Compiler Bug Isolation via Effective Witness Test Program Generation

Junjie Chen†
College of Intelligence and Computing, Tianjin University, China
junjiechen9208@gmail.com

Jiaqi Han
Peiyi Sun
HCST (Peking University), MoE, China
hanjiaqi.spy@pku.edu.cn

Lingming Zhang
University of Texas at Dallas, USA
zhanglm10@gmail.com

Dan Hao‡
Lu Hao§
HCST (Peking University), MoE, China
{haodan,zhangluces}@pku.edu.cn

ABSTRACT

Compiler bugs are extremely harmful, but are notoriously difficult to debug because compiler bugs usually produce few debugging information. Given a bug-triggering test program for a compiler, hundreds of compiler files are usually involved during compilation, and thus are suspect buggy files. Although there are lots of automated bug isolation techniques, they are not applicable to compilers due to the scalability or effectiveness problem. To solve this problem, in this paper, we transform the compiler bug isolation problem into a search problem, i.e., searching for a set of effective witness test programs that are able to eliminate innocent compiler files from suspects. Based on this intuition, we propose an automated compiler bug isolation technique, DiWi, which (1) proposes a heuristic-based search strategy to generate such a set of effective witness test programs by applying our designed witnessing mutation rules to the given failing test program, and (2) compares their coverage to isolate bugs following the practice of spectrum-based bug isolation. The experimental results on 90 real bugs from popular GCC and LLVM compilers show that DiWi effectively isolates 66.67%/79.89% bugs within Top-10/Top-20 compiler files, significantly outperforming state-of-the-art bug isolation techniques.

CCS CONCEPTS

• Software and its engineering → Software testing and debugging.

KEYWORDS

Compiler Debugging, Bug Isolation, Test Program Generation

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

ACM Reference Format:


1 INTRODUCTION

Software bugs in modern software systems can incur huge cost. For example, Tricentis.com studied software bugs in 363 companies all over the world, and reported that these bugs incurred almost $1.1 Trillion cost, and affected over 4.4 Billion customers in 2016. Among software bugs, compiler bugs are especially critical, since almost all software systems are compiled via compilers and a buggy compiler can potentially affect all the software systems built on it. Therefore, it is crucial to detect, isolate, and fix compiler bugs.

Although researchers have devoted dedicated efforts to compiler testing [15, 24, 29, 43, 53, 58, 82], compiler bug isolation and fixing are still a tedious and time-consuming process, since compilers are very complex and developers have to understand the root cause of a compiler bug and then determine the fixing strategy. In the literature, lots of automated bug isolation techniques (also known as fault localization, e.g., spectrum-based techniques or SBFL [22, 38, 52, 72, 77], slicing-based techniques [81], mutation-based techniques [35, 51, 60, 66, 67, 87]) have been proposed. However, these techniques can hardly be applicable to compilers due to their scalability or effectiveness problem. First, compilers like GCC are complex and large, making it extremely expensive to perform advanced static/dynamic analysis on a compiler. Therefore, program-analysis based techniques like slicing-based techniques and mutation-based techniques can hardly be used in compiler bug isolation. For example, we found that it can take over a month to execute only a limited set of GCC compiler mutants. Second, due to the complexity of compilers, the execution traces between passing and failing test programs tend to differ significantly. Therefore, it is hard for SBFL, which isolates compiler bugs by comparing coverage between passing and failing test programs, to isolate compiler bugs effectively, which is also demonstrated in Sections 2 and 5. In other words, automated compiler bug isolation is still challenging due to its inherent difficulty.

Given a bug-triggering test program (also called failing test program) for a compiler, developers need to manually isolate buggy compiler files (i.e., source files of compilers) among all the touched compiler files when compiling the given failing test program. However, there are usually at least hundreds of compiler files involved in
With this intention, we transform the problem of automated compiler bug isolation to a search problem, i.e., finding a set of effective witness test programs to help isolate compiler bugs precisely. However, it is challenging to find such a set of effective witness test programs. On the one hand, each witness test program in the set is required to have a large witness capability, i.e., it can eliminate innocent compiler files as many as possible from suspects. On the other hand, the witness capabilities of different witness test programs in the set are required to be diverse, i.e., each of them can eliminate different innocent compiler files from suspects, so that grouping them can be helpful to isolate the buggy compiler files precisely. To solve this problem, in this paper, we propose the first compiler bug isolation technique via searching for diversified witnesses, called DiWi (Diversified Witnesses). The main contribution of DiWi lies in how to find the set of effective witness test programs. To address the first challenge, we design a series of mutation rules for DiWi to generate witness test programs via slightly mutating the given failing test program. The key insight is that such minor changes are more likely to make the generated witness test program share a close compiler execution trace with the failing test program, so that the generated witness test program can eliminate more innocent compiler files from suspects. To address the second challenge, DiWi utilizes some heuristics to guide the construction of the set of effective witness test programs. More specifically, during the generation of each witness test program, DiWi considers the diversity of compiler execution traces between it and the already generated witness test programs, so that these generated witness test programs can eliminate different innocent compiler files from suspects. Finally, DiWi ranks the files executed by the failing test program based on coverage comparison between the set of effective witness test programs and the given failing test program like SBFL [74].

To evaluate the effectiveness of DiWi, we constructed an extensive dataset including 90 real-world reproducible bugs from the popular GCC [2] and LLVM [5] compilers. The experimental results on all the 90 studied bugs show that DiWi effectively isolates 10/37/60/71 bugs (out of 90) within Top-1/5/10/20 buggy files, significantly outperforming traditional SBFL. As the core of DiWi, we investigated the contribution of our search-based witness test programs compared with the developer-provided test suite, executing the buggy compiler (which includes test programs passing on this compiler) and the randomly generated passing test programs via Csmith [82] (the most widely-used random C program generation tool). The results show that our search-based witness test programs significantly outperform the latter two, e.g., isolating 150.00% and 66.67% more bugs than the developer-provided passing test programs and randomly generated passing test programs at the Top-1 position. We also investigated the contribution of our heuristic-based search strategy compared with the random search strategy during generation, and explored the synthesis of DiWi and developer-provided test suites. In summary, this paper makes the following contributions:

- **Idea.** An automated compiler bug isolation technique that transforms the problem of bug isolation to the problem of guided test program generation via search&mutation.

