

Noname manuscript No.  
(will be inserted by the editor)

## Bugs in Large Language Models Generated Code: An Empirical Study

Florian Tambon\* · Arghavan Moradi  
Dakhel\* · Amin Nikanjam · Foutse  
Khomh · Michel C. Desmarais ·  
Giuliano Antoniol

Received: date / Accepted: date

**Abstract** Large Language Models (LLMs) for code have gained significant attention recently. They can generate code in different programming languages based on provided prompts, fulfilling a long-lasting dream in Software Engineering (SE), i.e., automatic code generation. Similar to human-written code, LLM-generated code is prone to bugs, and these bugs have not yet been thoroughly examined by the community. Given the increasing adoption of LLM-based code generation tools (e.g., GitHub Copilot) in SE activities, it is critical to understand the characteristics of bugs contained in code generated by LLMs. This paper examines a sample of 333 bugs collected from code generated using three leading LLMs (i.e., CodeGen, PanGu-Coder, and Codex) and identifies the following 10 distinctive bug patterns: *Misinterpretations*, *Syntax Error*, *Silly Mistake*, *Prompt-biased code*, *Missing Corner Case*, *Wrong Input Type*, *Hallucinated Object*, *Wrong Attribute*, *Incomplete Generation*, and *Non-Prompted Consideration*. The bug patterns are presented in the form of a taxonomy. The identified bug patterns are validated using an online survey with 34 LLM practitioners and researchers. The surveyed participants generally asserted the significance and prevalence of the bug patterns. Researchers and practitioners can leverage these findings to develop effective quality as-

---

This work was supported by: Fonds de Recherche du Québec (FRQ), the Canadian Institute for Advanced Research (CIFAR) as well as the DEEL project CRDPJ 537462-18 funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Consortium for Research and Innovation in Aerospace in Québec (CRIAQ), together with its industrial partners Thales Canada inc, Bell Textron Canada Limited, CAE inc and Bombardier inc.

---

Florian Tambon · Arghavan Moradi Dakhel · Amin Nikanjam · Foutse Khomh · Michel C. Desmarais · Giuliano Antoniol  
Polytechnique Montréal, Montréal, Canada  
E-mail: {florian-2.tambon, arghavan.moradi-dakhel, amin.nikanjam, foute.khomh, michel.desmarais, giuliano.antoniol}@polymtl.ca  
“\*” authors contributed equally

surance techniques for LLM-generated code. This study sheds light on the distinctive characteristics of LLM-generated code.

**Keywords** Large Language Models · Bugs · Software Testing · Empirical Study

## 1 Introduction

The recent surge in the development of Large Language Models (LLMs) tailored for code generation has garnered significant attention [1, 2, 3]. Transformer-based models such as Codex [4] and Llama-2 [5], trained on large amounts of open-source code repositories, have demonstrated success in producing code across various programming languages, such as Python, Java, and C [6, 7, 8]. These models approach code generation as a transformation process, converting natural language descriptions (prompts) into executable programming language statements.

Positioned as potential AI pair programmers [2, 9, 10] in software projects, LLM-based code generators are poised to play a crucial role in the quality of the overall software project. However, similar to human-written code, LLM-generated code is prone to errors [9]. Nonetheless, Asare et al. [11] who recently examined vulnerabilities contained in LLM-generated code (using Copilot) compared to those in human-written code, reported that LLM models do not generate the same vulnerabilities as humans. In parallel, several studies [12, 13] reported that LLMs and human developers may not focus on the same part of the prompt to solve a coding task. This raises an important question: Do LLMs generate faults similar to those produced by human developers? This question is crucial because the effectiveness of popular quality assurance techniques like mutation testing depends on a precise characterization of faults occurring in the code under test [14].

To illustrate the potential differences between bug patterns in LLM-generated code and those in human-written code, let's consider the example from Listing 1 which presents the code generated by Codex [4], a well-known LLM to generate code, for the task of “returning the flags present in the Python ArgumentParser object” (shown in green as a docstring in the figure).

While the code seems reasonably correct at first glance, the attentive reader will notice in the last line (Line 14) that the LLM suddenly decided to sort the flags it extracted using the ‘sorted’ function. This step, not requested in the prompt, leads to the code being faulty for the requested use case. This kind of snippet is interesting as it highlights two key points: 1) LLMs might generate a seemingly correct code at first glance, which could deceive non-experienced users particularly if no (or not enough) testing is performed, and 2) LLMs may add a non-prompted feature to the code, leading to an error, something that a human developer would rarely do, which shows that LLMs bugs might not be exactly similar to human-made bugs.

While there exist studies that investigated bugs in LLM-generated code [15, 16, 17], to the best of our knowledge, none of them examined the bug

Listing 1: An example of a code generated by Codex [4]. The LLM is given the function signature and docstring (shown in green) as a prompt and should generate the code to address the task in the prompt. We labeled this buggy code as “Non-Prompted Consideration (NPC)” in our study.

---

```

1 def parser_flags(parser):
2     """
3         Given an argparse.ArgumentParser instance, return its argument flags in a space-separated
4         string.
5     """
6     flags = []
7     for action in parser._actions:
8         if isinstance(action, argparse._SubParsersAction):
9             for choice, subparser in action.choices.items():
10                 flags.append(choice)
11                 flags += parser_flags(subparser)
12             else:
13                 flags.append(action.option_strings[0])
14     return " ".join(sorted(flags))

```

---

patterns in code generated by LLMs, when applied to real programming tasks (i.e., involving the implementation of functionalities).

This paper fills this gap in the literature by examining the characteristics of bugs contained in code generated by three leading LLMs: CodeGen, PanGu-Coder, and Codex. The identified characteristics are then organized in a taxonomy and validated via a survey of practitioners and researchers. Our investigation is conducted using the following Research Questions (RQs):

**RQ1** What are the characteristics of bugs occurring in code generated by LLMs for real-world project tasks?

**RQ2** To what extent are the identified bug patterns in LLM-generated code relevant for software practitioners and researchers working with LLMs?

We conducted our study using CoderEval [6], which is a dataset of 230 Python functions selected from 10 real-world open-source projects hosted on GitHub. Each of the selected LLMs was used to generate code for the different tasks described in CoderEval. We identified the bugs contained in the obtained code and manually analyzed the characteristics of a random sample of 333 bugs. This led us to organize our findings in a taxonomy composed of 10 categories of bugs: *Misinterpretations*, *Syntax Error*, *Silly Mistake*, *Prompt-biased code*, *Missing Corner Case*, *Wrong Input Type*, *Hallucinated Object*, *Wrong Attribute*, *Incomplete Generation*, and *Non-Prompted Consideration*. To assess the validity of the obtained bug patterns, we performed an online survey with 34 software practitioners and researchers working with LLMs. The obtained results show that respondents encountered the different bug patterns we observed, in a proportion similar to that obtained from our sample set.

This paper makes the following contributions:

- We study bug patterns occurring in LLM-generated code and their prevalence.

- We propose a taxonomy from the characteristics of the bug patterns in LLM-generated code.
- We validate our findings using a survey and provide insights for researchers and practitioners.
- We make the dataset used in this study publicly available online [18] for other researchers and practitioners to replicate or build upon our work.

**The rest of the paper is organized as follows.** We present the background of the study in Section 2. Next, we detail the methodology of the study in Section 3 for both the empirical study of bugs and the validation survey. We then report our empirical investigation of bugs in LLM-generated code in Section 4.1. The validation survey results are presented in Section 4.2. We provide a discussion on our findings in Section 5. Then, we discuss the related works and threats to the validity of this study in Section 6 and Section 7, respectively. Finally, we present the conclusion and future work in Section 8.

## 2 Background

In this section, we introduce the three LLMs used in our study and describe the open coding methodology followed in this paper.

### 2.1 LLMs for code generation

Different LLMs have been introduced for automatic code generation in Software Engineering (SE). One highly potent LLM is OpenAI’s Codex [4], which has been extensively employed in diverse code-related SE tasks [19, 20, 21]. This closed-source GPT-based auto-regressive LLM [22] boasts up from 12 million to 12 billion parameters and undergoes fine-tuning on 54 million public repositories from GitHub. Notably, Codex is the model behind GitHub Copilot<sup>1</sup>, an in-IDE developer coding assistant, proficient at generating code based on user-provided context with the maximum input length of 4,096 tokens. A recent addition to GitHub Copilot is “Copilot Chat”<sup>2</sup>, which is tuned with human feedback for dialog use cases. Copilot Chat proves versatile for a broad range of code-related tasks. It excels at generating code fragments, describing code snippets in natural language, generating unit tests, and fixing buggy code, all tailored to the specific context of the task at hand.

PanGu-Coder [23] is an open-source pre-trained LLM for generating code from text. This model is based on PanGu- $\alpha$  architecture [24], a unidirectional decoder-only transformer with an additional layer for querying added on top of it. PanGu-Coder was trained in two stages using Python programs extracted from GitHub. In the first stage (unsupervised), the model is pre-trained on raw programming language data containing docstrings/inline comments written in

<sup>1</sup> <https://copilot.github.com/>

<sup>2</sup> <https://docs.github.com/en/copilot/github-copilot-chat>

natural language. Next, this model is fine-tuned to generate code from text in a second stage (supervised). Two versions of this model were published and leveraged for code generation recently [25]. These versions have respectively 317 million and 2.6 billion parameters, with a maximum input length of 1,024 tokens.

CodeGen [26] is a family of LLMs trained for text-to-code generation. Based on a decoder-only architecture, this family of models is designed for multi-turn program synthesis, in which a user engages with the LLM by gradually feeding specifications in natural language to obtain a corresponding program. This collaborative interaction allows the user, in conjunction with the LLM, to iteratively build and refine a program in multiple steps. CodeGen models were trained on a natural language corpus and code extracted from GitHub (i.e., a multilingual code dataset, and a dataset of Python programs). Various CodeGen models come with 350 million to 16.1 billion parameters and a maximum input length of 2,048 tokens [26]. CodeGen open-source models have been employed for code generation tasks in many recent studies [25, 27].

## 2.2 Open Coding

Open coding, also known as postformed code, is a qualitative research method used to analyze textual data (e.g., survey responses, and interview transcripts). In open coding, researchers start with a relatively unstructured approach, inspecting the data line by line and generating codes or labels for themes and concepts identified in the data. The generated codes are subject to dynamic adjustments, i.e., they can be added, removed, or merged to develop a final codebook. This flexibility allows for a more adaptive and nuanced analysis compared to the rigid structure of preformed codes [28]. Open coding has proven to be more useful than the preformed code approach in SE studies [29, 30].

In open coding, two human reviewers independently start the process of coding on the same set of samples. During this phase, the reviewers explore patterns, similarities, and categories within the samples relevant to the quantitative variables [28], i.e., characteristics or root causes of bugs in our study. They assign discrete categories to each sample. The two reviewers then engage in a discussion session to exchange and deliberate on their identified categories, forming the initial codebook.

In each discussion, a third reviewer, not involved in the coding process, intervenes in cases of diverging opinions between the two reviewers, to finalize the labels. The two reviewers continue to independently label the remaining samples and address similarly any conflicts that arise during subsequent discussion sessions [30]. Through this iterative process, the two reviewers independently label all the samples. At the end of the labeling process, they report the agreement reached through these independent labeling processes before entering into any negotiations [30]. Different studies in SE adapted

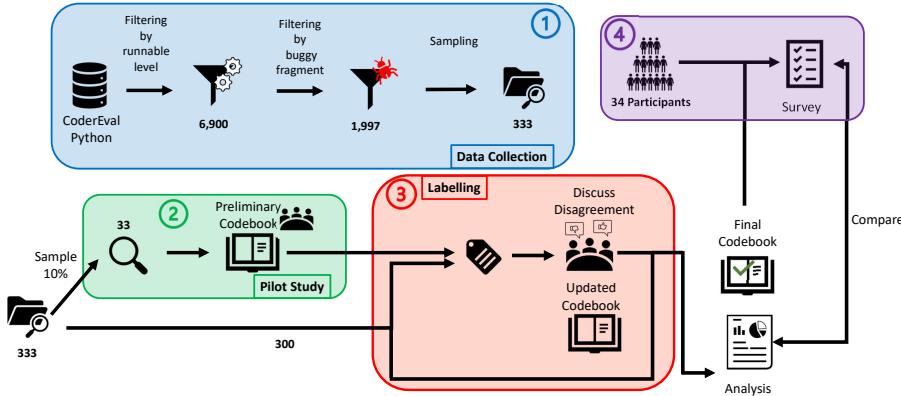


Fig. 1: High-level view of the followed methodology.

this technique to conduct qualitative research for different purposes including constructing a taxonomy [29, 30].

### 3 Study Design

In this section, we outline our methodology designed for investigating bugs in code generated by LLMs. We first describe our method for studying bug patterns in code generated by LLMs. Next, we present the methodology followed to validate the results of our empirical analysis. Specifically, we address the following research questions:

- RQ1:** What are the characteristics of bugs occurring in code generated by LLMs for real-world project tasks?
- RQ2:** To what extent are the identified bug patterns in LLM-generated code relevant for software practitioners and researchers working with LLMs?

Figure 1 depicts the study methodology. We started by collecting data from a benchmark dataset (CoderEval) and selected a sample set. Next, we established a preliminary codebook by labeling 10% of our sample set (pilot study). Subsequently, we iteratively completed the labeling process and our codebook. We finally organized our findings to obtain a taxonomy and surveyed practitioners to validate the bug patterns in our taxonomy. In the remainder of this section, we delve into further details of our methodology to answer each RQ.

### 3.1 RQ1: What are the characteristics of bugs occurring in code generated by LLMs for real-world project tasks?

#### 3.1.1 Data collection

We used a benchmark dataset named CoderEval [6] to conduct our study. This dataset consists of 230 Python functions from 43 open-source projects and 230 Java functions from 10 open-source projects selected from GitHub. In this study, we focused on Python functions as it is a very common and widely used programming language across different domains<sup>3</sup>. Each function in CoderEval is constituted of the function signature, a docstring describing the function, and the oracle code (i.e., correct reference solution to the function). Additional information is available in this benchmark such as the context of the function (like other functions being used, and files/projects in which the function is contained). CoderEval also contains a human-written description of each function which was obtained by showing the oracle code to a group of developers without the docstring. Finally, for each of the functions, the CoderEval dataset contains 10 samples of code generated by each of the three LLMs under test for both types of prompt (docstring and human-written description). In our study, we focused on the code generated by LLMs using the docstring, which is a similar setting to other code generation benchmarks such as HumanEval [31] or MBPP [32].

