

# An Insight into Security Code Review with LLMs: Capabilities, Obstacles and Influential Factors

JIAJIN YU, School of Computer Science, Wuhan University, China

PENG LIANG, School of Computer Science, Wuhan University, China

YUJIA FU, School of Computer Science, Wuhan University, China

AMJED TAHIR, School of Mathematical and Computational Sciences, Massey University, New Zealand

MOJTABA SHAHIN, School of Computing Technologies, RMIT University, Australia

CHONG WANG, School of Computer Science, Wuhan University, China

YANGXIAO CAI, School of Computer Science, Wuhan University, China

Security code review is a time-consuming and labor-intensive process typically requiring integration with automated security defect detection tools. However, existing security analysis tools struggle with poor generalization, high false positive rates, and coarse detection granularity. Large Language Models (LLMs) have been considered promising candidates for addressing those challenges. In this study, we conducted an empirical study to explore the potential of LLMs in detecting security defects during code review. Specifically, we evaluated the performance of six LLMs under five different prompts and compared them with state-of-the-art static analysis tools. We also performed linguistic and regression analyses for the best-performing LLM to identify quality problems in its responses and factors influencing its performance. Our findings show that: (1) existing pre-trained LLMs have limited capability in security code review but significantly outperform the state-of-the-art static analysis tools. (2) GPT-4 performs best among all LLMs when provided with a CWE list for reference. (3) GPT-4 frequently generates responses that are verbose or not compliant with the task requirements given in the prompts. (4) GPT-4 is more adept at identifying security defects in code files with fewer tokens, containing functional logic, or written by developers with less involvement in the project.

CCS Concepts: • Software and its engineering → Software post-development issues.

Additional Key Words and Phrases: Large Language Model, Code Review, Security Analysis

## ACM Reference Format:

Jiaxin Yu, Peng Liang, Yujia Fu, Amjed Tahir, Mojtaba Shahin, Chong Wang, and Yangxiao Cai. 2024. An Insight into Security Code Review with LLMs: Capabilities, Obstacles and Influential Factors. *J. ACM* 0, 0, Article 0 (2024), 26 pages. <https://doi.org/XXXXXX.XXXXXXX>

## 1 INTRODUCTION

Security defects in software programs can have severe consequences, including data breaches, financial losses, and service disruptions [12, 67]. The longer a security defect stays in the program,

---

Authors' addresses: Jiaxin Yu, School of Computer Science, Wuhan University, Wuhan, China, jiaxinyu@whu.edu.cn; Peng Liang, School of Computer Science, Wuhan University, Wuhan, China, liangp@whu.edu.cn; Yujia Fu, School of Computer Science, Wuhan University, Wuhan, China, yujia\_fu@whu.edu.cn; Amjed Tahir, School of Mathematical and Computational Sciences, Massey University, New Zealand, a.tahir@massey.ac.nz; Mojtaba Shahin, School of Computing Technologies, RMIT University, Australia, mojtaba.shahin@rmit.edu.au; Chong Wang, School of Computer Science, Wuhan University, Wuhan, China, cwang@whu.edu.cn; Yangxiao Cai, School of Computer Science, Wuhan University, Wuhan, China, yangxiaocai@whu.edu.cn.

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

0004-5411/2024/0-ART0 \$15.00

<https://doi.org/XXXXXX.XXXXXXX>

the higher its associated fixing and maintenance costs are [56]. Thus, many organizations are shifting security analysis to earlier stages of software development, typically at code review time [24]. Code review is a human-intensive process in which the reviewer examines the code submitted by the developer to detect bugs, verify the implementation of the specification, ensure compliance with guidelines, and ensure quality. During this process, the reviewer raises issues in the code, discusses them with the developer, and provides recommendations. Adopting security analysis in code review can incorporate diverse viewpoints from both reviewers and developers, and effectively prevent security defects created by programmers working alone, as they naturally have a limited perspective of potential security risks [70], from being merged into the source code repository. In this study, we focus on security analysis in the context of code review, referring to it as *security code review* [20].

Security code review is increasingly integrated into the development pipelines by project teams [55]. Its resource-intensive nature, which requires significant human effort and time to review and revise the code, poses a notable challenge, particularly in popular large-scale open-source projects with numerous contributions [26]. Therefore, it has been desired to develop effective automated tools to assist code reviewers in security code review [22]. Despite various static/dynamic/hybrid program analysis tools being proposed, they each encounter practical challenges, such as excessive imprecision [50], high false positive rates [35, 66], input range limitations [27] and weak scalability [50].

Several data-driven approaches based on machine learning (ML)/deep learning (DL) have been developed for vulnerability detection and repair, which can somewhat handle the challenges encountered by traditional program analysis tools [13]. However, due to the lack of sizable and realistic training datasets [52], those approaches inevitably lacked robustness and tended to fall short when implemented in unfamiliar, real-world projects [49]. Nong *et al.* found that even using the training dataset augmented by the latest vulnerability injection technique [51], the highest F1 score and the top-1 accuracy that state-of-the-art DL-based vulnerability detection and repair tools can achieve on the dataset constructed from real-world projects are still limited, at 20.01% and 21.05% respectively. Therefore, it remains essential to further explore techniques with better generalization and robustness to address vulnerabilities in real-world projects.

Recently, pre-trained LLMs have demonstrated promising performances across a broad spectrum of software engineering tasks, such as program repair [34], test generation [38, 61], and specification generation [71]. Such models show remarkable capability to understand and generate explanatory natural language responses, thus considered a potential candidate for security code review. We expect that LLM can act as a code reviewer, not only identifying security defects in code files, but also explaining the specific defect scenarios in detail to developers in a human-readable way. Applying LLMs to support security code review can help detect security defects earlier and improve productivity by helping code reviewers check their code faster and more efficiently. However, the application of LLMs in the specific field of security code review remains largely unexplored.

Many studies have focused on leveraging LLMs for vulnerability detection. Zhou *et al.* evaluated the capabilities of GPT-3.5 and GPT-4 in vulnerability detection under few-shot learning prompts [79], while Purba *et al.* compared LLMs with static analysis tools in detecting software vulnerabilities [58]. These studies mainly conducted a coarse-grained assessment of LLMs in binary classification tasks, which was to judge whether there were security defects in the code. They did not require LLMs to provide more detailed information such as the location and type of vulnerability. Researchers have also developed prompting strategies to enhance the effectiveness of LLMs in vulnerability detection, such as the DL-based prompting framework proposed by Yang *et al.* [72] and the vulnerability-semantic guided prompting approach formulated by Nong *et al.* [49]. However,

the datasets utilized in these studies are synthetic and CVE datasets, which do not fully represent real-world codebases.

**Motivated by the above-mentioned limitations**, our aim is to bridge the knowledge gap by conducting an empirical study to comprehensively explore the potential of LLMs in fine-grained security code review on a practical dataset constructed from real-world code repositories. Specifically, in this study, LLMs were requested to provide detailed information on the identified security defect, including its line number, type, specific description and suggested fix. The dataset we utilized consists of 534 code review files obtained from four open-source projects (namely, OpenStack Nova and Neutron, and Qt Base and Creator). We evaluated the performance of LLMs, analyzed existing quality problems, and examined factors influencing their performance to offer insights into the real-world applicability and limitations of LLM in security code review. First, we identified five prompt templates to enhance LLMs (based on the prompting strategies formulated by Zhang *et al.* [76]). Then, we compared the performance of 6 open-source and closed-source LLMs with static analysis tools across these prompt templates in detecting security defects identified by reviewers. After selecting the responses generated by the best-performing LLM-prompt combination, we first manually extracted quality problems from these responses. We then used these responses to construct a cumulative link mixed model to explore the impact of 10 factors on the performance of the LLM. Our **findings** reveal that: (1) current popular general-purpose LLMs exhibit limited capability in detecting security defects identified by reviewers in code reviews. (2) *GPT-4* performs best among all LLMs when provided with a CWE list for reference. (3) *Inconcise* and *non-compliant* responses to prompts are common problems encountered by *GPT-4* in security code review. (4) *GPT-4* is more likely to identify security defects in code files with fewer tokens, containing functional logic, and written by developers with less involvement in the project. In summary, this study makes the following **contributions**:

- We conducted a fine-grained evaluation of the capability of the currently popular LLMs in security code review.
- We measured the impact of the randomness of LLMs on the consistency of security code review.
- We identified quality problems in the responses generated by the top-performing LLM to present its challenges in security code review.
- We carried out the first analysis of the factors that may influence the performance of the LLM in security code review.

**Paper Organization.** Section 2 surveys the related work. Section 3 describes the methodology employed in this study. Section 4 presents the results, followed by a discussion of the implications of our study results in Section 5. Section 6 clarifies the threats to validity. Finally, Section 7 concludes this work and outlines future directions. The replication package of this study has been provided online [75].

## 2 RELATED WORK

Since our study focuses on using LLMs to assist reviewers in identifying security defects during code review, in this section, we limit our coverage to studies related to security defect detection, excluding those studies on security defect remediation or repair.

Charoenwet *et al.* conducted an empirical study on the application of static application security testing tools (SASTs) in security code review and found that a single SAST tool could produce warnings for 52% of vulnerability-contributing commits (VCCs) [15]. By combining several tools, warnings could be produced for 78% of VCCs. However, at least 76% of these warnings are irrelevant to the vulnerability in VCCs, highlighting the challenges that SAST tools face in providing detailed

information about the vulnerability during security code review. Inspired by this work, **our work** investigated the application of LLMs in security code review, aiming to overcome the limitations of SAST tools to some extent.

Several studies have explored the performance of LLMs in detecting security defects. The majority of these studies conducted a coarse-grained assessment of LLMs in binary classification tasks that determine whether the code contains any security defects without specifying the corresponding defect type. Zhou *et al.* utilized Common Weakness Enumeration (CWE), project information, and a few-shot learning approach to design prompts and evaluated the capability of two popular LLMs, GPT-3.5 and GPT-4, in binary judgement under various prompts [79]. They found that GPT-3.5 achieved competitive performance with the prior state-of-the-art security defect detection approaches, while GPT-4 consistently outperformed the state-of-the-art. Purba *et al.* compared the performance of LLMs with popular static analysis tools in binary detection. Their work demonstrates that although LLMs achieved a good recall rate, their false positive rate was significantly higher than that of static analysis tools [58]. Unlike previous studies, **our work** provides a more fine-grained evaluation of the capability of LLMs in security code reviews by including additional insights such as defect localization, description and suggestions for defect remediation.

Previous studies also investigated whether LLMs could provide more information in security defect detection. Bakhshandeh *et al.* provided LLMs with a list of CWEs and the inspection results of static analysis tools on the prompts, requesting the LLMs to generate specific CWE names along with the corresponding code line numbers[7]. Their study shows that LLMs effectively reduce false positives of static tools and demonstrate competitive or even superior performance in defect localization compared to these tools. Yin *et al.* evaluated the performance of LLMs across multiple tasks in security analysis, including defect detection, localization, and description [73], found that the existing state-of-the-art approaches such as LineVul [21] and SVulD [48] generally outperformed LLMs. Specifically, in defect localization and description tasks, LLMs exhibited limited overall accuracy and varied performance across different CWE types. These studies have all been limited to quantitatively evaluating LLM performance. In comparison, **our work** performed an in-depth analysis of the quality problems in LLM-generated responses and factors affecting LLM performance to gain a comprehensive understanding of the potential and challenges LLMs face in security code review.

