

# Actionable Warning Is Not Enough: Recommending Valid Actionable Warnings with Weak Supervision

Zhipeng Xue

The State Key Laboratory of  
Blockchain and Data Security,  
Zhejiang University  
Hangzhou, China  
zhipengxue@zju.edu.cn

Zhipeng Gao

Shanghai Institute for Advanced  
Study, Zhejiang University  
Shanghai, China  
zhipeng.gao@zju.edu.cn

Tongtong Xu

Huawei  
Hangzhou, China  
xutongtong9@huawei.com

Xing Hu

The State Key Laboratory of  
Blockchain and Data Security,  
Zhejiang University  
Hangzhou, China  
xinghu@zju.edu.cn

Xin Xia\*

The State Key Laboratory of  
Blockchain and Data Security,  
Zhejiang University  
Hangzhou, China  
xin.xia@acm.org

Shanping Li

The State Key Laboratory of  
Blockchain and Data Security,  
Zhejiang University  
Hangzhou, China  
shan@zju.edu.cn

## Abstract

The use of static analysis tools has gained increasing popularity among developers in the last few years. However, the widespread adoption of static analysis tools is hindered by their high false alarm rates (up to 90%). Previous studies have introduced the concept of actionable warnings and built machine-learning method to distinguish actionable warnings from false alarms. However, according to our empirical observation, the current assumption used for actionable warning(s) collection is rather shaky and inaccurate, leading to a large number of invalid actionable warnings. To address this problem, in this study, we build the first large actionable warning dataset by mining 68,274 reversions from Top-500 GitHub C repositories, we then take one step further by assigning each actionable warning a weak label regarding its likelihood of being a real bug. Following that, we propose a two-stage framework called ACWRECOMMENDER to automatically recommend the actionable warnings with high probability to be real bugs (AWHB). Our approach warms up the pre-trained model UniXcoder by identifying actionable warnings task (coarse-grained detection stage) and rerank AWHB to the top by weakly supervised learning (fine-grained reranking stage). Experimental results show that our proposed model outperforms several baselines by a large margin in terms of nDCG and MRR for AWHB recommendation. Moreover, we ran our tool on 6 randomly selected projects and manually checked the top-ranked warnings from 2,197 reported warnings, we reported top-10 recommended warnings to developers, 27 of them were already confirmed by developers as real bugs. Developers can quickly find real bugs among the massive amount of reported warnings, which verifies the practical usage of our tool.

\*Corresponding Author

## CCS Concepts

• Software and its engineering → Software maintenance tools.

## Keywords

Actionable warning recommendation, Static analysis, Weak supervision, Data mining

## ACM Reference Format:

Zhipeng Xue, Zhipeng Gao, Tongtong Xu, Xing Hu, Xin Xia, and Shanping Li. 2026. Actionable Warning Is Not Enough: Recommending Valid Actionable Warnings with Weak Supervision. In *2026 IEEE/ACM 48th International Conference on Software Engineering (ICSE '26)*, April 12–18, 2026, Rio de Janeiro, Brazil. ACM, New York, NY, USA, 13 pages. <https://doi.org/10.1145/3744916.3773133>

## 1 Introduction

The static analysis tools have been utilized for a considerable period of time [17, 70, 71]. These tools are widely used by software developers and companies to detect potential bugs and report warnings in recent years [20, 28, 35, 59, 62, 76]. For example, Facebook has developed *Infer* [19, 23], a static code analysis tool for checking generic bug patterns (e.g, null pointer exceptions, memory leaks, and race conditions) in their Android and iOS apps (including the main Facebook, Whatsapp, Instagram app, and many others). Due to the lightweight analysis and low computational cost, these static analysis tools have gained popularity among developers.

In spite of the success of static analysis tools, their widespread adoption in various software projects is hindered by two major problems. First, a high false alarm rate is observed. The warnings generated by static analysis tools have a very high false-positive rate (e.g., range up to 90%) [32]. In other words, only a small fraction of the reported warnings are real bugs and/or need to be acted on. Second, developers often experience information overloading while using static analysis tools. That is, even if the real bugs are reported, developers can become flooded with too many generated warnings. They may fail to filter out the false alarms and tend to get lost in massive amounts of irrelevant information.



This work is licensed under a Creative Commons Attribution 4.0 International License. *ICSE '26, Rio de Janeiro, Brazil*

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2025-3/26/04

<https://doi.org/10.1145/3744916.3773133>

To fill these gaps and help developers to better make use of static analysis tools, previous researchers have introduced the concept of **actionable warnings**, namely the warnings that need to be acted on by developers [14, 33]. Particularly, warnings that are acted on by developers are called actionable warnings. In practice, a warning is identified as “actionable” if and only if a warning is present in a revision and disappears in a subsequent revision. More formally, given a sequence of commits  $C = c_1, c_2, \dots, c_n$ , let  $W(i)$  denote the set of warnings reported by a static analysis tool on commit  $c_i$  ( $1 \leq i \leq n$ ). A warning  $w \in W(i)$  is defined as *actionable* if it disappears in any subsequent commit between  $c_i$  and  $c_n$ . The software engineering researchers have proposed different techniques for the task of actionable warning identification (AWI) [40, 67, 77].

Nonetheless, two significant limitations exist within these studies: Firstly, according to our in-depth analysis, **the current data collection process of obtaining actionable warnings is inaccurate and unreliable**. In other words, a considerable proportion of actionable warnings are invalid and not real bugs [16, 42]. The use of such uncurated datasets can be a major threat when models are trained with such data. Unfortunately, many automatic approaches proposed thus far [29, 30, 44] were trained and evaluated on these unreliable actionable warnings, without paying attention to the overall quality of the dataset and the correctness of warning labels. Secondly, regarding actionable warning identification, the existing approaches commonly use machine learning (ML) techniques for this task. These approaches train ML classifiers by using a set of hand-crafted features to determine whether warnings are actionable or not. These features are manually designed by human experts and critically affect the model’s performance. The warning features determination is **laborious and error-prone and heavily relies on the domain knowledge of the experts**. Moreover, there is no guarantee that these hand-crafted features are well-designed (different studies may design redundant features [30, 31, 38, 56]) and helpful (features have little contribution to the warning labels [67]). Therefore, the research problem we aim to tackle in this paper is: Given the reported warnings, can we accurately recommend the warnings that are more likely to be real bugs without resorting to manually pre-defined warning features?

In this study, we aim to rank **actionable warnings** produced by static analysis tools (*Infer* and *Flawfinder* in this study) and recommend the Actionable Warning with **High** probability to be real Bug (AWHB). To be more specific, we build the first large actionable warnings dataset under weak supervision and propose a two-stage framework, named **ACWRECOMMENDER (Actionable Warning Recommender)**, to automate this task. Our data collection involves two steps. First, we collect actionable warnings from the Top 500 popular C projects on GitHub, including 1,889 actionable warnings and 39,052 false alarms. Second, to identify real bugs from actionable warnings, we propose a weak supervision label strategy where each actionable warning is assigned a weak label using our semantic and structural matching rules. The label estimates the likelihood of the actionable warning being a real bug. To recommend AWHB, ACWRECOMMENDER is mainly divided into two stages, including a coarse-grained warm-up stage and a fine-grained reranking stage. In the coarse-grained detection stage, an actionable warnings dataset is used to warm-up a neural network.

In the fine-grained reranking stage, we further fine-tune the model to rank the AWHB to the top by weakly supervised learning [48, 49].

To evaluate the effectiveness of our weak supervision label strategy and ACWRECOMMENDER, we manually check the label accuracy and conduct extensive experiments on the AWHB recommendation task. By comparing with several baselines, the superiority of our proposed model is demonstrated. The experimental results show that: 1) According to our manual validation, 79.5% of AWHBs correctly identified real bugs, while only 19.5% of actionable warnings were true positives that revealed actual issues. 2) Regarding the AWHB recommendation task, our reranker performs better than its three baselines in terms of nDCG and MRR; 3) We further conducted an in-the-wild evaluation, and we report top-10 actionable warnings recommended by our tool to the developers of 6 popular Github projects, 27 of them have been confirmed by developers as real bugs, which further justifies the practical usage of our approach. In summary, this work makes the following contributions:

- 1) We build a large dataset for checking static analysis tools’ actionable warnings from popular GitHub C repositories, which contains 1,889 actionable warnings and 39,052 false alarms generated by *Infer* and *Flawfinder*. Then we propose a weak supervision label strategy to identify Actionable Warnings with **High** probability to be real Bugs (AWHB). Our manual verification reveals that 81% of AWHBs are real bugs.
- 2) We propose a novel two-stage model, ACWRECOMMENDER to automate the AWHB recommendation task. ACWRECOMMENDER fine-tuned the large pre-train model UniXcoder with text and code features, avoiding the laborious and error-prone process of designing warning features manually. This work builds on our preliminary demo tool [69] by extending it into a two-stage modeling framework and conducting a comprehensive evaluation to validate its effectiveness and practical value.
- 3) We extensively evaluate our ACWRECOMMENDER with several baseline methods. Evaluation results show that our model can significantly outperform the baselines in AWHB recommendation. Moreover, we have conducted an in-the-wild evaluation with 6 Github projects, we submitted the top-10 reported warnings to the developers, and 27 of them have been confirmed by developers as real bugs.
- 4) We have released our replication package [8], including the dataset and the source code of ACWRECOMMENDER. As the first attempt for the AWHB recommendation, we hope our research lays a good foundation for follow-up works and facilitates other researchers and practitioners to verify their ideas.

## 2 Motivation

In this section, we first show several motivating examples of actionable warnings from real-world software projects.

