

Code Refactoring with LLM: A Comprehensive Evaluation With Few-Shot Settings

Md. Raihan Tapader, Md. Mostafizer Rahman, Ariful Islam Shiplu, Md Faizul Ibne Amin, and Yutaka Watanobe

Abstract—In today’s world, the focus of programmers has shifted from writing complex, error-prone code to prioritizing simple, clear, efficient, and sustainable code that makes programs easier to understand. Code refactoring plays a critical role in this transition by improving structural organization and optimizing performance. However, existing refactoring methods are limited in their ability to generalize across multiple programming languages and coding styles, as they often rely on manually crafted transformation rules. The objectives of this study are to (i) develop an Large Language Models (LLMs) -based framework capable of performing accurate and efficient code refactoring across multiple languages (C, C++, C#, Python, Java), (ii) investigate the impact of prompt engineering (Temperature, Different shot algorithm) and instruction fine-tuning on refactoring effectiveness, and (iii) evaluate the quality improvements (Compilability, Correctness, Distance, Similarity, Number of Lines, Token, Character, Cyclomatic Complexity) in refactored code through empirical metrics and human assessment. To accomplish these goals, we propose a fine-tuned prompt-engineering-based model combined with few-shot learning for multilingual code refactoring. Experimental results indicate that Java achieves the highest overall correctness up to 99.99% the 10-shot setting, records the highest average compilability of 94.78% compared to the original source code and maintains high similarity ($\approx 53\text{--}54\%$) and thus demonstrates a strong balance between structural modifications and semantic preservation. Python exhibits the lowest structural distance across all shots ($\approx 277\text{--}294$) while achieving moderate similarity ($\approx 44\text{--}48\%$) that indicates consistent and minimally disruptive refactoring. C# consistently preserves semantic content, achieving up to 57.63% similarity at 10-shot, despite larger structural changes. C and C++ show moderate distances ($\approx 266\text{--}338$) and variable similarity (36–55%), reflecting more unpredictable outcomes.

Index Terms—Large Language Model, prompt engineering, code refactoring, API, fine-tuning, zero-shot prompting, few-shot prompting.

I. INTRODUCTION

Now-a-days, there is intense competition among programmers to write highly optimized code. The goal is to make code simple, concise, and easy to understand so that everyone from beginners to experts can read and maintain it effortlessly.

Md. Raihan Tapader is with the Department of Computer Science and Engineering, Dhaka University of Engineering & Technology, Gazipur, Bangladesh (e-mail: raihantapader@gmail.com).

Md. Mostafizer Rahman is with the Department of Computer Science, Tulane University, New Orleans, LA, USA (e-mail: mostafiz26@gmail.com, mrahman9@tulane.edu).

Ariful Islam Shiplu is with the Department of Computer Science and Engineering, Dhaka University of Engineering & Technology, Gazipur, Bangladesh (e-mail: shipluarifulislam@gmail.com).

Md Faizul Ibne Amin is with the Department of Computer and Information Systems, University of Aizu, Japan (e-mail: aminfazul007@gmail.com).

Yutaka Watanobe is with the Department of Computer and Information Systems, University of Aizu, Japan (e-mail: yutaka@u-aizu.ac.jp).

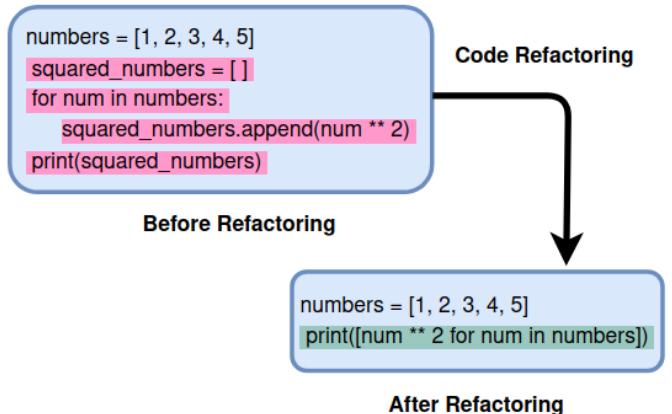


Fig. 1: A Readable Code Refactoring Example

Another major focus is on improving time and space efficiency. However, achieving this balance has become significant challenge because reducing execution time often increases memory uses, and vice versa. To address this, programmers continuously work on refactoring their code, improving earlier versions by removing redundancies, reducing the number of lines and tokens, and making the logic clearer. Code refactoring not only enhances performance but also improves code readability, maintainability, and adaptability to future changes as shown in Figure 1.

Over the few decades, code refactoring has evolved from being a purely manual and experience-driven practice to a systematic process supported by automated tools and frameworks. Early refactoring methods focused mainly on syntactic transformations, such as renaming variables, extracting methods, or eliminating duplicate code. Tools like Eclipse, IntelliJ IDEA, and other integrated development environments (IDEs) have helped standardize these transformation for making them accessible to a broader range of developers. However, as software systems have grown in size and complexity, the limitations of purely rule-based approaches have become apparent. Researchers have explored program analysis techniques such as abstract syntax trees (ASTs), control flow graphs (CFGs), and static analysis to detect code smells and suggest improvements.

More recently, machine learning and deep learning models have been applied to learn refactoring patterns from large-scale repositories to move beyond the surface-level changes toward deeper structural and semantic optimization [1]–[3]. Despite showing promising potential, current automated refactoring models struggle to generalize across languages and coding

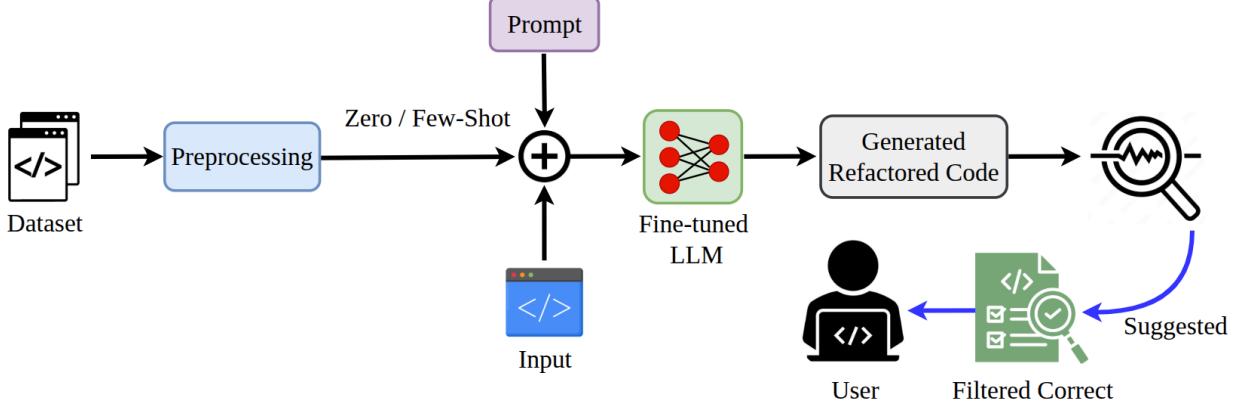


Fig. 2: Proposed approach

styles because they are typically trained on limited, domain-specific datasets. Moreover many existing approaches primarily target easily detectable code smells [4] or syntactic issues, rather than capturing the deeper semantic intent and architectural considerations embedded in complex code. Additionally, the inherent complexity of source code, which involves hierarchical structures, control flow dependencies, and data flow relationships poses significant challenges for standard sequence-based or graph-based models to fully understand and process.