- **Implementation.** A practical open-source tool implementing the proposed compiler bug isolation technique based on the LLVM Clang infrastructure.

- **Benchmark.** An open-source dataset containing 90 reproducible real bugs from GCC and LLVM compilers for future research on compiler bug detection, isolation, and fixing.

- **Study.** An extensive study on all 90 bugs from our dataset demonstrating that the proposed technique is able to significantly outperform existing techniques; our search-based witness test programs (the core of DiWi) have higher quality than the developer-provided test suite and the randomly generated passing test programs for compiler bug isolation; lastly, components from developer-provided test suites can further boost DiWi effectiveness.

### 2 MOTIVATION

Here, we use an example to illustrate the motivation of this paper. Figure 1 shows an example from LLVM bug ID 24482, where the left is the reported bug-triggering test program. When the buggy compiler (revision 245195 in LLVM trunk) compiles the failing program, it produces different outputs under the compilation options “-Os” and “-O1”. The bug occurs at the file "DeadStoreElimination.cpp" due to dead store elimination across basic blocks.

It is tedious for developers to manually find the buggy file since this revision contains a large number of files (i.e., 3,581). Although traditional SBFL is reported to be effective in other software systems [68], it may not be effective in compilers due to their characteristics. We applied Ochiai [7] (one of the most effective formulae in SBFL) based on the failing test program and the developer-provided...
test suite, and found the buggy file is ranked at the 659th position out of 3,581 files. That is, with traditional SBFL developers need to examine 658 innocent compiler files before finding the buggy one, indicating the inferior effectiveness of traditional SBFL in compilers.

To isolate compiler bugs effectively following SBFL, instead of the developer-provided passing test programs, a set of effective witness test programs are desirable. Intuitively, witness test programs sharing similar execution traces (except the buggy file) with the given failing test program are more helpful to eliminate innocent files from suspects. With this intuition, we generated 20 witness test programs by randomly introducing various minor changes into the failing test program, making them share similar execution trace with the failing one. Then we compared their coverage to isolate the bug like SBFL [74]. In this way, the buggy file is ranked at the 51st position, demonstrating that these generated witness test programs by slightly changing the failing test program can help isolate the bug effectively.

We further analyzed the isolation result using each witness test program and the given failing test program (namely a pair of test programs) to learn the performance of each individual witness test program. Among the 20 pairs, 1 pair achieves the optimal result (i.e., Top-1), 13 pairs rank the buggy file within the first 10 positions, and 7 pairs rank the buggy file after the 90th positions. Overall, there are many effective pairs and a few low-quality pairs, by changing the failing test program minorly. For example, Figure 1b shows the optimal witness test program that ranks the buggy file at the 1st position, which is generated by changing one variable of the failing test program in Figure 1a; while Figure 1c shows a low-quality witness test program that ranks the buggy file at the 33rd position, which is generated by replacing “b. f1” with “a. f1”. Since we use coverage comparison between test programs to isolate compiler bugs, a witness test program eliminating more innocent files from suspects tends to mean that it shares a more similar compiler execution trace with the failing test program, and also means that it has higher quality.

Based on the above analysis, we have the following observations: First, a set of high-quality witness test programs can help isolate compiler bugs with the given failing test program; Second, even if we change the failing test program minorly, the obtained witness test programs are of various quality.

3 RELATED WORK
We discuss the most closely related work to compiler bug isolation. Automated Debugging. In the literature, there are a huge amount of work on automated debugging, e.g., fault localization [27, 45–47, 55, 56, 80, 85] and program repair [21, 28, 48, 59, 62, 65], where our work targets the former. As presented in Section 1, the existing fault localization techniques cannot work well on compilers.

In the literature there also exists mutation-based fault localization [50, 51, 60, 66, 87], which aims to mutate source programs to check the impact of each code element on the test outcomes. However, our work is to introduce the idea of mutation to slightly change test programs so as to generate a set of effective witness test programs. Besides, some work focuses on improving fault localization via test generation [8, 10, 11, 54, 71] or test selection [25, 26, 61, 70] for ordinary programs. However, compiler inputs are programs and compilers are extremely complex and huge, making none of these existing test-generation and test-selection techniques for better debugging directly applicable here.

In automated debugging, our work is mostly related to compiler debugging. Most compiler debugging work focused on providing debugging messages/visualization [13, 31, 41, 64, 73]. Some work focused on reducing bug-triggering tests to facilitate debugging [12, 32, 69, 76, 84]. In our work, the provided bug-triggering tests are already the reduced ones, as required by compiler developers. Besides, Chen et al. [19] proposed a technique to rank test programs triggering bugs such that test programs triggering distinct bugs are early in the list. Holmes and Groce [34] further proposed a new metric to determine similarity of failing test programs to facilitate debugging. In contrast, our work aims to automatically isolate compiler bugs via effective witness test program generation. In particular, Zeller [83] proposed to produce an entire cause-effect chain in GCC from input to result for facilitating compiler debugging. Actually, cause-effect chain and our technique are complementary: 1) the former produces fault-diagnosis information at the program-state level while the latter does this at the source-code level, 2) the former manipulates in memory and may not handle external states, 3) the latter is more lightweight.

Mutation Testing. Mutation testing is one of the most effective methods to measure test-suite quality [36, 39, 86]. It deliberately seeds bugs into the original source program to simulate the bugs that developers often make in practice. Different from it, our work aims to conduct test program mutation to generate a set of effective witness test programs for facilitating compiler bug isolation. Here, we both apply existing mutation rules from mutation testing and also design new mutation rules for compiler bug isolation.

Compiler Testing. Compiler testing usually happens before compiler bug isolation (the target of our work). Most research on compiler testing focuses on test program generation [20, 82], test oracle construction [43, 57], and test execution acceleration [14–16, 18]. The general idea of mutation is also applied to compiler testing [20, 33, 42, 44], which aims to generate different test programs as much as possible by faster diverging from the given seed program for detecting deep compiler bugs. Different from them, our work aims to utilize mutation for compiler bug isolation, and the design goal of our mutation is to flip the compiler execution results (i.e., from failing to passing) to generate a set of witness test programs close to a given failing test program.

