

Who is Smarter? An Empirical Study of AI-based Smart Contract Creation

Rabimba Karanjai*, Edward Li*, Lei Xu[†], Weidong Shi*

*University Of Houston, TX, USA

{rkaranjai, wshi3}@uh.edu, edward.w.li@outlook.com

[†]Kent State University, OH, USA

xuleimath@gmail.com

Abstract—The introduction of large language models (LLMs) like ChatGPT and Google Palm2 for smart contract generation seems to be the first well-established instance of an AI pair programmer. LLMs have access to a large number of open-source smart contracts, enabling them to utilize more extensive code in Solidity than other code generation tools. Although the initial and informal assessments of LLMs for smart contract generation are promising, a systematic evaluation is needed to explore the limits and benefits of these models. The main objective of this study is to assess the quality of generated code provided by LLMs for smart contracts. We also aim to evaluate the impact of the quality and variety of input parameters fed to LLMs. To achieve this aim, we created an experimental setup for evaluating the generated code in terms of validity, correctness, and efficiency. Our study finds crucial evidence of security bugs getting introduced in the generated smart contracts as well as the overall quality and correctness of the code getting impacted. However, we also identified the areas where it can be improved. The paper also proposes several potential research directions to improve the process, quality and safety of generated smart contract codes.

Index Terms—GPT, smart contract, code generation, large language models, AI

I. INTRODUCTION

In recent years, large language models (LLMs) such as generative pre-trained transformer (GPT) models [1] and Google PaLM2 [2] are becoming more mature and ready for deployment for real applications [1], [3]. Among all the applications enabled by LLMs, machine-aided source code generation from natural language is fascinating and widely believed to have the potential to revolutionize the way humans create programs. For instance, ChatGPT and its derivatives (e.g., Github CoPilot [4]) have been used to generate source code of multiple programming languages based on natural language inputs from the programmer. But there have also been concerns about these being used for malicious purposes [5].

One of the major concerns of LLM-generated source code is its safety because there is a lack of rigorous understanding of the logic that the source code is generated. Many works have been done on analyzing the security features of the generated code for languages such as Python, Golang, and Javascript. In most cases, LLMs rely on the training dataset to “understand” natural language instructions to create source code and are programming language specific, i.e., they can only create source code of programming languages that have been used to train the model. Accordingly, the security analysis works are also language specific. And most security assessment results on one programming language may not be valid for another language.

Recently, decentralized ledgers and smart contracts running atop it are finding more applications, especially in the financial sector. For instance, the Uniswap smart contracts process about \$7.17 billion per day in 2021 [6]. Because of the popularity of smart contracts and their importance in various scenarios like Confidential Computing [7], [8], Decentralized Serverless Functions [9], Event based Transactions [10], [11], it is a natural question to ask whether LLMs can do a good

job at creating smart contracts based on user instructions and how secure the generated smart contracts are. To fill the gap, we investigate two popular LLMs, ChatGPT and Google PaLM2 [2], for generating smart contract code in Solidity, which is a mainstream smart contract programming language. We compare the code generation process and the generated smart contracts, and conduct systematic evaluation on the contracts. To systematically evaluate GPT 3.5 and PaLM2, we propose to construct an experimental setup to assess the generated code in terms of validity, correctness, and efficiency. In this context, we defined the following research questions:

- **RQ1** What is the quality of the code generated by LLMs like ChatGPT and Google Palm2?
 - **RQ1.1** Code Validity
 - **RQ1.2** Code Correctness
- **RQ2** Can better code be generated using different prompting techniques
- **RQ3** Are the generated codes secure?

In summary, the main contributions of the paper are: (i) We have explored and evaluated the feasibility and quality of the models for creating Solidity smart contract code; (ii) We tested and compared two popular source code generation tools, ChatGPT and Google PALM2 for smart contract code generation; (iii) We systematically assessed and evaluated the contract code generated by the GPT models and discussed insights obtained from the experiments; and (iv) In addition, we highlight possible directions for further research.

II. BACKGROUND

Assessing the quality of the software and identifying bugs as well as security vulnerabilities before deployment can prevent potential harm and unexpected expenses. Nevertheless, automatically detecting security-critical bugs in code is a difficult task in reality. This also applies to code that is generated by the AI models, particularly due to the complex and opaque nature of such models (treated as a black box). When it comes to Solidity, it is more important to investigate both code correctness, any kind of bugs, and performance-related issues like bytecode size and gas fee cost that any AI-based automated systems can create.

A. Ethereum, Smart Contract and Its Security

Ethereum is a widely used public, decentralized ledger that keeps immutable and transparent records that can be programmed on the ledger. These programmable records are known as smart contracts. Smart contracts enable the sharing and execution of program logic among multiple parties. They are typically written in Solidity, which is a programming language specifically designed for smart contracts. Solidity follows an object-oriented approach, uses static typing, and compiles to bytecode that can be executed on Ethereum Virtual

Machine (EVM). As Solidity compilers change rapidly, the latest version may not be backward-compatible with all smart contracts. In this case, the acceptable compiler versions are between 0.8.6 and 0.9.0.

The Solidity language supports various built-in types, including `uint`, `byte`, and `string`. It also has special types such as addresses. The address is a unique global identifier for all EOAs (Externally Owned Accounts) and smart contracts. All Ether transfers and contract executions must target a specific address. Solidity also has a built-in key-value dictionary called "mapping".

