AutoCodeRover: Autonomous Program Improvement

Yuntong Zhang
yuntong@comp.nus.edu.sg
National University of Singapore, Singapore

Haifeng Ruan
hruan@comp.nus.edu.sg
National University of Singapore, Singapore

Zhiyu Fan
zhiyufan@comp.nus.edu.sg
National University of Singapore, Singapore

Abhik Roychoudhury
abhik@comp.nus.edu.sg
National University of Singapore, Singapore

ABSTRACT

Researchers have made significant progress in automating the software development process in the past decades. Automated techniques for issue summarization, bug reproduction, fault localization, and program repair have been built to ease the workload of developers. Recent progress in Large Language Models (LLMs) has significantly impacted the development process, where developers can use LLM-based programming assistants to achieve automated coding. Nevertheless software engineering involves the process of program improvement apart from coding, specifically to enable software maintenance (e.g. program repair to fix bugs) and software evolution (e.g. feature additions). In this paper, we propose an automated approach for solving Github issues to autonomously achieve program improvement. In our approach called AutoCodeRover, LLMs are combined with sophisticated code search capabilities, ultimately leading to a program modification or patch. In contrast to recent LLM agent approaches from AI researchers and practitioners, our outlook is more software engineering oriented. We work on a program representation (abstract syntax tree) as opposed to viewing a software project as a mere collection of files. Our code search exploits the program structure in the form of classes/methods to enhance LLM's understanding of the issue's root cause, and effectively retrieve a context via iterative search. The use of spectrum based fault localization using tests, further sharpens the context, as long as a test-suite is available. We consider the recently proposed SWE-bench-lite which consists of 300 real-life Github issues involving bug fixing and feature additions. Experiments on SWE-bench-lite show increased efficacy in solving Github issues (22-23% on SWE-bench-lite), as compared to recent efforts from the AI community. Interestingly, our approach resolved 67 GitHub issues in less than 12 minutes each, whereas developers spent more than 2.77 days on average. On the full SWE-bench consisting of 2294 GitHub issues, AutoCodeRover solved around 16% of issues, which is higher than the efficacy of the recently reported AI software engineer Devin from Cognition Labs, while taking time comparable to Devin. We posit that our workflow enables autonomous software engineering, where, in future, auto-generated code from LLMs can be autonomously improved.

CCS CONCEPTS

• **Software and its engineering** → **Automatic programming**; **Software testing and debugging**; • **Computing methodologies** → **Natural language processing**.

1 BEYOND AUTOMATIC PROGRAMMING

Automating software engineering tasks has long been a vision among software engineering researchers and practitioners. One of the key challenges has been the handling of ambiguous natural language requirements, in the process of automatic programming. In addition, there has been progress in some other software engineering activities such as automated test generation [4, 5], automated program repair [9], and so on.

Recent progress in large language models (LLMs) and the appearance of tools like Github Copilot [26] hold significant promise in automatic programming. This progress immediately raises the question of whether such automatically generated code can be trusted to be integrated into software projects, and if not, what improvements to the technology are needed. One possibility is to automatically repair generated code to achieve *trust*. This brings out the importance of automating program repair tasks towards achieving the vision of autonomous software engineering.

Given this motivation of automating program repair, and the large number of hours developers often spend manually fixing bugs, we looked into the possibility of fully autonomous program improvement. Specifically, we feel that bug fixing and feature addition are the two key categories of tasks that a development team may focus on when maintaining an existing software project. To achieve this goal, we proposed an approach that augments LLM with context knowledge from the code repository. We call our tool AutoCodeRover.

Technically our solution works as follows. Given a real-life GitHub issue, LLM first analyzes the attached natural language description to extract keywords that may represent files/classes/methods/code snippets in the codebase. Once these keywords are identified, we employ a stratified strategy for the LLM agent to retrieve code context by invoking multiple necessary code search APIs at one time with the keyword combinations as arguments (e.g., search_method_in_file). These code search APIs are running locally based on AST analysis and are responsible for retrieving code context such as class signatures and method implementation details from a particular location in the codebase. By collecting the project context with code search APIs, LLM refines its understanding of the issue based on the currently available context. Note that our code context retrieval will proceed in an iterative fashion. The LLM agent directs the navigation and decides which code search APIs to use (i.e. where/what to retrieve code) in each iteration based on the current available context returned from the previous API calls. AutoCodeRover then enquires whether there is sufficient project context, and subsequently uses the collected context to derive the buggy locations. The patch construction is then handled by another

LLM agent which considers the buggy locations as well as all the context collected so far for those locations.

AutoCodeRover also can leverage debugging techniques such as spectrum-based fault localization (SBFL) [35] to decide more precise code search APIs for context retrieval if a test suite accompanying the project is available. SBFL primarily considers the control flow of the passing and failing tests and assigns a suspiciousness score to the different methods of the program. The LLM agent may prioritize retrieving context from particular methods and classes if the fault localization result is provided, e.g., when a method appears both in the issue description and in the output of fault localization. In the last step, AutoCodeRover may perform patch validation using available tests, to determine whether the patch produced by AutoCodeRover passes the tests in the given test-suite. Otherwise, AutoCodeRover will rerun the patch generation with a retry limit until a correct patch is found.

Contributions. Our contribution lies in the effective use of code search to make software engineering processes like program repair autonomous. We demonstrate this capability by autonomously solving GitHub issues. We report favorable experimental results on the SWE-bench dataset [13]. We achieve 15.95% efficacy on the full SWE-bench with 2294 GitHub issues, and 22.33% efficacy on the SWE-bench-lite subset with 300 GitHub issues. Overall, we see the need to endow a *software engineering oriented outlook* to the recent flurry of activity on LLM agents for software engineering, which mostly has only an AI flavor. We believe that the software engineering angle can be conveyed via (at least) the following five dimensions, and can enhance the capabilities of LLM agents.

- We work on program representations (abstract syntax tree or AST) as opposed to viewing a software project as a collection of files. We posit that working on program representations like AST will be useful in autonomous software engineering workflows.
- To solve GitHub issues, we focus on code search in a way that resembles the activity of a human software engineer. So we try to use the program structure - classes, methods, code snippets - in searching for relevant code context. This leads to a more effective usage of the context provided to LLM.
- We posit that higher efficacy of automated repair is more important than time efficiency, *as long as the time is within a threshold.* It is well-known in empirical software engineering research that time limits of 30-60 minutes for automated repair are tolerable, based on extensive developer surveys in the field [25]. We thus report 22% efficacy on SWE-bench-lite in solving GitHub issues, within 12 minutes. We can compare this timelimit of 12 minutes to the average time of 2.77 days to fix the GitHub issues manually.
- One should be able to exploit debugging techniques like test-based fault localization to guide the search for code in resolving GitHub issues. We show that use of fault localization in setting the code context leads to an increase in efficacy of AutoCodeRover in solving GitHub issues.
- Finally, it is also useful to examine how many of the solved GitHub issues are producing acceptable patches. We study the patches produced by AutoCodeRover and report that 2/3 of the autonomously produced patches from AutoCodeRover are correct and acceptable. We note that this aspect has not been reported by Devin [18] and SWE-agent [38].

2 RELEVANT LITERATURE

2.1 Program Repair

Test-suite based automated program repair (APR) has attracted significant attention in the last decade [9]. These techniques aim to generate a patch for a buggy program to pass a given test-suite. APR techniques typically include search-based, semantic-based, and pattern/learning-based APR. Search-based APR techniques like GenProg [33] take a buggy program and generate patches using predefined code mutation operators, or search for a patch over the patch space that passes the given test suite. Semantics-based APR techniques [22, 24] generate patches by formulating a repair constraint that needs to be satisfied based on a given test suite specification, and then solving the repair constraint to generate patches. Learning-based APR techniques [20, 30, 39] often train a deep learning model with large code repositories and are guided by a specific representation of code syntax and semantics to predict the next tokens that are most likely to be correct patch. There are also works [17, 23, 31] that tried to leverage GitHub issues and bug reports to improve APR effectiveness.

Recent work [7, 11, 37] have shown the use of LLMs for automated program repair. This line of work often assumes the buggy program statements are given (i.e. perfect fault localization assumption) and focuses on constructing APR-specific prompts that guide LLM to generate a patch for the selected buggy program statements multiple times until a patch that passes all tests is found. However, obtaining buggy locations for a large project is an essential and challenging task in resolving real-life bug reports.