Now, in the era of LLM, there is renewed interest in transforming how code refactoring is approached. Unlike earlier models that focused mainly on shallow syntactic changes or rule-based transformation, LLMs offer a deeper contextual understanding of both natural language and programming language [5], [6]. Studies [7] have shown that prompt design significantly affects model performance in code-related tasks, while Chen et al. [8] demonstrated that natural language prompts can successfully instruct models like Codex to produce syntactically and semantically correct code. Pornprasit et al. [9] use GPT-3.5 with zero-shot for code review automation and achieve impressive result. White et al. [10] work with Generative AI prompt pattern for code refactoring that help software engineering activities.

To the best of our knowledge, existing research does not adequately address several critical aspects of code refactoring. These include effectively highlighting key changes in refactored code, such as modifications in Variables and functions, estimating the success rates of refactoring operations, enabling seamless code transformation between different programming styles or paradigms, and ensuring bug-free correction of faulty code. Furthermore, there remains significant gap in supporting refactoring across diverse programming languages and in systematically exploring refactoring tasks that have yet to be thoroughly investigated. This lack of comprehensive tooling and methodological support ultimately limits programmer's and professional's ability to assess and improve code correctness, readability, maintainability, and security.

Scope: This research focuses on developing a domain specific fine-tuned prompt engineering model, further integrated with specialized few-shot learning algorithms, and evaluate

its effectiveness on a dataset spanning various programming languages.

The key contributions of this paper are as follows:

- We have created a dataset that includes code written in five different programming languages (C, C++, C#, Python, Java). In addition, we designed ten specialized prompts to guide the LLM in generating reliable code.
- Developed a model that combines fine-tuned prompt engineering with few-shot learning techniques to improve code refactoring performance.
- Experimental result Shows that Java programming language achieved the best result 99.99% correctness in 10-shot, and also achieved average 94.78% compilability compared with source code.

II. RELATED WORKS

Recent advances in LLMs have significantly transformed automated code generation tasks. Several studies have explored prompt engineering as an effective approach to guide LLMs in producing syntactically and semantically correct code. Work such as [7], [11] demonstrated that carefully crafted prompts can substantially improve model outputs without extensive retraining. However, these approaches often rely on manually designed prompts and do not address domain-specific adaptation.

In parallel, few-shot learning techniques have emerged to reduce the need for large labeled datasets by enabling models to generalize from a limited number of examples [12]. Prior research has shown promising results in applying few-shot learning to programming tasks. Yet most studies focus on single-language datasets or specific programming paradigms which is limit their applicability to diverse real-world environments.

Moreover, while instruction fine-tuning has gained attention for aligning LLMs with specific tasks, its potential in the domain of code refactoring remains underexplored. Studies such as [13]–[16] show that instruction-tuned models achieve better performance in code-related tasks (code editing, code generation) when explicitly guided by structured instructions. However, there is limited empirical research combining instruction fine-tuning with prompt engineering to assess their

joint impact on code quality improvement across multiple languages.

In the context of code refactoring, recent works have proposed neural approaches that attempt to automate code transformation. While these models offer performance improvements over rule-based systems, they often lack explainability, generalizability, and especially in multilingual settings. Metrics like cyclomatic complexity, code similarity, and code readability have been utilized to evaluate refactoring outcomes, but human evaluation remains essential to assess semantic preservation and maintainability.

Our work builds on this foundation by proposing an LLM-based multilingual refactoring framework, enriched with prompt engineering and instruction fine-tuning, and evaluates its effectiveness using both quantitative metrics (e.g., token length, character count, code similarity, cyclomatic complexity) and human-centered assessments.

III. METHODOLOGY

In this section, we provide a detailed description of our automated code refactoring model, which combines fine-tuned prompt engineering with few-shot learning to improve accuracy, efficiency, and generalizability. Leveraging prompt engineering allows the model to generate precise, context-aware code transformations, while few-shot learning facilitates adaptation to multiple programming languages and coding styles with minimal training data. The overall framework is depicted in Figure 2.

A. Overview

In recent years, extensive research by Shirafuji et al. [12] has focused on improving refactoring techniques, exploring various methods to boost readability, optimize performance, and reduce complexity. Advanced large language models (LLMs), like OpenAI Codex, are being utilized to experiment with sophisticated refactoring approaches. These models have demonstrated the ability to improve code organization, clarity, and efficiency, enabling developers to produce higher-quality, more efficient code. Figure 3 presents a real instance of refactored code produced by the model is demonstrated to showcase its practical effectiveness.

In this study, we first define the prompting module by designing ten distinct prompts. Each prompt consists of three key components: a custom instruction that highlights specific features as illustrated in Figure 4, along with a complex code snippet and its corresponding refactored version. Next, we outline the prompting strategies employed in our methodology, which include zero-shot, two-shot, four-shot, and few-shot prompting. In zero-shot prompting, only the instruction is provided without any accompanying examples. Conversely, two-shot, four-shot, and few-shot prompting incorporate two, four, and several relevant examples respectively, alongside the instruction. During the learning phase, the input code is supplied to a large language model (e.g., GPT-3.5-turbo) together with few-shot examples for instruction fine-tuning. The model learns the transformation patterns from the prompts