According to the security analysis of Ethereum [12], the majority of vulnerabilities on the platform are found in the application layer, particularly in smart contracts where they should be avoided. Smart contract vulnerabilities can range from common issues such as integer overflow to more specific problems related to the blockchain, such as re-entrancy. Despite being challenging to exploit, blockchain-specific vulnerabilities have resulted in severe attacks such as the well-known TheDAO [13] attack, largely because they were previously unknown.

B. Tools for Evaluating Security Issues

Several security testing techniques can be utilized to effectively discover software vulnerabilities and avoid bugs during the operation of a deployed system [14]–[16]. To accomplish this objective, these techniques aim to identify various programming errors, inadequate coding style, deprecated functionalities, or possible memory safety violations (such as unauthorized access to unsafe memory that can be exploited after deployment or outdated cryptographic schemes that are shown to be insecure [17]–[19]). Broadly, current methods for assessing the security of software can be divided into two categories: static [14], [20] and dynamic analysis [21], [22]. Static analysis examines the code of a given program to detect potential vulnerabilities, while the latter approach executes the code. For instance, fuzz testing (fuzzing) produces random program executions to trigger bugs in the program.

For our work, we have taken the approach of using static analysis tools to check for bugs in the GPT model-generated contract code. The detected bugs have been categorized into Low, Medium, and High severity categories based on their security implications of them. We have used a mix of open-source and commercial static analysis tools as described in Section IV-B1.

C. Large Language Models and Prompts

The natural language processing field has been advanced by LLMs, which have demonstrated significant progress in tasks such as question answering, translation, and reading comprehension [1], [23]. This progress has been made possible by increasing the model size from hundreds of millions [24] to hundreds of billions [1], using self-supervised objective functions, and utilizing large corpora of text training data. These models are often developed by large corporations and made available as pre-trained models. Brown et al. [1] have demonstrated that these models can be applied to various tasks with just a few input examples, without needing to adjust the model parameters. Users can provide a few-shot prompt template to guide the models toward generating the desired output for a particular task.

D. LLM and Code Generation

There is a growing interest in using large language models to understand and generate source code [25]–[27]. Feng et al. [28] and Guo et al. [29] have proposed encoder-only models with a modified objective function, focusing on code classification, code retrieval, and program repair [28], [29]. Ahmad et al. [30] and Wang et al. [27] have

TABLE I: Comparing Relevant Code Generation Tools

Features	ChatGPT	Google Plam2
IDE Support	No IDE Support	No IDE Support
First Release Time	Nov 30, 2022	May 10, 2023
Developer	OpenAI	Google
Providing References to Suggestions	NO	YES
Explanation of Suggestions	YES	YES
Providing Multiple Suggestions	NO	YES
Training Data Source	GitHub Repositories	"Google licensed code"
Multipurpose (other than programming)	YES	YES
Subscription	ChatGPT Free ChatGPT Plus (\$20 per month)	Insider Beta Access
Can be Used Offline?	NO	NO
Can it Access Local Files?	NO	YES

used an encoder-decoder architecture to address code-to-code and code-to-text generation tasks, such as program translation, program repair, and code summarization. Recently, decoder-only models have shown promise in generating programs in a left-to-right sequence [25], [31]. These models can be used for zero-shot and few-shots program generation tasks, such as code completion, code infilling, and text-to-code conversion [25], [26], [31].

III. CODE GENERATION IN SMART CONTRACT

Prompt engineering for creating issue-free Solidity smart contracts poses several challenges. One of the main challenges is the need for a deep understanding of the Solidity programming model and its associated issues including security risks. Solidity has its unique features and syntax, making it different from other programming languages, and thus requires specific expertise and insight to develop high-quality code. Additionally, Solidity smart contracts can suffer from various vulnerabilities, such as integer overflow/underflow, re-entrancy, and logic errors, among others. Another challenge is the difficulty in defining clear and concise prompts that accurately capture the intended functionality of the smart contract. The lack of clarity in prompts may lead to misunderstandings, which can result in incorrect implementation, inefficiency, and subsequent vulnerabilities. Addressing these challenges will require expertise in both Solidity programming and prompt engineering to ensure that the resulting smart contracts are issue-free, acceptable quality, and secure.

For the scope of the paper, we take a guided approach to generating prompts which give us a better performance as described in Section V-A.

IV. METHODOLOGY

In this section, we present our experimental setup. We commence by elaborating on the details of the experimental framework. After that, we present how we assess the effectiveness and scalability of our proposed method. We also explore the transferability of the prompts generated by our approach among various models. Additionally, we demonstrate that our approach can detect security vulnerabilities in the code generated by the GPT models.

A. Dataset

For our experiment, we could not use any of the existing datasets like HumanEval [32] since they primarily concentrate on Python as a programming language. To evaluate we had to create our dataset. To evaluate the correctness of our code we carefully curated a set of source files from GitHub under open source permissive license.

We leveraged GitHub APIs to construct our scraper that collects Solidity code from different projects. There were several factors we took into account when collecting the corpus:

- The project has to have 50 stars on GitHub. This would ensure the project has enough interest, is a fairly popular project, and has a higher chance of having human audience code
- The projects should have "Solidity" as a programming tag available for further filtering
- They have a sufficient amount of comments present in them. This helps us design and develop prompts in scale and create the experimentation pipeline.

Once these projects were collected, they are further processed to exclude any files which were not Solidity. A brief overview of the files we have collected is given below.

Raw code corpus collection. GitHub serves as an excellent resource for accessing a wide range of publicly accessible source code written in various programming languages. In our process, we replicated the most widely recognized repositories that were labeled with the Solidity language tag and garnered at least 50 stars. From each project, we extracted all the files that were authored in Solidity, thereby generating the primary training dataset.