APR techniques have been successfully deployed in industries for domain-specific bug fixing [3, 21, 34]. However, a long-standing challenge for the APR techniques is to resolve general real-life software issues from scratch. The above APR techniques rely on a high-quality test suite which is not always available in the real world and they do not leverage the valuable natural language specification from the original problem description. To address these challenges and achieve autonomous software engineering, we focus on resolving GitHub issues from a real-life dataset.

2.2 Dataset

SWE-bench [13] is a benchmark that aims to evaluate the capabilities of large language models in resolving end-to-end real-life software engineering tasks. The benchmark consists of 2,294 real-life software engineering task instances collected from the repositories of 12 popular large Python projects [1] (e.g., django, sympy). Each SWE-bench task instance contains a pair of Github issue and corresponding pull requests. The GitHub issue either reports a bug to be fixed or requests to implement new features. The pull request includes the code changes made by human developers to resolve the issue and test cases that prevent the issue. Unlike traditional code generation tasks in HumanEval [6] and MBPP [19] benchmark, resolving a SWE-bench instance is particularly challenging because it requires automatically generating code changes that address the problem in a GitHub issue for a matured large code repository based only on the issue description. More specifically, the process may involve a series of complex tasks like reasoning the target bug

[1] https://www.swebench.com/index.html

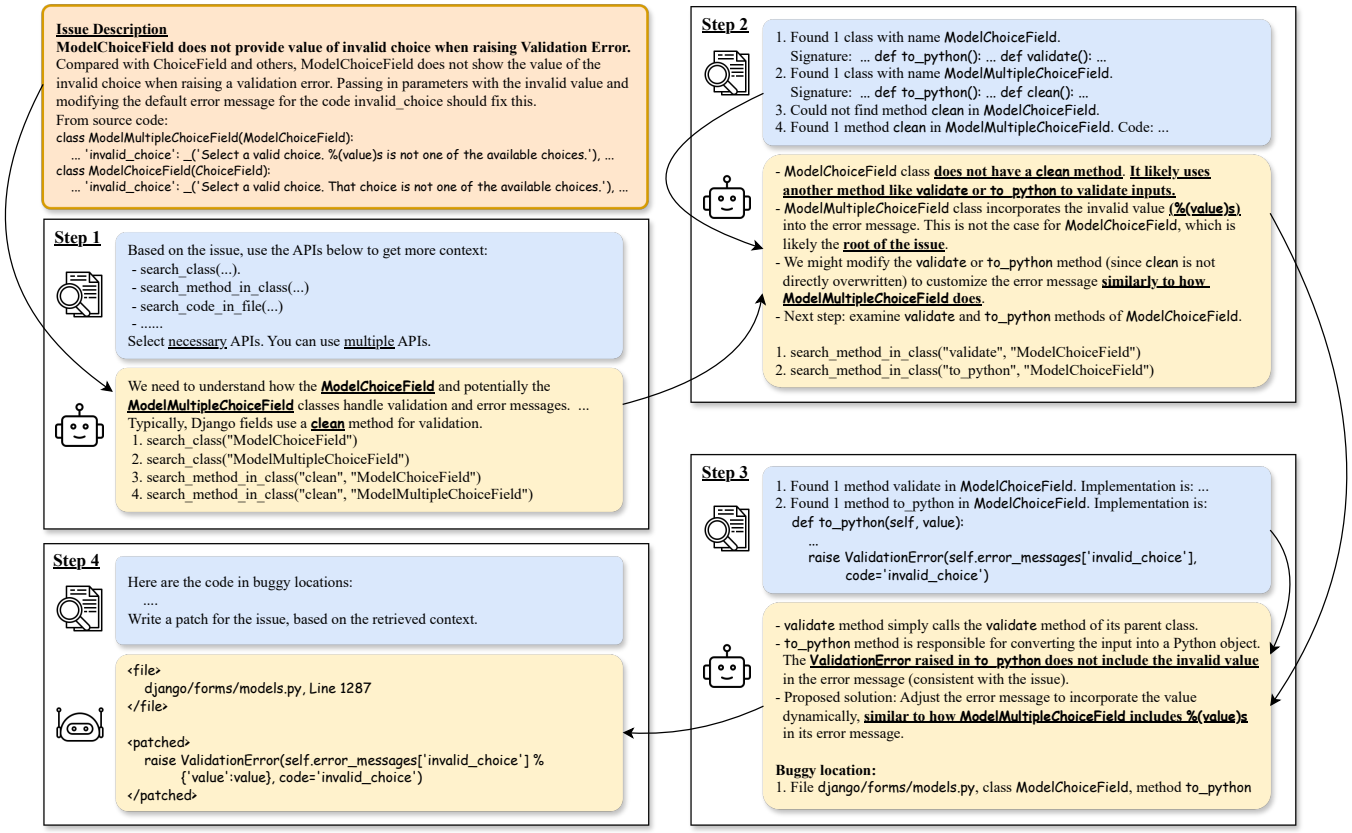


Figure 1: Issue description of django-13933 and AutoCodeRover's workflow on it.

location across files in the code repository, analyzing the root cause of the issue, proposing bug-fixing strategies, and eventually writing a patch that passes all the test cases added in the pull request.

SWE-bench team also constructed a lite version of SWE-bench (SWE-bench lite) due to the high cost of evaluating in the complete SWE-bench. SWE-bench lite includes 300 task instances sampled from SWE-bench, following a similar repository distribution. SWE-bench lite removed task instances with meta files such as images, external links, and instances without proper description (e.g., short problem statement which is fewer than 40 words).

3 MOTIVATING EXAMPLE

In this section, we illustrate how our tool AutoCodeRover can collect code context and generate a patch from an issue description and the corresponding project code-base. We show an example of a feature addition task. Figure 1 demonstrates the workflow of AutoCodeRover on an issue submitted to the Django issue tracker[2]. This issue is classified as "New feature" in the issue tracker, and is included in SWE-bench lite with the id "django-13933". The first part of Figure 1 shows the issue description (the code part is simplified for brevity). This issue requests adding support to the `ModelChoiceField` class, so that it "shows the value of the invalid choice when raising a validation error".

With this issue description, AutoCodeRover operates in two stages - *context retrieval* (Step 1-3 in Figure 1) and *patch generation* (Step 4). First, a context retrieval LLM agent is instructed to collect the relevant code context related to this issue, from a local copy of the Django project codebase. This retrieval is achieved by inferring relevant names (e.g. `ModelChoiceField`) and searching for them in the Abstract Syntax Tree (AST) of the project. A set of retrieval APIs are provided to the agent. In Step 1, the agent identifies the classes `ModelChoiceField` and `ModelMultipleChoiceField` might be relevant based on the issue description, and infers that a `clean` method might be also of interest for validation-related issues. It then invokes the `search_class` and `search_method_in_class` APIs to retrieve more information about them.

In Step 2, API invocations return the signature of the classes and implementation of the methods found. Results of invocation #3 reveals the absence of method `clean` in the class `ModelChoiceField`, which helps to *refine* the agent's understanding from Step 1. The returned class signature from invocation #1 also reveals two more relevant methods `to_python` and `validate`, which were not mentioned in the issue description. This suggests the retrieval should be performed *iteratively* in multiple steps, so that results from a previous search can become arguments of the following search. In this example, the agent then iteratively invokes `search_method_in_class`