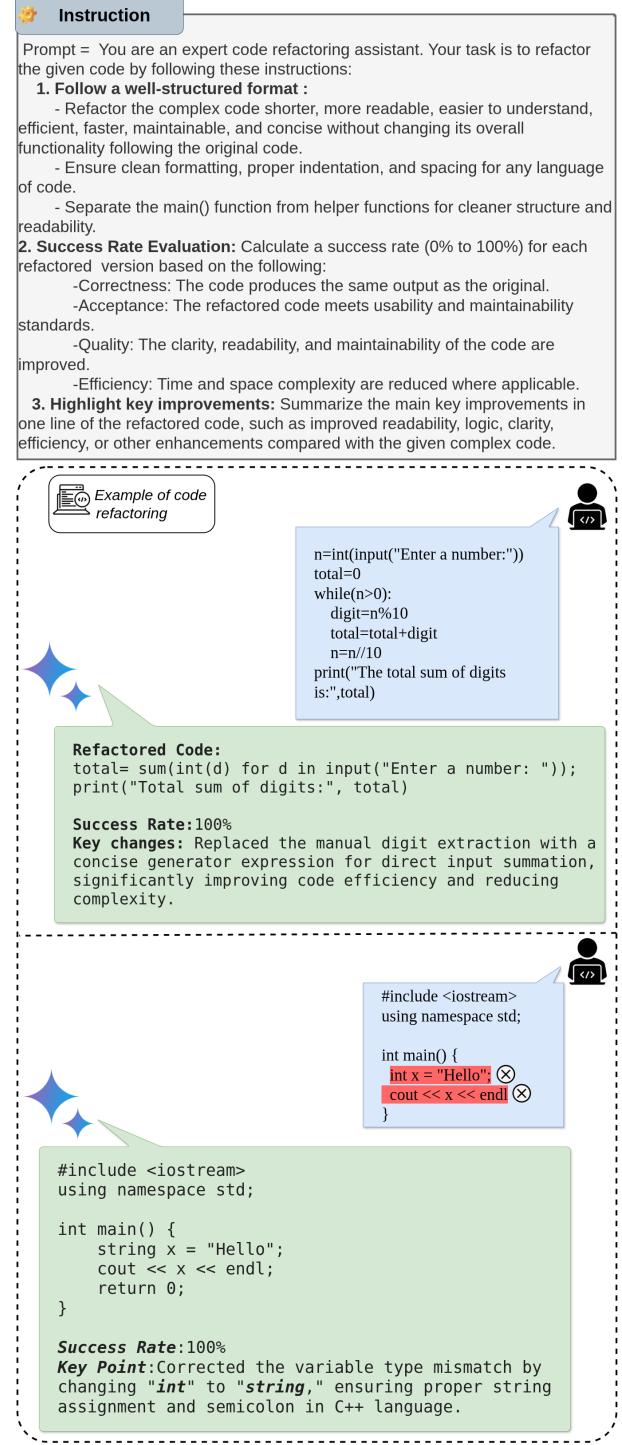


Fig. 3: Demonstration of model behavior: (1) instruction interpretation, (2) refactoring responses to input program based on system instructions, and (3) error handling proficiency when refactoring erroneous program. The model exhibits excellent performance.

and examples and produces a newly fine-tuned model specifically optimized for multilingual code refactoring. Finally, the fine-tuned model generates multiple refactored versions of the input code. These outputs undergo rigorous validation through

online compilers (e.g., Programiz), integrated development environments (IDEs) such as Visual Studio and Code::Blocks, as well as human evaluation to ensure syntactic correctness, semantic preservation, and enhanced maintainability. This workflow is further outlined in Algorithm 1.

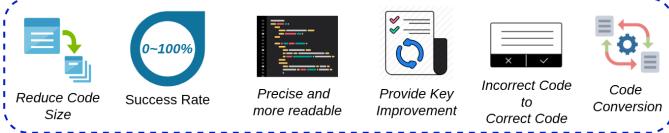


Fig. 4: Showing the special features of code refactoring which used to design prpmpt

B. Dataset

We curated a dataset \mathcal{D} consisting of 10 code samples from five programming languages: C, C++, C#, Python, and Java. Each sample contains an `Instruction`, a `Complex Code` snippet, and its corresponding `Refactored Code`. The dataset includes 2 samples each for C and C++, 3 for Python, and 1 each for C# and Java. Figure 5 depicts distribution of dataset samples. The instructions were systematically designed to ensure clarity, consistency, and relevance for refactoring tasks.

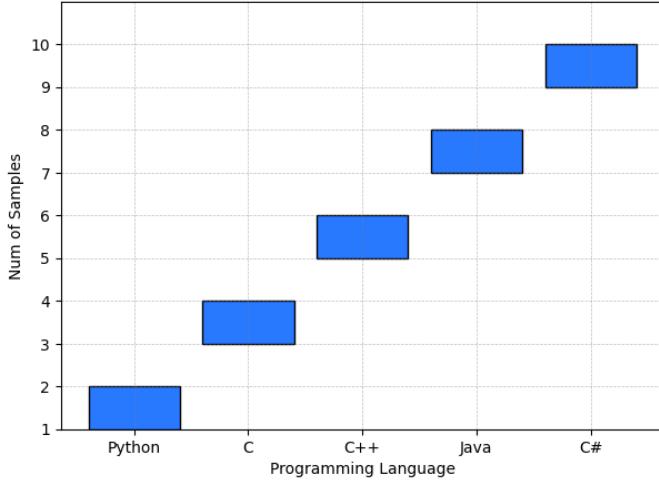


Fig. 5: Dataset Samples Distribution

Raw data in CSV format was converted into OpenAI's JSON format with roles `system`, `user`, and `assistant`. We rigorously validated the dataset to exclude malformed entries, those exceeding token limits, or missing fields, using the `tiktoken` library to enforce a 4096-token maximum. The preprocessing pipeline is depicted in Figure 6.

C. Refactoring Evaluation and Prompting Strategy

The dataset $\mathcal{D} = \{\text{Instruction}, \text{Complex Code}, \text{Refactored Code}\}$ supports evaluation via few-shot prompting. The number of examples n in the prompt varies by strategy:

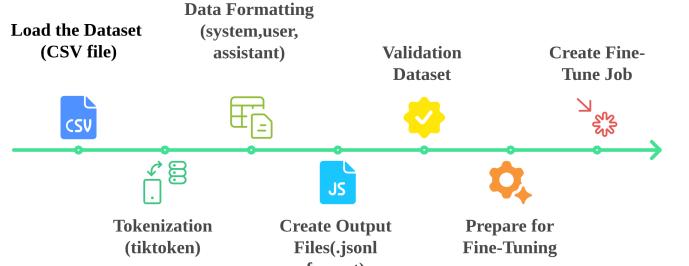


Fig. 6: End-to-End pipeline for data preprocessing and model fine-tuning

Algorithm 1 LLM-Based Multilingual Code Refactoring with Prompt Design and Few-Shot Learning

- 1: **Input:** Annotated code dataset with instructions, complex code, and refactored code
 - 2: **Output:** Multiple refactored code versions with success metrics from fine-tuned LLM
 - 3: **Dataset Preparation:** Load and convert dataset to OpenAI JSON format; discard invalid or oversized examples
 - 4: **Prompt Engineering:** Design zero-shot to few-shot prompts with selected example pairs embedded alongside the test code
 - 5: **Model Fine-Tuning:** Fine-tune GPT-3.5-Turbo on prepared dataset to obtain model M
 - 6: **Inference:** For each input code, generate prompts and query M to produce multiple diverse refactored outputs
 - 7: **Evaluation:** Assess each refactoring for Correctness, Cyclomatic Complexity, Line of Code, Compilability, Chars, Tokens, Similarity, Distance, and efficiency; summarize key improvements
 - 8: **Validation:** Verify refactored code via online compilation, IDE testing, and human review
-

$$n = \begin{cases} 0 & \text{Zero-shot} \\ 2 & \text{Two-shot} \\ 4 & \text{Four-shot}, \quad n \in \{0, \dots, 10\} \\ \vdots & \text{Few-shot} \end{cases}$$

For each complex code snippet C_i in the set $\mathcal{P} = \{C_1, C_2, \dots, C_N\}$, the prompt P_i is constructed as:

$$P_i = \{\mathcal{I}_i, \mathcal{E}_i, C_i\}$$

where \mathcal{I}_i is the instruction and $\mathcal{E}_i = \{(C^{(k)}, R^{(k)})\}_{k=1}^n$ are example pairs of complex and refactored code. The fine-tuned GPT-3.5-turbo model M generates $k = 5$ refactoring candidates per prompt input:

$$\mathcal{R}_i = M(P_i, \text{temperature} = t, n = k) = \{R_{i1}, \dots, R_{ik}\}$$

D. Model Configuration

We utilize the GPT-3.5-turbo-0125 model, leveraging its strong foundation in instruction tuning and code understanding derived from Codex and RLHF.