Due to the high false alarm rate of static analysis tools, software engineering researchers have proposed pipelines to collect actionable warnings [14, 33, 64], namely warnings that need to be acted on by developers. Upon investigation of these actionable warnings, we have observed that **actionable warnings collected by the current pipeline are inaccurate and may not necessarily represent real bugs**. This observation is also consistent with the latest empirical findings [36]. The underlying reason is the current pipeline regards warnings that exist in one revision and disappear in

**Table 1: The Examples of Actionable Warning-fix Commits**

<b>EX.1: Repo/Commit:</b> open62541/d52786e [6]	
367	<code>static UA_StatusCode</code> <code>UA_NodeMap_replaceNode(UA_Node *node) {</code> ...
373 -	<code>UA_NodeMapSlot *slot = findOccupiedSlot(</code> <code>ns, &amp;node-&gt;nodeId);</code>
+	<code>UA_NodeMapSlot *slot = findOccupiedSlot(</code> <code>ns, &amp;node-&gt;head.nodeId);</code>
374	<code>if(!slot){</code>
376	<code>return UA_STATUSCODE_BADNODEID;</code>
380	<code>UA_NodeMapEntry *oldEntry = slot-&gt;entry;</code>
<b>Commit Message:</b> refactor(server): Use a union to avoid casting of node classes	
<b>Warning Type:</b> Null Dereference	
<b>Warning Qualifier:</b> pointer 'slot' last assigned on line 373 could be null and is dereferenced at line 380	
<b>EX.2: Repo/Commit:</b> libevhttp/d13b72b [5]	
3651	<code>fd = socket(sa-&gt;sa_family, SOCK_STREAM, 0);</code> ...
+	<code>if (fd != -1)</code>
+	<code>evutil_closesocket(fd);</code>
3673	<code>return evhttp_accept_socket(http, fd, backlog);</code>
3674	<code>}</code>
<b>Commit Message:</b> FIX: Socket leakage on error #6. Cleanup open file descriptors when bind_sockaddr fails.	
<b>Warning Type:</b> Resource Leak	
<b>Warning Qualifier:</b> Resource acquired to 'fd' by call to 'socket()' at line 3651 is not released after line 3673.	

later revisions as actionable warnings. **However, this assumption is rather shaky because the disappearance of such warnings can be caused by a non-relevant fix/commit**, leading to the introduction of invalid actionable warnings. Table 1 demonstrates two actionable warnings, an invalid actionable warning (Ex.1) and a real bug warning addressed by developers (Ex.2).

Ex.1 shows an example of invalid actionable warnings. *Infer* reported a **Null Dereference** warning on line 373 because `slot` could be null and dereferenced. However, this warning is invalid because there is a null-checker for `slot`, and it will never be dereferenced if it is null. This warning disappeared in a non-relevant commit and was mistakenly extracted as an actionable warning. In contrast, Ex.2 shows a genuine bug warning (**Resource Leak** at line 3651) reported by *Infer*. To obtain a more accurate actionable warning dataset, it is necessary to determine whether the given actionable warning is invalid (i.e., Ex.1) or a genuine bug warning (i.e., Ex.2). Upon analyzing the “disappeared revision” of the invalid actionable warning and the genuine bug warning, we deduce that the genuine bug warning can be correctly identified by considering two key factors from the fix revision: semantic factor (e.g., commit message) and structural factor (e.g., code change context). The commit message conveys the semantic intention of the commit [26, 47, 65], while the code change conveys the syntactic structural information regarding the behavior of the commit [56, 66]. For example, the commit message of the genuine bug warning (Ex.2) validates the warning type (e.g., socket leakage), and the code change is a common code pattern for patching resource leakage, suggesting a high probability of this warning being related to a real bug. On the contrary, for the invalid actionable warning (Ex.1), the commit

message and code change of its disappeared revision have no correlation with the reported **Null Dereference** warning, which implies the likelihood of being a real bug of this warning is relatively low.

Guided by the motivating example, we propose a weak supervision labeling strategy including semantic (commit message) and syntactic (code change pattern) matching rules to distinguish AWHB from noisy actionable warnings, and use these labels to fine-tune the pre-trained model.

### 3 Approach

The overall framework is illustrated in Fig. 1. We present details of our approach design as follows.

#### 3.1 Actionable Warnings Collection

The goal of this step is to collect all the actionable warnings from a project and identify potential bug-fix revisions for each actionable warning. A commonly used method is to run static analysis tools on the compiled code of each revision and generate a list of warnings for each revision [45, 46]. Then for each warning, check whether it is closed in later revisions (label as actionable) or presented until the last revision (label as a false alarm). However, the previous method can be time-consuming and resource-intensive, especially for large projects with a long history of revisions.

In this study, we propose a graph-based binary search algorithm to efficiently identify actionable warnings and their corresponding bug-fix revisions, while skipping irrelevant commits during the search process. Our graph-based binary search algorithm includes two parts: graph linearization and linear binary search. Considering the real-world git histories often contain branches and merges and are organized in graph structures, the graph linearization is responsible for converting a git commit graph into a series of linear git histories. Following that, the linear binary search algorithm analyzes each linear git flow to identify actionable warnings and their corresponding bug-fix revisions within this linear git history.

The graph linearization algorithm takes a git commit history graph  $G$  as input, and outputs all possible linear commit histories within  $G$ . Each node in  $G$  represents a specific software reversion, while each edge represents a git commit connecting the previous reversion and its subsequent reversion. Fig. 2 illustrates an example of a Git commit graph where several reported warnings (marked as A, B, C, and D) were fixed and its linear commit histories extracted by our graph linearization algorithm. We define *start nodes* as nodes without parent nodes or nodes with more than one child node (such as node 1, node 2, node 4), and *end nodes* as nodes without child nodes (such as node 11). To perform the linearization, we first remove all the merge commits (e.g., edges marked with gray color) to avoid misleading results, such as falsely attributing a fix to the current branch when it actually originated from another branch via a merge. In this context, all the nodes before a merge commit become *end nodes* (such as node 8, node 9 and node 7), and all the nodes after a merge commit become *start nodes* (such as node 10). We then obtain all possible linear git commit histories by tracking from a *start node* (including node 1, node 2, node 4, and node 10) until reaching any *end nodes* (including node 6, node 7, node 8 and node 11). As illustrated in Fig. 2, the git commit graph

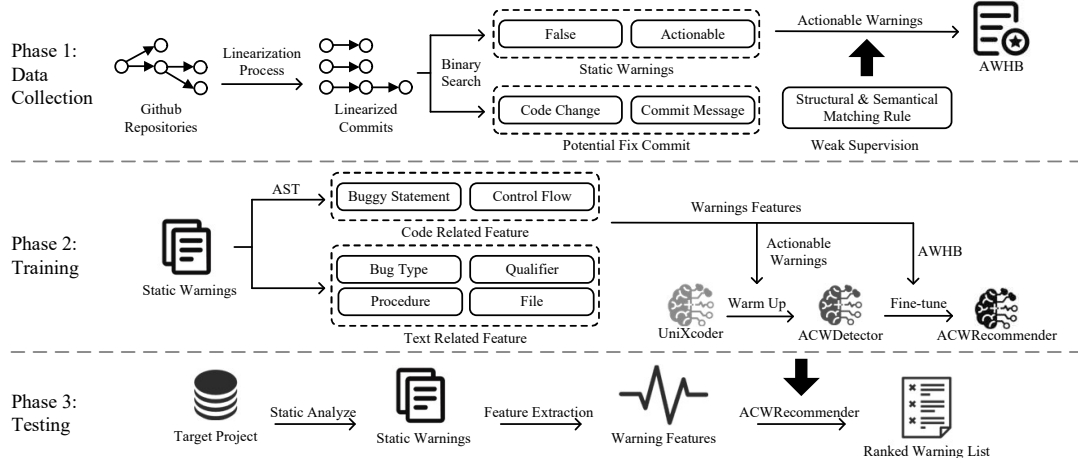


Figure 1: Overview of Our Approach

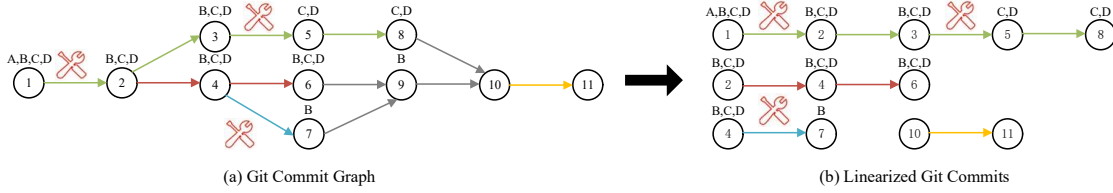


Figure 2: Example of the Git Commit Graph

will be converted to four linear git commit histories (marked with green/red/blue/yellow color) respectively.

After partitioning the git commit graph into a series of linear git histories, we utilize the linear binary search algorithm to detect actionable warnings and potential fix revisions on each linear git history. Specifically, we collect actionable warnings if and only if a warning present in a revision and disappears in its adjacent child revision. In this case, the child revision is regarded as a potential bug-fix revision. For each linear commit history, we set the left pointer point to the *start node* and the right pointer point to the *end node*. If the *start node* and *end node* have identical warning lists, then we hypothesize that the reversions between the *start node* and *end node* contain no actionable warning(s) and can be unchecked. Otherwise, there are actionable warnings between the *start node* and the *end node*. If the *start node* is the parent of the *end node*, we collect all warnings that vanished from the *start node* as actionable warnings. The process continues until all possible actionable warnings of this git linear history are identified.

Similar to previous research [67], if a warning has not been addressed for more than two years since its first occurrence, it is considered as a false warning.