[2] https://code.djangoproject.com/ticket/32347

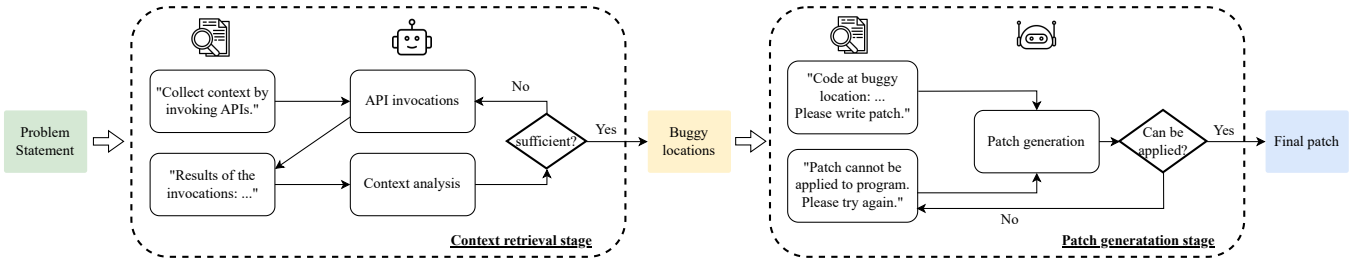


Figure 2: Overall workflow of AutoCodeRover.

```
""" AutoCodeRover patch. """
except (ValueError, TypeError, self.queryset.model.DoesNotExist):
-    raise ValidationError(self.error_messages['invalid_choice'],
-        code='invalid_choice')
+    % Include the invalid value in the error message
+    raise ValidationError(
+        self.error_messages['invalid_choice'] % {'value': value},
+        code='invalid_choice'
+    )

""" Developer patch. """
except (ValueError, TypeError, self.queryset.model.DoesNotExist):
-    raise ValidationError(self.error_messages['invalid_choice'],
-        code='invalid_choice')
+    raise ValidationError(
+        self.error_messages['invalid_choice'],
+        code='invalid_choice',
+        params={'value': value},
+    )
```

Figure 3: AutoCodeRover and developer patch for Django-13933.

on the two newly revealed methods. Furthermore, by referencing results from multiple invocations, the agent can infer that `ModelMultipleChoiceField` incorporates the invalid value into the message with `%(value)s`, and methods in `ModelChoiceField` can be modified *similarly* to `ModelMultipleChoiceField`.

In Step 3, the agent receives the implementation of `validate` and `to_python` methods. Among the two methods, it selects `to_python` to be the more suitable place to make changes, since `to_python` raises the relevent exception and does not include the invalid value. At this point, the retrieval agent deems the collected code context as sufficient for understanding the issue and drafting the patch. The identified buggy location, together with the gathered context and analysis so far, is passed to another patch generation agent. This agent is instructed to write patches following the format described in Step 4 (see yellow box) of Figure 2. In Step 4, a patch is written to allow a `value` to be integrated into the error message, by utilizing %-formatting in Python. Although written in a different way compared to the developer patch, this patch achieves similar effect and passes the developer provided test-suite for this issue.

4 AI PROGRAM IMPROVEMENT FRAMEWORK

In this section, we discuss the design of AutoCodeRover. AutoCodeRover is a system incorporating AI agents for program improvement tasks in large software projects. AutoCodeRover is designed to work in a realistic software development lifecycle, in which users submit issue reports to a software repository describing a bug, and the project maintainers craft a patch to resolve the issue. With a submitted issue, AutoCodeRover autonomously analyzes the submitted issue, retrieves the relevant code context in the software project, and generates a patch. This patch can then be vetted by human developers. If such a tool can automatically handle a certain percentage of the issues awaiting developers' attention, manual efforts are reduced.

4.1 Overview

We first describe the overall stages AutoCodeRover operates in, and will proceed in more details in the rest of this section. The overall workflow of AutoCodeRover is shown in Figure 2. AutoCodeRover takes in as input a problem statement P of the issue to be resolved, and a codebase C of the corresponding software project. This problem statement P contains the title and description of the issue, as shown in Section 3. From a problem statement written in natural language, AutoCodeRover analyzes the requirement from it and proceeds in two main stages, which are *context retrieval* and *patch generation*.

In the *context retrieval* stage, AutoCodeRover employs an LLM agent to navigate through a potentially large codebase C and extract the relevant code snippets relevant to P. This navigation is facilitated by a set of *context retrieval APIs* (Section 4.2), which enables an LLM to retrieve information about the project (e.g. class signatures) and actual code snippets (e.g. method implementations). The context retrieval agent directs this navigation, and decides which retrieval APIs to use based on the current available context. In order to make the LLM-directed navigation more "controlled", we devise a stratified stategy of invoking the retrieval APIs (Section 4.3). This stratified strategy instructs the LLM to only invoke necessary retrieval APIs based on the available information, and iteratively changes the set of retrieval APIs used when more code-related context are returned from the previous API calls.

Once the context retrieval agent has gathered sufficient context about the issue, AutoCodeRover proceeds to the *patch generation* stage (Section 4.5). In this stage, AutoCodeRover employs another LLM agent to extract more precise code snippets from the retrieved context, and craft a patch based on the extracted code snippets. This patch generation agent is instructed to craft a patch in a specific format, and if the produced patch does not follow the format specification or cannot be applied to the original codebase, the agent enters a retry-loop which terminates after a pre-configured number

Table 1: List of Context Retrieval APIs.

API name	Description	Output
search_class (cls)	Search for class `cls` in the codebase.	Signature of the searched class.
search_class_in_file (cls, f)	Search for class `cls` in file `f`.	Signature of the searched class.
search_method (m)	Search for method `m` in the codebase.	Implementation of the searched method.
search_method_in_class (m, cls)	Search for method `m` in class `cls`.	Implementation of the searched method.
search_method_in_file (m, f)	Search for method `m` in file `f`.	Implementation of the searched method.
search_code (c)	Search for code snippet `c` in the codebase.	Region of code surrounding the searched snippet.
search_code_in_file (c, f)	Search for code snippet `c` in file `f`.	Region of code surrounding the searched snippet.

(Issue sympy-12481)
`Permutation` constructor fails with non-disjoint cycles.

Calling `Permutation([[0,1],[0,1]])` raises a `ValueError` instead of constructing the identity permutation. If the cycles passed in are non-disjoint, they should be applied in left-to-right order and the resulting permutation should be returned.

This should be easy to compute. I don't see a reason why non-disjoint cycles should be forbidden.

(Issue django-13230)
Add support for `item_comments` to syndication framework.

Add comments argument to `feed.add_item()` in `syndication.views` so that `item_comments` can be defined directly without having to take the detour via `item_extra_kwargs`.
Additionally, comments is already explicitly mentioned in the `feedparser`, but not implemented in the view.

Figure 4: Issue description of sympy-12481 and django-13230. "Hints" are highlighted.

of attempts. Finally, AutoCodeRover outputs a patch that attempts to resolve the original issue.

We note that the workflow of AutoCodeRover discussed so far only requires the problem statement P and codebase C as input, and do not require any program specifications such as testcases. However, when testcases are available (e.g. provided by developers or generated from another tool), execution-based information and program analysis techniques can be integrated into the AutoCodeRover framework (Section 4.4). For example, statistical fault localization [35] tool can be used to reveal more relevant methods beyond those mentioned in the problem statement. The additionally revealed code context can thereby influence the set of retrieval APIs invoked in the *context retrieval* stage. Moreover, with testcases available, the patch generation agent can employ an additional patch validation step in its retry-loop when crafting a patch. In the remainder of this section, we discuss in greater detail the components of AutoCodeRover and their design considerations.

4.2 Context Retrieval APIs

In the *context retrieval* stage, the basic components are a set of APIs which an LLM agent can use to gather relevant code context and snippets from the codebase. For typical software project issues, we observe that users often mention some "hints" on which part of the codebase is relevant. These hints can be the names of the relevant methods, classes, or files, and sometimes also contain short code snippets. Although these hints may not directly point to the precise location for code modification, they often reveal code context in the project that is relevant to the current issue. Figure 4 illustrates two real-world issues submitted to the sympy and django projects, respectively, and the "hints" are highlighted. In sympy-12481, the class `Permutation` is mentioned twice; in django-13230, multiple hints are mentioned, such as code snippet `feed.add_item()`, package path `syndication.views`, method name `item_extra_kwargs` and etc. Based on the type of project and code related hints, we design a set of APIs for an LLM agent to retrieve code context from these hints. The current set of APIs in AutoCodeRover and their outputs are shown in Table 1. Once invoked by the LLM agent, the retrieval APIs search for classes, methods and code snippets in the code-base, and return the results back to the agent. To avoid forming very lengthy code context that creates distraction during patch generation, we return only necessary information as API outputs. For example, since the complete definition of a class can be lengthy in large projects, we only return signature of the class as output for `search_class` and `search_class_in_file`. Returning the signature to shorten cntext, is a better approach than cutting off the context at a certain bound. Upon receiving the class signature, the agent could then invoke another API to search for the relevant methods / snippets inside the class.

