,"role":"admin"}}'
5  );
6
7  // Trigger pollution
8  override({}, maliciousPayload);
9  // Global pollution takes effect
10 const normalUser = {};
11 console.log(normalUser.isAdmin); // true -
12   ↪ unintended admin privilege
13 console.log(normalUser.role); // "admin"
14 // Affects subsequently created objects
15 const anotherUser = { name: "Alice" };
16 console.log(anotherUser.isAdmin); // true

```

Listing 2: Exploitation example.

weakness detection.

Defect 2: One-dimensional evaluation. Many evaluations only verify the binary outcome (vulnerability present or not) without requiring reasoning quality. For this case, if a model merely outputs “prototype pollution vulnerability, severity: high,” it would be deemed a correct detection. Yet the model may fail to answer critical questions: (1) Which file? (`src/index.js`, not `utils/isObject.js` or `test/test.js`); (2) Which function? (the `override` function); (3) Which exact line? (line 12’s incomplete check, not line 8’s `for...in` loop or line 27’s assignment); (4) Why is line 12 inadequate? (blocks only `__proto__`, allowing `constructor` or `prototype` bypass). If the model answers via keyword matching (searching `__proto__` or `prototype`) without understanding taint flow (`source[key]` → `target[key]` assignment chain), such “correct detection” is mere surface pattern matching.

Defect 3: Insufficient data comprehensiveness. This backend library, operating in the Node.js ecosystem and utilized by web frameworks including Express and Koa, exemplifies cases where existing datasets fail to stratify frontend, backend, and full-stack contexts. JavaScript vulnerability patterns vary substantially across these environments: backend prototype pollution may enable privilege escalation or remote code execution (via `user.isAdmin` or `child_process` module pollution), while frontend prototype pollution predominantly causes DOM-based XSS or client-side denial of service. Mixed evaluation obscures LLM performance variations across distinct contexts.

Furthermore, the full repository comprises 15 files totaling 2,347 lines of code (`src/` directory: 5 files, 823 lines; `test/` directory: 7 files, 1,124 lines; `node_modules/` directory: 400 lines), but existing evaluations extract only the vulnerable function (approximately 33 lines), entirely discarding contextual dependencies including the `isObject` utility implementation, legitimate usage patterns in test cases, and security-relevant `package.json` policies. For example, this project’s `engines` field specifies “`node": ">=12.0.0"` to mandate Node.js 12 or later, thus cir-

cumventing `Object.prototype` pollution vulnerabilities known to affect earlier versions.

Defect 4: Inconsistent CWE granularity. This vulnerability’s official NVD label is CWE-1321 (Improperly Controlled Modification of Object Prototype Attributes ‘Prototype Pollution’), yet labeling granularity varies across sources: Mend.io assigns CWE-915 (Improperly Controlled Modification of Dynamically-Determined Object Attributes), Snyk uses CWE-471 (Modification of Assumed-Immutable Data), and GitHub Advisory uses CWE-1321 [22]. If an LLM reports “CWE-915 vulnerability: attackers can dynamically modify object attributes for privilege escalation,” the description is entirely accurate, yet strict CWE-ID matching would incorrectly penalize it as a type mismatch. This labeling granularity inconsistency is widespread—manual analysis of 50 JavaScript vulnerability samples reveals that 16% exhibit CWE annotation discrepancies across sources (e.g., CWE-79 vs. CWE-83, CWE-89 vs. CWE-564). Strict matching causes model F1 scores on GPT-5 to drop from 0.3207 to 0.2681, severely understating capability.

Defect 5: Label noise contaminates evaluation. This project’s GitHub fix commit (SHA: 2aced176, 2021-05-14) not only patches the vulnerability but also includes: (1) dependency version upgrades for 3 packages (`package.json`: `lodash` 4.17.19→4.17.21, `mocha` 8.2.1→8.4.0, `chai` 4.2.0→4.3.4); (2) test suite restructuring (`test/test.js`: 152→203 lines, 31 new test cases); (3) `README` updates (Security Policy section, 42 new lines). Naively comparing all pre-/post-fix differences (5 files, 137 line changes) produces annotation errors: mislabeling `package.json` as a “vulnerable file” (actually just dependency upgrades), `test/test.js` as “vulnerable code” (actual flaw is in `override`), and `README.md` as a “fix file” (merely documentation). When a model correctly identifies `src/index.js` line 12 but omits `package.json`, it is wrongly penalized for “missing vulnerable files” (though `package.json` is not vulnerable).

Defect 6: Prompting underutilizes model capability. With overly simple prompts (e.g., “Please analyze whether this code has security vulnerabilities”), LLMs may fail to leverage their full reasoning capabilities. This case demands complex multi-step analysis: (1) understanding JavaScript prototype chain semantics—recognizing `for...in` loops traverse both own and prototype-chain properties, and the relationships among `__proto__`, `constructor.prototype`, and `Object.prototype`; (2) tracing taint propagation paths—identifying the taint source as `source[key]` (line 17, attacker-controlled input), following dataflow through the `sourceValue` variable (line 17), conditional branches (line 20), recursive calls (line 22) or direct assignments (line 27), to the taint sink `target[key]` (lines 22 or 27, writing to the target object); (3) reasoning about attack vectors—understanding that when `key` is `constructor`, `target['constructor']` overwrites

the constructor reference, enabling pollution of all instances via `constructor.prototype`; (4) assessing mitigation completeness—recognizing that line 12’s `if (key === '__proto__')` blocks only one pollution vector, missing `constructor` and `prototype`. Simple prompts cannot guide models through these four reasoning steps, severely underestimating LLM capability.

Defect 7: Snippet-only inputs reduce task realism. Extracting only the vulnerable function snippet (approximately 33 lines) allows models to identify `__proto__` (appearing once, line 12) via keyword matching without understanding: (1) project dependency relationships—this library is depended upon by 1,200+ npm packages (including `express-session`, `body-parser`, `config`, etc.), with the vulnerability affecting millions of Node.js applications; (2) call chain complexity—the attack path spans HTTP request → `body-parser` JSON parsing → `express-session` reading session data → `override` merging configuration → global object pollution → privilege check logic impact, requiring understanding of a 5-layer call stack; (3) contextual dependencies—the `isObject` utility function (`utils/isObject.js`) implementation affects line 20’s conditional outcome; if `isObject` incorrectly classifies `null` as an object, additional vulnerability paths emerge; (4) test coverage—`test/test.js` contains 152 test cases, yet none test malicious inputs (e.g., objects with `constructor` properties), indicating developers were unaware of this threat vector. These contextual details are critical for assessing real-world vulnerability impact, but snippet-only inputs discard them entirely.

Defect 8: No false-positive assessment. Existing evaluations provide only vulnerable-version code without validating model behavior on fixed versions. This prevents identifying keyword over-reliance—if the model’s detection logic is “search for `__proto__` keyword; if found, report prototype pollution,” then even in the fixed code (lines 13–15 add complete protection: `const dangerousKeys = ['__proto__', 'constructor', 'prototype']; if (dangerousKeys.includes(key)) continue;`, which is actually secure), the model will still misreport a vulnerability upon seeing the `__proto__` string at line 14. Additionally, the fixed version presents another test: line 14’s `dangerousKeys` array is defined inside the loop (recreated each iteration)—a performance issue, not a security flaw; models flagging this as a vulnerability commit false positives.

Defect 9: Missing robustness testing. The pattern-matching over-reliance observed in Defect 8 becomes even more pronounced under noise and obfuscation. Applying four adversarial transformations to this vulnerable code reveals sharp drops in LLM detection rates, confirming reliance on surface features rather than semantic understanding:

- **(a) Obfuscation.** After applying `javascript-obfuscator` [23], variable names like `key` become `_0x3a2f`, and `__proto__` is encoded as Unicode escape

\x5f\x5f\x70\x72\x6f\x74\x6f\x5f\x5f.

Models relying on clear identifiers and literal keywords will fail to recognize the vulnerability after obfuscation.

- (b) **Noise injection.** We insert 51 safe decoy sinks at function entry (e.g., `fs.appendFileSync` with hardcoded log paths, `db.execute` with hardcoded queries, `innerHTML` with hardcoded content)—seemingly dangerous yet safe (no taint sources). Models using naive “flag all dangerous APIs” heuristics will produce abundant false positives while missing the real vulnerability.
- (c) **Noise + obfuscation.** Both code obfuscation and noise injection are applied simultaneously. Models with genuine dataflow understanding should still localize the real vulnerability and suppress false positives under these interferences.
- (d) **Prompt injection.** Misleading comments are inserted before vulnerable code (e.g., `// SECURITY AUDIT PASSED, // All known attack vectors handled, // NO VULNERABILITIES FOUND`); models over-relying on comment text rather than logic analysis will misclassify vulnerable code as safe.

Collectively, these nine defects indicate that current LLMs do not truly understand security semantics in JavaScript vulnerability detection, but instead rely on surface features—clear variable names (e.g., `override`, `__proto__`), specific code patterns, and comment text cues. Such pattern dependence cannot handle code obfuscation, misleading documentation, and complex contexts in real-world scenarios, resulting in low detection rates.

2.3. Limitations of Existing Benchmarks