At the time this study was started, CoderEval included code generated using three different LLMs<sup>4</sup>: CodeGen [26], PanGu-Coder [23], and Codex [4]. Two of these LLMs, CodeGen and PanGu-Coder, are open-source and have been used in previous studies that harness the power of LLMs for code generation tasks [25, 27]. The third one, Codex [4], is the model behind Copilot<sup>5</sup>. The code snippets generated by the LLMs were used to replace the oracle function sequentially. Upon replacement, the entry point code was executed, and the running output was captured, including return values, exceptions, errors, and console output. In case of no errors or exceptions, CoderEval compared the actual output with the expected output using unit tests. A difference between actual and expected output was considered a test failure. Conversely, an unexpected termination during execution was also treated as a test failure. The code snippets that led to a test failure were flagged as buggy in CoderEval. We used those flagged buggy samples in our empirical study.

The samples in CoderEval exhibit different “runnable levels”, indicating the extent of dependency a function requires for effective execution. Out of the 6,900 Python samples in CoderEval (2,300 per LLM), our study was restricted to code fragments falling into one of the following runnable levels:

<sup>3</sup> <https://spectrum.ieee.org/top-programming-language-2020>

<sup>4</sup> Note that at the start of this study, CoderEval contained the above-mentioned LLMs, which has since been updated later. As such, the version of the dataset we used in our study is: <https://github.com/CoderEval/CoderEval/tree/ec1177750cf10b5faa414a0e76d1430e75141a44>

<sup>5</sup> <https://github.com/features/copilot>

“*self\_runnable*” (without external dependencies), “*slib\_runnable*” (dependent only on Python’s standard libraries such as “os” or “json”), or “*plib\_runnable*” (dependent only on Python’s standard and public libraries such as “numpy”). We made this decision both to reduce the number of potential samples as well as because higher “runnable levels” would be more likely to result in the generation of code that fails to address the intended task, as highlighted in CoderEval’s paper [6]. Indeed, the higher this level, the more information is needed to properly understand how the function should work. This information is less likely to be correctly encompassed in the docstring for higher “runnable levels”, to the point that even a programmer not privy to the actual function would struggle to code the functionality. As such, any bugs in the code generated from the prompt of those categories would not necessarily reflect a weakness of LLM in generating code but just a lack of proper information. Therefore, we did not consider them to mitigate potential bias in our results. Finally, as our focus was on observing bugs generated by LLMs, we considered only those generated code fragments that were flagged as buggy by CoderEval. These filterings resulted in 1,997 buggy code samples.

To streamline the manual labeling effort, we opted to sample our data. At a 95% confidence interval and an error rate of 5%, this necessitated the analysis of 323 samples. We sampled **333** code fragments and ensured that the distribution of bugs across the three LLMs was balanced (highlighted as ① in Figure 1).

### 3.1.2 Manual labeling and classification

As no previous categorization exists for bugs in LLM-generated code, we conducted a manual analysis of our buggy samples employing an open coding procedure [28], similar to methodologies applied in prior studies on bug categorization [29, 30, 33]. The aim was to inductively construct bug categories in a bottom-up manner through manual analysis of buggy samples. This procedure allows us to identify various types of bugs based on our observations from the buggy samples. We used a shared document including the link to all buggy samples making sure that all reviewers could work together during the manual labeling process. The manual construction of the taxonomy involved three human reviewers, two of whom are senior Ph.D. students with experience in SE research. The third reviewer is a senior researcher with 10 years of research experience in the field of SE and AI.

We developed the codebook of our bug patterns in multiple rounds. First, a preliminary pilot study was conducted. The first two authors independently coded 10% of the sample set to reduce the subjective bias. Each coder was tasked to analyze each buggy sample and label it with one or more descriptive labels defined on their own, based on the bug pattern(s) observed in the sample. At the end of this round, the entire team convened to discuss the labels created, resolve conflicts, and merge the individual labels to create a consolidated set of categories, which served as the initial codebook for the next phase.

Table 1: Coding process: initial agreements and conflicts throughout the rounds.

| Round | Analyzed Samples | Initial Agreements | Conflicts | New Categories      |
|-------|------------------|--------------------|-----------|---------------------|
| Pilot | 33               | 26                 | 7         | 8                   |
| 1     | 50               | 37                 | 13        | 1 (Wrong Attribute) |
| 2     | 50               | 42                 | 8         | 1 (NPC)             |
| 3     | 50               | 34                 | 16        | 0                   |
| 4     | 50               | 37                 | 13        | 0                   |
| 5     | 50               | 43                 | 7         | 0                   |
| 6     | 50               | 39                 | 11        | 0                   |
| Total | 333              | 258                | 75        | 10                  |

In case of disagreement between the two coders, the third author would act as a tie-breaker (highlighted as ② in Figure 1).

In the second phase, the remaining buggy samples were divided into six parts, each containing 15% of the remaining buggy codes, and we repeated the previous process: the two coders manually investigated all samples in each part independently, and then the entire team met to cross-check the labels and address conflicts while resolving discrepancies. If a new category emerged (i.e., a bug could not be classified under the existing categories) or a previous label was refined, the entire team gathered to discuss and incorporate the change into the codebook. We also made sure to re-label the affected samples from the previous rounds, if any. If a sample contained multiple types of bugs, it was classified into all identified bug patterns (i.e., using multiple labels). We had no constraint on the number of labels that could be assigned to a particular buggy snippet. The entire process required approximately 108 person-hours and resulted in a taxonomy comprising 10 bug patterns (highlighted as ③ in Figure 1).

Given that we did not have any prior-defined labels and since we do not have a predefined number of labels for every buggy code snippet in our dataset, it is not possible to compute inter-rater agreement levels; which is typical for open coding. However, after finalizing categories, we realized that **78.2%** of the bugs were labeled similarly during independent labeling by the two reviewers before solving disagreements. Any conflicts were then thoroughly discussed until a 100% agreement was reached [30]. We retained all the initial labels and comments on our shared document to facilitate any subsequent discussions. We provide in Table 1 each step of our process. These labels and comments are shared in our replication package available at [18].

### 3.2 RQ2: To what extent are the identified bug patterns in LLM-generated code relevant for software practitioners and researchers working with LLMs?

To assess the relevance of the bug patterns identified in **RQ1**, we surveyed software practitioners and researchers working with LLMs (④ in Figure 1).

### 3.2.1 Participants' selection

To identify participants for our survey, we collected the contacts (email addresses) of GitHub users who collaborated on repositories that contain code generated using various LLMs. We followed the methodology proposed by Yujia et al. [34] to identify GitHub repositories containing code generated using LLMs. Yujia, et al. [34] employed various search keywords to collect GitHub repositories containing code generated by GitHub Copilot. We leveraged their shared dataset and broadened our list of participants, by adopting their methodology to collect additional repositories containing code-generated using Codex, Llama, Llama-2, CodeLlama, and ChatGPT. This process allowed us to collect a total of 113 repositories. Next, we employed the PyDriller [35] library to extract the email addresses of the collaborators on each repository. In total, we obtained 200 unique email addresses of practitioners who collaborated on GitHub repositories containing LLM-generated code.

We complemented this list of practitioners with a list of researchers who have published on code generation using LLMs. To obtain this list of researchers, we searched relevant LLM-based code generation papers over Google Scholar<sup>6</sup> using two keywords: “LLMs” and “code generation”. We collected the first 75 papers returned by Google Scholar and manually inspected each of them to ensure their relevance i.e., that they leveraged LLMs for automatic code generation. This inspection process allowed us to identify 56 relevant papers that focused on code generation using LLMs. Next, we extracted the email addresses of the authors of each of these papers (who are assumed to be researchers working on generating code with LLMs). We obtained a total of 182 emails after removing duplicates. Overall, we successfully sent the survey questionnaire to 382 unique email addresses. We discuss the details about the participants and the response rate in Section 4.2.

We also posted our survey questionnaire on two relevant Reddit channels: *LocalLLaMA* and *MachineLearning*. Figure 5 in Appendix A shows our post on Reddit.

The questionnaire was dispatched along with a detailed message outlining the purpose, scope, and estimated completion time of the survey (which is around 10 minutes). We made it clear to the participants that the survey was anonymous, though participants had the option to provide their emails for additional communication and to receive a summary of the study.

### 3.2.2 Survey with Practitioners

The survey was open for two weeks, during which we sent a reminder one week after the initial email. Both the survey form and the anonymized responses that we collected are available in our replication package [18]. The survey is divided into two parts and is inspired by similar empirical studies of bug patterns [29, 33, 36]. In the first part, we asked demographic questions to participants.

---

<sup>6</sup> <https://scholar.google.com>

We then asked more specific questions about bugs in LLMs generated code regarding their frequencies and complexities associated with detecting or fixing the bugs (see Figure 6 in Appendix B.1 for more details).

In the second part, we wanted to gather the participants' feedback on several aspects of each pattern of bug identified. To do so, we defined the questions similarly for each pattern of bugs in the taxonomy. Specifically, for each pattern of bug, we present participants with a description of the bugs belonging to the pattern and provide an example code snippet; we use the same examples that will be presented in Section 4.1. Next, we presented the participants with several questions to be answered using a 5-level Likert scale. The full questionnaire is presented in Appendix B.1 in Figure 7.

The questions start by asking about the frequency at which the participants encountered the bug pattern (ranging from 1-Never to 5-Always). Next, we defined the following concepts to the participants:

- i. Diagnosing: How hard would it be for the respondents to diagnose such an error in an LLM-generated code? Could they distinguish it easily or would it require running the code or using a debugger rendering the diagnostic hard? (1-Easy to 5-Hard)
- ii. Complexity: Is the type of error rather trivial (i.e. not something a human developer with decent programming knowledge would do) or, on the contrary, does it denote some degree of complexity? (1-Trivial to 5-Complex)
- iii. Fixing: If the respondents had to fix such a mistake, would it be hard? Would it just require adjusting some part of the code (i.e. easy), or would it require extensive refactoring (i.e. hard)? (1-Easy to 5-Hard)

Participants were then asked to answer from their own perception/experience of each bug pattern.

The survey also contained optional comment fields at the end of each section and the end of the survey. Participants had the opportunity to provide suggestions such as additional information related to the bug patterns in the study or bug patterns not described in the study.

### *3.2.3 Analysis of Survey Results*

Once the results were collected for each bug pattern by the respondents, we aggregated the results (see Section 4.2). To do so, we followed a methodology used in similar studies [33] leveraging weighted average. We processed the results for the frequency/diagnosing/complexity/fixing of each bug pattern reported by the participants by multiplying the value of the Likert Scale with each percentage. That is to say, for instance, we would obtain a score of 2.79 for a bug pattern where respondents replied the following: 1 - (17.6%), 2 (26.5%), 3 (23.5%), 4 (23.5%), and 5 - (8.8%). We also complemented the explanation of the results with comments from respondents relevant to the description of the results. The goal was to capture the perceived difficulty in fixing/diagnosing and the complexity of the different bug patterns according to practitioners. Additionally, we performed an analysis comparing the frequency

of reported bug patterns among respondents with those in our sample set. We wanted to assess whether the bug patterns within our sample set (extracted from CoderEval) matched LLM practitioners' experience when dealing with bugs. The idea is to provide further evidence that our labeled categories are encountered in similar proportions by the respondents who used LLMs, not on a dataset such as CoderEval but as a tool in their development activities. Specifically, we evaluated if there was a correlation between the weighted average of the frequency of bug patterns reported by the participants and the proportion of each bug pattern found in the studied sample. To do so, we calculated the Spearman Rho between those scores and the distribution of bug patterns. The results are reported in Table 2 (Section 4.1).

## 4 Empirical Results

In this section, we delve into the bug patterns obtained from our analysis as well as the respondents' answers from our survey. We examine the nature, frequency, and distribution of these bugs across the three different LLMs. All collected artifacts and data generated during our study are accessible in our replication package [18].

### 4.1 RQ1: What are the characteristics of bugs occurring in code generated by LLMs for real-world project tasks?

#### 4.1.1 Taxonomy

We organized our observations as a taxonomy. The taxonomy is organized into 10 categories. The full taxonomy is shown in Figure 2 with the percentage of code samples assigned to each category during manual labeling.

In the following, we explain each of the 10 bug patterns observed; providing representative examples.

**1. Misinterpretation:** Misinterpretation occurs when the generated code deviates from the intention of the prompt. This happens when the model struggles to capture the specifications of the task expressed in the prompt. In Listing 2, you can find an example task, “int\_to\_string”, along with its corresponding reference solution (top). The function declaration and the docstring at the beginning of the solution (lines 1 to 6) represent the prompt and the rest (lines 7 to 15) is the reference solution. A code generated by PanGu-Coder for this prompt is presented at the bottom of Listing 2 (lines 17 and 18). This code is misaligned with the prompt since it simply treats the variable “number” as an index and returns the item at that index in the list “alphabet”. However, the prompt describes a task to convert the variable “number” to a string using the given list “alphabet”. Consequently, we label the contained bug as *Misinterpretation*. This bug pattern is among the most common in our analyzed dataset across all three models, with 20.77% of buggy code samples categorized as *Misinterpretation*.

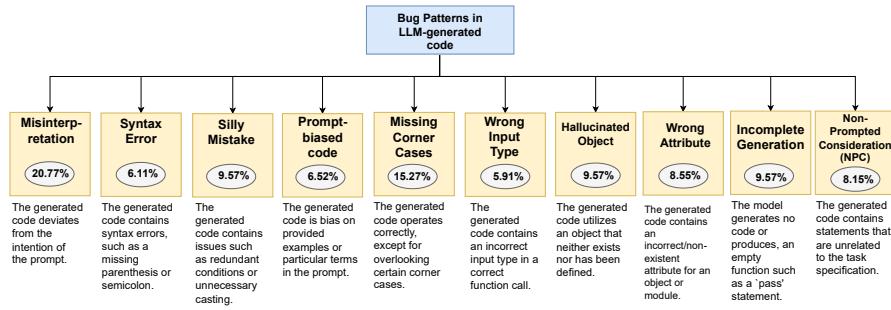


Fig. 2: Final taxonomy of bug patterns in code generated by LLM. The number following each category represents the percentage of code samples assigned to that category during manual labeling.