Interfacing with LLM. For the LLM agent to invoke the context retrieval APIs, the description and expected output of them are presented to it as part of prompt. When the agent decides to invoke a set of retrieval APIs, it responds with the list of API call names and the corresponding arguments. These retrieval API requests are processed locally by parsing a local codebase of the project into AST and searching over it. Results of locally executing these APIs are returned to the agent, forming the code context.

4.3 Stratified Context Search

The set of context retrieval APIs listed in Table 1 serves as building blocks for searching relevant code context. With the context retrieval APIs and the LLM-identified keyword "hints" from the problem statement, a set of possible *API invocations* can be derived by using the identified keywords as API parameters. We discuss a few observations on using these API invocations to gather code context, and propose a *stratified* context retrieval process.

Our first observation is that the context retrieval should not be restricted to a single API invocation. For example, in the issue django-13230 mentioned in Figure 4, if the retrieval starts from

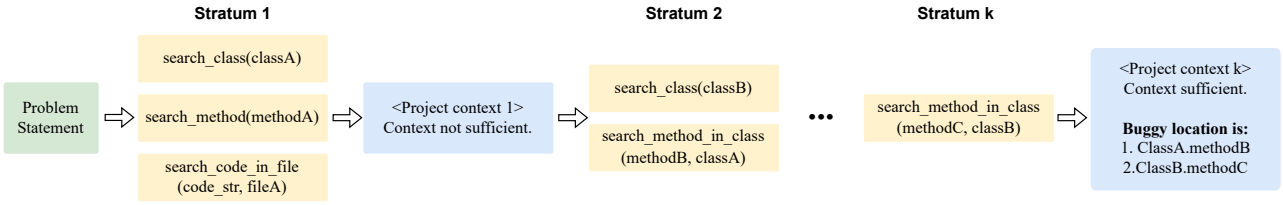


Figure 5: Stratified search with retrieval APIs for context gathering.

the invocation `search_method("add_item")`, the implementation of method `add_item` can be considered by the LLM agent as a sufficient context, since it appears to be relevant to the problem statement. However, searching from only one method can lead to incomplete context for the agent to reason about root cause of the problem. On the other hand, if all API invocations are executed at once, a large code context can be retrieved, especially when the problem statement mentions many class and method names. This large code context can be difficult for an LLM to comprehend, or may even exceed its context window.

The second observation is that some of the API invocation results provide more elements to build new possible API invocations, which means the process of invoking retrieval APIs should be iterative. For example, the result of a `search_class` call returns the method signatures within the searched class, and an LLM can iteratively invoke method-related APIs afterwards.

With these two observations, we propose a *stratified* search process for invoking context retrieval APIs, as illustrated in Figure 5. From a problem statement, stratified search iteratively invokes retrieval APIs to gather project code context, and finally outputs a list of potentially buggy locations to be fixed. In each stratum, we prompt the LLM agent to select a set of *necessary* API invocations, based on the current context. In stratum 1, the current context only contains the problem statement; in the following strata, the context contains both the problem statement and the code searched so far. By allowing LLM to select more than one API invocations and also instructing it to only select the necessary ones, we make the best use of the context, building what we deem an *optimal* context.

After the API invocations in a stratum are executed, the newly retrieved code snippets are added to the current context. The LLM agent is then prompted to analyze whether the current context is sufficient for understanding the issue, thereby deciding whether (a) we continue the iterative search process, or (b) we decide on the buggy locations which will be considered for fixing.

4.4 Analysis-Augmented Context Retrieval

We also investigate how program debugging techniques can augment our workflow. Specifically, we integrate the *Spectrum-based Fault Localization* (SBFL) analysis into AutoCodeRover to study the effect of such a test-based dynamic analysis. We make a test-suite T available to AutoCodeRover, in addition to the problem statement P and codebase C.

Spectrum-based Fault Localization. The goal of SBFL is to identify the location of software faults [36]. Given a test-suite T containing both passing and failing tests, SBFL considers control-flow differences in the passing and failing test executions, and assigns a suspiciousness score to different program locations. This suspiciousness score can be computed with various metrics such as Tarantula [15] and Ochiai [2]. Program elements (e.g. statements/basic blocks) with the highest suspiciousness scores are identified as likely fault locations. SBFL can be performed at different granularities of program elements, such as statements or basic blocks. Since the LLM can process reasonably long code snippets, we use *method-level* SBFL in AutoCodeRover. Given an test-suite, SBFL can be used to directly output a few program locations to be repaired. However, the accuracy of SBFL highly relies on the quality of the test-suite [16] - since the SBFL results are effectively an abstraction of the differential of control flows between passing tests and failing tests. Therefore, instead of replacing the stratified context retrieval with SBFL, we use the SBFL identified methods to augment the search process. Before AutoCodeRover enters the context retrieval stage, we provide the SBFL-identified methods to the LLM agent as "results from an external analysis tool that identifies suspicous code". The main role of SBFL-identified methods is to reveal more "hints" on relevant classes and methods beyond those mentioned in the problem statement. The LLM agent can then use the context retrieval APIs to examine these methods. Since the SBFL-identified methods are presented to the agent together with the problem statements, the agent can then cross-reference between these two sources of information. For example, if one of the SBFL-identified method names is more closly related to the problem statement, the LLM is more likely to invoke the `search_method` API on this name. We will demonstrate this observation in Section 6.2.

4.5 Patch Generation

In the *patch generation* stage, AutoCodeRover employs a *patch generation agent* to use the collected code context to write a patch for the problem statement. This agent is given the problem statement, the identified buggy locations/methods, and the history of context retrieval, including the invoked APIs, the API results, as well as the previous analysis on the code context made by the context retrieval agent.

As a first step, the patch generation agent retrieves the precise code snippets at the buggy locations from the codebase. Based on the precise code snippets and other relevant code context, the agent enters a retry-loop of generating patches. If a generated patch does not follow the specified patch format or could not be applied syntactically to the original program, the agent is prompted to retry. We also employs a linter to indentify Python-specific syntax issues such as indentation errors in this retry-loop. The agent is allowed to

retry up to a pre-configured attempt limits (currently set to three), after which the best patch so far is returned as output.

5 EXPERIMENT SETUP

To evaluate the capabilities of AutoCodeRover in resolving real-life software issues, we answer the following research questions.

RQ1: To what extent can AutoCodeRover automate software issues like human developers?
RQ2: Can existing debugging / analysis techniques assist AutoCodeRover?
RQ3: What are the challenges for AutoCodeRover and fully automated program improvement in future?

Benchmark. We evaluate AutoCodeRover in recently proposed benchmarks SWE-bench and SWE-bench lite [13], comprising 2294 and 300 real-life GitHub issues, respectively. The only input is the natural language description in the original GitHub issue and its corresponding buggy codebase. Details of SWE-bench and SWE-bench lite appear in Section 2.2.

Baseline and Evaluation Metric. We selected two LLM-based agent systems Devin [18] and Swe-agent [38] as baselines and compare their performance against AutoCodeRover. Swe-agent is publicly available as a GitHub repository[3], so we replicated it as Swe-agent-rep in SWE-bench lite with the default setting based on the provided scripts. In contrast to Swe-agent, we do not have access to Devin, so we take the most relevant reported result from their technical report [32]. To avoid the natural randomness of LLM, we repeat our experiments with AutoCodeRover and Swe-agent-rep three times when possible, and report the average and total numbers across three repetitions. We use (1) the percentage of resolved instances, (2) average time cost, and (3) average token cost to evaluate the effectiveness of the tools. These evaluation metrics represent overall effectiveness, time efficiency, and economic efficacy in resolving real-world GitHub issues.

Implementation and Parameters. We use the state-of-the-art OpenAI GPT-4 (gpt-4-0125-preview) as foundation inference model for both AutoCodeRover and Swe-agent-rep. In AutoCodeRover, the GPT-4 model is responsible for selecting search APIs to retrieve codebase context, refining the issue description, and writing a final patch. For parameters of GPT-4, we set a low temperature=0.2, max_tokens=1024 to produce relatively deterministic results and enable sufficient reasoning length for AutoCodeRover, and all other parameters remain as per default. For Swe-agent-rep, we increased the default cost limit of interacting with GPT-4 model (for resolving a task instace) to two USD and kept other settings as default. Note that, both two tools do not have a time limit, AutoCodeRover terminates either a patch is generated or the context retrieval stage repeats ten times. Swe-agent-rep terminates either a patch is generated or the cost of resolving an issue reaches two USD. Experiment results and artifacts for AutoCodeRover and Swe-agent-rep can be found at https://github.com/nus-apr/auto-code-rover.