To sum up, the existing fault localization techniques cannot work well for large-scale compiler systems, while the existing applications of mutation cannot be directly used to solve the problem of compiler bug isolation. Therefore, this work makes the first attempt to isolate compiler bugs via search-based mutation.

4 APPROACH
Following the observations in Section 2, we propose a novel technique, named DiWi (Diversified Witnesses), to isolate compiler bugs. In DiWi, we first deliberately generate a set of effective witness test programs, and then compare them with the given failing test program like SBFL to isolate compiler bugs [74]. From Section 2, we know that it is hard to ensure each generated witness test program to be effective, and thus we use an aggregation mechanism to minimize the impact of low-quality witness test programs in DiWi. Therefore, there are two key issues to the success of DiWi. First,
we need to carefully construct each witness test program to avoid introducing many low-quality witness test programs; otherwise, we have too much noise. Second, the witness test programs should evenly eliminate buggy suspects; otherwise, the aggregation process is prone to biases. To help construct such a set of effective witness test programs, we define the following two criteria.

\( C_1 \): Each test program in the set of effective witness test programs should share a similar compiler execution trace with the failing test program.

\( C_2 \): Test programs in the set of effective witness test programs should have great diversity in their compiler execution traces.

However, since the space of witness test programs is extremely huge (in fact infinite to be precise), efficiently generating such a set of effective witness test programs satisfying the two criteria is challenging. To satisfy the first criterion, we design a series of mutation rules for DiWi to generate witness test programs by slightly mutating the failing test program. Intuitively, such minor changes are likely to make the generated witness test program share a close compiler execution trace with the failing test program. Here, we call our mutation witnessing mutation, aiming to generate witness test programs with a large witness capability. To satisfy the second criterion, DiWi utilizes some heuristics to guide the construction of the set of effective witness test programs. That is, during the generation of a new witness test program, DiWi considers the diversity of compiler execution traces between the new one and the already generated witness test programs, aiming to make their witness capabilities diversified.

In the following, we introduce the designed witnessing mutation in Section 4.1, heuristic-based witness test program construction in Section 4.2, and the aggregation mechanism for compiler bug isolation in Section 4.3.

### 4.1 Witnessing Mutation

As presented in Section 3, the existing applications of mutation cannot be directly used for compiler bug isolation. In our context, the goal of test program mutation is to introduce slightly different control- and data-flow information to flip the compiler execution results (i.e., from failing to passing) to generate witness test programs. Therefore, we design a series of mutation rules specific to our mutation goal.

**Mutation Rules.** To achieve the mutation goal of flipping the failing test program, we manually analyzed historical compiler bugs to investigate why these test programs are able to trigger bugs and which characteristics of test programs can sensitively impact compiler execution results. Also, to make the changes minor, we design our mutation, based on the learned knowledge, at the level of fine-grained program elements, including variables, operators, and constants. Therefore, we propose three mutation categories: variable mutation, operator mutation, and constant mutation. For variable mutation, each program variable can potentially be changed into another compatible variable or type, since program variables are the core of data dependencies and variable types can impact many compiler optimizations. For operator mutation, each program operator (i.e., arithmetic, logical, relational, and unary) can potentially be changed into another compatible operator, since program operators can significantly impact data- and control-dependencies. For constant mutation, each program constant can potentially be changed into another value, since many compiler bugs are triggered under some specific values. Here, we both apply existing mutation rules from traditional mutation testing [39] and design new mutation rules for compiler bug isolation (e.g., variable type mutation).

More formally, following the presentation of prior work [89] we view a test program as a syntactic skeletal structure \( P \) with “holes” that can be configured with various variables (denoted as hole \( \square_v \)), operators (denoted as hole \( \square_o \)), and constant values (denoted as hole \( \square_c \)). In this way, we can fill each hole with mutated content to derive a new mutated program. For the ease of presentation, let us consider a WHILE-style language that has been widely used in the program analysis research [63, 89]. The program syntax rules for the WHILE-style language are shown in Figure 2. In the rules, non-terminals \( a, b, S \) denote arithmetic expressions, boolean expressions, and program statements, respectively; terminals \( op_a, op_b, op_p \), and \( op_c \) denote arithmetic, logical, relational, and unary operators, respectively; terminals \( x \) and \( n \) denote program variables and constants. Note that we use the WHILE-style language for the ease of presentation, and our technique applies to a full-fledged language such as C. To obtain program mutations, we recursively apply a mutation transformation to the WHILE syntax rules. Figure 3 presents the syntax rules for all our three mutation categories. We denote all the three categories of transformations and program holes as \( \square \) and \( \square_r \) respectively, i.e., \( \square = \square_v \cup \square_c \cup \square_r \). In this way, we formally define our mutation process.

DEFINITION 1 (Mutation Skeleton). Given any test program \( P \), we say \( P \) is a mutation skeleton of \( P \) if the abstract syntax tree (denoted as \( T_P \)) of \( P \) is the same as the transformed abstract syntax tree (denoted as \( [T_P] \)) of \( P \), i.e., \( T_P = [T_P] \).

DEFINITION 2 (First-order Mutation). Given a test program \( P \) and its mutation skeleton \( P \), for all the \( n \) holes (e.g., \( \{ \square_1, \square_2, \ldots, \square_n \} \)) within \( P \), first-order mutation fills each hole with the same content as \( P \) except filling one hole with a mutated content.

DEFINITION 3 (High-order Mutation). Given a test program \( P \) and its mutation skeleton \( P \), for all the \( n \) holes (e.g., \( \{ \square_1, \square_2, \ldots, \square_n \} \)) within \( P \), high-order mutation fills \( i (2 \leq i \leq n) \) holes with mutated contents, while filling the rest holes with the same content as \( P \).