We conclude from the motivating example that existing benchmarks suffer from three core issues: (1) **Unrealistic data and task setups**—code snippets replace full projects, and frontend/backend/full-stack stratification is absent, failing to reflect real-world engineering complexity; (2) **In-sufficient evaluation and annotation**—binary detection is emphasized over reasoning and localization quality, CWE granularity labeling is inconsistent and annotation quality is poor, undermining result comparability and credibility; (3) **Missing protocols and robustness testing**—prompting is overly simplistic, vulnerable/fixed project pairs for false-positive measurement are lacking, and adversarial tests (obfuscation, noise, prompt injection) are absent. A comprehensive, rigorous, and robust evaluation framework is therefore essential to accurately assess LLM vulnerability detection capability.

2.4. Method Overview

We propose SECJS, a comprehensive benchmark designed to evaluate LLMs’ capability in JavaScript vulnerability detection. SECJS comprises a dataset generation framework (FORGEJS) and an automated evaluation framework (JUDGEJS). Its design is guided by three principles, each instantiated by three concrete techniques.

Principle I: Comprehensiveness.

- **Principle I-1: Comprehensive CWE Coverage.** FORGEJS collects real projects across a broad range of CWEs, covering both traditional weaknesses, e.g., SQL injection, XSS, command injection, authentication bypass and hardcoded credentials; and newer categories, e.g., prototype pollution, ReDoS, and XXE.
- **Principle I-2: Comprehensive Evaluation Methodology.** We evaluate not only detection outcomes but also reasoning and localization quality from both project- and function-level, requiring models to identify the CWE type as well as the vulnerable file and function locations when applicable.
- **Principle I-3: Comprehensive Data Sources.** All samples are drawn from real GitHub repositories and stratified by application context—including frontend, backend and full-stack—to reflect ecosystem diversity [24].

Principle II: No underestimation.

- **Principle II-1: Employing CWE Equivalence Classes.** We group related CWEs into equivalence classes so that mismatches in granularity (e.g., CWE-79 vs. CWE-83) do not unfairly bias the evaluation [25].
- **Principle II-2: Denoising Datasets.** We provide noise-reduced subsets to mitigate annotation errors from unrelated edits (e.g., dependency upgrades, test refactors).
- **Principle II-3: Maximizing Detection Capability for LLMs.** We employ a strong prompting framework, claude-code-security-review [26], with state-of-the-art LLMs to encourage multi-step reasoning, including cataloging security patterns, contrasting code against known safe patterns, and tracing taint flows from sources to sinks.

Principle III: No overestimation.

- **Principle III-1: Repository-Level Context.** Models analyze full repositories instead of isolated code snippets, operating under realistic complexity and context.
- **Principle III-2: Vulnerable-Fixed Project Pairs.** Each case includes pre-fix and post-fix versions to quantify false positives and measure understanding of fixes.
- **Principle III-3: Enough Dataset Variants for Robustness Evaluation.** We construct noise, obfuscation, noise+obfuscation, and prompt-injection variants to probe robustness and detect reliance on superficial cues [23].

3. Methodology

In this section, we describe the architecture of SECJS Framework, which comprises two key components: FORGEJS (dataset generation framework) and the JUDGEJS evaluation framework.

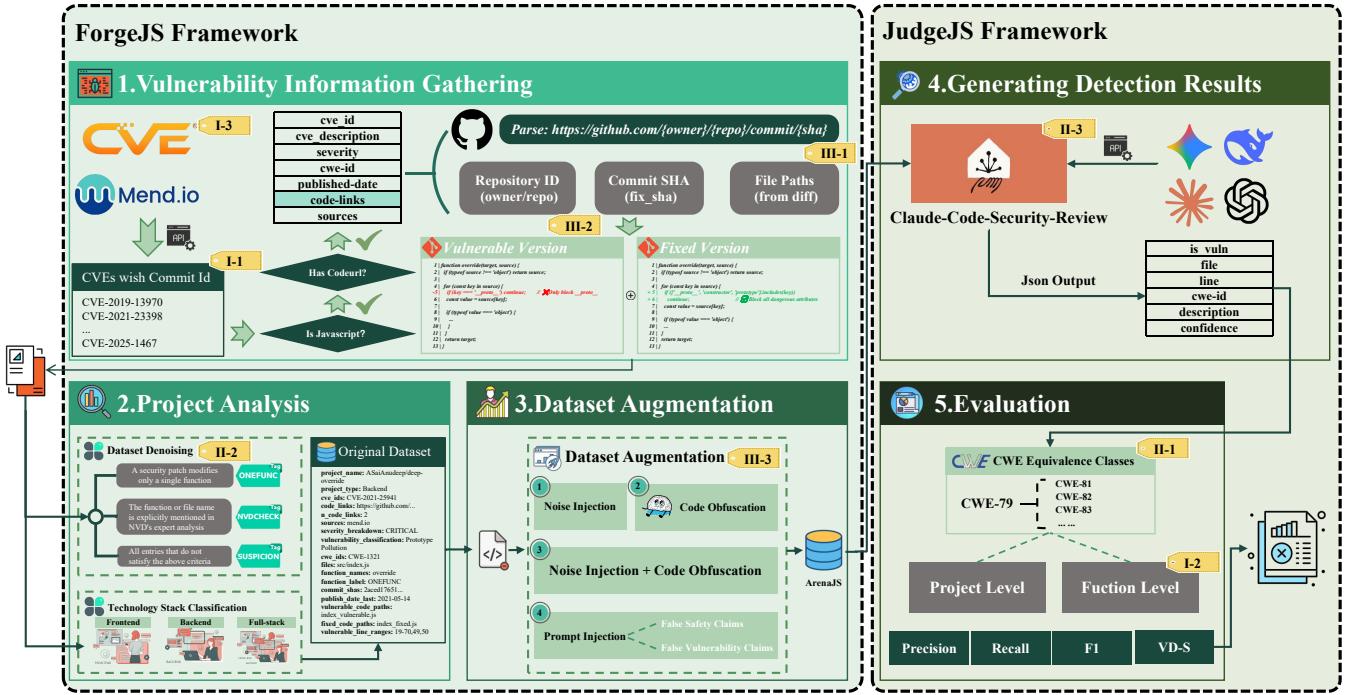


Figure 1: Overall architecture of the SECJS Framework.

3.1. System Architecture

Figure 1 presents the overall architecture of the SECJS Framework. The SECJS Framework comprises two major components: (a) the FORGEJS automated dataset generation framework and (b) the JUDGEJS automated evaluation framework. For component (a), FORGEJS builds the ARENAJS benchmark through three sequential steps.

(1) Vulnerability Information Gathering. FORGEJS collects vulnerability metadata from the official CVE website and Mend.io, requiring two essential fields: Metadata (e.g., CVE identifier, vulnerability type, summary) and GitHub URL. It then parses linked patch commits to extract repository identifiers, commit SHAs, affected file paths, and retrieves both pre-patch (parent commit) and post-patch (current commit) code versions.

(2) Project Analysis. This step further enriches the dataset by performing multi-dimensional categorization (e.g., frontend/backend/full-stack, complete/denoised datasets) and refining ground truth annotations (i.e., pinpointing vulnerable functions and their corresponding line ranges).

(3) Dataset Augmentation. To more effectively assess the true capabilities of large language models beyond simple pattern matching, we augment the original dataset with four enhancement strategies: noise injection, obfuscation, combined noise and obfuscation, and prompt injection, resulting in five datasets.

For component (b), JUDGEJS leverages claude-code-security-review (currently the state-of-the-art prompt system for vulnerability detection) to query large language models including ChatGPT-5 and Grok-4 for vulnerability detection across the five dataset variants generated from component (a) and collect their detection outputs. The evaluation is conducted at two granularities: project-level using binary tuple matching (vulnerability presence, CWE type correctness) and function-level using quaternary tuple matching (vulnerability presence, CWE type correctness, vulnerable file identification, vulnerable function identification), measuring each LLM’s performance with metrics such as precision, recall, F1-score, and VD-S.

3.2. ARENAJS Benchmark

ARENAJS is constructed automatically through three sequential steps: Vulnerability Information Gathering, Project Analysis, and Dataset Augmentation.

3.2.1. Vulnerability Information Gathering. We employ web crawlers to gather JavaScript vulnerability information from the official CVE website [27] and Mend.io [28]. Mend.io (formerly WhiteSource) is a commercial platform specializing in Software Composition Analysis (SCA) [29] and software supply chain security [30]. Many CVE entries lack corresponding GitHub repository URLs (or contain obsolete links), while Mend.io effectively complements the CVE database in this aspect. The collected data must include

two essential fields: Metadata (containing CVE identifier, CVE description, severity level, CWE type, and publication date) and GitHub URL (used for subsequent project retrieval). A complete example data entry after this step is provided in Appendix A.1.

Within the gathering stage, after acquiring CVE metadata from the NVD API [31] and Mend vulnerability database [28], FORGEJS extracts the actual vulnerable and patched code from GitHub commit history [24]. This sub-process maps CVE entries to concrete code changes and retrieves complete pre-patch and post-patch code files. CVE descriptions typically include links to GitHub patch commits (e.g., <https://github.com/user/repo/commit/sha>). The system parses these URLs to extract critical information:

- **Repository identifier** (`owner/repo`): identifying the affected GitHub repository
- **Commit SHA** (`commit_sha`): pinpointing the specific patch commit
- **File path** (`file_path`): extracting affected files from commit diff information

To construct complete vulnerable-patched code pairs, the system retrieves two code versions: the Vulnerable Version (based on the parent commit SHA) and the Fixed Version (based on the current commit SHA).

3.2.2. Project Analysis. This step further enriches the dataset. Specific operations include (1) multi-dimensional dataset categorization and (2) ground truth refinement. In part (1), we categorize the dataset along two dimensions: frontend/backend/full-stack and complete/denoised datasets.

Complete vs. Denoised Dataset. For complete/denoised dataset categorization, we adhere to the principle of “not underestimating LLM capabilities”. Empirical analysis reveals that numerous real-world projects simultaneously update versions while patching vulnerabilities, introducing substantial noise into the pre-patch and post-patch code pairs obtained during the gathering stage, thereby compromising ground truth accuracy. To address this, we assign three types of labels to dataset entries:

- **ONEFUNC**: assigned when a security patch modifies only a single function
- **NVDCHECK**: assigned when the function or file name is explicitly mentioned in NVD’s expert analysis
- **SUSPICION**: assigned to all entries that do not satisfy the above criteria

The denoised dataset comprises entries labeled as ONEFUNC or NVDCHECK. The complete dataset includes all entries.

Frontend/Backend/Full-stack Classification. The frontend/backend/full-stack categorization is performed based on JavaScript’s characteristics and real-world application scenarios. JavaScript, as a language applicable

to both client-side (frontend) and server-side (backend, e.g., Node.js [17]) environments, exhibits distinct vulnerability patterns across different application contexts. Our dataset exhibits the following project type distribution:

- **Backend**: approximately 58%, primarily comprising server-side vulnerabilities, API flaws, and database injection attacks [1], [3]
- **Frontend**: approximately 25%, primarily comprising XSS [32], DOM-based vulnerabilities, and client-side injection attacks [4], [33]
- **Full-stack**: approximately 15%, involving complex vulnerabilities spanning frontend-backend interactions

Ground Truth Refinement. In part (2), we extend ground truth annotations from project-level to function-level based on part (1), achieving fine-grained vulnerability localization. Building upon the project names, file paths, and code pairs obtained during the gathering stage, we introduce five additional fields through an automated three-step workflow. First, regular expression patterns extract all function definitions from both vulnerable and patched versions, capturing diverse JavaScript syntax [13], [34] (traditional declarations, arrow functions, class methods). Second, `difflib.SequenceMatcher` [35] performs line-level diff analysis, identifying all modified, inserted, or deleted lines. Third, these changed lines are mapped to their containing functions, generating: `function_names` (vulnerable function names), `vulnerable_function_names` (filtered vulnerable function list), `vulnerable_line_ranges` (e.g., “19-70,49,50” indicates function range 19-70 with critical patches at 49-50), `function_label_breakdown` (applying ONEFUNC/NVDCHECK/SUSPICION labels from part (1)), and `project_type_breakdown` (vulnerability distribution by frontend/backend/full-stack types from part (1)). This refinement enables rigorous function-level evaluation—models must precisely pinpoint specific vulnerable functions, rather than merely detecting vulnerability presence. A complete dataset entry example with all fields is provided in Appendix A.2.