System Environment. All experiments are conducted on an x86_64 Linux server with Ubuntu 20.04 installed.

[3]https://github.com/princeton-nlp/SWE-agent

6 EVALUATION

6.1 RQ1: Overall Effectiveness on SWE-bench

We first measure the overall effectiveness of AutoCodeRover and baselines with the number of resolved task instances in SWE-bench. With the goal of understanding to what extent the current AI systems can automatically resolve real-life software issues, the only inputs we provided are the natural language issue description and a local code repository checked out at the erroneous version. We repeated the experiment of AutoCodeRover three times and presented the average and total number of resolved software issues across the three runs. The average and total results are denoted as ACR-avg and ACR-all respectively (for brevity, we use ACR to denote AutoCodeRover in this section). When reporting time and token/cost for ACR-all, we report the time and cost required for running each task three times. Since Devin was evaluated on a random 25% subset of SWE-bench [18], we also report results of AutoCodeRover on this subset (we refer to this as "SWE-bench Devin subset"). Table 2 shows the overall result in full SWE-bench, SWE-bench Devin subset, SWE-bench lite respectively. Figure 6 shows a visual summary of AutoCodeRover's comparison with Swe-agent and Devin.

Results on SWE-bench. In the full SWE-bench, ACR-all (union of from the three ACR runs) resolved 15.95% task instances taking a time of 701 seconds per task (less than 12 minutes). In comparison, Swe-agent resolved 12.29% tasks in full SWE-bench, according to their report [12] (we did not replicate Swe-agent-rep on the full SWE-bench due to the high cost).

Comparison with Devin. The state-of-the-art closed-source baseline tool Devin [18] is evaluated in a random 25% subset of SWE-bench. To compare AutoCodeRover with Devin, we report AutoCodeRover's result on the 570 task instances Devin was evaluated on. The results of AutoCodeRover are taken from the ACR runs on full SWE-bench. In the SWE-bench Devin subset, the union of three runs of AutoCodeRover successfully resolved 15.79% of the task instances, which is higher than Devin. Figure 7a provides a more detailed exposition of the resolved tasks. Besides, the times taken by AutoCodeRover and Devin are comparable.

Results on SWE-bench lite. We performed another round of experiments with AutoCodeRover on SWE-bench lite (300 instances). The results reported in Table 2 indicate that on average ACR-avg can resolve 16.11% task instances in SWE-bench lite, which is at par with the reported results from Swe-agent [12]. We also investigated the union of all resolved tasks in the three runs, in which the percentage of resolved task instances increased to 22.33%.

Detailed Comparison with Swe-agent. Since Swe-agent is publicly available, we also attempted to run Swe-agent on SWE-bench lite. We replicated Swe-agent with two USD as cost budget for conversation with LLM per task instance in our environment (denote as Swe-agent-rep, it terminates either a patch is generated or reaches the two USD budget). Table 3 shows that when considering the union of all resolved tasks across three repetitions, AutoCodeRover resolved 22.33% out of the 300 tasks, whereas Swe-agent-rep resolved 14.67%. We further analyzed the commonly and uniquely resolved instances between AutoCodeRover and

Table 2: Overall Result of AutoCodeRover (ACR) and baselines on full SWE-bench, SWE-bench Devin subset, and SWE-bench lite. "-" indicates data not available.

Tools	Resolved Tasks	Avg Time	Avg Tokens
Reported result on full SWE-bench (size=2294)			
SWE-agent [12]	**12.29%** (282)	93	-
ACR-avg	10.59% (243)	234	40177 ($0.464)
ACR-all	**15.95%** (366)	701	120530 ($1.392)
Reported result on SWE-bench Devin subset (size=570)			
Devin [18]	**13.86%** (79)	> 600	-
ACR-avg	10.47% (59.7)	231	39068 ($0.452)
ACR-all	**15.79%** (90)	692	117204 ($1.357)
Reported result on SWE-bench lite (size=300)			
SWE-agent [12]	**17.00%** (51)	93	69976 ($0.739)[4]
ACR-avg	16.11% (48.3)	173	37602 ($0.435)
ACR-all	**22.33%** (67)	520	112806 ($1.304)

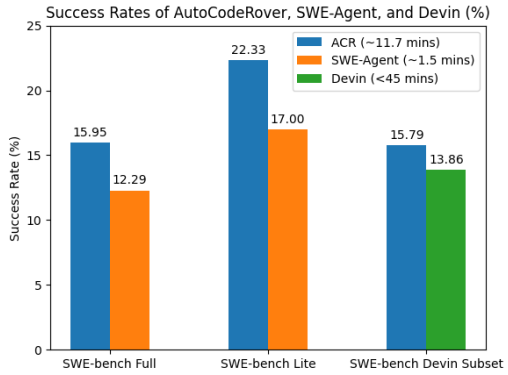


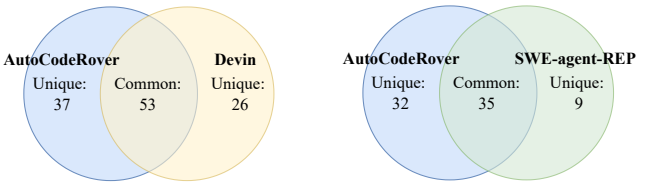
Figure 6: Summary of Results

SWE-agent-rep in Figure 7b, and we found that AutoCodeRover and SWE-agent-rep complement each other in different scenarios. AutoCodeRover uniquely resolved 32 task instances, which benefited from the fine-grained code context search at the AST level to precisely locate the bug locations (e.g., django-13401 searches three necessary methods at one time). On the other hand, the main reason that AutoCodeRover failed on the 9 unique instances resolved by SWE-agent-rep, is unimplemented search APIs (e.g., search_file invoked in django-12286). In such cases, AutoCodeRover can generate invalid search results when unimplemented APIs are invoked. This implies more robust search APIs are desired for future improvement. Since the evaluation environment is important to the result, to make the experimental replication complete and thorough, we also evaluated the generated patches of 300 task instances of SWE-bench lite within the SWE-agent docker environment which has been publicly released. However, Table 3 shows that the number of resolved issues for both ACR and SWE-agent-rep decreased, indicating that the docker released by SWE-bench team may need more work.

[4]This cell denotes the average cost per instance in our replication experiment (SWE-bench lite)

Table 3: The percentage of resolved task instances of AutoCodeRover and SWE-agent-rep on SWE-bench lite. The column "In SWE-agent Docker" represents the evaluation result under the docker environment released publicly by SWE-agent team on 6th April 2024.

Replication result on SWE-bench lite (size=300)				
	In our environment		In SWE-agent Docker	
	ACR	Agent-Rep	ACR	Agent-Rep
Run 1	16.00% (48)	9.33% (28)	10.00% (30)	6.67% (20)
Run 2	15.67% (47)	11.00% (33)	10.33% (31)	7.00% (21)
Run 3	16.67% (50)	9.33% (28)	10.67% (32)	6.00% (18)
All	**22.33% (67)**	14.67% (44)	14.00% (42)	9.00% (27)



(a) AutoCodeRover and Devin on SWE-bench Devin subset. **(b) AutoCodeRover and SWE-agent-rep on SWE-bench lite.**

Figure 7: Venn diagrams of resolved tasks instances.

Time / Token Cost. We are also interested in assessing the feasibility of deploying AutoCodeRover in the real world in terms of time and economic cost. On average, AutoCodeRover takes 234 seconds and 40177 tokens (equivalent to 0.464 USD) to resolve one task instance in SWE-bench. In comparison, our replication experiments with SWE-agent-rep cost 69976 tokens (equivalent to 0.739 USD) per task instance. When considering the combined three repetitions, AutoCodeRover takes 701 seconds (~11.7 minutes) per task, which is below the 30-60 minute time limit deemed acceptable by developers for automated repair tools [25]. Looking into the 67 issues resolved by AutoCodeRover in SWE-bench lite, it costs on average ~2.77 days for developers to create pull requests for 57 issues, and the other 10 issues take even longer to be closed by developers (ranging from 34 - 4023 days). The short response time and low cost show the significant potential for AutoCodeRover to act as a first step in future autonomous bug fixing.