The temperature parameter is tuned between 0.1 and 1.5 to balance creativity and consistency, with 0.3 yielding optimal performance. The maximum token limit for output generation is set to 2,048, accommodating average input lengths of approximately 110 tokens.

A unified system instruction guides the model to refactor code to be shorter, more readable, efficient, and maintainable without altering functionality. Key directives include:

- Producing multiple distinct refactoring variants per input.
- Maintaining clean formatting, proper indentation, and descriptive naming.
- Separating the main function from helper functions for modularity.
- Evaluating refactorings on correctness, usability, quality, and efficiency.
- Summarizing key improvements succinctly for each candidate.

All inputs instruction, complex code, and refactored code are formatted as code blocks enclosed in triple backticks to maintain prompt clarity.

IV. EXPERIMENTAL RESULTS

In this section, we present the evaluation metrics employed to assess the quality of the refactored code and discuss the experimental results obtained from applying our proposed approach.

A. Metrics

1) *Pass@k*: is a metric for evaluating the *functional correctness* of generated programs by estimating the probability, introduced by Chen et al. [8]. It is defined in Equation 1, where n is the total samples, c is the correct samples, and k is the evaluated samples, satisfying $n \geq k$ and $c \leq n$ rules. Specifically, Pass@1 measures how often the first output is correct, while Pass@10 checks if at least one of the top 10 outputs is correct [12].

$$\text{Pass}@k := \mathbb{E} \left[1 - \frac{\binom{n-c}{k}}{\binom{n}{k}} \right] \quad (1)$$

2) *Correct@k*: In this paper, we create *Correct@k* metric as the ratio of correct outputs among the top- k generated results using few-shot prompting, as shown in Equation 2. This metric is specifically designed to assess the correctness of the k generated samples, where n represents the total number of generated outputs, e denotes the number of erroneous ones, and $(n - e)$ indicates the number of correct outputs. The parameter k controls the level at which correctness is measured.

$$\text{Correct}@k = \begin{cases} 100\%, & k \leq (n - e), \\ \frac{n-e}{k}, & k > (n - e). \end{cases} \quad (2)$$

For each new input code snippet C_i , we generate $n = 5$ refactored versions and evaluate correctness using *Correct@k*, where $k \in \{1, 2, \dots, 5\}$. In this work, *Correct@1* refers to the proportion in which at least one of the generated outputs

is correct, whereas *Correct@5* reflects the cases where all five outputs are correct. Subsequently, if a generated code R_i satisfies the correctness criteria defined by the equation, the success rate is computed accordingly.

3) *McCabe's Cyclomatic Complexity(CC)*: is a software metric used to measure the number of linearly independent paths through a program's source code. It quantifies the program complexity by counting the number of decision points(e.g., if statements, while loops, for loops, switch statements) in the code [17]. A lower CC is preferable and makes the code easier to understand and maintain in software development. We employed a custom AST-based library to accurately compute the CC for both correct and incorrect code samples.

4) *Compilability*: refers to whether a generated program is syntactically correct and successfully compiles, reflecting the model's capabilities to produce valid and executable code. It is calculated as equation 3:

$$\text{Compilability} = \left(\frac{P}{P+E} \right) \times 100\% \quad (3)$$

where P denotes the number of programs that compile successfully, and E denotes the number of programs that fail to compile. In our case, the compilability of the complex(original) programs used for refactoring is 100%, as all of them were correct and problem-solved.

5) *Lines of Code (LOC)*: is a software metric that measures the number of lines in a source code. In this work, we use Source Lines of Code (SLOC), which excludes comments, blank lines, and structural braces (e.g., { }), focusing solely on the actual executable statements. For refactoring analysis, SLOC is preferred over raw LOC because it avoids inflated counts from non-functional lines and gives a more accurate picture of the actual number of lines and logic implemented rather than comments or empty lines.

6) *Chars*: refers to the total length of how much character is used in a code on average. This matrix sometimes yields anomalous values due to long string literals, comments, and excessive whitespace. To accurately measure the character length, we normalize the refactored code by removing such anomalies. For example, consider the string literals:

$A = \text{'Enter a num:'}$ and $B = \text{"H"}$; these would initially result in different character counts. However, after normalization, both are treated equivalently as $A = '$ and $B = "$ ', leading to consistent and fair character length comparisons.

7) *Tokens*: refers to the total length of how many tokens are used in a code on average. Tokens are the fundamental components of a program that the compiler processes. In this work, we used an official Python tokenizer for Python code, while for other languages, we employed a regex-based tokenizer. To ensure accurate token counts, all code snippets are normalized by removing comments, newlines, and indentation before tokenization.

8) *Levenshtein Distance*: also known as edit distance, is a string metric that quantifies the difference between two sequences (typically strings) by counting the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one string into the other. In this study,

TABLE I: Quantitative comparison of Compilability, CC, LOC, character count, and token count for refactored programs across five programming languages under varying shot settings.

