CodeEditorBench: EVALUATING CODE EDITING CAPABILITY OF LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) for code are rapidly evolving, with code editing emerging as a critical capability. We introduce *CodeEditorBench*, a pioneering evaluation framework designed to rigorously assess the performance of LLMs in code editing tasks, including debugging, translating, polishing, and requirement switching. Unlike existing benchmarks focusing solely on code generation, *CodeEditorBench* emphasizes real-world scenarios and practical aspects of software development. We curated diverse coding challenges and scenarios from five sources, covering various programming languages, complexity levels, and editing tasks. Evaluating 19 LLMs revealed that despite the relative consistency observed between the models' code editing and code generation abilities, notable differences persist.The results highlight the models' limitations in code polishing and code rewriting as required and also indicate that models specifically tailored for code feedback capabilities show significant improvements in code editing tasks. *CodeEditorBench* aims to catalyze advancements in LLMs by providing a robust platform for assessing code editing capabilities. We will release the dataset and evaluation code to enable the community to study code editing tasks of LLMs.[1].

1 INTRODUCTION

Recent advancements in LLMs (Touvron et al., 2023; Achiam et al., 2023) underscore the importance of their coding capabilities, extending beyond mere programming assistance (Tian et al., 2023; Nijkamp et al., 2023a) to encompass various tool-using applications (Qin et al., 2023; Cai et al., 2023). Specifically, code LLMs (Rozière et al., 2024; Guo et al., 2024; Zheng et al., 2024) are deployed across various tasks, such as code repair (Olausson et al., 2023), code optimization (Shypula et al., 2023).

Coding involves a wide range of skills, with code editing playing a pivotal role in software development, encompassing tasks such as optimization, refactoring, and bug fixing. Despite the growing use of LLMs as programming aids, existing evaluation methods, such as HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021), primarily focus on code generation, neglecting the crucial aspect of code editing in software development.

To bridge such a significant gap in evaluation, we advocate for developing a new benchmark to *comprehensively assess the code editing abilities of LLMs*. To this end, we introduce *CodeEditorBench*, a pioneering evaluation framework designed to assess the performance of LLMs in editing code rigorously, where the overview is described in Figure 1. The categorization of code editing problems helps to understand and evaluate the performance of code LLMs systematically. Based on the SDLC definitions[2], we evaluate code LLMs in four scenarios: *Code Debug*, *Code Translate*, *Code Polish*, and *Code Requirement Switch*, which are categorized to reflect the most common and critical types of tasks in the code editing process. See subsection A.1 for classification basis. ***(i) Code Debug***: Debugging is the process of locating and fixing errors in code. ***(ii) Code Translate***: Translating means converting code from one programming language to another. ***(iii) Code Polish***: Polishing refers to optimizing code without changing its functionality. ***(iv) Code Requirement Switch***: Requirement switching is adapting code to new or modified requirements.