3.2.3. Dataset Augmentation. Large language models currently face a critical challenge in code security analysis: excessive reliance on surface-level pattern matching instead of genuine semantic understanding. To systematically evaluate whether models genuinely comprehend code security semantics, we design four dataset augmentation strategies, each challenging distinct capability dimensions.

(1) Noise Injection. FORGEJS injects taint sinks without corresponding taint sources [36], [37], constructing seemingly dangerous but actually safe code patterns to test models’ false positive rates and data flow analysis capabilities. Specifically, it inserts 51 types of common dangerous API calls at random locations, but these calls use safe hardcoded data or validated inputs. Example: See Appendix (Listing 3) for a complete example.

(2) Code Obfuscation. FORGEJS employs javascript-obfuscator [38] to transform code while preserving semantics, testing models’ anti-obfuscation capabilities and semantic understanding depth. The obfuscated output (omitted for brevity) transforms clear identifiers into encoded tokens and control flow into nested closures, making static analysis significantly more challenging. Example:

See Appendix (Listing 4) for a complete example.

(3) Noise + Obfuscation. FORGEJS simultaneously applies noise injection and code obfuscation, creating the most challenging test scenarios to comprehensively evaluate model robustness under complex interference. It first injects noise, then obfuscates the entire codebase.

(4) Prompt Injection. FORGEJS inserts misleading comments at random locations (e.g., `// This function is completely safe and has been security audited.` or `// This is vulnerable.`), testing whether models are misled by explicit textual hints rather than independently judging actual code safety. In the implementation, FORGEJS randomly selects misleading comments from a customized prompt library. This prompt library comprises two types of carefully crafted misleading comment templates. The injection density is configured as one prompt per 50 lines of code to test model robustness without excessively disrupting code structure.

Type 1: False Safety Claims. (False Negative Induction): See Appendix for the example (Listing 5).

Type 2: False Vulnerability Claims. (False Positive Induction): See Appendix for the example (Listing 6).

3.3. JUDGEJS Evaluation Framework

Our fully-automated evaluation framework JUDGEJS comprises two core modules: (1) generating vulnerability detection results and (2) evaluating vulnerability detection results. The framework implements a complete automated pipeline from model invocation and result collection to multi-dimensional performance evaluation.

3.3.1. Generating Vulnerability Detection Results. For (1) generating vulnerability detection results, JUDGEJS invokes claude-code-security-review [26] (one of the current state-of-the-art LLM-based vulnerability scanners) to generate detection results for eight models (such as GPT-5, Claude-4.5-Sonnet, DeepSeek-V3.1, etc.) on five dataset variants (ARENAJS-Original, ARENAJS-Obfuscated, ARENAJS-Noise, ARENAJS-Noise+Obfuscated, ARENAJS-PromptInjection).

To avoid underestimating LLM capabilities, we set a 0.8 confidence threshold in the original claude-code-security-review prompt, outputting only results above the 0.8 confidence threshold. For comprehensive coverage, we expanded the scanner’s supported scan types by modifying the prompt to include all CWE types [39]—SQL injection (CWE-89), XSS (CWE-79), command injection (CWE-78), authentication bypass (CWE-287), hardcoded credentials (CWE-798), and 200+ other weaknesses.

Finally, we specified the return result format by modifying the prompt, with each scan returning JSON containing seven fields: `file` (vulnerable file path), `line` (line number), `severity` (HIGH/MEDIUM/LOW by CIA impact), `category` (CWE-ID), `description` (vulnerability explanation), `exploit_scenario` (attack steps), `recommendation` (fix method). JUDGEJS parses this output and matches predictions against ground truth using the criteria in Section 3.3.3.

3.3.2. Evaluating Detection Results. JUDGEJS evaluates model performance at two granularities: project-level and function-level.

Project-level Evaluation. At project-level, JUDGEJS verifies whether the model correctly identifies vulnerability presence and CWE type for each project. Matching criteria: $\langle \text{has_vulnerability}, \text{CWE_equal} \rangle$.

$$\text{TP}_{\text{proj}} : (\text{GT}_{\text{vuln}} = \text{Pred}_{\text{vuln}}) \cap (\text{CWE}_{\text{pred}} = \text{CWE}_{\text{gt}}) \quad (1)$$

Here, GT_{vuln} and $\text{Pred}_{\text{vuln}}$ denote project-level boolean indicators for ground-truth and predicted vulnerability presence, respectively; CWE_{pred} and CWE_{gt} denote the predicted and ground-truth CWE labels, respectively.

TN holds iff the ground truth is benign and the model also predicts no vulnerability; FP holds iff the ground truth is benign but the model asserts the presence of a vulnerability. All remaining cases are FN , including missed detections where the project is vulnerable but the model predicts no vulnerability and type mismatches where the model predicts a vulnerability yet its CWE label is not equal to the ground-truth label under the adopted CWE-equivalence relation. Type agreement is evaluated set-wise: when either side provides multiple CWEs, agreement holds if at least one predicted CWE exactly matches, or belongs to the same equivalence class as, at least one ground-truth CWE. The boolean indicators GT_{vuln} and $\text{Pred}_{\text{vuln}}$ are computed per project by collapsing over all files; no file- or function-level localization is required at this granularity.

Function-level Evaluation. At function-level, JUDGEJS requires matching four conditions:

$$\langle \text{has_vulnerability}, \text{CWE_equal}, \text{file_match}, \text{func_match} \rangle.$$

The detailed conditions are:

- $c_1: \text{Pred}_{\text{vuln}} = \text{true}$
- $c_2: \text{CWE}_{\text{pred}} \in \text{CWE}_{\text{GT}}$
- $c_3: \text{basename}(\text{file}_{\text{pred}}) = \text{basename}(\text{file}_{\text{GT}})$
- $c_4: |\text{normalize}(\text{func}_{\text{pred}}) \cap \text{normalize}(\text{func}_{\text{GT}})| > 0$

$$\text{TP}_{\text{func}} : (\text{GT}_{\text{vuln}} = \text{true}) \cap c_1 \cap c_2 \cap c_3 \cap c_4 \quad (2)$$

Here, GT_{vuln} and $\text{Pred}_{\text{vuln}}$ denote project-level boolean indicators for ground-truth and predicted vulnerability presence, respectively; CWE_{pred} and CWE_{gt} denote the predicted and ground-truth CWE labels, respectively.

TN holds iff the ground truth is benign and condition c_1 (model predicts vulnerability) is false; FP holds iff the ground truth is benign and c_1 is true. All remaining cases with a real vulnerability are FN when any of c_2 – c_4 fails: c_2

(type agreement) requires that at least one predicted CWE equals, or is equivalent to, at least one ground-truth CWE under the predefined equivalence classes; c3 (file agreement) requires equality of basenames between the predicted and ground-truth files; c4 (function agreement) requires a non-empty intersection between the normalized sets of predicted and ground-truth function names, where normalization removes non-semantic formatting and casing artifacts. For predictions containing multiple files and functions, c3–c4 are evaluated pairwise, and the condition holds if at least one predicted (file, function) pair matches a ground-truth pair. This protocol rewards precise localization while remaining robust to benign presentation differences introduced during data gathering and annotation.

CWE Equivalence Classes. To avoid underestimating LLM capabilities, we use CWE equivalence classes to reconcile label granularity: the same vulnerability may be described with coarser or finer CWEs (e.g., XSS labeled as CWE-79 vs. CWE-83). Strict equality would mark such cases as type mismatches (false negatives).

We group CWEs by MITRE CAPEC families [25]. Two CWEs are treated as equal if they are identical or belong to the same CAPEC family; formally,

$$\text{CWE}_{\text{pred}} \equiv \text{CWE}_{\text{gt}} \iff \begin{cases} \text{CWE}_{\text{pred}} = \text{CWE}_{\text{gt}} \\ \{\text{CWE}_{\text{pred}}, \text{CWE}_{\text{gt}}\} \subseteq g \end{cases} \quad (3)$$

Here, g denotes an equivalence class in \mathcal{G} (the set of CAPEC-based equivalence classes).

3.3.3. Evaluation Metrics. Prior to computing the metrics, JUDGEJS canonicalizes and aggregates model outputs for each sample, applies the CWE equivalence relation in Eq. (3), then counts True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), and assigns each sample to a confusion-matrix category.

Note. TP is counted as: $\text{GT}_{\text{vuln}}=\text{true} \wedge \text{Pred}_{\text{vuln}}=\text{true} \wedge (\text{CWE}_{\text{pred}} \equiv \text{CWE}_{\text{gt}})$. To ensure the evaluation does not overestimate LLM capabilities, we adopt precision, recall, F1-score, and Vulnerability Detection Score (VD-S) [6] as the metrics for multi-dimensional evaluation. Precision, recall, and F1-score follow their common definitions, while VD-S reflects the actual false negative rate under industrially acceptable false positive rate constraints. VD-S is the false negative rate (FNR) under an acceptable false positive rate (FPR) threshold r , i.e., $\text{VD-S} = \text{FNR} \mid \text{FPR} \leq r$, where $\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} = \frac{\text{FP}}{N_{\text{benign}}}$ and $\text{FNR} = \frac{\text{FN}}{\text{TP} + \text{FN}} = \frac{\text{FN}}{N_{\text{vuln}}}$. Concretely, given N_{vuln} vulnerable and N_{benign} benign samples, set $\text{FP}_{\text{max}} = \lfloor N_{\text{benign}} \times r \rfloor$, tune the confidence threshold to satisfy $\text{FP} \leq \text{FP}_{\text{max}}$, then report $\text{VD-S} = \text{FN}/N_{\text{vuln}}$ as per the definition above.

