

Exploring and Unleashing the Power of Large Language Models in Automated Code Translation

ZHEN YANG, Shandong University, China

FANG LIU*, Beihang University, China

ZHONGXING YU*, Shandong University, China

JACKY WAI KEUNG, City University of Hong Kong, China

JIA LI, Peking University, China

SHUO LIU, YIFAN HONG, and XIAOXUE MA, City University of Hong Kong, China

ZHI JIN and GE LI, Peking University, China

Code translation tools, namely transpilers, are developed for automatic source-to-source translation. Latest learning-based transpilers have shown impressive enhancement against rule-based counterparts in both translation accuracy and readability, owing to their task-specific pre-training on extensive monolingual corpora. Nevertheless, their current performance still remains unsatisfactory for practical deployment, and the associated training resources are also prohibitively expensive. Large Language Models (LLMs), pre-trained on huge amounts of human-written code/text, have shown remarkable performance in many code intelligence tasks due to their powerful generality, even without task-specific re-training/fine-tuning. Thus, LLMs can potentially circumvent the above limitations, but they have not been exhaustively explored yet. This paper investigates diverse LLMs and learning-based transpilers for automated code translation tasks, finding that: although certain LLMs have outperformed current transpilers, they still have some accuracy issues, where most of the failures are induced by a lack of comprehension of source programs (38.51%), missing clear instructions on I/O types in translation (14.94%), and ignoring discrepancies between source and target programs (41.38%).

Enlightened by the above findings, we further propose **UniTrans**, a **Unified code Translation** framework, applicable to various LLMs, for unleashing their power in this field. Specifically, **UniTrans** first crafts a series of test cases for target programs with the assistance of source programs. Next, it harnesses the above auto-generated test cases to augment the code translation and then evaluate their correctness via execution. Afterward, **UniTrans** further (iteratively) repairs incorrectly translated programs prompted by test case execution results. Extensive experiments are conducted on six settings of translation datasets between Python, Java, and C++. Three recent LLMs of diverse sizes, including GPT-3.5 and LLaMA-13B/7B, are tested with **UniTrans**, and all achieve substantial improvements in terms of computational accuracy and exact match accuracy among almost all translation settings, showing the universal effectiveness of **UniTrans** in practice.

CCS Concepts: • **Software and its engineering** → **Genetic programming**.

Additional Key Words and Phrases: Automated Code Translation, Large Language Models, Transformer

*Corresponding Author

Authors' addresses: Zhen Yang, Shandong University, Qingdao, China, zhenyang@sdu.edu.cn; Fang Liu, Beihang University, Beijing, China, fangliu@buaa.edu.cn; Zhongxing Yu, Shandong University, Qingdao, China, zhongxing.yu@sdu.edu.cn; Jacky Wai Keung, City University of Hong Kong, Hong Kong, China, Jacky.Keung@cityu.edu.hk; Jia Li, Peking University, Beijing, China, lijia@stu.pku.edu.cn; Shuo Liu, sliu273-c@my.cityu.edu.hk; Yifan Hong, yifanhong7-c@my.cityu.edu.hk; Xiaoxue Ma, xiaoxuema3-c@my.cityu.edu.hk, City University of Hong Kong, Hong Kong, China; Zhi Jin, zhijin@pku.edu.cn; Ge Li, lige@pku.edu.cn, Peking University, Beijing, China.

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

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM 2994-970X/2024/7-ART71

<https://doi.org/10.1145/3660778>

ACM Reference Format:

Zhen Yang, Fang Liu, Zhongxing Yu, Jacky Wai Keung, Jia Li, Shuo Liu, Yifan Hong, Xiaoxue Ma, Zhi Jin, and Ge Li. 2024. Exploring and Unleashing the Power of Large Language Models in Automated Code Translation. *Proc. ACM Softw. Eng.* 1, FSE, Article 71 (July 2024), 24 pages. <https://doi.org/10.1145/3660778>

1 INTRODUCTION

With the advancement and prosperity of software, various Programming Languages (PLs) were invented to deal with diverse development scenarios and needs, such as desktop applications, websites, and mobile applications. More and more software developed in one PL also has the necessity to be ported to other languages to satisfy the extension of business platforms [33, 38, 53]. Therefore, different transpilers [9, 25, 40, 46, 47] aiming at source-to-source translations emerged during the last several decades to improve the codebase migration efficiency.

Traditional rule-based transpilers require not only expertise in both source and target languages but also considerable effort and time-cost in rule design [46]. For example, the Commonwealth Bank of Australia invested approximately \$750 million and dedicated five years to migrate their platform from COBOL to Java [66]. Meanwhile, the translation performance in readability and correctness are also poor [46]. Therefore, a series of learning-based transpilers were proposed to improve the translation efficiency and efficacy with Neural Machine Translation (NMT) methods [14, 46–48], which normally leverage diverse task-specific pre-training on huge amounts of monolingual corpora. Previous studies have demonstrated the impressive improvement of these learning-based transpilers, but their current performance is still unsatisfactory for practical deployment, and the training resources are also unaffordable. For example, TransCoder [46], one of the state-of-the-art learning-based transpilers, was trained on 32 V100 GPUs for 12 days, whereas its translation performance among various mainstream PL pairs all cannot reach 50% in terms of computational accuracy according to our empirical results. Recent studies reveal that Large Language Models (LLMs), pre-trained on billions of text/code tokens, bypass the need for re-training/fine-tuning but demonstrate the powerful generality of various code-related tasks, such as code generation [13, 17, 29, 31, 34, 59], program repair [18, 54], and code summarization [8, 20]. However, as an alternative solution in automated code translation, their potential has not been exhaustively investigated yet.

Empirical Study. This work performs an empirical study on five recent LLMs, including GPT-3.5 [43], LLaMA-33B/13B/7B [50], and CodeGen [41], for automated code translation task and compare their performance with the three state-of-the-art learning-based transpilers, including TransCoder [46], TransCoder-IR [48], and TransCoder-ST [47]. We first manually clean the widely-used code translation dataset released by Roziere et al. [46] due to many errors and inconsistencies among its parallel corpus. Based on the cleaned dataset, we evaluate the above models under the metrics of Computational Accuracy (CA) and Exact Match Accuracy (EM Acc), where four settings of translation datasets (i.e., C++ to Java, Java to C++, C++ to Python, and Python to C++) are involved in this empirical study. Results demonstrate certain LLMs have outperformed state-of-the-art learning-based transpilers, and LLMs with more parameters tend to carry more powerful translation capabilities, showing that automated code translation with LLMs is promising. Nevertheless, LLMs still suffer some accuracy issues. To delve deep into these issues and find rescues to improve their performance further, taking GPT-3.5 as an example, we manually analyze 174 failures it made and partition them into various categories, e.g., failures concerning Syntax, Logic, API, etc., where 38.51% of failures are induced by a lack of comprehension of the source programs, 14.94% of them are brought by the missing instructions of explicit I/O types, and 41.38% of them are caused by the ignorance of the discrepancies between source and target programs.

UniTrans. Enlightened by the above findings, we propose a **Unified code Translation** framework, namely **UniTrans**, to unleash various LLMs' capabilities in this field. In general, **UniTrans** exploits auto-generated test cases¹ as extra information for LLMs to alleviate the aforementioned drawbacks. On the one hand, test cases imply the requirements of programs for code comprehension. Besides, I/O types can be easily labeled on test cases to complement the translation objective. In addition, test cases can also be executed to double-check or offer hints (e.g., error messages) to repair the translated programs, thereby alleviating failures brought about by neglecting discrepancies. Specifically, **UniTrans** consists of three phases, i.e., (1) the Test Case Generation Phase, (2) the Translation Augmentation Phase, and (3) the Translation Repair Phase. The Test Case Generation Phase leverages LLMs to generate a series of inputs with the *Input Generation Prompt* for source programs. Afterward, valid inputs can be selected, and their corresponding outputs can also be obtained via the execution of source programs, thereby gathering test cases. Since test cases are composed of simple I/O pairs, practitioners can easily decorate I/O types on their corresponding positions and convert them to fit target programs via heuristics. As such, the Translation Augmentation Phase can exploit these test cases to improve code translation quality with the *Translation Augmentation Prompt*. Following that, the translated programs are double-checked by the above test cases via execution, and unpassed ones are shipped to the Translation Repair Phase to extract error information and further repair with the *Repair Prompt*. **UniTrans** also provides an option of iterative repair for users to fix bugs in multiple rounds based on feedback from test cases.

To evaluate the effectiveness and universal applicability of **UniTrans**, we experiment with three LLMs, namely GPT-3.5, LLaMA-13B, and LLaMA-7B, on all six settings of translation datasets between Python, Java, and C++, a total of 568 samples for each PL, where the translation between Python and Java is newly introduced to verify the generality of our findings among unseen translation datasets. Extensive experimental results demonstrate **UniTrans** substantially boosts the code translation efficacy of three tested LLMs on almost all translation datasets. To be specific, GPT-3.5 obtains average improvements of 4.02% in terms of CA and 13.28% in terms of EM Acc. LLaMA-13B achieves average improvements of 19.20% and 36.42% in terms of CA and EM Acc, respectively. LLaMA-7B demonstrates average improvements of 28.58% and 71.22% in terms of each metric in order. Furthermore, we carry out a series of ablation studies and discussion experiments to investigate the contribution and influence of **UniTrans**'s each module with various LLMs, showing that test cases throughout the whole life-cycle of **UniTrans** are critically important. The contributions of this paper can be summarized below:

- We manually and rigorously cleaned the widely-used code translation dataset, including parallel PLs of Python, Java, and C++, and made an explicit breakdown to record our cleaning process. The cleaned dataset and the breakdown table have been published in [2].
- We carry out an empirical study to investigate the performance of recent LLMs on automated code translation and carefully analyze their prospects and limitations. Meanwhile, a series of invaluable findings are summarized.
- We propose and evaluate **UniTrans**, a **Unified code Translation** framework applicable to various LLMs to unleash their power in code translation. Motivated by our findings, **UniTrans** introduces auto-generated test cases throughout the whole code translation framework via translation augmentation and repair. Comprehensive evaluations are conducted to examine the effectiveness of **UniTrans** both quantitatively and qualitatively, including ablation studies, discussion experiments, and case studies. We open-source **UniTrans** in [2].

¹In this work, we generate test cases via **UniTrans**, which should be discriminated from the evaluation-purpose unit test/test suite provided by the dataset.

2 BACKGROUND AND RELATED WORK

2.1 Automated Code Translation

Automated code translation tools aim to construct a function that approximates f such that given the source program x , the target program $y = f(x)$. Early studies proposed various rule-based approaches, such as C2Rust [6] and CxGo [4] to carry out translations from C to Rust and Go with manually crafted rules. However, they are language-dependent and extremely labor-intensive, as developers have to implement an enormous amount of translation rules for every function, object, and standard library of every language pair. Besides, the translated programs also suffer low readability and correctness [33, 37, 39]. To this end, learning-based transpilers have been successively proposed during the last several years [14, 25, 40, 42]. Owing to the unavailability of bilingual samples, state-of-the-art learning-based approaches [46–48] leverage unsupervised/weakly supervised techniques to train their model with massive monolingual data. Although they achieved remarkable improvement against previous rule-based approaches, their limitations are still evident, i.e., (1) their performance is still insufficient for practical deployment, and (2) their training expenditures are very resource-consuming and expensive. To combat the above limitations, LLMs are considered in this work due to their powerful generality on a broad spectrum of code-related tasks without the necessity of re-training/fine-tuning [36, 62, 65]. Consequently, we explore the ability of various LLMs in the code translation task, along with summarizing their strengths and weaknesses. Moreover, we propose **UniTrans** to further unleash LLMs' code translation performance.