Plausible / Correct patches. Overfitting is a well-known challenge in the Automated Program Repair community [8]. A program patch that passes the given test suite is said to be *plausible*. However, a plausible patch is deemed as *overfitting* if it fails to conform to the developer's intent. Otherwise, it is deemed as *correct*. To further understand the patch quality of AutoCodeRover and baselines, we manually verify the correctness of task-resolving (i.e. plausible) patches in SWE-bench lite. Since three repetitions are performed, we consider a task to have a *correct* patch if any of the three repetitions produced a *correct* patch. A plausible patch is correct if it is semantically equivalent to the developer patch. In this verification process, at least two authors of the paper cross-validated each patch, and any disagreement was resolved with another author. Overall,

Table 4: Result of ACR-val, ACR-val-sbfl on SWE-bench lite, *in one run only.*

Tools	Resolved Tasks	Avg Time	Avg Tokens
ACR-avg	16.11% (48.3)	173	37602 ($0.435)
ACR-val	17.00% (51)	234	46629 ($0.538)
ACR-val-sbfl	**20.33%** (61)	250	40545 ($0.470)

on SWE-bench lite, ACR has a correctness rate of 65.7% (44 correct/67 plausible). SWE-AGENT-REP has a slightly higher correctness rate of 72.7% (32/44), but the absolute number of correctly resolved tasks is smaller than ACR. Finally, the correctness rate of DEVIN on SWE-bench Devin subset is 53.2% (42/79). We observed that vast majority of ACR's overfitting patches (all but 2 of the overfitting patches) modify the same methods as the developer patches, but the code modifications are wrong. This means that even the overfitting patches from ACR are useful to the developer, since it helps in localization. The main causes of wrong modifications are the limits of the LLM's capability or insufficient context. Apart from these, we noticed two other interesting causes of overfitting. One cause is that the issue creator gives a preliminary patch in the description. This patch can be different from the final developer patch, misleading the LLM. The other interesting cause is that the issue creator mentioned a case that needs to be handled. The LLM only fixes this mentioned case, but the developer fixed other similar cases as well. The two causes indicate that the issue description, just like test suites, can be an incomplete specification.

6.2 RQ2: Effect of SBFL

In this research question, we want to understand whether program analysis techniques such as Spectrum-based fault localization (SBFL) can benefit the workflow of AUTOCODEROVER. Different from RQ1, here we construct a common scenario in program repair that AUTOCODEROVER has access to the complete test-suite of the target task instance. We use the developer-written test cases for each task instance (provided in SWE-bench lite) as the test-suite. To understand the effect of SBFL, we perform two sets of experiments: (1) AUTOCODEROVER and use the test-suite for patch validation during the patch generation retry-loop; (2) additionally provide SBFL results (top-5 suspicious methods) to AUTOCODEROVER at the beginning of the context retrieval stage compared to (1). We denote these two settings as ACR-val and ACR-val-sbfl. The patch validation is added as follows: when a patch is generated by the LLM agent, the test-suite is executed on the patched program; if the patch fails to pass the complete test-suite, AUTOCODEROVER re-invokes the patch generation agent to write a new patch. This validation loop is configured to run at most three times.

Results. Table 4 shows that, with the additional information provided by SBFL, the number of resolved tasks increased from 51 to 61 (i.e. 17.00% to 20.33% resolved rate on SWE-bench lite). Compared to the validation-only setting, additionally adding the SBFL component helps to resolve 20 unique task instances. Moreover, when comparing with task instances resolved in all other runs combined (i.e. ACR-all and ACR-val) on SWE-bench lite, ACR-val-sbfl still uniquely resolves **7** task instances that are not resolved in any of the other runs.

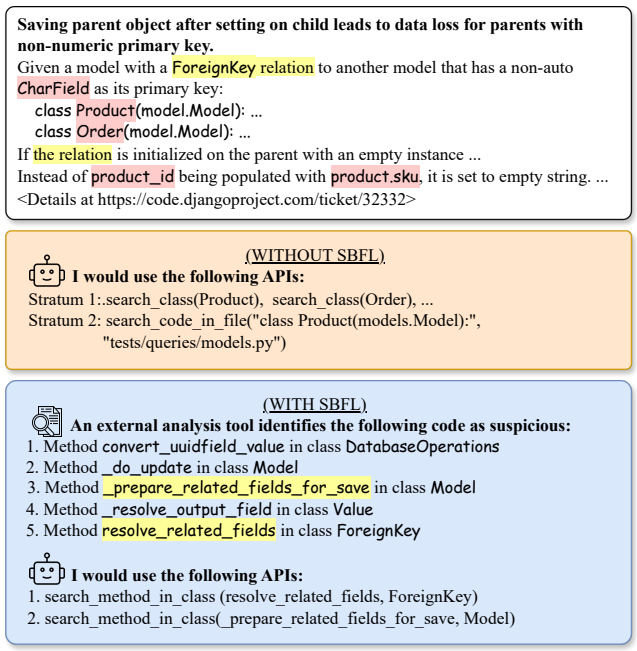


Figure 8: Issue description and AUTOCODEROVER's context retrieval (w. and w.o. SBFL), on django-13964.

Case study. We present a case study on one of the tasks uniquely resolved by ACR-val-sbfl. Figure 8 shows the issue description of django-13964[5] and how AUTOCODEROVER attempts to retrieve code context with and without SBFL. This issue reported a bug when saving django models to a database. A simplified version of the issue description is shown in the first part of Figure 8, in which important parts are highlighted. In this issue, some "hints" (highlighted in red) mentioned are actually distracting factors for a context retrieval agent. For example, the `Product` and `Order` classes describe how the bug can be reproduced, and are not classes that cause the bug. Retrieving code context from these two classes (as shown in Figure 8 part 2 "WITHOUT SBFL") does not yield useful results for forming the context and resolving the issue.

On the other hand, the SBFL component can provide extra hints for context retrieval agent (as shown in Figure 8 part 3 "WITH SBFL"). This is because SBFL considers test execution differences, which in this case revealed a few more methods in the codebase that are related to the issue. With these newly revealed hints, the agent decides to invoke APIs to search for the `resolve_related_fields` and `_prepare_related_fields_for_save` methods (the latter method is actually where the developer chose to fix this bug[6]). Moreover, we observe that the agent does not solely rely on the SBFL results to make API invocations. Instead of searching for methods ranked as top-1 in the SBFL results, the agent searched for the 3rd and 5th ranked methods. These methods are more related

[5] https://code.djangoproject.com/ticket/32332
[6] https://github.com/django/django/pull/13964/files

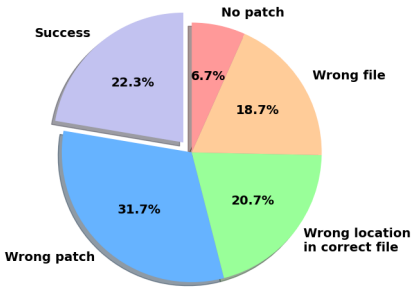


Figure 9: Taxonomy of Challenges in SWE-bench lite.

to some other hints mentioned in the issue description (highlighted in yellow), and the LLM agent is able to exploit this correlation between the natural language descriptions and the SBFL analysis results. With the correct context collected, AutoCodeRover is then able to draft a patch that resolves this issue. This suggests that an execution-based analysis can complement the agent workflow by revealing information not included in the issue description.

6.3 RQ3: Challenges on real-life tasks

In this research question, we analyze the task instances in SWE-bench lite that AutoCodeRover failed to resolve, and provide a taxonomy of the issue characteristics to highlight the practical challenges in achieving fully automated software improvement. Our taxonomy consists of challenges in the fault localization stage and patch generation stage. Specifically, for each task, we analyze the best run in the three repetitions, and classify each of the 300 tasks into one of the following:

- Success: The generated patch resolves the issue.
- Wrong patch: The generated patch modifies all methods that are modified in the developer patch. This means the patch content is wrong but the patch location(s) are correct.
- Wrong location in correct file: The generated patch that modifies the correct file but wrong location(s) in the file.
- Wrong file: The generated patch modifies the wrong file.
- No patch: No patch is generated from the retrieved context.