Employing specific prompting frameworks to enhance the performances of LLMs in identifying security defects has also been explored. Yang *et al.* introduced a framework utilizing DL models, integrating context learning and Chain-of-Thought (CoT) to augment prompts in security defect detection [72]. Their framework yielded superior performance compared to state-of-the-art prompting techniques. Nong *et al.* proposed a vulnerability-semantics-guided prompting (VSP) approach that combines CoT with various auxiliary information to evaluate the performance of LLMs in identifying security defects (including their classification and resolution) [49]. The research concerns and employed datasets of the above two studies [49, 72] differ significantly from ours. These studies focus on designing benchmarks to enhance LLM in vulnerability detection, utilizing synthetic datasets and CVE datasets composed of publicly disclosed vulnerabilities to validate their improvements. Compared with them, **our work** focuses on the empirical evaluation of LLMs in the context of security code review, employing a dataset of security defects collected from code reviews of real-world software projects.

It is worth noting that the aforementioned studies focus on general vulnerability detection by LLMs, whereas our work targets the specific area of security code review by LLMs. Inspired by these related works, we aim to leverage the potential of LLMs in security analysis to identify the obstacles, application strategies, and improvement directions for integrating LLMs into security code review.

### 3 METHODOLOGY

#### 3.1 Research Questions

The goal of this work is to comprehensively assess the capability of LLMs to assist with security code review by automatically detecting security defects in the given code. To achieve this goal, we divided our study into three Research Questions (RQs), as shown below. The research procedure for each RQ is illustrated in Fig. 1.

**RQ1:** How effectively do LLMs detect security defects during code review?

RQ1.1: How effectively do LLMs perform security code reviews under the basic prompt?

RQ1.2: Can prompting with auxiliary information improve the performance of LLMs?

RQ1.3: Can the Chain-of-Thought (CoT) prompting approach improve the performance of LLMs?

RQ1.4: What is the impact of the randomness of LLMs on the consistency of security code review?

**Motivation.** Many studies have focused on the performance of LLMs in detecting security defects, but they often suffer from low-quality datasets collected automatically and a narrow detection scope primarily confined to function-level code [77]. To address these limitations, RQ1 utilizes a manually curated dataset to ensure data quality. It comprehensively evaluates the capability of LLMs to review a complete code file from real-world code reviews for security defects. Additionally, due to the inherent randomness of LLMs, the responses of LLMs may exhibit considerable divergence for the same prompt. However, this aspect of LLMs has not yet been studied when applied to security tasks. Hence, we conducted a quantitative analysis of the consistency of LLM responses to explore the impact of LLM randomness on the task of security code review.

**RQ2:** What are the quality problems in LLM-generated responses during security code review?

**Motivation.** Kabir *et al.* noted that the responses of ChatGPT to programming questions may be verbose, inconsistent, and contain conceptual or logical errors [36], negatively impacting the quality of the answers. However, for security code review, the quality problems present in responses generated by LLMs have not been thoroughly analyzed in previous studies. Hence, RQ2 conducts an in-depth linguistic analysis of LLM responses for quality problems and examines the types and distribution of these problems to bridge the knowledge gap and indicate potential areas for improvement.

**RQ3:** Which factors influence the performance of LLMs in security code review?

**Motivation.** Previous studies have identified potential factors influencing the effectiveness of LLMs in detecting security defects [16, 36, 78], primarily through case studies or observations. However, there is a lack of evidence substantiating such claims. RQ3 employs a regression analysis to validate whether these factors significantly impact the performance of LLMs. This RQ helps pinpoint areas of concern in applying LLMs to security code review and highlights possible improvement directions.

#### 3.2 Data Collection

We leveraged the code review dataset constructed by Yu *et al.* [74] and examined the capability of LLMs in security code review. The dataset contains 614 review comments, each identifying a specific security defect across 15 predefined types from four projects: OpenStack Nova and Neutron (primarily written in Python), and Qt Base and Creator (primarily written in C++). These projects are the most active within OpenStack and Qt [33] and have been widely used in code review-related studies [28, 29, 64]. For each security defect, we could not include the code from the entire patchset

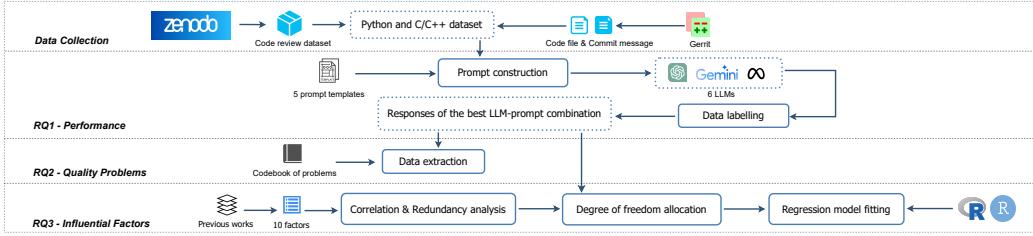


Fig. 1. An overview of the research procedure for investigating the three RQs

in our prompt, which is always too long and exceeds the input token limits of LLMs. Instead, we included only the code from the single file, where the reviewer commented to identify a security defect. The 614 security-related review comments were reorganized as follows:

First, we filtered out comments published at the patchset level, as such data cannot be located in a specific file to construct a prompt. Then, for each code file to which these comments correspond, we consolidated all security-related comments in the file and the corresponding security defect records into a single data item. Next, we utilized the Gerrit API to retrieve the content of each code file and the commit messages of the patchsets to which they belong. During this procedure, if we cannot obtain the content of the code files through the API, we ignore those files. Lastly, since 97% files in this dataset are Python, C and C++ files, we decided to remove any other files (all are configuration and scripting files) as those only formed  $\approx 3\%$  of the files and thus had a negligible impact on our results.

Our final dataset consists of 534 code files (258 Python and 276 C/C++ files), each including the source code, the commit message of the associated patchset, and detailed information of the **target security defects** as identified by the reviewers.

### 3.3 Research Procedure of RQ1

Table 1. Versions and hyper-parameters of 6 LLMs studied.

<b>Model</b>	<b>Version</b>	<b>Context Length</b>	<b>Parameter</b>	
			temperature	top_p
GPT-3.5	May 2024 version	-	-	-
GPT-4	May 2024 version	-	-	-
GPT-4 Turbo	gpt-4-1106-preview	128k	1.0	1.0
Gemini Pro	gemini-1.0-pro	32k	0.9	1.0
Llama 2 7B	llama-2-7b-chat	4k	1.0	0.9
Llama 2 70B	llama-2-70b-chat	4k	1.0	0.9

**3.3.1 LLM Selection.** To fully evaluate the performance of LLMs in security defect detection, we selected four representative LLMs from the proprietary GPT and Gemini families, along with two from the open-source Llama 2 family, which are the most popularly used LLMs when we started our experiments in December 2023. Specifically, for proprietary LLMs, we chose *GPT-3.5* and *GPT-4* provided by the ChatGPT platform (May 2024 version), and *GPT-4 Turbo* and *Gemini Pro* accessed through the API. For open-source LLMs, we deployed the smallest and largest chat versions of models in Llama 2 series—namely, *Llama 2 7B* and *Llama 2 70B*—on a server with eight NVIDIA GeForce RTX 4090 GPUs. Throughout the experiments, we kept all hyper-parameters of these

LLMs at their default settings. The versions, token constraints, and parameter settings of LLMs used in our study are summarized in Table 1.

Notably, *GPT-3.5* and *GPT-4* can be accessed through both API and the ChatGPT platform. During our preliminary exploration stage, we compared the performances of *GPT-4* and *GPT-3.5* accessed by these two ways. The two models provided by the ChatGPT platform performed better and thus were employed in our study.

**3.3.2 Prompt Design.** Prompt engineering is a process of designing prompts to optimize model performance on downstream tasks [41]. To design prompts, we followed the best practices published by OpenAI [53] and the three prompt design strategies in Zhang *et al.* [76]: basic prompt, enhanced prompt with auxiliary information, and CoT prompt. During our preliminary exploration stage, we tried prompts with various auxiliary information and different CoT intermediate reasoning steps. To fully harness the capabilities of LLMs in detecting security defects, we finally selected five best-performing prompts. Fig. 2 illustrates the construction strategies for the five prompts.

- (1) **Prompt 1 ( $\mathbf{P}_b$ ): the basic prompt.** We instruct the LLM to review the provided code for security defects and output their description, code line numbers, and suggested fix. If no defects are found, the LLM should output a fixed sentence, ‘No security defects are detected in the code.’
- (2) **Prompt 2 ( $\mathbf{P}_r$ ):  $\mathbf{P}_b + \text{Project Information}$ .**  $\mathbf{P}_r$  is designed based on  $\mathbf{P}_b$  but includes the name of the source project as auxiliary information. The LLM is asked to act as a code reviewer for the project using the provided code in the prompt.
- (3) **Prompt 3 ( $\mathbf{P}_c$ ):  $\mathbf{P}_b + \text{General CWE Instruction}$ .**  $\mathbf{P}_c$  directly instructs the LLM to use CWE as a reference for identifying security defects without providing specific CWE details.
- (4) **Prompt 4 ( $\mathbf{P}_{cid}$ ):  $\mathbf{P}_b + \text{Specific CWE Instruction}$ .**  $\mathbf{P}_{cid}$  provides a specific list of the first level CWEs in VIEW-1000 (Research Concepts)<sup>1</sup>.
- (5) **Prompt 5 ( $\mathbf{P}_{cot}$ ):  $\mathbf{P}_{cot-1} + \mathbf{P}_{cot-2}$ .** Since we only included the code of a single file in the prompt instead of an entire patchset, we designed  $\mathbf{P}_{cot}$  to analyze the impact of missing contextual information on the performance of LLMs. The task of security code review is split into two steps. In  $\mathbf{P}_{cot-1}$ , given that we only provide a single code file from a complete project in the prompt, we include the corresponding commit message of the file and instruct the LLM to generate code context according to the commit message. Then  $\mathbf{P}_{cot-2}$  guides the LLM to review the provided code for security defects with reference to the generated code context.

**3.3.3 Evaluation Metrics.** During the experiments, the responses of LLMs were provided in the form of natural language text, making it necessary to manually inspect these responses to ensure the accuracy of the performance measurements. We considered that LLM-generated responses could be evaluated in a manner similar to the approach adopted by Mahajan *et al.* [43] for measuring responses in Stack Overflow. We adapted the evaluation metrics formulated in their work [43] as follows: First, we excluded the ‘Unavailable’ category defined by Mahajan *et al.* because LLMs do not have a case of no response. Next, the definitions of each category were adjusted to match the task of security code review. Specifically, each response was categorized into one of the four rating categories detailed in Table 2. The corresponding evaluation metrics are I-Score ( $\frac{I}{I+H+U+M} \times 100\%$ ), IH-Score ( $\frac{I+H}{I+H+U+M} \times 100\%$ ), and M-Score ( $\frac{M}{I+H+U+M} \times 100\%$ ). Higher I-Score and IH-Score indicate a stronger ability of the model to detect security defects identified by reviewers. At the same time, a lower M-Score signifies less misleading results produced during security defect detection. Given

<sup>1</sup>VIEW-1000 is a weaknesses classification framework designed for academic research, widely adopted in prior work (e.g., [54]). It organizes CWEs into a hierarchical structure based on parent-child relationships among CWE entries. The first level of CWEs in VIEW-1000 encompasses a broad range of commonly encountered security defects, thus applied in Prompt 4 ( $\mathbf{P}_{cid}$ ).