Listing 2: Reference Solution for the task “int\_to\_string”, and code generated by PanGu-Coder labeled as “Misinterpretation”.

```

1 def int_to_string(number: int, alphabet: List[str], padding: Optional[int] = None) -> str:
2     """
3         Convert a number to a string, using the given alphabet.
4
5         The output has the most significant digit first.
6         """
7     output = ""
8     alpha_len = len(alphabet)
9     while number:
10         number, digit = divmod(number, alpha_len)
11         output += alphabet[digit]
12     if padding:
13         remainder = max(padding - len(output), 0)
14         output = output + alphabet[0] * remainder
15     return output[::-1]
16 -----
17 def int_to_string(number: int, alphabet: str) -> str:
18     return alphabet[number]

```

**2. Syntax Error:** Any syntax error, such as a missing parenthesis or semicolon, is categorized as a *Syntax Error*. Listing 3 showcases a code example with a syntax error in the first line generated by CodeGen for the same task, “int\_to\_string”, as in Listing 2. In this case, the LLM failed to complete the function declaration (adding the closing parenthesis) at the end of the function signature which resulted in a syntax error. It is worth noticing that *Syntax Error* is one of the labels for the code example in Listing 3. Although this bug pattern is not very common across the models in our study, one of the models in our analysis generated more buggy code with specific syntax errors, as further discussed in Section 4.1.2.

---

Listing 3: Code generated by CodeGen and labeled as “Syntax Error”.

---

```

1 def int_to_string(number:
2     """
3         Convert a number to a string, using the given alphabet.
4
5     The output has the most significant digit first.
6     """
7     alphabet_string = ""
8     for i in range(10):
9         alphabet_string += chr(number % 10 + ord('A'))
10        number //= 10
11
12    return alphabet_string

```

---

**3. Silly Mistake:** We labeled issues such as redundant conditions or unnecessary casting as *Silly Mistake*. While these types of issues may not introduce a bug on their own, they have the potential to easily lead to one. In Listing 4, there is an example generated by CodeGen where both the *if* and *else* conditions have the same statement. This is labeled as *Silly Mistake*. This bug pattern ranks among the top three most common patterns across the models.

---

Listing 4: Code generated by CodeGen and labeled as “Silly Mistake”.

---

```

1 if not logfile.exists():
2     logging.basicConfig(
3         filename=str(logf),
4         format='%(asctime)s %(levelname)s %(name)s %(message)s',
5         datefmt='%Y-%m-%d %H:%M:%S',
6         level=logging.DEBUG,
7     )
8 else:
9     logging.basicConfig(
10         filename=str(logf),
11         format='%(asctime)s %(levelname)s %(name)s %(message)s',
12         datefmt='%Y-%m-%d %H:%M:%S',
13         level=logging.DEBUG,
14     )

```

---

**4. Prompt-biased code:** This issue occurs when the LLM excessively relies on provided examples or particular terms in the prompt while implementing a function and it sometimes hinders the generalization or correctness of the generated code. Listing 5 illustrates this issue. Lines 1 to 9 constitute the prompt, while lines 10 to 13 show the code generated by Codex for the task described in the prompt. The code should return the polygon for given vertices. The length of the vertices can be varied. However, lines 5 to 8 contain an example of vertices with length 4 and the list of its expected outputs, given

to clarify the task provided in the prompt. In the code in Listing 5 generated by Codex, the code is accurate only for the specific example provided in the prompt, when the length of the vertex is 4, and will fail in other cases. This code is labeled as a “Prompt-biased code”. This type of pattern is less common across the models; however, CodeGen and Codex generated more buggy code biased on the prompt compared to PanGu-Coder. We discussed more details in Section 4.1.2.

Listing 5: Code generated by Codex and labeled as “Prompt-biased Code”.

---

```

1 def vertex3tuple(vertices):
2     """returns 3 points for each vertex of the polygon. This will include the vertex and
3     the 2 points on both sides of the vertex:
4
5     polygon with vertices ABCD
6     Will return
7     DAB, ABC, BCD, CDA -> returns 3tuples
8     #A B C D -> of vertices
9     """
10    return [(vertices[-1], vertices[0], vertices[1]),
11              (vertices[0], vertices[1], vertices[2]),
12              (vertices[1], vertices[2], vertices[3]),
13              (vertices[2], vertices[3], vertices[0])]
```

---

**5. Missing Corner Case:** This bug occurs when the generated code operates correctly, except for overlooking certain corner cases. Listing 6 provides an example of this bug pattern along with its corresponding reference solution (top). Lines 17 to 24 represent the prompt and line 25 shows the code generated by Codex. The task involves implementing a function that checks if a given host is the localhost. Although the Codex-generated code checks for various representations of localhost, the reference solution is expected to verify additional potential options, such as `socket.gethostname()`, which were present in the oracle given for the task described in the prompt. Listing 6 is labeled as *Missing Corner Case*. This bug pattern is the second most common pattern with 15.27% of buggy code samples categorized as *Missing Corner Case*.

**6. Wrong Input Type:** We use this label when LLM uses an incorrect input type in a correct function call. For example in Listing 7, the task describes implementing a function to reduce a given list of classes to its ordered minimum equivalent. Lines 5 to 8 represent the code generated by CodeGen to address this task. In line 8, the Python built-in function `min` is called on a List of ClassType inputs which raises an error. We labeled this buggy code example as *Wrong Input Type*. It is worth noticing that this is not the only label for this buggy code example in Listing 7. *Wrong Input Type* is the least frequent bug pattern across the buggy code generated by three models.

**7. Hallucinated Object:** We apply this label when the LLM hallucinates by attempting to utilize an object that neither exists nor has been defined. For instance, Listing 8 represents the code generated by Codex for the

Listing 6: Reference Solution for the task “is\_local” and code generated by Codex, labeled as “Missing Corner Case”.

---

```

1 import socket
2 def is_local(host):
3     """
4         Checks if the host is the localhost
5
6         :param host: The hostname or ip
7         :return: True if the host is the localhost
8         """
9     return host in ["127.0.0.1",
10                     "localhost",
11                     socket.gethostname(),
12                     # just in case socket.gethostname() does not work
13                     platform.node(),
14                     socket.gethostbyaddr(socket.gethostname())[0]
15                 ]
16
17 import socket
18 def is_local(host):
19     """
20         Checks if the host is the localhost
21
22         :param host: The hostname or ip
23         :return: True if the host is the localhost
24         """
25     return host in ['localhost', '127.0.0.1', '::1']

```

---

Listing 7: Code generated by CodeGen and labeled as “Wrong Input Type”.

---

```

1 def minimalBases(classes):
2     """
3         Reduce a list of base classes to its ordered minimum equivalent
4         """
5     if len(classes) == 1:
6         return classes[0]
7     else:
8         return min(classes)

```

---

task “make\_find\_paths” that involves transforming all the paths given in the find-paths variable into glob patterns. The generated function calls a function named “find\_path\_to\_glob” (line 14), which is not defined. *Hallucinated Object* is among the third common bug patterns across the models.

**8. Wrong Attribute:** This label is applied when the LLM utilizes an incorrect/non-existent attribute for an object or module. For example, in Listing 9, there is code generated by PanGu-Coder for the same task presented in Listing 1 in the introduction, named “parser\_flag”. The prompt describes the task as “Given an argparse.ArgumentParser instance, returns its argument

---

Listing 8: Code generated by Codex and labeled as “Hallucinated Object”.

```

1 def make_find_paths(find_paths):
2     """
3         Given a sequence of path fragments or patterns as passed to '--find', transform all
4         path fragments into glob patterns. Pass through existing patterns untouched.
5
6     For example, given find_paths of:
7
8         ['foo.txt', 'pp:root/somedir']
9
10    ... transform that into:
11
12        ['sh:**/*foo.txt**/*', 'pp:root/somedir']
13
14    return [find_path_to_glob(x) for x in find_paths]

```

---

flags in a space-separated string.” The code generated by PanGu-Coder tries to use ‘flags’ as an attribute of the input ‘parser’, which is not a correct attribute according to the information provided in the prompt. The code sample in Listing 9 is labeled as “Wrong Attribute”. This bug pattern is ranked as the fourth most common pattern, closely following the bug patterns in the third position.

---

Listing 9: Code generated by PanGu-Coder and labeled as “Wrong Attribute”.

```

1 def parser_flags(parser):
2     """
3         Given an argparse.ArgumentParser instance, return its argument flags in a
4         space-separated string.
5
6     return''.join('--{0}'.format(f) for f in parser._flags)

```

---

**9. Incomplete Generation:** This label is applied when the LLM generates no code or produces, for example, a ‘pass’ statement or an empty function. This bug pattern also includes cases where the LLM stops completing the code in the middle of the function either because of token limits or because it reached an <end\_of\_sequence> token without properly finishing the code. Listing 10 generated by CodeGen is an example of this category. *Incomplete Generation* with *Hallucinated Object* and *Silly Mistake* are the third most frequent bug patterns we observed.

**10. Non-Prompted Consideration (NPC):** This label is applied when the LLM adds statements to the code that are unrelated to the task specified in the prompt, leading to errors. For instance, in Listing 1, which was presented as an example in the introduction, there is a code generated by Codex, for the task named “parser\_flag”. The prompt involves “Given an arg-

Listing 10: Code generated by CodeGen and labeled as “Incomplete Generation”.

---

```

1 def write_configuration(config_filename, rendered_config, mode=0o600, overwrite=False):
2     """
3         Given a target config filename and rendered config YAML, write it out to file. Create any
4         containing directories as needed. But if the file already exists and overwrite is False,
5         abort before writing anything.
6         """
7     if os.path.exists(config_filename) and overwrite:
8         print("Overwriting " +
9             #

```

---

`parse.ArgumentParser` instance, returns its argument flags in a space-separated string.” As we already discussed in the introduction, although the prompt does not specify sorting the flags, at line 14, the generated function sorts the list of flags before joining them, which is not in line with the given prompt. *NPC* with a slight variation from the *Wrong Attribute* type, stands at the fifth position with 8.15% of buggy code examples across all models.

#### 4.1.2 Analysis of the Bug Patterns

The bug patterns have been defined based on the bug patterns observed in our studied samples. As mentioned in Section 3, some of the buggy code samples are labeled in multiple categories due to the presence of different types of bugs or because a bug could be labeled as an overlap of multiple categories: **62.16%** of the samples are labeled in a single category, **29.13%** with two, **8.41%** with three categories, and only one instance of buggy code is assigned four labels.

Table 2 presents the distribution of buggy code across different categories and by various LLMs. Some categories are more prevalent in specific models. For instance, in the case of a robust model like Codex, the most common bug pattern is *Missing Corner Case* (**23.53%** of the buggy samples). *Missing Corner Case* occurs when the model generates code that addresses the task description but fails on one or a few exceptional test inputs due to a minor oversight, such as forgetting some particular cases. Listing 6 presents an example of this type. Conversely, in the case of relatively weaker and open-source models like PanGu-Coder or CodeGen, the most common category is *Misinterpretation*. This bug pattern occurs when the code generated by the model does not align with the prompt and deviates from the intended purpose of the task.

Among all models, CodeGen has the highest number of buggy samples in the *Silly Mistake* and *Syntax Error* categories. Codex, as a stronger model, exhibits the highest number of samples in the *Non-Prompted Consideration* category. This category pertains to situations where the model considers additional conditions or statements that were not requested in the prompt, leading to bugs. For example, the model might convert the final output of a function

Table 2: The distribution of bug patterns in the generated code by three different LLMs. We put in bold the top categories per LLM.

| Bug Pattern           | CodeGen       | PanGu-Coder   | Codex         | Total         |
|-----------------------|---------------|---------------|---------------|---------------|
| Misinterpretation     | <b>22.84%</b> | <b>24.11%</b> | <b>15.03%</b> | <b>20.77%</b> |
| Syntax Error          | <b>9.14%</b>  | 4.96%         | 3.27%         | 6.11%         |
| Silly Mistake         | <b>14.72%</b> | 4.26%         | 7.84%         | <b>9.57%</b>  |
| Prompt-biased code    | 8.12%         | 3.55%         | 7.19%         | 6.52%         |
| Missing Corner Case   | 5.58%         | <b>19.86%</b> | <b>23.53%</b> | <b>15.27%</b> |
| Wrong Input Type      | 6.60%         | 8.51%         | 2.61%         | 5.91%         |
| Hallucinated Object   | 5.08%         | <b>14.89%</b> | 10.46%        | <b>9.57%</b>  |
| Wrong Attribute       | 7.61%         | 9.93%         | 8.50%         | 8.55%         |
| Incomplete Generation | <b>13.71%</b> | 3.55%         | 9.80%         | <b>9.57%</b>  |
| NPC                   | 6.60%         | 6.38%         | <b>11.76%</b> | 8.15%         |
| <b>Total</b>          | 100%          | 100%          | 100%          | 100%          |

into an integer, which was not required in the prompt, or check if the output is a string and then apply further functionality (see Listing 1).

The third most common bug pattern in code generated by PanGu-Coder is *Hallucinated Object*. This occurs, for example, when the model erroneously hallucinates the existence of a function that does not exist (see Listing 8). Our observations reveal that when the model struggles to address the prompt, it sometimes resorts to calling an auxiliary function that is not defined, assuming that this function accomplishes the primary objective of the task. Usually, the function’s name is generated based on the words used in the prompt.

When considering all three studied models, *Misinterpretation* is the most common bug pattern (**20.77%**). The second place belongs to *Missing Corner Cases*, *Hallucinated Object*, with *Incomplete Generation* and *Silly Mistake* following closely in the third position. It is noteworthy that the last three mentioned bug patterns should rarely be observed in code written by humans. For example, human developers typically do not frequently call undefined functions (*Hallucinated Object*) or access attributes on objects that do not exist (*Wrong Attribute*). Some of these bug patterns can be detected by Python Linters embedded in various IDEs, such as VS Code, which further reduces the likelihood of encountering them in human-written code. For the *Silly Mistake* category, a common mistake is when the model generates an if/else condition but uses the same statement in both branches, which is also a rare mistake in human-written code.

**Findings 1:** While buggy samples in CoderEval exhibit *Misinterpretation* and *Missing Corner Cases* bug patterns (that would lead to an assertion error during unit testing), there are also a non-trivial number of more severe bug patterns such as *Hallucinated Object* or *Wrong Attribute*. Those mistakes are not typical of human developers, either because they show strange behavior (e.g., *Silly Mistake* with duplicated if/else condition) or because they would be easily detected using an IDE.

To further study bug patterns in LLM-generated code on the CoderEval benchmark, we look into what types of bugs are present, for each task in the benchmark, across LLMs. The goal is to identify if some tasks in the dataset are more likely to lead to a specific bug pattern across all LLMs which could indicate some potential shortcomings in the prompts or if bug patterns are evenly distributed across different tasks in the CoderEval benchmark. To do so, we counted in our sample set, for each task, the number of times a particular bug pattern was made, regardless of the employed LLM. We then normalized values per task to know what bug patterns per task was more prevalent. Results are presented as a heatmap in Figure 3.