### 3.2 Label under Weak Supervision

All of the previous research [14, 33, 64] focused on detecting actionable warnings only, however, as shown in Fig. 1(a), the current collected actionable warnings are inaccurate and unreliable. In this study, we aim to take one more important step by assigning actionable warnings with different probability scores under a weak supervision labeling strategy. The higher probability scores indicate

actionable warning(s) are more likely to be real bug(s) (referred to **AWHB** in this study), which should be inspected at the beginning. As mentioned in Section 2, the mismatch between warnings and their bug-fix reversions results in **invalid actionable warnings**. To gather more accurate actionable warnings, we estimate the “matching degree” of each actionable warning and its bug-fix commit in terms of two perspectives, i.e., semantical matching rule (using commit message) and structural matching rule (using code change context). Based on each matching rule, we assign a value to the actionable warning. In this study, we focus on four types of warnings (i.e., **Uninitialized Variable**, **Resource Leak**, **Null Dereference**, and **Dead Store**) reported by the default settings of *Infer* and the **Buffer Overflow** warning detected by *Flawfinder*. For each warning type, the first author reviewed 20 warning-fix commits and summarized the most common keywords and fix patterns for this warning type.

**3.2.1 Semantic Matching Rule.** The commit message of a reversion typically provides a semantic summary. If the bug-fix reversion’s commit message relates to the actionable warning semantically, the actionable warning is likely to be a real bug and being fixed by the bug-fix reversion. In this step, we extract the commit message from bug-fix reversion and estimate a semantic matching score  $CM(x)$  of the actionable warning  $x$  by following Equ.1:

$$CM(x) = \max \begin{pmatrix} 3, & \text{WTK in } m \\ 2, & \text{WCK in } m \\ 1, & \text{CK in } m \\ 0, & \text{No matching} \end{pmatrix} \quad (1)$$

**Table 2: Semantic Matching Rule**

Warning Type	Warning Qualifier Template	Warning Type Keyword	Warning Context Keyword	Common Keyword
Uninitialized Variable	The value read from <i>variable</i> was never initialized	initial, define, assign, declare	<i>variable</i>	fix, repair, bug, warning, solve, problem, handle, eliminate, address, issue, fail, error, exception, patch, crash
Null Dereference	<i>pointer</i> last assigned on line # could be null and is dereferenced at line #	dereference, null pointer, null check, NullPointerException	<i>pointer</i>	
Resource Leak	Resource acquired to <i>variable</i> by call to <i>function</i> at line # is not released after line #	resource, leak, release, cleanup, alloc, clear, close, free, destroy, terminate, end	<i>variable, function</i>	
Dead Store	The value written to <i>variable</i> is never used	dead store, unused, redundant	<i>variable</i>	
Buffer Overflow	without a limit specification, <i>function</i> permits buffer overflows	buffer, overflow	<i>function</i>	

Given an actionable warning  $x$ ,  $m$  refers to the commit message of the potential fix revision of  $x$ . We define three types of keywords for semantic matching, namely **Warning Type Keywords (WTK)**, **Warning Context Keywords (WCK)**, and **Common Keywords (CK)**. **WTK** includes the keyword related to the specific warning type. **WCK** includes the critical phrase in the warning qualifier template. **Common Keyword (CK)** includes the common word about warning fixing, as shown in Table 2. In particular, we perform our semantic matching as follows: (a) Specifically, if the commit message  $m$  contains words matching the warning type keywords (e.g., initial, resource, leak), we consider this actionable warning **highly likely** to be a real bug, and assign it a semantic score of 3. (b) The second column of Table 2 shows the warning qualifier template of each warning type, we extract the context keyword from the qualifier template as shown in the fourth column. If such warning context keywords are mentioned in the commit message  $m$ , we assume the actionable warning is **likely** to be a real bug and assign it a semantic score of 2. (c) We assign a semantic score of 1 to an actionable warning when commit message  $m$  only contains fix-related common keywords (e.g., fix, repair, bug), because the fixed bug **may not** match the corresponding actionable warning. (d) In the last, if the commit message  $m$  does not match any keyword in the matching rule, it is **unlikely** the actionable warning relates to a real bug, we assign it a semantic score of 0. Especially, if the commit message  $m$  matches several rules at the same time, we retain the maximum value of the semantic matching score.

**3.2.2 Structural Matching Rule.** Besides using semantic information, we also incorporate structural information (i.e., code change context) to determine if an actionable warning is fixed by the code change of the bug-fix reversion correspondingly. Similarly, for each actionable warning, we obtain the code changes from the corresponding potential bug-fix commit and assign a structural matching score based on our code change matching rules. A high score indicates the code change is likely to fix the actionable warning, and the actionable warning is likely to be a real bug. We first manually summarize the common fix pattern for each warning type, as listed in the second column in Table 3. If the bug-fix reversion's code change matches the fix pattern of a specific warning type, the code change is likely to fix the actionable warning and we assign a structural matching score  $CC(x)$  to this actionable warning  $x$  by

**Table 3: Structural Matching Rule**

Warning Type	Fix Pattern	Scope Pattern
Uninitialized Variable	assign value by assignment or reference	before warning
Null Dereference	add a null-check	before warning
Resource Leak	invoke resource free-related function	after warning
Dead Store	use the variable, remove assignment	after warning
Buffer Overflow	use the safe function, add a boundary check	before warning

following Equ. 2.

$$CC(x) = \max \left( \begin{cases} 2, & c \text{ match Fix Pattern} \\ 1, & c \text{ match Scope Pattern} \\ 0, & \text{No matching} \end{cases} \right) \quad (2)$$

Given an actionable warning  $x$ ,  $c$  refers to the code change of the potential fix revision of  $x$ . We define two types of patterns of structural matching, namely fix pattern and scope pattern. Fix pattern refers to our summarised bug-fix patterns for each warning type, and scope pattern refers to the expected location of the bug-fix code change. In particular: (a) If the commit code change matches the summarized fix pattern in the second column of Table 3, we consider this code change is responsible for resolving the corresponding warnings, and assign a structural matching score of 2 to this actionable warning to indicate the actionable warning is very likely to be fixed by its associated bug-fix reversion. (b) In addition to using the fix pattern as the strict checking rule, we also consider the scope pattern as a loosened standard for our study. The scope pattern refers to the expected location of the bug-fix code change (shown in the last column of Table 3). If the code changes fail to match the fix pattern while falling within the expected bug-fix scope, we consider the actionable warning as a potential bug and assign a score of 1. (c) For the actionable warning that does not match any structural matching rule, we assign it a score of 0.

**3.2.3 Aggregation of Weak Labels.** To obtain a more reliable and robust label for each actionable warning, we use majority voting to

combine the above semantic matching score and structural matching score, as demonstrated in Equ. 3.

$$Label(x) = \begin{cases} \text{VTB}, & CM(x) + CC(x) > 3 \\ \text{LTB}, & 2 \leq CM(x) + CC(x) \leq 3 \\ \text{UTB}, & 0 \leq CM(x) + CC(x) < 2 \end{cases} \quad (3)$$

Given an actionable warning  $x$ ,  $CM(x)$  refers to the semantic score for  $x$  using the commit message matching rule, and  $CC(x)$  refers to the structural score for  $x$  using the code change matching rule. (a) If the sum of  $CM(x)$  and  $CC(x)$  is greater than 3, it means the actionable warning is very likely to be a real bug from both semantic and structural aspects and we label  $x$  as **VTB (Very Likely To be Bugs)**. For example, when the commit message matches the bug type (i.e.,  $CM(x) = 3$ ) and the code change matches the fix pattern (i.e.,  $CC(x) = 2$ ) at the same time, we can almost ensure the actionable warning  $x$  is a real bug. (b) Similarly, if the sum of  $CM(x)$  and  $CC(x)$  falls between 2 and 3, it means the warning  $x$  matches the bug-fix reversion from either the semantic or structural aspect and is likely to be a real bug, we label  $x$  as **LTB (Likely To be Bugs)**. (c) Lastly, if the sum of  $CM(x)$  and  $CC(x)$  is less than 2, it means the actionable warning  $x$  mismatches the bug-fix reversion and is unlikely to be a real bug, we label such instances as **UTB (Unlikely To be Bugs)**.

So far, we assign each actionable warning with a weak label representing its likelihood of being a real bug. We then define the warning  $x$  whose  $Label(x)$  is **VTB** or **LTB** as **AWHB (Actionable Warning with High probability to be real Bug)**. In other words, actionable warnings that are likely or very likely to be real bugs are referred to as **AWHB** in this study, which is defined as follows:

$$\text{AWHB} = \{x | Label(x) \in \{\text{VTB}, \text{LTB}\}\} \quad (4)$$

### 3.3 Two-stage Model

So far, we have collected actionable warnings and further labeled the AWHB under weak supervision. Due to the data imbalance between the false positive warnings, actionable warnings and AWHB, we propose a two-stage model which includes the coarse-grained detection stage and fine-grained reranking stage. In the coarse-grained detection stage, the actionable warnings dataset is used to warm-up and let the model learn how to distinguish actionable warnings from false alarms and acquire a basic semantic understanding of static analysis warnings. In the fine-grained reranking stage, we further fine-tune the model to rerank the AWHB to the top by weakly supervised learning.

**3.3.1 Warm-up the Detector Model.** At this stage, the model is warmed up by learning to distinguish actionable warnings from false alarms, which helps it acquire a fundamental semantic understanding of static analysis warnings.

For a given reported warning, two types of key information are prepared as input for UniXcoder, i.e., text related input (e.g., bug type) and code related input (e.g., AST). UniXcoder is well-suited for our encoding task because it combines multi-modal data (comment and AST) during training. In particular, regarding the text related input, we extract the bug type, qualifier, procedure, and filename from warning reports. Bug type refers to the category of software issue, qualifier provides additional information, procedure outlines

the specific function warning occurred, and filename identifies the location of the bug in the code. Regarding the code related input, we pinpoint the bug location of the reported warning and extract the associated buggy statement for this warning. We then check the parent node of the buggy statement from AST and add control flow information of the buggy statement from AST as code context. We feed the text related input and code related input into UniXcoder and obtain the embedding by directly concatenating the outputs of the pre-trained model.