Shots	Programming Language	Compilability	CC	LOC	Chars	Tokens
0-shot	C	90.00	4.99 (± 0.07)	6.04 (± 0.83)	197.56 (± 11.40)	150.25 (± 9.41)
	C++	92.67	4.94 (± 0.27)	8.15 (± 1.51)	232.13 (± 18.13)	156.36 (± 14.71)
	C#	88.67	5.01 (± 0.27)	6.13 (± 2.19)	223.08 (± 12.71)	146.61 (± 9.74)
	Python	91.33	2.10 (± 0.97)	6.06 (± 0.48)	287.89 (± 26.79)	123.71 (± 6.62)
	Java	90.00	3.32 (± 1.10)	8.34 (± 3.16)	229.19 (± 39.94)	118.69 (± 31.99)
2-shot	C	83.00	4.95 (± 0.46)	6.17 (± 1.32)	217.43 (± 17.18)	168.79 (± 16.92)
	C++	90.00	4.99 (± 0.17)	7.51 (± 1.10)	238.74 (± 39.69)	165.69 (± 18.37)
	C#	94.00	4.99 (± 0.07)	11.02 (± 1.23)	255.74 (± 9.04)	164.53 (± 5.59)
	Python	94.67	2.11 (± 0.54)	7.23 (± 0.98)	246.65 (± 24.43)	123.77 (± 6.11)
	Java	94.00	4.92 (± 0.62)	8.59 (± 1.10)	323.83 (± 32.12)	185.90 (± 24.52)
4-shot	C	87.00	4.88 (± 0.33)	7.43 (± 2.21)	229.91 (± 19.59)	156.73 (± 15.23)
	C++	92.67	4.97 (± 0.13)	9.17 (± 1.98)	265.78 (± 16.64)	160.19 (± 9.54)
	C#	97.33	4.94 (± 0.18)	11.64 (± 1.28)	269.29 (± 13.04)	161.45 (± 7.86)
	Python	93.33	1.79 (± 1.05)	6.47 (± 1.20)	269.77 (± 28.52)	120.05 (± 7.73)
	Java	93.33	4.37 (± 0.89)	8.34 (± 1.60)	300.25 (± 46.57)	167.35 (± 34.74)
6-shot	C	93.00	4.91 (± 0.38)	12.29 (± 2.58)	252.03 (± 20.88)	159.07 (± 16.49)
	C++	94.00	2.58 (± 0.68)	9.89 (± 1.62)	223.39 (± 24.46)	112.15 (± 15.79)
	C#	96.67	4.94 (± 0.22)	10.84 (± 1.60)	273.36 (± 16.97)	163.35 (± 12.40)
	Python	87.33	2.50 (± 1.16)	7.29 (± 1.71)	237.57 (± 43.22)	121.39 (± 10.13)
	Java	96.00	2.33 (± 0.62)	6.67 (± 1.55)	190.10 (± 28.23)	83.93 (± 20.01)
8-shot	C	93.33	4.95 (± 0.25)	11.67 (± 3.19)	260.99 (± 26.76)	160.63 (± 14.55)
	C++	91.33	2.93 (± 1.10)	9.92 (± 2.01)	225.71 (± 42.30)	116.18 (± 27.93)
	C#	95.33	4.73 (± 0.49)	9.20 (± 1.53)	272.13 (± 28.72)	162.07 (± 21.53)
	Python	87.33	1.88 (± 1.08)	5.84 (± 1.65)	294.47 (± 36.28)	120.63 (± 10.53)
	Java	96.67	4.99 (± 0.07)	8.97 (± 1.59)	344.15 (± 14.53)	195.21 (± 13.95)
10-shot	C	96.00	4.97 (± 0.21)	12.05 (± 2.79)	259.00 (± 19.05)	158.39 (± 12.58)
	C++	92.00	4.93 (± 0.29)	12.14 (± 2.23)	303.55 (± 21.78)	162.53 (± 13.89)
	C#	95.33	4.91 (± 0.29)	10.91 (± 2.42)	307.29 (± 32.90)	162.00 (± 15.84)
	Python	90.67	3.32 (± 1.16)	8.95 (± 1.91)	271.87 (± 31.99)	112.81 (± 9.98)
	Java	98.67	4.92 (± 0.18)	8.81 (± 1.46)	344.39 (± 18.07)	190.32 (± 15.01)

we do not normalize the code before comparison in order to showing the actual distance btween original and refactored code. For example, consider the following two strings:

A: "int x+y+z", and B: "int y-z". The total number of edits needed to convert A into B is 3, so the Levenshtein distance is 3.

9) *Similarity*: is a metric used to measure how alike the original, and refactored code are, based on the Levenshtein distance. A higher distance implies lower similarity, and vice versa. In this study, Similarity between two programs a,b, with lengths $|a|$, $|b|$, is defined as Equation 4:

$$\text{Similarly}(a, b) = \left(1 - \frac{\text{Distance}(a, b)}{\max(|a|, |b|)}\right) \quad (4)$$

Here, Distance (a,b) represents the Levenshtein distance between the two programs.

B. Results

We conducted comprehensive experiments across multiple programming languages using a fine-tuned large language model with few-shot prompting to assess its effectiveness in automated code refactoring. Table I summarizes key code quality metrics—Compilability, CC, LOC, Chars, and token count for five programming languages (C, C++, C#, Python, and Java) evaluated under different few-shot prompting scenarios ranging from 0-shot to 10-shot.

In the 0-shot baseline, compilability ranges narrowly between 88.67% (C#) and 92.67% (C++) which indicates that even without examples, the model maintains reasonable syntactic consistency. CC is relatively stable for C, C++, and C#

around 5.0, while Python and Java generate simpler control flows with lower complexity 2.10 and 3.32, respectively. The refactored programs at this stage are compact, mostly between 6–8 lines of code and 200–290 characters, with Python being the most concise in token count 123.71. Introducing two example prompts significantly alters the compilability trends across languages. While C experiences a decline to 83% compilability, higher-level managed languages such as C#, Python, and Java demonstrate notable improvements, reaching compilability rates of 94% and above. Concurrently, LOC increase modestly for all languages, with particularly pronounced growth in C# (11.02 lines) and Java (8.59 lines), reflecting longer and potentially more detailed function implementations. Corresponding increases are also observed in character and token counts, consistent with the enhanced code elaboration. At the 4-shot prompting level, compilability stabilizes across languages, with C# achieving the highest rate of 97.33%. Python's CC further decreases to 1.79, indicating the generation of highly linear and simplified code structures. Meanwhile, C++ and Java exhibit increased verbosity, each surpassing 260 characters on average. C# continues to show the greatest increase in LOC, reaching 11.64 lines, suggesting that the model tends to produce more descriptive or boilerplate code when provided with additional contextual examples. A notable shift in trends emerges at the 6-shot prompting level. Java attains a compilability of 96%, closely followed by C# at 96.67%. In contrast, Python experiences a significant decline to 87.33%, reversing its previous stability. Interestingly, C++ exhibits a marked reduction in cyclomatic complexity to 2.58, indicating that additional examples encourage the model to