Data pre-processing. The detailed data pre-processing strategy comparison with other models is analyzed in TABLE I.

All our available codes were passed through a parser that would strip out the functions from the Solidity files, **however keeping the pragma, variables and other code information intact. Including comments.** These processed files are called *Code templates* in the remainder of the paper.

B. Code Analysis

We analyze the generated codes based on our predefined parameters. To answer our different RQ's we have carefully designed an experimental workflow defined in Fig. 2.

Once the code generation part finishes, we pass it through the compilation stage and different code quality analysis steps. Once they are done we also do static analysis on them.

1) *Tools for Analyzing Smart Contracts:* This section introduces the tools we used for our smart contract analysis.

Mythril [33], which is a Python-based command-line tool. It conducts an interactive analysis of EVM bytecode and presents the Control Flow Graph (CFG) visually. Nodes in the CFG display disassembled code and edges are labeled by path formulas. The tool uses the Z3 SMT solver to prune the search space and compute concrete values to exploit potential vulnerabilities. Detailed documentation of the checked vulnerabilities is available online. Developed and maintained by ConsenSys, Mythril has been available on GitHub under an MIT license since September 2017.

Slither [12] is a tool designed for the static analysis of Solidity source code that is version 0.4 or higher. It offers APIs for users to customize their analysis of the source code and generates its own intermediate representation, known as SlitherR, from the Solidity source codes. SlitherR provides in-depth information on inheritance relationships and control flows for manual and automated inspection.

We used *MythX* to analyze source code. MythX [34] is a web-based platform that provides three tools for analyzing smart contracts: Mythril [33], Harvey [35], and Maru. It analyzes both the source code and compiled bytecode of smart contracts and combines the results of the three tools to improve accuracy. Mythril uses symbolic execution to explore all possible paths of a smart contract, while also applying taint analysis to more accurately identify known vulnerabilities. Harvey is a smart contract fuzzer that generates random inputs to test for

unexpected behavior, and Maru performs static analysis on smart contracts.

V. EXPERIMENT DESIGN

A. Designing Prompts

Prompt is a function of the LLM that provides a set of instructions to the program to customize, enhance, or refine the program.

Prompt plays a crucial role in shaping subsequent interactions and the output generated from an LLM by setting specific rules and guidelines for an LLM conversation. The prompt provides initial rules, sets the context for the conversation, and informs the LLM about what information is important, as well as what the desired output form and content should be. Prompt engineering is the process used to program LLMs using prompts.

The concept of prompt patterns is similar to that of software patterns in that it provides a codified approach to solving specific problems, but it focuses specifically on the context of output generation from LLMs such as ChatGPT. Prompt patterns provide a means to customize the output and interactions of LLMs, similar to how software patterns provide a codified approach to common software development challenges. By documenting and leveraging prompt patterns, users and teams can enforce constraints on the generated output, ensure that relevant information is included, and change the format of interaction with the LLM to better solve the problems that they are facing.

Each prompt pattern follows a similar format to classic software patterns, but with slight modifications to match the context of output generation with LLMs. It includes a name and classification, the intent and context, the motivation, the structure and key ideas, an example implementation, and consequences. The prompt pattern name uniquely identifies the pattern and indicates the problem being addressed, while the classification includes categories of pattern types such as output customization, error identification, prompt improvement, interaction, and context control. The intent and context describe the problem that the prompt pattern solves, and the goals that it achieves, while the motivation provides the rationale for the problem and explains why solving it is important. The structure and key ideas section describe the fundamental contextual information that the prompt pattern provides to the LLM, while the example implementation demonstrates how the prompt pattern is worded in practice. Finally, the consequences section summarizes the pros and cons of applying the pattern and may guide how to adapt the prompt to different contexts.

For our experiments, we used a guided approach to generate our prompts. For a generation, we used the following formula for generating prompts:

Step 1: All the extracted codes were passed through a Python parser that would extract the code comments from each of the Solidity files along with names. These were then saved in an intermediate file with the file name. **Step 2:** We would construct prompts based on the previously generated *code template* and *Step 1*.

- **Google Palm2:** For Palm2 the recipe we followed was: *Can you write a secure program that does filename on the code template.*
- **ChatGPT:** For GPT the recipe we followed was: *Can you give example of a secure program that does filename on the code template.*

B. Experimental Workflow

In Figure 2, we provide a diagram of how the experiment was conducted. First, the dataset has been collected from Github and then pruned using a data cleaner script. This has been de-duplicated to include only unique files. Then we use a combination of custom python script and solhint to extract comments from the .sol files. For

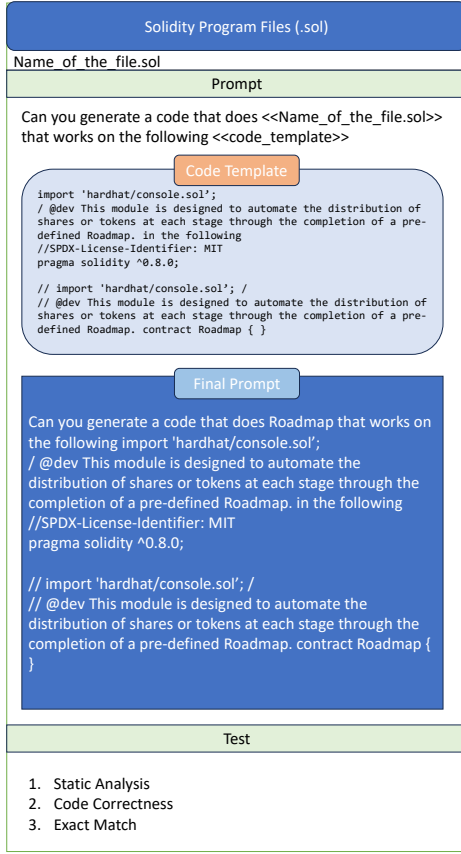


Fig. 1: Prompt generation example.

