

Improving Compiler Bug Isolation by Leveraging Large Language Models

Yixian Qi

Tianjin University

Tianjin, China

qiyixian@tju.edu.cn

Bowen Chen

Tianjin University

Tianjin, China

3022244160@tju.edu.cn

Jiajun Jiang

Tianjin University

Tianjin, China

jiangjiajun@tju.edu.cn

Hongyu Zhang

Chongqing University

Chongqing, China

hongyujohn@gmail.com

Fengjie Li

Tianjin University

Tianjin, China

fengjie@tju.edu.cn

Junjie Chen

Tianjin University

Tianjin, China

junjiechen@tju.edu.cn

Abstract

Compilers play a foundational role in building reliable software systems, and bugs within them can lead to catastrophic consequences. The compilation process typically involves hundreds of files, making traditional automated bug isolation techniques inapplicable due to scalability or effectiveness issues. Current mainstream compiler bug localization techniques have limitations in test program mutation and resource consumption. Inspired by the recent advances of pre-trained Large Language Models (LLMs), we propose an innovative approach named AutoCBI, which (1) uses LLMs to summarize compiler file functions and (2) employs specialized prompts to guide LLM in reordering suspicious file rankings. This approach leverages four types of information: the failing test program, source file function summaries, lists of suspicious files identified through analyzing test coverage, as well as compilation configurations with related output messages, resulting in a refined ranking of suspicious files. Our evaluation of AutoCBI against state-of-the-art approaches (DiWi, RecBi and FuseFL) on 120 real-world bugs from the widely-used GCC and LLVM compilers demonstrates its effectiveness. Specifically, AutoCBI isolates 66.67%/69.23%, 300%/340%, and 100%/57.14% more bugs than RecBi, DiWi, and FuseFL, respectively, in the Top-1 ranked results for GCC/LLVM. Additionally, the ablation study underscores the significance of each component in our approach.

Keywords

Compiler Debugging, Bug Isolation, Large Language Models

1 Introduction

Compiler bugs present critical challenges in software systems, potentially leading to catastrophic consequences [10] such as system crashes [35], erroneous program behaviors [38], and security vulnerabilities [14, 34]. Ensuring the robustness and reliability of compilers is crucial, as they are foundational tools that take the responsibility to translate high-level source code into low-level executable programs. Compiler bugs may propagate to application programs compiled by the compilers due to incorrect translations, which will seriously aggravate the difficulty of the downstream program debugging process since it is hard to determine whether the unexpected behaviors in application programs are from the used compilers or the program themselves. Consequently, the identification and debugging of compiler bugs are vital tasks in maintaining software integrity and reliability.

To tackle this issue, a variety of compiler testing methods have been proposed and developed in the literature [3, 13, 26, 39, 47], aiming at uncovering potential bugs by generating diverse and comprehensive test cases. These approaches leverage techniques such as random testing [15], fuzz testing [17] and deep learning [25], which have shown effectiveness in discovering subtle and critical flaws in compilers. Empirical studies have demonstrated the efficacy of these methods in identifying bugs across widely used compilers (i.e., GCC [16] and LLVM [28]). However, despite these advancements, the debugging process following the identification of test failures remains time-consuming and labor-intensive. The challenge is underpinned by two primary factors: First, due to the inherent complexity of compilers, failing test programs often involve numerous interdependent files, all of which are potential sources of failures. As a result, the intricate interrelation of compiler components complicates the isolation of the root cause, often requiring significant manual effort to localize the actual responsible source files of the failure. Second, existing test programs are frequently too weak to distinctly separate faulty files from non-faulty ones, resulting in reduced effectiveness of existing coverage-based fault localization techniques [6, 29, 43, 47, 48].

To tackle these challenges, existing compiler bug isolation approaches enhance test coverage differentiation through two main techniques. The first technique involves mutating compiler configurations to achieve varied coverage of compiler files, aiding in the identification of bugs [45, 51]. However, this method can be too coarse-grained, as an individual optimization may interact with multiple source files, complicating the accurate pinpointing of specific buggy locations. The second technique focuses on generating additional passing test programs that share similar execution paths with failing tests, enabling the isolation of faulty files through coverage analysis [8, 9, 36]. Nonetheless, the complexities of compiler programs and their intricate execution path constraints make it difficult to produce effective passing test cases with limited mutations [8, 9]. Although these methods are effective in isolating compiler bugs, they remain constrained by their dependence on test coverage alone.

Additionally, existing approaches often fail to fully utilize the insights developers gain during manual debugging, such as the characteristics of failing test programs, error messages, and compiler semantics. Developers typically draw on their deep understanding of code functions, synthesizing information from various sources to

correlate test features, observed failures, and relevant code components [21]. This informed approach helps narrow the search space, allowing for more effective identification and resolution of bugs. Unfortunately, current compiler bug isolation methods overlooked these insights, which may hinder their effectiveness. While information retrieval (IR)-based techniques [50] have proven successful in fault localization, their reliance on high-quality bug reports [37], which are often lacking for compiler bugs [32], limits their applicability and effectiveness in this context.

Inspired by the recent advances of large language models (LLMs), which have demonstrated remarkable effectiveness in various software engineering [20, 22, 49] and natural language processing tasks [30], this paper proposes an innovative compiler bug isolation approach, named AutoCBI, which aims at improving the compiler bug isolation performance by leveraging LLMs’ capabilities in both natural language and code understanding for integrating additional contextual information into the debugging process. This integration provides us a new way to significantly improve the performance of compiler bug isolation by considering a broader spectrum of data, including both code and documents, which are critical for accurate diagnosis. Specifically, AutoCBI currently integrates four types of information, i.e., the failing test program, function summaries of implicated source files, lists of suspicious files identified through analyzing test coverage, as well as compilation configurations and related output messages with a specially designed prompt.

To assess the effectiveness of AutoCBI, we have conducted an extensive experiment using a widely-used dataset consisting of 120 real-world reproducible bugs from two prominent compilers, i.e., GCC [16] and LLVM [28]. We compare AutoCBI with state-of-the-art baselines. The experimental results demonstrated that AutoCBI successfully localized the faulty files within the top 1/5/10/20 positions for 42/63/78/95 bugs, significantly outperforming the best-performing baseline (i.e., RecBi) with improvements of at least 68.00%/21.15%/9.86%/6.74%, thereby highlighting the effectiveness of our approach. Additionally, our ablation study indicated that each type of information utilized in AutoCBI significantly contributes to its effectiveness. Interestingly, the results show that not all types of information are necessary for localizing certain bugs, and in some cases, the inclusion of certain information can even negatively impact the localization performance. This finding suggests the need for further investigation and the development of adaptive methods for combining various information optimally.

In summary, this work makes the following major contributions:

- This work provides a new way of improving compiler bug isolation by integrating data from multiple sources beyond the coverage information of test execution.
- This work introduces the first compiler bug isolation method utilizing LLM to incorporate various context features (named AutoCBI).
- We have conducted extensive experiments to evaluate the performance of our method against baseline approaches, highlighting its superiority and the significance of each component within our method.
- We have open-sourced all our implementations and experimental results to facilitate replication and support future research. <https://anonymous.4open.science/r/AutoCBI-D823/>

```

1 int a = 1;
2 int b;
3 int d;
4 short e;
5
6 int
7 main ()
8 {
9     for ( ; b; b++)
10    {
11        ;
12    }
13    short f = a;
14    int g = 15;
15    e = f ? f : 1 << g;
16    int h = e;
17    d = h == 83647 ? 0 : h;
18    if (d != 1)
19    {
20        __builtin_abort ();
21    }
22    return 0;
23 }
```

(a) Failing test program.

```

1 int a = 1;
2 int b;
3 short int d;
4 short e;
5
6 int
7 short f = a;
8 int g = 15;
9 e = f ? f : 1 << f;
10 int h = e;
11 d = h == 83647 ? 0 : h;
12 if (d != 1)
13 {
14     __builtin_abort ();
15 }
16 short f = a;
17 int g = 15;
18 e = f ? f : 0 << g;
19 int h = e;
20 d = h == 83647 ? 0 : h;
21 if (d != 1)
22 {
23     __builtin_abort ();
}
```

(b) Three passing test programs generated by RecBi.

Figure 1: The failing test program triggering GCC bug (ID: 59221 [1]) and passing test programs by mutating the failing one. The mutated lines are highlighted.

2 Motivation

In this section, we use an example to illustrate the limitations of existing compiler bug isolation methods and motivate our approach. Fig. 1 presents an example from the GCC compiler (Bug ID: 59221 [1]), including the reported bug-triggering test program. The bug is located in the source file of “tree-ssa-threadupdate.c”. In this case, when the compiler (revision 205097 in GCC trunk) compiles the test program, an issue with the Value Range Propagation (VRP) optimization leads to incorrect code generation at optimization levels -O2 and -O3 on x86_64-linux-gnu systems. As introduced, existing approaches perform bug isolation by either mutating failing test programs to generate passing ones or mutating the compiling optimization configurations.