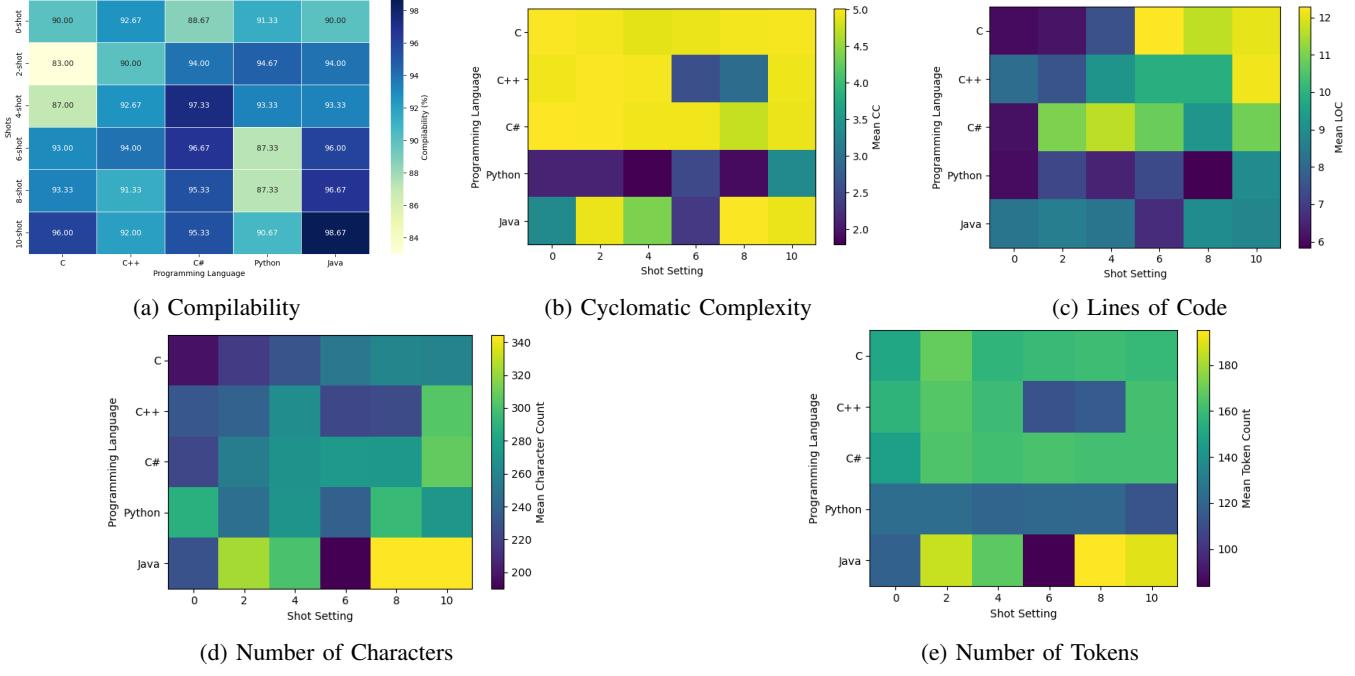


Fig. 7: Visualization of program metrics across five programming languages (C, C++, C#, Python, Java) and varying shot settings. Subplots show (a) Compilability (%), (b) Cyclomatic Complexity, (c) Lines of Code, (d) Number of Characters, and (e) Number of Tokens, highlighting trends and variations across 0-shot to 10-shot settings.

TABLE II: Quantitative comparison between original and LLM-refactored programs across five programming languages, evaluated on Compilability, Cyclomatic Complexity, Lines of Code, character count, and token count. Each pair of rows reports the baseline properties of the original implementation and the corresponding refactored variant generated by the model.

Programming Language	#	Compilability	CC	LOC	Chars	Tokens
C	Original	99.99	5.00 (± 0.00)	19 (± 0.00)	387 (± 0.00)	176 (± 0.00)
	Refactored	90.39	4.94 (± 0.28)	9.28 (± 2.15)	236.15 (± 19.14)	158.98 (± 14.20)
C++	Original	99.99	5.00 (± 0.00)	20 (± 0.00)	419 (± 0.00)	185 (± 0.00)
	Refactored	92.11	4.22 (± 0.44)	9.46 (± 1.74)	248.22 (± 27.17)	145.52 (± 16.71)
C#	Original	99.99	5.00 (± 0.00)	20 (± 0.00)	426 (± 0.00)	181 (± 0.00)
	Refactored	94.56	4.92 (± 0.25)	9.96 (± 1.71)	266.81 (± 18.90)	160.00 (± 12.16)
Python	Original	99.99	5.00 (± 0.00)	18 (± 0.00)	407 (± 0.00)	135 (± 0.00)
	Refactored	90.78	2.28 (± 0.99)	6.97 (± 1.32)	268.04 (± 31.87)	120.39 (± 8.52)
Java	Original	99.99	5.00 (± 0.00)	21 (± 0.00)	487 (± 0.00)	198 (± 0.00)
	Refactored	94.78	4.14 (± 0.58)	8.29 (± 1.74)	288.65 (± 29.91)	156.90 (± 23.37)

generate code with fewer branching paths. Meanwhile, C reaches its highest lines of code count at 12.29, suggesting that the model begins producing more extended and detailed low-level code structures at this stage. At 8-shot, performance remains high for all except Python still 87.33%. Java reaches 96.67%, while C and C# maintain 95%. Complexity levels remain low for Python 1.88 and C++ 2.93. Java again produces the most verbose outputs, reaching 344.15 characters and 195.21 tokens, suggesting that the model begins over-elicitating large output structures. The best compilability across the entire table is achieved here, with 98.67% for Java and 96.00% for C, while all other languages score above 92%. CC values converge toward 5.0 for C, C++, and C#, whereas Python and Java retain moderate complexity. LOC increases further, with C++ (12.14) and C# (10.91) yielding longer functions, while Python remains compact 8.95 LOC. Token count reaches its overall peak in Java 190.32 tokens, reinforcing that Java scales in verbosity proportionally with more examples. Figure

7 shows the heatmap visualization of all metrics.

Table II presents a quantitative comparison between original and LLM-refactored code across five programming languages (C, C++, C#, Python, and Java) evaluated on Compilability, CC, LOC, character count, and token count. Across all languages, the original code exhibits perfect or near-perfect compilability (99.99%), serving as a strong ground truth baseline. In refactored versions, Java demonstrates the highest post-refactoring reliability (94.78%), closely followed by C# (94.56%) which indicates that strongly-typed, object-oriented languages remain structurally resilient under automated transformations. Conversely, C and Python experience the greatest degradation, suggesting that syntax rigidity in C and formatting sensitivity in Python make them more vulnerable to syntactic inconsistencies during refactoring. In terms of Cyclomatic Complexity, refactored code tends to preserve or slightly reduce structural complexity across most languages. For example, C++ decreases from 5.00 to 4.22, and Java from

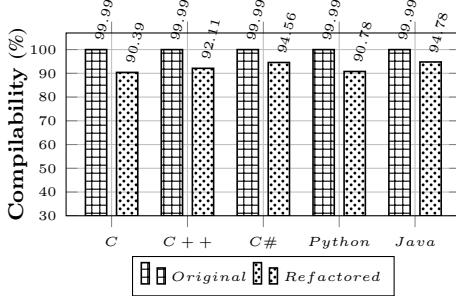


Fig. 8: Compilability accuracy across original and refactored code

5.00 to 4.14, suggesting that the model often favors simpler control flow. Python shows the largest drop, from 5.00 to 2.28, implying that LLM refactoring may employ more linear or flattened logic structures in dynamic languages. A major change is observed in code length metrics. All refactored programs show a significant reduction in Lines of Code, with average LOC dropping from 18–21 lines to approximately 7–10 lines across all languages. This compression effect is further reflected in character count and token count, both of which decrease sharply. For instance, Java reduces from 487 to 288 characters, and Python from 407 to 268, while token counts drop from \approx 180 down to \approx 150 or lower. Interestingly, C# and Java retain higher CC values while achieving moderate LOC reductions which suggests that the model preserves structural richness while shortening syntax. Conversely, Python shows aggressive simplification both in structure and length, reflecting language-specific biases in refactoring behavior. Figure 8 illustrates the comparison between original and our proposed model performance.

Table III presents the average Correct@1–5 scores across five programming languages—C, C++, C#, Python, and Java—evaluated under varying shot configurations ranging from 0-shot to 10-shot prompting. Overall, the results reveal that all languages maintain consistently high correctness, typically surpassing 95% accuracy regardless of the number of examples provided. Python exhibits the lowest performance under zero-shot prompting (78.66%), indicating its stronger reliance on in-context examples compared to other languages. However, it rapidly improves once few-shot guidance is introduced, peaking at 99.32% in the 2-shot setting. Java and C demonstrate exceptional stability, frequently exceeding 98–99% correctness across multiple shot counts, while C# achieves its highest accuracy with 4-shot prompting (99.32%) before experiencing minor fluctuations. Interestingly, increasing shots does not always correlate with higher performance; certain languages achieve their peak correctness at mid-range shot levels (2–6 shots) rather than at the highest shot count. Figure 9 illustrates the correctness accuracy of various programming language in different shot settings.