<b>Prompt 1 (<math>P_b</math>):</b> the basic prompt
Please review the code below to detect security defects. If any are found, please describe the security defect in detail and indicate the corresponding line number of code and solution. If none are found, state: 'No security defects are detected in the code'.
original code
<b>Prompt 2 (<math>P_r</math>):</b> $P_b$ + Project Information
I want you to act as a code reviewer of <b>PROJECT</b> in <b>COMMUNITY</b> . Please review the code below to detect security defects. If any are found, please describe the security defect in detail and indicate the corresponding line number of code and solution. If none are found, state: 'No security defects are detected in the code'.
original code
<b>Prompt 3 (<math>P_c</math>):</b> $P_b$ + General CWE Instruction
Please review the code below for security defects using the CWE (Common Weakness Enumeration) as a reference. If any are found, please describe the security defect in detail and indicate the corresponding line number of code and solution. If none are found, state: 'No security defects are detected in the code'.
original code
<b>Prompt 4 (<math>P_{cid}</math>):</b> $P_b$ + Specific CWE Instruction
Please review the code below for security defects. You can consider defect types in terms of:
<ul style="list-style-type: none"> <li>• CWE-284 (Improper Access Control)</li> <li>• CWE-435 (Improper Interaction Between Multiple Entities)</li> <li>• CWE-664 (Improper Control of a Resource Through its Lifetime)</li> <li>• CWE-682 (Incorrect Calculation)</li> <li>• CWE-691 (Insufficient Control Flow Management)</li> <li>• CWE-693 (Protection Mechanism Failure)</li> <li>• CWE-697 (Insufficient Comparison)</li> <li>• CWE-703 (Improper Check or Handling of Exceptional Conditions)</li> <li>• CWE-707 (Improper Neutralization)</li> <li>• CWE-710 (Improper Adherence to Coding Standards)</li> </ul>
If any are found, please describe the security defect in detail and indicate the corresponding line number of code and solution. If none are found, state: 'No security defects are detected in the code'.
original code
<b>Prompt 5 (<math>P_{cot}</math>):</b> $P_{cot-1} + P_{cot-2}$
<b>P<sub>cot-1</sub>:</b> Based on the given code from a commit, please generate supplementary code files according to the commit message. #####commit message
COMMIT MESSAGE
#####code
original code
<b>P<sub>cot-2</sub>:</b> In the context of the generated files, analyze the original code for security defects. If any are found, please describe the security defect in detail and indicate the corresponding line number of code and solution. If none are detected, state: 'No security defects are detected in the code'.

Fig. 2. Construction templates for the five prompts

Table 2. Adjusted IHMU-category for LLM responses.

Classification	Definition
Instrumental (I)	The response explicitly indicates the existence of the security defect identified by the reviewer and provides a fully accurate description.
Helpful (H)	The response raises concerns related to the security defect identified by the reviewer, but their descriptions may not be entirely accurate or specific enough.
Misleading (M)	The response does not provide any helpful information, e.g., claiming no defects were found or failing to perform the detection task.
Uncertain (U)	The response does not mention any defects identified by the reviewer but instead points out the existence of other security defects. Due to the lack of expertise of the researchers about the code and its context, the actual existence of these security defects cannot be confirmed.

that LLMs exhibit considerable randomness, we repeated the experiment for each LLM-prompt combination three times to mitigate the impact of such randomness. We calculated the average of the three performance scores as the final result.

**3.3.4 Baselines.** We tried different tools and learning techniques for detecting security defects to select appropriate baselines for our work.

To compare with LLMs, we focused on techniques providing fine-grained detection, including classification, localization, and description of defects. This led us to exclude techniques with insufficient detection granularity, such as ProRLearn [59], GRACE [42] and LineVul [21]. Given that we evaluated the performance of LLMs on the C/C++ and Python datasets separately, for language-specific static analysis tools, we selected CppCheck [19] for C/C++ and Bandit [1] for Python. Furthermore, we also used CodeQL [23] and SonarQube [63] to analyze Python files. These two tools can directly analyze code files written in dynamically-typed languages like Python but require code files in statically-typed languages like C/C++ to be compiled first. Therefore, we could not apply them to our C/C++ dataset, which is comprised of uncompiled and isolated C/C++ code files. Additionally, since Semgrep [3] supports multiple programming languages and can analyze individual code files, we adopted it for both C/C++ and Python datasets. To obtain comprehensive analysis results, CodeQL adopted the Python-security-and-quality.qls test suite, covering a wide range of issues from basic code structure and naming conventions to advanced security and performance vulnerabilities. Other tools utilized their default rule sets for analysis.

**3.3.5 Data Labelling.** Prompts were constructed using five templates and then fed into six LLMs to collect their responses. Since the responses of LLMs were provided in the form of natural language text, to rate the responses accurately, we manually inspected the content of LLM-generated responses, their corresponding source code, and security defects identified by reviewers to categorize them into the four categories formulated in Table 2. To mitigate bias, pilot data labelling was conducted, in which we randomly selected 100 out of 534 code files from our dataset, constructed five prompts for each file and fed these prompts into LLMs to collect responses. Using the definitions of categories in Table 2 as the labelling criteria, the first and third authors labelled the collected responses separately. Discrepancies in labelling results were discussed with the second author to reach a consensus. The inter-rater reliability was calculated using Cohen's Kappa coefficient [18], yielding a value of 0.89. Then, the first and third authors labelled all the rest of the responses and discussed any ambiguous cases with the second author to finalize the categorization. Based on the labelling results, we calculated the performance scores of each LLM-prompt combination in the

Python and C/C++ dataset. We found that the combination of *GPT-4* and prompt  $\mathbf{P}_{cid}$  outperformed the others (as detailed in Section 4.1). Therefore, we aggregated the responses generated by *GPT-4* with  $\mathbf{P}_{cid}$  across the Python and C/C++ datasets, resulting in a total of 421 responses (some code files exceeded the token limitation of *GPT-4*) for each of the three repetitive experiments, to further analyze RQ2 and RQ3.

**3.3.6 Consistency Calculation.** To measure the randomness of LLMs, we adopted the metrics in Chang *et al.*'s work [14] and utilized entropy to measure the consistency of responses generated by LLMs across three repetitive experiments. Specifically, an LLM generated three responses for a given code file under a specific prompt template. We then calculated the entropy for the four response categories in the three responses. We determined the overall consistency by averaging the entropy across all code files using the following formula:  $\frac{1}{R} \sum_{r \in R} \text{entropy}(p_1^r, \dots, p_4^r)$ , where  $R$  is the response count in each round of the experiment. A higher average entropy suggests the LLM has higher inconsistency in repetitions of experiments.

### 3.4 Research Procedure of RQ2

We manually inspected the responses of *GPT-4* with  $\mathbf{P}_{cid}$  across the three repetitive experiments to analyze the quality problems present in these responses. We employed the open coding and constant comparative method [25] with a predefined category by the MAXQDA tool for qualitatively analyzing the responses.

Based on the problem categories formulated by Kabir *et al.* [36] and the problems of LLM-generated text in various domains identified by previous studies [8, 37, 46], two of the authors firstly inspected a few *GPT-4* responses independently and recorded their observations. Then they collaboratively reviewed the inspected responses and discussed to establish an initial codebook, resulting in four themes: *Correctness*, *Understandability*, *Conciseness*, and *Compliance*, each with a series of problem types. In particular, we excluded the *Consistency* theme defined in Kabir *et al.* [36] from our study. The reason is that the consistency between the LLM's responses and the reviewers' assessments (i.e., the ground truth) has been investigated in RQ1. Following the codebook, we conducted a pilot data analysis across two iterations to improve the agreement between the two authors. In each round, ten responses were randomly selected for the two authors to analyze their quality problems. The results were then compared and discussed, with partial adjustments made to the codebook. Since a single response could contain multiple problems, two authors completed the remaining data analysis task where the labels are not mutually exclusive. Therefore, we calculated the agreement level using Krippendorff's alpha [39] with Jaccard distance, which was improved from 0.40 in the first round to 0.80 in the second, indicating a high level of agreement. Then the remaining data analysis was completed by the two authors and verified by the second author. Throughout the entire procedure, we employed a negotiated agreement approach [10]: any conflicts were consulted and addressed by three authors, ensuring the reliability of data analysis. For each type of problem, we computed its average frequency in three repetitive experiments as the final occurrence rate of this problem in *GPT-4* responses.

### 3.5 Research Procedure of RQ3

To answer RQ3, we adopted the approach formulated by Harrell [30] to construct a regression model. From Table 2, we can see that the response variables in our dataset are the ratings of I, H, U, and M. In terms of the helpfulness of the response, by definition, I is more important than H, and so forth. Accounting for the ordinal nature of the response variable and the randomness of the LLM in three rounds of experiments, we ultimately aggregated the data from all three experiments to fit a cumulative link mixed model [57]. This model supports ordered categorical response variables

and allows us to include random effects to group responses in each round together. The specific construction process is described in the following subsections.

Table 3. Factors that may influence the performance of LLM in security code review

Factor	Source*	Definition	Rationale
Token	Adjusted from [45, 55]	Number of tokens of the code file. Computed using the TikToken [4] tokenizer.	Larger code files may be more difficult for LLMs to analyze.
FileType	Created	Type of the code file, i.e., <i>Source</i> or <i>Auxiliary</i> . <i>Source</i> indicates files that directly contain the logic of the application (i.e., .cpp, .c and .py) while <i>Auxiliary</i> files are utilized for declarations (i.e., .h and .hpp).	The understanding capability of LLMs and the scenarios of security defects may vary across different types of files.
SecurityDefectType	Transformed from [16]	Type of the security defect that the reviewers identified in this code file, i.e., one of the 15 types of security defect formulated by Yu <i>et al.</i> [74].	ChatGPT is proven to display distinct detection performances across various vulnerabilities [16, 78]. Thus, we assume that LLMs may be more adept at detecting certain specific types of security defects in code.
Commit	Collected from [55]	Number of historical commits for the code file in its code change.	Excessive history commits may lead developers to fatigue or overconfidence, resulting in their ignorance of security considerations in simpler scenarios.
Complexity	Collected from [45, 55]	McCabe's Cyclomatic Complexity [44]. Computed using the Lizard [2] analyzer.	Code comprehensibility may become more challenging for LLMs as its cyclomatic complexity increases [55].
Community	Created	Name of the community source of the code file, i.e., OpenStack or Qt.	OpenStack and Qt differ significantly in terms of their structure, technology stack, project function, and main components. The LLMs' performance may differ on data derived from these two communities.
AuthorExperience	Collected from [55]	Number of closed code commits the author has submitted prior to the commit this code file corresponding to.	Experienced authors may write more rigorous and standardized code, while the security vulnerabilities in the code are also relatively more subtle and difficult to detect.
AnnotationRatio	Created	The ratio of manual annotations in the code file. If the token number of all annotations in the file $f$ is $a_f$ and the total token number of this file is $f_t$ , then $AR = a_f/f_t$ .	The capability of LLMs to comprehend manual annotations and code may vary.
AnnotationHasCode	Transformed from [16]	Whether the annotations in the code file include code.	Code in annotations was abandoned by developers but often interpreted as actual code by LLMs [16], leading to incorrect security defect detection.
AnnotationisSecurityRelated	Created	The relevance of annotations to the security defects identified by reviewers in this code file, i.e., 1 (annotations mentioned target security defects identified by reviewers), 0 (Annotations were unrelated to security) and -1 (Annotations mentioned non-target security defects).	Keywords in the description related to target security defect in annotations can act as cues for LLMs, while the descriptions of non-target security defects often involve the historical fixing or security mechanisms, which may cause interference to LLMs.

\* The sources of factors correspond to the descriptions in **bold** in Section 3.5.1.