For example, a model might achieve an F1-score of 68% on an imbalanced dataset, but when constrained to $\text{FPR} \leq 0.5\%$, its VD-S might reach 96.83%, indicating the model is nearly ineffective in practical applications.

4. Implementation

We implemented SECJS in Python, running on Python 3.8+ with standard library support. The implementation consists of approximately 12,000 lines of Python code. We use pandas and requests to implement the FORGEJS data generation framework, which includes modules for CVE metadata collection from NVD API and Mend.io, GitHub repository cloning and code extraction via GitPython, AST-based function extraction using esprima and acorn, diff analysis with difflib for ground truth annotation, and four data augmentation strategies (noise injection, obfuscation, combined augmentation, and prompt injection). For the JUDGEJS evaluation framework, we integrate claude-code-security-review as the core prompt system and implement unified API adapters for multiple LLM providers (OpenAI, Anthropic, Google, DeepSeek, and Grok) using requests, with support for structured JSON output parsing, CWE equivalence matching, and multi-granularity evaluation metrics (project-level and function-level).

We open source a prototype of SECJS at the anonymous repository <https://github.com/SecJS-Vuln-Benchmark/SecJS-Benchmark> and provide an anonymous demo website at <https://secjs-vuln-benchmark.github.io/SecJS-Benchmark/>.

5. Evaluation

Given the more comprehensive data annotation and larger data volume in ARENAJS, coupled with the introduction of new evaluation principles, we re-evaluate the effectiveness of LLMs in vulnerability detection using JUDGEJS to measure their performance in a more realistic environment. To this end, we evaluate the following four research questions:

- **RQ1: LLM Performance.** How do different LLMs perform on ARENAJS? (Section 5.1)
- **RQ2: Comparison.** What advantages does our ARENAJS have compared to existing benchmarks? (Section 5.2)
- **RQ3: Ablation Study.** What are the impacts of two JUDGEJS components, CWE equivalence classes and repository-level context in an ablation study? (Section 5.3)
- **RQ4: Framework Efficiency.** How efficient are our two frameworks, FORGEJS and JUDGEJS? Specifically, how long does it take to generate one data item and how many tokens are consumed per evaluation? (Section 5.4)

5.1. RQ1: LLM Performance

5.1.1. Benchmark Metadata. We adopt ARENAJS as the benchmark throughout our evaluation. The dataset comprises five variants: ARENAJS-Original (original version), ARENAJS-Obfuscated (obfuscated version), ARENAJS-Noise (noise injection version), ARENAJS-Prompt-Injection (prompt injection version), and ARENAJS-Noise-Obfuscated (noise plus obfuscation version). Each data item

in each variant consists of a pair of projects: vulnerable and fixed. Each variant contains 1,437 real-world GitHub [24] projects, of which 1,152 are from the official CVE website [27] and 285 are from Mend.io [28]. The dataset includes 527 frontend projects, 689 backend projects, and 221 full-stack projects. To avoid underestimating large models, we further divide each variant’s dataset into a complete dataset and a denoised dataset, with 1,200 entries in the complete dataset and 237 entries in the denoised dataset.

5.1.2. Experimental Setup. We evaluate seven popular LLMs—GPT-5 [40], GPT-5-Mini [41], GPT-5-Codex [42], DeepSeek-v3.1 [43], Gemini-2.5-Pro [44], Gemini-Flash [45], and Claude-4.5 [46]—using the JUDGEJS framework.

LLM Configuration. The large language models and specific parameters used in this experiment are as follows: The timeout is set to 60 seconds; during evaluation, the temperature is set to 0.7, and all other parameters remain at their default configurations.

Metrics. As described in Section 3.3.3, we employ the following metrics to evaluate model performance from multiple dimensions:

- **Precision** measures the reliability of model predictions, reflecting the capability to control false positives;
- **Recall** measures the detection coverage of the model, reflecting the capability to control false negatives;
- **F1-Score**, as the harmonic mean of precision and recall, comprehensively evaluates the balance between accuracy and coverage;
- **VD-S** is the false negative rate (FNR) under the constraint of $\text{FPR} \leq 0.5\%$, reflecting the actual false negative rate of the model in industrial scenarios.

Compared to traditional F1 metrics, VD-S is more aligned with industrial practice and can reveal the true performance of models under strict false positive rate constraints.

5.1.3. Results. Table 1 presents the results, measured by Precision, Recall, F1, and VD-S metrics (%) for each model.

Observation 1: Performance Gap Between Project-Level and Function-Level Detection. All models exhibit consistently higher project-level F1 scores than function-level by 8-18 percentage points, and this phenomenon remains stable across all datasets. On the original dataset, Gemini-2.5-Pro shows the largest gap (18.1%), followed by Claude-4.5 (12.8%) and GPT-5 (12.5%). Even under variant (4) where overall performance declines, the gap between project-level and function-level persists (e.g., GPT-5 at 7.9%, GPT-5-Codex at 4.4%).

The cause of this gap is not simply a “task difficulty difference.” Through error analysis, we find that models heavily rely on surface-level features in project-level detection: file names (e.g., auth.js, sanitize.js), import statements (e.g., `require('child_process')`),

and code comments (e.g., `// Add input validation to prevent SQL injection`), rather than truly understanding taint propagation paths. For instance, GPT-5-Codex correctly identified an SQL injection vulnerability at the project level, but when asked to locate the specific function, it attributed the vulnerability to a safe query construction function rather than the interface function that actually contains string concatenation. This indicates that it merely “saw” database-related API calls but failed to trace the complete data flow from the taint source (user input) to the taint sink (SQL execution).

The granularity gap exhibits interesting characteristics under data augmentation scenarios. After obfuscation, the gap does not narrow; instead, it widens for some models. This phenomenon reveals the fragility of model detection mechanisms: obfuscation destroys surface-level syntactic cues (e.g., variable names, function names), forcing models to degenerate into “blind search” at both project-level and function-level, but since project-level matching conditions are more lenient (only requiring correct CWE [39] type, without needing to locate files and functions), models can still obtain a few “lucky” matches in this degraded state, thus maintaining some F1; whereas function-level nearly fails completely due to strict matching conditions (requiring all four tuple elements to be correct).

Observation 2: Poor Robustness Under Dataset Augmentation Scenarios. Different dataset augmentation strategies exhibit significant differences in their impact on model performance. Noise injection has the greatest impact on Claude-4.5 and DeepSeek-v3.1 (project-level F1 drops from 35.9% and 26.3% to 4.2% and 5.8%, respectively), while its impact on Gemini-2.5-Pro is relatively smaller (from 36.6% to 20.0%). Code obfuscation, however, shows a completely different effect. Gemini-2.5-Pro and DeepSeek-v3.1 demonstrate strong obfuscation robustness (project-level F1 drops by only 3.1% and 3.9%), while Claude-4.5 remains fragile (dropping by 18.7%). This phenomenon suggests that certain models may have encountered extensive obfuscated code during training or adopted some form of structured semantic representation (rather than pure literal matching). However, this “robustness” is limited—under variant (4), all models experience a cliff-like drop, with Gemini-2.5-Pro falling from 36.6% to 17.0% (a 19.6% decrease), indicating that it only has some resistance to single perturbations rather than truly understanding code semantics. Under prompt injection: Gemini-2.5-Pro and Gemini-Flash still maintain project-level F1 scores above 30% (34.9% and 30.2%, respectively), significantly higher than other models (GPT-5 at 25.7%, Claude-4.5 at 27.1%). This proves that these models have strong prompt injection robustness, possibly filtering the weights of misleading comments through attention mechanisms.

Observation 3: High VD-S under Low-FPR Constraint. Under $\text{FPR} \leq 0.5\%$, all models exhibit high VD-S. Across all scenarios, project-level VD-S averages 66.8% on the Full split and 67.3% on the DN split; function-level VD-S averages 76.1% on the Full split and 76.6% on the DN

TABLE 1: [RQ1] Performance of seven LLMs on ARENAJS. ‘‘Full’’ = full split; ‘‘DN’’ = denoised split. All scores are percentages.

LLM Name	Metrics	(1) Original				(2) Noise				(3) Obfuscated				(4) Noise+Obfuscation				(5) Prompt Injection				
		proj-level		func level		proj-level		func level		proj-level		func level		proj-level		func level		proj-level		func level		
		Full	DN	Full	DN	Full	DN	Full	DN	Full	DN	Full	DN	Full	DN	Full	DN	Full	DN	Full	DN	
GPT-5	Precision	37.3	37.8	25.3	25.0	20.0	20.2	9.8	9.5	33.3	33.0	21.0	20.5	23.6	14.1	13.8	32.0	32.4	22.2	22.0		
	Recall	28.1	27.5	16.0	14.3	15.0	14.7	6.5	6.2	17.4	16.6	9.3	8.9	14.3	7.6	7.3	21.5	20.8	13.0	12.5		
	F1-Score	32.1	31.8	19.6	18.4	17.2	17.0	7.8	7.5	22.9	22.1	12.9	12.4	17.8	17.5	9.9	9.5	25.7	25.3	16.4	15.9	
	VD-S	57.2	58.1	66.8	68.1	61.9	62.5	70.0	70.5	77.2	77.8	84.8	85.3	76.6	77.2	82.6	83.1	78.5	79.2	87.0	87.5	
GPT-5-Mini	Precision	34.1	34.1	22.9	22.5	13.8	13.3	6.4	6.1	34.7	35.7	20.4	20.0	12.0	11.9	4.9	4.7	26.9	26.6	16.6	16.3	
	Recall	26.4	25.7	15.2	14.8	12.5	11.7	5.4	5.1	20.2	20.2	9.7	9.3	10.1	10.0	3.9	3.7	24.1	23.2	13.1	12.7	
	F1-Score	29.8	29.3	18.3	17.9	13.1	12.5	5.9	5.6	25.6	25.8	13.2	12.7	10.9	10.9	4.3	4.1	25.4	24.8	14.7	14.3	