Warm-up is a crucial step in the fine-tuning process of the pre-trained model, as it helps to stabilize the model's weights, reduce overfitting, and allow it to adapt to the new task more effectively [60]. In this work, we began by warming up the pre-trained model UniXcoder through identifying actionable warnings. The actionable warning identification task can be viewed as a binary classification problem. That is, for a given reported warning  $x$ , we use the model  $f(x; \theta)$  to determine whether  $x$  is actionable or not. The actionable warning dataset without weak supervision is used to warm up  $f(x; \theta)$  and the optimization goal is defined as follows:

$$\min_{\theta} \frac{1}{N} \sum_{x \in \mathcal{X}} \mathcal{L}(f(x; \theta), y_x) \quad (5)$$

where  $\mathcal{L}$  denotes a loss function and  $y_x$  is the actionable warning label without weak supervision. Any loss function suitable for a classification task can be used in the warm-up process, and in this study, we use the Binary Cross Entropy Loss.

**3.3.2 Fine-tune the Reranker Model.** To rank different levels of actionable warnings, we transform the ranking problem into a multiclass classification task. That is, for a given reported warning  $x$ , we aim to predict  $x$  as V/L/U/False Warning based on our actionable warnings dataset under weak supervision. The optimization problem is defined as follows:

$$\min_{\theta} \frac{1}{N} \sum_{x \in \mathcal{X}} \mathcal{J}(g(f(x; \theta)), \tilde{y}_x) \quad (6)$$

The  $\tilde{y}_x$  is the label aggregated from weak supervision.  $g(x)$  is the softmax function to compute the probability of each class for  $x$ , and  $\mathcal{J}$  is Cross Entropy Loss, which is suitable for the multiclass classification task. When the optimization is done, the final ranking score for each warning  $x$  can be inferred as follows:

$$S(x) = \begin{cases} class(x) + g_{\tilde{y}}(x), & \tilde{y} \in \text{VTB}, \text{LTB}, \text{UTB} \\ class(x) - g_{\tilde{y}}(x), & \tilde{y} \in \text{False Warning} \end{cases} \quad (7)$$

where  $class(x)$  maps each warning  $x$  to a base class score 0/1/2/3 if the predicted class is False Warning/UTB/LTB/VTB,  $\tilde{y}$  denotes the predicted class of  $x$  and  $g_{\tilde{y}}(x)$  denotes the probability of the predicted class. Suppose  $x_1$  is predicted as VTB with a probability of 0.6, then the final ranking score  $S(x_1) = 3.6$ ; if  $x_2$  is predicted as False Warning with a probability of 0.7, then its final ranking score  $S(x_2) = -0.7$ . Finally, all the actionable warnings are ranked by their final ranking scores for recommendation.

## 4 Evaluation

### 4.1 Data Preparation

We build our actionable warning dataset by collecting data from the top 500 repositories (ordered by the number of stars) in GitHub for C



repositories. To the best of our knowledge, this is the first actionable warning dataset collected by mining the histories of popular GitHub C repositories. In this study, we adopt two static analysis tools, i.e., *Infer* and *Flawfinder*, to detect potential warnings in software systems. For *Infer*, we use the default settings to generate warnings on our dataset. In total, *Infer* reported 31,380 warnings, among which 99.1% (31,098 warnings) belong to the top four warning types: **Uninitialized Variable**, **Resource Leak**, **Null Dereference**, and **Dead Store**. Therefore, we focus on these four categories in our study. For the *Flawfinder*, we set the priority level to 4, which indicates the vulnerability with high risks, and collect the **Buffer Overflow** type warnings.

To save time and resources, we have implemented a filtering process for our projects. First, we exclude any projects that require more than half an hour to compile and test using SA tools. Additionally, we filter out projects whose latest revisions cannot be compiled successfully. Finally, we collect both actionable and false warnings from 68,274 revisions of the 394 projects. To avoid the double counts of warnings, following the setting of *Infer*, we deduplicated warnings according to the unique hash value of each warning, which is generated by bug type, located file, located procedure, and context code. Moreover, by comparing the reported bug location, we make sure there are no warnings duplicated reported by *Infer* and *Flawfinder*. The detail of the collected warning dataset is listed in Table 4.

Table 5 illustrates the commit message and weak supervision label of the collected 1,889 actionable warnings and the aggregated labels of them based on the Equ. 3. Following the settings of previous work [67, 73, 75], the current existing warnings that haven't been acted by developers for more than two years are regarded as false warnings. As a result, 39,052 warnings in our study are assigned with labels of **False Warnings**. We regard the actionable warning whose aggregated labels are **VTB** (Very Likely To be Bugs) and **LTB** (Likely To be Bugs) as **AWHB**. As can be seen, AWHB only makes up a small proportion (287/40,941) of the total reported warnings.

To address the data imbalance in our dataset, in the warm-up stage, we intentionally preserved the original distribution of warnings to allow the model to learn the general characteristics of both actionable and false warnings. In the fine-tuning stage, we first oversampled the actionable warnings (10×) to balance the training dataset. Then, we employed a weighted cross-entropy loss, where each actionable warning was assigned a weight based on its weakly supervised labeled score, which helps the model pay more attention to the AWHB.

## 4.2 Baselines

To demonstrate the effectiveness of our proposed model, ACWRECOMMENDER, we compared it to the following selected baselines:

**Golden Feature-RF:** The golden feature [67] is currently the state-of-the-art approach for actionable warning identification, which applies a random forest classifier to the 23 most influential features. We followed the experimental settings of it after removing 5 features that may introduce data leakage [36].

**Golden Feature-SVM:** Further research [73] on the Golden Features of Wang et al. [67] showed that a linear SVM algorithm was an optimal choice for identifying actionable warnings.

**Table 4: Data Statistics**

Warning Label	Warning Type	Count
Actionable Warning	Uninitialized Variable	164
	Null Dereference	238
	Resource Leak	50
	Dead Store	338
	Buffer Overflow	1,099
	Total Warning	1,889
False Warning	Uninitialized Variable	7,660
	Null Dereference	10,544
	Resource Leak	842
	Dead Store	11,544
	Buffer Overflow	8,462
	Total Warning	39,052

**Table 5: Data with Weak Supervision Statistics**

Warning Label	Warning Type	Count
Commit Message Matching Rule	Warning Type Keyword	143
	Warning Context Keyword	44
	Common Keyword	550
	No Matching	1,152
Code Change Matching Rule	Patch Pattern	228
	Scope Patter	396
	No Matching	1265
Aggregated Label	VTB	151
	LTB	136
	UTB	1,602
	False Warning	39,052

**Random Forest:** According to our preliminary experiment among more than 10 machine learning models based on the warning feature embedding from UniXcoder, we selected Random Forest as a baseline, which performs best in actionable warning identification.

**Random Ranking:** In the AWHB recommendation task, the Random ranking strategy shuffles the warning dataset randomly.

## 4.3 Evaluation Metrics

To evaluate the performance of AWHB recommendation, we use the metrics including MRR and nDCG@K, which are widely-used metrics in information retrieval and recommender system [25, 27, 56, 61]. Mean Reciprocal Rank (MRR) measures how quickly the first AWHB is found in the recommended warning list. The normalized discounted cumulative gain at K (nDCG@K) measures the quality of the recommendation by comparing the recommended warnings list to a ground truth warnings list. nDCG@K takes into account both the relevance of AWHB and its position in the recommended warnings list. A higher MRR and nDCG@K value indicates a better recommendation result.

## 4.4 Quantitative Analysis

**4.4.1 RQ1: The Effectiveness of Weak Supervised Labeling.** To evaluate the effectiveness of our weak supervision labeling strategy in identifying real bugs, we manually inspected a sample of the actionable warnings and our labeled data respectively. Specifically, we randomly sampled 200 actionable warnings and AWHBs from our

**Table 6: Labeling Effectiveness Evaluation**

Warning Label		Uninitialized Variable	Null Dereference	Resource Leak	Dead Store	Buffer Overflow	Total
Actionable Warning	# of Samples	28	35	14	51	72	200
	# of Real Bugs	5	10	2	7	15	39
	Real Bugs Ratio	5/28	10/35	2/14	7/51	15/72	19.5%
AWHB	# of Samples	28	35	14	51	72	200
	# of Real Bugs	25	31	11	29	63	159
	Real Bugs Ratio	25/28	31/35	11/14	29/51	63/72	79.5%

study, respectively. We then invited three software developers, each with at least five years of C programming experience, to review each sample and identify if the warning is valid and represents a real bug. All developers were asked to independently review each warning, following clear instructions that clarified the evaluation criteria: *Please determine whether the given commit fixed the warning reported by the static analysis tool based on: (1) the warning report; (2) the warning's surrounding code context; (3) the related code changes; and (4) the commit messages.* To minimize potential order effects, all warnings were presented to reviewers in a randomized order. Additionally, no additional information such as CWE labels or model predictions, was provided to avoid confirmation bias. After collecting results from each developer, the first author engaged in and had a discussion with three developers when different opinions were met. Finally, each warning is assigned with a unique label (i.e., real bug or false alarm) after discussions. The results of the manual validation are shown in Table 6:

**Only 19.5% of actionable warnings represent real bugs**, validating our empirical findings in Sec.2 that the actionable warning identified by prior assumption may not necessarily indicate real issues. An example is shown in Table 1 Ex.1, *Infer* falsely reported a **Null Dereference** warning in a previous revision and this warning was recognized as an actionable warning. According to previous empirical studies [32], 35% to 91% warnings reported by SA tools are spurious false alarms.

**Our weak supervised labeling method is effective for identifying real bugs, 79.5% of AWHBs are verified as true bugs by developers.** For example, 25 out of 28 **Uninitialized Variable** warnings are validated as realistic bugs. The matching from semantic and structural aspects allows our weak supervised labeling process to accurately identify a real **Uninitialized Variable** bug.