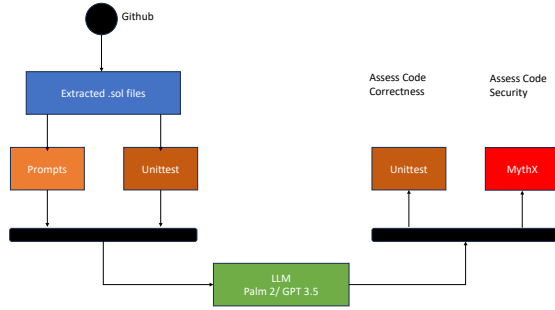


Fig. 2: A step-by-step illustration of the experiment's workflow.

each of the projects we evaluated, we also look at any unittest folder present in the repo. These are then used for generating prompt pairs using template. The unittest are kept in separate folders and tracked by Repo ID. The prompt generation has been detailed in Section V-A. Given the non-deterministic nature of the code generation tools, we inspect the prompts and then use Palm2 and Gpt 3.5 API to generate the codes with temperature 0. Once this is completed, we start the assessment phase by running unit tests on the codes to assess validity and correctness. After this, we utilize Mythx [34] to detect any potential bugs that might have been introduced. We detect the bugs and code smells separate, since the code smells are mostly different from bugs, they do not necessarily cause the code to be incorrect,

but introduce some uneasiness in the code which can cause more problems in the future. We save the individual assessment results related to the problem for each step of the assessment phase.

VI. EVALUATION

A. Code Metrics

We have assessed our results based on several metrics related to the quality of the generated Solidity code, including code validity, code correctness, code security, code reliability, and code maintainability.

To evaluate code validity, we use a binary metric with two possible values, 0 and 1, indicating whether the generated code is valid or not. We assess code validity by checking how well the code adheres to the syntax rules of Solidity and whether any errors occur during runtime. We use the solc compiler to check for code validity since our dataset is constructed for the Solidity programming language. After initiating the program, we record any errors that could be raised during runtime.

For code correctness, we measure the extent to which the generated code performs as intended. We use problem-specific unit tests from our GitHub Dataset to assess code correctness. On average, each problem comes with 7 unit tests, and we calculate code correctness as the number of passed unit tests divided by the total number of unit tests for a specific problem. We believe that this is the most convenient way to assess code correctness, given their abundance.

We also evaluate the average code correctness, which is measured as the sum of all code correctness scores divided by the problem count.

In addition to the metrics mentioned above, we also evaluated the following:

- *Code security*: We used MythX to assess the security of the generated code. MythX is a static analysis tool that can identify security vulnerabilities in code.
- *Code reliability*: We used the number of bugs detected by MythX to assess the reliability of the generated code. Bugs are errors in code that can cause the code to malfunction. Code smells [36] are patterns in code that can indicate potential problems, but they do not necessarily cause the code to be incorrect.
- *Code maintainability*: We used the number of comments in the generated code to assess the maintainability of the code. Comments are used to explain the code, which can make it easier to understand and maintain.

We show our calculation methods for Code Correctness and Average Code Correctness as defined by [36] below. *CCS* stands for Code Correctness Score, and CCS_i is the Code Correctness Score for the i^{th} problem. The range of i is 0 to n .

$$\text{Code Correctness} = \frac{\sum_{i=0}^n CCS_i [CCS_i = 1]}{n}$$

$$\text{Average Code Correctness} = \frac{\sum_{i=0}^n CCS_i}{n}$$

Here $n=86$ for the final dataset for the codes that were compilable for both ChatGPT and Palm2. However, we can generalize the formula to other scenarios.

In the final stages of our evaluation, we use MythX to assess the security, reliability, and maintainability of the generated code. We define a vulnerability as a weakness in the code that could be exploited by an attacker to gain unauthorized access to the system. Vulnerabilities can introduce potential risks, whether the code is deployed as standalone software or as part of a larger software system.

VII. RESULTS

A. Code Validity (RQ1.1)

The use of solc compiler made it easy for us to evaluate code validity just by trying to execute the generated codes. We were able to log all successful and unsuccessful generations and analyze them. Compile errors were also logged for guided code generation for our future work.

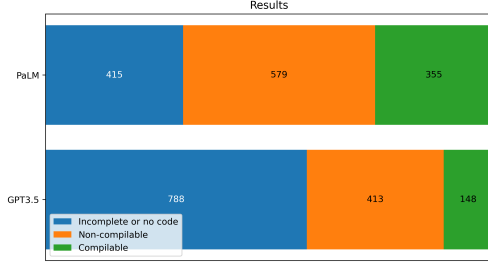


Fig. 3: Code compile statistics.

B. Code Correctness (RQ1.2)

We used the number of passed unit tests divided by all unit tests to calculate the success percentage of the code for each problem, as defined in Section VI-A.

We also measured the average code correctness score by dividing the summation of all code correctness scores by the number of all problems, as defined in Section VI-A.