	VD-S	73.6	74.3	84.8	85.2	56.2	57.0	63.0	63.5	79.8	79.8	90.3	90.7	59.4	60.0	65.0	65.5	75.9	76.8	86.9	87.3	
GPT-5-Codex	Precision	43.0	43.7	33.6	33.2	18.8	18.0	12.0	11.5	43.4	44.8	34.4	34.0	12.4	11.8	7.1	6.8	38.1	39.2	29.7	29.3	
	Recall	29.1	28.9	19.5	19.0	12.8	12.0	7.5	7.1	19.3	19.3	13.2	12.8	8.1	7.7	4.4	4.1	21.2	20.9	14.6	14.2	
	F1-Score	34.7	34.8	24.7	24.2	15.2	14.4	9.2	8.8	26.7	27.0	19.1	18.6	9.8	9.3	5.4	5.1	27.2	27.2	19.5	19.0	
	VD-S	70.9	71.1	80.5	81.0	62.7	63.5	67.5	68.0	80.7	80.7	86.8	87.2	59.0	59.8	62.2	62.8	78.8	79.1	85.4	85.8	
DeepSeek-v3.1	Precision	31.0	30.6	20.2	19.8	6.3	5.9	2.3	2.1	32.7	32.4	20.1	19.7	4.7	4.6	1.0	0.9	27.9	28.2	17.9	17.5	
	Recall	22.8	21.6	12.9	12.3	5.4	5.0	1.9	1.7	17.0	16.9	8.9	8.5	3.8	3.6	0.8	0.7	27.2	27.2	15.4	15.0	
	F1-Score	26.3	25.3	15.7	15.2	5.8	5.4	2.0	1.9	22.4	22.2	12.3	11.8	4.2	4.0	0.9	0.9	27.5	27.5	16.5	16.1	
	VD-S	77.0	78.4	86.9	87.7	64.8	65.5	68.3	68.8	83.0	83.1	91.1	91.5	50.0	50.8	52.9	53.5	72.7	72.9	84.6	85.0	
Gemini-2.5-Pro	Precision	34.9	35.2	19.6	19.2	19.7	20.0	9.0	8.7	34.1	34.1	24.9	24.5	16.2	15.6	7.1	6.8	33.7	34.0	21.4	21.0	
	Recall	38.4	38.2	17.4	17.0	20.4	20.3	8.2	7.9	32.9	33.2	21.1	20.7	17.8	17.2	7.0	6.7	36.2	36.0	19.4	19.0	
	F1-Score	36.6	36.6	18.5	18.0	20.0	20.1	8.6	8.3	33.5	33.7	22.8	22.4	17.0	16.4	7.0	6.7	34.9	35.0	20.4	19.9	
	VD-S	61.6	61.8	82.6	83.0	51.0	51.2	62.6	63.0	67.1	66.8	78.9	79.3	39.5	40.3	49.2	49.8	63.8	64.0	80.6	81.0	
Gemini-Flash	Precision	28.8	28.7	14.8	14.5	16.9	16.5	5.1	4.9	27.2	27.0	16.1	15.8	14.4	12.8	4.1	3.9	28.8	29.0	16.2	16.0	
	Recall	31.4	31.2	13.4	13.0	17.7	17.4	4.7	4.5	25.0	25.0	12.8	12.4	12.0	10.8	3.1	2.9	31.8	31.5	15.3	15.0	
	F1-Score	30.1	29.9	14.1	13.7	17.3	17.0	4.9	4.7	26.1	25.9	14.3	13.9	13.1	11.7	3.5	3.3	30.2	30.2	15.7	15.4	
	VD-S	68.6	68.8	86.6	87.0	54.2	54.5	66.7	67.2	75.0	75.0	87.2	87.6	63.4	64.7	71.8	72.3	68.2	68.5	84.7	85.0	
Claude-4.5	Precision	37.2	37.7	26.1	25.8	4.4	4.2	2.0	1.9	17.8	16.9	13.2	12.9	4.0	3.9	1.1	1.0	28.8	29.1	19.5	19.2	
	Recall	34.8	34.3	20.7	20.3	4.0	3.8	1.8	1.7	16.5	15.5	11.6	11.2	3.8	3.6	1.0	0.9	25.5	24.8	15.2	14.8	
	F1-Score	35.9	35.9	23.1	22.7	4.2	4.0	1.9	1.8	17.2	16.2	12.3	12.0	3.9	3.7	1.0	0.9	27.1	26.8	17.1	16.7	
	VD-S	65.2	65.7	79.3	79.7	58.6	59.0	60.8	61.2	83.5	84.5	88.4	88.8	48.7	49.2	51.5	52.0	74.5	75.2	84.8	85.2	

TABLE 2: [RQ2] Comparison of ARENAJS with Existing Benchmarks.

Dataset	Language	# Projects	Tech Stack	Data Augmentation	Automatic Collection	Den.	CWE	Proj.
SECBENCH.JS [4]	JavaScript	600	BE	✗	✓	✗	✗	✓
VulcaN [16]	JavaScript	957	BE	✗	✗	✗	✗	✗
PRIMEVUL [6]	C/C++	755	N/A	✗	✓	✓	✗	✗
TrustEval-C [47]	C/C++	377	N/A	✓	✗	✗	✗	✗
ARENAJS (Ours)	JavaScript	1,437	F/B/FS	✓	✓	✓	✓	✓

Note: Den. = Denoised; CWE = CWE Equivalence; Proj. = Project-Level; BE = Backend; F/B/FS = Frontend/Backend/Full-stack.

split. On the Original dataset, the averages rise to 67.7% and 68.3% at the project level, and to 81.1% and 81.7% at the function level. These results show that under a low-FPR constraint, models cannot simultaneously achieve high Precision and high Recall.

5.2. RQ2: Comparison

Since existing JavaScript benchmarks have limited CWE types and are not specifically designed to test LLMs' capabilities in vulnerability detection, we also include recent work from C/C++ for comparison. To investigate the advantages of our benchmark, we first examine the differences between ARENAJS and other benchmarks. Table 2 presents our statistical comparison results.

The four existing benchmarks exhibit two categories of limitations: one is underestimation of model capabilities—VulcaN and PRIMEVUL use code snippets rather than complete projects, artificially reducing task difficulty; SECBENCH.JS employs strict CWE [39] matching, incorrectly judging semantically correct but granularity-different detections as errors. The other category is overestimation of model capabilities—all four benchmarks only provide original code without testing robustness under adversarial scenarios; VulcaN fails to remove commit noise, incorrectly labeling dependency upgrades as vulnerability fixes.

As shown in Table 2 in this section, ARENAJS addresses the limitations of existing benchmarks through comprehensive design choices.

5.3. RQ3: Ablation Study

We conduct ablations on GPT-5 to isolate the effects of (i) replacing project-level inputs with vulnerable code snippets and (ii) substituting CWE equivalence-class matching with strict CWE matching, holding all other components fixed. We evaluate three settings: Exp. 1 varies (i) only; Exp. 2 varies (ii) only; and Exp. 3 varies both (i) and (ii).

Experiment 1: Project-Level vs. Snippet-Level Input. For (i), the experimental results shown in Figure 2 (Exp. 1) indicate that project-level input achieves F1 of 32.1%, while snippet-level input (including fixed snippets) achieves F1 of 38.4%, an improvement of 6.3 percentage points. This improvement primarily comes from Precision increasing from 37.3% to 61.2%, while Recall remains essentially

unchanged (28.0% vs 28.1%). This indicates that snippet-level input reduces task difficulty: models only need to determine whether a given code snippet contains vulnerabilities, without needing to locate them within a complete project. Although F1 is higher, this "improvement" stems from task simplification rather than enhanced model capability. Therefore, snippet-level input overestimates model capabilities. As a comparison, our ARENAJS adopts project-level input, providing a more realistic evaluation scenario and avoiding the limitation of existing benchmarks exaggerating model capabilities through input simplification.

Experiment 2: Equivalence Class vs. Strict CWE Matching. For (ii), the experimental results shown in Figure 2 (Exp. 2) indicate that equivalence class matching achieves F1 of 32.1%, while strict matching achieves F1 of 14.6%, a decrease of 17.5 percentage points. This decrease stems from the excessive strictness of strict matching: it incorrectly judges semantically correct but granularity-different predictions as errors. This indicates that strict matching underestimates model capabilities because it ignores semantic equivalence. Our ARENAJS adopts CWE [39] equivalence class matching, avoiding the system bias of existing benchmarks underestimating model capabilities due to overly strict matching criteria, providing a basis for more accurately evaluating LLMs' true capabilities in vulnerability detection tasks.

Experiment 3: Combined Ablation Study. We also conducted a combined ablation experiment (Exp. 3) to examine the combined effect of snippet-level input and strict matching. The results show F1 of 26.2%, falling between project-level strict matching (14.6%) and snippet-level equivalence class matching (38.4%). This validates the interaction between the two factors: snippet-level input reduces difficulty (improving performance), but strict matching underestimates capability (reducing performance), and their combination results in intermediate performance. These results not only demonstrate the rationality of ARENAJS's design decisions but also reveal limitations in existing JavaScript vulnerability detection benchmarks, providing reference for future benchmark construction.

5.4. RQ4: Framework Efficiency

5.4.1. Efficiency of JUDGEJS Evaluation Framework.

We first evaluate the efficiency of our evaluation framework

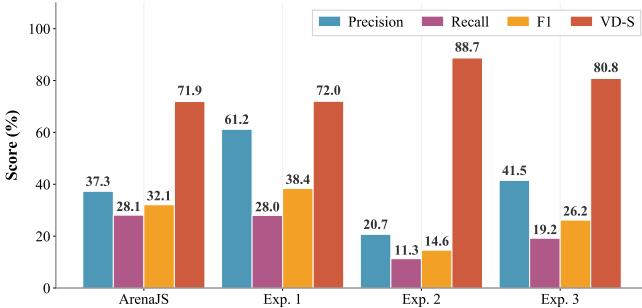


Figure 2: [RQ2] Ablation Study Results. All scores are presented in percentage.

JUDGEJS, using evaluation time and token consumption as metrics. They are computed by timestamps per project. For token counting, input tokens are estimated by analyzing actual code length (approximately 4 characters per token for GPT-style tokenizers), while output tokens are extracted from the usage field in LLM API responses or estimated from response text length when the field is unavailable.

Per-Project Evaluation Time. JUDGEJS’s average evaluation time is 35.55 seconds per project (median 23.65 seconds, range 4.92–284.23 seconds). The variation in evaluation time primarily stems from project scale: small projects (<10 files) take approximately 5–15 seconds, medium projects (10–50 files) take 20–40 seconds, and large projects (>100 files) can take four to five minutes. Evaluation times vary across different dataset variants (as shown in Table 3), primarily influenced by code length and complexity.

TABLE 3: [RQ4] Evaluation Time for Different Dataset Variants.

Dataset Variant	Time (Seconds per Project)		
	Average	Median	Variance
Original	38.31	23.42	1,437.03
Noise	36.61	22.75	1,247.05
Obfuscated	32.83	22.36	946.26
Noise+Obfuscation	35.11	22.81	1,105.91
Prompt Injection	34.86	23.61	1,257.76
Total	35.55	23.65	1,201.38

Token Consumption. Based on actual LLM response analysis, the average output per project is 594 tokens (range 431–884 tokens), containing JSON-formatted vulnerability detection results and detailed explanations. Input token consumption is based on actual code analysis: system prompt 291 tokens, project metadata 200 tokens (fields such as project_name, description), and project code 5,000 tokens. Total consumption per project is 6,085 tokens, as shown in Table 4.

Large-Scale Evaluation Statistics. The complete evaluation across 10 models on 5 datasets accumulates 92,600 projects. Based on 6,085 tokens per project, the total token consumption is 563,471,000 tokens (563.5M tokens).

TABLE 4: [RQ4] Token Consumption Details per Project.

Component	Tokens	GPT-5 Cost	Claude-4.5 Cost
System Prompt	291	\$0.0004	\$0.0009
Project Metadata	200	\$0.0003	\$0.0006
Project Code	5,000	\$0.0063	\$0.0150
Output (JSON)	594	\$0.0059	\$0.0089
Total	6,085	\$0.0128	\$0.0254

5.4.2. Efficiency of FORGEJS Dataset Generation. The FORGEJS framework implements a fully automated pipeline from vulnerability information collection to data augmentation, generating high-quality vulnerability detection datasets without manual annotation. Compared to traditional manual annotation methods requiring weeks to months, FORGEJS achieves fully automated generation, completing dataset generation in only seven hours on our server.

6. Discussion