**Our weak supervised labeling method performs consistently well for different SA tools and warning types.** For example, in terms of *Infer*, the real bug ratio of four warning types (**Uninitialized Variable**, **Null Dereference**, **Resource Leak**, and **Dead Store**) is 75% (96/128), in terms of *Flawfinder*, the **Buffer Overflow** warning real bug ratio is 87.5% (63/72), which shows the robustness of our method.

**4.4.2 RQ2: The Recommendation Effectiveness Evaluation.** Considering that more than 80% of AWHBs are validated as real bugs in RQ1, we regard the AWHBs as warnings that should be equally recommended and handled earlier. In this research question, we evaluate the AWHB recommendation effectiveness of our fine-tuned reranker model. During the evaluation, we built 100 test samples by randomly selecting 1,000 warnings from the testing set. For each test sample, we ensured there were at least 5 AWHBs within the

**Table 7: Recommendation Effectiveness Evaluation**

Measure	nDCG@1	nDCG@3	nDCG@5	MRR
Random Ranking	0.020	0.031	0.052	0.007
Golden Feature-SVM	0.074	0.095	0.137	0.064
Golden Feature-RF	0.087	0.104	0.152	0.121
Random Forest	0.143	0.211	0.239	0.204
<b>ACWRECOMMENDER</b>	<b>0.356</b>	<b>0.426</b>	<b>0.477</b>	<b>0.452</b>

1,000 sampled warnings (considering the actionable warning ratio is around 1:20), each test sample was then shuffled for evaluation. For each test sample, we aim to evaluate how often the AWHBs are ranked higher among other warnings. Thus, the widely-accepted metrics nDCG@K and MRR are used to measure different methods' ranking performance. In particular, regarding baseline methods, we use the actionable warnings' probability scores outputted by the baseline methods for ranking. Regarding our proposed model, we use the final ranking score of **ACWRECOMMENDER** (defined in Equ 7) for ranking. The evaluation result is listed in Table 7. From the table, we can observe the following points:

Although initially designed for identifying actionable warnings, we attempted to rank the warnings based on their score from Golden Feature-based approaches and utilized it as a baseline method in recommendation evaluation. Similar to the actionable warning identification, the Golden Feature-based approaches (i.e., **Golden Feature-SVM** and **Golden Feature-RF**) outperform **Random Ranking** in recommending AWHB. However, their suboptimal performance across all metrics suggests that it is inadequate for addressing the recommendation task. Similar to the result in RQ1, the better performance of **Random Forest** also suggests the features extracted by UniXcoder are more suitable for actionable warning recommendation than the selected features used in **Golden Feature**.

**Our proposed model is effective for AWHB recommendation and outperforms all the baseline methods.** The nDCG@1 value exceeds 0.3, indicating that over three-tenths of the queries can accurately identify AWHB in the topmost position of the recommended warning list with a high degree of certainty. The MRR value is approximately 0.45, suggesting that on average, the first AWHB can be found at the third position in the recommended warning list. We attribute this performance to the following reasons: First, the weak label from our matching rules can be efficient in distinguishing the real bugs from actionable warnings. Second, the semantic features employed in **ACWRECOMMENDER** are suitable for AWHB recommendation. To be specific, the commit message matching rule utilizes text information of warnings, and the code



**Table 8: Type-Wise Recommendation Evaluation**

Measure	nDCG@1	nDCG@3	nDCG@5	MRR
Uninitialized Variable	0.380	0.492	0.531	0.511
Null Dereference	0.413	0.507	0.557	0.520
Resource Leak	0.485	0.534	0.591	0.566
Dead Store	0.263	0.395	0.419	0.410
Buffer Overflow	0.312	0.425	0.487	0.434

change matching rule relies on detecting similar patterns in the code information of warnings.

**4.4.3 RQ3: Type-Wise Evaluation.** To gain a deeper understanding of how our model performs on different static analysis tools and different warning types, in this research question we conduct a type-wise evaluation. Specifically, we calculated ACWRECOMMENDER’s performance regarding five warning types (i.e., **Uninitialized Variable**, **Null Dereference**, **Resource Leak**, **Dead Store** and **Buffer Overflow**) of two static analysis tools (i.e., *Infer* and *Flawfinder*). The evaluation results are presented in Table 8. From these tables, we have the following observations:

**Our approach can generalize well to different static analysis tools and different warning types.** To verify the generalizability of our approach, we adopt two unique static analysis tools, i.e., *Infer* and *Flawfinder*, to cover different types of warnings. It is clear that ACWRECOMMENDER performs consistently well for most types of warnings, such as **Uninitialized Variable**, **Null Dereference** and **Resource Leak** warning reported by *Infer* and **Buffer Overflow** warning reported by *Flawfinder*. We attribute the generalization ability of our approach to the following reasons: (i) Our base model UnixCoder can automatically and effectively learn features from the warning code and warning report without any reworking. Different static analysis tools’ output can be easily incorporated into the pipeline of our approach. (ii) Our weak supervised labeling method establishes the structural and semantic matching rule for AWHBs, these rules are applicable for general warning types.

## 4.5 In-the-wild Evaluation

Our ultimate goal is to help developers find real bugs reported by SA tools by inspecting as few warnings as possible. To do this, we further conduct an in-the-wild evaluation to evaluate the practical value of our ACWRECOMMENDER for checking real bugs in real-world GitHub repositories. We began by randomly selecting 10 C repositories hosted on GitHub that had more than 500 stars and were last updated within a week. We ensured that the selected repositories were not used in the previous training and testing set. Next, we ran *Infer* on the latest revision of each repository and obtained the *Infer* reported warnings for each repository. Then ACWRECOMMENDER reranks the *Infer* warnings and generates a warning list for checking. After removing 4 projects that can not be compiled, we obtained 2,197 *Infer* reported warnings from 6 selected repositories (namely libevent, flac, vim, xrdp, netcdf-c and radare2).

**Developer’s Feedback of Pull Request** Theoretically, we can submit all of our top-ranked warnings to ask developers to confirm if these warnings are real bugs. However, the workload is too

heavy if developers need to check every warning candidate. To reduce the burden of developers, we first checked Top30% (743) reported warnings and validated 23 warnings as potential bugs. Following that, we manually fixed each validated warning and submitted a warning fix PR (Pull Request) directly to its corresponding GitHub repository. Among the 23 submitted validated warnings, 21 of them have already been confirmed by the developers, 19 of them have been merged, and 2 of them have been approved by code reviewers. It is worth mentioning that **more than half of the confirmed bugs (12/21) are identified within the Top 10% of the recommended warnings, which justifies the effectiveness of our tool for alleviating the effort for finding real bugs in a massive amount of warnings.**

In general, developers appreciated the contributions we made to their repositories. For example, after confirming our validated **Null Dereference** warning [1], one developer of vim project responded “Thanks. Also realized another issue with make\_node while reviewing the changes.” The *Infer* reported 239 warnings for this revision, it is time-consuming and costly for developers to check these warnings one by one, after adopting our tool, the developer can easily find this bug by only checking 3 top warnings. Similarly, another developer from libevent confirmed our pull-request [2] commented after our further explanation, and commented that “Yes, missed the err label, you are right.” The response shows the AWHB recommended by our tool can assist developers in finding bugs that they may have overlooked.

**Developer’s Feedback of Submitted Issues.** Because we lack the programming context and domain expertise with respect to each test project, our validated warnings are by no means the entire warnings that need to be fixed. In other words, there are still warnings that we are unsure about. To alleviate the bias of our manual validation, for each top-10 reported warning in the above GitHub repositories, we formed the warning as an issue report and submitted the issue to its GitHub repository. In particular, we submitted 50 warnings (excluding 10 warnings already confirmed through pull requests) to 6 GitHub repositories. 10 issues have been confirmed by developers as real bugs. Finally, **among these top-60 warnings reported by ACWRECOMMENDER, 27 warnings (including 10 warnings submitted by pull requests and 17 warnings submitted by issue reports) are identified as real bugs**, which further confirms the effectiveness of our tool for recommending valid actionable warnings.

The developers of radare2 validated and fixed all the committed top-5 warnings[3]. The developers of xrdp validated two **Dead Store** warnings[4]. Some of our recommended warnings are true defects but do not affect the usage or safety of software, therefore the developers are reluctant to act on these warnings. For example, even xrdp developers confirmed our reported **Dead Store** warnings, they consider these warnings “does not matter” and replied “I can’t see that any of these are of any concern from a security perspective.” The developers’ perspective of **Dead Store** warnings has led to a low-quality label for this warning type, resulting in real bugs being labeled as false alarms. It greatly hinders the learning performance of our tool on this warning type and further explains the relatively poor performance of our approach on **Dead Store** warnings. Some of our recommended warnings are considered by developers as false alarms by overestimating program behaviors. For example,

**Table 9: Labeling Effectiveness on SonarQube Warnings**

Count	Actionable Warnings	AWHB	Real Bugs	Ratio
Uninitialized Variable	69	29	25	25/29
Null Dereference	47	19	15	15/19
Resource Leak	28	5	5	5/5
Dead Store	14	0	0	-
Buffer Overflow	11	3	2	2/3
Total	169	56	47	83.9%

for a **Null Dereference** warning reported by our tool in `xrdp`, the developer responded "*Potentially a NULL pointer dereference but `xrdp` is full of instances of this idiom.*"

## 5 Discussion

### 5.1 Generalizability of Heuristic Rules

In this work, our weakly-supervised labeling strategy is based on heuristically defined semantic and structural matching rules. These heuristic rules are successfully applied to *Infer* and *Flawfinder* to verify their effectiveness. To demonstrate whether these rules are generalizable to other SA tools, we extend these summarized rules to an additional static analysis tool, *SonarQube*.