Bug isolation by mutating test programs: The complexity of compiler programs makes generating suitable passing test programs challenging, undermining fault localization effectiveness. For example, in this example (GCC bug ID: 59221 [1]), 277 source files are involved in compiling the failing test program. To isolate the faulty file, passing programs should cover the other 276 files, but intricate dependencies between them make this almost impossible. Mutated test programs are often too similar to the failing one, covering the same files and leading to identical execution failures. Specifically, RecBi [9], a state-of-the-art compiler bug isolation method, generated only three valid passing test programs in an hour (see Fig. 1b, partial test programs with mutations marked in blue). Existing methods typically produce very few passing test programs within a time limit, limiting their effectiveness and often resulting in inaccurate identification of buggy files, as shown by this example where the faulty file ranked outside the top 20 positions.

Bug isolation by mutating optimizations: Optimization configurations determine which components of a compiler execute during the compilation of a test program. Existing approaches modify

these options or their order in an effort to make a failing test program pass, bypassing the faulty files. As a result, files not covered by successful test programs become likely candidates for faults. However, inherent dependencies among source files complicate this bypassing. When successful, significant coverage differences typically emerge between failing and passing executions, making accurate isolation challenging. For instance, in the example shown in Fig. 1, bypassing optimizations related to VRP to make the failing test program pass involves multiple source files, any of which could be faulty.

Consequently, the performance of existing approaches remains limited due to the inherent complexity of compiler bug isolation. Importantly, these limitations are not isolated instances. Previous evaluations [9, 45, 51] indicate that state-of-the-art methods from these two categories can accurately localize at most 22.50% and 24.16% of compiler bugs within the top 1 position. Given thousands of bug reports related to GCC and LLVM compilers, there is an urgent need to develop more effective methods for compiler bug isolation due to the fundamental role of compilers.

Consider the typical debugging process undertaken by developers, who often draw upon a deep understanding of test failures and the functions of the involved code components. In these situations, developers may analyze the error messages generated by the test program and the characteristics exhibited by the failing test. This analysis guides them toward identifying the components that might produce such messages or are responsible for handling the corresponding code features. For the example depicted in Fig. 1, the control flow structure (associated with the for loop) causes the failure. Specifically, during Value Range Propagation (VRP) optimization at the -O2 and -O3 optimization levels, the VRP process mishandles the control flow pattern introduced by the for loop, leading to a compilation crash. By checking the code semantics developers may understand that the source file “tree-ssa-threadupdate.c” is closely related to the control flow updating as described by “*Thread edges through blocks and update the control flow and SSA graphs*” in the corresponding document of the source file. As a result, developers may have more confidence to believe that this file has a high possibility to be the root cause. However, existing approaches have never utilized such information for compiler bug isolation. Nevertheless, accurately understanding the code features causing the compilation failure and further correlating the failure with the source file are indeed challenging. To the best of our knowledge, we are the first to explore the capability of further incorporating such external information for improving compiler bug isolation by leveraging LLMs. Actually, our approach complements existing compiler bug isolation methods and provides a general framework that can combine existing methods (offering extra features) for further enhancement. As a result, the faulty file for the GCC compiler bug ([ID: 59221](#)) is accurately localized at top 1 by our approach.

3 The Design of AUTOBBI

This section outlines the detailed processes in our workflow and the design of the prompts, which aim to optimize compiler bug localization by integrating multi-source information with the capabilities of LLMs in both natural language and code understanding. Fig. 2 presents the overall workflow of our tool, AUTOBBI (Automated

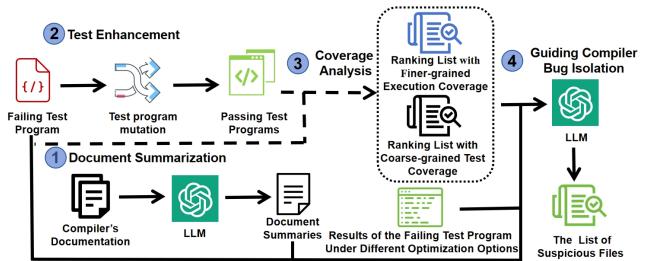


Figure 2: The framework of AUTOBBI

Compiler Bug Isolation), encompassing four core components. ① Document summarization – converts rich-text documents of source code into concise summaries, preserving key insights about code function to minimize noise. ② Test enhancement – generates a set of passing test programs to improve the distinguishing ability of test coverage; ③ Coverage analysis – effectively utilizes test coverage features to enhance fault localization; ④ Prompt engineering – develops specialized LLM prompts to integrate multi-source data for guiding compiler bug isolation. In the following, we will delve into the details of each component.

3.1 Document Summarization

As highlighted in Section 2, understanding the function of source code is crucial for developers to identify the root causes of test failures during debugging. Documents accompanying source files are pivotal in this comprehension process. Therefore, effectively leveraging these documents to enhance function understanding and improve bug isolation performance is essential.

However, automatically interpreting these documents is challenging, as they often include rich-text formats with natural language and code snippets, leading to lengthy descriptions. Given the constraints on LLM input length and the need to integrate multiple sources, it’s vital to refine these documents before processing them with LLMs. To address this, AUTOBBI implements a preprocessing step to refine the associated source code documents prior to bug isolation. Specifically, AUTOBBI leverages LLMs to convert the rich-text documents into summaries. For example, Fig. 3 shows the document excerpt of the source file “tree-ssa-threadupdate.c”, where the faulty code reside for the GCC compiler bug triggered by the failing test program shown in Fig. 1a. Specifically, the document often comprises two major information: (1) *source code*, including the declarations of header files, the definitions of data structures, macros, methods, and variables; (2) *natural language*, explaining the function of each method defined in the file. Consequently, directly feeding such rich-text formatted documents into LLMs, along with other data sources, can potentially exceed their maximum input length and introduce noise, hindering program semantic understanding. This underscores the need for a document summarization component in AUTOBBI. This component aims to generate precise summaries for each source file by emphasizing core insights and omitting irrelevant details.

Notably, this is a one-time process; the generated summaries can be reused unless the document is updated. Fig. 3 illustrates the prompt designed for producing these document summaries using LLMs (The text highlighted in the prompt will be replaced for

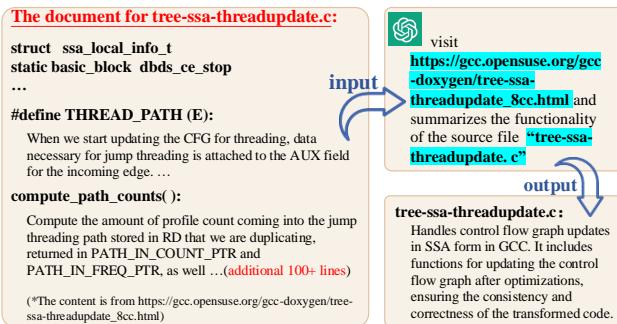


Figure 3: The simplified document for the source file “tree-ssa-threadupdate.c” from the GCC compiler (on the left) and its summary generated by ChatGPT-4o along with the used prompt (on the right).

different source files). Specifically, AutoCBI leverages ChatGPT-4o for generating document summaries, as it can automatically access a given website link and interpret its contents. More specifically, we first used a crawler to collect the document links to all the source files for the specified compilers from their official websites. Subsequently, we established mapping relationships between the files and their corresponding links, which will be supplied to the LLM for summary generation. In this figure, we also present the resulting summary for the given document (shown in the figure), demonstrating how it effectively captures the core function of the source file in natural language by removing extraneous information, such as code definitions. As explained in Section 2, the test failure is closely correlated with the control structure (i.e., the for loop), which is mishandled during performing the value range propagation optimization, and thus produce compilation failures. It is evident that the summary produced by the LLM shown in the figure can effectively facilitate the understanding of the code function, and thus improve the localization performance. In our experiment, the desired source file “tree-ssa-threadupdate.c” was successfully localized at the top 1 position by our approach, reflecting the importance of the summaries. In contrast, existing approaches failed to generate the effective passing test programs that help accurately isolate this faulty file from the others, making the faulty file localized outside the top 20 positions. In our final experiment, we will also evaluate the contribution of the document summarization component to the overall effectiveness of AutoCBI in an ablation study.