2.2 Automated Test Case Generation

Numerous literature has put forward many approaches to automatically generate test cases for a focal method during the last several decades. Traditional tools, such as Randoop [44], EvoSuite [19], and MOSA [45], exploit search-based heuristics to generate test cases, suffering limitations on diversity and quantity. A series of follow-up studies [16, 51] proposed various learning-based approaches to overcome the above limitations, but they require massive training resources. Recent research leverages the generality of LLMs to generate test cases via zero-shot/few-shot learning, which has drawn lots of attention in academia and industry owing to their impressive performance and lightweight characteristic [12, 28, 49, 61]. Nonetheless, owing to the nature of code translation, source programs are known. Therefore, we leverage source programs to validate the generated program inputs and gather corresponding outputs, thereby creating exactly correct test cases.

2.3 Automated Program Repair

Automated Program Repair (APR) tools are designed to patch buggy code given the original code and corresponding buggy location. Traditional APR tools [10, 21, 27, 35] are based on human-crafted heuristics or templates, leading to a lack of generality on unseen types of bugs. Subsequently, learning-based APR tools, such as Recoder [67], DeepFix [23], and CODIT [11], emerged to generate more diverse and expressive patches. Nevertheless, owing to their enormous reliance on historical bug-fixing data, researchers nowadays tend to explore more lightweight approaches using LLMs without re-training/fine-tuning [54–56]. Different from the above approaches, given the auto-generated test cases for evaluation, our idea can employ the specific error information fetched from the program execution results to assist the program repair during the Translation Repair Phase.

3 MOTIVATION

This section introduces the motivation of **Unitrans**, which is derived from an in-depth analysis of failed cases of LLMs. In fact, we first conducted an empirical study on four translation datasets, i.e., C++ to Java, Java to C++, C++ to Python, and Python to C++, with two evaluation metrics, i.e.,

Computational Accuracy (CA) and Exact Match Accuracy (EM Acc) mentioned in Section 5.3, to explore the performance of various recent LLMs against state-of-the-art learning-based transpilers. The detailed experimental setting, dataset, and results are presented in Section 5 and 6.1. During the empirical study, we take the best-performing LLM (i.e., GPT-3.5) as an example and carry out the in-depth analysis based on a series of failed cases sampled from GPT-3.5’s results, thereby exploring potential improvement directions for LLMs in code translation.

Specifically, to ensure a 95% confidence level and 5% confidence interval when sampling, we follow previous work [15, 26, 57, 64] to randomly sample a total of 195 failed cases² from GPT-3.5’s empirical results, where 52 are from C++ to Python, 43 are from Python to C++, 30 are from Java to C++, and 70 are from C++ to Java. Particularly, 21 cases extracted from the dataset of C++ to Java are ignored, as their mistakes are owing to the misintroduction of “import” statements or wrapped by “class”, which can be easily eliminated by regular expressions. Eventually, 174 failed cases remained. Subsequently, we made a systematic taxonomy according to the cause of each failure. In detail, independent labeling is conducted by the first and second authors, followed by a double-check through a review process, and then a final taxonomy decision is reached through in-depth discussions. Following the above rigorous procedure, the remaining 174 failed cases are categorized into six classes.

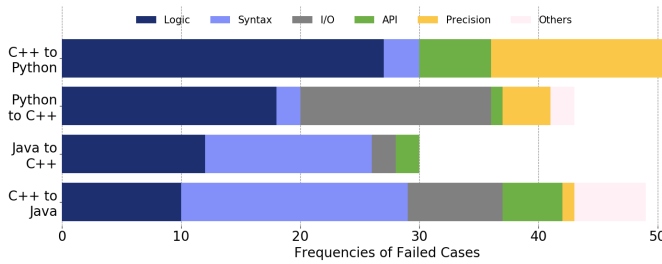


Fig. 1. Frequencies of Failed Cases in Each Translation Dataset

The detailed taxonomy, examples, and corresponding numbers are shown in Table 1. Moreover, we present Figure 1 to showcase the frequencies of various failures in each translation dataset. Apparently, **Logic failure** is the most prominent category, accounting for 20.41% to 51.92% of all failures across diverse translation datasets, indicating that lack of comprehension of the source program is a primary weakness of GPT-3.5. The **I/O failure** occurs most frequently (37.21%) when translating from Python to C++. Because dynamically typed PLs (e.g., Python) without explicit type declaration can hardly infer the parameters/return values’ type of statically typed PLs (e.g., C++), showing that prompting I/O types is necessary when code translation. On the contrary, when translating from statically typed PLs to dynamically typed PLs, **mistakes concerning Precision** become apparent (28.85%) owing to their differences in data types (e.g., C++ has double/float types, while Python only has the float type.) and operations (e.g., integer division as shown in Table 1). Additionally, the **Syntax failure** mainly appears when translating between Java and C++, as these two PLs’ syntaxes are more complicated compared with Python, leading to GPT-3.5 ignoring some discrepancies between source and target PLs. Similarly, **mistakes of API**, such as API mis-migration or misuse, are still typically due to the unfamiliarity of both PLs when translation. In summary, we attribute failures concerning Logic to the lack of comprehension of the source program (38.51%). Besides, we argue the I/O failure is induced by missing explicit I/O type declaration (14.94%). Finally, we credit mistakes relevant to Syntax, API, and Precision to the ignorance of the discrepancies between source and target PLs (41.38%).

²Here, we define the failed cases as those that cannot pass all given unit tests, i.e., using CA as the assessment criterion.

Enlightened by the above findings in the manual analysis, we carry an idea of introducing test cases as extra information to alleviate the limitations of LLMs during their code translation, and the reasons are three-fold. Firstly, test cases imply the requirements of programs, facilitating LLMs' comprehension of the program logic. Besides, I/O types can be easily decorated on test cases, thereby mitigating the mistakes induced by missing I/O type instructions. Thirdly, executing test cases can double-check or provide hints (e.g., error messages) to rectify incorrectly translated programs, thereby alleviating mistakes brought by neglecting discrepancies between PLs. Consequently, **UniTrans** is proposed in this work, and its detailed introduction is elaborated in the following sections.

4 UNITRANS

UniTrans consists of three phases, namely (1) the Test Case Generation Phase, which leverages LLMs and source programs to generate test cases, (2) the Translation Augmentation Phase, employing auto-generated test cases as extra information to augment translation and inspect their correctness, and (3) the Translation Repair Phase, which further repairs incorrectly translated programs assisted by test case execution results. The overview of **UniTrans** is shown in Figure 2. We elaborate on each of its components as follows.

Table 1. Failure Cases Taxonomy

Category	Description and Examples			Amount
Logic	The translated program bears logical inconsistencies against the ground truth program.			67
	Source Program	Translated Program	Ground Truth Program	
	return * max_element(arr, arr+n); //source C++ code //gets the max value of the top-n items.	return Arrays.stream(arr).max().getAsInt(); Arrays.sort(arr, 0, n); //incorrect Java code //gets the max value of the whole array.	Arrays.sort(arr, 0, n); return arr[n-1]; //ground truth Java code	
Syntax	Incorrectly introduces the syntax of the source PL to the target PL or makes the syntactic errors in the target PL.			38
	Source Program	Translated Program	Ground Truth Program	
	return (!(x/z)) ? x : z; //source C++ code	return (!(x/z)) ? x : z; //incorrect Java code //!" is not applicable for digits in Java.	return ((x/z) == 0) ? x : z; //ground truth Java code	
I/O	The translated program carries inconsistent I/O types against the ground truth program.			26
	Source Program	Translated Program	Ground Truth Program	
	def evenlength(n): #source Python code	int evenlength(int n){ //incorrect C++ code //incorrect input/output types.	string evenlength (string n) { //ground truth C++ code	
API	Incorrectly duplicates the API symbols of the source PL to the target PL or misuses APIs of the target PL.			14
	Source Program	Translated Program	Ground Truth Program	
	stack <int> s; ... int tp = s.top(); //source C++ code	Stack<Integer> s = new Stack<>(); ... int tp = s.top();//incorrect Java code //stack:top() is only available in C++	Stack<Integer> s = new Stack<>(); ... int tp = s.peek(); //ground truth Java code	
Precision	The translated program returns digits of inconsistent precision against the ground truth program.			20
	Source Program	Translated Program	Ground Truth Program	
	return (b/m-1)*(b/m)/2; #source C++ code	return (b/m-1)*(b/m)/2 #incorrect Python code #as 'b' is an integer parameter, integer #division ('//') should be used instead.	return (b/m-1)*(b/m)//2 #ground truth Python code	
Others	Other mistakes.			9
Total				174

^Φ Only code fragments are listed in examples. Comments highlighted in red are the specific errors of the translated programs.

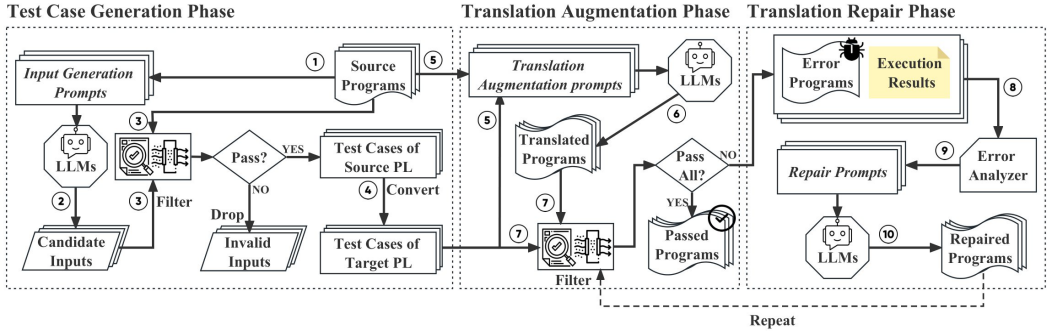


Fig. 2. UniTrans

4.1 Test Case Generation

Test Case Generation Phase is presented on the left-hand side of Figure 2. Instead of directly generating test cases for focal methods, we leverage LLMs to generate methods' candidate inputs first (Step 1-2) with the *Input Generation Prompt*. This prompt can be formally defined as “ $\${prog}_{src}$ Please generate ten groups of differentiated valid inputs for the above focal method of $\${pl}_{src}$ language, in the format of [Input_1]\n[Input_2]\n...[Input_10]. Finally, use END_OF_CASE to finish your answer.”, where $\${prog}_{src}$ and $\${pl}_{src}$ are placeholders for a source program and the name of source PL (e.g., Java and Python). An example is shown in Figure 3, where the text in bold and italic format are natural language instructions, the same as the follow-up prompts shown in Figures 4, 5, and 7. Particularly, we instruct LLMs to generate 10 candidate inputs for each inference, explicitly stating the requirement that inputs need to be differentiated. Afterward, we use source programs as calibrations to filter valid inputs and obtain their corresponding outputs via execution (Step 3), thereby constructing collections of test cases tc_{src} . Before pipelining to the second phase, we convert their formats to fit the target PLs via heuristic rules, thereby formulating tc_{tar} (Step 4). tc_{tar} can be arranged as “Inputs:\n $\${inp}_{tar}$ Outputs:\n $\${out}_{tar}$ ”, where $\${inp}_{tar}$ and $\${out}_{tar}$ are the placeholders of inputs and outputs of a test case in the format of the target PL. In particular, for target PLs with statically typed characteristics, such as Java and C++, we complement tc_{tar} with input/output types, i.e., “Inputs:\n $\${inp}_{tar}^*$ Outputs ($\${type}_{tar}$):\n $\${out}_{tar}$ ”, where $\${inp}_{tar}^*$ is the placeholder of a test case's inputs with explicit variable types, while $\${type}_{tar}$ is the placeholder of outputs' type of a test case. Both of them are in the target PL's format.