Since we need diverse execution traces representing different ways to flip the failing test program for effective compiler bug isolation, for each mutation category we design a plurality of mutation rules. The detailed mutation rules for each mutation category are shown in Table 1. Based on the three types of holes, we design 132 specific mutation rules in total and the full list can be found in the project webpage. Here we regard each specific mutation operation on one type of program holes as an individual mutation rule, e.g., replacing “\( x=0 \)” with “\( y=0 \)” and replacing “\( x=\alpha \)” with “\( x=1 \)” are two different rules, while replacing “\( x=0 \)” with “\( y=0 \)” and replacing “\( x=\alpha \)” with “\( z=\alpha \)” belong to the same rule.
Different much as possible. In DiWi, given a failing test program, di
generate a witness test program that di
select a subset of e
Due to huge search space and limited computational resources, we

4.2 Heuristic-Based Test Program Generation
Due to huge search space and limited computational resources, we
cannot generate all witness test programs via mutation and then select
a subset of effective witness test programs from them. One of the most
cost-effective ways is that, during each generation, we generate a witness
program that differs from existing ones as much as possible. In DiWi, given a failing test program, different
performs differently for different initial failing test programs. Therefore, we carefully design an adaptive procedure to select mutation rules for constructing a set of effective witness test programs. Intuitively, if a mutation rule can more frequently generate witness test programs that have greater diversity with the existing ones, the mutation rule should be selected with higher probability for further mutations. Based on the intuition, we compute a priority score for each mutation rule MR as $\text{Score}(MR) = \frac{1}{m} \sum_{i=1}^{m} \text{Dist}(P, P_i) * \text{Rate}_s$, where: (1) $m$ is the number of existing witness test programs in the set; (2) $P$ is the new witness test program generated by using $\text{MR}$; (3) $\text{Dist}(.)$ is the coverage distance, computed by Formula 1; (4) $\text{Rate}_s$ is the success rate of generating accepted witness test programs, i.e., the ratio of the number of times the witness test program generated by $\text{MR}$ is accepted into the witness set to the number of times $\text{MR}$ is selected for mutations.

Mutation rules can be ranked according to the descending order of the computed priority scores. However, we cannot directly select the mutation rule ranked at the $1_{st}$ position for next mutation, since the ranking is based on historical results of these mutation rules and can hardly perfectly predict future results. Therefore, each mutation rule should have some probability to be selected for next mutation, and a mutation rule ranked higher in the ranking list should have a larger probability to be selected. That is, the problem of mutation rule selection in DiWi can be regarded as the sampling problem from a probability distribution.

Here, since which mutation rule to be selected depends on the most recent behavior of each mutation rule, it is actually a typical Markov Chain (MC). Therefore, to solve the sampling problem from a probability distribution, DiWi adopts the Metropolis-Hastings (MH) algorithm [49], the most popular Markov Chain Monte Carlo method, as heuristic by assuming the desired distribution to be equilibrium distribution [23]. Here, MH obtains random samples from a probability distribution. In our context, it samples the next mutation rule (denoted as $\text{MR}_k$) based on the current mutation rule (denoted as $\text{MR}_a$) according to a probability distribution. If $\text{MR}_k$ is better than $\text{MR}_a$, i.e., the priority score of $\text{MR}_k$ is larger than that of $\text{MR}_a$, $\text{MR}_k$ is definitely accepted; If not, $\text{MR}_k$ still has some probability to be accepted. Following the existing work [20], we set the probability distribution to be the geometric distribution, which is the probability distribution of the number of X of Bernoulli trials needed to obtain one success. If the success probability on each trial is $p$, the probability the $k_{th}$ trial being the first success can be computed as $P_{s}(X=k) = (1-p)^{k-1}p$.

During the process, mutation rules are selected randomly, and thus the proposal distribution is symmetric. Therefore, the probability of accepting $\text{MR}_k$ given $\text{MR}_a$ is computed as $P_a(\text{MR}_k|\text{MR}_a) = P_a(\text{MR}_k) / P_a(\text{MR}_a)$, where $k_a$ and $k_b$ are the positions of $\text{MR}_a$ and $\text{MR}_b$ in the ranking list of mutation rules according to the descending order of their priority scores. Please note that, when $\text{MR}_a$ is better than $\text{MR}_b$ (i.e., $P_a(\text{MR}_a) > P_a(\text{MR}_b)$), $P_a(\text{MR}_b|\text{MR}_a) = 1$. After acquiring the result of $\text{MR}_b$ (i.e., accept or reject), DiWi updates its score and re-ranks these mutation rules for next iteration.

4.2.3 Overall Algorithm. We formally present the generation process of DiWi in Algorithm 1. The initial set of seed programs contains only the given failing test program $P_f$, and the priority score of each mutation rule is 0. In this algorithm, Line 1 randomly selects a mutation rule as the current one $\text{MR}_a$. Lines 3-21 construct a set of witness test programs until achieving a terminating condition. Line 4 selects a seed program $P$ to mutate. Line 5 gets the position of $\text{MR}_a$ in the ranking list of mutation rules. Lines 6-10 acquire the next mutation rule $\text{MR}_b$. Line 11 applies $\text{MR}_b$ to mutate $P$ to generate a new program $P'$. Lines 12-17 determine whether $P'$ is accepted by $P$ and $S'$. Based on its execution results and coverage distances with the existing witness programs in $P$, Lines 18-20 update the score of $\text{MR}_b$ and re-rank these mutation rules for next iteration. If $P'$ is accepted into $P$, the score is updated by changing the two items in the formula Score; Otherwise, the score is updated by just changing the latter item in the formula Score, since the witness set and the coverage distances are not changed. Lines 22-28 determine whether terminating the construction process. If there is no witness program constructed via lower order mutation, the construction process (Lines 2-21) is repeated by using higher order mutation (i.e., using higher order failing programs as seed programs).