### 3.5.1 Attributes Construction.

To comprehensively collect factors influencing the detection of security defects by LLMs, excluding the prompt and the model itself, we undertook the following steps:

First, we **collected** the attributes from previous works [45, 55, 68] that were utilized in analyzing relationships related to code review and applicable to the code file data in our work. Given that LLMs understand and generate text token by token, we **adjusted** the factor adopted from McIntosh *et al.*'s work [45] – ‘Size’, to ‘Token’ (see the first row of Table 3), as larger code files typically contain more tokens. Next, we extracted characteristics of data that were explicitly indicated to affect or correlate with the capabilities of LLMs in Chen *et al.*'s work [16], **transforming** them into factors

for our experiment. Lastly, after thoroughly observing and analyzing the distribution of responses, we further **created** and adopted several potentially influential factors based on our observations. In the end, we obtained a list of 10 factors in Table 3, along with the source, definition, and rationale for each factor.

Since acquiring a factor like ‘AuthorExperience’ entails querying the code review history of projects, we utilized Gerrit API to retrieve detailed information for all patchsets under each code change. We used Python scripts complemented with manual analysis to query, compute, and label these factors. Besides, since we adopted ‘SecurityDefectType’ in our factor list, we only considered data where the code file contains a single security defect to avoid the case in which ‘SecurityDefectType’ could have multiple values. After filtering out code files with multiple security defects, we aggregated the dataset from three repetitive experiments. We obtained a total of 1,215 responses generated by the GPT-4 and  $P_{cid}$  combination. This comprised 405 responses from each experiment and their factor values ready for model fitting.

### 3.5.2 Regression Model Construction.

**Correlation and Redundancy Analysis:** If explanatory variables are highly correlated, it can cause interference in model construction and analysis. As the predefined explanatory variables in our study include both categorical and continuous variables, we utilized a refined Pearson’s hypothesis test of independence – Phi\_K [6], to measure the correlation between each factor. The correlation coefficients obtained are illustrated in Fig. 3. We chose  $|\rho| > 0.7$  as the threshold because it is recommended as the threshold for strong correlation [32]. From Fig. 3, where an abbreviation denotes each factor, it can be seen that the correlation coefficients between each pair of factors are all below 0.7, suggesting that there is no strong correlation between these factors. However, two factors without high correlation can still be duplicates, distorting the relationship between explanatory and response variables in model fitting. We encoded categorical variables as dummy variables for redundancy analysis [40]. Utilizing the *redund* function in the *rms* package, we assessed potential redundant variables with its default threshold of  $R^2 \geq 0.9$ , ultimately finding none.

**Degrees of Freedom Allocation:** To assign degrees of freedom to each explanatory variable, we measured the potential of nonlinear relationships between each factor and the response variable by calculating the Spearman’s rank correlations ( $\rho^2$ ) between them. As Harrell [30] recommended, explanatory variables with larger  $\rho^2$  can be assigned more degrees of freedom within the range of 3 to 5. According to the computation results shown in Fig. 4, for interval variables, we considered using the *rcs* function from the *rms* package to allocate 3 degrees of freedom for the factor ‘Token’ with the highest  $\rho^2$ , while the other factors are allocated 1 degree of freedom. Category variables should be converted into factor variables, and their corresponding degrees of freedom are automatically allocated during model fitting.

**Model Fitting:** Next, we fitted the 1,215 responses and their rating categories onto a cumulative link mixed model using the *clmm* function from the *ordinal* R package [17]. We tested models with and without allocating 3 degrees of freedom to the ‘Token’ factor to balance model complexity and fit. The model that did not allocate these degrees of freedom, which had a lower Akaike Information Criterion (AIC), was then selected as our final fitting result [9].

All the results and scripts utilized in data labelling, extraction, and regression analysis are provided in our replication package [75].

## 4 RESULTS

### 4.1 Performance of LLMs (RQ1)

Based on the evaluation metrics defined in Section 3.3.3, we evaluated the capability of six popular LLMs in security code review across five different prompts. We compared the performance of LLMs

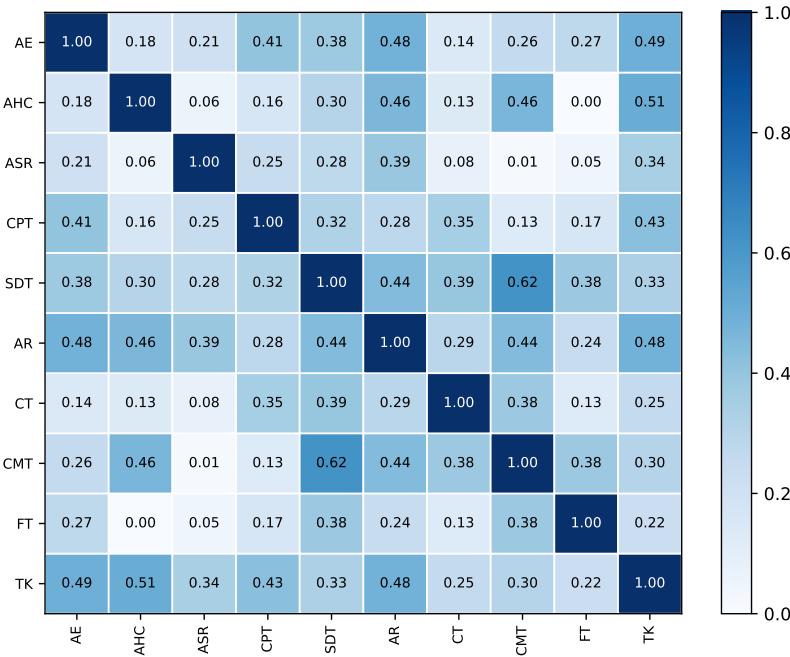
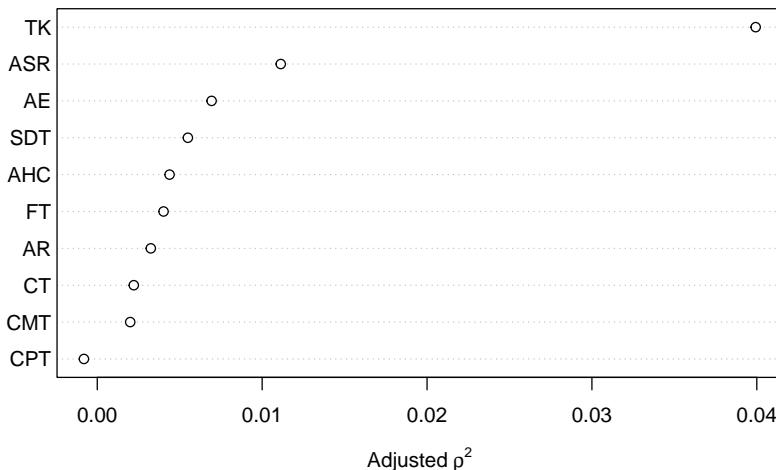


Fig. 3. The Phi\_K correlation coefficients of explanatory variables

Fig. 4. Dotplot of the Spearman  $\rho^2$  between explanatory and response variables across three repetitive experiments

with baseline tools in the C/C++ and Python datasets, respectively. If two of the three metrics were superior (i.e., higher I-Score or IH-Score, lower M-Score), we deemed that the current combination outperforms the others. Due to differences in the token limit of each LLM and the inherent token count of each prompt template, the number of responses successfully generated by each LLM-prompt combination also varied. To ensure rigorous comparisons, we selected the intersection

of source code files in our complete dataset so that all LLM-prompt combinations could generate responses, resulting in 57 instances. As a result, two sets of scores are provided: (1) the response count and performance scores of each LLM-prompt combination on the complete Python and C/C++ dataset (see Table 4 and Table 5), and (2) the performance scores of each LLM-prompt combination on the selected 57 instances (see Table 6). The best values for each score are highlighted in bold and marked with asterisks in the tables.

From Table 4 and Table 5, it is evident that, on the Python dataset, the combination of *GPT-4* with  $\mathbf{P}_{cid}$  achieved the best performance. However, on the C/C++ dataset, the combination of *Llama 2 70B* with  $\mathbf{P}_{cid}$  performs the best. However, Table 6 demonstrated that in the selected 57 cases, *Llama 2 70B* scored lower than *GPT-4*. We speculate that this discrepancy could be attributed to the smaller volume of responses obtained by *Llama 2 70B*, leading to inflated performance scores of *Llama 2 70B* in the complete datasets. On the whole, we consider *GPT-4* paired with  $\mathbf{P}_{cid}$  as the top-performing LLM-prompt combination. As of the time of our experiment (May 2024), the capabilities of currently popular LLMs in conducting security code reviews are still limited. Notably, with enhanced prompts, the best performance of each LLM in security code review is significantly superior to that of static analysis tools. This could be attributed to two reasons: (1) for context-sensitive tools such as CodeQL, static analysis can be impacted by the lack of code context in our dataset [60]; (2) there is a relatively high frequency of security defects related to multi-threading and asynchronous programming, such as *race conditions*. As such features can only be accurately modelled using dynamic analysis, these security defects are challenging for static analysis tools to detect [15].

**4.1.1 RQ1.1.** When using basic prompts constructed by  $\mathbf{P}_b$  (see Table 4 and Table 5), *GPT-4* performs the best, obtaining the highest I-Score (Python: 4.76%, C/C++: 4.45%), IH-Score (Python: 9.34%, C/C++: 7.96%), and the lowest M-Score (Python: 61.31%, C/C++: 55.47%). Following closely behind is *Llama 2 70B* and *GPT-3.5*. *GPT-4 Turbo*, *Gemini Pro*, and *Llama 2 7B* performed the worst, with their performance in the Python dataset even inferior to that of the baseline tool, Bandit. Similar results are observed in Table 6. Specifically, in the selected 57 instances, under the basic prompt  $\mathbf{P}_b$ , *GPT-4* achieved I-Score = 8.77%, IH-Score = 20.46%, and M-Score = 43.86%. It is evident that *GPT-4*'s performance on the selected 57 instances exceeds its performance across the entire dataset. Given that these 57 code files are instances with relatively fewer tokens within our dataset, it could be speculated that the quantity of tokens input to LLM affects its performance in detecting security defects, which has been further explored and analyzed in RQ3.

**4.1.2 RQ1.2.** As shown in Table 4 and Table 5, the changes of the performance scores of each LLM under prompt  $\mathbf{P}_r$  compared with that under  $\mathbf{P}_b$  is generally small. Specifically, *GPT-3.5*, *GPT-4*, and *GPT-4 Turbo* show subtle increases and decreases in performance on the Python and C/C++ datasets, respectively. *Gemini Pro* and *Llama 2 70B* demonstrate minor performance decreases on both datasets, while *Llama 2 7B* shows a slight improvement. A possible reason is that the name of components or packages utilized may already imply the source project of the code file, thereby diminishing the impact of project information provided in  $\mathbf{P}_r$ . Furthermore, the improvement of emphasizing the task through adding a persona in the prompt is LLM-specific. We speculate that some LLMs may already understand the task described in the prompt quite well, and thus, introducing a persona could add unnecessary complexity to the prompt and lead to distractions.