<i>Input Generation Prompt</i>	<i>The Output of GPT3.5</i>
<pre>static int breakSum (int n) { if (n == 0 n == 1) return n ; return Math . max ((breakSum (n / 2) + breakSum (n / 3) + breakSum (n / 4)) , n) ; } </pre> <p>Please generate ten groups of differentiated valid inputs for the above focal method of Java language, in the format of [Input_1]\n[Input_2]\n...[Input_10]. Finally, use END_OF_CASE to finish your answer.</p>	<pre>Inputs1: int x=0; Inputs2: int x=1; Inputs3: int x=2; (other results ...) END_OF_CASE </pre>

Fig. 3. An Example of Input Generation Prompt

4.2 Translation Augmentation

After the Test Case Generation Phase, *Translation Augmentation Prompts* can be composed with prepared test cases and source programs (Step 5), as an example shown in Figure 4. The prompt template can be formally defined as “Given $\${pl}_{src}$ code:\n $\${prog}_{src}$ \nGiven test cases: $\${TC}_{tar}$ \nTranslate given $\${pl}_{src}$ code to $\${pl}_{tar}$ code, and ensure the translated $\${pl}_{tar}$ code can pass all given test cases. Use END_OF_CASE to finish your answer.”, where $\${TC}_{tar}$ consists of a series of prepared

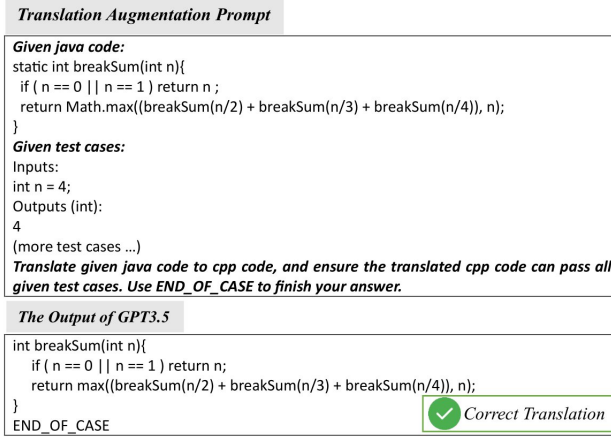


Fig. 4. An Example of Translation Augmentation Prompt

test cases tc_{tar} , $\{pl_{tar}\}$ is the placeholder for the name of the target PL. Hence, LLMs can leverage the above prompt to augment the code translation with extra test case information (Step 6). Afterward, we further execute the translated programs with prepared test cases to double-check their correctness as a preliminary inspection (Step 7). A program that passes all given test cases is deemed correct, and we return it to users. Otherwise, we will pipeline this sample to the last phase for repair. The Translation Augmentation Phase is listed in the middle of Figure 2.

4.3 Translation Repair

Translated programs once failed on certain test cases, their compilation/runtime errors will be thrown via the stack backtraces, while logic errors will list the explicit discrepancies between the expected outputs and actual outputs to identify the output inconsistencies. To this end, as shown in the right-hand side of Figure 2, we design an Error Analyzer (EA) to extract error information from the above Execution Results (Step 8), including error lines and error messages. Taking the buggy Java program in Figure 5 as an example, Figure 6 is its Execution Result throwing a compilation error. Following our designed regular expressions, EA can extract the error line number (“11”) and error message (“cannot find symbol”) in Line 2, thereby providing necessary information in the TRP for program repair (highlighted in yellow). It should be noted that, as each translated program needs to be inserted into the testing template file (e.g., tmp.java and tmp.py, refer to [2] for details) for evaluation, “11” here is the buggy line number in tmp.java rather than in the translated program. Thus, we must subtract 6 prefix lines of our translated program (i.e., $11-6=5$) in tmp.java. As such, we can precisely specify the buggy line in the translated program, as shown in Figure 5. Besides, Execution Results of different PLs and errors (e.g., compilation, runtime, and logic errors) are slightly different. To that end, we tailor different regular expressions towards various PLs and error types in EA and have published them in [2].

Afterward, the extracted information can be embedded into the *Repair Prompts*, facilitating LLMs to fix corresponding error programs and refine the translation (Step 9-10). The prompt for compilation/runtime error can be formally defined as “Given buggy $\{pl_{tar}\}$ code:
 $\{prog_{tar}^*\}$
Given test case:
 $\{tc_{tar}^*\}$
Error message: $\{err_msg\}$
Fix the buggy line (marked $\{com_sym_{tar}\}$ <Buggy Line>) in the buggy $\{pl_{tar}\}$ code according to the given error message. Use END_OF_CASE to finish your answer.”, where $\{prog_{tar}^*\}$, $\{tc_{tar}^*\}$, $\{err_msg\}$, and $\{com_sym_{tar}\}$ are the placeholders of a translated program with a buggy line specified, a failed tc_{tar} , an error message, and the comment symbol of the target PL (e.g., “//” for Java and C++ while “#” for Python) in order. It should be

Repair Prompt	The Output of GPT3.5
<p>Given buggy java code:</p> <pre>static int findRotations(String str) { String tmp = str + str; int n = str.length(); for (int i = 1; i <= n; i++) { String substring = tmp.substring(i, str.size()); //<Buggy Line> if (str.equals(substring)) return i; } return n; } </pre> <p>Given test case: Inputs: String str = "computer"; Outputs (int): 8</p> <p>Error message: cannot find symbol Fix the buggy line (marked //<Buggy Line>) in the buggy java code according to the given error message. Use END_OF_CASE to finish your answer.</p>	<pre>static int findRotations(String str) { String tmp = str + str; int n = str.length(); for (int i = 1; i <= n; i++) { String substring = tmp.substring(i, str.length()); if (str.equals(substring)) return i; } return n; } END_OF_CASE </pre> <div style="text-align: right; margin-top: 10px;"> Correct Translation </div>

Fig. 5. An Example of Repair Prompt

noted that even though each translated program can be evaluated with more than one test case, we only feed the first failed case for repair due to the restriction of the prompt token length. The definition of prompt template of logic error is slightly different, which is: “Given buggy $\{pl_{tar}\}$ code:\n $\{prog_{tar}\}$ \nGiven test case:\n $\{tc_{tar}^*\}$ \nError message: $\{err_{msg}\}$ \nFix the buggy $\{pl_{tar}\}$ code according to the given error message. Use END_OF_CASE to finish your answer.”, where $\{prog_{tar}\}$ is the placeholder of an incorrectly translated program without a specified buggy line, as logic errors are hard to locate. Besides, it is also optional for users to repeatedly evaluate the repaired programs with auto-generated test cases to fetch updated error information and repair until reaching the pre-defined maximum iteration, namely iterative repair.

1	Compilation Error:
2	./cleaned_data/test_case_gen_scripts/tmp_scripts/tmp.java:11: error: cannot find symbol
3	String substring = tmp.substring(i, str.size());
4	^
5	symbol: method size()
6	location: variable str of type String
7	1 error
8	error: compilation failed

Fig. 6. An Execution Result Example

5 EXPERIMENTAL SETTING

This section elaborates on the whole experimental setting across this work, including the dataset, models, metrics, implementation, research questions, and evaluation methodology.

5.1 Data Cleaning

To assess the effectiveness of TransCoder, Roziere et al. [46] released a widely-used translation dataset composed of 948 parallel functions in Python, Java, and C++. However, only 568 of them contain evaluation-purpose unit tests for at least one PL, where 464 unit tests are for Python, 482 are for Java, and 467 are for C++. As such, we focus on this part of the dataset in our experiments. The dataset was collected from GeeksforGeeks [3], which is an online platform containing many coding problems and solutions in various PLs. However, based on our observation of the preliminary experiments on GPT-3.5, lots of failures are induced by data noises in unit test scripts or inconsistencies among parallel functions. To reveal the actual ability of different approaches, four authors with more than three years of programming experience in Python, Java, and C++ are involved in cleaning the dataset manually. All their corrected samples are required to be cross-checked by two participators, thereby assuring the data quality. For clarification, we divide those identified noises into different categories, such as logic inconsistency, runtime/compilation error, and unit test error. In summary, a total of 252 errors were found, where 132 errors in Python, 58 errors in Java, and 62 errors in C++. The detailed breakdown of these data noises is attached in our replication package due to the page limit. All these noises are eliminated after the data cleaning.

5.2 Studied Models

We introduce five recent LLMs of diverse sizes and families as well as three state-of-the-art learning-based transpilers for empirical study. Their detailed information is listed below.

- **GPT-3.5 [43]**: A set of models improved on GPT-3, fine-tuned with RLHF techniques for understanding human instructions and generating natural language and code. We use OpenAI's APIs to access its latest version, i.e., gpt-3.5-turbo-0613.
- **LLaMA [50]**: A family of multilingual large language models ranging from 7B to 65B in parameter size, trained from various open-source datasets of up to 1.4 trillion tokens, where 4.5% are code from Github. We include three versions of 7B/13B/33B for experiments.
- **CodeGen [41]**: A collection of language models trained on up to 354.7 billion English tokens and 222.5 billion code tokens. It also has diverse sizes in parameters ranging from 350M to 16.1B. During the experiment, we use codegen-6B-multi.
- **TransCoder [46]**: It is an unsupervised machine translation-based transpiler pre-trained with cross-lingual language modeling, denoising auto-encoding, and back-translation. As such, a vast amount of monolingual samples can be leveraged for training. TransCoder's translation ability covers Python, Java, and C++.
- **TransCoder-IR [48]**: It is an incremental work of TransCoder, introducing low-level compiler Intermediate Representation (IR) to improve the code translation performance. Along with the pre-training tasks of TransCoder, TransCoder-IR involves extra pre-training tasks, such as translation language modeling, translation auto-encoding, and IR generation. TransCoder-IR was trained for translation among Java, C++, Rust, and Go.
- **TransCoder-ST [47]**: It is another augmented version based on TransCoder, which leverages automatically generated test cases to filter out invalid translations from the back-translation process, thereby improving the performance of unsupervised machine translation approaches in the code translation task.

5.3 Evaluation Metrics

Following the previous studies in the code translation field [33, 46–48], we adopt Computational Accuracy (CA) and Exact Match Accuracy (EM Acc) to assess the performance of each model.

Exact Match Accuracy (EM Acc): It computes the ratio of translations that exactly match ground truths, which can be formally defined as:

$$EM\ Acc = \frac{\sum_{k=1}^N em(y_k, \hat{y}_k)}{N}, \text{ where } em(y_k, \hat{y}_k) = \begin{cases} 1 & y_k = \hat{y}_k \\ 0 & y_k \neq \hat{y}_k \end{cases} \quad (1)$$

where N denotes the total number of translation samples, y_k denotes the ground truth of the k -th sample, and \hat{y}_k denotes the translated program via a certain transpiler for the k -th sample. For example, given y_k and \hat{y}_k , only if they are exactly identical, they can be deemed a correct translation in EM Acc. Thus, EM Acc concludes the lower-bound of the effectiveness of transpilers.

Computational Accuracy (CA): It computes the ratio that the translated programs can produce the same execution result as the ground truths, given the same inputs. Thus, this metric considers the semantic equivalency of programs, which can be formally defined as:

$$CA = \frac{\sum_{k=1}^N ca(y_k, \hat{y}_k)}{N}, \text{ where } ca(y_k, \hat{y}_k) = \begin{cases} 1 & Exec_k(y_k) = Exec_k(\hat{y}_k) \\ 0 & Exec_k(y_k) \neq Exec_k(\hat{y}_k) \end{cases} \quad (2)$$

where N , y_k , and \hat{y}_k carry the same meaning as those in EM Acc, $Exec_k(\cdot)$ denotes the execution result of a program with the test suite of the k -th sample. For example, even if y_k and \hat{y}_k are not identical literally, they are considered a correct translation in CA, as long as $Exec_k(y_k) = Exec_k(\hat{y}_k)$.