Overall, the various bug patterns are evenly distributed across different tasks, as indicated by the darker areas, less than 0.5, in the heatmap, Figure 3. This suggests that the tasks are not biased toward a particular bug pattern that we identified. However, the few very light spots on the heatmap (greater than 0.9) show the majority of code samples for a specific task, categorized under a particular bug pattern. Based on our results, only two tasks, all the buggy code snippets were classified under a single bug pattern. For example, for the task named “fix\_namespace\_prefix\_w” where all the buggy code snippets generated by different LLMs are categorized as *Misinterpretation*. This classification stems from the fact that the prompt associated with this task, “Convert text that defaults to ‘w:st=’ ’ to ‘w-st=’ ’, does not accurately represent the functionality revealed by the oracle and corresponding test cases. However, all models generated at least one correct solution based on the same prompt, which passed all the test cases.

Additionally, as shown in Figure 3, only for a few tasks, the majority of bugs are concentrated in a specific type. For example, in the “force\_string” task, where up to 72% of the bug patterns generated by different models are categorized as *Missing Corner Case* due to an additional newline character in the prompt. The prompt for this task reads, “This function returns the bytes object corresponding to ‘obj’ / in case it is a string using UTF-8.”. The content after the backslash appears as a new line in the prompt, creating ambiguity in the prompt’s objective for all LLMs. In the buggy code snippets, the function attempts to return the bytes object corresponding to ‘obj’. In cases where ‘obj’ is a string, it decodes it to UTF-8. However, upon checking the oracle, it becomes evident that the intended purpose of the prompt is to return the corresponding UTF-8 if ‘obj’ is an instance of a byte object.

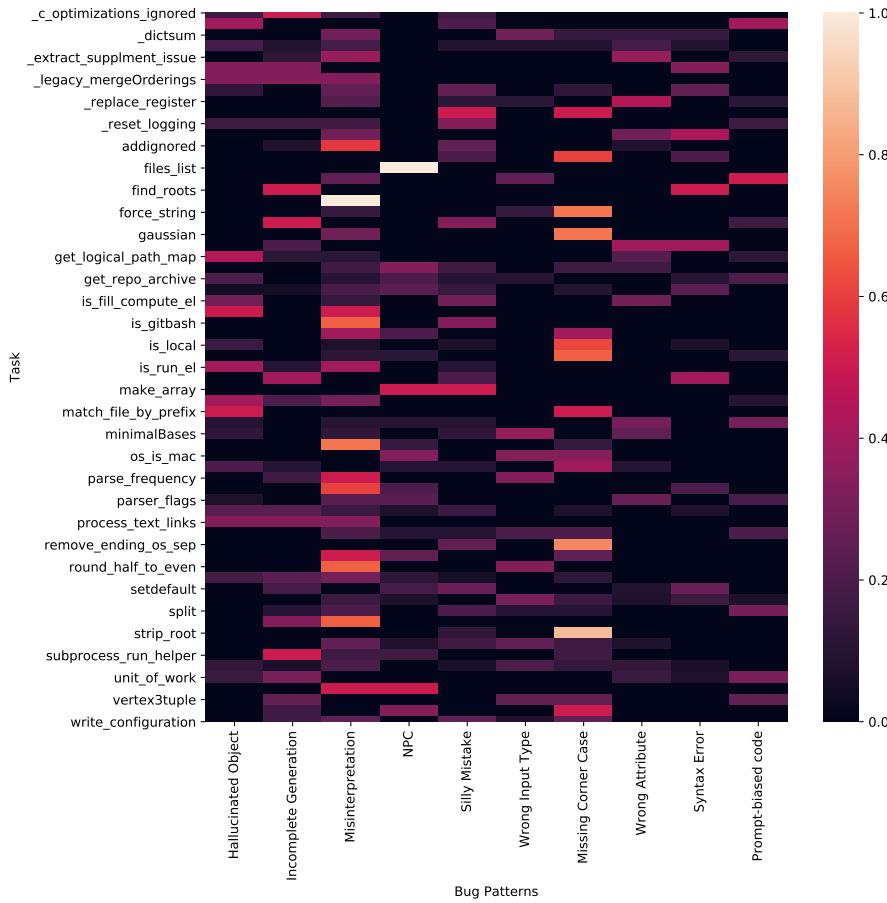


Fig. 3: The heatmap illustrates the distribution of bug patterns on various tasks across all LLMs. The values are normalized over lines per task. The brighter areas represent a higher number of code samples of a specific task categorized into a particular bug pattern. The black areas indicate the lack of code samples in the bug pattern for a specific task.

As discussed in Section 3.1.1, our result is limited to tasks with runnable levels no higher than “*plibRunnable*” (dependent only on Python’s standard and public libraries, such as “*numpy*”). The goal of this limitation is to mitigate potential biases in our results that may arise from the lack of information on dependencies not covered in the prompt. However, to further our analysis, we assess whether the three selected runnable levels in our sample set cause any bias towards specific bug patterns. To conduct this analysis, we report the distribution of bug patterns in different runnable levels of tasks across LLMs, as shown in Table 3.

Table 3: Distribution of bug patterns across different runnable levels (levels of task dependencies).

| Bug Pattern           | self_runnable | slib_runnable | plib_runnable | Total         |
|-----------------------|---------------|---------------|---------------|---------------|
| Misinterpretation     | <b>16.29%</b> | <b>24.86%</b> | <b>23.53%</b> | <b>20.77%</b> |
| Syntax Error          | <b>7.69%</b>  | 4.32%         | 5.88%         | 6.11%         |
| Silly Mistake         | 8.60%         | <b>11.35%</b> | 8.24%         | 9.57%         |
| Prompt-biased Code    | 8.14%         | 6.49%         | 2.35%         | 6.52%         |
| Missing Corner Case   | 10.41%        | <b>17.84%</b> | <b>22.35%</b> | <b>15.27%</b> |
| Wrong Input Type      | <b>6.79%</b>  | 4.86%         | 5.88%         | 5.91%         |
| Hallucinated Object   | 9.95%         | 9.73%         | 8.24%         | 9.57%         |
| Wrong Attribute       | <b>13.12%</b> | 2.16%         | <b>10.59%</b> | 8.55%         |
| Incomplete Generation | 9.50%         | 10.27%        | 8.24%         | 9.57%         |
| NPC                   | 9.50%         | 8.11%         | 4.71%         | 8.15%         |
| <b>Total</b>          | 100%          | 100%          | 100%          | 100%          |

The results, as indicated in Table 3, show a varied distribution of bug patterns across different runnable levels. Similar to the result in Table 2, the *Misinterpretation* category contains the maximum number of buggy code snippets in each dependency level, with the second-highest occurrences being *Missing Corner Case* for “*slib\_runnable*” and “*plib\_runnable*”, and *Wrong Attribute* for “*self\_runnable*”. While higher runnable levels are more likely to cause generating buggy code by LLMs, as highlighted in CoderEval benchmark [6], the results in Table 3 do not demonstrate that increasing the dependency level of the task up to “*plib\_runnable*” impacts the bug patterns in code snippets for most of the bug patterns.

We observed only the following trend in the case of three bug patterns: *Missing Corner Case* (an increasing trend with runnable level) and *Prompt-biased code/Non-Prompted Consideration* (a decreasing trend with runnable level). For the *Missing Corner Case*, the results suggest that increasing the runnable level might make the model more likely to miss some expected cases because the prompt might lack information on how the library is expected to be used to tackle the problem. On the contrary, the amount of *Prompt-biased Code* and *NPC* increased when the runnable level decreased (a decreasing trend with runnable level) since the models might rely more on the details in the prompt for tasks in the CoderEval dataset with a lower level of dependencies on the libraries. Similar trends are not observable for other types of bugs at different runnable levels. Therefore, considering the runnable levels in our dataset, our proposed bug patterns are not influenced by a specific runnable level.

**Findings 2:** The bug patterns are evenly distributed over different tasks in the CoderEval dataset. The bug patterns are also evenly distributed over different runnable levels (dependency levels) of tasks in this dataset. However, our results reveal that the ambiguity present in the description of a task in this dataset may impact the distribution of buggy samples within a specific task between different bug patterns across different LLMs.

#### 4.2 RQ2: To what extent are the identified bug patterns in LLM-generated code relevant for software practitioners and researchers working with LLMs?

The survey was open for two weeks resulting in 34 answers, i.e., we achieved a return rate of 8.9%, which is in line with many SE surveys conducted outside specific companies [36, 37]. We first give results of the questions concerning the demographics and experience of the participants. We then detail results regarding the frequency of reported bug patterns, as well as the analysis of the reported difficulty of diagnosing/fixing and the perceived complexity of the bug patterns (see Table 4).

##### 4.2.1 Participants' demographics and experience with LLM bugs

All the participants answered the demographic questions. We obtain the following answers:

- For the academic field: 12 Ph.D. students, 4 Researchers, 8 Undergraduate/Graduate students and 1 Lecturer. For the industry field: 6 Developers, 2 Data scientists and a CTO.
- ChatGPT was used by **31** participants. **67%** of the participants reported only non-open-source LLMs such as ChatGPT, Bard [38] or Claude [39]. **10** participants cited open-source LLMs such as LLama or CodeGen.
- In terms of programming language used for LLMs, besides Python, the participants responded: **47%** for JavaScript, **27%** for C++ and Java and **24%** for C. These languages are among the most used programming languages according to different programming language indexes [40, 41]. Other programming languages were reported in smaller proportion: Rust (6%), HTML/CSS (6%), SQL (9%), C# (6%), or Go (3%). Three respondents (all PhD students) reported exclusively using LLMs only to program in Python language.

Note that the sum of responses for LLMs and programming languages goes above the number of responses as one could provide multiple answers.

Before asking participants specific questions about the categories of bugs, we collected information about their general experience dealing with bugs in LLM-generated code. **80%** of the participants reported that they often or always experience bugs when using LLMs to generate code. No respondent said they *never* faced bugs in LLM-generated code. We asked the participants about the complexity of the bugs they encountered in their LLM code generation activities and **68%** of them mentioned that the bugs had a medium difficulty to fix, with only **21%** of participants reporting that their encountered bugs were *easy* to fix. We also asked the participants how they proceeded to fix the encountered bugs and **30%** of them reported that they fixed the bugs manually by reading the stack traces and googling the error messages. The remaining participants reported using a combination of manual inspection with re-prompting the LLM to correct the mistake. No participant declared only

using re-prompting based on the error message, similar to approaches presented in [42]. This highlights the fact that debugging LLM-generated code is still predominantly a manual task. We believe that the bug patterns proposed in this paper are an important first step toward the development of efficient automatic debugging tools for LLM-generated code.

#### 4.2.2 Frequency of bugs patterns

Figure 4 summarizes the obtained aggregated results in the “Frequency” column. A table with the results for the answer per Likert scale score of the respondents is also available in the Appendix B.2. For each category of bug, we report the following: the frequency at which participants declared they encountered the bug pattern, what is the perceived difficulty in diagnosing such a bug pattern, what is the perceived complexity of the bug pattern (i.e., is it a trivial mistake a competent developer would not do or does it denotes some complexity), and what is the perceived difficulty to fix such a bug pattern. For each of those metrics, we report the weighted average results giving the proportion of respondents that answered 1/2/3/4/5 given the Likert scale as described in Section 3.2.3. For instance, the *Missing Corner Cases* bug pattern has a score of **3.23** and is among the most frequent bug patterns encountered by participants. All categories in the taxonomy have been encountered and acknowledged by the participants. The most encountered categories are *Wrong Attributes*, *Prompt-biased Code*, or *Missing Corner Cases* with an aggregated score above 3. Thus, participants report that they often encounter bugs from these categories. One participant reported that they observed *Prompt-biased Code* bugs when asking LLMs to generate code using a prompt with several examples of unit tests: “*I face this problem when docstring given too many examples, and the prompt requires LLMs to pass at least all the test cases*”. The categories of bugs that were encountered the least are *Syntax Error* and *Silly Mistake* with scores respectively of **2.03** and **2.35**. Thus, most participants reported that they never or rarely encountered these two categories of bugs. This distribution is in line with the results obtained in RQ1, for Codex (one of the most popular LLMs) using the CoderEval dataset: *Syntax Error*, *Silly Mistake*, and *Wrong Input Type* are the least frequent bug categories in Codex-generated code (see Table 2). Moreover, the *Missing Corner Case* is the most frequent bug category in code generated using Codex. To confirm this observation, we conducted a correlation analysis as described in Section 3.2.2 between the frequency of occurrence of bug patterns reported by respondents and the frequency of bug patterns observed in our sample set. The highest correlation is obtained for Codex ( $\rho = 0.47$ , medium correlation) which seems plausible as we have seen at the beginning of this section that a large majority of the respondents worked using ChatGPT and/or Codex. The  $\rho$  for PanGu and CodeGen are **0.28** and **-0.18** respectively. As such, the frequency of the bug patterns reported by the respondents is similar to the distribution of the bug patterns sampled for the Codex model in our sample set.