Figure 9 shows the distribution of the 300 tasks in SWE-bench lite. AutoCodeRover resolves 22.3% of the issues ("Success"), as mentioned in Section 6.1. The fail-to-resolve cases are included in the remaining four categories. In 31.7% of the tasks, AutoCodeRover correctly decided on all patch locations (at the method-level), but did not produce a correct patch ("Wrong patch"). More fine-grained intra-procedural analysis and specification inference techniques can play a significant role in improving these cases, by providing the patch generation agent with more method-level repair guidance. In the other three categories, the fault localization could not pinpoint all the locations to be modified. In 20.7% of the tasks, a patch is generated in the correct file, but at wrong methods / classes in the file ("Wrong location in correct file"). In some of these runs, the developer patch modifies multiple methods, but the generated patch did not modify all of them. In the other categories, a patch could not be generated at the correct file - in 18.7% of the tasks a patch is generated in wrong files, and in 6.7% of the tasks there is no applicable patch ("Wrong file" and "No patch"). We manually inspected some tasks in these two categories, and observed that their issue description mentions few methods / classes / files in the codebase. Instead, some of them contain short examples to reproduce the issue. For these tasks, one possibility is to generate a comprehensive test-suite based on the issue description, and then use execution information of the test-suite (e.g. SBFL) to reveal suspicious program locations. On the other hand, some other tasks do not contain reproducible examples and only consist of natural language descriptions. For these tasks some human involvement might be helpful. The developers could focus on these tasks.

7 DISCUSSION ON IMPROVEMENTS

In this section, we discuss a few possible directions for future improvements on AutoCodeRover.

Issue Reproducer. In Section 4, we described two scenarios that AutoCodeRover can work on, (1) issue description only, (2) aided by SBFL when the test suite is available. Although the second scenario is not always practical in real-life software development, the issue description sometimes contains a concrete reproduction script from the user that reports the bug. In the future, it is possible to design an LLM agent specifically to generate a bug reproduction test based on the GitHub issue description. The bug reproduction test can then be used to validate the correctness of AutoCodeRover's generated patch and possibly enable a regeneration process if a patch fails to pass the reproduction test. However, we also note that it is not an easy task, because large projects often use different automated testing frameworks, a test that cannot be properly executed may even confuse the main LLM agent on the GitHub issue and lead to a negative result.

Semantic Artifacts. During context retrieval, AutoCodeRover navigates through the codebase by visiting code entities such as classes and methods, which has shown to be effective. The idea of utilizing code-specific structure for context retrieval can be taken further by considering artifacts from program semantics. For example, from an initial set of methods identified in the issue description, a static call graph analysis [28, 29] can be used to collect additional relevant methods when testcases are absent. There are also potential in integrating a language server [10] for codebase navigation, so that the context retrieval agent can perform more code-specific actions such as "jump-to-definition" from a method invocation. Finally, one can use forward data dependence analysis [14] from the methods in the issue description to search for other relevant methods.

Human Involvement. Currently, AutoCodeRover left all decision-making processes to its foundation LLM — e.g. deciding context retrieving locations and whether to terminate the retrieval stage. However, leaving it to the LLM may not always be enough. Note that it is common to have multiple rounds of discussion even between the project maintainers before a pull request is created. Hence, a flexible interface and human involvement criteria between human developers and LLM agents are desired. When an LLM agent gets stuck in any stage of the workflow, the human developers can be asked to provide additional insights on the issue which may steer

the LLM agent to the correct reasoning direction and eventually resolve more challenging tasks.

8 THREATS TO VALIDITY

Despite the high efficacy AutoCodeRover achieved in SWE-bench, we report a few potential threats to our approach and experiments and discuss how we addressed them. First, LLM is known for its randomness to generate different results in different runs, which may threaten the validity of AutoCodeRover's performance. We address this by repeating our main evaluation experiment three times and releasing a replication package for practitioners. Second, the operating system environment may affect the evaluation result for the project subjects in SWE-bench (e.g., matplotlib). We share the evaluation environment in Section 5 and provide a docker environment for usage by the community. Third, we consider a task instance in SWE-bench as resolved if the generated patch passes all the test cases added in the pull request for the issue. However, this patch may be overfitting. We addressed this threat by manually verifying whether the patches are semantically equivalent to developers' patches among the authors. Finally, our result of Swe-agent-rep is lower than reported by SWE-bench team. This might be due to the different OS environments, different LLM, and parameter setups. To address this, we replicated Swe-agent-rep with the default setting in the provided scripts, and we used the same LLM (gpt-4-0125-preview) on Swe-agent-rep as AutoCodeRover. We also ran experiments in the docker provided by SWE-bench team (Table 3); this docker was publicly released on 6 April 2024.

9 PERSPECTIVES

AI-based software engineering is currently a topic of study among researchers, innovators, and entrepreneurs. Part of the trigger for this interest has been provided by efforts like GitHub Copilot [1] in the last few years. These efforts show significant promise in the use of Large Language Models (LLMs) to automatically generate code. At the same time, code generated by LLMs may be incorrect [7] or vulnerable [27]. Thus, we need autonomous processes which allow for code improvements, such as bug fixes and feature additions.

In this paper, we suggest a LLM-based solution AutoCodeRover for autonomous code improvement. The key distinguishing feature of AutoCodeRover is its conscious attempt to inject a software engineering outlook by integrating (a) use of program representations such as ASTs instead of files, (b) iterative code search by exploiting program structure and (c) use of test-based fault localization when tests can be constructed. AutoCodeRover shows significant efficacy in terms of solving real-life Github issues. AutoCodeRover shows that relying on the GitHub issue description to guide code search for patches / modifications can be misleading. It shows the need for use of program representation and structure to enable autonomous code modifications.

Today, the code LLMs cannot produce safe and secure code which can be trusted enough to be integrated into real software projects. There is thus a need to autonomously improve code (both automatically generated and manually written), for which LLMs can play a role. Future work needs to focus further on the appropriate points when tools like AutoCodeRover may converse with the human programmer. Developers may need to shift to playing different roles at the same time in *future software industry* - vetting different conversations with LLM-based tools like AutoCodeRover to enable a variety of software engineering activities. This would contrast with *today's software industry* where a person has specific roles like programmer/tester/architect/requirements-engineer. Thus, apart from full-stack engineers, we may see more *full-lifecycle software engineers* in future, who are comfortable to work with the entire lifecycle of (the components of) a software system. Furthermore, the focus of LLM oriented workflows is on *scale* today. This focus may shift to engendering of *trust* if LLM oriented autonomous software engineering becomes commonplace. Future software engineers may focus more on greater trust, instead of larger scale.

DATA AVAILABILITY

We share full public access to (1) AutoCodeRover's implementation, (2) all patches generated during our experiment and conversation history with LLM, (3) replication scripts for our experiments, (4) replication scripts and logs for Swe-agent-rep at the following website: https://github.com/nus-apr/auto-code-rover

ACKNOWLEDGMENTS

This work was partially supported by a Singapore Ministry of Education (MoE) Tier 3 grant "Automated Program Repair", MOE-MOET32021-0001.

REFERENCES

[1] 2022. GitHub Copilot, your AI pair programmer. https://github.com/features/copilot/

[2] Rui Abreu, Peter Zoeteweij, and Arjan J.C. 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, IEEE, 89–98. https://doi.org/10.1109/taic.part.2007.13

[3] Johannes Bader, Andrew Scott, Michael Pradel, and Satish Chandra. 2019. Getafix: Learning to fix bugs automatically. *Proceedings of the ACM on Programming Languages* 3, OOPSLA (2019), 1–27.

[4] Marcel Böhme, Cristian Cadar, and Abhik Roychoudhury. 2021. Fuzzing: Challenges and Reflections. *IEEE Software* 38, 3 (2021).

[5] Cristian Cadar and Koushik Sen. 2013. Symbolic execution for software testing: three decades later. *Commun. ACM* 56, 2 (2013).

[6] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating Large Language Models Trained on Code. *arXiv.org* abs/2107.03374 (7 2021). arXiv:2107.03374 https://arxiv.org/abs/2107.03374

[7] Zhiyu Fan, Xiang Gao, Martin Mirchev, Abhik Roychoudhury, and Shin Hwei Tan. 2023. Automated Repair of Programs from Large Language Models.. In *45th IEEE/ACM International Conference on Software Engineering, ICSE 2023, Melbourne, Australia, May 14-20, 2023*. IEEE, IEEE, 1469–1481. https://doi.org/10.1109/icse48619.2023.00128

[8] Xiang Gao, Sergey Mechtaev, and Abhik Roychoudhury. 2019. Crash-avoiding program repair. In *Proceedings of the 28th ACM SIGSOFT International Symposium on Software Testing and Analysis*. 8–18.