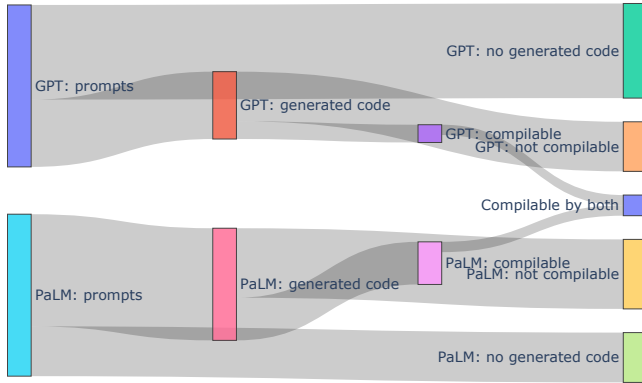


Fig. 4: Division of different generated codes based on their correctness.

C. Code Generation Improvement (RQ2)

In this part, as explained in Section V, we prompted ChatGPT and Palm2 to generate code for the same problems, this time with placeholder function names instead of meaningful, and informative function names, and also without any comment information, without code template. We replaced the function names with ‘waldo’.

Our code validity score increased to 93.9% for Palm2 and decreased to 89.6% for ChatGPT.

For code correctness, if we compare the results of the two experiments for ChatGPT, the rate of correctly generated code dropped by 4.2%. The incorrectly generated code percentage increased from 0.6%, and the partially correctly generated code percentage increased from 4.6%. For Palm2, the rate of correctly generated code dropped by 3.6%. The incorrectly generated code percentage increased from 0.6%, and the partially correctly generated code percentage increased from 3%.

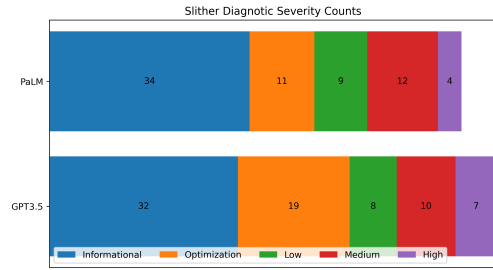


Fig. 5: Security incidents reported by Slither.

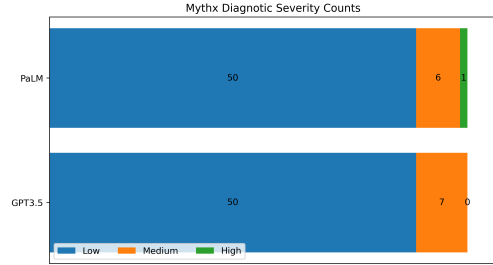


Fig. 6: Security incidents reported by MythX.

D. Code Security(RQ3)

As explained in Section VIII, the generated codes had different bugs that got reported by both Slither and Mythx. Fig. 5 and Fig. 6 show the distribution of different severity bugs among the generated code between two LLMs.

A detailed distribution of the security incidents has been described in TABLE II and TABLE III.

Surprisingly it seems Slither catches more High Severity bugs in code produced by ChatGPT than MythX. However, MythX shows a marginal increase in severe bugs produced in Palm2-generated Code.

TABLE II: Bugs found by Slither

		GPT3.5	PaLM2
Severity	Lint		
Informational	dead-code	24	27
	similar-names	0	2
	assembly	4	0
	reentrancy-unlimited-gas	1	0
	low-level-calls	1	1
	too-many-digits	1	1
	unused-state	1	3
Optimization	constable-states	7	7
	immutable-states	12	4
Low	reentrancy-benign	1	0
	calls-loop	1	1
	missing-zero-check	2	1
	reentrancy-events	0	1
	timestamp	3	3
	shadowing-local	1	3
	erc20-interface	0	5
Medium	incorrect-equality	5	4
	locked-ether	2	3
	tautology	3	0
High	uninitialized-state	4	0
	weak-prng	3	4

TABLE III: Bugs found by MythX

Severity	Lint	GPT3.5	PaLM2
Low	Assert Violation	25	28
	State Variable Default Visibility	5	15
	Weak Sources of Randomness from Chain Attributes	13	3
	Authorization through tx.origin	5	3
	Timestamp Dependence	1	2
	Reentrancy	1	0
Medium	Requirement Violation	0	1
	Incorrect Constructor Name	6	5
	DoS with Failed Call	1	1
High	Integer Overflow and Underflow	2	3

VIII. SECURITY ANALYSIS

A. Security Analysis

In this section, we demonstrate several code segments that were generated and have one or more specific security issues. We highlight some of the bugs we detected in these codes in the following sections. For all of these checks we used the tools from Section IV-B1.

Timestamp dependency. The concept of timestamp dependency refers to the use of the current timestamp of the Ethereum block by smart contracts to process time-dependent events. However, since the timestamp can be manipulated to some extent by miners, malicious miners may take advantage of this vulnerability to alter the timestamp for their own benefit. This bug was used in the governmental Ponzi scheme attack. As a result, developers should not rely solely on the precision of the block's timestamp.

```
pragma solidity ^0.8.0;
contract TimestampDependency {
    uint256 public lastTime;
    uint256 public counter;
    function increment() public {
        require(block.timestamp > lastTime + 1
            minutes, "Must wait at least 1 minute");
        lastTime = block.timestamp;
        counter++;
    }
}
```

Listing 1: Timestamp dependency

In Listing 1, the increment function increases a counter variable, but it only allows one increment per minute. The function uses the `block.timestamp` value to check when the last increment occurred.

However, this approach is vulnerable to manipulation by miners. A malicious miner could modify the timestamp of the block to allow multiple increments within one minute, breaking the intended behavior of the contract.