We can see from Table 4 and Table 5 that compared to using the basic prompt  $\mathbf{P}_b$ , the performance of *GPT-3.5* significantly decreases under  $\mathbf{P}_c$  and  $\mathbf{P}_{cid}$  in both Python and C/C++ datasets. The performances of *GPT-4* with  $\mathbf{P}_c$  exhibit different changes across different datasets, with an improvement in the C/C++ dataset but a decrease in the Python dataset. However, all other LLMs achieved higher performance scores across both datasets. Moreover, *GPT-4*, *GPT-4 Turbo*, *Gemini*

Table 4. Performance scores and entropy for each LLM-Prompt combination on the Python dataset

Python					
Model & Prompt		P <sub>b</sub>	P <sub>r</sub>	P <sub>c</sub>	P <sub>cid</sub>
GPT-3.5	Resp. Count	123	123	123	123
	I-Score	1.08%	1.35%	0.27%	0.27%
	IH-Score	2.71%	3.80%	1.35%	0.27%
	M-Score	89.37%	84.28%	92.41%	97.29%
	Entropy	0.2572	0.3692	0.2070	0.0746
GPT-4	Resp. Count	175	175	175	174
	I-Score	4.76%	3.43%	3.24%	5.17*%
	IH-Score	9.34%	9.52%	8.38%	14.56*%
	M-Score	61.31%	60.95%	46.47%	44.06%
	Entropy	0.7426	0.7283	0.8600	0.9416
GPT-4 Turbo	Resp. Count	201	201	201	201
	I-Score	0.66%	0.33%	0.66	0.83
	IH-Score	0.66	0.83	2.32	1.99
	M-Score	97.18%	96.35%	85.74%	92.87%
	Entropy	0.1757	0.0746	0.3370	0.6125
Gemini Pro	Resp. Count	162	161	160	160
	I-Score	0.00%	0.00%	0.63%	1.25%
	IH-Score	0.41%	0.00%	0.84%	3.54%
	M-Score	90.54%	97.09%	84.04%	55.21%
	Entropy	0.0777	0.0901	0.2424	0.1873
Llama 2 7B	Resp. Count	62	62	61	61
	I-Score	0.00%	0.54%	1.62%	2.19%
	IH-Score	1.61%	1.61%	5.40%	8.74%
	M-Score	87.10%	83.33%	56.78%	40.98%
	Entropy	0.2625	0.3070	0.4766	0.4351
Llama 2 70B	Resp. Count	62	62	62	61
	I-Score	0.54%	1.07%	2.15%	3.83%
	IH-Score	4.84%	4.30%	6.45%	11.48%
	M-Score	79.57%	84.41%	58.60%	43.17%
	Entropy	0.4470	0.3474	0.3514	0.3832
Baselines	SAST	SonarQube	CodeQL	Bandit	Semgrep
	Resp. Count	258	258	258	258
	I-Score	0.00%	0.39%	0.00%	0.00%
	IH-Score	0.00%	1.16%	1.16%	0.00%
	M-Score	91.47%	82.56%	78.29%	90.31%

*Pro*, *Llama 2 7B*, and *Llama 2 70B* all achieved their best performance under P<sub>cid</sub>. Therefore, it can be inferred that adding information related to CWE to the prompt generally enhances the capabilities of security defect detection of various LLMs in code review. It is worth noting that under P<sub>cid</sub>, the decrease in GPT-4's I-Score (Python: 4.76% → 3.24%, C/C++: 4.45% → 3.91%) and in M-Score (Python: 61.31% → 46.47%, C/C++: 55.47% → 44.53%) reflect that guiding LLMs to reference CWE in the prompt indeed enables GPT-4 to detect more security defects. Still, it does not guarantee the accuracy of the detection results. Under P<sub>cid</sub>, providing a specific CWE list as a reference in the prompt leads to significant improvements across all metrics and datasets for GPT-4. Hence, how to leverage CWE information in the prompt is worthy of consideration.

Table 5. Performance scores and entropy for each LLM-Prompt combination on the C/C++ dataset

C/C++					
Model & Prompt		P <sub>b</sub>	P <sub>r</sub>	P <sub>c</sub>	P <sub>cid</sub>
GPT-3.5	Resp. Count	141	141	141	139
	I-Score	1.89%	2.13%	0.00%	0.24%
	IH-Score	5.67%	3.31%	0.47%	0.96%
	M-Score	76.60%	86.05%	99.29%	95.44%
	Entropy	0.4860	0.2439	<b>0.0195*</b>	0.1153
GPT-4	Resp. Count	247	247	247	247
	I-Score	4.45%	3.91%	3.91%	5.40%
	IH-Score	7.96%	8.01%	9.31%	11.74%
	M-Score	55.47%	64.24%	44.53%	39.00%
	Entropy	0.6215	0.5329	0.6505	0.6955
GPT-4 Turbo	Resp. Count	269	269	269	269
	I-Score	0.25%	0.12%	0.25%	0.50%
	IH-Score	0.25%	0.37%	1.24%	1.61%
	M-Score	98.02%	98.64%	91.08%	93.43%
	Entropy	0.1430	0.0457	0.2493	0.5413
Gemini Pro	Resp. Count	223	221	221	221
	I-Score	0.00%	0.15%	0.30%	1.66%
	IH-Score	0.45%	0.15%	0.90%	3.77%
	M-Score	92.97%	98.19%	89.44%	61.54%
	Entropy	0.0512	0.0332	0.1401	0.1576
Llama 2 7B	Resp. Count	57	57	57	51
	I-Score	0.58%	0.00%	0.00%	2.61%
	IH-Score	2.34%	3.51%	5.85%	4.58%
	M-Score	90.06%	87.13%	81.29%	52.94%
	Entropy	0.2650	0.2167	0.2167	0.4092
Llama 2 70B	Resp. Count	57	57	57	51
	I-Score	1.17%	1.17%	1.75%	<b>6.54*%</b>
	IH-Score	4.68%	2.92%	3.51%	<b>15.03*%</b>
	M-Score	67.25%	80.12%	66.08%	40.52%
	Entropy	0.5667	0.4145	0.3895	0.5515
Baselines	SAST	CppCheck		Semgrep	
	Resp. Count	276	276		
	I-Score	0.00%	0.00%		
	IH-Score	0.72%	0.36%		
	M-Score	98.19%	96.74%		

4.1.3 *RQ1.3.* Table 4 and Table 5 illustrate that P<sub>cot</sub> significantly enhances the ability to detect security defects for *Llama 2 7B* and *Llama 2 70B* in both Python and C/C++ datasets. However, for *Gemini Pro* and *GPT-4 Turbo*, there is no improvement and even a detrimental effect to these two LLMs. As for *GPT-3.5* and *GPT-4*, they respectively exhibit improvement and decrease in their performances across different datasets. It can be seen that the impact of supplementing missing context by CoT prompting on the performance of LLMs is multifaceted and unstable. A possible explanation is that different LLMs vary in their ability to process and understand input. Moreover, we instructed the LLM to generate missing code context based on the corresponding commit message in P<sub>cot</sub>. On the one hand, the generated context may contain the critical information

Table 6. Performance scores for each LLM-Prompt combination on 57 instances that all LLM-prompt combinations could generate responses

Model & Prompt		$P_b$	$P_r$	$P_c$	$P_{cid}$	$P_{cot}$
GPT-3.5	I-Score	3.51%	3.51%	0.58%	0.58%	2.34%
	IH-Score	7.60%	8.77%	1.17%	0.58%	4.09%
	M-Score	81.87%	79.53%	94.74%	95.32%	77.19%
GPT-4	I-Score	8.77%	8.77%	5.85%	12.87*	9.94
	IH-Score	20.46%	19.30%	14.04%	23.98*	19.88
	M-Score	43.86%	42.11%	45.61%	40.35%	41.52%
GPT-4 Turbo	I-Score	0.58%	1.75%	2.34%	2.34%	0.00%
	IH-Score	0.58%	3.51%	5.26%	4.09%	1.75%
	M-Score	97.08%	91.23%	75.44%	88.30%	92.98%
Gemini Pro	I-Score	0.58%	0.00%	0.00%	0.58%	0.00%
	IH-Score	1.17%	0.00%	0.58%	2.92%	0.00%
	M-Score	96.49%	100.00%	88.30%	76.61%	100.00%
Llama 2 7B	I-Score	0.58%	1.17%	1.17%	0.58%	1.75%
	IH-Score	1.75%	2.34%	4.09%	2.34%	7.60%
	M-Score	87.13%	84.21%	60.82%	53.80%	25.73%
Llama 2 70B	I-Score	2.34%	1.17%	2.92%	5.26%	4.09%
	IH-Score	4.68%	2.34%	7.02%	17.54%	4.68%
	M-Score	70.18%	80.12%	42.69%	20.47*	44.44%

needed to analyze the target security defect, which can help the LLM in security code review. On the other hand, the consistency between this generated context and the actual code context cannot be guaranteed. We observed that in some cases, the code context generated under Prompt  $P_{cot-1}$  was inconsistent with the actual code context, resulting in the original security defect no longer being identified as a defect within the generated context. Consequently, when Prompt  $P_{cot-2}$  was applied, the detection of security defects in the provided code was misled by the generated code context and yielded erroneous results.

**4.1.4 RQ1.4.** According to the mean entropies of various LLM-prompt combinations in Table 4 and Table 5, *GPT-3.5*, *GPT-4 Turbo*, and *Gemini Pro* exhibit relatively higher consistency in their responses across three repetitive experiments, followed by *Llama 2 7B* and *Llama 2 70B*, with *GPT-4* showing the least consistency. Specifically, on the Python dataset, the highest consistency of responses was observed under  $P_{cot}$  with *GPT-4 Turbo*, with a mean entropy of 0.0404. On the C/C++ dataset, the combination of *GPT-3.5* and  $P_c$  demonstrates the lowest mean entropy of 0.0195. Given these two LLM-prompt combinations' poor security defect detection capabilities, their high consistency may be due to their tendency not to detect any security defects in the code. *GPT-4* paired with  $P_{cid}$  is considered the top-performing combination in security code review but shows high mean entropies on both Python and C/C++ datasets, with 0.9416 and 0.6955 respectively. These results suggest that LLMs with more robust security defect detection capabilities may also exhibit greater randomness.

## 4.2 Quality Problems in Responses (RQ2)

The average distribution of 12 problem types across four themes in responses to three repetitive experiments is illustrated in Fig. 5, where the same color represents types within the same theme, with darker shades indicating the themes.

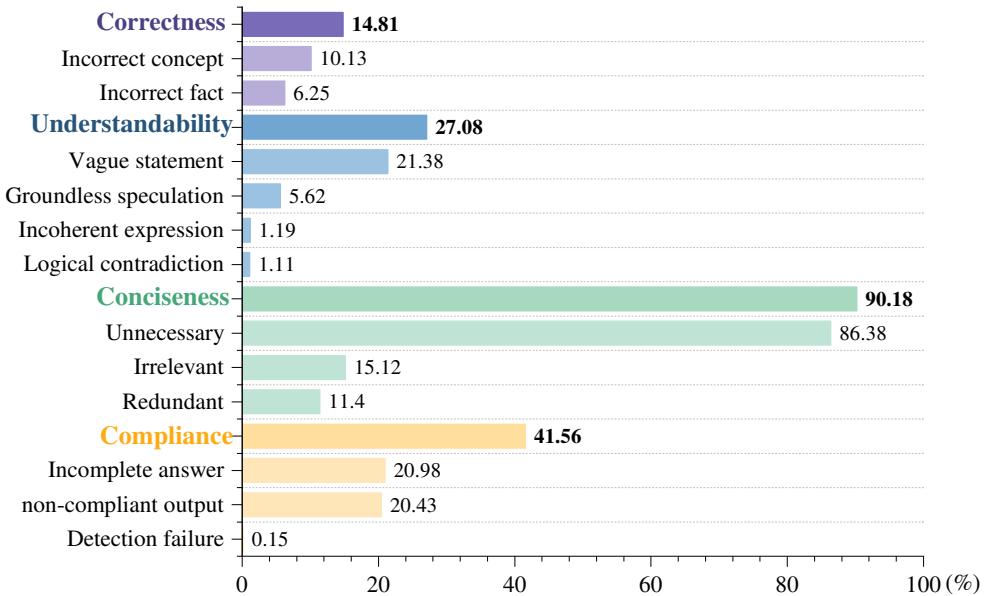


Fig. 5. Average proportion of each problem type present in responses generated by GPT-4

From the perspective of the four themes, most responses contain problems related to **Conciseness**, accounting for 90.18%. Nearly half of the responses (41.56%) involve **Compliance** related problems, which indicate the requirements of the task described in the prompt are not met. Moreover, 27.08% of the responses show problems of **Understandability** due to inappropriate logic, expression, or attribution. Notably, problems related to **Correctness**, such as deviation from objective facts or common knowledge in the field of security, are encountered less frequently than other themes (14.08%). The typical types of problems within each theme are described in detail below.