[9] Claire Le Goues, Michael Pradel, and Abhik Roychoudhury. 2019. Automated program repair. *Commun. ACM* 62 (11 2019), 56–65. Issue 12.

[10] Nadeeshaan Gunasinghe and Nipuna Marcus. 2021. *Language Server Protocol and Implementation*. Springer.

[11] Nan Jiang, Kevin Liu, Thibaud Lutellier, and Lin Tan. 2023. Impact of Code Language Models on Automated Program Repair., In 45th IEEE/ACM International

Conference on Software Engineering, ICSE 2023, Melbourne, Australia, May 14-20, 2023 (Melbourne, Victoria, Australia). *International Conference on Software Engineering*, 1430–1442. https://doi.org/10.1109/icse48619.2023.00125

[12] Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2024. Leaderboard results on SWE-bench. Retrieved April 8, 2024 from https://www.swebench.com/

[13] Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. 2024. SWE-bench: Can Language Models Resolve Real-world Github Issues?. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=VTF8yNQM66

[14] Wuxia Jin et al. 2024. PyAnalyzer: An Effective and Practical Approach for Dependency Extraction from Python Code. In *International Conference on Software Engineering (ICSE)*.

[15] James A Jones, Mary Jean Harrold, and John Stasko. 2002. Visualization of test information to assist fault localization. In *Proceedings of the 24th international conference on Software engineering*. 467–477.

[16] Fabian Keller, Lars Grunske, Simon Heiden, Antonio Filieri, Andre van Hoorn, and David Lo. 2017. A critical evaluation of spectrum-based fault localization techniques on a large-scale software system. In *2017 IEEE International Conference on Software Quality, Reliability and Security (QRS)*. IEEE, 114–125.

[17] Anil Koyuncu, Kui Liu, Tegawendé F Bissyandé, Dongsun Kim, Martin Monperrus, Jacques Klein, and Yves Le Traon. 2019. iFixR: Bug report driven program repair. In *Proceedings of the 2019 27th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering*. 314–325.

[18] Cognition Labs. 2024. Devin, AI software engineer. Retrieved April 12, 2024 from https://www.cognition-labs.com/introducing-devin

[19] Yujia Li, David H. Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. 2022. Competition-Level Code Generation with AlphaCode. *Science* abs/2203.07814, 6624 (12 2022), 1092–1097. https://doi.org/10.48550/arxiv.2203.07814

[20] Thibaud Lutellier, Hung Viet Pham, Lawrence Pang, Yitong Li, Moshi Wei, and Lin Tan. 2020. CoCoNuT: combining context-aware neural translation models using ensemble for program repair., In ISSTA '20: 29th ACM SIGSOFT International Symposium on Software Testing and Analysis, Virtual Event, USA, July 18-22, 2020, Sarfraz Khurshid and Corina S. Pasareanu (Eds.). *International Symposium on Software Testing and Analysis*, 101–114.

[21] Alexandru Marginean, Johannes Bader, Satish Chandra, Mark Harman, Yue Jia, Ke Mao, Alexander Mols, and Andrew Scott. 2019. Sapfix: Automated end-to-end repair at scale. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE, 269–278.

[22] Sergey Mechtaev, Jooyong Yi, and Abhik Roychoudhury. 2016. Angelix: scalable multiline program patch synthesis via symbolic analysis., In Proceedings of the 38th International Conference on Software Engineering, ICSE 2016, Austin, TX, USA, May 14-22, 2016, Laura K. Dillon, Willem Visser, and Laurie Williams (Eds.). *International Conference on Software Engineering*, 691–701.

[23] Manish Motwani and Yuriy Brun. 2023. Better automatic program repair by using bug reports and tests together. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 1225–1237.

[24] Hoang Duong Thien Nguyen, Dawei Qi, Abhik Roychoudhury, and Satish Chandra. 2013. SemFix: program repair via semantic analysis., In 35th International Conference on Software Engineering, ICSE '13, San Francisco, CA, USA, May 18-26, 2013 (San Francisco, CA, USA), David Notkin, Betty H. C. Cheng, and Klaus Pohl (Eds.). *International Conference on Software Engineering*, 772–781. https://doi.org/10.1109/icse.2013.6606623

[25] Yannic Noller, Ridwan Shariffdeen, Xiang Gao, and Abhik Roychoudhury. 2022. Trust Enhancement Issues in Program Repair. In *IEEE/ACM 44th International Conference on Software Engineering (ICSE)*.

[26] Brayan Stiven Torrres Ovalle. 2023. GitHub Copilot. https://doi.org/10.26507/paper.2300

[27] H Pearce, B Tan, B Ahmad, R Karri, and B Dolan-Gavitt. 2023. Examining zero-shot vulnerability repair with large language models. In *IEEE Symposium on Security and Privacy (SP)*.

[28] Barbara G Ryder. 1979. Constructing the call graph of a program. *IEEE Transactions on Software Engineering* 3 (1979), 216–226.

[29] Vitalis Salis, Thodoris Sotiropoulos, Panos Louridas, Diomidis Spinellis, and Dimitris Mitropoulos. 2021. Pycg: Practical call graph generation in python. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. IEEE, 1646–1657.

[30] Edward K. Smith, Earl T. Barr, Claire Le Goues, and Yuriy Brun. 2015. Is the cure worse than the disease? overfitting in automated program repair., In Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering, ESEC/FSE 2015, Bergamo, Italy, August 30 - September 4, 2015, Elisabetta Di Nitto, Mark Harman, and Patrick Heymans (Eds.). *ESEC/SIGSOFT FSE*, 532–543. http://people.cs.umass.edu/%7Ebrun/pubs/pubs/Smith15fse.pdf

[31] Shin Hwei Tan, Ziqiang Li, and Lu Yan. 2024. CrossFix: Resolution of GitHub issues via similar bugs recommendation. *Journal of Software: Evolution and Process* 36, 4 (2024), e2554.

[32] The Cognition Team. 2024. SWE-bench technical report (Devin). Retrieved April 12, 2024 from https://www.cognition-labs.com/post/swe-bench-technical-report

[33] Westley Weimer, ThanhVu Nguyen, Claire Le Goues, and Stephanie Forrest. 2009. Automatically finding patches using genetic programming., In 31st International Conference on Software Engineering, ICSE 2009, May 16-24, 2009, Vancouver, Canada, Proceedings. *2009 IEEE 31st International Conference on Software Engineering*, 364–374. https://doi.org/10.1109/icse.2009.5070536

[34] David Williams, James Callan, Serkan Kirbas, Sergey Mechtaev, Justyna Petke, Thomas Prideaux-Ghee, and Federica Sarro. 2023. User-Centric Deployment of Automated Program Repair at Bloomberg. *arXiv preprint arXiv:2311.10516* (2023).

[35] WE Wong, R Gao, Y Li, R Abreu, and F Wotawa. 2016. A survey on software fault localization. *IEEE Transactions on Software Engineering* (2016), 707–740. Issue 8.

[36] W Eric Wong, Ruizhi Gao, Yihao Li, Rui Abreu, and Franz Wotawa. 2016. A survey on software fault localization. *IEEE Transactions on Software Engineering* 42, 8 (2016), 707–740.

[37] Chunqiu Steven Xia, Yuxiang Wei, and Lingming Zhang. 2023. Automated program repair in the era of large pre-trained language models. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 1482–1494.

[38] John Yang, Carlos E. Jimenez, Alexander Wettig, Shunyu Yao, Karthik Narasimhan, and Ofir Press. 2024. SWE-agent: Agent Computer Interfaces Enable Software Engineering Language Models.

[39] Qihao Zhu, Zeyu Sun, Yuan-an Xiao, Wenjie Zhang, Kang Yuan, Yingfei Xiong, and Lu Zhang. 2021. A syntax-guided edit decoder for neural program repair., In ESEC/FSE '21: 29th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Athens, Greece, August 23-28, 2021, Diomidis Spinellis, Georgios Gousios, Marsha Chechik, and Massimiliano Di Penta (Eds.). *ESEC/SIGSOFT FSE*, 341–353. https://arxiv.org/pdf/2106.08253