3.2 Test Enhancement

Given the large number of source files involved in a test failure, relying solely on the failing test program often fails to provide adequate insights into the cause of the failure. Studies have demonstrated that test coverage can effectively narrow down the search space for faulty files by assigning each file a likelihood of being faulty [33]. To enhance this process, AutoCBI includes a test enhancement component that aims to generate a set of passing test programs closely related to the failing one, thereby improving the discriminative ability of test programs.

Currently, AutoCBI utilizes RecBi [9] for generating passing test programs, as it has proven effective in producing high-quality test programs that significantly aid in compiler bug isolation through structure mutation operations. Importantly, the test enhancement component in AutoCBI is flexible and can be replaced with any established test program generation method, as it operates independently of the core methodology.

3.3 Coverage Analysis

It is well-known that existing coverage-based fault localization methods rely heavily on the quality of test programs. Accurately isolating faulty files is challenging, especially when many source files are involved in failing test executions. For instance, spectrum-based fault localization (SBFL) assigns suspiciousness scores to source files based on their execution across test programs. This task becomes even more difficult with a limited number of test programs, as seen in compiler bug isolation.

For instance, the widely-used SBFL method, Ochiai [4], calculates the likelihood of a source file being faulty using the formula presented in Formula 1. In this formula, $\text{failed}(f)$ and $\text{passed}(f)$ represent the numbers of failing and passing test programs covering the source file f , while totalfailed is the total number of failed test programs. Thus, Ochiai uses this formula to assess the fault probability for each file f . Consequently, the effectiveness of fault localization significantly depends on the quantity and quality of the test programs.

$$\text{sus}(f) = \frac{\text{failed}(f)}{\sqrt{\text{totalfailed} \times (\text{failed}(f) + \text{passed}(f))}} \quad (1)$$

Consider the GCC bug (ID: 59221) as an example, where only three passing test programs were generated even with the state-of-the-art test enhancement method, RecBi [9]. Notably, all passing test programs covered the faulty file, resulting in a suspicious score of $1/\sqrt{1 \times (1+3)} = 0.5$, ranking it 23rd on the list of suspicious files. In compiler bug isolation, it is common for totalfailed and $\text{failed}(f)$ to be 1, since typically there is only one failing test program per bug. Thus, the performance of traditional SBFL methods is heavily dependent on the number of passing test programs, which remains a challenge as previously mentioned. Given the limited number of test programs and the multitude of source files involved in the failing execution, exploring new methods to more effectively utilize test programs is crucial. Our analysis reveals that, although passing test programs often cover the same source files as the failing test program, the number of executions varies due to recursive method calls (i.e., one source file could be executed multiple times by one test program). This observation encourages us to explore the potential of using finer-grained execution coverage, as opposed to traditional test coverage used by SBFL, to enhance fault localization in compiler bug isolation scenarios.

Preliminary Study: To investigate the fault localization performance using finer-grained execution coverage, we conducted a preliminary study employing the same SBFL formula but incorporating this detailed level of coverage. We selected the Ochiai formula as a representative due to its widespread use in previous studies and proven effectiveness. In this context, $\text{failed}(f)$ and $\text{passed}(f)$ denote the number of failing and passing test executions that cover the

source file f , respectively, while totalfailed represents the number of failing test executions involving the file. Our experiment used a widely-referenced compiler bug dataset [8, 9], containing 120 compiler bugs from GCC and LLVM. We compared these results with those obtained from traditional SBFL methods. Fig. 4 shows the analysis results, focusing on bugs that could be localized within the top 40 positions for clarity and improved readability. In the figure, the x -axis represents the positions of faulty files in the localization results using original coarse-grained test coverage, while the y -axis indicates the positions using finer-grained execution coverage. Each point in the figure corresponds to a specific compiler bug from either GCC or LLVM.

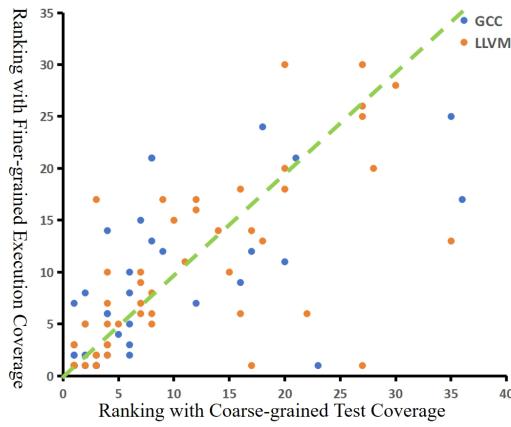


Figure 4: Fault localization results by Ochiai when using execution coverage against test coverage.

The results indicate that finer-grained execution coverage can indeed enhance fault localization accuracy for some compiler bugs—It improved the localization results for 47/120 compiler bugs compared to coarse-grained test coverage. For instance, in the GCC bug ([ID: 59221 \[1\]](#)), the target faulty file “tree-saa-threadupdate.c” rose from the 23rd position with coarse-grained test coverage to the 1st position with finer-grained execution coverage. However, this approach can also reverse results for other bugs. Notably, finer-grained execution coverage worsened fault localization outcomes for 34/120 compiler bugs. For example, the faulty file “DAGCombiner.cpp” for the LLVM bug ([ID: 25154 \[2\]](#)) dropped from 1st to 10th position when using finer-grained coverage. Hence, both coverage types have their strengths and weaknesses, and they complement each other. Based on this conclusion, AutoCBI integrates both types of coverage information to enhance fault localization performance. AutoCBI computes suspicious scores for each source file with both coverage types, generating two ranking lists of potential faulty files, which are then fed into LLMs for integration and further refinement together with comprehensive information, such as aforementioned document summaries.

3.4 Prompt Engineering

To leverage LLMs for integrating comprehensive information in compiler bug isolation, we have crafted a specialized prompt. As

motivated in the Introduction, AutoCBI currently incorporates four types of information: document summaries of source files involved in the failing test execution, suspicious file lists calculated by both coarse-grained test coverage and finer-grained execution coverage, the failing test program, and the compilation configuration along with related output messages.

The prompt is designed to enable LLMs to understand the file function and assist in re-ranking suspicious files using this comprehensive information, leading to a more precise file ranking. This step is central to the entire AutoCBI workflow, leveraging the powerful reasoning and contextual comprehension capabilities of LLMs. By providing detailed prompt information, the model can make nuanced assessments of suspicious files within various contexts. For example, by analyzing test outputs that trigger the fault under different optimization configurations, the LLM may identify the root cause of the failure and how these optimization strategies impact the program’s execution, further narrowing down the range of potential faulty files. Meanwhile, source file summaries can help LLMs refine the ranking of suspicious files by correlating the root cause with the functions of the source files.

```
[source-code-start]
//Source code of the failing test
program
[source-code-end]
[rankfile-start]
//Suspicious file list generated
by traditional SBFL
[rankfile-end]
[result-start]
//Compilation outputs under
different compilation
configurations
[result-end]
[executed-file-start]
//Suspicious file list generated
by finer-grained execution
coverage
[executed-file-end]
```

Reorder these suspicious files between [rankfile-start] and [rankfile-end] according to the functional description of the corresponding [\[GCC/LLVM\]](#) source code, code features of the source code between [source-code-start] and [source-code-end], the list of the executed files and the result of the source code running under different optimization options. According to the degree of suspicion of the defect, 0 represents the lowest degree of suspicion, and 10 represents the highest degree of suspicion. Assign a score to each file between [rankfile-start] and [rankfile-end], add the score to the end of each line. The files will be sorted in descending order according to the score.

Figure 5: The prompt used by AutoCBI for integrating multiple data sources.