Table IV presents a comprehensive evaluation of Distance and Similarity metrics for refactored code. Distance measures the structural deviation between the refactored code and the original, while Similarity quantifies the degree of lexical or semantic alignment. Across all shot settings, Python consis-

TABLE III: Average Correct@1–5 scores for five programming languages across different few-shot settings, illustrating the effect of shot count on code generation correctness.

Shots	C	C++	C#	Python	Java
0-shot	97.66	97.41	95.63	78.66	97.16
2-shot	96.07	96.32	97.49	99.32	98.49
4-shot	97.88	96.38	99.32	98.66	96.75
6-shot	99.32	99.32	98.49	95.99	99.32
8-shot	97.91	98.16	93.27	93.88	98.91
10-shot	97.82	97.07	93.08	96.66	99.99

tently achieves the lowest mean distances, ranging approximately from 276.93 (2-shot) to 293.67 (10-shot), indicating that refactored Python code preserves structural fidelity better than other languages. In contrast, Java and C# frequently exhibit the highest distances, particularly at higher shots, reaching up to 509.93 for Java at 6-shot that reflects more substantial structural modifications during refactoring. C and C++ generally maintain moderate distances, suggesting that while the structural changes are noticeable, they are not as extreme as those observed for Java or C#. In terms of similarity, C# consistently achieves the highest mean values, peaking at 57.63% under 10-shot prompting that indicates that despite larger structural changes, the semantic and lexical content is largely retained. Java also maintains high similarity values at higher shots (\approx 53–54%) that highlights its ability to balance structural changes with semantic preservation. Python exhibits moderate similarity values (\approx 44–48%), which, combined with its low distance, that indicates that the refactoring is consistent and minimally disruptive. C and C++ display variable similarity across shots, with mean values ranging from 36% to 55% that reflects less predictable outcomes depending on the shot configuration. Examining shot-wise trends, low-shot settings (0–2 shots) generally result in higher distances and lower similarity for most languages which suggests that minimal examples are insufficient to guide the model effectively. Moderate-shot settings (4–6 shots) tend to improve both metrics and achieve a better balance between structural fidelity and semantic preservation. At high-shot settings (8–10 shots), Python maintains low distance, while C# and Java continue to achieve high similarity which highlights the optimal shot configuration depends on the language and desired outcome. Overall, Python is the most structurally consistent across all shots, C# preserves semantic content most effectively, and Java provides a balanced trade-off between structural modification and semantic similarity at higher shots, emphasizing the importance of language-specific strategies and shot configurations in automated code refactoring.

V. CONCLUSION

This study presents a comprehensive investigation into the effectiveness of LLMs for multilingual code refactoring across C, C++, C#, Python, and Java. By designing a prompting module with ten distinct prompts and employing zero-shot, few-shot, and multi-shot strategies, we fine-tuned the model to generate high-quality refactored code while preserving semantic correctness. Quantitative evaluation across multiple metrics including compilability, cyclomatic complexity, lines

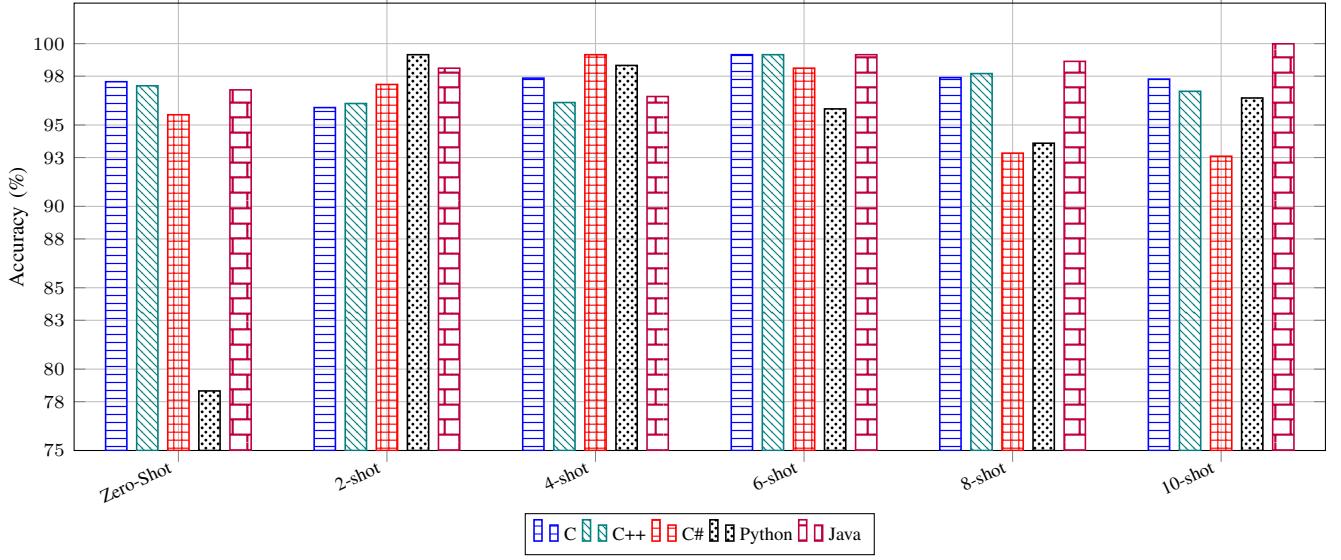


Fig. 9: Correct@1–5 averages for five programming languages across few-shot settings.

TABLE IV: Distance and Similarity metrics for refactored code across five programming languages under varying shot settings, showing mean, standard deviation, and min–max ranges to capture structural and semantic changes.