5.4 Implementation

Regarding LLMs, we implement GPT-3.5 by invoking OpenAI’s API [7]. For open-source LLMs, such as LLaMA and CodeGen, we instantiate them with their replication packages and load their weights from HuggingFace [5]. The default settings of LLMs are the same, using nucleus sampling [24] with $top_p=0.95$, $temperature=0.8$, and 10 samples per translation. Besides, we design a *Basic Prompt* to evaluate LLMs’ code translation capability, which can be formally defined as “Given $\{pl_{src}\}$ code: $\backslash n\{prog_{src}\}\backslash n$ Translate given $\{pl_{src}\}$ code to $\{pl_{tar}\}$ code. Use END_OF_CASE to finish your answer.” An example of the *Basic Prompt* and a generated result is shown in Figure 7. In experiments, we adopt a one-shot learning setting to make every LLM understand more about the task and only evaluate their first outputs. For learning-based translators, we reproduce them using their published source code and load them with the optimal weights for each translation dataset. Following that, we fix their $beam_sizes=10$ and fetch their first outputs for re-evaluation during inference on the cleaned dataset.

Towards **UniTrans**, the settings of its backbone LLMs are kept as default. In the Translation Augmentation Phase, considering the trade-off between performance and efficiency, we randomly select 3 generated test cases to construct *Translation Augmentation Prompts*. Besides, in the Translation Repair Phase, we leverage LLMs to repair error programs for 1 round as the default setting, considering the efficiency issue.

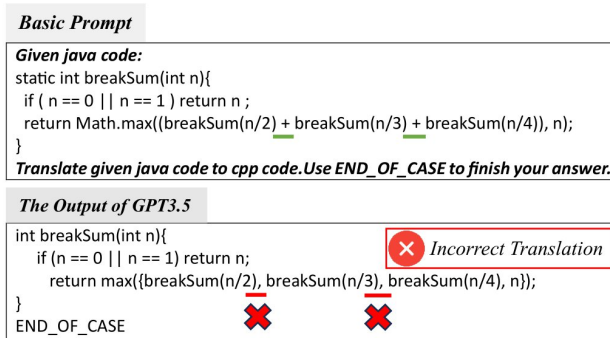


Fig. 7. Basic Prompt

5.5 Research Questions and Evaluation Methodology

Here, we introduce six research questions of this work and propose our evaluation methodology.

RQ1: What is the performance of recent LLMs against state-of-the-art learning-based transpilers in code translation? This RQ extensively assesses the code translation performance of recent LLMs. In this empirical study, we evaluate the models mentioned in Section 5.2 on four translation datasets, i.e., C++ to Python, Python to C++, Java to C++, and C++ to Java, leaving the remaining two datasets, i.e., Python to Java and Java to Python, intact so as to examine the generality of **UniTrans** on translation scenarios that have not been analyzed in follow-up experiments. As such, we can quantitatively explore the practicality of LLMs in the code translation task and make a fair comparison with state-of-the-art learning-based transpilers.

RQ2: How does UniTrans perform with different LLMs? This RQ aims to examine the effectiveness of our proposal and its generalization capability on various LLMs. In experiments, different LLMs of diverse sizes are selected for investigation, including LLaMA-7B, LLaMA-13B, and GPT-3.5. For experimental datasets, along with the four pairs of translations, i.e., C++ to Python, Python to C++, Java to C++, and C++ to Java, used in Section 6.2 (RQ1), we also newly include two translation datasets between Python and Java. Because the former four pairs of translation datasets have been analyzed in Sections 3 and 6.2, and we tailored specific methods to overcome

LLMs' failure on them, thereby proposing UniTrans. If we only conduct experiments on these four datasets, we cannot know if our proposal has generality or not to other translation pairs.

RQ3: What are the contributions of each component in UniTrans? UniTrans consists of two components to boost its code translation performance, i.e., (1) the Translation Augmentation Phase (TAP) and (2) the Translation Repair Phase (TRP). Based on the experimental setting of RQ2, we further deconstruct UniTrans and successively add the TAP and TRP to LLMs³, including LLaMA-7B, LLaMA-13B and GPT-3.5, to investigate their contributions in terms of CA and EM Acc on each translation dataset.

RQ4: How does the number of test cases influence the performance of UniTrans? Incorporating more test cases in the *Translation Augmentation Prompt* (i.e., in the TAP) indeed provides more references but occupies the valid prompt capacity and induces unexpected redundancy. Besides, this effect also carries certain impacts on the subsequent TRP, as potentially more vulnerabilities can be detected and repaired by adding more test cases. As such, to uncover an appropriate setting of this component, we investigate the performance of different LLMs with different numbers of test cases in this RQ. Due to the page limit, we consider Java as the target PL and divide four trials, where we randomly select {0, 1, 3, 5} Java test cases for each group, respectively, from the generated test case pool of each experimented LLM (i.e., LLaMA-7B and GPT-3.5). We do not continue sampling more test cases because the line coverage increment has been gradually leveling off, as tested by a Java coverage tool, namely EMMA [1]. Following that, we conduct experiments on the TAP and TRP, using diverse settings of test cases, and analyze the outcomes.

RQ5: What is the performance of LLMs on the valid input generation task? Efficiently generating valid inputs for programs facilitates the follow-up test case gathering, translation augmentation, and repair. Hence, this RQ aims to explore the performance of LLMs in generating valid inputs given limited attempts, thereby concluding the proper number of attempts for the guidance of practical deployment. Following the above experimental setting, we further discuss the LLMs' capabilities to generate at least {1, 3, 5} valid inputs for each PL (i.e., Python, Java, and C++) given limited attempts of inference (i.e., {1, 2, 3}). As we mentioned in Section 4.1, for each inference, LLMs are instructed to generate 10 candidate inputs with the *Input Generation Prompt*. Thus, they will totally generate {10, 20, 30} candidate inputs in theory for each trial. Both LLaMA-7B and GPT-3.5 with the default setting are included in this experiment.

RQ6: How do the rounds of iterative repair affect the performance of UniTrans? Iteratively repairing might bring the buggy code closer to the ground truths but also may lead it to degenerate. Besides, endlessly repairing incurs extra unaffordable time costs. Hence, this RQ intends to unveil an appropriate maximum repair iteration of UniTrans. For this experiment, we pre-define four maximum repair iterations (i.e., {0, 1, 2, 3}) and include two LLMs (i.e., LLaMA-7B and GPT-3.5). In detail, we instruct LLMs to consecutively repair 3 rounds and record their repair results sequentially on each translation dataset. Besides, we exclude EM Acc from the evaluation metrics because it underestimates the efficacy of the repair process and makes it hard to observe the performance differences, as discussed in Section 6.3. Meanwhile, CA also cannot capture all the repair details. For example, if a post-repair program can pass⁴ more unit tests than the pre-repair program, the repair should be deemed effective even if it cannot pass the whole test suite. Whereas CA does not consider such improvement.

Thus, Apart from CA, we also introduce a new evaluation metric, namely Pass Rate (PR), to assess the performance variation of iterative repair in a more fine-grained manner. Specifically, PR

³Before equipping with TAP and TRP, we have already employed the Test Case Generation Phase for LLMs to generate their respective test cases with the same setting as RQ2.

⁴Passing a unit test means the translated program can produce the same output as the ground truth program, given the same input.

measures the average percentage of unit tests that are passed by repaired programs. We consider PR as an auxiliary metric of CA in this experiment, as CA considers the whole test suite as an entity while PR can focus on individuals of the test suite. Therefore, when two counterparts achieve the same CA score, we use PR to distinguish their difference in the granularity of individual unit tests.

Table 2. Empirical Results of Each Model for Code Translation


Models	C++ to Python		Python to C++		Java to C++		C++ to Java		Average	
	CA	EM Acc	CA	EM Acc	CA	EM Acc	CA	EM Acc	CA	EM Acc
Large Language Models										
CodeGen	30.17%	4.40%	22.70%	1.94%	35.12%	7.57%	11.20%	2.46%	24.80%	4.09%
LLaMA-7B	23.49%	3.17%	16.49%	1.41%	32.98%	10.04%	21.16%	9.86%	23.53%	6.12%
LLaMA-13B	33.19%	4.05%	32.98%	2.46%	37.90%	10.39%	40.25%	12.15%	36.08%	7.26%
LLaMA-33B	47.63%	4.75%	38.12%	1.94%	59.53%	17.43%	45.23%	12.50%	47.63%	9.16%
GPT-3.5	87.07%	11.44%	89.51%	6.69%	92.93%	27.46%	82.16%	26.58%	87.92%	18.04%
Learning-based Transpilers										
TransCoder	36.64%	2.29%	30.40%	0.88%	27.84%	5.81%	49.77%	14.39%	36.16%	5.84%
TransCoder-IR	/	/	/	/	40.99%	14.26%	50.53%	18.28%	45.76%	16.27%
TransCoder-ST	46.34%	2.64%	47.85%	1.06%	49.68%	9.68%	64.73%	17.43%	52.15%	7.70%

^Φ For clarification, we adopt "-" to concatenate LLaMA and each of its corresponding parameter sizes for discrimination. For example, LLaMA with 7B parameters is dubbed LLaMA-7B.

6 EXPERIMENTAL RESULTS

6.1 RQ1: What is the performance of recent LLMs against state-of-the-art learning-based transpilers in code translation?

Table 2 demonstrates the empirical results of each model on four translation datasets in terms of CA and EM Acc. (1) As evident from the results, irrespective of the translation dataset experimented, GPT-3.5 consistently performs the best, achieving 87.92% in terms of CA and 18.04% in terms of EM Acc on average among four translation datasets. Besides, LLaMA-33B also performs better than the state-of-the-art learning-based transpilers on most translation datasets in terms of CA and EM Acc. (2) In general, with the increment of parameter size, the translation performance of LLMs is improved gradually. (3) Comparing CodeGen and LLaMA-7B with similar parameter sizes but different pre-training resources and model implementations, their translation abilities are shown to be neck-to-neck in performance, where, on average, CodeGen performs better in terms of CA while LLaMA-7B obtains a higher EM Acc score. (4) It is noticeable that TransCoder-IR outperforms TransCoder-ST in terms of EM Acc on average. In contrast, TransCoder-ST obtains a higher score in terms of CA on average. A potential explanation is TransCoder-ST trained with samples filtered by unit tests, which makes it comparatively insensitive to the lexical matching but focuses more on semantic equivalency.

 **Answer to RQ1:** LLMs have achieved impressive performance in the code translation task compared with state-of-the-art transpilers, showing a promising prospect. Nonetheless, they still suffer some accuracy issues. Even the GPT-3.5 cannot carry out the perfect translation, as highlighted by the multiple failures discussed in Section 3.