Re-entrancy. In Solidity, contracts can have external calls that can be accessed by attackers to execute a function within the same contract multiple times, leading to unexpected operations within the contract. The DAO attack in 2016 was a notable instance that exploited this vulnerability.

```
pragma solidity ^0.8.0;
contract Bank {
    mapping(address => uint) private balances;
    function deposit() public payable {
        balances[msg.sender] += msg.value;
    }
    function withdraw(uint amount) public {
        require(balances[msg.sender] >= amount, "Insufficient balance");
        (bool success,) = msg.sender.call{value: amount}("");
    }
}
```

```
require(success, "Transaction failed");
balances[msg.sender] -= amount;
}
function getBalance() public view returns (uint)
{
    return balances[msg.sender];
}
```

Listing 2: Re-entrancy

In this code Listing 2, the withdraw function allows a user to withdraw a certain amount of Ether from their account balance. However, the function contains a re-entrancy bug because it performs an external call to `msg.sender` before updating the balance, which allows an attacker to call the withdraw function repeatedly before the balance is updated, potentially causing them to withdraw more Ether than they have in their account.

Unchecked send. The vulnerability known as "Unauthorized Ether transfer" occurs when certain functions in a contract are visible to the public and allow non-zero sends, allowing external users to transfer Ether from the vulnerable contract even if they lack the necessary credentials. As a result, unauthorized individuals can call these functions and transfer Ether without proper authorization.

We have noticed this multiple times in ChatGPT-generated code, especially for Dutch auction.

```
pragma solidity ^0.8.0;
contract UncheckedSend {
    mapping(address => uint) balances;
    function deposit() public payable {
        balances[msg.sender] += msg.value;
    }
    function withdraw(uint amount) public {
        require(balances[msg.sender] >= amount, "Insufficient balance");
        msg.sender.call{value: amount}("");
        balances[msg.sender] -= amount;
    }
}
```

Listing 3: Unchecked Send

In the withdraw function, the contract transfers the requested amount of Ether to the caller using the call function. However, this send operation is unchecked, which means that if the call function fails for some reason, such as the receiver not having enough gas to execute its fallback function, the transaction will not be reverted and the Ether will be lost. This can result in unauthorized Ether transfers if the function is called by an attacker with insufficient balance or other invalid parameters.

All of these have been reported in the LLMs generated code among a plethora of other bugs as described in Section VII-D in detail.

B. Performance and Smart Contract Specific Analysis

In addition to correctness and security assessment, we have evaluated GPT-generated contracts using certain smart contract-specific metrics like contract size and gas fee cost.

We measured the bytecode size of the smart contracts generated by the GPT models. Compared with the handwritten contracts, the contracts generated by the GPT models may have smaller size. For instance, a manually written auction contract has bytecode size of 4.747KB. The ChatGPT generated version has size of 2.286KB.

Since the GPT models were trained with the source code dataset publicly available, when asked to generate a contract that is likely contained in the training dataset, the GPT models may generate contract code similar to the original code in the training dataset (with different variable names and/or function names). For instance, we have tested the GPT models to generate contract for Black-Scholes. The contract size by ChatGPT is 2.371KB, which is very close to the publicly available version that has 2.399KB bytecode size.

TABLE IV: Gas cost comparison for auction contracts.

	Contracts			
	ChatGPT		Manual	
	Min	Max	Min	Max
bid	2678	43075	0	66135
highestBid	526	526	0	816
highestBidder	486	486	0	596
endPrice	504	504	0	658

To evaluate the run-time performance of the GPT model-generated smart contracts, we measured gas cost for running smart contract functions. For certain tests, running AI-generated smart contract functions cost less gas than the manually written functions used for comparison. To collect the gas fee for each function, we applied a Solidity contract fuzzer [37]. The fuzzer, implemented on top of PyEVM, can analyze the source code and generate test cases for each function in a contract. At the same time, it measured gas fee cost, each time a function was called.

Table IV shows the gas cost results for two auction contracts, one created by ChatGPT and the other one manually written. The table lists min and max gas fees for four functions of each contract. Handwritten functions used more gas on average, due to the presence of additional code and abort logic. ChatGPT-generated functions did not contain this additional code, resulting in lower gas costs.

We observed that GPT models produce contract code that is typically smaller and less gas-intensive than code written by human developers.

IX. THREATS TO VALIDITY OF RESEARCH DESIGN

While our approach shares some similarities with previous studies, this work is more thorough than any other approach we are aware of for Solidity code. It is important to note, however, that current experiments have some limitations.

One-shot code generation. When we used code generators to generate code, we provided them with the function names, parameters, and a docstring explaining the function. We also provided them with a few test cases. However, we did not provide them with any additional code snippets that would clarify our intent for the specific problem. In most cases, the success rate of the code generators could be improved if we provided them with hints in the form of code snippets.

Reproduction of the generations. When we ran our experiments, we noticed that the code generators were not deterministic. This means that they would generate different outputs for the same input in different trials. To avoid this, we generated code for each problem in one iteration, saved the code, and then evaluated the saved code. However, even with this precaution, our results may not be fully reproducible because the underlying LLMs of the code generators are being retrained over time.

Metrics. As outlined in Section VI-A, we assessed the generated code utilizing the Code Validity, Correctness, and Security metrics. Nevertheless, we acknowledge the possibility of incorporating supplementary metrics, including Readability, Cyclomatic Complexity, and Reusability, to comprehensively evaluate the generated code.

X. RELATED WORKS

Since the time when the GPT models were made open to the public, the models have attracted significant interest from the research community. Researchers have evaluated these models for generating source code in particular from a security vulnerability perspective [38], [39]. [40] evaluated the security of the programs generated by GitHub Copilot. They found that 40% of the generated programs were vulnerable. These results differ from our results, which may be due to the characteristics of our dataset.