**Correctness:** The responses reveal two types of incorrectness: *Incorrect fact* (6.25%) and *Incorrect concept* (10.13%). The responses with *Incorrect fact* include: incorrect line numbers of the code, function names not matching the source code provided, names not corresponding to the CWE-ID, and the described code content not matching the actual scenario. *Incorrect concept* indicates a misunderstanding of the concept of a certain CWE-ID or a specific security defect type. As shown in the example below (Example 1), the inaccurate comparison logic described in the response was mistakenly categorized as CWE-682, thus the response is considered to exhibit the *Incorrect concept* problem.

```
Example 1 - Incorrect concept
GPT-4: ...**CWE-682 (Incorrect Calculation):**
**Issue:** The method `inventory_has_changed` uses a comparison logic that might not be
entirely accurate in all scenarios...
```

**Understandability:** Problems that reduce understandability of responses include *Vague statement* (21.38%), *Incoherent expression* (1.19%), *Logical contradiction* (1.11%), and *Groundless speculation* (5.26%). Among these, *Vague statement* is the most prevalent one. It refers to a situation where the description of the detected security defect is not specific enough, failing to pinpoint the defect to a specific location or providing a overly general fix suggestion. An example (Example 2) of *Vague statement* is shown below, in which *GPT-4* did not provide a concrete location or line number for the identified security defect in the response.

**Example 2 - Vague statement****GPT-4:** ...

\*\*CWE-703(Improper Check or Handling of Exceptional Conditions)\*\*

\*\*General Concern\*\*: Throughout the code, there are instances where error conditions might not be handled comprehensively...

**Conciseness:** In all responses, 11.4% contain *Redundant* content, 15.12% include information *Irrelevant* to the task of security defect detection, and 86.38% contain *Unnecessary* information, all of which are problems related to conciseness of responses. Among them, *Unnecessary* information, which refers to information that is somewhat related to security code review but does not contribute to detecting security defects, frequently appears in the responses generated by *GPT-4*. Such information includes the reasons for the absence of security defects in the current code and general suggestions on security measures and review processes. As illustrated in the example below (Example 3), the response indicates that no defect related to CWE-284 exists in the code and further elaborates on the reason. However, this is not the content we required *GPT-4* to generate in the prompt, hence is deemed to be *Unnecessary* information.

**Example 3 - Unnecessary****GPT-4:** ...\*\*CWE-284 (Improper Access Control)\*\*:

- There doesn't seem to be any direct issue related to improper access control in the given code...

**Compliance:** Three types of non-compliance were identified regarding whether the responses satisfied the requirements in the prompts. Among them, *Detection failure* is observed in very few responses (0.15%), where *GPT-4* explicitly claimed that it could not perform security check in its response or misinterpreted requirements in the prompts for other tasks. *Non-compliant output* was present in 20.43% of responses, in which *GPT-4* fails to include the fixed phrase ‘No security defects are detected in the code’, when no defects are found. *Incomplete answer* is the most common problem (20.98%) under this theme, referring to cases in which *GPT-4*’s responses lack the line numbers and fix suggestions of detected security defects requested in the prompt. From the example below (Example 4), we can see that while the potential memory management issues caused by temporary files and directories are described and a fix suggestion is provided, no specific line numbers are given. Hence, it is categorized as an *Incomplete answer*.

**Example 4 - Incomplete answer****GPT-4:** ### 3. CWE-664

\*\*Potential Issue:\*\* The creation of temporary files and directories (e.g., in function `createQmlrcFile`) ... can potentially lead to resource management issues.

\*\*Solution:\*\* Implement thorough checks and error handling...

While Kabir *et al.* [36] also evaluated the quality problems in LLM-generated responses, the application domain of LLMs and the distribution of identified problems in their work differ from ours. Specifically, Kabir *et al.* requested LLMs to answer programming questions on Stack Overflow, finding that over half (52%) of the responses suffered from correctness issues. In contrast, our study applied LLMs to security code review, revealing a much lower incidence of correctness issues at 14.81%. This indicates that LLMs face different challenges across various software engineering tasks, highlighting the need to optimize and customize LLMs for specific areas of application.

### 4.3 Factors Influencing LLMs (RQ3)

**4.3.1 Model Stability.** To assess the adequacy of the fit of our regression model, we utilized Nagelkerke’s Pseudo  $R^2$  [47]. This yielded a value of 0.158, which is relatively lower than the value of regression analysis in other fields reported in e.g., [55, 69]. A possible explanation is that although we have considered the randomness of LLM responses during model fitting, the random effects incorporated into our model may not be sufficient to capture all the random variation

Table 7. Explanatory powers of factors to the performance of GPT-4 under  $P_{cid}$  in security code review.

Factor (:Reference)		Odds Ratio	D.F.	$\chi^2$	Pr(>Chi)
Token		0.92	1	48.07	< .0001***
Community	Qt:Openstack	1.94	1	19.03	< .0001***
FileType	Source:Auxiliary	2.82	1	18.28	< .0001***
AnnotationisSecurityRelated	-1:0	0.97	2	9.95	0.0069**
	1:0	2.69			
AuthorExperience		0.87	1	4.89	0.0271*
Commit		1.01	1	3.06	0.0801
SecurityDefectType	Buffer Overflow:Race Condition	0.68	10	13.98	0.1742
	Crash:Race Condition	0.86			
	Deadlock:Race Condition	0.92			
	DoS:Race Condition	0.87			
	Encryption:Race Condition	1.18			
	Improper Access:Race Condition	0.68			
	Integer Overflow:Race Condition	0.86			
	Resource Leak:Race Condition	1.41			
	Use After Free:Race Condition	0.37			
	Extra:Race Condition	1.43			
Complexity		1.03	1	1.74	0.1868
AnnotationRatio		0.56	1	1.55	0.2126
AnnotationHasCode		1.11	1	1.09	0.2973

Statistical significance of explanatory power according to Wald  $\chi^2$  test:

\* $p < 0.05$ ; \*\* $p < 0.01$ ; \*\*\* $p < 0.001$

in the dataset. Nonetheless, through a log-likelihood ratio test using the *anova* function of the *ordinal* R package [17], we found that this regression model exhibits significant explanatory power ( $L_R\ chi^2 = 129.57$ ,  $P_r(> chi^2) < 0.0001$ ) over a null model. Given that our goal is to construct an inferential regression model rather than a predictive one, this model with a relatively lower  $R^2$  can still provide valid insights into the relationship between explanatory variables and response variables [5].

**4.3.2 Explanatory Power of Each Factor.** The Wald statistic is employed to evaluate the explanatory power exerted by each factor on the response variable. A larger Wald  $\chi^2$  with a smaller *p-value* indicates that the variable has more significant predictive power for the regression model. We calculated the  $\chi^2$  and *p-value* of each factor and adopted the commonly used significance level of 0.05 to evaluate their impact. As shown in Table 7, ‘Token’ yields the highest predictive power on the fitted model, with the highest  $\chi^2$  (48.07) and the lowest *p-value* (<.0001). Trailing closely behind are ‘Community’ and ‘FileType’ factors, with correspondingly significant contributions ( $\chi^2=19.03$ ,  $p<.0001$ ) and ( $\chi^2=18.28$ ,  $p<.0001$ ). Adding to this, ‘AnnotationisSecurityRelated’ and ‘AuthorExperience’ displayed less influence but remained noteworthy, with  $\chi^2$  values of 9.95 and 4.89, respectively. However, the remaining factors, i.e., ‘Commit’, ‘SecurityDefectType’, ‘Complexity’, ‘AnnotationRatio’, and ‘AnnotationHasCode’ are deemed not to largely influence the performance of LLMs in detecting security defects in code reviews, as their *p-values* fall below the 0.05 threshold. Notably, ‘SecurityDefectType’ has a relatively high  $\chi^2$  value (13.98), which could be attributed to the fact that this factor is a nominal variable with multiple categories of security defects, thus automatically allocated a high number of degrees of freedom. As a result, even if this factor is not significant, its relative contribution ( $\chi^2$ ) remains elevated.

**4.3.3 Relationships between Factors and Responses.** The odds ratio (OR) is commonly used to measure the relationships between explanatory and response variables [55, 62, 68]. In this study, the OR represents the odds of LLM generating a better response in security code review versus not generating a better response. For categorical explanatory variables, the OR compares each category to a reference category. For interval variables, the OR reflects the effect of a one-unit increase in the explanatory variable. We utilized the *summary* function of the *rms* package [31] to calculate the OR for each factor (see Table 7). The OR of ‘FileType’ exceeds 1, being 2.82, while the OR of ‘Token’ and ‘AuthorExperience’ are both less than 1, at 0.92 and 0.87, respectively. This suggests that the LLM is more likely to detect security defects in code files that have fewer tokens, contain functional logic, and were written by developers with less involvement in the project. Furthermore, the OR of ‘Community’, which is greater than 1, indicates that compared to OpenStack, the LLM performs better in security code reviews from Qt. In addition, for ‘AnnotationisSecurityRelated’, when annotations in code files contain descriptions related to the target security defect (labelled as 1), the OR is 2.69, indicating that with related security annotations, the LLM is more likely to identify the target security defect. Conversely, when annotations contain descriptions related to non-target security defects (labelled as -1), the OR is 0.97. This suggests that the LLM may be disrupted by annotations and less likely to detect the target security defect.

## 5 DISCUSSION

**A multi-layered review strategy is suggested to apply LLMs to security code review.** As the results of RQ1 indicate, the existing general-purpose LLMs still fall significantly short of reaching the effectiveness of manual security code review. However, LLMs can serve as auxiliary tools to assist reviewers. The findings of RQ3 suggest that LLMs are more adept at identifying relatively simple security defects in shorter, logic-centric code written by developers with less project experience. We recommend a multi-layered review strategy to fully leverage the capability of LLMs - reviewers can first use LLMs to conduct an initial automatic review of those short and simple commits for security defects, then review both the submitted code and the LLM’s detection results to improve efficiency. This would allow reviewers to concentrate on an in-depth review of those complex commits, including lengthy code written by more experienced developers, thereby reducing the risk of missing sophisticated security defects. Additionally, where LLMs currently show limitations, these complex commits should be more heavily incorporated into datasets used for fine-tuning LLMs to improve their effectiveness in security code review.

**Providing a task scope is crucial for enhancing the performance of LLMs.** Our findings indicate that incorporating a CWE list into prompts significantly improves the performance of various LLMs in security defect detection, which aligns with the results of Steenhoek *et al.* [65]. Providing a CWE list can effectively narrow the problem domain, guiding LLMs to focus on key areas, thereby increasing detection accuracy. However, in real-world security reviews, although with the experience and domain knowledge, The reviewers lack prior knowledge of all potential types of security defects that may be present in the code. Therefore, constructing a comprehensive range of CWEs covering commonly encountered security defect types is necessary. It should also be noted that an overly exhaustive scope can diminish the accuracy of LLM responses. In Bakhshandeh *et al.*’s work [7], providing a list of all 75 CWEs in their prompts resulted in a decrease in the accuracy of security defect detection and location by the LLMs. This drop in accuracy may be attributed to the overly lengthy and detailed list of CWEs provided, which interferes with the LLM’s capability to focus effectively on the target defect, yielding adverse effects.