6.2 RQ2: How does UniTrans perform with different LLMs?

Table 3 demonstrates the experimental results of **UniTrans** embedded with various LLMs, where each row of "Improvement" denotes the improvement of **UniTrans** against its corresponding underlying LLM. As can be seen, **UniTrans** consistently boosts the performance of each LLM in the code translation task. Among all translation pairs, **UniTrans** brings improvements of 28.58% in

Table 3. Experimental Results of **UniTrans** with Different LLMs


Models	Java to Python		Python to Java		C++ to Python		Python to C++		Java to C++		C++ to Java		Average	
	CA	EM Acc	CA	EM Acc	CA	EM Acc	CA	EM Acc	CA	EM Acc	CA	EM Acc	CA	EM Acc
LLaMA-7B	27.16%	3.35%	15.15%	1.23%	23.49%	3.17%	16.49%	1.41%	32.98%	10.04%	21.16%	9.86%	22.74%	4.84%
UniTrans	31.90%	3.52%	23.24%	3.52%	26.08%	4.23%	20.34%	2.64%	35.76%	13.56%	33.40%	17.78%	28.45%	7.54%
Improvement	17.45%	5.07%	53.40%	186.18%	11.03%	33.44%	23.35%	87.23%	8.43%	35.06%	57.84%	80.32%	28.58%	71.22%
LLaMA-13B	39.01%	4.58%	36.31%	1.94%	33.19%	4.05%	32.98%	2.46%	37.90%	10.39%	40.25%	12.15%	36.61%	5.93%
UniTrans	42.89%	4.75%	38.80%	2.82%	40.73%	5.28%	43.25%	4.23%	47.54%	12.85%	47.93%	17.43%	43.52%	7.89%
Improvement	9.95%	3.71%	6.86%	45.36%	22.72%	30.37%	31.14%	71.95%	25.44%	23.68%	19.08%	43.46%	19.20%	36.42%
GPT-3.5	89.22%	10.74%	74.89%	6.16%	87.07%	11.44%	89.51%	6.69%	92.93%	27.46%	82.16%	26.58%	85.96%	14.85%
UniTrans	91.16%	11.44%	81.33%	8.10%	88.79%	12.68%	94.22%	7.57%	94.86%	29.41%	85.48%	29.40%	89.31%	16.43%
Improvement	2.17%	6.52%	8.60%	31.49%	1.98%	10.84%	5.26%	13.15%	2.08%	7.10%	4.04%	10.57%	4.02%	13.28%

[†] Table cells with light cyan background denote the highest improvement in terms of CA/EM Acc of each LLM among various translation datasets. Digits in **bold** denote the highest improvement in terms of CA/EM Acc among all LLMs and all results.

terms of CA and 71.22% in terms of EM Acc on average for LLaMA-7B, which is the most significant among all three experimented LLMs. Particularly, the translation of Python to Java and C++ to Java respectively achieves the highest improvement in terms of EM Acc (186.18%) and CA (57.84%) among all results. Besides, on average, it enhances LLaMA-13B by 19.20% and 36.42% in terms of CA and EM Acc, where the translation of Python to C++ obtains the highest enhancement for LLaMA-13B. Moreover, it also boosts GPT-3.5 by 4.02% and 13.28%, on average, in terms of each evaluation metric in order, and translating from Python to Java gains the most.

To investigate whether the improvement of **UniTrans** against each backbone is statistically significant, we perform the Wilcoxon Signed-Rank Test (WSRT) [52] with a confidence level of 95% to make pairwise comparisons for them across six translation datasets. Subsequently, we further conduct Cliff's Delta analysis [22] to measure the effect size in pairs. Results demonstrate that the improvements are all statistically significant and non-negligible⁵, which further proves the superior performance of **UniTrans** in unleashing the power of LLMs in code translation.

Although the translation between Python and Java was not extensively investigated in Section 3, **UniTrans** still improves each LLM's performance on these translation datasets significantly, showing that our idea has the potential to be extended to other unseen translation pairs, and the effectiveness is also remarkable.

 **Answer to RQ2:** **UniTrans** can bring consistent and substantial improvements for different LLMs of diverse parameter sizes on various translation datasets, showing a powerful generality on various models and programming language pairs.

6.3 RQ3: What are the contributions of each component in **UniTrans**?

We present Table 4 to demonstrate the contribution of TAP and TRP of **UniTrans** on various LLMs, where their positive improvements in terms of CA and EM Acc are highlighted in light cyan. In general, adding TAP and TRP are both effective in enhancing LLMs' code translation ability, where TAP brings average improvements on CA and EM Acc by 23.77% and 67.05% for LLaMA-7B, by 14.67% and 29.79% for LLaMA-13B, and by 2.06% and 13.02% for GPT-3.5. Meanwhile, adding TRP enhances the metrics of CA and EM Acc on average by 3.77% and 2.80% for LLaMA-7B, by 4.05% and 4.76% for LLaMA-13B, and by 1.93% and 0.25% for GPT-3.5. In summary, smaller models obtain more enhancement as their room for improvement is relatively larger.

It is noticeable that GPT-3.5 suffers a little performance decline in terms of CA when using TAP for the translation from Java to Python. We manually inspect those failed cases and find most of the mistakes are tiny errors, such as mis-interchanging operands and operator misuse. We speculate

⁵Detailed statistical test results are recorded in [2]

that LLMs still might overlook some details even with test cases for augmentation. On the other hand, GPT-3.5 has already obtained quite good performance, leading to little room for further improvement via test cases. However, since the performance decline is very tiny (-0.48%) and the EM Acc score here is improved (+4.93%), we argue adding test cases as extra information is harmless for GPT-3.5 to translate code from Java to Python.


In addition, it is also worth noting that, compared with TAP, introducing TRP brings less improvement to EM Acc, sometimes even zero. A potential explanation is that many repair processes are targeted to fix specific lines based on initial translated programs. Thus, once an initial translated program has a different implementation thought from the ground truth program, even the successfully fixed program will have difficulty matching the ground truth perfectly, causing the underestimation of the improvement. Hence, it is more important to consider the performance variation in terms of CA rather than EM Acc for TRP. Moreover, we find that TRP accounts for a larger proportion of the whole improvement in terms of CA with the increment of model size (LLaMA-7B: 3.77/23.77 < LLaMA-13B: 4.05/14.67 < GPT-3.5: 1.93/2.06). A potential explanation is larger models carry more powerful generality to program repair. Thus, even with less improvement room, they boost more than adding test cases as translation augmentation. Therefore, we stress that the TRP is also necessary for **UniTrans** and is mutually complementary to TAP.

Finally, following the procedure in RQ2, we further conduct WSRT and Cliff's Delta analysis to explore whether the improvements of TAP and TRP are significant. Results demonstrate sequentially adding TAP and TRP modules for each backbone obtains statistically significant improvements, and the effects are all non-negligible ⁶.

Table 4. Ablation Study of **UniTrans** with Different LLMs

Models	Java to Python		Python to Java		C++ to Python		Python to C++		Java to C++		C++ to Java		Average	
	CA	EM Acc	CA	EM Acc	CA	EM Acc	CA	EM Acc	CA	EM Acc	CA	EM Acc	CA	EM Acc
LLaMA-7B	27.16%	3.35%	15.15%	1.23%	23.49%	3.17%	16.49%	1.41%	32.98%	10.04%	21.16%	9.86%	22.74%	4.84%
+ TAP	31.25%	3.35%	22.20%	3.52%	25.43%	4.23%	19.49%	2.46%	34.48%	13.38%	31.74%	17.25%	27.43%	7.37%
Improvement	15.06%	0.00%	46.53%	186.18%	8.26%	33.44%	18.19%	74.47%	4.55%	33.27%	50.00%	74.95%	23.77%	67.05%
+ TRP	31.90%	3.52%	23.24%	3.52%	26.08%	4.23%	20.34%	2.64%	35.76%	13.56%	33.40%	17.78%	28.45%	7.54%
Improvement	2.08%	5.07%	4.68%	0.00%	2.56%	0.00%	4.36%	7.32%	3.71%	1.35%	5.23%	3.07%	3.77%	2.80%
LLaMA-13B	39.01%	4.58%	36.31%	1.94%	33.19%	4.05%	32.98%	2.46%	37.90%	10.39%	40.25%	12.15%	36.61%	5.93%
+ TAP	42.03%	4.75%	36.72%	2.64%	39.44%	4.93%	42.18%	3.87%	46.47%	12.85%	44.19%	16.55%	41.84%	7.60%
Improvement	7.74%	3.71%	1.13%	36.08%	18.83%	21.73%	27.90%	57.32%	22.61%	23.68%	9.79%	36.21%	14.67%	29.79%
+ TRP	42.89%	4.75%	38.80%	2.82%	40.73%	5.28%	43.25%	4.23%	47.54%	12.85%	47.93%	17.43%	43.52%	7.89%
Improvement	2.05%	0.00%	5.66%	6.82%	3.27%	7.10%	2.54%	9.30%	2.30%	0.00%	8.46%	5.32%	4.05%	4.76%
GPT-3.5	89.22%	10.74%	74.89%	6.16%	87.07%	11.44%	89.51%	6.69%	92.93%	27.46%	82.16%	26.58%	85.96%	14.85%
+ TAP	88.79%	11.27%	79.67%	8.10%	87.28%	12.68%	92.72%	7.57%	94.43%	29.41%	82.99%	29.39%	87.65%	16.40%
Improvement	-0.48%	4.93%	6.38%	31.49%	0.24%	10.84%	3.59%	13.15%	1.61%	7.10%	1.01%	10.57%	2.06%	13.02%
+ TRP	91.16%	11.44%	81.33%	8.10%	88.79%	12.68%	94.22%	7.57%	94.86%	29.41%	85.48%	29.40%	89.31%	16.43%
Improvement	2.67%	1.51%	2.08%	0.00%	1.73%	0.00%	1.62%	0.00%	0.46%	0.00%	3.00%	0.00%	1.93%	0.25%

^Φ TAP and TRP are the abbreviations of the Translation Augmentation Phase and Translation Repair Phase, which is the same as Table 5, 7 and 8.

 **Answer to RQ3:** Each component, i.e., (1) the Translation Augmentation Phase and (2) the Translation Repair Phase, is essential for **UniTrans**. Three LLMs have been progressively improved with the addition of each component gradually.

⁶Detailed statistical test results are recorded in [2]. It should be noted that we did not conduct the statistical tests between TRP and TAP in terms of EM Acc because, as we analyzed in Section 6.3, EM Acc can hardly manifest the improvement of TRP against TAP inherently.

Table 5. Test Case Number Analysis for UniTrans with LLaMA-7B/GPT-3.5

# Test Cases	Line Coverage	Python to Java (TAP)		Python to Java (TRP)		C++ to Java (TAP)		C++ to Java (TRP)	
		CA	EM Acc	CA	EM Acc	CA	EM Acc	CA	EM Acc
UniTrans with LLaMA-7B									
0	0	15.15%	1.23%	/	/	21.16%	9.86%	/	/
1	78.84%	21.58%	3.52%	23.03%	3.70%	29.88%	18.13%	30.91%	18.49%
3	85.20%	22.20%	3.52%	23.24%	3.52%	31.74%	17.25%	33.40%	17.78%
5	87.20%	20.33%	2.99%	21.58%	2.99%	27.59%	16.20%	29.67%	17.43%
UniTrans with GPT-3.5									
0	0	74.89%	6.16%	/	/	82.16%	26.58%	/	/
1	82.57%	79.25%	7.92%	80.91%	7.80%	82.34%	25.53%	84.54%	25.63%
3	92.14%	79.67%	8.10%	81.33%	8.10%	82.99%	29.39%	85.48%	29.40%
5	94.50%	76.35%	8.63%	79.38%	8.45%	79.88%	28.35%	82.76%	28.87%

^Φ We label the translation pair with "(TAP)" or "(TRP)" to indicate their performance after each phase. Since it is impossible to identify error programs without test cases, we do not apply TRP for trials with zero test cases.

6.4 RQ4: How does the number of test cases influence the performance of UniTrans?

Table 5 demonstrates the experimental results of test case number analysis on the TAP and TRP of UniTrans, respectively. Initially, with the increment of the test case sampling budget, line coverage rates gradually increase, and the performance of LLMs is also progressively enhanced, which means more test cases can provide a broader vision for LLMs to comprehend the code. However, the improvements in both models do not last long. LLaMA-7B achieves its best when given test cases no more than 3, while GPT-3.5 can further improve its performance in some situations with 5 test cases. A potential explanation for this variation is even though more test cases offer higher code coverage rates and more chances to learn the code, the marginal returns are gradually decreasing. In this case, apart from the benefits mentioned above, excessive test cases likewise introduce more redundancies, leading to both backbone models being distracted to different extents in the Translation Augmentation Phase, which further hinders the improvement of the Translation Repair Phase.