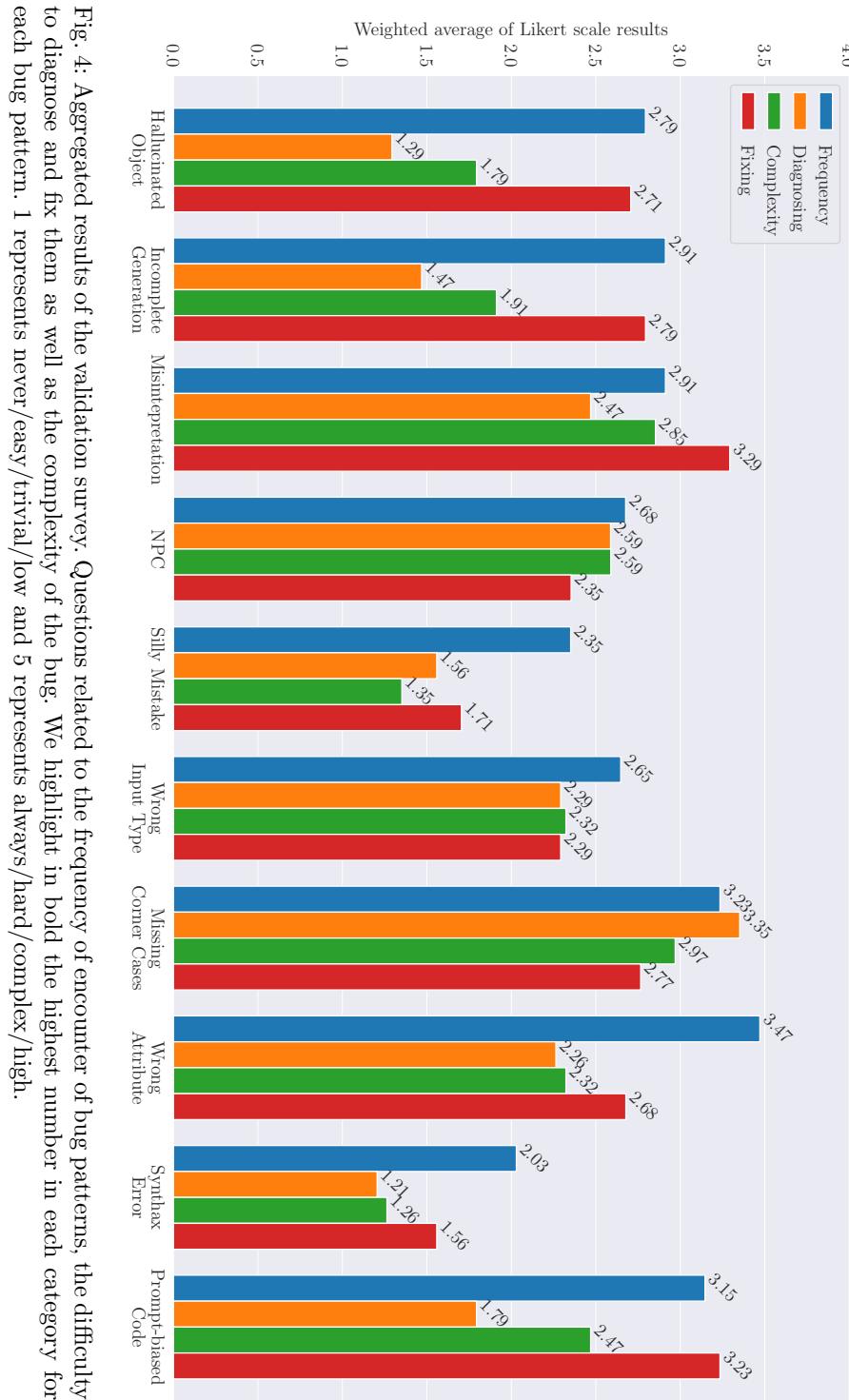


Fig. 4: Aggregated results of the validation survey. Questions related to the frequency of encounter of bug patterns, the difficulty to diagnose and fix them as well as the complexity of the bug. We highlight in bold the highest number in each category for each bug pattern. 1 represents never/easy/trivial/low and 5 represents always/hard/complex/high.

We do note some discrepancies such as *Prompt-biased Code* or *Wrong Attribute* type being more prevalent in respondents' answers than what we observed in CoderEval or *Misinterpretation* not being as prevalent. This can be explained by the type of prompt used to generate code in CoderEval. Indeed, most of the prompts were extracted from real projects so information in the docstring might be limited which impacts the LLMs' generation (“*Usually more outlining or instructions can be added after a docstring to help guide the model*”). This can be, for instance, because the developers who designed the code understood it. Thus, they put a limited docstring as it was sufficiently understandable for them, yet it might not be enough for an LLM. Similarly, CoderEval's prompt generally does not include an example of the task at hand which might limit *Prompt-biased Code* prevalence. Nonetheless, even with a limited docstring, some errors such as *Wrong Attribute*, *Hallucinated Objects* or *Wrong Input Type* are unexpected as they are more linked to expertise in a programming language and library used rather than an understanding of the task at hand.

#### 4.2.3 Diagnosing, Complexity, and Fixing of bug patterns

Figure 4 summarizes the obtained results in the “Diagnosing”, “Complexity” and “Fixing” columns which were defined to participants (see Section 3.2.2). A table with the results for the answer per Likert scale score of the respondents is also available in the Appendix B.2. In terms of diagnosing the bugs, most bug patterns were reported by the respondents as easy to diagnose (score < 2), especially as several respondents pointed out that: [the bugs are] “*Easy to diagnose when using an IDE with real-time code checking*”. According to our survey respondents, the only bug patterns that are not as easy to diagnose are *Misinterpretation*, *Non-Prompted Consideration*, and *Missing Corner Cases*. In the case of *Non-Prompted Consideration*, a respondent highlighted that: “*I have observed some difficult cases that non-requested behaviour is in the middle of the generated code*”. *Missing Corner Cases* is the only category reported as being quite hard to diagnose by respondents (score > 3). This result could be expected since the *Missing Corner Cases* category encompasses code that will compile correctly, but break unit tests, as highlighted by one of the participants: “*It is a risk to become complacent to LLM code generation because sometimes code looks reasonable (in this case). Still, you might be unaware of all the ways it could fail*”. Another respondent emphasized the importance of efficient test cases when addressing bugs from the *Missing Corner Cases* category, noting that: “*It is very dependent on the evaluation test case set. From its short natural language description, like docstring, it's hard to cover all its possible corner cases*”.

Regarding how likely a developer is to make such a mistake (i.e., “complexity” of the bug patterns), respondents mostly consider the identified bug patterns to have low complexity, with *Syntax Error* and *Silly Mistake* types being the least complex. This adds to our intuition that most bug patterns in LLM-generated code are different from human-made bugs. Indeed, as most

developers use IDE, *Syntax Errors* or similar mistakes are pretty easy to catch. Similarly *Silly Mistakes*, such as the example presented in Listing 4, are not a bug that developers would likely do, and, even if done, they are easy to detect. On the contrary, *Misinterpretations* and *Missing Corner Cases* bug patterns are more complex and more similar to a type of error that a developer would make. This is highlighted by the respondents, as both categories have the highest aggregated score. One survey participant commented that: “*For example, I would very much forget about IPv6 (don’t we all?:)) and miss that corner case myself. So it is both my lack of knowledge and implicitly trusting the LLM output that can make this such an insidious bug to diagnose and solve*”. Thus, the complexity of such bugs could arise from both the nature of the error and developers’ excessive reliance on LLMs, which amplifies the impact and intricacy of the issue.

Regarding the difficulty of fixing the bugs from the different categories, the survey participants consider the bugs to be moderately difficult to fix; mirroring their responses in the first part of the survey, when we asked about the difficulty of fixing LLM-generated bugs. Only *Syntax Errors* and *Silly Mistakes* are considered easy to fix with an aggregated score of < 2. Bugs from the *Misinterpretation* category are reported to be more difficult to fix than the other categories by the survey’s participants. This result is consistent with the observation we made during the labeling phase. *Misinterpretation* bugs generally lead to a code that is way different compared to the prompt specification. In that case, re-prompting or extensive manual effort is needed to fix the generated code fragment. For some bug categories, like *Hallucinated Objects*, participants suggested that providing additional information in the prompt can often help the LLMs to correct the mistake: “*Most often time after providing a header or function definition, the model can correctly create the method or function*”. However, in several cases, especially for bugs in the *Misinterpretation* category, it is not enough to provide additional information: “*It depends on whether you know what you are expecting the code to look like. If so, this is easy to spot, and probably you’re better off manually writing the code. I have tried to prompt-engineer, but I find that it’s a waste of my time. Sometimes, you will get code generated that misinterprets the prompt, but maybe you are not aware of what the generated solution looks like so you have to read the generated code intently or figure out the library calls that are generated to understand that it misunderstood the prompt*”.

Finally, as we reported in Section 4.1, some LLMs like Codex (and so Copilot), tend to generate additional code that is not explicitly requested in the task’s description. This may be linked to the different settings of the LLM: an IDE plugin for Copilot/Codex or a web browser/standalone for CodeGen and PanGu-Coder. This was echoed by one of the respondents of our validation survey who also observed this kind of behavior being more prevalent in Copilot-based models compared to the others: “*There is a difference in LLM-generated code quality between using an in-IDE plugin such as copilot, and using a browser-based tool such as chat gpt or google bard. I find that copilot*

*has more of a tendency to recommend code that I don't need, or code that is redundant; [...]”.*

**Findings 3:** The surveyed practitioners encountered the different bug patterns in our empirical results. In particular, their reported frequency of bug patterns correlates with the frequency observed for *Codex* in our sample set. They reported that bug patterns are generally easy to diagnose and fix but it requires manual effort. Those LLMs bug patterns are generally different from human-made bugs. *Misinterpretation* and *Missing Corner Case* bug patterns are the only exceptions: being harder to fix and diagnose, as well as denoting faults that are more akin to human-made.

## 5 Lessons Learned

This study shed light on the nature of bugs occurring in code generated using LLMs. We observed that bugs contained in LLMs-generated code differ from bugs in human-written code. Our analysis and the feedback received from survey participants also reveal that LLM-generated code can appear correct at first glance and deceive users depending on the complexity of the code and the experience of said users. To the light of our manual labeling and validation survey, in the following, we provide several lessons learned for both LLMs' users and researchers.

### 5.1 Lessons Learned for LLM Users

#### Limited information in the prompt can drastically lead the model astray:

CoderEval's tasks are composed of a function signature with a docstring extracted from actual GitHub repositories. While this might be enough to some extent for a human, LLMs seem to struggle with this somewhat limited information. One issue we observed in our sample set concerns the library or even the Python version. Indeed, there is little control over which version of a given programming language or library the LLMs will use in the generated code if it is not indicated explicitly in the prompt. We came across examples of generated code that would only work for a certain version of Python while the prompt and the oracle code were pointing to another version. For instance, when parsing XML files, the LLMs would sometimes use the function “`getchildren()`” which is deprecated for certain versions of Python. This is something that some respondents in our survey reported: (“*Sometimes, the LLM would generate functions that are deprecated. It is not necessarily hallucinating, it was just trained on an older version of the Python package you are using*”). Interestingly, by looking at other samples of the same task not necessarily in our sample set, we found that the usage of this function by LLMs was not

systematic and some samples were correctly using the expected (more recent) function. Hence, having an older version in the generated code can not only be explained by deprecated training data. Similarly, on tasks requiring a different coding between Python 2 and 3, the LLMs would often only consider one of them (e.g., using the API for Python 3 for both the Python 2 and Python 3 cases).

*Recommendation:* Re-prompting or adding additional information (for instance Chain Of Thoughts [43, 44] or examples of test cases [45]) can help the LLM in the code generation process. This is noted by a respondent of our survey: “*Most often time after providing a header or function definition, the model can correctly create the method or function*”. For instance, in the cases we observed in our sample set, with the mismatch between the desired version of the library by the user and the obtained version in the LLM’s code, providing the concrete version of the libraries can help. By doing so in the prompt, the user can constrain the LLM to generate accordingly. For instance, Guilherme et al. [46] generated Java test cases for a particular Java/JUnit version by explicitly mentioning it in the prompt.

**More testing, less trusting:** Even when given enough information, LLMs might generate faulty codes. In that case, having access to test cases to properly test the generated code is important. This however can be quite a manual effort. In a regression test setting, using automated test case generation tools such as EvoSuite [47] can help. Leveraging LLMs themselves was also shown to be a potential way to generate unit tests [8, 48]. Nonetheless, one should be cautious even when test cases are used. For instance, in the case of *Misinterpretation* or *Missing Corner Case*, the error might not be obvious, and potential existing test cases might not catch the error in the LLMs’ generated code. For instance, Liu et al. [8] showed that existing test cases in HumanEval dataset [4], which encompasses simpler tasks compared to CoderEval, would not catch all existing mistakes in LLMs’ generated code. Worse, if no tests are available, users either have to trust the LLMs’ generated code or manually read and analyze the obtained code to make sure there are no errors. The first option introduces a high risk depending on the context in which the code should be used while the second one adds a tedious and time-consuming task for the developers, proportional to the code complexity. With the previous points mentioned, it is obvious that LLM-generated code should not be too trusted as it can easily be error-prone. If more experienced developers are less likely to fall into this trap, novice users or people unfamiliar with the particular task requested in the prompt can easily trust generated code, especially if it “looks” correct, as highlighted in one comment from our survey participants: “*It is a risk to become complacent to LLM code generation, because sometimes code looks reasonable (in this case), but you might be unaware of all the ways*

*it could fail*". Some previous studies also pointed out that more novice users tend to over-rely on the code [9, 49, 50] which can have serious negative effects.

*Recommendation:* When possible, users should make use of test cases. Both in the prompt to improve the generation process of the LLMs [45], as well as to check the generated code for potential errors. In any case, users (especially novice users) should be mindful when reusing LLMs generated code by trusting the solution is correct. One could draw an analogy with StackOverflow where several studies [51, 52, 53] showed that users tend to rely on those forums while the code present here might not be always reliable.

**Knowing how and when to use LLMs:** In some cases, if too many details are needed for the LLMs to generate a proper code, it might be more straightforward to write the function directly. For instance, Vaithilingam et al. [54] showed that users leveraging Copilot for coding tasks were not necessarily faster than users without Copilot for the reasons we observed in our study: the code generated might be buggy, oftentimes not in a straightforward way (*Misinterpretation/Missing Corner Case*), or the users might have a hard time understanding the logic of the generated code. This is best put by one of our survey's respondents: "*It depends on whether you know what you are expecting the code to look like. If so, this is easy to spot, and probably you're better off manually writing the code. I have tried to prompt and engineer, but I find that it's a waste of my time. Sometimes, you will get code generated that misinterprets the prompt, but maybe you are not aware of what the generated solution looks like so you have to read the generated code intently or figure out the library calls that are generated to understand that it misunderstood the prompt*". Moreover, straight code generation might not be recommended in all cases, such as we analyzed in our study. An alternative would be to relax the problem to instead do *code completion* as commented by a survey's participant: "*In most of my usage scenarios, I generate code with LLM based on existing code. It could be different from letting LLMs generate code from scratch.*". In that case, the user can start coding the functionality and then use the LLM to complete some lines to lower the workload. The additional context would help the model in making more accurate and less error-prone generations. Even in the case of the generated code containing an error, it will likely be easier for a developer to debug as part of the logic was implemented by said developer and the generated part should also follow a similar logic. This, however, requires some context to work with, which might hamper the usefulness of the method. Another way to mitigate full code generation would be to use LLMs to kick-start coding, saving time on more redundant or simple parts. This has the advantage of not necessarily needing code context and letting the users have more control over the code. This was another point mentioned in the study of Vaithilingam et al. [54]: even when they did not observe a significant time advantage while using LLMs, developers still preferred them because they

could kick-start the coding process and avoid searching for solutions on the Internet.