Shots	Programming Language	Distance			Similarity		
		Mean(Std)	Min ~ Max		Mean(Std)	Min ~ Max	
0-shot	C	327.19 (\pm 20.74)	228.00	400.00	45.10% (\pm 3.48%)	32.89%	61.74%
	C++	338.70 (\pm 23.44)	262.00	497.00	45.19% (\pm 3.79%)	19.58%	57.61%
	C#	465.27 (\pm 40.03)	335.00	607.00	40.12% (\pm 5.15%)	21.88%	56.89%
	Python	294.92 (\pm 20.85)	252.00	411.00	44.39% (\pm 3.89%)	22.45%	52.45%
	Java	484.62 (\pm 41.08)	336.00	603.00	33.98% (\pm 5.60%)	17.85%	54.22%
2-shot	C	327.99 (\pm 21.14)	218.00	413.00	44.96% (\pm 3.55%)	30.70%	63.42%
	C++	321.71 (\pm 30.15)	256.00	814.00	48.28% (\pm 3.58%)	18.52%	58.58%
	C#	340.23 (\pm 30.90)	258.00	525.00	56.21% (\pm 3.98%)	32.43%	66.80%
	Python	276.93 (\pm 22.39)	243.00	371.00	47.75% (\pm 4.22%)	30.00%	54.15%
	Java	366.52 (\pm 35.86)	297.00	541.00	50.13% (\pm 4.90%)	26.29%	59.54%
4-shot	C	303.27 (\pm 20.76)	229.00	387.00	49.12% (\pm 3.48%)	35.07%	61.58%
	C++	298.04 (\pm 17.27)	259.00	460.00	51.77% (\pm 2.79%)	25.57%	58.09%
	C#	336.52 (\pm 29.40)	234.00	639.00	56.72% (\pm 3.80%)	17.76%	69.88%
	Python	293.85 (\pm 19.29)	240.00	384.00	44.57% (\pm 3.61%)	30.05%	54.72%
	Java	381.26 (\pm 56.08)	283.00	585.00	48.06% (\pm 7.64%)	20.30%	61.44%
6-shot	C	278.35 (\pm 32.28)	223.00	439.00	53.30% (\pm 5.41%)	26.34%	62.58%
	C++	420.29 (\pm 18.11)	231.00	485.00	31.99% (\pm 2.93%)	21.52%	62.62%
	C#	337.33 (\pm 40.19)	235.00	564.00	56.59% (\pm 5.17%)	27.41%	69.76%
	Python	292.47 (\pm 30.15)	229.00	435.00	44.81% (\pm 5.69%)	17.92%	56.79%
	Java	509.93 (\pm 22.27)	350.00	552.00	30.53% (\pm 3.03%)	24.80%	52.32%
8-shot	C	266.33 (\pm 29.19)	206.00	450.00	55.34% (\pm 4.93%)	24.50%	68.50%
	C++	395.13 (\pm 50.76)	229.00	466.00	36.06% (\pm 8.21%)	24.60%	62.94%
	C#	384.07 (\pm 46.00)	269.00	643.00	50.57% (\pm 5.92%)	17.25%	65.38%
	Python	301.74 (\pm 26.39)	192.00	401.00	43.07% (\pm 4.98%)	24.34%	63.77%
	Java	339.45 (\pm 16.89)	303.00	433.00	53.75% (\pm 2.30%)	41.01%	58.72%
10-shot	C	269.11 (\pm 23.97)	192.00	476.00	54.85% (\pm 4.02%)	20.13%	67.79%
	C++	306.83 (\pm 32.44)	234.00	439.00	50.35% (\pm 5.25%)	28.96%	62.14%
	C#	329.25 (\pm 44.95)	204.00	665.00	57.63% (\pm 5.78%)	14.41%	73.75%
	Python	293.67 (\pm 21.15)	240.00	421.00	44.64% (\pm 3.92%)	20.57%	54.72%
	Java	339.33 (\pm 26.12)	255.00	524.00	53.77% (\pm 3.56%)	28.61%	65.26%

of code, character and token counts that demonstrates the model’s ability to produce syntactically correct and maintainable code with notable improvements in structural simplicity. The Results show that Java stands out as the top-performing language, achieving not only the highest average compilability rate of 94.78% but also the best correctness score of 99.99% in the 10-shot setting. This demonstrates that the model’s strong reliability when provided with sufficient contextual examples. Python, on the other hand, consistently yields the lowest

cyclomatic complexity and suggests that the refactored outputs tend toward simpler and more maintainable structures. Across all shot configurations, the Correct@k performance remains consistently high across languages, indicating that even minimal prompting is sufficient for generating functionally accurate code. The distance and similarity analyses further confirm that the refactored programs preserve core semantic behavior while introducing meaningful structural improvements. Additionally, the comparative results between original and LLM-refactored

implementations clearly show that the model is capable of reducing code complexity and lines of code without sacrificing functional integrity.

REFERENCES

- [1] M. Aniche, E. Maziero, R. Durelli, and V. H. Durelli, “The effectiveness of supervised machine learning algorithms in predicting software refactoring,” *IEEE Transactions on Software Engineering*, vol. 48, no. 4, pp. 1432–1450, 2020.
- [2] B. K. Sidhu, K. Singh, and N. Sharma, “A machine learning approach to software model refactoring,” *International Journal of Computers and Applications*, vol. 44, no. 2, pp. 166–177, 2022.
- [3] P. Naik, S. Nelaballi, V. S. Pusuluri, and D.-K. Kim, “Deep learning-based code refactoring: A review of current knowledge,” *Journal of Computer Information Systems*, vol. 64, no. 2, pp. 314–328, 2024.
- [4] A. Kaur and M. Kaur, “Analysis of code refactoring impact on software quality,” in *MATEC web of conferences*, vol. 57, p. 02012, EDP Sciences, 2016.
- [5] M. M. Rahman, A. I. Shiplu, Y. Watanobe, and M. A. Alam, “Roberta-bilstm: A context-aware hybrid model for sentiment analysis,” *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2025.
- [6] A. Shirafuji, M. M. Rahman, M. F. I. Amin, and Y. Watanobe, “Program repair with minimal edits using codet5,” in *2023 12th International Conference on Awareness Science and Technology (iCAST)*, pp. 178–184, IEEE, 2023.
- [7] J. Shin, C. Tang, T. Mohati, M. Nayebi, S. Wang, and H. Hemmati, “Prompt engineering or fine-tuning: An empirical assessment of llms for code,” in *2025 IEEE/ACM 22nd International Conference on Mining Software Repositories (MSR)*, pp. 490–502, IEEE, 2025.
- [8] 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.
- [9] C. Pornprasit and C. Tantithamthavorn, “Fine-tuning and prompt engineering for large language models-based code review automation,” *Information and Software Technology*, vol. 175, p. 107523, 2024.
- [10] J. White, S. Hays, Q. Fu, J. Spencer-Smith, and D. C. Schmidt, “Chatgpt prompt patterns for improving code quality, refactoring, requirements elicitation, and software design,” in *Generative AI for Effective Software Development*, pp. 71–108, Springer, 2024.
- [11] J. Cordeiro, S. Noei, and Y. Zou, “An empirical study on the code refactoring capability of large language models,” *arXiv preprint arXiv:2411.02320*, 2024.
- [12] A. Shirafuji, Y. Oda, J. Suzuki, M. Morishita, and Y. Watanobe, “Refactoring programs using large language models with few-shot examples,” in *2023 30th Asia-Pacific Software Engineering Conference (APSEC)*, pp. 151–160, IEEE, 2023.
- [13] K. Li, Q. Hu, X. Zhao, H. Chen, Y. Xie, T. Liu, Q. Xie, and J. He, “Instructcoder: Instruction tuning large language models for code editing,” *arXiv preprint arXiv:2310.20329*, 2023.
- [14] X. Zhang, Z. Z. Chen, X. Ye, X. Yang, L. Chen, W. Y. Wang, and L. R. Petzold, “Unveiling the impact of coding data instruction fine-tuning on large language models reasoning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 25949–25957, 2025.
- [15] Z. Yuan, J. Liu, Q. Zi, M. Liu, X. Peng, and Y. Lou, “Evaluating instruction-tuned large language models on code comprehension and generation,” *arXiv preprint arXiv:2308.01240*, 2023.
- [16] Z. Ma, H. Guo, J. Chen, G. Peng, Z. Cao, Y. Ma, and Y.-J. Gong, “Llamoco: Instruction tuning of large language models for optimization code generation,” *arXiv preprint arXiv:2403.01131*, 2024.
- [17] T. J. McCabe, “A complexity measure,” *IEEE Transactions on software Engineering*, no. 4, pp. 308–320, 1976.