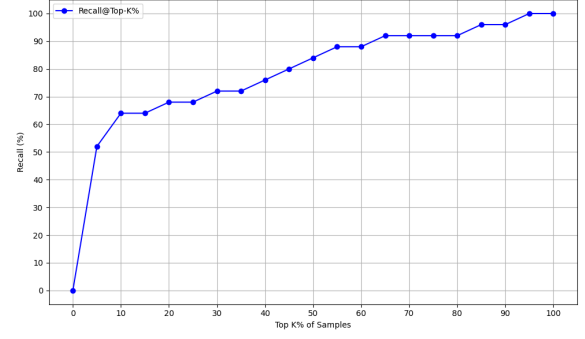
**Experimental Setup.** We randomly sampled 50 projects from our dataset, for each project, we ran *SonarQube* on its compiled code of each commit version to collect actionable warnings (introduced in Section 3). As a result, we collected 169 actionable warnings across the aforementioned five warning types. After that, we apply our heuristic rules to these actionable warnings to identify AWHBs. Finally, 56 out of 169 actionable warnings are labeled as AWHBs. After that, we follow the same experimental settings (detailed in Section 4.4.1) to manually review each of the 169 actionable warnings as real bugs or not.

**Experimental Results.** The experimental results are shown in Table 9. From the table, we can see that: (1) 83.9% AWHBs (i.e., 47 out of 56) were confirmed by developers as real bugs. The effectiveness of our heuristic rules is consistent across different static analysis tools (i.e., *Infer*, *Flawfinder*, and *SonarQube*), demonstrating the effectiveness and robustness of our weak supervision labeling strategy. (2) It is worth mentioning that we use warning descriptions and ASTs to define these heuristic rules, without relying on any specific static analysis tools. This design ensures our heuristic rules are highly generalizable and scalable, which can be easily extended to other static analysis tools.

### 5.2 Practical Application Settings

Our approach re-ranked AWHBs higher up for developers' investigation. To assess the trade-off between inspection effort and the tool's effectiveness, we analyze the practical application settings.

**Experimental Setup.** To assess the trade-off between inspection effort and effectiveness of identifying real bugs, we evaluate the *Recall@K* metric using our testing dataset, where *K* refers to the top-*K*% ranked warnings, and *Recall@K* represents the proportion of real bugs that appear in top-*K*% ranked warnings.

**Figure 3: Recall@Top-K% curve**

**Experimental Results.** The evaluation result is shown in Fig. 3. From the figure, we can see that: (1) Inspecting the top 5% of ranked warnings can already cover over 50% of AWHB. Therefore, in practical usage, we recommend that developers review the top 5% of recommended warnings to achieve a good balance between inspection cost and bug detection effectiveness. (2) As shown in our in-the-wild experiments, we submitted the top 10 ranked warnings identified by our approach, and 27/60 submitted warnings were confirmed as real bugs by project developers, demonstrating strong real-world applicability. (3) For deployment, our approach can be easily implemented as a plugin or extension for popular IDEs (e.g., VSCode, IntelliJ). Once integrated, it can analyze static analysis warnings in real time and highlight only the AWHB. It can also be deployed as a post-processing step in CI pipelines to filter SA outputs before they reach developers. This helps developers better utilize SA tools and further improve software quality.

### 5.3 Reducing Manual Efforts with LLMs

Our approach relies on heuristic rules in the weak supervision labeling process, while defining these rules involves manual inspection, we believe the effort is both cost-effective and efficient. For semantic rules, relevant keywords for each warning type can be easily extracted from warning descriptions, making this step straightforward and low-effort. For structural rules, although summarizing fix patterns from ASTs requires domain expertise, this manual work is a one-time cost. Once established, these patterns can be reused across projects with minimal maintenance.

**Experimental Setup.** Inspired by the capabilities of LLMs [21, 22, 50, 51, 70, 72], we further explored how LLMs can be utilized to reduce manual efforts during this process. Specifically, for each type of warning, we leverage ChatGPT to automatically generate relevant keywords (i.e., semantic rules) and candidate fix patterns (i.e., structural rules) by providing LLMs with reported warnings.

**Experimental Results.** The detailed prompts and LLMs' outputs are demonstrated in our recorded chat histories [9–13]. We manually reviewed the LLM outputs to assess their consistency with human-summarized patterns. Our manual analysis reveals that: (1) Regarding semantic rules, the keywords produced by the LLM cover most of the human summarized keywords across all five types of warnings. (2) Regarding structural rules, three types of warnings (e.g., Dead Code, Null Dereference, and Uninitialized

Variable) exactly match our human-defined patterns and even suggest additional fix patterns. For Resource Leak and Buffer Overflow warnings, the LLM generates specific fix patterns tailored to typical cases. For example, they recommend using `fclose()` for `_IO_FILE` leaks but may overlook user-defined resource-closing functions. In contrast, more comprehensive function call patterns, e.g., *close* and *free*, are summarized by human experts. (3) Overall, LLMs show promising potential in reducing manual efforts for summarizing heuristic rules, and how to utilize LLMs to generate comprehensive and validated rules remains an open challenge and is beyond the scope of this work.

## 5.4 Threats to Validity

**Threats to internal validity.** The internal validity relates to potential errors in model implementation and experimental settings. Our approach relies on heuristic rules for data labeling, which may introduce labeling noise. To mitigate this, we manually reviewed the automatically labeled data and found that over 80% of the labels were accurate. Our heuristic-based method also suffered from overlooking false negatives. Warnings present in the latest commit, but potentially fixed in the future, are not labeled as actionable under our current definition. These false negatives are excluded from our analysis because they have not yet met the actionable criterion. Additionally, some false negatives arise from the incompleteness of our manually summarized heuristic rules, which may fail to capture all possible fix patterns. In the future, we plan to enhance our labeling strategy to better capture delayed fixes.

**Threats to external validity.** External validity concerns the generalizability and representativeness of our results. Our experiments are conducted on five common C/C++ vulnerability types detected by two widely used static analysis tools. In terms of representativeness, our study focuses on five representative warning types. Four of them (Uninitialized Variable, Null Dereference, Resource Leak, Dead Store) account for over 99% of Infer’s reported warnings, while Buffer Overflow is the most frequent warning in Flawfinder. Regarding severity, several selected types correspond to top-ranked CWEs [7] and have been widely studied [41, 43, 52, 63]. In terms of generalizability, our framework is designed to be both generalizable and scalable, and the consistent effectiveness across different SA tools (i.e., *Infer*, *Flawfinder*, and *SonarQube*) further demonstrates the generalizability of our approach.

**Threats to construct validity.** The Construct validity is introduced by the confirmation bias in manual evaluations, especially when assessing the accuracy of our weak supervision labeling strategy. To mitigate this risk, we recruited three experienced developers and presented the warnings in a randomized order, asking each to determine whether each warning corresponded to a real bug independently.

## 6 Related Work

Based on the assumption that the warnings that are not acted upon by developers are regarded as false alarms. The goal of warning identification is to identify actionable warnings and false alarms. Several studies, leveraged machine learning to distinguish actionable and non-actionable alarms based on manually predefined feature [14, 29, 33, 39, 40, 44, 64, 78, 79]. Zhang et al. [79] extracted

variable-level features and showed that they outperform file/class-level ones. Wang et al. [67] identified 20 “Golden Features,” achieving state-of-the-art performance in actionable warning identification. However, some researchers doubted the result of the Golden Feature. Yang et al. [74] investigated the superior efficacy of the Golden Feature in SVM, as compared to CNN, due to the intrinsically simple data. Kang et al. [36] revealed a data leakage in Golden Feature, and reported a pool performance for it after fixing the data leakage and data duplication. They concluded that the Golden Feature is inadequate for the warning identification task.

The research on warning recommendation aims to prioritize warnings that are more likely to be true bugs and ordered up in the list. The key insight of these studies is to predict the probability of static analysis warnings. Some of them combined the results from several SA tools [24, 53, 55, 58]. For example, Ribeiro et al. [58] using AdaBoost on features from three tools, and Flynn et al. [24] gathered warnings from a suite of software assurance tools. However, they require analyzing the same code by multiple tools, and this increases the code analysis time. Some researchers recommend the warnings involving additional human-in-the-loop [34, 37, 54, 57]. Mangal et al. [57] used Bayesian inference on a probabilistic model derived from the derivation graph, and regarded the warning inspected by the developer as the ground truth. DRAKE [34], for instance, updated weights iteratively based on user feedback. It updated the weights based on the user feedback each round. A number of studies recommend the warnings based on the history-aware information [15, 18, 38, 68]. Aman et al. [15] estimated the lifetimes of alarms by using the survival analysis method, and assigned higher priority to alarms that have shorter lifetimes.

All the prior approaches focus on handling actionable warnings. While our approach first proposed a weakly supervised labeling method to assign different labels to the actionable warnings based on their probabilities to be real bugs. This research is based on our preliminary tool demo [69] by extending it to a two-stage modeling, conducting a comprehensive evaluation to verify its effectiveness and value.

## 7 Conclusion

This research aims to identify actionable warnings produced by static analysis tool and recommend the Actionable Warning with High probability to be real Bug. To address these task, we first collect actionable warnings from top-500 GitHub repositories. We then propose a weakly supervised strategy to identify Actionable Warning with High probability to be real Bugs (AWHBs). We propose an approach, namely ACWRECOMMENDER, which leverages UniXcoder to recommend AWHBs. Comprehensive evaluation results show the effectiveness of our approach on this task. Extensive experiments on the real-world GitHub repositories have demonstrated its practical value in improving the utility of SA tools.

## Acknowledgments

This research is supported by the National Science Foundation of China (No. 62572322). This research is partially sponsored by the Shanghai Sailing Program (23YF1446900) and the CCF-Huawei Populus Grove Fund. We also thank the anonymous reviewers for their insightful comments and suggestions.