4.3 Aggregation-Based Compiler Bug Isolation

After constructing a set of witness test programs, DiWi isolates compiler bugs by analyzing the set of witness test programs and the given failing test program. Following SBFL [7], DiWi computes the suspicious value for each statement within the touched code when compiling the given failing test program. Here DiWi adopts Ochiai [7], one of the most effective formulae in SBFL, to compute the suspicious value for each touched statement. The formula is $s_{\ast}(s) = \frac{e_{fs}}{\sqrt{e_{fs} + n_{fs}e_{fs} + e_{ps}}}$, where $e_{fs}$ and $n_{fs}$ represent the number of failing test programs that execute and do not execute statement $s$, and $e_{ps}$ represents the number of passing test programs executing $s$. Here, we have only one given failing test program and just consider the statements touched by the failing test program, and thus $e_{fs}$ is 1 and $n_{fs}$ is 0. Therefore, $s_{\ast}(s) = \frac{1}{\sqrt{1+e_{ps}}}$ in DiWi.
Then, similar to method-level aggregation [74], DiWi computes the suspicious value of each file by aggregating the suspicious values of all the touched statements in the file. When a failing test program is mutated to a set of witness test programs, the coverage for the buggy file would be changed more than that for the bug-free files on the whole. Therefore, to compute the suspicious value of each file, we use the formula \( SUS(f) = \frac{\sum_{i=1}^{n_f} sus(s_i)}{n_f} \), where \( n_f \) is the number of touched statements when compiling the failing test program in the file \( f \).

5 EVALUATION

In this study, we address the following research questions:

- **RQ1**: How does DiWi perform on compiler bug isolation?
- **RQ2**: Do our search-based witness test programs outperform the developer-provided test suite and the randomly generated passing test programs?
- **RQ3**: Does our heuristic-based search strategy outperform the random search strategy during mutation?
- **RQ4**: Can DiWi take advantage of the developer-provided test suite (which is an exploration to further boost DiWi)?

5.1 Benchmark

We used GCC and LLVM as subjects, which cover almost all popular C compilers used in the existing work [15, 17, 43, 44, 82]. To investigate the effectiveness of DiWi, from the bug repositories [3, 6] of GCC and LLVM, we manually collected 90 bugs in total, each of which is required to satisfy the following conditions: 1) the bug has the equipped failing test program and the compilation options triggering it in the bug report; 2) the bug has been fixed; 3) the bug can be reproduced in our experimental environment. We manually collected bugs reported after 2013 following these conditions until we had 45 bugs for each compiler since the manual process is costly. Also, we manually identified the buggy locations (i.e., buggy files) for each bug, which serve as the ground truth to evaluate the effectiveness of bug isolation in our study. On average, a GCC buggy version has 1,588 files with 1,414K source lines of code (SLOC), while a LLVM buggy version has 3,507 files with 1,431K SLOC\(^{\text{1}}\). We release the benchmark to facilitate the future research on compiler bug detection, isolation, and fixing, and welcome more researchers to contribute to this benchmark. In our benchmark, each bug has: ① buggy compiler version; ② failing test program; ③ compilation options for reproducing the bug; ④ buggy location; ⑤ fixed version. Our benchmark and code are available at the project webpage: https://github.com/JunjieChen/DiWi.

5.2 Implementation and Configuration

DiWi utilizes Clang Libtooling library [1] to parse a test program to an abstract syntax tree (AST) and then mutates it at the AST. DiWi utilizes Gcov [4] to collect compiler test coverage. Here we set the success probability of each Bernoulli trial \( p \) in Algorithm 1 to be 0.023 by satisfying the following conditions: ① \( 0.95 \leq \sum_{k=2}^{32} (p^{X=k}) \leq 1 \); ② \( p \geq \frac{1}{12} \); ③ \( (1 - p)^{132 - 1} p < \varepsilon \), where \( \varepsilon \) is a quite small deviation (e.g., 0.001). We set the terminating condition of DiWi to be one hour limit in our study. For any technique involving randomness, we repeated it 5 times and use the median results. Our study is conducted on a workstation with four-core CPU, 120G memory, and Ubuntu 14.04 operating system.

5.3 Measurements

To evaluate the effectiveness of bug isolation, we measure the position of each buggy file in the ranking list produced by a bug isolation technique. If more than two files have the same suspiciousness, we use the worst ranking following the existing work [37, 51, 68]. We compute the following widely used metrics [9, 51, 60, 74]:

- **Top-n** refers to the number of successfully isolated bugs within the Top-n position (i.e., \( n \in \{1, 5, 10, 20\} \) in our study) in the ranking list. Larger is better.
- **Mean First Rank (MFR)** refers to the mean of the first buggy file rank for each bug. This metric emphasizes fast isolation of the first buggy element to ease debugging. Smaller is better.
- **Mean Average Rank (MAR)** refers to the mean of the average rank of all buggy files for each bug. Different from MFR, MAR emphasizes precise isolation for all buggy elements.

5.4 Compared Techniques

Spectrum-based bug isolation (SBFL) [80] is the most widely-studied bug isolation technique among traditional bug isolation techniques. It is interesting to evaluate the effectiveness of traditional SBFL on compilers. It first records the coverage status of each program element (i.e., each compiler file in the study) during each test execution and the test outcomes (i.e., passing or failing). Then SBFL computes a suspicious value for each program element using some formula, and finally ranks the program elements based on their suspicious values. Researchers have made dedicated efforts to design various formulas on suspiciousness computation. We evaluated eight popular formulas following the existing work [68, 81, 88], including SBI, Ochiai, Taranandu, Jaccard, Ochiai2, Kulczynski2, \( O^P \), and \( D^2 \), and found they achieved extremely similar results for compiler bug isolation in our study. Due to space limitation, we use the most effective Ochiai as the representative. For each compiler bug, we use the given failing test program as the failing test program, and then use the developer-provided test suite as the passing test programs.

As the core of DiWi is to generate effective witness test programs, it is interesting to investigate the impact of generated witness test programs on bug isolation. To achieve this goal, we replace the set of our search-based witness test programs with the developer-provided test suite, and then use the aggregation-based SBFL to isolate bugs. We call this technique **SBFL\text{dev}**. Besides, randomly generated test programs via test-program generation tools like Csmith [82] have been demonstrated to be quite effective for detecting compiler bugs [15, 82]. Therefore, we also replace the set of our search-based witness test programs with a set of randomly generated passing test programs via Csmith [82], and then use the aggregation-based SBFL to isolate bugs. Similarly, we call this technique **SBFL\text{rand}**.