*Recommendation:* On one hand, if the user has a precise idea of what needs to be done, the advantage of generating the code brought by the LLMs might not offset the time it takes to correctly prompt the LLMs and to debug/understand the generated code by the user. In that case, manual coding can be more recommended. Alternatively, code completion can be used, as the additional code context provided by the users can both guide the LLMs and lead to a code that is more understandable by the users. On the other hand, if the user has no idea how to tackle the task, LLMs can serve as a code kickstart.

## 5.2 Lessons Learned for Researchers

Following our results and analysis, we provide additional lessons learned that could be future potential research venues for the community regarding bugs in the code generated by LLMs:

**Repairing the code generated by LLMs:** One way to mitigate bugs in LLM-generated code would be to directly repair them, ideally before the code is returned to the user. Automatic Program Repair (APR) tools [55] have been used to fix traditional software programs and several studies [56, 57] started applying them to code generated by LLMs. These studies, however, made use of simpler datasets (e.g., HumanEval [31] or LeetCode [58], which are not based on programming tasks extracted from real-world projects) and still failed to find several mistakes. Hence, we recommend analyzing existing bugs in practical tasks similarly to what we did, which could be beneficial to guide such repair. This could help improve existing APR tools for LLMs generated code or complement prompt for LLMs to self-check generated code against identified bug patterns.

**Proposing Code features related Benchmarks for Testing code LLMs:** We observed that LLMs-generated code can contain different bug patterns and that, depending on the code dependency level, LLMs might be more prone to certain errors. Thus, creating standardized benchmark datasets and evaluation metrics specifically tailored for assessing code-related features, the performance of bug detection and triaging methods for LLM-generated code. This will enable fair comparison and evaluation of different testing approaches. Existing benchmarks already focusing on code repair such as CodeXGlue [59] or HumanEvalPack [60], on dependency level such as CoderEval [6] or ClassEval [61] or even on code efficiency such as EffiBench [62] are important to be considered. Hence, promoting new benchmarks or complementing the above benchmark with additional features such as different bug patterns, dependency levels or other code-related features could help in better assessing LLMs shortcomings.

**Interpretable AI for Code Generation:** Besides potential APR techniques, fixing and diagnosing the bugs in LLMs generated code remain manual tasks. This task can be quite intensive depending on the nature of the bug patterns introduced such as *Missing Corner Case* or *Misinterpretation* as we described in our study. As such, employing interpretability techniques for LLMs could help in providing insights into potential fixes or assisting in bug detection. Such techniques analyze how input (i.e., prompt) variations affect a model’s output [12]. By monitoring the effect of presence or absence of different combinations of input elements, these approaches facilitate uncovering the relative contribution of each element to the output, enabling model analysis without necessitating access to its internal architecture [63]. Thus, applying interpretable AI techniques to LLMs to enhance transparency in code generation is also an important research direction. For instance, Ji et al. [64] analyzed how prompt changes impact code generation.

## 6 Related Works

In this section, we review related works and highlight their findings relevant to buggy code generated by LLMs. While various studies have incorporated LLMs for diverse programming tasks, our focus is specifically on studies that investigated bugs in LLM-generated code. We also discuss studies proposing LLM-based programming assistant tools and studies that examine the quality of code generated by LLMs.

Vaithilingam et al. [54] conducted a human study involving 24 participants to investigate the user experience of using Copilot for completing programming tasks. Their study revealed that participants using Copilot had a lower success rate in accurately completing these tasks compared to those using Intellisense<sup>7</sup>. This occurred primarily because the participants struggled to detect and correct the errors contained in the code generated by Copilot. Participants faced difficulties correcting the bugs generated by Copilot, to the point that they preferred writing the code from scratch instead of repairing the bugs contained in Copilot’s code.

Moradi et al. [9] assessed the quality of code generated by Copilot for solving different fundamental algorithmic problems such as searching and sorting. Moradi et al. also compared the quality of code generated by Copilot as an AI pair programming assistant for certain programming tasks with the quality of human-written code. They also examined the effort required to repair bugs in code generated by Copilot using an APR tool, comparing it with effort required to fix bugs in human-written code. Their results highlight some differences in the cost of repairing buggy Copilot-generated code compared to those generated by humans. Their results also suggest that this difference could be due to Copilot occasionally overlooking specific details in the prompt.

Mastropaoletti et al. [65] studied the robustness of LLM-based programming assistant tools and investigated the extent to which the input provided to

---

<sup>7</sup> <https://visualstudio.microsoft.com/services/intellicode/>

Copilot as a prompt affects the generated output. To conduct this study, they employed a Deep Learning based paraphrasing technique to rephrase task descriptions and assessed the correctness of the rephrased descriptions through manual evaluation by the authors. Their results indicate that rephrasing the task description influences the distribution of solutions generated by Copilot across categories of *failing test*, *passing test*, *syntax error*, and *no solution*. Thus, there is a potential loss in accuracy (moving generated code snippets from *passing test* category to *failing test* or *syntax error*) associated with the description of the prompt.

While the findings of these studies shed light on the bug-proneness of code generated by LLMs, none of them have thoroughly investigated the bug patterns and/or the characteristics of such bugs. Only a handful of studies have delved into the types of bugs observed in the code generated by LLMs, and we will discuss them in the following.

Honarvar et al. [66] conducted a study to assess the robustness of four different LLMs in addressing what they termed as instances or neighborhood questions. To create a set of neighborhoods for a template question, they replaced variables in 60 different programming tasks with input tests collected from their test oracle and generated different instances for a single programming question. In the evaluation process, rather than just labeling a code as buggy, they assigned a correctness score to indicate the number of instances of a template question successfully handled by the code generated by an LLM. Their findings revealed that there are buggy code snippets generated by LLMs that failed only in a few instances. Subsequently, they categorized the reasons behind failures based on observed error types into classes such as *syntax error*, *runtime error*, *assertion error*, and *wrong number of arguments*.

Fan et al. [15] aimed to enhance the reliability of code generated by Codex by fixing its generated buggy code. They employed both an APR tool and also leveraged Codex to repair the buggy code. To assess the feasibility of applying APR to repair buggy code from Codex, they initially analyzed common mistakes found in solutions generated by Codex. Two authors manually attempted to fix the bugs in the Codex-generated code, creating fixing patches by referencing other human-provided solutions for the same task on Leetcode [58]. The authors categorized buggy snippets using a predefined set of defect categories, derived from a benchmark study on human buggy code for programming competition [67]. This classification was inspired by the type of fix required for the buggy code, such as *operator mutation* or *variable mutation*. They concluded that the buggy code generated by Codex shared similar mutant operators for fixing as those found in human buggy code. However, their results on repairing buggy code revealed that Codex outperformed APR tools (which are also inspired by human bugs) in fixing its own buggy code, and they recommended enhancements to APR tools to overcome this limitation. Their conclusions are not confirmed by our study. Indeed, while *Misinterpretation* and *Missing Corner Case* bug patterns, which relate more to something human developers would do (as validated by the survey in Section 4.2), are still present, multiple non-human like mistakes occur even on a more advanced model like Codex

(e.g., *Wrong Attribute*, *Hallucinated Object*). This can be explained by the fact that CoderEval is based on actual programs mined from GitHub and not coding competition snippets such as those extracted from LeetCode contests.

Liu et al. [16] conducted an analysis of code quality issues in the code generated by ChatGPT for a specific set of LeetCode problems [58]. They utilized static analysis tools, such as Pylint for Python, to examine common characteristics of the generated code. They applied open card sorting to discuss and group the ChatGPT-generated code based on their results on test cases and the outcome of static analysis tools. After grouping the code snippets they assigned a name to each group based on common error patterns, leading to four different categories of quality issues in ChatGPT-generated code: *compilation and runtime errors*, *wrong output* (linked to assertion errors), *code style and maintainability*, and *performance and efficiency*. Within each category, they counted error types from the messages raised by static analysis tools and reported the distribution of error types. Their results revealed that code snippets passing all test cases exhibited fewer quality issues, although they were not entirely free of such issues. While we did not study quality issues (as we focused on bugs rather than code quality issues), the prevalence of *Misinterpretation* and *Missing Corner Case* categories in our dataset are in line with the prevalence of *Wrong Output* category in their study. Thus, we reached similar observations in more practical tasks (compared to LeetCode competitions which are simpler). Moreover, their *Compilation and Runtime Error* in the LeetCode tasks match one or more of our categories, for instance, the *Wrong Attribute* or *Wrong Input Type*. Many bug categories found in our study were not found in their study, for example, *Hallucinated Object* or *Non-Prompted Consideration*. This difference is likely due to the nature of the studied tasks: CoderEval tasks are more practical (as they were directly extracted from GitHub) than the programming competition tasks on the LeetCode. This gives way to more complex errors such as *Prompt-biased Code* or *Hallucinated Object* as the docstring is less detailed and the task more complex in CoderEval. We believe that our study and presented bug patterns not only confirm their findings but also improve their results by exploring bug patterns in more practical tasks and not limited to the error types reported by static analysis tools.

Jesse et al. [68] examined the proneness of LLMs in generating Simple, Stupid Bugs (SStuBs). Stupid bugs are bugs that require single statement fixes but are challenging to locate. To conduct the study, they used a dataset of SStuBs in Java. They used the code before the error line as a prompt and presented it to the LLM and examined whether the model completed the code by regenerating the bug or generating the fix. The results revealed that the LLMs used in their study produced more SStuBs than the correct code. Additionally, SStuBs generated by Codex took significantly longer to be fixed, underscoring the challenge posed by the prevalence of SStuBs and the potential difficulty in distinguishing between natural and incorrect outputs of LLMs. The study also demonstrated that adding comments to the code can, to some extent, assist the model in generating fewer SStuBs. While the focus

of this study was on Java, not all equivalent SStuBs in Python were observed in our sample set. However, we categorized bugs with similar characteristics as *Missing Corner Cases*, where a small change or a single statement modification can transform the buggy code into correct code.

The study by Pan et al. [17] is focused on analyzing the effectiveness of LLMs in code translation and is the only study that as part of their results developed a systematic bug taxonomy of buggy code generated by LLMs when employed for code translation tasks. The characteristics outlined in their taxonomy are particularly relevant to the challenges associated with translating code from one language to another such as *Mismatch of API behavior between source and target*, *Removal of logic from the source code*, and *Mismatch of behavior after replacing API call*.

To the best of our knowledge, our study is the first to systematically analyze the characteristics of LLM-generated buggy code and construct a bug taxonomy based on the observed bug patterns. Furthermore, we are the first to evaluate the appropriateness of our proposed taxonomy on LLM-generated buggy code through a survey study targeting users of LLMs. In contrast to prior studies that based their analyses on programming tasks collected from platforms like LeetCode, our analysis consists of the buggy code generated by three different LLMs using real-world programming tasks collected from the CoderEval benchmark dataset. We assert that our findings contribute to a more comprehensive understanding of bugs in LLM-generated code.

## 7 Threats to Validity

*Construct validity:* Our methodology may pose a potential threat as the process of collecting and labeling buggy samples could influence our results. While our methodology is similar to that of existing works that categorized bugs [29, 33, 36], we addressed this threat by meticulously describing our approach, to allow for external validations. Another limitation could come from the dataset we used: Python functions and the three examined LLMs. Python is a widely used programming language in various domains, and the examined LLMs have been used in previous studies for code generation tasks [19, 20, 21, 25, 27]. In the absence of a pre-existing taxonomy for categorizing bugs, we recognized the risk of introducing bias into our classification. To mitigate this risk, we followed an open coding procedure. Each rater independently evaluated buggy samples, and conflicts were discussed until a consensus was reached. Additionally, to validate our categories, we surveyed practitioners who employed LLMs for code generation. Finally, removing bugs classified as runnable level above *plib\_runnable* could introduce a loss of information regarding the bug patterns identified. Nonetheless, any bugs in the code generated from the prompt of those categories would not necessarily reflect a weakness of LLM in generating code but just a lack of proper information. As such, we preferred removing them and potentially losing some information rather than adding noises to our study.

*Internal Validity:* One source of internal threats to validity is the potential bias in manual inspection and labeling of buggy samples. To alleviate this threat, the first two authors (two senior Ph.D. candidates who have experience in SE research and programming) labeled buggy code samples independently. After a series of meetings with the third author (a senior researcher with 10 years of research experience in SE and AI), a consensus was reached on the labeling process and the criteria for categorizing different bug patterns. Another threat arises from the sampling method used to select buggy snippets from CoderEval. To address potential sampling bias, we followed similar statistical approaches [69, 70] utilizing 95% confidence levels and 5% confidence intervals. Furthermore, we validated through a survey, following existing studies [29, 33, 36], to verify that our obtained categories are representative of bug patterns faced by practitioners. The response rate, 8.9%, was in line with recent survey studies in SE that received response rates of 8.4% to 13.3% [36, 37, 71]. In particular, we showed in Section 4.2 that there is a correlation between the bug patterns encountered by respondents and the distribution of bug patterns sampled for Codex, which further comforts us that our sampling is representative. Finally, no new bug patterns were reported by survey respondents.

*External Validity:* One external threat to the validity is the LLMs used to generate the code in the dataset. In this study, three LLMs were used, Code-Gen, PanGu-Coder, and Codex: the first two are open-source and have been used in previous studies that harness the power of LLMs for code generation tasks [21, 25, 27, 72], and the third one, Codex, is the model behind Copilot. Another threat is that the tasks for which the LLMs generated code may not be representative of real tasks which could limit the relevance of the bug patterns found. The choice of the CoderEval dataset mitigates this issue, as it is based on real tasks mined from GitHub. Nonetheless, future works should consider expanding our study to cover a more diverse set of LLMs and functions (problem and prompt) to generate code. Another threat to external validity stems from our focus on *Python* code. Since *Python* is one of the most popular languages for development, we believe that our findings are relevant to the majority of programming languages. However, some language-specific bugs can not be detected in Python because of its properties. For instance, type-based variable declaration or memory management issues would not be something we could easily obtain in a Python program. Thus, future works should consider expanding our study to include other programming languages. The last threat is regarding the rapid development of LLMs and the long-term relevance of the proposed taxonomy. While future studies could expand this taxonomy based on the evolving landscape of LLM advancements, we aimed to draw attention to the distinctive nature of bugs in code generated by LLMs compared to human bugs, even in common types like syntax errors. Moreover, we believe the bug patterns remain broad and will not be overshadowed by the advancements in LLMs over time. Categories such as *Prompt-biased* code, *Hallucinated Objects*, or *Wrong Input Type* are linked to the limitations arising from the generative nature of LLMs. While they may see a reduction in frequency with improved model performance, they are likely to persist.