Specifically, we first input the compiler source file summaries into the LLM to ensure the model can understand the specific role of each file in the compiler. This step is crucial because a compiler is a complex system with hundreds of files, each serving unique functions. By providing these function summaries, the model can better understand which files play critical roles in the compilation process, thus more effectively evaluating each file’s potential relevance and suspiciousness in the fault localization task. The prompt for providing document summaries is structured as follows: “[summary-start]...[summary-end]\ The functional description between [summary-start] and [summary-end] is related to the source code files of [\[GCC/LLVM\]](#). Learn the functions of these files and complete the following tasks based on the functions of these files”. Each file’s summary is organized as illustrated in Fig. 3. Building on this, the detailed prompt is shown in Fig. 5. It is structured into five parts,

providing the model with rich contextual information, helping the LLM combine clues from different sources to localize the bug. The specific prompt content is as follows:

- (1) **Source code of the failing test program.** We first provide the source code of the failing test program that triggers the compilation failure. This information helps the LLM identify the code features in the test program that cause the abnormal behavior or failure during the compilation process, serving as the starting point for subsequent fault localization. In the prompt, we use **[source-code-start]** and **[source-code-end]** to mark the beginning and end of the failing test program information.
- (2) **Suspicious file list generated by traditional SBFL.** After enhancing the test programs, AUTO-CBI leverages traditional SBFL techniques to generate a list of suspicious files. This list serves as the target for refinement by the LLM, which synthesizes the insights from various contextual sources to improve localization accuracy. In the prompt, we use **[rankfile-start]** and **[rankfile-end]** to mark the beginning and end of the file list.
- (3) **Compilation outputs under different compilation configurations.** Compiler optimization options can significantly influence the output of test programs. To assist in this analysis, we provide the outputs of the failing test program under different configurations. These results enable the LLM to better understand the abnormal behavior observed during compilation and to assess how various optimization strategies impact the code execution path. This, in turn, helps in narrowing down the range of potentially faulty files. In the prompt, we use **[result-start]** and **[result-end]** to mark the beginning and end of the compilation output under different configurations.
- (4) **Suspicious file list generated by finer-grained execution coverage.** As demonstrated in the preliminary study presented in Section 3.3, finer-grained execution coverage information complements the coarse-grained test coverage employed by traditional SBFL methods. Consequently, we also provide the LLM with a list of suspicious files derived from finer-grained coverage data. The expectation is that the LLM will further assess the relationship between these files and the compiler fault based on their execution frequency and coverage. Files with higher coverage may have a higher likelihood of being likely to the bug. In the prompt, we use **[executed-file-start]** and **[executed-file-end]** to mark the beginning and end of the file list calculated by utilizing the finer-grained execution coverage.
- (5) **Task description.** Finally, we offer a task description that outlines the purpose of each piece of information provided, aiding the LLM in comprehending the components of the prompt and their relevance to the fault localization process. This description includes detailed explanations for each data segment, enabling the model to effectively integrate and analyze data from multiple sources, such as the failing test program, the list of suspicious files, and compilation results. The task description is localized at the end of the prompt as illustrated in Fig. 5.

Table 1: SBFL formulas using different coverage.

Name	Formula	
	w/ Test Coverage	w/ Execution Coverage
Wong2	$\frac{\text{failed}(f) - \text{passed}(f)}{\text{failed}(f)}$	$\frac{C_f(f) - C_p(f)}{C_f(f)}$
Ochiai	$\frac{\sqrt{\text{total failed} \times (\text{failed}(f) + \text{passed}(f))}}{\text{failed}(f)^2}$	$\frac{\sqrt{C_f(f) \times (C_f(f) + C_p(f))}}{C_f(f)^2}$
DStar2	$\frac{\text{passed}(f) + (\text{total failed} - \text{failed}(f))}{\text{passed}(f) + (\text{total failed} - \text{failed}(f))}$	$\frac{C_p(f)}{C_p(f) + C_f(f)}$
Barinel	$1 - \frac{\text{passed}(f) - \text{failed}(f)}{\text{failed}(f) / \text{total failed}}$	$1 - \frac{C_p(f) + C_f(f)}{C_p(f) + C_f(f)}$
Tarantula	$\frac{\text{failed}(f) / \text{total failed} + \text{passed}(f) / \text{total passed}}{\text{failed}(f) / \text{total failed} + \text{passed}(f) / \text{total passed}}$	$\frac{C_p(f) + C_f(f)}{C_p(f) + C_f(f)}$

* $C_f(f)$: execution times of source file f in the failing/passing test runs;

* $C_p(f)$: average execution times among passing test programs that cover source file f ;

4 Evaluation

4.1 Research Questions

- **RQ1:** How effective is AUTO-CBI in compiler bug isolation?
- **RQ2:** What is the impact of different types of information on AUTO-CBI’s performance?
- **RQ3:** What is the impact of different SBFL formulas and LLMs used in AUTO-CBI on its performance?

4.2 Implementation and Configuration

To assess the performance of our approach, we developed an automated compiler bug isolation tool, named AUTO-CBI. Specifically, AUTO-CBI integrates the test generation capabilities of the state-of-the-art RecBi [9] for test enhancement, which has been shown to be effective by using structural mutations for compiler test program generation. Notably, our approach can be combined with any test generation technique since it functions independently of the test enhancement component.

For the suspicious file list based on coarse-grained test coverage, we adopted the default SBFL formula (i.e., Ochiai) used by RecBi [9], as it has performed well in previous studies. In contrast, for the list based on finer-grained execution coverage, we employed the SBFL formula Wong2 [42] by default, which achieved slightly better fault localization results. Additionally, we empirically explore the impact of different formulas on localization results for both coverage types in the final evaluation, and the results show that the formulas have marginal influence to the effectiveness of AUTO-CBI. Table 1 presents the formulas that will be examined. In particular, each formula has two versions, corresponding to the use of different coverage types: test coverage and execution coverage as illustrated in Table 1.

Furthermore, AUTO-CBI utilizes ChatGPT-4o [31] as the default LLM for generating document summaries and making localization inference. This ensures that the model can effectively respond to prompts and perform complex reasoning based on multi-source input data, thereby improving the accuracy of bug localization.

4.3 Benchmark

To ensure the comprehensiveness and representativeness of our study, we utilized a commonly-used benchmark dataset containing 120 real-world compiler bugs, in line with previous research [8, 9, 36, 45, 51]. This dataset includes an equal distribution of 60 GCC [16] bugs and 60 LLVM [28] bugs, covering various types of issues within these compilers. Specifically, each bug linked to the specific compiler version that caused it, along with the failing test programs and compilation configurations that trigger the bug, as well as the actual files responsible for the bug (serving as the

ground truth for our compiler faulty file localization task). This data provides a robust foundation for evaluating the effectiveness of AUTO-CBI.

4.4 Evaluation Metrics

To evaluate the performance of AUTO-CBI in compiler bug isolation, we employed several commonly used metrics to measure fault localization effectiveness [8, 9, 36, 45, 51]. The specific evaluation metrics are detailed below. Additionally, we will also discuss the **cost** of using LLMs in our evaluation.

- **Top-N:** This metric measures the number of bugs successfully isolated among the top N positions in the suspicious file list, with N values of 1, 5, 10, and 20. A higher Top-N value indicates better performance, reflecting that more actual faulty files are found within the top N, thus improving bug isolation efficiency.
- **Mean First Rank (MFR):** This metric reflects the average position of the first identified faulty file in the suspicious file list. A lower MFR indicates quicker identification of the first faulty file, helping developers narrow their debugging scope. Thus, a smaller MFR value signifies better performance by finding faulty files earlier, reducing debugging time and effort.
- **Mean Average Rank (MAR):** When a bug involves multiple faulty files, MAR indicates the average position of all these files in the suspicious file list. For multiple bugs, it averages these values across all cases. A smaller MAR is preferable, as it signifies better accuracy in locating all faulty files, not just the first, and suggests consistent performance when addressing multiple bugs.

4.5 Baselines

To provide a comprehensive evaluation of our approach’s performance, we aimed to compare it with all state-of-the-art compiler bug isolation methods. However, because some existing methods have not open-sourced their implementations and we were unable to replicate their results based on the original papers (including ODFL [45], LocSeq [51], and LLM4CBI [36]), we ultimately selected RecBi [9], DiWi [8] and FuseFL [40] (an LLM-based method) – three widely used state-of-the-art baselines in previous studies – as the baselines for our evaluation. Specifically, we perform fault localization at the file level by following existing studies [8, 9, 36, 45, 51]. To ensure a fair comparison, we adhered strictly to their experimental settings when running their open-source implementations. Each tool was executed in five iterations for each bug, selecting the best-performing results to replicate the outcomes reported in their original papers. Particularly, we adopted the same ChatGPT-4o (more powerful than ChatGPT-3.5 originally used by FuseFL) in FuseFL for a fair comparison.

In addition, to evaluate the **contribution of each component** to the effectiveness of our approach, we also developed a set of variants of AUTO-CBI for comparison. We use AUTO-CBI-*src* to represent the variant of AUTO-CBI by removing the component of *src* from it. In particular, $src \in \{summary, execov, testcov, compile, llm, failtest\}$, which respectively represents the document summary, the suspicious file list derived from finer-grained execution coverage, the suspicious file list derived from coarse-grained test coverage, the compilation results of failing test programs, the large language

model (i.e., relying solely on finer-grained execution coverage after removing LLM), and the failing test program.

Furthermore, we assess the **impact of different formulas** on the effectiveness of AUTO-CBI by creating another two sets of variants that use different formulas for calculating suspicious scores. Specifically, AUTO-CBI+*FORMULA*[*exec*] utilizes *FORMULA* for finer-grained execution coverage, while AUTO-CBI+*FORMULA*[*test*] applies it for coarse-grained coverage. The formulas considered include Ochiai [4], Wong2 [42], DStar2 [41], Barinel [5], and Tarantula [23], with the default configurations being AUTO-CBI+Wong2[*exec*] and AUTO-CBI+Ochiai[*test*]. Finally, to evaluate the **impact of different LLMs** on the effectiveness of our approach, we developed another two variants of AUTO-CBI, i.e., AUTO-CBI+*deepseekv3* and AUTO-CBI+*deepseekr1*, which represent replacing the default LLM ChatGPT-4o in AUTO-CBI with DeepSeek-v3-671B [11] and DeepSeek-R1-671B [12], respectively.