Besides, to investigate the impact of our heuristic-based search strategy on bug isolation, we replace this strategy with the random search strategy. That is, we do not have any guidance for constructing witness test programs via mutation. We call this variant of DiWi **DiWi\text{rand}** (random search). Note that we used the same terminating condition for all compared techniques for fair comparison.

---

\( ^{\text{1}}\)Since GCC and LLVM are implemented using C and C++ respectively, we consider all the C files for the former and C++ files for the latter.
5.5 Threats to Validity
First, the findings in this work may not generalize to other compiler bugs. To reduce this threat, we tried our best to construct a new dataset including 90 real-world compiler bugs. Note that the process was extremely time consuming and tedious; to our knowledge this is the largest dataset for reproducible compiler bugs. Second, besides the used two kinds of passing test programs, there are other kinds of test programs, e.g., programs generated via swarm testing [29]. In the future, we will use more kinds of passing programs for comparison. Third, the settings (e.g., parameters in our search-based strategy and terminating condition of DiWi) may impact our study. In the future, we will explore their impacts.

5.6 Results and Analysis
5.6.1 Overall effectiveness of DiWi. Rows “DiWi” in Table 2 present the effectiveness of DiWi. Overall, DiWi successfully isolates 10/37/60/71 bugs (out of 90 bugs) within Top-1/5/10/20 buggy files, demonstrating its effectiveness on compiler bug isolation. That is, about 66.67% and 78.89% bugs are effectively isolated within 10 and 20 compiler files, respectively. We further analyzed the effectiveness of DiWi for different compiler systems. Intuitively, LLVM has a much larger number of files (shown in Section 5.1), and should be harder to perform bug isolation. However, interestingly, shown in Table 2, DiWi achieves quite similar effectiveness on GCC and LLVM, and DiWi even isolates 3 more bugs within Top-5 on LLVM than GCC. Moreover, we find that other studied techniques indeed perform worse on LLVM than GCC. Therefore, that demonstrates the scalability of DiWi — DiWi’s effectiveness does not decrease dramatically when facing larger compiler systems.

We also analyzed the comparison effectiveness between DiWi and traditional SBFL shown in Rows “SBFL” in Table 2. We find that traditional SBFL performs poorly on the studied compilers. For example, SBFL ranks the GCC buggy files as the 276.22(MFR) and 14.76(MAR), localizing compiler buggy files within 14.27(MFR) and 14.76(MAR), respectively. We further analyzed the effectiveness of SBFL for different compiler systems. Intuitively, LLVM has a much larger number of files (shown in Section 5.1), and should be harder to perform bug isolation. However, interestingly, shown in Table 2, DiWi achieves quite similar effectiveness on GCC and LLVM, and DiWi even isolates 3 more bugs within Top-5 on LLVM than GCC. Moreover, we find that other studied techniques indeed perform worse on LLVM than GCC. Therefore, that demonstrates the scalability of DiWi — DiWi’s effectiveness does not decrease dramatically when facing larger compiler systems.

We also analyzed the comparison effectiveness between DiWi and traditional SBFL shown in Rows “SBFL” in Table 2. We find that traditional SBFL performs poorly on the studied compilers. For example, SBFL ranks the GCC buggy files as the 276.22(MFR) and 14.76(MAR), localizing compiler buggy files within 14.27(MFR) and 14.76(MAR), respectively. We further analyzed the effectiveness of SBFL for different compiler systems. Intuitively, LLVM has a much larger number of files (shown in Section 5.1), and should be harder to perform bug isolation. However, interestingly, shown in Table 2, DiWi achieves quite similar effectiveness on GCC and LLVM, and DiWi even isolates 3 more bugs within Top-5 on LLVM than GCC. Moreover, we find that other studied techniques indeed perform worse on LLVM than GCC. Therefore, that demonstrates the scalability of DiWi — DiWi’s effectiveness does not decrease dramatically when facing larger compiler systems.

We also analyzed the comparison effectiveness between DiWi and traditional SBFL shown in Rows “SBFL” in Table 2. We find that traditional SBFL performs poorly on the studied compilers. For example, SBFL ranks the GCC buggy files as the 276.22(MFR) and 14.76(MAR), localizing compiler buggy files within 14.27(MFR) and 14.76(MAR), respectively. We further analyzed the effectiveness of SBFL for different compiler systems. Intuitively, LLVM has a much larger number of files (shown in Section 5.1), and should be harder to perform bug isolation. However, interestingly, shown in Table 2, DiWi achieves quite similar effectiveness on GCC and LLVM, and DiWi even isolates 3 more bugs within Top-5 on LLVM than GCC. Moreover, we find that other studied techniques indeed perform worse on LLVM than GCC. Therefore, that demonstrates the scalability of DiWi — DiWi’s effectiveness does not decrease dramatically when facing larger compiler systems.

Qualitative Analysis. We further conducted qualitative analysis using two examples. Figure 5 shows an example LLVM bug, where the left is the failing test program and the right is one of witness test programs generated by DiWi. The test program (in Figure 5a) is miscompiled by the LLVM trunk (revision 229830) at -01 and above. The bug occurs at the file “ScalarEvolution.cpp”, which incorrectly promotes the 16-bit add into a 32-bit add. After just one mutation shown in Figure 5b, the mutated program does not trigger the bug, demonstrating the power of our designed mutation rules. We computed the coverage distance between the two programs, and its value is only 0.013, which confirms our assumption that minor changes are likely to make them share close compiler execution traces. In particular, DiWi ranks the buggy file at the 6th of all files.

Figure 6 shows an example for a GCC bug. The failing test program in Figure 6a is miscompiled by the GCC trunk (revision 206472) at -Os and above. The bug occurs at the file “tree-ssa-sink.c”, because it does not well handle the case where a second eliminated

extension requires widening a copy created for elimination of a prior extension. After adding “volatile” as shown in Figure 6b, the mutated program does not trigger the bug anymore. We computed their coverage distance, and its value is 0.019, which is also very small. In particular, DiWi ranks the buggy file at the 1st of all files.