*Conclusion Validity:* The conclusions drawn in this study have certain limitations, notably the potential for wrong/missing bug patterns. We conducted manual inspections of 333 buggy snippets from a pool of over 1,997 in the dataset. Similarly, buggy snippets were generated by 3 different LLMs to ensure the reliability of our findings. In both instances, the sample sizes were deemed substantial enough to be representative and avoid leading to misguided conclusions. Additionally, to minimize potential errors, bug labeling was independently conducted by two raters and subsequently discussed. Moreover, a validation survey was used to compare our results to users' experience and feedback. Lastly, we have provided a replication package [18] to facilitate the reproducibility of our findings, enabling other researchers to expand upon our study.

## 8 Conclusion and Future Work

In this paper, we empirically investigated the bug patterns observed in code generated by LLMs. Our analysis provides insights into the nature, frequency, and distribution of these bug patterns. We organized our examination of the patterns in a taxonomy for buggy code generated by LLMs. These analyses lead to interesting observations on different bug patterns found in LLM-generated code, such as *Prompt-biased Code*, *Wrong Attribute* and *Hallucinated Object*, which are not common in human buggy code. The cause of several bug patterns is associated with the natural language description of the prompts. Our results indicate that the absence of certain details or even the inclusion of an extra newline in the prompt—typically not problematic to humans, can cause ambiguity for LLMs. Consequently, this ambiguity may lead to the generation of buggy code by LLMs, such as *Missing Corner Case* or *Prompt-biased Code* patterns.

In addition, certain bug patterns originate from the stochastic generative nature of LLMs such as *Hallucinated Object* or *Wrong Attribute* which are also not common in human-written buggy code. While the restricted information of the prompt might explain the existence of *Misinterpretation* and *Missing Corner Case* bug patterns, there is still a non-trivial number of bug patterns (64% of our sample set) such as *Wrong Attribute* or *Hallucinated Object*, which should not be linked to a lack of information in the prompt. For instance, in Listing 7, LLM should not apply the ‘min’ function on a list of classes, irrespective of whether or not the docstring describes the expected function enough, as it is simply not applicable to this situation.

We evaluated the relevance of our findings by conducting an online survey targeting software practitioners and researchers involved in LLMs. The respondents of the validation survey mostly encountered the bug patterns we observed. They assessed that, while the majority of the bug patterns can be relatively easy to diagnose and fix, several categories such as *Misinterpretation* or *Non-Prompted Consideration* are difficult to diagnose and fix. Moreover, the

patterns of bugs in LLMs generated code are different from the bug patterns of human developers even in categories such as *Syntax Error*.

Our results serve as an initial taxonomy of bug patterns observed in LLM-generated code. While further investigation is required to compare their appearance in human-written code and other LLMs, the findings in this study combined with the validation survey represent the first steps in characterizing bugs in LLM-generated code. This opens up avenues for future studies on enhancing the quality of code generated by LLM-based programming assistant tools. Our findings also can be leveraged to improve different SE tools that rely on characteristics of human buggy code, such as mutant operators in MT or repair rules in APR tools, as well as potential development in Computer Science (CS) education studies.

## Acknowledgment

We would like to thank the participants who contributed to the survey questionnaire.

## Data Availability Statement

The dataset generated during the current study is available in the replication package, which is accessible via [18].

## References

1. D. Wong, A. Kothig, and P. Lam, “Exploring the verifiability of code generated by github copilot,” *arXiv preprint arXiv:2209.01766*, 2022.
2. S. Imai, “Is github copilot a substitute for human pair-programming? an empirical study,” in *Proceedings of the ACM/IEEE 44th International Conference on Software Engineering: Companion Proceedings*, 2022, pp. 319–321.
3. F. A. Sakib, S. H. Khan, and A. Karim, “Extending the frontier of chatgpt: Code generation and debugging,” *arXiv preprint arXiv:2307.08260*, 2023.
4. M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman *et al.*, “Evaluating large language models trained on code,” *arXiv preprint arXiv:2107.03374*, 2021.
5. H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” *arXiv preprint arXiv:2307.09288*, 2023.
6. H. Yu, B. Shen, D. Ran, J. Zhang, Q. Zhang, Y. Ma, G. Liang, Y. Li, T. Xie, and Q. Wang, “Codereval: A benchmark of pragmatic

- code generation with generative pre-trained models,” *arXiv preprint arXiv:2302.00288v1*, 2023.
7. M. Jin, S. Shahriar, M. Tufano, X. Shi, S. Lu, N. Sundaresan, and A. Svyatkovskiy, “Inferfix: End-to-end program repair with llms,” in *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ser. ESEC/FSE 2023. New York, NY, USA: Association for Computing Machinery, 2023, p. 1646–1656. [Online]. Available: <https://doi.org/10.1145/3611643.3613892>
  8. J. Liu, C. S. Xia, Y. Wang, and L. Zhang, “Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation,” *arXiv preprint arXiv:2305.01210*, 2023.
  9. A. Moradi Dakhel, V. Majdinasab, A. Nikanjam, F. Khomh, M. C. Desmarais, and Z. M. J. Jiang, “Github copilot ai pair programmer: Asset or liability?” *Journal of Systems and Software*, vol. 203, p. 111734, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0164121223001292>
  10. C. Bird, D. Ford, T. Zimmermann, N. Forsgren, E. Kalliamvakou, T. Lowdermilk, and I. Gazit, “Taking flight with copilot: Early insights and opportunities of ai-powered pair-programming tools,” *Queue*, vol. 20, no. 6, pp. 35–57, 2022.
  11. O. Asare, M. Nagappan, and N. Asokan, “Is github’s copilot as bad as humans at introducing vulnerabilities in code?” *Empirical Softw. Engg.*, vol. 28, no. 6, sep 2023. [Online]. Available: <https://doi.org/10.1007/s10664-023-10380-1>
  12. J. Li, Y. Zhang, Z. Karas, C. McMillan, K. Leach, and Y. Huang, “Do machines and humans focus on similar code? exploring explainability of large language models in code summarization,” *arXiv preprint arXiv:2402.14182*, 2024.
  13. B. Kou, S. Chen, Z. Wang, L. Ma, and T. Zhang, “Is model attention aligned with human attention? an empirical study on large language models for code generation,” 2023.
  14. R. Just, D. Jalali, L. Inozemtseva, M. D. Ernst, R. Holmes, and G. Fraser, “Are mutants a valid substitute for real faults in software testing?” in *Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering*, ser. FSE 2014. New York, NY, USA: Association for Computing Machinery, 2014, p. 654–665. [Online]. Available: <https://doi.org/10.1145/2635868.2635929>
  15. Z. Fan, X. Gao, M. Mirchev, A. Roychoudhury, and S. H. Tan, “Automated repair of programs from large language models,” in *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 2023, pp. 1469–1481.
  16. Y. Liu, T. Le-Cong, R. Widyasari, C. Tantithamthavorn, L. Li, X.-B. D. Le, and D. Lo, “Refining chatgpt-generated code: Characterizing and mitigating code quality issues,” 2023.

17. R. Pan, A. R. Ibrahimzada, R. Krishna, D. Sankar, L. P. Wassi, M. Merler, B. Sobolev, R. Pavuluri, S. Sinha, and R. Jabbarvand, “Lost in translation: A study of bugs introduced by large language models while translating code,” *arXiv preprint arXiv:2308.03109v3*, 2023.
18. “The replication package,” <https://github.com/FlowSs/BugsInLLMs>, 2023.
19. N. Nguyen and S. Nadi, “An empirical evaluation of github copilot’s code suggestions,” in *Proceedings of the 19th International Conference on Mining Software Repositories*, 2022, pp. 1–5.
20. C. Lemieux, J. P. Inala, S. K. Lahiri, and S. Sen, “Codamosa: Escaping coverage plateaus in test generation with pre-trained large language models,” in *Accepted by 45th International Conference on Software Engineering (ICSE)*, 2023.
21. X. Chen, M. Lin, N. Schärli, and D. Zhou, “Teaching large language models to self-debug,” *arXiv preprint arXiv:2304.05128*, 2023.
22. T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
23. F. Christopoulou, G. Lampouras, M. Gritta, G. Zhang, Y. Guo, Z. Li, Q. Zhang, M. Xiao, B. Shen, L. Li *et al.*, “Pangu-coder: Program synthesis with function-level language modeling,” *arXiv preprint arXiv:2207.11280*, 2022.
24. W. Zeng, X. Ren, T. Su, H. Wang, Y. Liao, Z. Wang, X. Jiang, Z. Yang, K. Wang, X. Zhang *et al.*, “Pangu- $\alpha$ : Large-scale autoregressive pretrained chinese language models with auto-parallel computation,” *arXiv preprint arXiv:2104.12369*, 2021.
25. D. Zan, B. Chen, F. Zhang, D. Lu, B. Wu, B. Guan, Y. Wang, and J.-G. Lou, “When neural model meets nl2code: A survey,” *arXiv preprint arXiv:2212.09420*, 2022.
26. E. Nijkamp, B. Pang, H. Hayashi, L. Tu, H. Wang, Y. Zhou, S. Savarese, and C. Xiong, “Codegen: An open large language model for code with multi-turn program synthesis,” *arXiv preprint arXiv:2203.13474*, 2022.
27. B. Chen, F. Zhang, A. Nguyen, D. Zan, Z. Lin, J.-G. Lou, and W. Chen, “Codet: Code generation with generated tests,” *arXiv preprint arXiv:2207.10397*, 2022.
28. C. B. Seaman, “Qualitative methods in empirical studies of software engineering,” *IEEE Transactions on software engineering*, vol. 25, no. 4, pp. 557–572, 1999.
29. N. Humbatova, G. Jahangirova, G. Bavota, V. Riccio, A. Stocco, and P. Tonella, “Taxonomy of real faults in deep learning systems,” in *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, 2020, pp. 1110–1121.
30. M. X. Liu, A. Sarkar, C. Negreanu, B. Zorn, J. Williams, N. Toronto, and A. D. Gordon, ““what it wants me to say”: Bridging the abstraction gap between end-user programmers and code-generating large language

- models,” in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–31.
31. M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. de Oliveira Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. P. Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever, and W. Zaremba, “Evaluating large language models trained on code,” 2021.
  32. J. Austin, A. Odena, M. Nye, M. Bosma, H. Michalewski, D. Dohan, E. Jiang, C. Cai, M. Terry, Q. Le, and C. Sutton, “Program synthesis with large language models,” 2021.
  33. F. Tambon, A. Nikanjam, L. An, F. Khomh, and G. Antoniol, “Silent bugs in deep learning frameworks: an empirical study of keras and tensorflow,” *Empirical Software Engineering*, vol. 29, no. 1, p. 10, 2024.
  34. Y. Fu, P. Liang, A. Tahir, Z. Li, M. Shahin, and J. Yu, “Security weaknesses of copilot generated code in github,” *arXiv preprint arXiv:2310.02059*, 2023.
  35. D. Spadini, M. Aniche, and A. Bacchelli, “Pydriller: Python framework for mining software repositories,” in *Proceedings of the 2018 26th ACM Joint meeting on european software engineering conference and symposium on the foundations of software engineering*, 2018, pp. 908–911.
  36. A. Nikanjam, M. M. Morovati, F. Khomh, and H. Ben Braiek, “Faults in deep reinforcement learning programs: a taxonomy and a detection approach,” *Automated Software Engineering*, vol. 29, 2021.
  37. V. Nardone, B. Muse, M. Abidi, F. Khomh, and M. Di Penta, “Video game bad smells: What they are and how developers perceive them,” *ACM Trans. Softw. Eng. Methodol.*, vol. 32, no. 4, may 2023. [Online]. Available: <https://doi.org/10.1145/3563214>
  38. A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann *et al.*, “Palm: Scaling language modeling with pathways,” *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.
  39. Anthropic, “Model card and evaluations for claude models,” 2013. [Online]. Available: <https://www.anthropic.com/index/introducing-claude>
  40. “gitindex,” 2023, <https://github.info/>.
  41. “tiobeindex,” 2023, <https://www.tiobe.com/tiobe-index/>.
  42. M. Schäfer, S. Nadi, A. Eghbali, and F. Tip, “Adaptive test generation using a large language model,” *arXiv preprint arXiv:2302.06527*, 2023.
  43. J. Wei, X. Wang, D. Schuurmans, M. Bosma, b.ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou, “Chain-of-thought prompting elicits reasoning in large language models,” in *Advances in Neural Information Processing*

- Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 24 824–24 837. [Online]. Available: [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf)
- 44. J. Li, G. Li, Y. Li, and Z. Jin, “Structured chain-of-thought prompting for code generation,” *arXiv preprint arXiv:2305.06599*, 2023.
  - 45. C. Liu, S. D. Zhang, and R. Jabbarvand, “Codemind: A framework to challenge large language models for code reasoning,” 2024.
  - 46. V. Guilherme and A. Vincenzi, “An initial investigation of chatgpt unit test generation capability,” in *Proceedings of the 8th Brazilian Symposium on Systematic and Automated Software Testing*, ser. SAST ’23. New York, NY, USA: Association for Computing Machinery, 2023, p. 15–24. [Online]. Available: <https://doi.org/10.1145/3624032.3624035>
  - 47. G. Fraser and A. Arcuri, “A large-scale evaluation of automated unit test generation using evosuite,” *ACM Trans. Softw. Eng. Methodol.*, vol. 24, no. 2, dec 2014. [Online]. Available: <https://doi.org/10.1145/2685612>
  - 48. Y. Tang, Z. Liu, Z. Zhou, and X. Luo, “Chatgpt vs sbst: A comparative assessment of unit test suite generation,” 2023.
  - 49. G. L. Scoccia, “Exploring early adopters’ perceptions of chatgpt as a code generation tool,” in *2023 38th IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW)*, 2023, pp. 88–93.
  - 50. M. Kazemitabaar, X. Hou, A. Henley, B. J. Ericson, D. Weintrop, and T. Grossman, “How novices use llm-based code generators to solve cs1 coding tasks in a self-paced learning environment,” *arXiv preprint arXiv:2309.14049*, 2023.
  - 51. F. Fischer, K. Böttinger, H. Xiao, C. Stransky, Y. Acar, M. Backes, and S. Fahl, “Stack overflow considered harmful? the impact of copy&paste on android application security,” in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 121–136.
  - 52. T. Zhang, G. Upadhyaya, A. Reinhardt, H. Rajan, and M. Kim, “Are code examples on an online q&a forum reliable?: A study of api misuse on stack overflow,” in *2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE)*, 2018, pp. 886–896.
  - 53. M. Verdi, A. Sami, J. Akhondali, F. Khomh, G. Uddin, and A. K. Motlagh, “An empirical study of c++ vulnerabilities in crowd-sourced code examples,” *IEEE Trans. Softw. Eng.*, vol. 48, no. 5, p. 1497–1514, may 2022. [Online]. Available: <https://doi.org/10.1109/TSE.2020.3023664>
  - 54. P. Vaithilingam, T. Zhang, and E. L. Glassman, “Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models,” in *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 2022, pp. 1–7.
  - 55. M. Monperrus, “Automatic software repair: A bibliography,” *ACM Comput. Surv.*, vol. 51, no. 1, jan 2018. [Online]. Available: <https://doi.org/10.1145/3105906>
  - 56. Z. Li, C. Wang, Z. Liu, H. Wang, D. Chen, S. Wang, and C. Gao, “Cctest: Testing and repairing code completion systems,” in *2023 IEEE/ACM*