Vaithilingam et al [41] conducted a user study to assess the effects of GitHub Copilot. [42] evaluated GitHub Copilot using 33 LeetCode questions and four programming languages: Python, Java, JavaScript, and C. This also fundamentally differs from this work based on the programming language. [43] conducted an empirical study to investigate the effect of semantic-preserving changes in natural language on the generated code function of GitHub Copilot. They provided GitHub Copilot with 892 non-trivial Java method descriptions. In addition, researchers also explored LLMs for test code generation for languages such as Java [44].

However, most of the existing work has been focused on languages such as Python, JavaScript, etc. There was a gap in exploring the GPT models for smart contract generation in a comprehensive manner. This preliminary study attempted to fill this gap and highlighted some initial findings. Our research scratched the surface of this emerging research area that contains several potential directions for further investigation as pointed out in the analysis section.

XI. CONCLUSION

This work investigated two state-of-the-art code generation LLMs, namely OpenAI ChatGPT and Google Palm2 for generating smart contract code in Solidity. We evaluated the quality of the generated code in terms of correctness, validity, and security. We propose and demonstrate a validation framework for evaluating code quality for smart contract code, as well as a pipeline using which code verification can be mostly automated.

In addition, we proposed a code template-based code generation approach using prompts that help us achieve better compilable code than prompts without context.

Based on the findings, for now, Palm2 produces consistently more compilable code than ChatGPT. However, they are matched when it comes to introducing bugs in the generated code. This poses a problem for recommending any of these tools to actual programmers.

Future work will consist of improving upon the verification platform so that it can handle a much bigger dataset instead of a curated popular dataset. Furthermore, we aim to extend our template-based code generation to provide a more robust, universal, and safe way to generate smart contracts using LLMs.

AUTHOR CONTRIBUTIONS

Conceptualization: R.K; **Data curation:** R.K, E.L; **Formal analysis:** R.K; **Methodology:** R.K; **Resources:** W.S; **Supervision:** L.X and W.S; **Validation:** R.K, L.X and W.S; **Visualization:** R.K, E.L; **Writing – original draft:** R.K; **Writing – review & editing:** R.K, L.X and W.S;

REFERENCES

- [1] 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.
- [2] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen *et al.*, “Palm 2 technical report,” *arXiv preprint arXiv:2305.10403*, 2023.
- [3] J. Jang, S. Kim, S. Ye, D. Kim, L. Logeswaran, M. Lee, K. Lee, and M. Seo, “Exploring the benefits of training expert language models over instruction tuning,” 2023. [Online]. Available: <https://arxiv.org/abs/2302.03202>
- [4] J. Finnie-Ansley, P. Denny, B. A. Becker, A. Luxton-Reilly, and J. Prather, “The robots are coming: Exploring the implications of openai codex on introductory programming,” in *Proceedings of the 24th Australasian Computing Education Conference*, ser. ACE ’22. New York, NY, USA: Association for Computing Machinery, 2022, p. 10–19. [Online]. Available: <https://doi.org/10.1145/3511861.3511863>