**Using LLMs in security code review requires developers to standardize their code commenting practices.** According to RQ3, annotations containing information related to non-target security defects, which were often already resolved by developers, negatively impact the LLM’s

capability in security code review. This suggests that LLMs could mistakenly interpret these resolved security defects as existing defects. Moreover, although the *p-value* of ‘AnnotationRatio’ (0.21) is more than 0.05, we still consider its OR value (0.56) to be somewhat informative. The more annotations in the code, the more challenging it is for LLMs to detect security defects, reflecting the disruptive effect and poor quality of annotations written by developers. Hence, we recommend that developers write code comments in a standardized way to clarify the purpose of each comment, thus better-supporting LLMs in code review tasks.

***With the development of LLMs, while expanding its context window, it is also necessary to ensure that LLMs can accurately capture details from long inputs.*** The emergence of LLMs that support large amounts of tokens has alleviated the challenge of inputting code context information in applying LLMs to code-related tasks. However, as evidenced by the results of RQ3, when dealing with code that contains more tokens, it is more difficult for LLMs to identify target security defects. This is because many security defects are introduced by small code snippets, and longer inputs may distract LLMs from overlooking certain details, thereby failing to detect security defects introduced by small code snippets. Additionally, according to RQ1, despite *GPT-4 Turbo* has a context of 128K tokens, its performance in detecting security defects is noticeably inferior to that of smaller-context LLMs like *GPT-4*, *GPT-3.5*, and *Llama 2*. Therefore, how to ensure the accuracy of security defect detection when expanding the LLMs’ context window is a promising direction.

***A key challenge in applying LLMs to security code review is ensuring comprehensive detection of security defects while maintaining consistent results.*** The results from RQ1.4 indicate that LLMs with stronger capabilities to detect security defects tend to exhibit poorer consistency in their responses. This may be because the outputs generated by LLMs with higher randomness are more diverse, making them more likely to include content related to target security defects compared to those generated by LLMs with lower randomness. Both detection capability and the stability of detection results are crucial for the practical application of LLMs in security code review. However, current studies related to LLMs in vulnerability detection have largely overlooked the importance of the stability of results. Therefore, for researchers of LLMs, exploring approaches to balance detection capabilities with the randomness inherent in LLMs represents a valuable direction for future improvements.

## 6 THREATS TO VALIDITY

**Internal Validity.** One threat could be **data leakage**. The dataset used in our study covers two large-scale open-source projects, which may overlap with the data used for training LLMs. Since the training data cutoff dates for *GPT-3.5* and *GPT-4* are not disclosed and the dataset we used [74] predates the release of some of the LLMs we selected, we cannot avoid the overlap between the dataset we used and the training data of LLMs by filtering the data for a specific period of time. Moreover, according to the findings by Cao *et al.* [11], contaminated data does not always affect the results, and the model sometimes even performs better on data after the models’ cutoff date. Another threat is **the design of prompts**. We design our prompt according to best practices to evoke the best model performance. We selected words that are as unbiased, not suggestive, and unambiguous as possible and iteratively revised the prompts to avoid individual words from impacting the responses of LLMs during our preliminary exploration. In addition, we only included code from the single vulnerable file rather than the whole patchset in the prompt, which may affect the performance of LLMs due to **the lack of contextual information**. To specifically analyze its impact, we designed prompt  $P_{cot}$ , in which we instructed LLMs to generate code context according to the commit message before detecting security defects. Lastly, LLMs may generate varied responses for the same prompts due to their **inherent randomness**. To mitigate this threat, we conducted three repetitive experiments for RQ1. For RQ2, we calculated the average frequency

of each problem across these three experiments. In RQ3, we combined the responses from the three experiments to fit the model.

**Conclusion Validity.** In LLM selection, we tried to encompass a diverse range of popular LLMs (i.e., open-source and closed-source, large-scale and small-scale). In dataset collection, we used a dataset of code reviews collected from real-world open-source projects that are widely used in code review-related studies. We acknowledge that the dataset may not fully represent industrial projects, and projects from proprietary software development are needed to increase the generalizability of the results.

**External Validity.** With the active development of LLMs, both *GPT-3.5* and *GPT-4* adopted in our work are continually updated without public disclosure. These models are also not made publicly available via API, thereby hindering the reproducibility of this study. Therefore, we explicitly indicate the versions of these two models used in the experiments and provide all the responses generated by these models [75]. In addition, the data labeling and extraction processes in this study were manually conducted, which may introduce subjective judgements. Thus, we adopted an open coding and constant comparative approach, conducting pilot experiments before all formal manual tasks to ensure a consensus among multiple authors.

## 7 CONCLUSIONS AND FUTURE WORK

In our study, we conducted a comprehensive analysis of the practicality of current popular LLMs in the context of security code review. We investigated the capabilities of 6 pre-trained LLMs in detecting security defects similar to those identified by human reviewers during real-world code reviews. For the best-performing LLM, we conducted a linguistic analysis of the quality problems in the responses generated by this LLM. We performed a regression analysis to explore factors influencing its performance.

Our main findings are: (1) In security code review tasks, LLMs demonstrate better capabilities compared to state-of-the-art static analysis tools. However, their application remains limited. (2) *GPT-4* performs best among all LLMs when provided with a CWE list for reference. (3) During code review, *GPT-4* is more effective in identifying security defects in code files with fewer tokens, containing functional logic, and written by developers with less involvement in the project. (4) The most prevalent quality problems in the responses generated by *GPT-4* are verbosity and non-compliance with task requirements.

We plan to extend this work in the next step by incorporating code reviews from proprietary software projects into our dataset and attempting enhancement strategies such as few-shot learning and fine-tuning. In addition to applying LLMs to help reviewers detect security defects during code review, we also plan to analyze the ability of LLMs to fix security defects in code so that developers can comprehensively evaluate the potential of LLMs in the task of secure code review.

## ACKNOWLEDGMENTS

This work has been partially supported by the National Natural Science Foundation of China (NSFC) with Grant No. 62172311.

## REFERENCES

- [1] 2024. *Bandit*. <https://github.com/PyCQA/bandit/>.
- [2] 2024. *lizard*. <https://github.com/terryyin/lizard/>.
- [3] 2024. *Semgrep*. <https://semgrep.dev/>.
- [4] 2024. *tiktoken*. <https://github.com/openai/tiktoken/>.
- [5] Paul Allison. 2014. Prediction vs. Causation in Regression Analysis. *Statistical Horizons* 703 (2014).
- [6] Max Baak, Rose Koopman, Hella Snoek, and Sander Kloos. 2020. A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics. *Computational Statistics & Data Analysis* 152 (2020), 107043.

- [7] Atieh Bakhshandeh, Abdalsamad Keramatfar, Amir Norouzi, and Mohammad Mahdi Chekidehkhou. 2023. Using chatgpt as a static application security testing tool. *arXiv preprint arXiv:2308.14434* (2023).
- [8] Ali Borji. 2023. A categorical archive of chatgpt failures. *arXiv preprint arXiv:2302.03494* (2023).
- [9] Kenneth P Burnham and David R Anderson. 2004. Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research* 33, 2 (2004), 261–304.
- [10] John L Campbell, Charles Quincy, Jordan Osserman, and Ove K Pedersen. 2013. Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research* 42, 3 (2013), 294–320.
- [11] Jialun Cao, Wuqi Zhang, and Shing-Chi Cheung. 2024. Concerned with Data Contamination? Assessing Countermeasures in Code Language Model. *arXiv preprint arXiv:2403.16898* (2024).
- [12] Huseyin Cavusoglu, Birendra Mishra, and Srinivasan Raghunathan. 2004. The effect of internet security breach announcements on market value: Capital market reactions for breached firms and internet security developers. *International Journal of Electronic Commerce* 9, 1 (2004), 70–104.
- [13] Saikat Chakraborty, Rahul Krishna, Yangruibo Ding, and Baishakhi Ray. 2021. Deep learning based vulnerability detection: Are we there yet? *IEEE Transactions on Software Engineering* 48, 9 (2021), 3280–3296.
- [14] Chia-Hsuan Chang, Mary M Lucas, Yeawon Lee, Christopher C Yang, and Grace Lu-Yao. 2024. Beyond Self-Consistency: Ensemble Reasoning Boosts Consistency and Accuracy of LLMs in Cancer Staging. In *Proceedings of the 22nd International Conference on Artificial Intelligence in Medicine (AIME)*. Springer, 677–687.
- [15] Wachiraphan Charoenwet, Patanamon Thongtanunam, Van-Thuan Pham, and Christoph Treude. [n. d.]. An Empirical Study of Static Analysis Tools for Secure Code Review. In *Proceedings of the 42nd International Conference on Software Engineering (ICSE)*. ACM.
- [16] Chong Chen, Jianzhong Su, Jiachi Chen, Yanlin Wang, Tingting Bi, Yanli Wang, Xingwei Lin, Ting Chen, and Zibin Zheng. 2023. When ChatGPT meets smart contract vulnerability detection: How far are we? *arXiv preprint arXiv:2309.05520* (2023).
- [17] Rune Haubo Bojesen Christensen and Maintainer Rune Haubo Bojesen Christensen. 2015. Package ‘ordinal’. *Stand* 19, 2016 (2015).
- [18] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46.
- [19] Cppcheck Solutions AB 2024. *Cppcheck*. Cppcheck Solutions AB. <https://www.cppcheck.com/>.
- [20] Anne Edmundson, Brian Holtkamp, Emanuel Rivera, Matthew Finifter, Adrian Mettler, and David Wagner. 2013. An empirical study on the effectiveness of security code review. In *Proceedings of the 5th International Symposium on Engineering Secure Software and Systems (ESSoS)*. Springer, 197–212.
- [21] Michael Fu and Chakkrit Tantithamthavorn. 2022. LineVul: A transformer-based line-level vulnerability prediction. In *Proceedings of the 19th International Conference on Mining Software Repositories (MSR)*. ACM, 608–620.
- [22] Sadeq Gholami and Zeineb Amri. 2021. Automated secure code review for web-applications.
- [23] GitHub 2021. *CodeQL*. GitHub. <https://codeql.github.com/>.
- [24] GitLab 2022. *GitLab: Mapping the DevSecOps Landscape - 2022 Survey*. GitLab. <https://about.gitlab.com/developer-survey>.
- [25] Barney G Glaser. 1965. The constant comparative method of qualitative analysis. *Social Problems* 12, 4 (1965), 436–445.
- [26] Qi Guo, Junming Cao, Xiaofei Xie, Shangqing Liu, Xiaohong Li, Bihuan Chen, and Xin Peng. 2024. Exploring the potential of ChatGPT in automated code refinement: An empirical study. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering (ICSE)*. ACM, 1–13.
- [27] Istvan Haller, Asia Slowinska, Matthias Neugschwandtner, and Herbert Bos. 2013. Dowsing for {Overflows}: A Guided Fuzzer to Find Buffer Boundary Violations. In *Proceedings of the 22nd USENIX Security Symposium (USENIX Security)*. USENIX, 49–64.
- [28] Kazuki Hamasaki, Raula Gaikovina Kula, Norihiro Yoshida, AE Camargo Cruz, Kenji Fujiwara, and Hajimu Iida. 2013. Who does what during a code review? datasets of oss peer review repositories. In *Proceedings of the 10th Working Conference on Mining Software Repositories (MSR)*. IEEE, 49–52.
- [29] Xiaofeng Han, Amjad Tahir, Peng Liang, Steve Counsell, Kelly Blincoe, Bing Li, and Yajing Luo. 2022. Code smells detection via modern code review: A study of the OpenStack and Qt communities. *Empirical Software Engineering* 27, 6 (2022), 127.
- [30] Frank E Harrell. 2001. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Vol. 608. Springer.
- [31] Frank E Harrell Jr, Maintainer Frank E Harrell Jr, and Depends Hmisc. 2017. Package ‘rms’. *Vanderbilt University* 229, Q8 (2017).
- [32] Dennis E Hinkle, William Wiersma, and Stephen G Jurs. 2003. *Applied Statistics for the Behavioral Sciences*. Vol. 663. Houghton Mifflin Boston.