 **Answer to RQ4:** Incorporating more test cases indeed offers a more extensive vision to LLMs to translate code, but it also brings redundancies and distracts LLMs. Based on the experimental results above, we recommend larger LLMs can be fed with relatively more test cases, while smaller LLMs are suggested to offer fewer test cases.

Table 6. Valid Input Generation Analysis for UniTrans with LLaMA-7B/GPT-3.5

# Valid Inputs	Python			Java			C++			Average		
	1	2	3	1	2	3	1	2	3	1	2	3
UniTrans with LLaMA-7B												
1	<i>70.42%</i>	84.51%	95.07%	78.52%	89.79%	97.01%	69.37%	84.68%	95.77%	72.77%	86.33%	93.84%
3	66.55%	82.39%	92.78%	73.24%	87.15%	95.95%	63.73%	79.75%	92.08%	67.84%	83.10%	92.31%
5	63.38%	79.40%	91.55%	68.13%	85.04%	94.19%	59.15%	74.82%	88.38%	63.55%	79.75%	91.26%
UniTrans with GPT-3.5												
1	96.48%	97.01%	97.36%	96.83%	97.18%	97.18%	78.52%	83.80%	89.44%	90.61%	92.66%	96.77%
3	94.89%	97.01%	97.18%	96.30%	97.01%	97.18%	78.52%	83.80%	88.20%	89.90%	92.61%	95.48%
5	94.36%	96.48%	97.01%	95.95%	97.01%	97.01%	78.52%	83.80%	88.03%	89.61%	92.43%	94.13%

^Φ # Valid Inputs denotes the number of valid inputs measured per experiment, where each experiment carries out {1, 2, 3} attempts of inference.

6.5 RQ5: What is the performance of LLMs on the valid input generation task?

Table 6 records the rate of generating at least {1, 3, 5} valid inputs given {1, 2, 3} attempts of inference. For example, *70.42%* (highlighted in italics in Table 6) denotes LLaMA-7B can generate at

least one valid input for 70.42% of programs given one attempt of inference. Apparently, GPT-3.5 outperforms LLaMA-7B in most of the settings, as it carries more parameters and a more powerful generality. In the meantime, when provided with only three attempts of inference for each of the two models, GPT-3.5 is capable of generating at least 1/3/5 valid inputs for an average of 96.77%/95.48%/94.13% of programs, while LLaMA-7B achieves this for 93.84%/92.31%/91.26% of programs on average, showing that the LLMs' capabilities of valid input generation for programs are highly efficient. As obtaining valid inputs is equivalent to obtaining test cases, the above results further demonstrate the high efficiency of our proposed Test Case Generation Phase, proving the practicability of **UniTrans** for future deployment.


 **Answer to RQ5:** LLMs with various parameter sizes are capable of generating valid inputs for over 90% programs with few attempts, showing that **UniTrans**, a code translation framework that is highly dependent on test cases, is feasible and efficient in practice.

Table 7. Iterative Repair Analysis for **UniTrans** with LLaMA-7B/GPT-3.5

# Max Iter.	Java to Python		Python to Java		C++ to Python		Python to C++		Java to C++		C++ to Java		Average	
	CA	PR	CA	PR	CA	PR	CA	PR	CA	PR	CA	PR	CA	PR
UniTrans with LLaMA-7B														
+ Repair-0	31.25%	38.66%	22.20%	25.81%	25.43%	33.23%	19.49%	22.31%	34.48%	36.02%	31.74%	32.12%	27.43%	31.36%
+ Repair-1	31.90%	39.63%	23.24%	28.15%	26.08%	34.07%	20.34%	23.47%	35.76%	37.32%	33.40%	33.92%	28.45%	32.76%
+ Repair-2	32.11%	39.94%	23.65%	28.55%	26.29%	34.14%	20.56%	23.92%	35.97%	37.41%	33.82%	34.69%	28.73%	33.11%
+ Repair-3	31.82%	39.39%	24.38%	29.14%	26.94%	34.72%	20.12%	23.79%	36.19%	37.79%	34.02%	34.90%	28.91%	33.29%
UniTrans with GPT-3.5														
+ Repair-0	88.79%	91.87%	79.67%	81.84%	87.28%	91.12%	92.72%	95.76%	94.43%	95.89%	82.99%	83.88%	87.65%	90.06%
+ Repair-1	91.16%	93.73%	81.33%	82.93%	88.79%	92.41%	94.22%	96.81%	94.86%	95.87%	85.48%	86.51%	89.31%	91.38%
+ Repair-2	91.16%	93.77%	81.74%	83.88%	89.66%	93.00%	94.65%	96.92%	95.72%	96.83%	85.89%	86.72%	89.80%	91.85%
+ Repair-3	90.95%	93.51%	81.74%	83.80%	89.66%	93.02%	94.43%	96.92%	95.93%	96.40%	85.89%	86.54%	89.77%	91.70%

† + Repair- $\{1, 2, 3\}$ denotes iteratively evaluating the repaired programs and repeating the TRP for $\{1, 2, 3\}$ times in **UniTrans**. In particular, + Repair-0 denotes no TRP when constructing **UniTrans**.

6.6 RQ6: How do the rounds of iterative repair affect the performance of **UniTrans**?

Table 7 presents the experimental results of **UniTrans** with LLaMA-7B/GPT-3.5 under various maximum repair iterations in terms of CA and PR. For both LLaMA-7B and GPT-3.5, we highlight their best-performing repair iteration for each translation pair in light cyan, while the decisive metrics are highlighted in bold, following the evaluation criterion presented in Section 5.5. As can be seen, with more repair attempts, regardless of LLaMA-7B or GPT-3.5, the marginal returns of the performance decline gradually. In most situations, GPT-3.5 achieves its best performance when repeating the repair 2 times, while LLaMA-7B can further improve its performance even if the repair has been repeated 3 times, which is also consistent with their average performance. The explanation is intuitive, as GPT-3.5, with a way larger parameter size, carries a much more powerful generality to program repair than LLaMA-7B, leading it to fix the bugs with fewer steps. On the other hand, The remaining buggy code for GPT-3.5 is less than LLaMA-7B. Thus, there is also less room for GPT-3.5 to further improve. In addition, **UniTrans** only repairs those true error programs as they are filtered out by auto-generated test cases. Thus, their performance should be no change at worst in terms of CA in theory, whereas some of their performance, in fact, declines a little bit with more rounds of repair, such as repairing in the 3rd round with LLaMA-7B/GPT-3.5 for the translation dataset of Python to C++. We manually examine those affected samples and come to our conclusion: some error programs cannot be revealed by the evaluation-purpose test suite but are identified by our auto-generated test cases. However, **UniTrans** cannot correctly fix them.

In contrast, it makes them further away from the ground truths, leading to performance declines based on this set of evaluation metrics.


 **Answer to RQ6:** Iteratively repairing error programs can fix more bugs but also bring the risk of degeneration. We suggest practitioners using **UniTrans** with larger LLMs to perform fewer rounds of repair, while for smaller LLMs, we recommend performing repair for more rounds.

Table 8. The Efficacy of **UniTrans** on Various Categories of Failures, Taking GPT-3.5 as An Example

Statistics	Logic	Syntax	I/O	API	Precision	Others
Total Failures	67	38	26	14	20	9
Improvement by TAP	23 (34.32%)	17 (44.74%)	12 (46.15%)	5 (35.71%)	3 (15%)	4 (44.44%)
Total Failures after TAP	44	21	14	9	17	5
Improvement by TRP	8 (18.18%)	10 (47.62%)	4 (28.57%)	1 (11.11%)	2 (11.76%)	2 (40%)
Total Improvement	31 (46.27%)	27 (71.05%)	16 (61.54%)	6 (42.86%)	5 (25%)	6 (66.67%)

^φ Digits in light cyan and bold denote the most failure types that TAP/TRP addresses in quantity and proportion, respectively.

7 DISCUSSION

7.1 Case Study

In this section, we list two examples to qualitatively present the advantage of adding test cases for augmenting code translation and repairing error programs. (1) Figure 7 demonstrates a failed case of using *Basic Prompt*, where GPT-3.5 cannot fully capture the logic of the source program and make a Logic failure. In contrast, when we add several test cases for augmentation, as shown in Figure 4, GPT-3.5 rectifies its previous mistake and completes the translation correctly. (2) Another example can be found in Figure 5, a code translation sample has been shipped to the Translation Repair Phase due to its failure in the Translation Augmentation Phase. As can be seen, although given several test cases for translation augmentation, GPT-3.5 still cannot correctly translate the source program to the target program due to the Java API misuse, i.e., Java uses “length()” to get lengths of strings, rather than “size()”. By identifying the buggy lines and pointing out the error message, GPT-3.5 successfully repairs the error program.

7.2 Failure Taxonomy Look Back

As mentioned in Section 3, we partition LLMs’ failures into six categories, taking GPT-3.5 as an example. This section looks back at the failure taxonomy and discusses the enhancement of **UniTrans** against GPT-3.5 on each failure type so as to summarize the future direction. As seen from the results in Table 8, TAP solves 46.15% of I/O failures and 23 Logic failures, which are the highest in terms of proportion and quantity, respectively, among various failures. It demonstrates adding test cases with I/O types for translation augmentation is effective for LLMs’ to comprehend program logic and the requirement on I/O. Moreover, TAP also has a commendable ability to resolve other mistakes, such as Syntax (44.74%) and API (35.71%). For the remaining mistakes, TRP further repairs them based on the translated programs in TAP. Apparently, it exhibits the most superior ability in handling Syntax failures, where 10 such failures are addressed, accounting for 47.62% of the various types of mistakes. In general, both TAP and TRP are capable of resolving those mistakes as we expected, but they still fail in certain cases, especially on Precision failures (only 25% are solved in total). We speculate that these mistakes are so tiny that they result in only minor discrepancies with the ground truth programs, which would be extremely imperceptible to an LLM. As such, we emphasize that one of the future directions for code translation with LLMs is how to make LLMs perceive those minor discrepancies, thereby further improving the reliability of automated code translation.

7.3 Threats to Validity

In this section, we categorize threats to the validity of this work from three aspects.

Internal Validity: One threat to the internal validity comes from the potential data leakage of LLMs, which means there may be some overlaps between the training set of LLMs and the testing set used in this work. However, according to our compared learning-based transpilers [46–48] and open-source LLMs [41, 50], although those LLMs were also pre-trained on natural languages, as for training resources on code, all the above models were pre-trained based on GitHub public dataset available on Google BigQuery⁷. However, the evaluation dataset crafted by [46–48] was extracted from an independent online platform, namely GeeksforGeeks [3], and [46–48] claimed that there is no data overlap between the pre-training and evaluation datasets. As for GPT-3.5, a closed-source LLM, it is hard to know their pre-training resources. An alternative way to inspect the overlap between the training and testing set is to evaluate GPT-3.5 on a new testing set with minimal leakage possibility and check the performance variation between the original and new testing sets. To do this, we curated 82 C++ codes manually written by some undergraduate students. These codes were submitted via a university’s private Online Judge (OJ) platform from Nov. 2023 to Jan. 2024. Thus, the OJ platform is invisible from outside the university. Besides, GPT-3.5-turbo-0613 (we used version) is a snapshot before June 13th, 2023. Therefore, it cannot exploit codes written after its release date for training. In summary, this evaluation dataset’s possibility of data leakage is minimal. Afterward, we leverage GPT-3.5 to translate our collected 82 C++ code to Java and Python with a one-shot setting (same as the default setting in our manuscript), respectively. As the OJ platform already has the Java and Python test suites for evaluation but without the ground truth for comparison, we only assess GPT-3.5 in terms of CA, which is 81.71% on C++ to Python translation and 80.49% on C++ to Java translation. Compared to the experimental results in Section 6.1, the performance drops 6.16% on C++ to Python while drops 2.03% on C++ to Java. Considering the code snippets of the newly crafted dataset are longer than the C++ dataset used in our manuscript (17.23 v.s. 12.94), the performance’s slight degeneration is acceptable. Hence, we conclude that the threat to the training/testing overlap is very limited. All evaluation results and datasets used in this section can be found in [2].