By employing the above variants for comparison in our experiment, we can comprehensively evaluate the contribution of each component and their influence. This may also provide insights for further refinement in future research.

5 Results and Analysis

5.1 Overall Effectiveness of AUTO-CBI

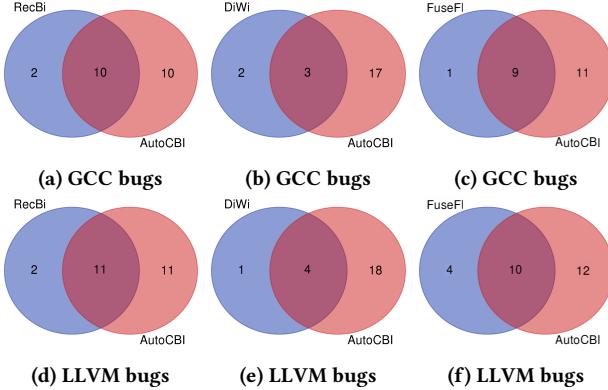
Table 2 presents the experimental results of our approach compared to baseline methods on the benchmark dataset. The findings demonstrate that AUTO-CBI significantly outperforms state-of-the-art baselines across all metrics. Specifically, AUTO-CBI successfully localized 20, 33, 39, and 45 GCC compiler bugs within the Top-1, Top-5, Top-10, and Top-20 positions, respectively, on the suspicious file list refined by the LLM after integrating multiple data sources. It improves over RecBi by 66.67%, 26.92%, 8.33%, and 9.76% in these rankings, indicating AUTO-CBI’s effectiveness in narrowing down the scope of faulty files, particularly excelling in the Top-1 and Top-5 positions. Notably, Top-1 and Top-5 are crucial metrics for localization techniques, as developers typically prefer to examine no more than five suggested locations provided by automated methods, as highlighted in previous studies [24]. Similarly, AUTO-CBI localized 22, 30, 39, and 50 LLVM compiler bugs within the Top-1, Top-5, Top-10, and Top-20 positions, achieving improvements of 69.23%, 15.38%, 11.43%, and 4.17% over RecBi. Compared to DiWi and FuseFL, the improvements in Top-1 accuracy range from 300% to 340% and from 57.14% to 100%. These results further demonstrate AUTO-CBI’s effectiveness and highlight the generality of our approach for localizing bugs across various compilers.

We also conducted a detailed analysis of the bugs that each method accurately localized within the Top-1 position and examined the overlaps among them. Fig. 6 illustrates the results. From the figure, we can observe that a substantial portion of the bugs localized by RecBi, DiWi and FuseFL can also be accurately localized by our approach, highlighting the remarkable performance of AUTO-CBI. One key reason for this is that AUTO-CBI incorporates the strengths of RecBi’s test program generation capabilities, which is a core aspect of RecBi. Furthermore, the results demonstrate that integrating multiple data sources significantly enhances the accuracy of compiler bug localization without largely diminishing the effectiveness of individual data sources.

Table 2: Overall effectiveness of AUTO-CBI compared with state-of-the-art baseline methods.

Subject	Approach	Top-1	\uparrow Top-1	Top-5	\uparrow Top-5	Top-10	\uparrow Top-10	Top-20	\uparrow Top-20	MFR	\uparrow MFR	MAR	\uparrow MAR
GCC	AUTO-CBI	20	/	33	/	39	/	45	/	7.71	/	9.22	/
	RecBi	12	66.67%	26	26.92%	36	8.33%	41	9.76%	13.69	43.68%	15.62	40.97%
	DiWi	5	300.00%	15	120.00%	26	50.00%	35	28.57%	20.77	62.88%	21.34	56.76%
	FuseFL	10	100.00%	22	50.00%	30	30.00%	37	21.62%	12.57	38.66%	14.49	36.37%
LLVM	AUTO-CBI	22	/	30	/	39	/	50	/	8.70	/	10.25	/
	RecBi	13	69.23%	26	15.38%	35	11.43%	48	4.17%	10.58	17.77%	12.1	15.29%
	DiWi	5	340.00%	19	57.89%	28	39.29%	38	31.58%	17.98	51.61%	18.84	45.59%
	FuseFL	14	57.14%	26	15.38%	33	18.18%	41	21.95%	13.31	34.64%	13.43	23.68%

*Columns " \uparrow " present the improvement rates of AUTO-CBI over a compared technique in terms of various measurements.

**Figure 6: Overlaps of compiler bugs localized at Top-1.**

The significant improvement can be attributed to AUTO-CBI's ability to integrate multiple data sources and utilize the robust reasoning and understanding capabilities of the LLM. This enables a more efficient analysis of the differences between failing and passing test executions, as well as the incorporation of contextual information like source file summaries. These additional inputs help the LLM more accurately assess the suspiciousness of each file, allowing for more accurate identification and localization of files most likely responsible for compiler bugs. For instance, as demonstrated in Section 2, the GCC compiler bug ([ID: 59221 \[1\]](#)) was initially localized at the 23rd position by the state-of-the-art RecBi. However, by utilizing finer-grained execution coverage, the faulty file was successfully localized at the 1st position. This allows the LLM to refine the results by considering all relevant factors.

Moreover, we also evaluated the efficiency and cost of our approach. On average, the bug isolation process takes approximately **275 seconds** per bug, excluding the test enhancement process, and the cost for bug isolation is about **\$ 0.075** per bug for using ChatGPT-4o. This overhead is relatively minor compared to the test generation and iterative testing processes employed by existing compiler bug isolation methods, and the cost is also relatively small. In summary, AUTO-CBI's performance demonstrates that integrating multiple data sources with the reasoning and understanding capabilities of large language models can substantially enhance the efficiency of compiler bug isolation. It not only excels in the GCC compiler but also achieves outstanding results in the LLVM

Table 3: AUTO-CBI's result when disabling each component.

Sub.	Approach	Top-1	\uparrow Top-1	Top-5	\uparrow Top-5	MFR	\uparrow MFR	MAR	\uparrow MAR
GCC	AUTO-CBI	20	/	33	/	7.71	/	9.22	/
	AUTO-CBI-summary	10	-50.00%	25	-24.24%	12.60	-63.42%	16.62	-80.26%
	AUTO-CBI-compile	9	-55.00%	26	-21.21%	13.41	-73.93%	17.2	-86.55%
	AUTO-CBI-excov	11	-45.00%	21	-36.36%	11.81	-53.18%	15.64	-69.63%
	AUTO-CBI-testcov	14	-30.00%	25	-24.24%	10.06	-30.48%	14.28	-54.88%
	AUTO-CBI-llm	14	-30.00%	23	-30.30%	12.08	-56.68%	16.06	-74.19%
LLVM	AUTO-CBI-failtest	9	-55.00%	21	-36.36%	12.56	-62.91%	16.60	-80.04%
	AUTO-CBI	22	/	30	/	8.70	/	10.25	/
	AUTO-CBI-summary	11	-50.00%	21	-30.00%	12.57	-44.48%	12.65	-23.41%
	AUTO-CBI-compile	14	-36.36%	24	-20.00%	10.82	-24.37%	11.53	-12.49%
	AUTO-CBI-excov	14	-36.36%	24	-20.00%	12.60	-44.83%	12.79	-24.78%
	AUTO-CBI-testcov	15	-31.82%	26	-13.33%	9.35	-7.47%	10.74	-4.78%
	AUTO-CBI-llm	16	-27.27%	25	-16.67%	9.73	-11.84%	11.12	-8.47%
	AUTO-CBI-failtest	10	-54.55%	21	-30.00%	12.80	-47.13%	13.05	-27.32%

*Columns " \uparrow " present the improvement rates of experiment variants over AUTO-CBI.

compiler. The more precise localization results can help significantly reduce debugging time, providing developers with a more efficient tool that advances intelligent debugging and maintenance of compilers.