## References

- [1] 2023. <https://github.com/vifm/vifm/pull/882>
- [2] 2023. <https://github.com/libevent/libevent/pull/1429>
- [3] 2023. <https://github.com/radareorg/radare2/issues/22313>
- [4] 2023. <https://github.com/neutrino-labs/xrdb/security/advisories/GHSA-f3qh-qc3p-254c#event-113510>
- [5] 2023. Actionable warning in libevhpt. <https://github.com/Yellow-Camper/libevhpt/commit/d13b72b3887096f54d3570b723f2f518cf0e903b>
- [6] 2023. Actionable warning in open62541. <https://github.com/open62541/open62541/commit/d52786e61878fd4b2b798c84de49c289b2bfff6f>
- [7] 2024. 2024 CWE Top 25 Most Dangerous Software Weaknesses. [https://cwe.mitre.org/top25/archive/2024/2024\\_cwe\\_top25.html](https://cwe.mitre.org/top25/archive/2024/2024_cwe_top25.html)
- [8] 2025. ACWR Recommender replication package. <https://zenodo.org/records/15550900>
- [9] 2025. Buffer Overflow Pattern from ChatGPT. <https://chatgpt.com/share/686f813f-e080-800d-ae2-2b94db9589a7>
- [10] 2025. Dead Store Pattern from ChatGPT. <https://chatgpt.com/share/686f7f58-98a4-800d-92f5-a02f51edf534>
- [11] 2025. Null Dereference Pattern from ChatGPT. <https://chatgpt.com/share/68387613-fdf8-800d-824f-92ea172ec4df>
- [12] 2025. Resource Leak Pattern from ChatGPT. <https://chatgpt.com/share/686f8015-3acc-800d-b187-cbe7554cdc55>
- [13] 2025. Uninitialized Variable Pattern from ChatGPT. <https://chatgpt.com/share/686f7e7c-006c-800d-89c2-2a60747cf4cf>
- [14] Enas A. Alihashashneh, Rajeev R. Raj, and James H. Hill. 2018. Using Machine Learning Techniques to Classify and Predict Static Code Analysis Tool Warnings. *2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)* (2018), 1–8.
- [15] Hirohisa Aman, Sousuke Amasaki, Tomoyuki Yokogawa, and Minoru Kawahara. 2019. A Survival Analysis-Based Prioritization of Code Checker Warning: A Case Study Using PMD. In *International Conference on Big Data, Cloud Computing, Data Science & Engineering*.
- [16] Pavel Avgustinov, Arthur I Baars, Anders S Henriksen, Greg Lavender, Galen Menzel, Oege De Moor, Max Schafer, and Julian Tibble. 2015. Tracking static analysis violations over time to capture developer characteristics. In *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, Vol. 1. IEEE, 437–447.
- [17] Guillaume Brat and Arnaud Venet. 2005. Precise and scalable static program analysis of NASA flight software. In *2005 IEEE Aerospace Conference*. IEEE, 1–10.
- [18] Aji Ery Burhandenny, Hirohisa Aman, and Minoru Kawahara. 2017. Investigation of Coding Violations Focusing on Authorships of Source Files. *2017 5th Intl Conf on Applied Computing and Information Technology/4th Intl Conf on Computational Science/Intelligence and Applied Informatics/2nd Intl Conf on Big Data, Cloud Computing, Data Science (ACIT-CSI-BCD)* (2017), 248–253.
- [19] Cristiano Calcagno, Dino Distefano, Jérémy Dubreil, Dominik Gabi, Pieter Hooimeijer, Martino Luca, Peter W. O'Hearn, Irene Papakonstantinou, Jim Purbrick, and Dulma Rodriguez. 2015. Moving Fast with Software Verification. In *NASA Formal Methods*. <https://api.semanticscholar.org/CorpusID:12281434>
- [20] Zirui Chen, Xing Hu, Xin Xia, Yi Gao, Tongtong Xu, David Lo, and Xiaohu Yang. 2024. Exploiting Library Vulnerability via Migration Based Automating Test Generation. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering (Lisbon, Portugal) (ICSE '24)*. Association for Computing Machinery, New York, NY, USA, Article 228, 12 pages. doi:10.1145/3597503.3639583
- [21] Zhenlong Dai, Bingrui Chen, Zhuolu Zhao, Xiu Tang, Sai Wu, Chang Yao, Zhipeng Gao, and Jingyuan Chen. 2025. Less is More: Adaptive Program Repair with Bug Localization and Preference Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 128–136.
- [22] Zhenlong Dai, Chang Yao, WenKang Han, Yuanying Yuan, Zhipeng Gao, and Jingyuan Chen. 2024. Mpcoder: Multi-user personalized code generator with explicit and implicit style representation learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 3765–3780.
- [23] Dino Distefano, Peter W. O'Hearn, and Hongseok Yang. 2006. A Local Shape Analysis Based on Separation Logic. In *International Conference on Tools and Algorithms for Construction and Analysis of Systems*. <https://api.semanticscholar.org/CorpusID:361092>
- [24] Lori Flynn, William Snively, David Svoboda, Nathan M. VanHoudnos, Richard Qin, Jennifer Burns, David Zubrow, Robert Stoddard, and Guillermo Marce-Santurio. 2018. Prioritizing Alerts from Multiple Static Analysis Tools, Using Classification Models. *2018 IEEE/ACM 1st International Workshop on Software Qualities and their Dependencies (SQUADE)* (2018), 13–20.
- [25] Zhipeng Gao, Xin Xia, David Lo, and John Grundy. 2020. Technical Q&A site answer recommendation via question boosting. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 30, 1 (2020), 1–34.
- [26] Zhipeng Gao, Xin Xia, David Lo, John Grundy, and Thomas Zimmermann. 2021. Automating the removal of obsolete TODO comments. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 218–229.
- [27] Zhipeng Gao, Xin Xia, D. Lo, John C. Grundy, Xindong Zhang, and Zhenchang Xing. 2022. I Know What You Are Searching For: Code Snippet Recommendation from Stack Overflow Posts. *ACM Transactions on Software Engineering and Methodology* (2022).
- [28] Zhipeng Gao, Zhipeng Xue, Xing Hu, Weiye Shang, and Xin Xia. 2024. Easy over Hard: A Simple Baseline for Test Failures Causes Prediction. In *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*. 306–317.
- [29] Quinn Hanam, Lin Tan, Reid Holmes, and Patrick Lam. 2014. Finding patterns in static analysis alerts: improving actionable alert ranking. In *Proceedings of the 11th working conference on mining software repositories*. 152–161.
- [30] Sarah Heckman and Laurie Williams. 2008. On establishing a benchmark for evaluating static analysis alert prioritization and classification techniques. In *Proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement*. 41–50.
- [31] Sarah Heckman and Laurie Williams. 2009. A model building process for identifying actionable static analysis alerts. In *2009 International conference on software testing verification and validation*. IEEE, 161–170.
- [32] Sarah Heckman and Laurie Williams. 2011. A systematic literature review of actionable alert identification techniques for automated static code analysis. *Information and Software Technology* 53, 4 (2011), 363–387. doi:10.1016/j.infsof.2010.12.007 Special section: Software Engineering track of the 24th Annual Symposium on Applied Computing.
- [33] Kihong Heo, Hakjoo Oh, and Kwangkeun Yi. 2017. Machine-Learning-Guided Selectively Unsound Static Analysis. *2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE)* (2017), 519–529.
- [34] Kihong Heo, Mukund Raghothaman, Xujie Si, and M. Naik. 2019. Continuously reasoning about programs using differential Bayesian inference. *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation* (2019).
- [35] Nasif Imtiaz, Akond Ashfaqur Ur Rahman, Effat Farhana, and Laurie A. Williams. 2019. Challenges with Responding to Static Analysis Tool Alerts. *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)* (2019), 245–249. <https://api.semanticscholar.org/CorpusID:195298514>
- [36] Hong Jin Kang, Khai Loong Aw, and David Lo. 2022. Detecting false alarms from automatic static analysis tools: how far are we?. In *Proceedings of the 44th International Conference on Software Engineering*. 698–709.
- [37] Hyunsu Kim, Mukund Raghothaman, and Kihong Heo. 2022. Learning probabilistic models for static analysis alarms. In *Proceedings of the 44th International Conference on Software Engineering*. 1282–1293.
- [38] Sunghun Kim and Michael D. Ernst. 2007. Prioritizing Warning Categories by Analyzing Software History. *Fourth International Workshop on Mining Software Repositories (MSR'07:ICSE Workshops 2007)* (2007), 27–27.
- [39] Ugur Koc, Parsa Saadatpanah, Jeffrey S. Foster, and Adam A. Porter. 2017. Learning a classifier for false positive error reports emitted by static code analysis tools. *Proceedings of the 1st ACM SIGPLAN International Workshop on Machine Learning and Programming Languages* (2017).
- [40] Seongmin Lee, Shin Hong, Jungbae Yi, Taeksu Kim, Chul-Joo Kim, and Shin Yoo. 2019. Classifying False Positive Static Checker Alarms in Continuous Integration Using Convolutional Neural Networks. *2019 12th IEEE Conference on Software Testing, Validation and Verification (ICST)* (2019), 391–401.
- [41] Haonan Li, Yu Hao, Yizhuo Zhai, and Zhiyun Qian. 2023. Assisting static analysis with large language models: A chatgpt experiment. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 2107–2111.
- [42] Junjie Li and Jinxiu Yang. 2024. Tracking the Evolution of Static Code Warnings: The State-of-the-Art and a Better Approach. *IEEE Transactions on Software Engineering* 50, 3 (2024), 534–550.
- [43] Tuo Li, Jia-Ju Bai, Yulei Sui, and Shi-Min Hu. 2022. Path-sensitive and alias-aware tpestate analysis for detecting OS bugs. In *Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*. 859–872.
- [44] Guangtai Liang, Lingjing Wu, Qian Wu, Qianxiang Wang, Tao Xie, and Hong Mei. 2010. Automatic construction of an effective training set for prioritizing static analysis warnings. *Proceedings of the 25th IEEE/ACM International Conference on Automated Software Engineering* (2010).
- [45] Kui Liu, Dongsun Kim, Tegawendé F. Bissyandé, Shin Yoo, and Yves Le Traon. 2021. Mining Fix Patterns for FindBugs Violations. *IEEE Transactions on Software Engineering* 47, 1 (2021), 165–188. <https://doi.org/10.1109/TSE.2018.2884955>
- [46] Kui Liu, Anil Koyuncu, Dongsun Kim, and Tegawendé F Bissyandé. 2019. Avatar: Fixing semantic bugs with fix patterns of static analysis violations. In *2019 IEEE 26th International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 1–12.
- [47] Zhongxin Liu, Xin Xia, Ahmed E Hassan, David Lo, Zhenchang Xing, and Xinyu Wang. 2018. Neural-machine-translation-based commit message generation: how far are we?. In *Proceedings of the 33rd ACM/IEEE international conference on automated software engineering*. 373–384.