- 45th International Conference on Software Engineering (ICSE), 2023, pp. 1238–1250.
- 57. S. Moon, Y. Song, H. Chae, D. Kang, T. Kwon, K. T. iunn Ong, S. won Hwang, and J. Yeo, “Coffee: Boost your code llms by fixing bugs with feedback,” 2023.
  - 58. “Leetcode contest,” <https://leetcode.com/contest/>, 2023.
  - 59. S. Lu, D. Guo, S. Ren, J. Huang, A. Svyatkovskiy, A. Blanco, C. Clement, D. Drain, D. Jiang, D. Tang *et al.*, “Codexglue: A machine learning benchmark dataset for code understanding and generation,” *arXiv preprint arXiv:2102.04664*, 2021.
  - 60. N. Muennighoff, Q. Liu, A. Zebaze, Q. Zheng, B. Hui, T. Y. Zhuo, S. Singh, X. Tang, L. von Werra, and S. Longpre, “Octopack: Instruction tuning code large language models,” *arXiv preprint arXiv:2308.07124*, 2023.
  - 61. X. Du, M. Liu, K. Wang, H. Wang, J. Liu, Y. Chen, J. Feng, C. Sha, X. Peng, and Y. Lou, “Classeval: A manually-crafted benchmark for evaluating llms on class-level code generation,” 2023.
  - 62. D. Huang, J. M. Zhang, Y. Qing, and H. Cui, “Effibench: Benchmarking the efficiency of automatically generated code,” 2024.
  - 63. Y. Liu, C. Tantithamthavorn, Y. Liu, and L. Li, “On the reliability and explainability of automated code generation approaches,” *arXiv preprint arXiv:2302.09587*, 2023.
  - 64. Z. Ji, P. Ma, Z. Li, and S. Wang, “Benchmarking and explaining large language model-based code generation: A causality-centric approach,” 2023.
  - 65. A. Mastropaoletti, L. Pasarella, E. Guglielmi, M. Ciniselli, S. Scalabrino, R. Oliveto, and G. Bavota, “On the robustness of code generation techniques: An empirical study on github copilot,” *arXiv preprint arXiv:2302.00438*, 2023.
  - 66. S. Honarvar, M. van der Wilk, and A. Donaldson, “Turbulence: Systematically and automatically testing instruction-tuned large language models for code,” *arXiv preprint arXiv:2312.14856*, 2023.
  - 67. S. H. Tan, J. Yi, S. Mechtaev, A. Roychoudhury *et al.*, “Codeflaws: a programming competition benchmark for evaluating automated program repair tools,” in *2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C)*. IEEE, 2017, pp. 180–182.
  - 68. K. Jesse, T. Ahmed, P. T. Devanbu, and E. Morgan, “Large language models and simple, stupid bugs,” in *2023 IEEE/ACM 20th International Conference on Mining Software Repositories (MSR)*. Los Alamitos, CA, USA: IEEE Computer Society, may 2023, pp. 563–575. [Online]. Available: <https://doi.ieeecomputersociety.org/10.1109/MSR59073.2023.00082>
  - 69. T. Zhang, C. Gao, L. Ma, M. Lyu, and M. Kim, “An empirical study of common challenges in developing deep learning applications,” in *2019 IEEE 30th International Symposium on Software Reliability Engineering (ISSRE)*, 2019, pp. 104–115.
  - 70. M. M. Morovati, A. Nikanjam, F. Tambon, F. Khomh, and Z. M. Jiang, “Bug characterization in machine learning-based systems,” *Empirical Software Engineering*, vol. 29, no. 1, p. 14, 2024.

71. A. Cater-Steel, M. Toleman, and T. Rout, “Addressing the challenges of replications of surveys in software engineering research,” in *2005 International Symposium on Empirical Software Engineering, 2005*. IEEE, 2005, pp. 10–pp.
72. J. Bogatinovski and O. Kao, “Auto-logging: AI-centred logging instrumentation,” in *2023 IEEE/ACM 45th International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER)*. IEEE, 2023, pp. 95–100.
73. “Google forms,” <https://www.google.ca/forms/about/>, 2022.
74. A. N. Oppenheim, *Questionnaire design, interviewing and attitude measurement*. Bloomsbury Publishing, 2000.

## Appendix

### A Reddit Post

We posted our survey questionnaire on two relevant Reddit channels: *LocalLLaMA* and *MachineLearning*. Figure 5 shows the posted survey on Reddit.

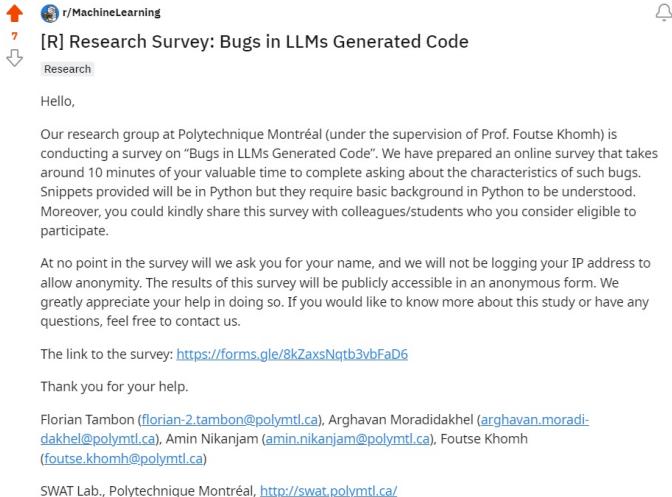


Fig. 5: The posted survey on Reddit.

### B Survey

#### B.1 The details of the questionnaire in the survey

To prepare the survey, we used Google Forms [73], a well-known online tool for creating and sharing surveys and quizzes. The survey has two parts inspired by previous empirical

**Bugs in Large Language Models Generated Code**

\* Indique une question obligatoire

**General LLM Usage**

This part is about general questions on the LLM usage and the potential bugs found in LLM generated code.

What is your current job title? (Developer, Researcher, ... students may indicate degree: PhD, Master) \*

Votre réponse

Which LLMs do you mostly use? Select as many as you use. You can enter one manually if it's not on the list. \*

Codex (Copilot)  
 ChatGPT  
 Llama  
 CodeGen  
 Autre : \_\_\_\_\_

Fig. 6: Survey received by the participant

studies on bug patterns [29, 33, 36]. In the first part, we asked demographic questions to participants i.e. what is your current job title, what LLMs they have used, and for what programming languages they have used LLMs to generate code, other than Python (since it is the programming language of the given code snippets), as presented in Figure 6. Then, we asked for more specific questions about bugs in LLMs generated code, more precisely, we asked:

- How often do you encounter mistakes in LLMs generated code you use? Those mistakes can be anything, from code that will not compile (variable not defined, function that does not exist) to more silly choices from the LLMs (multiple If condition checking the same thing or casting back and forth a variable for no reason).
- When you encounter mistakes in LLMs generated code, how complex is fixing those issues?
- To fix those issues, how do you proceed?

For the first three questions, we used a 5-level Likert scale [74] (Never to Always or Trivial to Very Hard). For, the third one, the respondents could choose between “Manually”, “Re-prompts the LLM”, “Combination of both” or “Other” (and they were given the possibility to write down how).

In the second part of the survey, we wanted to consolidate our bug patterns and gather the participants’ feedback on several aspects of each pattern of bug in our study. An example

**Hallucinated objects:** \*

LLMs hallucination such as using a function or built-in type that does not exist and was not defined previously but that the LLMs consider present within the context.

In the example below, the LLM (besides not doing what the docstring asks), uses a function named "find\_path\_to\_glob" which is not defined in the context.

```
...
Given a sequence of path fragments or patterns as passed to '--find', transform all path
fragments into glob patterns. Pass through existing patterns untouched.

For example, given find_paths of:

['foo.txt', 'pp:root/somedir']

... transform that into:

[sh:/**/foo.txt/**, 'pp:root/somedir']
...
def make_find_paths(find_paths):
    return [find_path_to_glob(x) for x in find_paths]
```

How often do you encounter such a type of mistake?

|       |                       |                       |                       |                       |                              |
|-------|-----------------------|-----------------------|-----------------------|-----------------------|------------------------------|
| 1     | 2                     | 3                     | 4                     | 5                     |                              |
| Never | <input type="radio"/> Always |

How do you assess diagnosing, severity, and fixing of such a bug? \*

If you never experienced such bug, please answer based on your understanding using the previous snippet as an example for the bug type.

|            | 1 (Easy/<br>Trivial)  | 2                     | 3                     | 4                     | 5 (Hard/<br>Complex)  |
|------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| Diagnosing | <input type="radio"/> |
| Complexity | <input type="radio"/> |
| Fixing     | <input type="radio"/> |

Fig. 7: Example of questions for a bug pattern in the survey. The description of the task is collected from the CoderEval dataset [6].

of questions in the second part is shown in Figure 7. The structure of the questions is similar for each pattern of bugs in the taxonomy.

## B.2 Survey Results

Table 4: Results of validating survey for the bug patterns. Questions related to the frequency of encounter of bug patterns, the difficulty to diagnose and fix them as well as the complexity of the bug. We highlight in bold the highest number in each category for each bug pattern. 1 represents never/easy/trivial/low and 5 represents always/hard/complex/high.

| Types of Bug          | Responses  |   |   |  |  |
|-----------------------|--|---|---|--|--|
|                       | Frequency<br>1 — 2 — 3 — 4 — 5                     | Diagnosing<br>1 — 2 — 3 — 4 — 5             | Complexity<br>1 — 2 — 3 — 4 — 5             | Fixing<br>1 — 2 — 3 — 4 — 5                        |  |
| Hallucinated object   | 17.6% - <b>26.5%</b> - 23.5% - 23.5% - 8.8%        | <b>79.4%</b> - 17.6% - 0% - 0% - 2.9%       | <b>47.1%</b> - 32.4% - 14.7% - 5.9% - 0%    | 17.6% - 26.5% - <b>29.4%</b> - 20.6% - 5.9%        |  |
| Incomplete Generation | 20.6% - 11.8% - 29.4% - <b>32.4%</b> - 5.9%        | <b>73.5%</b> - 14.7% - 2.9% - 8.8% - 0%     | <b>41.2%</b> - 38.2% - 8.8% - 11.8% - 0%    | 26.5% - 20.6% - 11.8% - <b>29.4%</b> - 11.8%       |  |
| Misinterpretation     | 5.9% - <b>35.3%</b> - 26.5% - 26.5% - 5.9%         | 29.4% - 17.6% - <b>32.4%</b> - 17.6% - 2.9% | 14.7% - 23.5% - 26.5% - <b>32.4%</b> - 2.9% | 5.9% - 11.8% - <b>38.2%</b> - 35.3% - 8.8%         |  |
| NPC                   | 17.6% - 23.5% - <b>35.3%</b> - 20.6% - 2.9%        | 23.5% - 23.5% - <b>26.5%</b> - 23.5% - 2.9% | 17.6% - <b>32.4%</b> - 29.4% - 14.7% - 5.9% | <b>29.4%</b> - <b>29.4%</b> - 23.5% - 11.8% - 5.9% |  |
| Silly Mistake         | <b>29.4%</b> - <b>29.4%</b> - 20.6% - 17.6% - 2.9% | <b>58.8%</b> - 26.5% - 14.7% - 0% - 0%      | <b>67.6%</b> - 29.4% - 2.9% - 0% - 0%       | <b>55.9%</b> - 26.5% - 8.8% - 8.8% - 0%            |  |
| Wrong Input Type      | 17.6% - <b>38.2%</b> - 14.7% - 20.6% - 8.8%        | <b>35.3%</b> - 23.5% - 20.6% - 17.6% - 2.9% | 17.6% - <b>50%</b> - 14.7% - 17.6% - 0%     | 26.5% - <b>38.2%</b> - 17.6% - 14.7% - 2.9%        |  |
| Missing Corner Cases  | 11.8% - 17.6% - 20.6% - <b>35.3%</b> - 14.7%       | 5.9% - 14.7% - 26.5% - <b>44.1%</b> - 8.8%  | 14.7% - 14.7% - <b>38.2%</b> - 23.5% - 8.8% | 17.6% - <b>29.4%</b> - 17.6% - <b>29.4%</b> - 5.9% |  |
| Wrong Attribute       | 0% - 17.6% - 26.5% - <b>47.1%</b> - 8.8%           | 20.6% - <b>44.1%</b> - 26.5% - 5.9% - 2.9%  | 17.6% - <b>50%</b> - 17.6% - 11.8% - 2.9%   | 8.8% - <b>38.2%</b> - 35.3% - 11.8% - 5.9%         |  |
| Syntax Error          | <b>47.1%</b> - 20.6% - 20.6% - 5.9% - 5.9%         | <b>85.3%</b> - 8.8% - 5.9% - 0% - 0%        | <b>82.4%</b> - 8.8% - 8.8% - 0% - 0%        | <b>70.6%</b> - 14.7% - 2.9% - 11.8% - 0%           |  |
| Prompt-biased Code    | 11.8% - 23.5% - 11.8% - <b>44.1%</b> - 8.8%        | <b>50.0%</b> - 26.5% - 17.6% - 5.9% - 0%    | 11.8% - <b>47.1%</b> - 26.5% - 11.8% - 2.9% | 5.9% - 23.5% - 26.5% - <b>29.4%</b> - 14.7%        |  |