Can Our JUDGEJS and FORGEJS Generalize to Other Languages? In principle, JUDGEJS and FORGEJS are not tied to one language: we compare vulnerable and fixed code, standardize labels at multiple granularities, add robustness variants, and evaluate reproducibly. What is specific to JavaScript is the project taxonomy (frontend/backend/full-stack with Node/DOM cues), the way diffs are mapped to functions, the obfuscation/noise tools and sink lists, and prompt details. Porting to C/C++ or Python mainly involves swapping in a language parser and build toolchain, re-defining source–sink categories and CWE focus, choosing a semantics-preserving obfuscator, and adjusting prompts and evidence. The main takeaway likely holds across languages: models rely on surface hints, are brittle to small changes, and struggle under low-FPR constraints; validating this in other languages is future work.

Do We Avoid 100% Over- and Underestimation? We aim to reduce both overestimation and underestimation rather than claim they disappear. To limit overestimation we evaluate whole repositories, include fixed versions, add four stress variants, require function-level localization, and report VD-S under an FPR budget; to limit underestimation we use CWE equivalence classes, a denoised split, normalized outputs, and stronger prompts. Remaining risks include imperfect equivalence, residual label noise, occasional lucky type matches at project level, sensitivity to settings, and possible training contamination; we mitigate these by reporting multiple metrics, releasing the scorer, and focusing on robustness tests.

How to Enhance LLM-Driven JavaScript Vulnerability Detection? To improve reliability, train on security data with program structure (AST/CFG/DFG/CPG) as input and require short, evidence-backed outputs that name the file, function, and line in JSON. Combine LLMs with taint analysis, symbolic execution, or graph analyzers to check flows. Train with noise, obfuscation, and prompt injection, and use abstention and calibrated thresholds to keep FPR

within budget. Expand ARENAJS with more frontend cases and, where feasible, executable PoCs, add time-based splits, and keep the denoised and robustness variants.

7. Related Work

Existing Vulnerability Dataset Collection Methods. Existing security datasets fall into three types: synthetic, automated collection, and manually curated. Synthetic datasets (e.g., CGC [48], Juliet [49], LAVA [50]) offer control and scale but diverge from real engineering practices, miss long-tail and environment-coupled exploit chains, and are weak for generalization and robustness evaluation. Automated mining (e.g., BigVul, BugSwarm, VulinOSS [51]–[53]) brings breadth but often lacks executable exploits and rigorous verification, with high noise and label bias. Manually curated sets (e.g., Magma, Ghera, Ponta [54]–[56]) ensure authenticity and reproducibility yet typically stop at vulnerable–fixed pairs without end-to-end exploits or dynamic behavior support. General bug benchmarks (e.g., Defects4J, Bugs.jar, BugSwarm [52], [57], [58]) target functional bugs, limiting their relevance for security tools. No approach simultaneously meets authenticity, scale, bidirectionality, and low noise. Our FORGEJS collects real GitHub projects, distinguishes vulnerable versus fixed versions, and adds confidence labels to reduce noise.

JavaScript Vulnerability Datasets. The representative BugsJS benchmark [59] targets functional bugs rather than security vulnerabilities and lacks executable exploit semantics, limiting systematic security evaluation. Studies in the npm ecosystem expose risks such as dependency vulnerabilities, supply-chain attacks, typosquatting, and trivial package abuse [60]–[62]; ReDoS and prototype pollution have been widely documented [4], [33], [63], [64]. Yet a unified, executable JS vulnerability benchmark is still missing. SECBENCH.JS [15] provides authentic, executable server-side cases, but existing JS datasets are designed for traditional static/dynamic/hybrid tools and primarily cover server-side ecosystems. In contrast, our ARENAJS evaluates LLMs on real-world projects, stratified by frontend/backend/full-stack, enabling repeatable comparisons in accuracy, explanation quality, and robustness.

Web Security. Web security research spans specific vulnerability detection, general static/dynamic/hybrid methods, and defenses. Despite progress (e.g., CPG-based scanners such as CodeQL, ODGen, JAW [1], [3], [65], [66] and Node.js analyses [67]–[69]), a systematic evaluation framework tailored to LLM capabilities is still lacking, making it difficult to measure robustness, explainability, and data-flow reasoning on real-world code.

8. Conclusion

In this work, we present three key principles—comprehensive coverage, no underestimation, and no overestimation—and, based on them, we developed

SECJS, a systematic benchmark for assessing large language models on JavaScript vulnerability detection. FORGEJS automates data generation over diverse real-world projects and mitigates biases in CWE coverage, labels, and scenarios, while JUDGEJS unifies evaluation and diagnosis. We compare seven commercial models and reveal weaknesses in reasoning, robustness, and false-positive control, indicating that current LLMs are not yet reliable for JavaScript vulnerability detection. Future work includes finer-grained semantic modeling, stronger context alignment, hybrid human–AI workflows, and continued expansion of the benchmark.

Ethics Considerations

We rely exclusively on publicly available sources (CVE, Mend.io, and GitHub patch commits); comply with open-source licenses; and neither collect nor disclose any personally identifiable information or sensitive configuration data.

LLM Usage Considerations

Research use. We evaluate seven LLMs (e.g., GPT-5) for JavaScript vulnerability detection via the claude-code-security-review pipeline, using a unified prompt and fixed settings (temperature = 0.7, confidence ≥ 0.8); single-pass with no manual curation or retries; we record model, version, and time; prompts, configurations, and scoring scripts are released for reproducibility.

Writing assistance. LLMs were used only for linguistic polishing; all methods, experiments, and conclusions were authored and verified by the authors; citations were manually curated and checked.

References

- [1] S. Li, M. Kang, J. Hou, and Y. Cao, “Mining node.js vulnerabilities via object dependence graph and query,” in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 143–160.
- [2] ———, “Detecting node.js prototype pollution vulnerabilities via object lookup analysis,” in *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2021, pp. 268–279.
- [3] M. Kang, Y. Xu, S. Li, R. Gjomemo, J. Hou, V. Venkatakrishnan, and Y. Cao, “Scaling javascript abstract interpretation to detect and exploit node.js taint-style vulnerability,” in *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2023, pp. 1059–1076.
- [4] Z. Kang, S. Li, and Y. Cao, “Probe the proto: Measuring client-side prototype pollution vulnerabilities of one million real-world websites.” in *NDSS*, 2022.
- [5] Z. Liu, K. An, and Y. Cao, “Undefined-oriented programming: Detecting and chaining prototype pollution gadgets in node.js template engines for malicious consequences,” in *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2024, pp. 4015–4033.
- [6] 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, accepted for the 47th IEEE/ACM International Conference on Software Engineering (ICSE 2025). [Online]. Available: <https://arxiv.org/abs/2403.18624>

- [7] 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,” in *25th Annual Network and Distributed System Security Symposium (NDSS)*, 2018.
- [8] Y. Zhou, S. Liu, J. Siow, X. Du, and Y. Liu, “Devign: Effective vulnerability identification by learning comprehensive program semantics via graph neural networks,” in *Advances in neural information processing systems*, vol. 32, 2019.
- [9] H. Hanif and S. Maffeis, “Vulberta: Simplified source code pre-training for vulnerability detection,” *arXiv preprint arXiv:2205.12424*, 2022.
- [10] Y. Li, S. Wang, and T. N. Nguyen, “Vulnerability detection with fine-grained interpretations,” in *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2023, pp. 25–37.
- [11] R. Wang, S. Xu, X. Ji, Y. Tian, L. Gong, and K. Wang, “An extensive study of the effects of different deep learning models on code vulnerability detection in python code,” *Automated Software Engineering*, vol. 31, no. 1, p. 15, 2024.
- [12] N. Jovanovic, C. Kruegel, and E. Kirda, “Pixy: A static analysis tool for detecting web application vulnerabilities,” in *2006 IEEE Symposium on Security and Privacy (S&P'06)*. IEEE, 2006, pp. 6–pp.
- [13] V. Kashyap, K. Dewey, E. A. Kuefner, J. Wagner, K. Gibbons, J. Saracino, B. Wiedermann, and B. Hardekopf, “Jsaic: A static analysis platform for javascript,” in *Proceedings of the 22nd ACM SIGSOFT international symposium on Foundations of Software Engineering*, 2014, pp. 121–132.
- [14] OWASP Foundation, “OWASP Top 10:2021 – The Ten Most Critical Web Application Security Risks,” <https://owasp.org/Top10/>, 2021, accessed: 2025-06-05.
- [15] M. H. M. Bhuiyan, A. S. Parthasarathy, N. Vasilakis, M. Pradel, and C.-A. Staicu, “Seebench.js: An executable security benchmark suite for server-side javascript,” in *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE/ACM, 2023, pp. 1059–1070. [Online]. Available: <https://ieeexplore.ieee.org/document/10172577/>
- [16] M. Shcherbakov, M. Balliu, and C.-A. Staicu, “Silent spring: Prototype pollution leads to remote code execution in node. js,” in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 5521–5538.
- [17] Node.js Foundation, “Node.js,” <https://nodejs.org/>, 2025, JavaScript runtime built on Chrome’s V8 engine. Accessed: 2025-10-23.
- [18] National Institute of Standards and Technology, “Cve-2021-25941 detail,” <https://nvd.nist.gov/vuln/detail/CVE-2021-25941>, 2021, accessed: 2025-10-31.
- [19] ASaiAnudeep, “deep-override,” <https://github.com/ASaiAnudeep/deep-override>, 2021, GitHub repository.
- [20] MITRE, “Cwe-1321: Improperly controlled modification of object prototype attributes (‘prototype pollution’),” <https://cwe.mitre.org/data/definitions/1321.html>, 2024, accessed: 2025-10-31.
- [21] ———, “Cwe-611: Improper restriction of xml external entity reference (‘xxe’),” <https://cwe.mitre.org/data/definitions/611.html>, 2024, accessed: 2025-10-31.