- [5] R. Karanjai, "Targeted phishing campaigns using large scale language models," *arXiv preprint arXiv:2301.00665*, 2022.
- [6] J. Benson, "Uniswap trading volume exploded by 450% to \$7 billion. here's why," 2021. [Online]. Available: <https://decrypt.co/63280/uniswap-trading-volume-exploded-7-billion-heres-why>
- [7] K. Rabimba, L. Xu, L. Chen, F. Zhang, Z. Gao, and W. Shi, "Lessons learned from blockchain applications of trusted execution environments and implications for future research," in *Workshop on Hardware and Architectural Support for Security and Privacy*, ser. HASP '21. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: <https://doi.org/10.1145/3505253.3505259>
- [8] R. Karanjai, Z. Gao, L. Chen, X. Fan, T. Suh, W. Shi, and L. Xu, "Dhtee: Decentralized infrastructure for heterogeneous tees," in *2023 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, 2023, pp. 1–3.
- [9] R. Karanjai, L. Xu, N. Diallo, L. Chen, and W. Shi, "Defaas: Decentralized function-as-a-service for emerging dapps and web3," in *2023 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, 2023, pp. 1–3.
- [10] R. Karanjai, L. Xu, Z. Gao, L. Chen, M. Kaleem, and W. Shi, "Privacy preserving event based transaction system in a decentralized environment," in *Proceedings of the 22nd International Middleware Conference*, ser. Middleware '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 286–297. [Online]. Available: <https://doi.org/10.1145/3464298.3493401>
- [11] —, "On conditional cryptocurrency with privacy," in *2021 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, 2021, pp. 1–3.
- [12] K. B. Kim and J. Lee, "Automated generation of test cases for smart contract security analyzers," *IEEE Access*, vol. 8, pp. 209 377–209 392, 2020.
- [13] (2017). [Online]. Available: <https://www.sec.gov/litigation/investreport/34-81207.pdf>
- [14] M. Beller, R. Bholanath, S. McIntosh, and A. Zaidman, "Analyzing the state of static analysis: A large-scale evaluation in open source software," in *2016 IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, vol. 1. IEEE, 2016, pp. 470–481.
- [15] G. Chatzieleftheriou and P. Katsaros, "Test-driving static analysis tools in search of c code vulnerabilities," in *2011 IEEE 35th annual computer software and applications conference workshops*. IEEE, 2011, pp. 96–103.
- [16] M. Christakis and C. Bird, "What developers want and need from program analysis: an empirical study," in *Proceedings of the 31st IEEE/ACM international conference on automated software engineering*, 2016, pp. 332–343.
- [17] A. Gosain and G. Sharma, "Static analysis: A survey of techniques and tools," in *Intelligent Computing and Applications: Proceedings of the International Conference on ICA, 22-24 December 2014*. Springer, 2015, pp. 581–591.
- [18] K. Goseva-Popstojanova and A. Perhinschi, "On the capability of static code analysis to detect security vulnerabilities," *Information and Software Technology*, vol. 68, pp. 18–33, 2015.
- [19] L. Szekeres, M. Payer, T. Wei, and D. Song, "Sok: Eternal war in memory," in *2013 IEEE Symposium on Security and Privacy*. IEEE, 2013, pp. 48–62.
- [20] N. Ayewah, W. Pugh, D. Hovemeyer, J. D. Morgenthaler, and J. Penix, "Using static analysis to find bugs," *IEEE software*, vol. 25, no. 5, pp. 22–29, 2008.
- [21] A. Fioraldi, D. Maier, H. Eißfeldt, and M. Heuse, "Afl++ combining incremental steps of fuzzing research," in *Proceedings of the 14th USENIX Conference on Offensive Technologies*, 2020, pp. 10–10.
- [22] Y. Shoshitaishvili, R. Wang, C. Salls, N. Stephens, M. Polino, A. Dutcher, J. Grosen, S. Feng, C. Hauser, C. Kruegel *et al.*, "Sok:(state of) the art of war: Offensive techniques in binary analysis," in *2016 IEEE symposium on security and privacy (SP)*. IEEE, 2016, pp. 138–157.
- [23] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.
- [24] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacl-HLT*, 2019, pp. 4171–4186.
- [25] 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.
- [26] D. Fried, A. Aghajanyan, J. Lin, S. Wang, E. Wallace, F. Shi, R. Zhong, W.-t. Yih, L. Zettlemoyer, and M. Lewis, "InCoder: A generative model for code infilling and synthesis," *arXiv preprint arXiv:2204.05999*, 2022.
- [27] Y. Wang, W. Wang, S. Joty, and S. C. Hoi, "Codet5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation," *arXiv preprint arXiv:2109.00859*, 2021.
- [28] Z. Feng, D. Guo, D. Tang, N. Duan, X. Feng, M. Gong, L. Shou, B. Qin, T. Liu, D. Jiang *et al.*, "Codebert: A pre-trained model for programming and natural languages," *arXiv preprint arXiv:2002.08155*, 2020.
- [29] D. Guo, S. Ren, S. Lu, Z. Feng, D. Tang, S. Liu, L. Zhou, N. Duan, A. Svyatkovskiy, S. Fu *et al.*, "Graphcodebert: Pre-training code representations with data flow," *arXiv preprint arXiv:2009.08366*, 2020.
- [30] W. U. Ahmad, S. Chakraborty, B. Ray, and K.-W. Chang, "Unified pre-training for program understanding and generation," *arXiv preprint arXiv:2103.06333*, 2021.
- [31] 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.
- [32] 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.
- [33] (2023) Consensys/mythril: Security analysis tool for evm bytecode. supports smart contracts built for ethereum, hedera, quorum, vechain, roostock, tron and other evm-compatible blockchains. [Online]. Available: <https://github.com/Consensys/mythril>
- [34] (2022) Mythx: Smart contract security service for ethereum. [Online]. Available: <https://mythx.io/>
- [35] V. Wüstholtz and M. Christakis, "Harvey: A greybox fuzzer for smart contracts," in *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2020, pp. 1398–1409.
- [36] B. Yetişiren, I. Özsoy, M. Ayerdem, and E. Tüzün, "Evaluating the code quality of ai-assisted code generation tools: An empirical study on github copilot, amazon codewhisperer, and chatgpt," *arXiv preprint arXiv:2304.10778*, 2023.
- [37] D. Soto, A. Bergel, and A. Hevia, "Fuzzing to estimate gas costs of ethereum contracts," in *2020 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, 2020, pp. 687–691.
- [38] H. Pearce, B. Ahmad, B. Tan, B. Dolan-Gavitt, and R. Karri, "Asleep at the keyboard? assessing the security of github copilot's code contributions," 2021. [Online]. Available: <https://arxiv.org/abs/2108.09293>
- [39] H. Hajipour, T. Holz, L. Schönherr, and M. Fritz, "Systematically finding security vulnerabilities in black-box code generation models," 2023. [Online]. Available: <https://arxiv.org/abs/2302.04012>
- [40] H. Pearce, B. Ahmad, B. Tan, B. Dolan-Gavitt, and R. Karri, "Asleep at the keyboard? assessing the security of github copilot's code contributions," 2021. [Online]. Available: <https://arxiv.org/abs/2108.09293>
- [41] P. Vaithilingam, T. Zhang, and E. L. Glassman, "Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models," in *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '22. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: <https://doi.org/10.1145/3491101.3519665>
- [42] N. Nguyen and S. Nadi, "An empirical evaluation of github copilot's code suggestions," in *2022 IEEE/ACM 19th International Conference on Mining Software Repositories (MSR)*, 2022, pp. 1–5.
- [43] A. Mastropaolo, L. Pascarella, E. Guglielmi, M. Ciniselli, S. Scalabrino, R. Oliveto, and G. Bavota, "On the robustness of code generation techniques: An empirical study on github copilot," 02 2023.
- [44] P. Nie, R. Banerjee, J. J. Li, R. J. Mooney, and M. Gligoric, "Learning deep semantics for test completion," 2023.