- [33] Toshiki Hirao, Shane McIntosh, Akinori Ihara, and Kenichi Matsumoto. 2020. Code reviews with divergent review scores: An empirical study of the OpenStack and Qt communities. *IEEE Transactions on Software Engineering* 48, 1 (2020), 69–81.
- [34] Nan Jiang, Kevin Liu, Thibaud Lutellier, and Lin Tan. 2023. Impact of Code Language Models on Automated Program Repair. *arXiv preprint arXiv:2302.05020* (2023).
- [35] Brittany Johnson, Yoonki Song, Emerson Murphy-Hill, and Robert Bowdidge. 2013. Why don't software developers use static analysis tools to find bugs?. In *Proceedings of the 35th International Conference on Software Engineering (ICSE)*. IEEE, 672–681.
- [36] Samia Kabir, David N Udo-Imeh, Bonan Kou, and Tianyi Zhang. 2024. Is Stack Overflow Obsolete? An Empirical Study of the Characteristics of ChatGPT Answers to Stack Overflow Questions. In *Proceedings of the 42nd CHI Conference on Human Factors in Computing Systems (CHI)*. ACM, Article No.: 935.
- [37] Dinesh Kalla and Sivaraju Kuraku. 2023. Advantages, disadvantages and risks associated with ChatGPT and AI on cybersecurity. *Journal of Emerging Technologies and Innovative Research* 10, 10 (2023).
- [38] Sungmin Kang, Juyeon Yoon, and Shin Yoo. 2023. Large language models are few-shot testers: Exploring llm-based general bug reproduction. In *Proceedings of the 45th International Conference on Software Engineering (ICSE)*. IEEE, 2312–2323.
- [39] Klaus Krippendorff. 2004. Measuring the reliability of qualitative text analysis data. *Quality and Quantity* 38 (2004), 787–800.
- [40] Pierre Legendre and Louis Legendre. 2012. *Numerical Ecology*. Elsevier.
- [41] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Comput. Surveys* 55, 9 (2023), 1–35.
- [42] Guilong Lu, Xiaolin Ju, Xiang Chen, Wenlong Pei, and Zhilong Cai. 2024. GRACE: Empowering LLM-based software vulnerability detection with graph structure and in-context learning. *Journal of Systems and Software* 212 (2024), 112031.
- [43] Sonal Mahajan, Negarsadat Abolhassani, and Mukul R Prasad. 2020. Recommending stack overflow posts for fixing runtime exceptions using failure scenario matching. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*. ACM, 1052–1064.
- [44] Thomas J McCabe. 1976. A complexity measure. *IEEE Transactions on Software Engineering* 4 (1976), 308–320.
- [45] Shane McIntosh, Yasutaka Kamei, Bram Adams, and Ahmed E Hassan. 2016. An empirical study of the impact of modern code review practices on software quality. *Empirical Software Engineering* 21 (2016), 2146–2189.
- [46] Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. 2023. Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text. *arXiv preprint arXiv:2301.13852* (2023).
- [47] Nico JD Nagelkerke et al. 1991. A note on a general definition of the coefficient of determination. *Biometrika* 78, 3 (1991), 691–692.
- [48] Chao Ni, Xin Yin, Kaiwen Yang, Dehai Zhao, Zhenchang Xing, and Xin Xia. 2023. Distinguishing Look-Alike Innocent and Vulnerable Code by Subtle Semantic Representation Learning and Explanation. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*. ACM, 1611–1622.
- [49] Yu Nong, Mohammed Aldeen, Long Cheng, Hongxin Hu, Feng Chen, and Haipeng Cai. 2024. Chain-of-Thought Prompting of Large Language Models for Discovering and Fixing Software Vulnerabilities. *arXiv preprint arXiv:2402.17230* (2024).
- [50] Yu Nong, Haipeng Cai, Pengfei Ye, Li Li, and Feng Chen. 2021. Evaluating and comparing memory error vulnerability detectors. *Information and Software Technology* 137 (2021), 106614.
- [51] Yu Nong, Richard Fang, Guangbei Yi, Kunsong Zhao, Xiapu Luo, Feng Chen, and Haipeng Cai. 2024. VGX: Large-Scale Sample Generation for Boosting Learning-Based Software Vulnerability Analyses. In *Proceedings of the 46th International Conference on Software Engineering (ICSE)*. ACM, 1–13.
- [52] Yu Nong, Yuzhe Ou, Michael Pradel, Feng Chen, and Haipeng Cai. 2022. Generating realistic vulnerabilities via neural code editing: an empirical study. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*. ACM, 1097–1109.
- [53] OpenAI 2023. *gpt-best-practices*. OpenAI. <https://platform.openai.com/docs/guides/prompt-engineering/>.
- [54] Shengyi Pan, Lingfeng Bao, Xin Xia, David Lo, and Shanping Li. 2023. Fine-grained commit-level vulnerability type prediction by CWE tree structure. In *Proceedings of the 45th International Conference on Software Engineering (ICSE)*. IEEE, 957–969.
- [55] Rajshakhar Paul, Asif Kamal Turzo, and Amiangshu Bosu. 2021. Why security defects go unnoticed during code reviews? a case-control study of the chromium os project. In *Proceedings of the 43rd International Conference on Software Engineering (ICSE)*. IEEE, 1373–1385.

- [56] Strategic Planning. 2002. The economic impacts of inadequate infrastructure for software testing. *National Institute of Standards and Technology* 1, 2002 (2002).
- [57] Daniel Powers and Yu Xie. 2008. *Statistical Methods for Categorical Data Analysis*. Emerald Group Publishing.
- [58] Moumita Das Purba, Arpita Ghosh, Benjamin J Radford, and Bill Chu. 2023. Software vulnerability detection using large language models. In *Proceedings of the 34th International Symposium on Software Reliability Engineering Workshops (ISSREW)*. IEEE, 112–119.
- [59] Zilong Ren, Xiaolin Ju, Xiang Chen, and Hao Shen. 2024. ProRLearn: boosting prompt tuning-based vulnerability detection by reinforcement learning. *Automated Software Engineering* 31, 2 (2024), Article No.: 38.
- [60] Alejandro Russo and Andrei Sabelfeld. 2010. Dynamic vs. static flow-sensitive security analysis. In *Proceedings of the 23rd IEEE Computer Security Foundations Symposium (CSF)*. IEEE, 186–199.
- [61] Max Schäfer, Sarah Nadi, Aryaz Eghbali, and Frank Tip. 2023. An empirical evaluation of using large language models for automated unit test generation. *IEEE Transactions on Software Engineering* (2023).
- [62] Maninder Singh Setia. 2016. Methodology series module 2: Case-control studies. *Indian Journal of Dermatology* 61, 2 (2016), 146–151.
- [63] SonarSource SA 2024. *SonarQube*. SonarSource SA. <https://www.sonarqube.org/>.
- [64] Davide Spadini, Maurício Aniche, Margaret-Anne Storey, Magiel Bruntink, and Alberto Bacchelli. 2018. When testing meets code review: Why and how developers review tests. In *Proceedings of the 40th International Conference on Software Engineering (ICSE)*. IEEE, 677–687.
- [65] Benjamin Steenhoek, Md Mahbubur Rahman, Monoshi Kumar Roy, Mirza Sanjida Alam, Earl T Barr, and Wei Le. 2024. A Comprehensive Study of the Capabilities of Large Language Models for Vulnerability Detection. *arXiv preprint arXiv:2403.17218* (2024).
- [66] Li Sui, Jens Dietrich, Amjad Tahir, and George Fournous. 2020. On the recall of static call graph construction in practice. In *Proceedings of the 33rd International Symposium on Software Testing and Analysis (ISSTA)*. ACM.
- [67] Rahul Telang and Sunil Wattal. 2007. An empirical analysis of the impact of software vulnerability announcements on firm stock price. *IEEE Transactions on Software Engineering* 33, 8 (2007), 544–557.
- [68] Patanamon Thongtanunam, Shane McIntosh, Ahmed E Hassan, and Hajimu Iida. 2017. Review participation in modern code review: An empirical study of the Android, Qt, and OpenStack projects. *Empirical Software Engineering* 22 (2017), 768–817.
- [69] Asif Kamal Turzo. 2022. Towards improving code review effectiveness through task automation. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. ACM, 1–5.
- [70] Charles Weir, Awais Rashid, and James Noble. 2017. *I'd Like to Have an Argument, Please: Using Dialectic for Effective App Security*. Internet Society.
- [71] Danning Xie, Byungwoo Yoo, Nan Jiang, Mijung Kim, Lin Tan, Xiangyu Zhang, and Judy S Lee. 2023. Impact of large language models on generating software specifications. *arXiv preprint arXiv:2306.03324* (2023).
- [72] Yanjing Yang, Xin Zhou, Runfeng Mao, Jinwei Xu, Lanxin Yang, Yu Zhangm, Haifeng Shen, and He Zhang. 2024. DLAP: A Deep Learning Augmented Large Language Model Prompting Framework for Software Vulnerability Detection. *arXiv preprint arXiv:2405.01202* (2024).
- [73] Xin Yin and Chao Ni. 2024. Multitask-based Evaluation of Open-Source LLM on Software Vulnerability. *arXiv preprint arXiv:2404.02056* (2024).
- [74] Jiaxin Yu, Liming Fu, Peng Liang, Amjad Tahir, and Mojtaba Shahin. 2023. Security Defect Detection via Code Review: A Study of the OpenStack and Qt Communities. In *Proceedings of the 17th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*. IEEE, 1–12.
- [75] Jiaxin Yu, Peng Liang, Yujia Fu, Amjad Tahir, Mojtaba Shahin, Chong Wang, and Yangxiao Cai. 2024. *Replication Package for the Paper: "An Insight into Security Code Review with LLMs: Capabilities, Obstacles and Influential Factors"*. <https://zenodo.org/doi/10.5281/zenodo.13771089>.
- [76] Chenyuan Zhang, Hao Liu, Jiutian Zeng, Kejing Yang, Yuhong Li, and Hui Li. 2023. Prompt-enhanced software vulnerability detection using chatgpt. *arXiv preprint arXiv:2308.12697* (2023).
- [77] Xin Zhou, Sicong Cao, Xiaobing Sun, and David Lo. 2024. Large Language Model for Vulnerability Detection and Repair: Literature Review and Roadmap. *arXiv preprint arXiv:2404.02525* (2024).
- [78] Xin Zhou, Kisub Kim, Bowen Xu, Jiakun Liu, DongGyun Han, and David Lo. 2023. The devil is in the tails: How long-tailed code distributions impact large language models. In *Proceedings of the 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 40–52.
- [79] Xin Zhou, Ting Zhang, and David Lo. 2024. Large Language Model for Vulnerability Detection: Emerging Results and Future Directions. *arXiv preprint arXiv:2401.15468* (2024).