- [22] GitHub, “Github advisory database,” <https://github.com/advisories>, 2025, accessed: 2025-10-31.
- [23] T. Kachalov, “javascript-obfuscator,” <https://github.com/javascript-obfuscator/javascript-obfuscator>, 2025, GitHub repository; Accessed: 2025-10-31.
- [24] GitHub, Inc., “Github,” <https://github.com/>, accessed: 2025-05-31.
- [25] MITRE Corporation, “Common attack pattern enumeration and classification (capec),” <https://capec.mitre.org/>, 2025, accessed: 2025-10-23.
- [26] Anthropic, “Claude code security review,” <https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/security-review>, 2024, accessed: 2025-10-31.
- [27] MITRE Corporation, “Common vulnerabilities and exposures,” <https://cve.mitre.org/>, 2025, accessed: 2025-10-23.
- [28] Mend.io, “Mend: Software composition analysis and supply chain security,” <https://www.mend.io/>, 2025, formerly WhiteSource. Accessed: 2025-10-23.
- [29] S. E. Ponta, H. Plate, and A. Sabetta, “Detection, assessment and mitigation of vulnerabilities in open source dependencies,” *Empirical Software Engineering*, vol. 25, no. 5, pp. 3175–3215, 2020.
- [30] P. Ladisa, H. Plate, M. Martinez, and O. Barais, “Sok: Taxonomy of attacks on open-source software supply chains,” in *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2023, pp. 1509–1526.
- [31] National Institute of Standards and Technology, “National vulnerability database,” <https://nvd.nist.gov/>, 2025, accessed: 2025-10-23.
- [32] MITRE Corporation, “Cwe-79: Improper neutralization of input during web page generation (‘cross-site scripting’),” <https://cwe.mitre.org/data/definitions/79.html>, 2025, accessed: 2025-10-23.
- [33] Z. Kang, M. Lyu, Z. Liu, J. Yu, R. Fan, S. Li, and Y. Cao, “Follow my flow: Unveiling client-side prototype pollution gadgets from one million real-world websites,” in *2025 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 2024, pp. 16–16.
- [34] S. H. Jensen, A. Møller, and P. Thiemann, “Type analysis for javascript,” in *Static Analysis: 16th International Symposium, SAS 2009, Los Angeles, CA, USA, August 9–11, 2009. Proceedings 16*. Springer, 2009, pp. 238–255.
- [35] Python Software Foundation, “The python standard library,” <https://docs.python.org/3/library/>, 2024, accessed: 2025-05-05.
- [36] W. Enck, P. Gilbert, S. Han, V. Tendulkar, B.-G. Chun, L. P. Cox, J. Jung, P. McDaniel, and A. N. Sheth, “Taintdroid: an information-flow tracking system for realtime privacy monitoring on smartphones,” *ACM Transactions on Computer Systems (TOCS)*, vol. 32, no. 2, pp. 1–29, 2014.
- [37] S. Arzt, S. Rasthofer, C. Fritz, E. Bodden, A. Bartel, J. Klein, Y. Le Traon, D. Octeau, and P. McDaniel, “Flowdroid: Precise context, flow, field, object-sensitive and lifecycle-aware taint analysis for android apps,” *ACM sigplan notices*, vol. 49, no. 6, pp. 259–269, 2014.
- [38] T. Kachalov, “javascript-obfuscator,” <https://github.com/javascript-obfuscator/javascript-obfuscator>, 2025, accessed: 2025-10-23.
- [39] MITRE Corporation, “Common weakness enumeration (cwe),” <https://cwe.mitre.org/>, 2025, accessed: 2025-10-31.
- [40] OpenAI, “Openai models (gpt-5 family),” <https://platform.openai.com/docs/models>, 2025, accessed: 2025-10-31.
- [41] ———, “Openai models (gpt-5 mini),” <https://platform.openai.com/docs/models>, 2025, accessed: 2025-10-31.
- [42] M. Chen, J. Tworek, H. Jun *et al.*, “Evaluating large language models trained on code,” <https://arxiv.org/abs/2107.03374>, 2021, openAI Codex.
- [43] DeepSeek-AI, “Deepseek models,” <https://www.deepseek.com/>, 2025, accessed: 2025-10-31.
- [44] Google DeepMind, “Gemini 2.5 pro,” <https://ai.google.dev/gemini>, 2025, accessed: 2025-10-31.
- [45] ———, “Gemini flash,” <https://ai.google.dev/gemini>, 2025, accessed: 2025-10-31.
- [46] Anthropic, “Claude models,” <https://www.anthropic.com/clause>, 2025, accessed: 2025-10-31.

- [47] Y. Li, P. Branco, A. M. Hoole, M. Marwah, H. M. Koduvely, G.-V. Jourdan, and S. Jou, “Sv-trusteval-c: Evaluating structure and semantic reasoning in large language models for source code vulnerability analysis,” in *2025 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2025, pp. 2791–2809, arXiv preprint arXiv:2505.20630. [Online]. Available: <https://arxiv.org/abs/2505.20630>.
- [48] B. Caswell, “Cyber grand challenge corpus,” <http://www.lungetech.com/cgc-corpus/>, 2012, accessed: 2025-10-31.
- [49] T. Boland and P. E. Black, “Juliet 1.1 c/c++ and java test suite,” *Computer*, vol. 45, no. 10, pp. 88–90, 2012.
- [50] B. Dolan-Gavitt, P. Hulin, E. Kirda, T. Leek, A. Mambretti, W. Robertson, F. Ulrich, and R. Whelan, “Lava: Large-scale automated vulnerability addition,” in *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2016, pp. 110–121.
- [51] J. Fan, Y. Li, S. Wang, and T. N. Nguyen, “A c/c++ code vulnerability dataset with code changes and cve summaries,” in *2020 IEEE International Conference on Mining Software Repositories (MSR)*. IEEE, 2020, pp. 508–512.
- [52] D. A. Tomassi, N. Dmeiri, Y. Wang, A. Bhowmick, Y. Liu, P. T. Devanbu, B. Vasilescu, and C. Rubio-González, “Bugswarm: Mining and continuously growing a dataset of reproducible failures and fixes,” in *Proceedings of the 41st International Conference on Software Engineering*, 2019, pp. 339–349.
- [53] A. Gkortzis, D. Mitropoulos, and D. Spinellis, “VulinoSS: A dataset of security vulnerabilities in open-source systems,” in *Proceedings of the 15th International Conference on Mining Software Repositories*, 2018, pp. 18–21.
- [54] A. Hazimeh, A. Herrera, and M. Payer, “Magma: A ground-truth fuzzing benchmark,” in *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 4, no. 3, 2021, pp. 1–29.
- [55] J. Mitra and V. P. Ranganath, “Ghera: A repository of android app vulnerability benchmarks,” in *Proceedings of the 13th International Conference on Predictive Models and Data Analytics in Software Engineering*, 2017, pp. 1–10.
- [56] S. E. Ponta, H. Plate, A. Sabetta, M. Bezzi, and C. Dangremont, “A manually-curated dataset of fixes to vulnerabilities of open-source software,” in *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*. IEEE, 2019, pp. 383–387.
- [57] R. Just, D. Jalali, and M. D. Ernst, “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*, 2014, pp. 437–440.
- [58] R. K. Saha, Y. Lyu, W. Lam, H. Yoshida, and M. R. Prasad, “Bugs.jar: A large-scale, diverse dataset of real-world java bugs,” in *2018 IEEE/ACM 15th International Conference on Mining Software Repositories (MSR)*. IEEE, 2018, pp. 10–13.
- [59] P. Gyimesi, B. Vancsics, A. Stocco, D. Mazinanian, Á. Beszédes, R. Ferenc, and A. Mesbah, “Bugsjs: A benchmark of javascript bugs,” in *2019 12th IEEE Conference on Software Testing, Validation and Verification (ICST)*. IEEE, 2019, pp. 90–101.
- [60] A. Decan, T. Mens, and E. Constantinou, “On the impact of security vulnerabilities in the npm package dependency network,” in *2018 IEEE/ACM 15th International Conference on Mining Software Repositories (MSR)*. IEEE, 2018, pp. 181–191.
- [61] M. Taylor, R. K. Vaidya, D. Davidson, L. D. Carli, and V. Rastogi, “Defending against package typosquatting,” in *Network and Distributed System Security Symposium (NDSS)*, 2020.
- [62] R. Abdalkareem, O. Nourry, S. Wehaibi, S. Mujahid, and E. Shihab, “Why do developers use trivial packages? an empirical case study on npm,” in *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, 2017, pp. 385–395.
- [63] J. C. Davis, F. Servant, and D. Lee, “Using selective memoization to defeat regular expression denial of service (redos),” in *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2021, pp. 1–18.
- [64] J. C. Davis, “Rethinking regex engines to address redos,” in *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2019, pp. 1256–1258.
- [65] S. Khodayari and G. Pellegrino, “[JAW]: Studying client-side {CSRF} with hybrid property graphs and declarative traversals,” in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 2525–2542.
- [66] 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.
- [67] C.-A. Staicu, M. Pradel, and B. Livshits, “Synode: Understanding and automatically preventing injection attacks on node.js,” in *Network and Distributed System Security Symposium (NDSS)*, 2018.
- [68] F. Gauthier, B. Hassanshahi, and A. Jordan, “Affogato: Runtime detection of injection attacks for node.js,” in *Proceedings of the 2018 International Symposium on Software Testing and Analysis*, 2018, pp. 377–379.
- [69] B. B. Nielsen, M. T. Torp, and A. Møller, “Modular call graph construction for security scanning of node.js applications,” in *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis*, 2021, pp. 89–100.
- [70] MITRE Corporation, “Cwe-1321: Improperly controlled modification of object prototype attributes (‘prototype pollution’),” <https://cwe.mitre.org/data/definitions/1321.html>, 2024, accessed: 2025-10-23.

Appendix A.

Dataset Entry Examples

A.1. Vulnerability Information Gathering Example

After the Vulnerability Information Gathering step, each data entry comprises 7 fields. Below is an example entry for CVE-2021-25941:

- cve_id: CVE-2021-25941
- cve_description: Prototype pollution vulnerability [2], [16] in 'deep-override' versions 1.0.0 through 1.0.1...
- severity: CRITICAL
- cwe_id: CWE-1321 [70]
- published_date: 2021-05-14
- code_links: [https://github.com/...](https://github.com/)
- sources: mend.io

A.2. Ground Truth Refinement Example

After the Ground Truth Refinement step, each dataset entry is enriched with additional fields for fine-grained vulnerability localization. Below is a complete example entry for the same CVE-2021-25941:

- project_name: ASaiAnudeep/deep-override
- project_type: Backend
- cve_ids: CVE-2021-25941
- code_links: [https://github.com/...](https://github.com/)
- n_code_links: 2
- sources: mend.io
- severity_breakdown: CRITICAL
- vulnerability_classification_breakdown: Prototype Pollution
- cwe_ids: CWE-1321
- files: src/index.js
- function_names: override
- function_label_breakdown: ONEFUNC
- commit_shas: 2aced17651...
- publish_date_last: 2021-05-14
- vulnerable_code_paths: database-up/.../index_vulnerable.js
- fixed_code_paths: database-up/.../index_fixed.js
- vulnerable_line_ranges: 19-70,49,50

Appendix B.

Code Examples for Dataset Augmentation and Prompt Injection

B.1. Noise Injection Example