5.2 The Contribution of Each Component

As mentioned in Section 4.5, to evaluate the contribution of each component in our approach, we have developed a set of variants of AUTO-CBI, and then we performed the same experiment with them and compared their results with the original AUTO-CBI. Table 3 presents these experimental results (We only report Top-1/5 to save space. The complete results can be found at [our project homepage](#)), along with the relative improvement of each variant compared to AUTO-CBI (negative numbers indicate performance degradation). The results indicate that all components of AUTO-CBI significantly enhance our approach's overall effectiveness. Specifically, removing one of the components in AUTO-CBI (i.e., source file summaries, compilation configurations and results, finer-grained execution coverage, coarse-grained test coverage, failing test program and the use of LLM) results in an average decrease of 50.00%, 45.24%, 40.48%, 30.95%, 54.76% and 28.57% in the number of bugs localized within the Top-1 position across the two compilers.

Specifically, without the inclusion of source file summaries or the failing test program, AUTO-CBI-summary and AUTO-CBI-failtest are able to localize only half (vs AUTO-CBI) of the compiler bugs within the Top-1 position, making the most significant decline in performance of AUTO-CBI compared with the other variants. This highlights that function summaries of source files are essential for

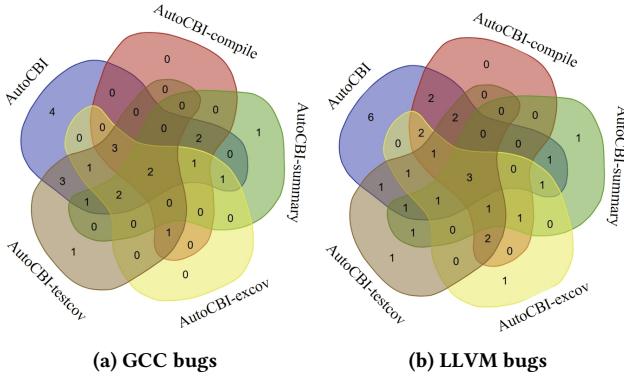


Figure 7: Overlaps of compiler bugs localized at Top-1 when using different variants of AutoCBI.

the LLM to comprehend the role and the function of each file within the compiler architecture. These summaries provide the LLM with critical background knowledge, enabling the model to not only rely on the surface-level coverage information but also assess a file’s importance and relevance from a functional perspective during the compilation process. This finding underscores that compiler file summaries have a pivotal impact on bug localization and can greatly assist the LLM in better identifying and isolating faulty files. Similarly, the failing test program offers the bug-triggering features, which are closely associated with a certain part of compiler implementations and thus can aid the LLM to effectively identify the most possibly faulty locations. Therefore, they should not be overlooked in the compiler bug localization process.

Additionally, the table reveals that when the use of LLM is removed and only finer-grained execution coverage is employed for fault localization, AutoCBI_{-llm} successfully localized 14 and 16 bugs within the Top-1 position across the two compilers, still outperforming the baseline methods. This underscores the importance of incorporating finer-grained coverage, particularly when working with a limited number of test programs for compiler bug isolation.

Finally, to investigate the complementariness of the data sources used in AutoCBI, we analyzed the overlaps of bugs that each source successfully localized within the Top-1 position. Fig. 7 depicts the findings, showing that different data sources complement one another, each contributing to the accurate localization of various compiler bugs. However, an intriguing phenomenon was also observed: adding certain data sources can potentially diminish localization performance for some bugs. For instance, removing finer-grained execution coverage led to five additional bugs being localized within the Top-1 position on the benchmark. This is consistent with our preliminary study in Section 3.3, where finer-grained execution coverage occasionally reduced fault localization accuracy. Taking the LLVM compiler bug (ID: 25154 [2]) as an example, the faulty file “DAGCombiner.cpp” was initially ranked 1st, but its rank dropped to 10th when using finer-grained execution coverage, which misled the LLM and hindered accurate localization. Similarly, other data sources might also provide inaccurate information affecting AutoCBI’s overall performance. For example, removing file summaries allowed five new bugs to be localized within the Top-1 position,

Table 4: AutoCBI’s result under different configurations.

Sub.	Approach	Top-1		Top-5		Top-5		MFR	↑MFR	MAR	↑MAR
		20	/	33	/	7.71	/				
GCC	AutoCBI	20	/	33	/	7.71	/	9.22	/		
	AUTO CBI+Ochiai[exec]	18	-10.00%	30	-9.09%	10.21	-32.43%	14.28	-54.88%		
	AUTO CBI+Barinel[exec]	18	-10.00%	28	-15.15%	9.75	-26.46%	11.76	-27.55%		
	AUTO CBI+DStar2[exec]	17	-15.00%	27	-18.18%	10.60	-37.48%	12.54	-36.01%		
	AUTO CBI+Tarantula[exec]	17	-15.00%	28	-15.15%	10.58	-37.22%	14.76	-60.09%		
	AUTO CBI+Wong2[test]	19	-5.00%	32	-3.03%	8.81	-14.27%	9.28	-0.56%		
	AUTO CBI+Barinel[test]	19	-5.00%	32	-3.03%	8.22	-6.61%	9.36	-1.52%		
	AUTO CBI+DStar2[test]	20	0.00%	32	-3.03%	8.33	-8.04%	9.72	-5.42%		
LLVM	AUTO CBI+Tarantula[test]	18	-10.00%	33	0.00%	8.08	-4.80%	9.64	-4.56%		
	AUTO CBI+deepseekv3	16	-20.00%	34	3.03%	10.76	-39.56%	12.13	-31.56%		
	AUTO CBI+deepseek1	16	-20.00%	35	6.06%	9.96	-29.18%	11.5	-24.73%		
	AutoCBI	22	/	30	/	8.70	/	10.25	/		
	AUTO CBI+Ochiai[exec]	15	-31.82%	27	-10.00%	9.04	-3.91%	9.88	3.61%		
	AUTO CBI+Barinel[exec]	18	-18.18%	30	0.00%	8.24	5.29%	8.29	19.12%		
	AUTO CBI+DStar2[exec]	19	-13.64%	29	-3.33%	9.15	-5.17%	9.16	10.63%		
	AUTO CBI+Tarantula[exec]	18	-18.18%	27	-10.00%	9.63	-10.69%	11.17	-8.98%		
LLVM	AUTO CBI+Wong2[test]	21	-4.55%	30	0.00%	8.72	-0.23%	10.28	-0.29%		
	AUTO CBI+Barinel[test]	22	0.00%	30	0.00%	8.46	2.76%	10.02	2.24%		
	AUTO CBI+DStar2[test]	21	-4.55%	30	0.00%	8.74	-0.46%	10.30	-0.49%		
	AUTO CBI+Tarantula[test]	21	-4.55%	29	-3.33%	8.63	0.80%	10.20	0.49%		
	AUTO CBI+deepseekv3	16	-27.27%	28	-6.67%	10.08	-15.86%	11.62	-13.37%		
	AUTO CBI+deepseek1	16	-27.27%	31	3.33%	9.67	-11.15%	11.05	-7.80%		

*Columns “↑” present the improvement rates of experiment variants over AutoCBI.

likely because the summaries were not informative enough or contained ambiguous descriptions that misled the LLM in correlating failures with source files.

In summary, removing one or more data sources contributed to nine additional bugs being localized within the Top-1 position, as shown in Fig. 7. This suggests that while multi-source information generally enhances model performance, the quality, accuracy, and completeness of this information are crucial for effective localization. In some cases, incomplete or biased information may lead to incorrect localization, and reducing such information can sometimes improve fault localization performance. This indicates the potential for developing an adaptive integration method that automatically determines the usefulness of each data source for fault localization. We plan to explore this approach in future work.

5.3 Performance under Different Configurations

As discussed in Section 4.2, the current implementation of AutoCBI uses the Ochiai and Wong2 formulas as default configurations to calculate suspicious file lists based on coarse-grained test coverage and finer-grained execution coverage, respectively. To examine the impact of these configurations, we developed two sets of AutoCBI variants, as outlined in Section 4.5, where each formula was replaced with an alternative one. Table 1 provides details on all the formulas studied in this experiment.

The experimental results are presented in Table 4, where we show the outcomes for each variant and its relative improvement over the default AutoCBI (equivalent to $\text{AUTO CBI+Wong2[exec]}$ and $\text{AUTO CBI+Ochiai[test]}$). Negative numbers indicate a degradation in fault localization performance. The results demonstrate that finer-grained execution coverage is more sensitive to the choice of formula than coarse-grained test coverage. Specifically, changing the formula for test coverage results in an average performance decline of 4.2% (i.e., one fewer bug localized) in terms of Top-1 accuracy, compared to a decline of up to 16.48% (i.e., about four fewer bugs localized) when altering the formula for execution coverage.

This is because the limited number of test programs and their similar execution paths tend to have a smaller impact on calculation results across different formulas. In contrast, the finer-grained execution coverage of different source files can vary more significantly than test coverage, making it more sensitive to coverage utilization methods. Despite this, the overall performance of AutoCBI with different formulas is not significantly affected, and it consistently outperforms baseline methods by a large margin.