Besides, the replication of the baselines and evaluation metrics are also two threats to the internal validity. To minimize these threats, we strictly followed their replication documents and directly utilized their source code for model construction. All baselines are consistently evaluated on our cleaned dataset. As for the replication of evaluation metrics, e.g., CA, EM Acc, and PR, we reused the published code from [30, 46] to implement them.

External Validity: The first threat to the external validity lies in the quality of the evaluation dataset, which was released by [46], and has been widely used in the code translation field owing to its multi-lingual samples and self-contained test suites. However, the dataset is flawed, as we mentioned in Section 5.1, and manually cleaning the dataset may not eliminate all noises or errors. To address this, we required the four authors to cross-check their cleaned results in pairs to reach a consensus for each group of parallel functions. Furthermore, we release the cleaned dataset for public evaluation.

The limited choice of experimental models is another threat. In this work, we experimented with the most widely examined LLMs in different families and sizes [12, 32, 54, 59]. For example, we select open-source LLaMA-7B/13B/33B to make a comparison among LLMs of the same family but different sizes, and we also include CodeGen-6B to compare with LLaMA-7B to explore the influence between different LLM families. Furthermore, we introduce GPT-3.5 for experiments because it is a typical closed-source LLM. More importantly, all the above selected LLMs can generate code

⁷<https://console.cloud.google.com/marketplace/details/github/github-repos>

given specific requirements, ensuring the accomplishment of the experiment. Besides, we select TransCoder, TransCoder-IR, and TransCoder-ST for comparison, as they are the state-of-the-art transpilers to date, ensuring an in-depth investigation and analysis. Therefore, we believe this threat is minimal and will include more relevant models for experiments in the future.

Construct Validity: The property of evaluation metrics is a primary threat [58, 60, 63]. In this work, we adopt CA and EM Acc to evaluate the code translation performance of different approaches from both lexical and semantic correctness as most of the relevant papers [33, 46–48]. Besides, we also include PR to conduct the measurement in a more fine-grained manner [30]. Therefore, we believe the evaluation is comprehensive.

8 CONCLUSION

This work aims to enhance the efficiency and reliability of codebase migration. To achieve this, we propose to leverage LLMs to substitute learning-based transpilers. Hence, we first conduct an extensive empirical study and an in-depth analysis to investigate the strengths and weaknesses of recent LLMs compared with current transpilers. Enlightened by our findings in the empirical study, we design a **Unified code Translation** framework for diverse LLMs, namely **UniTrans**, which incorporates test cases to augment code translation and repair bugs for incorrectly translated programs. Comprehensive experiments are carried out and demonstrate **UniTrans** improves various LLMs' capabilities in code translation by a significant margin. Below, we summarize the implications of this work for practitioners and researchers, respectively.

Implications for practitioners: **UniTrans** is an automated code translation framework that can effectively boost the code translation performance of various LLMs, as substantiated in Section 6.2. It exempts large-scale fine-tuning from LLMs and only needs to use a series of well-designed prompts and the execution of auto-generated test cases to effectively improve the performance of LLMs. Notably, Practitioners are only required to designate a handful of hyperparameters, such as the number of test cases in TAP and the round of repair in TRP, then **UniTrans** can be encapsulated into a complete system for codebase migration, showing its high practical utility and aligning with contemporary demands for efficiency and resource optimization in code translation processes.

Implications for researchers: This work empirically explored the prospects and limitations of recent LLMs in automated code translation. Subsequently, we conducted the inaugural failure taxonomy grounded in the translation results of the best-performing LLM and summarized a series of improvement directions. Researchers can further propose more effective methods to unleash the power of LLMs or even re-construct them based on our insightful findings. While the prompts meticulously devised in each phase of the **UniTrans** have been tailored for optimal performance, the possibility persists that alternative prompt designs may exist with superior efficacy. Thus, we call for researchers to explore more about prompt design based on the three-step (i.e., Test Case Generation Phase, Translation Augmentation Phase, and Translation Repair Phase) framework we proposed, thereby advancing the state-of-the-art in automated code translation.

ACKNOWLEDGMENTS

This work was partially supported by National Key R&D Program under Grant No.2023YFB4503801, National Natural Science Foundation of China (Grant No. 62102233, 62302021, 62192731, 62192730, 62072007, 62192733, 61832009), Shandong Province Overseas Outstanding Youth Fund (Grant No. 2022HWYQ-043), Qilu Young Scholar Program of Shandong University, the General Research Fund (GRF) of the Research Grants Council of Hong Kong, the industry research funds of City University of Hong Kong (7005217,9220097,9220103,9229029,9229098,9678149), and the Key Program of Hubei under Grant JD2023008.

REFERENCES

- [1] [n. d.]. EMMA: a free Java code coverage tool. <https://emma.sourceforge.net/>. (Accessed on 05/06/2024).
- [2] [n. d.]. FSE-24-UniTrans. <https://github.com/yz1019117968/FSE-24-UniTrans>. (Accessed on 04/19/2024).
- [3] [n. d.]. GeeksforGeeks. <https://www.geeksforgeeks.org/>. (Accessed on 05/06/2024).
- [4] [n. d.]. gotranspile/cxgo: Tool for transpiling C to Go. <https://github.com/gotranspile/cxgo>. (Accessed on 05/06/2024).
- [5] [n. d.]. Hugging Face – The AI community building the future. <https://huggingface.co/>. (Accessed on 05/06/2024).
- [6] [n. d.]. immunant/c2rust: Migrate C code to Rust. <https://github.com/immunant/c2rust>. (Accessed on 05/06/2024).
- [7] [n. d.]. platform.openai.com. <https://platform.openai.com/docs/models/gpt-3-5>. (Accessed on 05/06/2024).
- [8] Toufique Ahmed and Premkumar Devanbu. 2023. Few-Shot Training LLMs for Project-Specific Code-Summarization. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering* (Rochester, MI, USA) (ASE '22). Association for Computing Machinery, New York, NY, USA, Article 177, 5 pages. <https://doi.org/10.1145/3551349.3559555>
- [9] Miltiadis Allamanis, Earl T Barr, Premkumar Devanbu, and Charles Sutton. 2018. A survey of machine learning for big code and naturalness. *ACM Computing Surveys (CSUR)* 51, 4 (2018), 1–37. <https://doi.org/10.1145/3212695>
- [10] Johannes Bader, Andrew Scott, Michael Pradel, and Satish Chandra. 2019. Getafix: Learning to fix bugs automatically. *Proceedings of the ACM on Programming Languages* 3, OOPSLA (2019), 1–27. <https://doi.org/10.1145/3360585>
- [11] Saikat Chakraborty, Yangruibo Ding, Miltiadis Allamanis, and Baishakhi Ray. 2022. CODIT: Code Editing With Tree-Based Neural Models. *IEEE Transactions on Software Engineering* 48, 4 (2022), 1385–1399. <https://doi.org/10.1109/TSE.2020.3020502>
- [12] Bei Chen, Fengji Zhang, Anh Nguyen, Daoguang Zan, Zeqi Lin, Jian-Guang Lou, and Weizhu Chen. 2022. CodeT5: Code Generation with Generated Tests. In *The Eleventh International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.2207.10397>
- [13] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021). <https://doi.org/10.48550/arXiv.2107.03374>
- [14] Xinyun Chen, Chang Liu, and Dawn Song. 2018. Tree-to-tree neural networks for program translation. *Advances in neural information processing systems* 31 (2018). <https://doi.org/10.48550/arXiv.1802.03691>
- [15] Zhenpeng Chen, Yanbin Cao, Yuanqiang Liu, Haoyu Wang, Tao Xie, and Xuanzhe Liu. 2020. A comprehensive study on challenges in deploying deep learning based software. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 750–762. <https://doi.org/10.1145/3368089.3409759>
- [16] Elizabeth Dinella, Gabriel Ryan, Todd Mytkowicz, and Shuvendu K. Lahiri. 2022. TOGA: A Neural Method for Test Oracle Generation. In *Proceedings of the 44th International Conference on Software Engineering* (Pittsburgh, Pennsylvania) (ICSE '22). Association for Computing Machinery, New York, NY, USA, 2130–2141. <https://doi.org/10.1145/3510003.3510141>
- [17] Yihong Dong, Xue Jiang, Zhi Jin, and Ge Li. 2023. Self-collaboration Code Generation via ChatGPT. *arXiv preprint arXiv:2304.07590* (2023). <https://doi.org/10.1145/3672459>
- [18] Zhiyu Fan, Xiang Gao, Martin Mirchev, Abhik Roychoudhury, and Shin Hwei Tan. 2023. Automated Repair of Programs from Large Language Models. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. 1469–1481. <https://doi.org/10.1109/ICSE48619.2023.00128>
- [19] Gordon Fraser and Andrea Arcuri. 2011. Evosuite: automatic test suite generation for object-oriented software. In *Proceedings of the 19th ACM SIGSOFT symposium and the 13th European conference on Foundations of software engineering*. 416–419. <https://doi.org/10.1145/2025113.2025179>
- [20] Mingyang Geng, Shangwen Wang, Dezun Dong, Haotian Wang, Ge Li, Zhi Jin, Xiaoguang Mao, and Xiangke Liao. 2024. Large language models are few-shot summarizers: Multi-intent comment generation via in-context learning. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*. 1–13. <https://doi.org/10.1145/3597503.3608134>
- [21] Ali Ghanbari, Samuel Benton, and Lingming Zhang. 2019. Practical program repair via bytecode mutation. In *Proceedings of the 28th ACM SIGSOFT International Symposium on Software Testing and Analysis*. 19–30. <https://doi.org/10.1109/ase.2019.00116>
- [22] Robert J Grissom and John J Kim. 2005. *Effect sizes for research: A broad practical approach*. Lawrence Erlbaum Associates Publishers. <https://doi.org/10.1198/tech.2006.s437>
- [23] Rahul Gupta, Soham Pal, Aditya Kanade, and Shirish Shevade. 2017. Deepfix: Fixing common c language errors by deep learning. In *Proceedings of the aaai conference on artificial intelligence*, Vol. 31. <https://doi.org/10.1609/aaai.v31i1.10742>
- [24] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.1904.09751>
- [25] Svetoslav Karaivanov, Veselin Raychev, and Martin Vechev. 2014. Phrase-based statistical translation of programming languages. In *Proceedings of the 2014 ACM International Symposium on New Ideas, New Paradigms, and Reflections on*