Furthermore, DiWi also brings extra benefits. That is, these generated witness test programs via mutation provide some useful hints for the developers to facilitate bug diagnosis/fixing. For example, from Figure 5, the bug is not triggered again by replacing “+” with “+”, which means that “-1” or “+1” has an impact on the bug. That is true, because incorrect promotion from 16-bit add to 32-bit add leads to incorrect widening from “-1” to “65535”.

5.6.2 DiWi vs. SBFL$^a$ vs. SBFL$^b$ and SBFL$^c$ vs. SBFL$^d$ and. We compared DiWi with SBFL$^a$, and SBFL$^b$ and SBFL$^c$ and SBFL$^d$ and to investigate the impact of our search-based witness test programs. From Rows “DiWi”, “SBFL$^a$”, and “SBFL$^b$”, and “SBFL$^c$” and “SBFL$^d$” in Table 2, DiWi outperforms them in terms of all the used metrics. For example, DiWi isolates 150.00% 68.18%/33.33%/12.70% more bugs than SBFL$^a$ and 66.67%/54.17%/36.36%/18.33% more bugs than SBFL$^c$ and SBFL$^d$ and, within Top-1/5/10/20 for all the studied bugs. That demonstrates that DiWi achieves much more precise isolation results than SBFL$^a$ and SBFL$^c$ and SBFL$^d$ and. In particular, we conducted the Wilcoxon Signed-Rank Test [79] for their isolation ranks at the significance level of 0.05 to determine whether there are significant differences between them on all the studied bugs. The p-value is 0.0016 and 0.0021 respectively, demonstrating that DiWi does significantly outperform them. These results indicate that our search-based witness test programs are more powerful than both the developer-provided and randomly generated (via Csmith) witness test programs on isolating compiler bugs.
To further investigate why our search-based witness test programs can significantly outperform the others, we further quantitatively evaluated one basic assumption of DiWi, i.e., minor mutation changes are likely to make them share close compiler traces. We computed the mean coverage distance between the given failing test program and all witness test programs for each bug. Figure 7 shows the coverage distance comparison among DiWi (our search-based witness program generation), SBFL\textsubscript{dev} (the developer-provided test suite, denoted as Dev in this figure), and SBFL\textsubscript{rand} (the Csmith random program generation, denoted as Rand). In this figure, the violin plots show the density of coverage distances at different values, and the box plots show the median and interquartile ranges. From this figure, we find that our search-based witness programs have much smaller coverage distances with the given failing test program than the developer-provided witness programs and the randomly generated witness programs via Csmith. That validates our assumption. Interestingly, in GCC there are some cases where the coverage distances are obviously larger than other cases. We looked into the code and found that all these bugs are crash bugs, and various code regions not executed by the failing test program (due to crashes) can be executed by the witness test programs.

5.6.3 DiWi vs. DiWi\textsuperscript{rand}. We also compared DiWi and DiWi\textsuperscript{rand} to investigate the impact of our heuristic-based search strategy shown in Rows "DiWi\textsuperscript{rand}" in Table 2. From this table, DiWi outperforms DiWi\textsuperscript{rand} for both GCC and LLVM in terms of all the used metrics.

DiWi isolates 66.67%/27.59%/25.00%/7.58% more bugs within Top-1/5/10/20 than DiWi\textsuperscript{rand} for all the studied bugs. DiWi performs 24.38% and 23.56% better than DiWi\textsuperscript{rand} in terms of MFR and MAR. That demonstrates that DiWi performs more precise than DiWi\textsuperscript{rand} for isolating both GCC and LLVM bugs. We also conducted the Wilcoxon Signed-Rank Test for the isolation ranks of DiWi and DiWi\textsuperscript{rand}. The p-value is 1.345e-06, demonstrating the superiority of DiWi over DiWi\textsuperscript{rand}. The results indicate that our heuristic-based search strategy outperforms random search.

To investigate why the heuristic-based search strategy outperforms the random search strategy, we quantitatively evaluated another basic assumption of DiWi, i.e., witness test programs generated via our heuristic-based search strategy should have great diversity. We computed the coverage diversity among witness test programs. Here we first computed the minimum distance for each witness test program with others, and then computed the mean of all minimum distances for all witness test programs. We find that the mean coverage diversity of our heuristic-based strategy is about 214.72X and 14.10X greater than that of the random search strategy, validating our assumption.