We also analyzed the bugs successfully localized by AutoCBI under different configurations. The findings support a similar conclusion: different configurations (or formulas) complement each other, as each formula contributes to uniquely localizing different bugs in the Top-1 position. For instance, with finer-grained execution coverage, the Wong2 formula successfully localized five unique bugs that other formulas could not. Similarly, the Tarantula formula contributed to the unique localization of four additional bugs. This suggests that combining the strengths of various formulas, or developing more effective formulas tailored for finer-grained execution coverage (such as the formulas in Table 1), could further enhance the fault localization performance of AutoCBI.

Finally, to evaluate the performance of AutoCBI when utilizing different LLMs, we replaced the default ChatGPT-4o model with DeepSeek-V3 and DeepSeek-R1 as explained in Section 4.5. The results are presented in Table 4. The results demonstrate that their Top-5 performance is similar to ChatGPT-4o, while the Top-1 performance slightly decreased. Nevertheless, our approach always outperforms the baselines. In particular, the cost is much lower, with DeepSeek-V3 requiring only \$0.0045 per bug on average, and DeepSeek-R1 costing \$0.009 per bug.

6 Discussion

6.1 Threats to Validity

The threats to internal validity mainly arise from the implementation details of AutoCBI and its comparative methods – DiWi [8], RecBi [9], and FuseFL [40]. To address this threat, we thoroughly reviewed our implementation of AutoCBI. For the baseline methods, we directly used their open-source implementations and successfully replicated the results reported in their original papers. Additionally, we have open-sourced all our implementations to allow for further verification.

The threats to external validity are primarily related to the subjects and the LLMs used in our experiment. For the subjects, we adopted a widely-used benchmark dataset comprising 120 real-world bugs from two mature compilers, GCC and LLVM. While these bugs are diverse and representative, our approach may perform differently on entirely new datasets. Regarding the LLMs, We employed ChatGPT-4o as our primary LLM because of its strong inference capabilities. We also explored two other LLMs DeepSeek-V3 and DeepSeek-R1. The results indicate that while effectiveness varied slightly among LLMs, our approach consistently outperformed all baseline methods. Additionally, the potential for data leakage from LLM training involving similar compiler bugs could affect model performance and the fairness of our experiment. However, the distinct focus of our work compared to the training process helps mitigate this threat. Nevertheless, we plan to compile a new

dataset of recent compiler bugs to further evaluate our approach’s performance.

6.2 Future Work

Although our approach yielded promising results in the evaluation, we have identified several exciting directions that can be explored for further improving the performance of AutoCBI.

Interactive Bug Isolation: As discussed in Section 5.2, while multiple data sources enhance the overall effectiveness of AutoCBI, some may adversely impact fault localization for certain bugs. By interactively incorporating different data sources and using feedback to refine the combination strategy through methods like CoT (Chain-of-Thought) or ToT (Tree-of-Thought), AutoCBI can better leverage contextual information from different data sources, leading to further improvement.

Non-Prompt-Based Bug Isolation: The prompt provides the LLM with essential information, such as task description and context. An inaccurate prompt can lead to hallucination, and significantly impair bug isolation performance. Exploring methods that eliminate reliance on prompt engineering and instead allow the LLM to automatically request necessary information is promising. This approach offers the LLM greater flexibility, enabling it to actively acquire necessary supplementary data on demand during bug isolation. It reduces human intervention and can potentially further enhance automation and performance in complex scenarios.

7 Related Work

In this section, we discuss the research most relevant to our work. As mentioned in the introduction, existing compiler bug isolation techniques primarily fall into two categories: mutating test programs and mutating compilation configurations. In the first category, Chen et al. [8] introduced DiWi, which generates a set of passing test programs by applying mutation operations to the failing test program. Subsequently, Chen et al. [9] developed RecBi, which enhances DiWi by incorporating structural mutations and using reinforcement learning to guide the program generation process. Recently, Tu et al. [36] introduced LLM4CBI, a method that leverages LLMs to guide the generation of passing test programs for compiler bug isolation. For the second category, Yang et al. [45] introduced ODFL, which isolates compiler bugs by analyzing coverage differences of the failing test program across different compilation configurations (e.g., O2 vs O3). Similarly, Zhou et al. [51] proposed LocSeq, which targets compiler bugs caused by interactions in optimization sequences. It constructs optimization sequences that are similar to the bug-triggering sequence but do not trigger the bug, using coverage differences to identify buggy locations. Additionally, Zeller [46] introduced delta debugging, which isolates compiler bugs by examining runtime program states between passing and failing runs. Holmes and Groce [18, 19] proposed using mutation testing to locate compiler bugs; if two failing test programs for a compiler can be made to pass by applying the same mutation, they are likely due to the same fault. Chang et al. [7] suggested utilizing specially-designed type-based data-flow analysis to aid compiler bug debugging, while Lim and Debray [27] employed dynamic analysis techniques to identify errors in JIT compilers. Yang et al. [44]

proposed LLMAO, which leverages LLMs for localizing program bugs by taking source code as input while ignoring other features.

Unlike these approaches, our method seeks to enhance compiler bug isolation performance by leveraging LLMs to integrate multiple data sources, rather than relying solely on individual information such as test coverage or program states. Although the latest FuseFL [40] also leverages LLMs to combine multiple contextual features, the used features and method are still largely different. Our experimental results demonstrate that our approach significantly outperformed FuseFL. In fact, our approach complements these existing methods and can potentially be integrated with them by taking their results as extra input for additional improvements.

8 Conclusion

In this paper, we introduce AUTO-CBI, a novel compiler bug isolation method that leverages the power of LLMs to integrate multiple data sources, including failing test programs, source file summaries, lists of suspicious files identified through both coarse-grained and fine-grained test execution coverage, as well as compilation configurations with related output messages. This comprehensive contextual information guides the LLM in effectively localizing defects in compilers. Our experimental results using a widely-used benchmark dataset demonstrate that AUTO-CBI successfully isolated 42/63/78/95 out of 120 bugs within Top-1/5/10/20 positions. Remarkably, AUTO-CBI outperformed the best-performing RecBi [9] by isolating 66.67% and 26.92% more bugs regarding Top-1 and Top-5 accuracies for GCC, and 69.23% and 15.38% more bugs for LLVM. Furthermore, the results highlight the robustness of AUTO-CBI, which consistently maintains high effectiveness across compilers and configurations, surpassing existing advanced techniques.