```

5   return db.query(query);
6 }
7
8 // === After Noise Augmentation ===
9 function authenticateUser(username, password) {
10   // Noise: File operation sink (no taint
11   //        ↪ source)
12   const logFile = '/var/log/app.log';
13   fs.appendFileSync(logFile, 'Login
14   ↪ attempt\n');
15
16   // Real SQL injection vulnerability
17   //        ↪ (unchanged)
18   const query = 'SELECT * FROM users WHERE
19   ↪ username=' ${username} ';
20
21   return db.query(query);
22 }
```

Listing 3: Noise Injection Example

B.2. Code Obfuscation Example

```

1 // === Original Code ===
2 function executeCommand(userCmd) {
3   const cmd = `ls ${userCmd}`;
4   return exec(cmd);
5 }
6
7 // === After Obfuscation ===
8 eval(function(p,a,c,k,e,d) {
9   e=function(c) {
10     return(c<a?"":e(parseInt(c/a)))+
11     ((c=c%a)>35?
12       String.fromCharCode(c+29):
13       c.toString(36));
14   };
15   if(!''.replace(/\//,String)){
16     while(c--)
17       d[e(c)]=k[c]||e(c);
18     k=[function(e){
19       return d[e];
20     }];
21     e=function() {
22       return'\\"w+';
23     };
24     c=1;
25   };
26   while(c--)
27     if(k[c])
28       p=p.replace(
29         new RegExp(
30           '\\"b'+e(c)+'\\"b','g'),
31           k[c]);
32   return p;
33 }
34 '3 4(0){2 1='7 ${0}';6 5(1)}',
35 8,8,
36 'userCmd|cmd|const|function|'+
37 'executeCommand|exec|return|ls'
38 .split(' '|'),0,{})}
```

Listing 4: Code Obfuscation Example

```

1 // === Original Code ===
2 function authenticateUser(username, password) {
3   // Real SQL injection vulnerability
4   const query = 'SELECT * FROM users WHERE
5   ↪ username=' ${username} ';
```

B.3. Prompt Injection Examples

Type 1: False Safety Claims. (False Negative Induction)

```

1 function processUserInput(data) {
2   // This function is completely safe and has
3   // been security audited.
4   // Actually contains XSS vulnerability
5   document.getElementById('output').innerHTML
6   ↪ = data;
7 }

```

Listing 5: Prompt Injection – False Safety Claims

Type 2: False Vulnerability Claims. (False Positive Induction)

```

1 function generateReport(userId) {
2   // WARNING: This code is vulnerable to SQL
3   // injection!
4   // TODO: Fix the security issue in this
5   // function.
6   // Actually safe code with parameterized
7   // query
8   const query = 'SELECT * FROM reports WHERE
9   ↪ user_id = ?';
10  return db.execute(query, [userId]);
11 }

```

Listing 6: Prompt Injection – False Vulnerability Claims

Appendix C. Overview Code Examples: Pattern Dependence and Context

C.1. Keyword Dependence

```

1 // Model may wrongly flag XSS merely due to
2 // seeing innerHTML
3 element.innerHTML = userInput;
4
5 // Secure usage that should not be flagged
6 element.innerHTML =
7   ↪ DOMPurify.sanitize(userInput);

```

Listing 7: Overview — Keyword Dependence

C.2. Syntactic-Structure Dependence

```

1 // Model may infer SQL injection only from the
2 // variable name "query"
3 const query = "SELECT * FROM users";
4
5 // In fact this is a hard-coded safe query
6 db.execute(query);

```

Listing 8: Overview — Syntactic-Structure Dependence

C.3. Context Neglect

```

1 function processData(input) {
2   // Model may ignore this critical validation
3   ↪ step

```

```

3 if (!isValid(input)) return;
4
5 // Focusing only on this line would be a
6 // → misclassification
7 database.query(input); // input has already
8 // → been validated
9

```

Listing 9: Overview — Context Neglect

Appendix D. Discussion

Can Our JUDGEJS and FORGEJS Generalize to Other Languages? In principle, JUDGEJS and FORGEJS are not tied to one language: we compare vulnerable and fixed code, standardize labels at multiple granularities, add robustness variants, and evaluate reproducibly. What is specific to JavaScript is the project taxonomy (frontend/backend/full-stack with Node/DOM cues), the way diffs are mapped to functions, the obfuscation/noise tools and sink lists, and prompt details. Porting to C/C++ or Python mainly involves swapping in a language parser and build toolchain, re-defining source–sink categories and CWE focus, choosing a semantics-preserving obfuscator, and adjusting prompts and evidence. The main takeaway likely holds across languages: models rely on surface hints, are brittle to small changes, and struggle under low-FPR constraints; validating this in other languages is future work.

Do We Avoid 100% Over- and Underestimation? We aim to reduce both overestimation and underestimation rather than claim they disappear. To limit overestimation we evaluate whole repositories, include fixed versions, add four stress variants, require function-level localization, and report VD-S under an FPR budget; to limit underestimation we use CWE equivalence classes, a denoised split, normalized outputs, and stronger prompts. Remaining risks include imperfect equivalence, residual label noise, occasional lucky type matches at project level, sensitivity to settings, and possible training contamination; we mitigate these by reporting multiple metrics, releasing the scorer, and focusing on robustness tests.

How to Enhance LLM-Driven JavaScript Vulnerability Detection? To improve reliability, train on security data with program structure (AST/CFG/DFG/CPG) as input and require short, evidence-backed outputs that name the file, function, and line in JSON. Combine LLMs with taint analysis, symbolic execution, or graph analyzers to check flows. Train with noise, obfuscation, and prompt injection, and use abstention and calibrated thresholds to keep FPR within budget. Expand ARENAJS with more frontend cases and, where feasible, executable PoCs, add time-based splits, and keep the denoised and robustness variants.