5.6.4 Exploring DiWi with Developer Tests. Our search-based witness test programs have been demonstrated to outperform the other two. We further analyzed the cases where each type of witness programs performs well. Here we chose the bugs isolated within Top-5 to analyze. We found although DiWi isolates 37 bugs within Top-5, SBFL\textsubscript{dev} and SBFL\textsubscript{rand} can also isolate 11 additional bugs in total within Top-5. That is, the developer-provided programs and randomly generated programs via Csmith can complement our search-based witness programs to some degree. If we can effectively synthesize them, the compiler bug isolation effectiveness may be improved. We analyzed why randomly generated programs can isolate additional bugs, and found the reason to be that the tool Csmith cannot cover all C language features, causing some files always uncovered. When the bug occurs at these files, due to such an occasional factor, the files are easy to isolate using Csmith generated passing programs. Getting rid of this occasional factor, we made the first attempt in the direction by synthesizing our search-based witness programs and the developer-provided witness programs.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Technique</th>
<th>Top-1</th>
<th>Top-5</th>
<th>Top-10</th>
<th>Top-20</th>
<th>MFR</th>
<th>MAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCC</td>
<td>DiWi</td>
<td>5</td>
<td>17</td>
<td>30</td>
<td>36</td>
<td>13.93</td>
<td>14.76</td>
</tr>
<tr>
<td></td>
<td>SBFL</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>276.22</td>
<td>276.22</td>
</tr>
<tr>
<td></td>
<td>SBFL\textsuperscript{dev}</td>
<td>3</td>
<td>66.67</td>
<td>12</td>
<td>41.67</td>
<td>25</td>
<td>20.00</td>
</tr>
<tr>
<td></td>
<td>SBFL\textsuperscript{rand}</td>
<td>1</td>
<td>400.00</td>
<td>12</td>
<td>41.67</td>
<td>22</td>
<td>36.36</td>
</tr>
<tr>
<td></td>
<td>DiWi\textsuperscript{rand}</td>
<td>3</td>
<td>66.67</td>
<td>14</td>
<td>21.43</td>
<td>24</td>
<td>25.00</td>
</tr>
<tr>
<td>LLVM</td>
<td>DiWi</td>
<td>5</td>
<td>20</td>
<td>30</td>
<td>35</td>
<td>14.60</td>
<td>14.76</td>
</tr>
<tr>
<td></td>
<td>SBFL</td>
<td>1</td>
<td>400.00</td>
<td>3</td>
<td>566.67</td>
<td>3</td>
<td>900.00</td>
</tr>
<tr>
<td></td>
<td>SBFL\textsuperscript{dev}</td>
<td>1</td>
<td>400.00</td>
<td>10</td>
<td>100.00</td>
<td>20</td>
<td>50.00</td>
</tr>
<tr>
<td></td>
<td>SBFL\textsuperscript{rand}</td>
<td>5</td>
<td>-</td>
<td>12</td>
<td>66.67</td>
<td>22</td>
<td>36.36</td>
</tr>
<tr>
<td></td>
<td>DiWi\textsuperscript{rand}</td>
<td>3</td>
<td>66.67</td>
<td>15</td>
<td>33.33</td>
<td>24</td>
<td>25.00</td>
</tr>
<tr>
<td>ALL</td>
<td>DiWi</td>
<td>10</td>
<td>37</td>
<td>60</td>
<td>71</td>
<td>14.27</td>
<td>14.76</td>
</tr>
<tr>
<td></td>
<td>SBFL</td>
<td>1</td>
<td>900.00</td>
<td>3</td>
<td>1,133.33</td>
<td>3</td>
<td>1,900.00</td>
</tr>
<tr>
<td></td>
<td>SBFL\textsuperscript{dev}</td>
<td>4</td>
<td>150.00</td>
<td>22</td>
<td>68.18</td>
<td>45</td>
<td>33.33</td>
</tr>
<tr>
<td></td>
<td>SBFL\textsuperscript{rand}</td>
<td>6</td>
<td>66.67</td>
<td>24</td>
<td>54.17</td>
<td>44</td>
<td>36.36</td>
</tr>
<tr>
<td></td>
<td>DiWi\textsuperscript{rand}</td>
<td>6</td>
<td>66.67</td>
<td>29</td>
<td>27.59</td>
<td>48</td>
<td>25.00</td>
</tr>
</tbody>
</table>

* Columns "Top-n" present the improvement rates of DiWi over a compared technique in terms of various measurements.
What developers want. We conducted a survey by communicating with 7 compiler developers to understand their preferences and needs for compiler debugging tools. Developers expressed their willingness to improve the debugging process by identifying buggy program parts and providing effective witness programs. In such a way, DiWi contributes to providing a set of effective witness programs to isolate bugs. Furthermore, developers also expressed their willingness to improve the debugging process by identifying buggy program parts and providing effective witness programs. DiWi is practical and shows strong desire for DiWi by using the words “can’t wait to see” during the communications. Even the developer who does not first identify buggy files when debugging, also expresses his/her willingness to improve the debugging process by using DiWi. In the future, we will improve DiWi at finer granularity such as the method level.

### DISCUSSION

**What developers want.** To investigate the practicability of DiWi, we conducted a survey by communicating with 7 compiler developers (sending out 10 requests in total) from 4 international companies building their own compilers (including the LLVM team). Developers confirmed that their compiler bug debugging process starts from buggy files identification and this step is time-consuming, indicating the necessity of compiler bug isolation at the file level. Moreover, developers think the effectiveness of DiWi (shown in our study) is practical and show strong desire for DiWi by using the words “can’t wait to see” during the communications. Even the developer who does not first identify buggy files when debugging, also expresses his/her willingness to improve the debugging process by using DiWi. In the future, we will improve DiWi at finer granularity such as the method level.

#### Test oracle challenge.

We present the used test oracles to determine whether a mutated program is passing in Section 4.1. However, some oracles are not absolutely precise, especially for wrong-code bugs, DiWi treats the mutated program producing consistent results as a passing program, but it may still trigger the bug. For example, Figure 9a shows a failing program, where the outputs (of b) under -O0 and -O1 are different. The other two figures are two mutated passing programs under the used oracle. However, both of them are fake passing programs. In Figure 9b, the variable b in printf statement is mutated to be c. It still triggers the bug, but it is regarded as a passing program since c prints the same results under -O0 and -O1. In Figure 9c, the value of b (i.e., 0) is mutated to be 1, which is equal to the initial value of b. Such mutations make the output always the same (i.e., 1), causing the used oracle to miss the bug. If a mutated passing program may impact the isolation effectiveness. However, it is challenging to solve this oracle problem. To reduce its impact, DiWi avoids the mutations that directly change the oracle (i.e., printf statements), but cannot deal well with other cases such as Figure 9c. In the future, we will introduce advanced data- and control-flow analysis [63] to address this challenge.

**Undefined behavior challenge.** Another challenge lies in undefined behaviors for compilers, which mean the semantics of certain operations are undefined in the programming-languages standards [30]. If a program contains undefined behaviors, compilers may produce varied results. Identifying undefined behaviors is a difficult challenge in compiler research [49, 78]. It is also a threat in our work, since our mutation may introduce undefined behaviors. However, undefined behaviors tend to impact bug detection, since different results of a “failing” program may be caused by real bugs or undefined behaviors. In DiWi, we only kept the passing programs with the same results to isolate bugs, and thus the threat may not be serious. We will relieve this problem by adopting existing light-weight methods [49] for identifying undefined behaviors.

### Conclusion

In this paper, we propose a novel compiler bug isolation technique, DiWi, which proposes a heuristic-based search strategy to carefully generate a set of effective witness programs by performing our designed witness mutation rules on the given failing program. The results on 90 real bugs for GCC and LLVM show, DiWi isolates 78.89% of the studied bugs within 20 compiler files, significantly outperforming state-of-the-art SBFL. DiWi is general and not limited to compilers, we plan to apply it to other systems taking structurally complex test inputs, e.g., operating systems and browsers.
REFERENCES