- [48] Yingwei Ma, Rongyu Cao, Yongchang Cao, Yue Zhang, Jue Chen, Yibo Liu, Yuchen Liu, Binhua Li, Fei Huang, and Yongbin Li. 2025. SWE-GPT: A Process-Centric Language Model for Automated Software Improvement. *Proceedings of the ACM on Software Engineering* 2, ISSTA (2025), 2362–2383.
- [49] Yingwei Ma, Qingping Yang, Rongyu Cao, Binhua Li, Fei Huang, and Yongbin Li. 2025. Alibaba lingmaagent: Improving automated issue resolution via comprehensive repository exploration. In *Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering*. 238–249.
- [50] Yubo Mai, Zhipeng Gao, Xing Hu, Lingfeng Bao, Yu Liu, and Jianling Sun. 2024. Are human rules necessary? generating reusable apis with cot reasoning and in-context learning. *Proceedings of the ACM on Software Engineering* 1, FSE (2024), 2355–2377.
- [51] Yubo Mai, Zhipeng Gao, Haoye Wang, Tingting Bi, Xing Hu, Xin Xia, and Jianling Sun. 2025. Towards Better Answers: Automated Stack Overflow Post Updating. In *2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE)*. IEEE, 591–603.
- [52] Ivano Malavolta, Kishan Nirghin, Gian Luca Scoccia, Simone Romano, Salvatore Lombardi, Giuseppe Scanniello, and Patricia Lago. 2023. Javascript dead code identification, elimination, and empirical assessment. *IEEE Transactions on Software Engineering* 49, 7 (2023), 3692–3714.
- [53] Na Meng, Qianxiang Wang, Qian Wu, and Hong Mei. 2008. An Approach to Merge Results of Multiple Static Analysis Tools (Short Paper). *2008 The Eighth International Conference on Quality Software* (2008), 169–174.
- [54] Kien-Tuan Ngo, Dinh-Truong Do, Thu-Trang Nguyen, and Hieu Dinh Vo. 2021. Ranking Warnings of Static Analysis Tools Using Representation Learning. *2021 28th Asia-Pacific Software Engineering Conference (APSEC)* (2021), 327–337.
- [55] Paulo Jorge Costa Nunes, Ibéria Medeiros, José Fonseca, Nuno Ferreira Neves, Miguel Pupo Correia, and Marco Paulo Amorim Vieira. 2018. An empirical study on combining diverse static analysis tools for web security vulnerabilities based on development scenarios. *Computing* 101 (2018), 161–185.
- [56] Fangcheng Qiu, Zhipeng Gao, Xin Xia, David Lo, John Grundy, and Xinyu Wang. 2021. Deep just-in-time defect localization. *IEEE Transactions on Software Engineering* 48, 12 (2021), 5068–5086.
- [57] Mukund Raghothaman, Sulekha Kulkarni, Kihong Heo, and Mayur Naik. 2018. User-guided program reasoning using Bayesian inference. In *Proceedings of the 39th ACM SIGPLAN Conference on Programming Language Design and Implementation*. 722–735.
- [58] Athos Ribeiro, Paulo Meirelles, Nelson Lago, and Fabio Kon. 2019. Ranking warnings from multiple source code static analyzers via ensemble learning. In *Proceedings of the 15th International Symposium on Open Collaboration*. 1–10.
- [59] Justin Smith, Brittany Johnson, Emerson R. Murphy-Hill, Bill Chu, and Heather Richter Lipford. 2019. How Developers Diagnose Potential Security Vulnerabilities with a Static Analysis Tool. *IEEE Transactions on Software Engineering* 45 (2019), 877–897. <https://api.semanticscholar.org/CorpusID:64711495>
- [60] Linxin Song, Jieyu Zhang, Tianxiang Yang, and Masayuki Goto. 2022. Adaptive Ranking-based Sample Selection for Weakly Supervised Class-imbalanced Text Classification. In *Conference on Empirical Methods in Natural Language Processing*.
- [61] Yanqi Su, Zhenchang Xing, Xin Peng, Xin Xia, Chong Wang, Xiwei Xu, and Liming Zhu. 2021. Reducing bug triaging confusion by learning from mistakes with a bug tossing knowledge graph. In *2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 191–202.
- [62] Mohammad Tahaei, Kami Vaniea, Konstantin Beznosov, and Maria Klara Wolters. 2021. Security Notifications in Static Analysis Tools: Developers' Attitudes, Comprehension, and Ability to Act on Them. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021). <https://api.semanticscholar.org/CorpusID:233987670>
- [63] Haoxin Tu, Lingxiao Jiang, Debin Gao, and He Jiang. 2024. Beyond a joke: Dead code elimination can delete live code. In *Proceedings of the 2024 ACM/IEEE 44th International Conference on Software Engineering: New Ideas and Emerging Results*. 32–36.
- [64] Akshay Utture, Shuyang Liu, Christian Gram Kalhauge, and Jens Palsberg. 2022. Striking a Balance: Pruning False-Positives from Static Call Graphs. *2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE)* (2022), 2043–2055.
- [65] Haoye Wang, Zhipeng Gao, Tingting Bi, John Grundy, Xinyu Wang, Minghui Wu, and Xiaohu Yang. 2024. What makes a good todo comment? *ACM Transactions on Software Engineering and Methodology* 33, 6 (2024), 1–30.
- [66] Haoye Wang, Zhipeng Gao, Xing Hu, David Lo, John Grundy, and Xinyu Wang. 2024. Just-in-time todo-missed commits detection. *IEEE Transactions on Software Engineering* 50, 11 (2024), 2732–2752.
- [67] Junjie Wang, Song Wang, and Qing Wang. 2018. Is there a "golden" feature set for static warning identification? an experimental evaluation. In *Proceedings of the 12th ACM/IEEE international symposium on empirical software engineering and measurement*. 1–10.
- [68] Chadd C. Williams and Jeffrey K. Hollingsworth. 2005. Automatic mining of source code repositories to improve bug finding techniques. *IEEE Transactions on Software Engineering* 31 (2005), 466–480.
- [69] Zhipeng Xue, Zhipeng Gao, Xing Hu, and Shanping Li. 2023. ACWRRecommender: A Tool for Validating Actionable Warnings with Weak Supervision. In *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 1876–1880.
- [70] Zhipeng Xue, Zhipeng Gao, Shaohua Wang, Xing Hu, Xin Xia, and Shanping Li. 2024. Selfpico: Self-guided partial code execution with llms. In *Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis*. 1389–1401.
- [71] Zhipeng Xue, Xiaoting Zhang, Zhipeng Gao, Xing Hu, Shan Gao, Xin Xia, and Shanping Li. 2025. Clean Code, Better Models: Enhancing LLM Performance with Smell-Cleaned Dataset. *arXiv preprint arXiv:2508.11958* (2025).
- [72] Dapeng Yan, Zhipeng Gao, and Zhiming Liu. 2023. A closer look at different difficulty levels code generation abilities of chatgpt. In *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 1887–1898.
- [73] Xueqi Yang, Jianfeng Chen, Rahul Yedida, Zhe Yu, and Tim Menzies. 2021. Learning to recognize actionable static code warnings (is intrinsically easy). *Empirical Software Engineering* 26 (2021), 1–24.
- [74] Xueqi Yang, Jianfeng Chen, Rahul Yedida, Zhe Yu, and Tim Menzies. 2021. Learning to recognize actionable static code warnings (is intrinsically easy). *Empirical Software Engineering* 26 (2021).
- [75] Xueqi Yang, Zhe Yu, Junjie Wang, and Tim Menzies. 2021. Understanding static code warnings: An incremental AI approach. *Expert Systems with Applications* 167 (2021), 114134.
- [76] Yanming Yang, Xing Hu, Zhipeng Gao, Jinfu Chen, Chao Ni, Xin Xia, and David Lo. 2024. Federated learning for software engineering: A case study of code clone detection and defect prediction. *IEEE Transactions on Software Engineering* 50, 2 (2024), 296–321.
- [77] Ulas Yuksel and Hasan Sözer. 2013. Automated Classification of Static Code Analysis Alerts: A Case Study. *2013 IEEE International Conference on Software Maintenance* (2013), 532–535.
- [78] Ulas Yuksel and Hasan Sözer. 2013. Automated classification of static code analysis alerts: A case study. In *2013 IEEE International conference on software maintenance*. IEEE, 532–535.
- [79] Yuwei Zhang, Ying Xing, Yun zhan Gong, Dahai Jin, Honghui Li, and Feng Liu. 2019. A variable-level automated defect identification model based on machine learning. *Soft Computing* 24 (2019), 1045–1061.