References

- [1] 2013. GCC BUG ID 59221. https://gcc.gnu.org/bugzilla/show_bug.cgi?id=59221. Accessed: 2025.
- [2] 2015. LLVM BUG ID 25154. https://bugs.llvm.org/show_bug.cgi?id=25154. Accessed: 2025.
- [3] Rui Abreu, Peter Zoeteweij, and Arjan JC Van Gemund. 2007. On the accuracy of spectrum-based fault localization. In *Testing: Academic and industrial conference practice and research techniques-MUTATION (TAICPART-MUTATION 2007)*. IEEE, 89–98.
- [4] Rui Abreu, Peter Zoeteweij, and Arjan JC Van Gemund. 2007. On the accuracy of spectrum-based fault localization. In *Testing: Academic and industrial conference practice and research techniques-MUTATION (TAICPART-MUTATION 2007)*. IEEE, 89–98.
- [5] Rui Abreu, Peter Zoeteweij, and Arjan JC Van Gemund. 2009. Spectrum-based multiple fault localization. In *2009 IEEE/ACM International Conference on Automated Software Engineering*. IEEE, 88–99.
- [6] Samuel Benton, Xia Li, Yiling Lou, and Lingming Zhang. 2020. On the effectiveness of unified debugging: An extensive study on 16 program repair systems. In *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering*. 907–918.
- [7] Bor-Yuh Evan Chang, Adam Chlipala, George C Necula, and Robert R Schneck. 2005. Type-based verification of assembly language for compiler debugging. In *Proceedings of the 2005 ACM SIGPLAN international workshop on Types in languages design and implementation*. 91–102.
- [8] Junjie Chen, Jiaqi Han, Peiyi Sun, Lingming Zhang, Dan Hao, and Lu Zhang. 2019. Compiler bug isolation via effective witness test program generation. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 223–234.
- [9] Junjie Chen, Haoyang Ma, and Lingming Zhang. 2020. Enhanced compiler bug isolation via memoized search. In *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering*. 78–89.
- [10] Junjie Chen, Jibesh Patra, Michael Pradel, Yingfei Xiong, Hongyu Zhang, Dan Hao, and Lu Zhang. 2020. A survey of compiler testing. *ACM Computing Surveys (CSUR)* 53, 1 (2020), 1–36.
- [11] DeepSeek. 2024. DeepSeek-V3-671B. <https://api-docs.deepseek.com/zh-cn/news/news1226>. Accessed: 2025.
- [12] DeepSeek. 2025. DeepSeek-R1-671B. <https://api-docs.deepseek.com/zh-cn/news/news250120>. Accessed: 2025.
- [13] Nicholas DiGiuseppe and James A Jones. 2011. On the influence of multiple faults on coverage-based fault localization. In *Proceedings of the 2011 international symposium on software testing and analysis*. 210–220.
- [14] Vijay D'Silva, Mathias Payer, and Dawn Song. 2015. The correctness-security gap in compiler optimization. In *2015 IEEE Security and Privacy Workshops*. IEEE, 73–87.
- [15] Joe W Duran and Simeon C Ntafos. 1984. An evaluation of random testing. *IEEE transactions on Software Engineering* (1984), 438–444.
- [16] GCC. [n. d.] GCC. <https://gcc.gnu.org>. Accessed: 2024.
- [17] Patrice Godefroid, Michael Y Levin, David A Molnar, et al. 2008. Automated whitebox fuzz testing.. In *NDSS*, Vol. 8, 151–166.
- [18] Josie Holmes and Alex Groce. 2018. Causal distance-metric-based assistance for debugging after compiler fuzzing. In *2018 IEEE 29th International Symposium on Software Reliability Engineering (ISSRE)*. IEEE, 166–177.
- [19] Josie Holmes and Alex Groce. 2020. Using mutants to help developers distinguish and debug (compiler) faults. *Software Testing, Verification and Reliability* 30, 2 (2020), e1727.
- [20] Xinyi Hou, Yanjie Zhao, Yue Liu, Zhou Yang, Kailong Wang, Li Li, Xiapu Luo, David Lo, John Grundy, and Haoyu Wang. 2023. Large language models for software engineering: A systematic literature review. *ACM Transactions on Software Engineering and Methodology* (2023).
- [21] Jiajun Jiang, Yingfei Xiong, and Xin Xia. 2019. A manual inspection of Defects4J bugs and its implications for automatic program repair. *Science china information sciences* 62 (2019), 1–16.
- [22] Matthew Jin, Syed Shahriar, Michele Tufano, Xin Shi, Shuai Lu, Neel Sundaresan, and Alexey Svyatkovskiy. 2023. Inferfix: End-to-end program repair with llms. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 1646–1656.
- [23] James A Jones and Mary Jean Harrold. 2005. Empirical evaluation of the tarantula automatic fault-localization technique. In *Proceedings of the 20th IEEE/ACM international Conference on Automated software engineering*. 273–282.
- [24] Pavneet Singh Kochhar, Xin Xia, David Lo, and Shamping Li. 2016. Practitioners' expectations on automated fault localization. In *Proceedings of the 25th International Symposium on Software Testing and Analysis* (Saarbrücken, Germany) (ISTTA 2016). Association for Computing Machinery, New York, NY, USA, 165–176. [doi:10.1145/2931037.2931051](https://doi.org/10.1145/2931037.2931051)
- [25] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.
- [26] Xia Li and Lingming Zhang. 2017. Transforming programs and tests in tandem for fault localization. *Proceedings of the ACM on Programming Languages* 1, OOPSLA (2017), 1–30.
- [27] HeuiChan Lim and Saumya Debray. 2021. Automated bug localization in JIT compilers. In *Proceedings of the 17th ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments*. 153–164.
- [28] LLVM. [n. d.] LLVM. <https://llvm.org>. Accessed: 2024.
- [29] Yiling Lou, Ali Ghanbari, Xia Li, Lingming Zhang, Haotian Zhang, Dan Hao, and Lu Zhang. 2020. Can automated program repair refine fault localization? a unified debugging approach. In *Proceedings of the 29th ACM SIGSOFT International Symposium on Software Testing and Analysis*. 75–87.
- [30] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *Comput. Surveys* 56, 2 (2023), 1–40.
- [31] OpenAI. 2024. ChatGPT-4o-2024-08-06. <https://platform.openai.com/docs/models#gpt-4o>. Accessed: 2025.
- [32] Alan Romano, Xinyu Liu, Yonghui Kwon, and Weihang Wang. 2021. An empirical study of bugs in webassembly compilers. In *2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 42–54.
- [33] Raul Santelices, James A Jones, Yanbing Yu, and Mary Jean Harrold. 2009. Lightweight fault-localization using multiple coverage types. In *2009 IEEE 31st International Conference on Software Engineering*. IEEE, 56–66.
- [34] Huzafa Sidhpurwala. 2019. Security flaws caused by compiler optimizations. *Red Hat Blog* (2019).
- [35] Chengnian Sun, Vu Le, and Zhendong Su. 2016. Finding and analyzing compiler warning defects. In *Proceedings of the 38th International Conference on Software Engineering*. 203–213.
- [36] Haoxin Tu, Zhide Zhou, He Jiang, Imam Nur Bani Yusuf, Yuxian Li, and Lingxiao Jiang. 2024. Isolating Compiler Bugs by Generating Effective Witness Programs with Large Language Models. *IEEE Transactions on Software Engineering* (2024).
- [37] Qianqian Wang, Chris Parnin, and Alessandro Orso. 2015. Evaluating the usefulness of ir-based fault localization techniques. In *Proceedings of the 2015 international symposium on software testing and analysis*. 1–11.
- [38] Xi Wang, Nickolai Zeldovich, M Frans Kaashoek, and Armando Solar-Lezama. 2013. Towards optimization-safe systems: Analyzing the impact of undefined

- behavior. In *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles*. 260–275.
- [39] Ming Wen, Junjie Chen, Yongqiang Tian, Rongxin Wu, Dan Hao, Shi Han, and Shing-Chi Cheung. 2019. Historical spectrum based fault localization. *IEEE Transactions on Software Engineering* 47, 11 (2019), 2348–2368.
- [40] Ratnadira Widayarsi, Jia Wei Ang, Truong Giang Nguyen, Neil Sharma, and David Lo. 2024. Demystifying faulty code: Step-by-step reasoning for explainable fault localization. In *2024 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 568–579.
- [41] W Eric Wong, Vidroha Debroy, Ruizhi Gao, and Yihao Li. 2013. The DStar method for effective software fault localization. *IEEE Transactions on Reliability* 63, 1 (2013), 290–308.
- [42] W Eric Wong, Yu Qi, Lei Zhao, and Kai-Yuan Cai. 2007. Effective fault localization using code coverage. In *31st Annual International Computer Software and Applications Conference (COMPSAC 2007)*, Vol. 1. IEEE, 449–456.
- [43] Jifeng Xuan and Martin Monperrus. 2014. Test case purification for improving fault localization. In *Proceedings of the 22nd ACM SIGSOFT international symposium on foundations of software engineering*. 52–63.
- [44] Aidan ZH Yang, Claire Le Goues, Ruben Martins, and Vincent Hellendoorn. 2024. Large language models for test-free fault localization. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*. 1–12.
- [45] Jing Yang, Yibiao Yang, Maolin Sun, Ming Wen, Yuming Zhou, and Hai Jin. 2022. Isolating compiler optimization faults via differentiating finer-grained options. In *2022 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 481–491.
- [46] Andreas Zeller. 2002. Isolating cause-effect chains from computer programs. *ACM SIGSOFT Software Engineering Notes* 27, 6 (2002), 1–10.
- [47] Lingming Zhang, Miryung Kim, and Sarfraz Khurshid. 2011. Localizing failure-inducing program edits based on spectrum information. In *2011 27th IEEE International Conference on Software Maintenance (ICSM)*. IEEE, 23–32.
- [48] Lingming Zhang, Lu Zhang, and Sarfraz Khurshid. 2013. Injecting mechanical faults to localize developer faults for evolving software. *ACM SIGPLAN Notices* 48, 10 (2013), 765–784.
- [49] Yunlong Zhang, Haifeng Ruan, Zhiyi Fan, and Abhik Roychoudhury. 2024. Autocoderover: Autonomous program improvement. In *Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis*. 1592–1604.
- [50] Jian Zhou, Hongyu Zhang, and David Lo. 2012. Where should the bugs be fixed? More accurate information retrieval-based bug localization based on bug reports. In *2012 34th International Conference on Software Engineering (ICSE)*. 14–24. doi:[10.1109/ICSE.2012.6227210](https://doi.org/10.1109/ICSE.2012.6227210)
- [51] Zhida Zhou, He Jiang, Zhilei Ren, Yuting Chen, and Lei Qiao. 2022. Locseq: Automated localization for compiler optimization sequence bugs of LLVM. *IEEE Transactions on Reliability* 71, 2 (2022), 896–910.