- Programming & Software*. 173–184. <https://doi.org/10.1145/2661136.2661148>
- [26] JWKJW Kotrlík and CCHCC Higgins. 2001. Organizational research: Determining appropriate sample size in survey research. *Information technology, learning, and performance journal* 19, 1 (2001), 43.
- [27] Claire Le Goues, Michael Dewey-Vogt, Stephanie Forrest, and Westley Weimer. 2012. A systematic study of automated program repair: Fixing 55 out of 105 bugs for \$8 each. In *2012 34th International Conference on Software Engineering (ICSE)*. IEEE, 3–13. <https://doi.org/10.1109/icse.2012.6227211>
- [28] Caroline Lemieux, Jeevana Priya Inala, Shuvendu K Lahiri, and Siddhartha Sen. 2023. CODAMOSA: Escaping coverage plateaus in test generation with pre-trained large language models. In *International conference on software engineering (ICSE)*. <https://doi.org/10.1109/icse48619.2023.00085>
- [29] Jia Li, Yongmin Li, Ge Li, and Zhi Jin. 2023. Structured Chain-of-Thought Prompting for Code Generation. *CoRR* (2023). <https://doi.org/10.48550/arXiv.2305.06599>
- [30] Jia Li, Yongmin Li, Ge Li, Zhi Jin, Yiyang Hao, and Xing Hu. 2023. SKCODER: A Sketch-Based Approach for Automatic Code Generation. In *Proceedings of the 45th International Conference on Software Engineering (Melbourne, Victoria, Australia) (ICSE '23)*. IEEE Press, 2124–2135. <https://doi.org/10.1109/ICSE48619.2023.00179>
- [31] Jia Li, Yunfei Zhao, Yongmin Li, Ge Li, and Zhi Jin. 2023. AceCoder: Utilizing Existing Code to Enhance Code Generation. *CoRR* (2023). <https://doi.org/10.48550/arXiv.2303.17780>
- [32] Jia Li, Yunfei Zhao, Yongmin Li, Ge Li, and Zhi Jin. 2023. Towards Enhancing In-Context Learning for Code Generation. *arXiv preprint arXiv:2303.17780* (2023). <https://doi.org/10.48550/arXiv.2303.17780>
- [33] Fang Liu, Jia Li, and Li Zhang. 2023. Syntax and Domain Aware Model for Unsupervised Program Translation. In *Proceedings of the 45th International Conference on Software Engineering (Melbourne, Victoria, Australia) (ICSE '23)*. IEEE Press, 755–767. <https://doi.org/10.1109/ICSE48619.2023.00072>
- [34] Fang Liu, Yang Liu, Lin Shi, Houkun Huang, Ruifeng Wang, Zhen Yang, and Li Zhang. 2024. Exploring and Evaluating Hallucinations in LLM-Powered Code Generation. *arXiv preprint arXiv:2404.00971* (2024). <https://doi.org/10.48550/arXiv.2404.00971>
- [35] Matias Martinez and Martin Monperrus. 2016. Astor: A program repair library for java. In *Proceedings of the 25th International Symposium on Software Testing and Analysis*. 441–444. <https://doi.org/10.1145/2931037.2948705>
- [36] Noor Nashid, Mifta Sintaha, and Ali Mesbah. 2023. Retrieval-based prompt selection for code-related few-shot learning. In *Proceedings of the 45th International Conference on Software Engineering (ICSE '23)*. <https://doi.org/10.1109/icse48619.2023.00205>
- [37] Anh Tuan Nguyen, Tung Thanh Nguyen, and Tien N Nguyen. 2013. Lexical statistical machine translation for language migration. In *Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering*. 651–654. <https://doi.org/10.1145/2491411.2494584>
- [38] Anh Tuan Nguyen, Tung Thanh Nguyen, and Tien N. Nguyen. 2015. Divide-and-Conquer Approach for Multi-Phase Statistical Migration for Source Code. In *Proceedings of the 30th IEEE/ACM International Conference on Automated Software Engineering (Lincoln, Nebraska) (ASE '15)*. IEEE Press, 585–596. <https://doi.org/10.1109/ASE.2015.74>
- [39] Anh Tuan Nguyen, Tung Thanh Nguyen, and Tien N Nguyen. 2015. Divide-and-conquer approach for multi-phase statistical migration for source code (t). In *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 585–596. <https://doi.org/10.1109/ase.2015.74>
- [40] Trong Duc Nguyen, Anh Tuan Nguyen, and Tien N Nguyen. 2016. Mapping API elements for code migration with vector representations. In *Proceedings of the 38th international conference on software engineering companion*. 756–758. <https://doi.org/10.1145/2889160.2892661>
- [41] Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. 2022. CodeGen: An Open Large Language Model for Code with Multi-Turn Program Synthesis. In *The Eleventh International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.2203.13474>
- [42] Yusuke Oda, Hiroyuki Fudaba, Graham Neubig, Hideaki Hata, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2015. Learning to generate pseudo-code from source code using statistical machine translation. In *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 574–584. <https://doi.org/10.1109/ase.2015.36>
- [43] OpenAI. 2023. GPT3.5. <https://platform.openai.com/docs/guides/gpt/chat-completions-api>. (Accessed on 09/11/2023).
- [44] Carlos Pacheco, Shuvendu K Lahiri, Michael D Ernst, and Thomas Ball. 2007. Feedback-directed random test generation. In *29th International Conference on Software Engineering (ICSE '07)*. IEEE, 75–84. <https://doi.org/10.1109/icse.2007.37>
- [45] Annibale Panichella, Fitsum Meshesha Kifetew, and Paolo Tonella. 2015. Reformulating branch coverage as a many-objective optimization problem. In *2015 IEEE 8th international conference on software testing, verification and validation (ICST)*. IEEE, 1–10. <https://doi.org/10.1109/icst.2015.7102604>
- [46] Baptiste Roziere, Marie-Anne Lachaux, Lowik Chausson, and Guillaume Lample. 2020. Unsupervised translation of programming languages. *Advances in Neural Information Processing Systems* 33 (2020), 20601–20611. <https://doi.org/10.48550/arXiv.2006.03511>

- [47] Baptiste Roziere, Jie Zhang, Francois Charton, Mark Harman, Gabriel Synnaeve, and Guillaume Lample. 2021. Leveraging Automated Unit Tests for Unsupervised Code Translation. In *International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.2110.06773>
- [48] Marc Szafraniec, Baptiste Roziere, Hugh James Leather, Patrick Labatut, Francois Charton, and Gabriel Synnaeve. 2022. Code Translation with Compiler Representations. In *The Eleventh International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.2207.03578>
- [49] Yutian Tang, Zhijie Liu, Zhichao Zhou, and Xiapu Luo. 2023. ChatGPT vs SBST: A Comparative Assessment of Unit Test Suite Generation. *arXiv e-prints* (2023), arXiv–2307. <https://doi.org/10.1109/tse.2024.3382365>
- [50] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023). <https://doi.org/10.48550/arXiv.2302.13971>
- [51] Michele Tufano, Dawn Drain, Alexey Svyatkovskiy, Shao Kun Deng, and Neel Sundaresan. 2020. Unit test case generation with transformers and focal context. *arXiv preprint arXiv:2009.05617* (2020). <https://doi.org/10.48550/arXiv.2009.05617>
- [52] Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in Statistics: Methodology and Distribution*. Springer, 196–202. https://doi.org/10.1007/978-1-4612-4380-9_16
- [53] Wei Wu, Yann-Gaël Guéhéneuc, Giuliano Antoniol, and Miryung Kim. 2010. AURA: A Hybrid Approach to Identify Framework Evolution. In *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering - Volume 1* (Cape Town, South Africa) (ICSE '10). Association for Computing Machinery, New York, NY, USA, 325–334. <https://doi.org/10.1145/1806799.1806848>
- [54] Chunqiu Steven Xia, Yuxiang Wei, and Lingming Zhang. 2023. Automated program repair in the era of large pre-trained language models. In *Proceedings of the 45th International Conference on Software Engineering (ICSE 2023)*. Association for Computing Machinery. <https://doi.org/10.1109/icse48619.2023.00129>
- [55] Chunqiu Steven Xia and Lingming Zhang. 2022. Less training, more repairing please: revisiting automated program repair via zero-shot learning. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 959–971. <https://doi.org/10.1145/3540250.3549101>
- [56] Chunqiu Steven Xia and Lingming Zhang. 2023. Conversational automated program repair. *arXiv preprint arXiv:2301.13246* (2023). <https://doi.org/10.48550/arXiv.2301.13246>
- [57] Pengyu Xue, Linhao Wu, Zhongxing Yu, Zhi Jin, Zhen Yang, Xinyi Li, Zhenyu Yang, and Yue Tan. 2024. Automated Commit Message Generation with Large Language Models: An Empirical Study and Beyond. *arXiv preprint arXiv:2404.14824* (2024). <https://doi.org/10.48550/arXiv.2404.14824>
- [58] Zhen Yang, Jacky Keung, Xiao Yu, Xiaodong Gu, Zhengyuan Wei, Xiaoxue Ma, and Miao Zhang. 2021. A multi-modal transformer-based code summarization approach for smart contracts. In *2021 IEEE/ACM 29th International Conference on Program Comprehension (ICPC)*. IEEE, 1–12. <https://doi.org/10.1109/icpc52881.2021.00010>
- [59] Zhen Yang, Jacky Wai Keung, Zeyu Sun, Yunfei Zhao, Ge Li, Zhi Jin, Shuo Liu, and Yishu Li. 2024. Improving domain-specific neural code generation with few-shot meta-learning. *Information and Software Technology* 166 (2024), 107365. <https://doi.org/10.1016/j.infsof.2023.107365>
- [60] Zhen Yang, Jacky Wai Keung, Xiao Yu, Yan Xiao, Zhi Jin, and Jingyu Zhang. 2023. On the significance of category prediction for code-comment synchronization. *ACM Transactions on Software Engineering and Methodology* 32, 2 (2023), 1–41. <https://doi.org/10.1145/3534117>
- [61] Zhiqiang Yuan, Yiling Lou, Mingwei Liu, Shiji Ding, Kaixin Wang, Yixuan Chen, and Xin Peng. 2023. No More Manual Tests? Evaluating and Improving ChatGPT for Unit Test Generation. *arXiv preprint arXiv:2305.04207* (2023). <https://doi.org/10.48550/arXiv.2305.04207>
- [62] Fengji Zhang, Bei Chen, Yue Zhang, Jacky Keung, Jin Liu, Daoguang Zan, Yi Mao, Jian-Guang Lou, and Weizhu Chen. 2023. Repocoder: Repository-level code completion through iterative retrieval and generation. *arXiv preprint arXiv:2303.12570* (2023). <https://doi.org/10.18653/v1/2023.emnlp-main.151>
- [63] Fengji Zhang, Xiao Yu, Jacky Keung, Fuyang Li, Zhiwen Xie, Zhen Yang, Caoyuan Ma, and Zhimin Zhang. 2022. Improving Stack Overflow question title generation with copying enhanced CodeBERT model and bi-modal information. *Information and Software Technology* 148 (2022), 106922. <https://doi.org/10.1016/j.infsof.2022.106922>
- [64] Haoxiang Zhang, Shaowei Wang, Tse-Hsun Chen, Ying Zou, and Ahmed E Hassan. 2019. An empirical study of obsolete answers on stack overflow. *IEEE Transactions on Software Engineering* 47, 4 (2019), 850–862. <https://doi.org/10.1109/tse.2019.2906315>
- [65] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. <https://doi.org/10.48550/arXiv.2303.18223> arXiv:2303.18223 [cs.CL]

- [66] Hao Zhong, Suresh Thummalapenta, Tao Xie, Lu Zhang, and Qing Wang. 2010. Mining API Mapping for Language Migration. In *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering - Volume 1* (Cape Town, South Africa) (*ICSE '10*). Association for Computing Machinery, New York, NY, USA, 195–204. <https://doi.org/10.1145/1806799.1806831>
- [67] Qihao Zhu, Zeyu Sun, Yuan-an Xiao, Wenjie Zhang, Kang Yuan, Yingfei Xiong, and Lu Zhang. 2021. A syntax-guided edit decoder for neural program repair. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 341–353. <https://doi.org/10.1145/3468264.3468544>

Received 2023-09-28; accepted 2024-04-16