[1] https://anonymous.4open.science/r/CodeEditorBench-2813

[2] https://en.wikipedia.org/wiki/Software_development_process

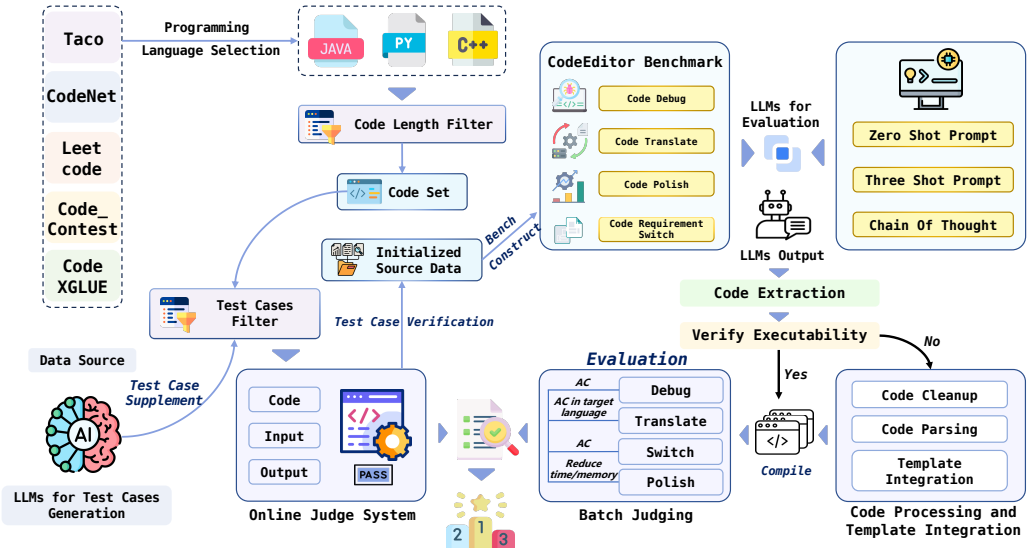


Figure 1: Overview of *CodeEditorBench*. *CodeEditorBench* evaluates programming languages by selecting initial data from five sources and filtering based on code length. It enriches the dataset with LLM generated test cases, which, along with all code, are verified by an online judge system. The benchmark is developed for four problem types using specific methodologies, described in Figure 3. Assessment of 19 LLMs involves crafting prompts for zero-shot, three-shot, and chain of thought settings. Outputs are filtered and integrated with templates for compilation. The OJ's batch judging determines the LLMs' scores, ensuring a rigorous evaluation process.

Additionally, we make significant efforts to manually construct test cases for each problem in various programming scenarios to precisely check the editing correctness. We further established an online evaluation system to facilitate easy evaluation of a broad set of code LLMs. Consequently, we compiled a dataset containing 7,961 code editing tasks, each with an average of 44 test cases (a minimum of 8 and a maximum of 446). Inspired by LiveCodeBench (Jain et al., 2024), we implemented a timestamp-based filtering process to consider the data pollution phenomenon. This process led to a refined dataset called CodeEditorBench_Plus. The original dataset was thereafter designated as CodeEditorBench_Primary.

To complement the introduction of a comprehensive benchmark, we aim to delineate the current array of available models through a code editing leaderboard. We assess 6 base models and 13 models that undergo instruction tuning across four distinct scenarios, utilizing the same experimental framework, and employ two evaluative approaches: zero-shot and three-shot.

We summarize our contributions as follows:

- We provide a unified framework for assessment, including tools for visualization, training, and additional analyses. We will also make all the data involved in the evaluations publicly available to foster further examination of LLM characteristics. Furthermore, we plan to incorporate more evaluation metrics in the future.
- We extend previous evaluations of code editing capabilities, which were limited in scope. For example, FixEval (Haque et al., 2023) assessed only two models for code debugging, DebugBench (Tian et al., 2024) evaluated five, and AVATAR (Ahmad et al., 2021) did not test any LLMs for code translation. Pie4Perf (Shypula et al., 2023)'s focus was limited to the CodeLlama and GPT series for code polishing.Our more comprehensive evaluation demonstrates a clear correlation between models' code editing and code generation performance.
- We highlight the models' limitations in code polishing and code rewriting as required. Based on this comprehensive evaluation, we have also identified discrepancies in the rankings compared to code generation tasks. Additionally, The results indicate that models specifically tailored for code feedback capabilities show significant improvements in code editing tasks.

2 RELATED WORK

Code LLMs The field witnesses significant growth in developing code LLMs to address the challenges in code understanding and generation. This trend starts with the introduction of Codex (Chen et al., 2021) by OpenAI, followed by the emergence of many influential models including CodeGen (Nijkamp et al., 2023b), CodeT5 (Wang et al., 2021; 2023), and InCoder (Fried et al., 2023). Recent popular open-source models like CodeLLaMa (Rozière et al., 2024), DeepSeek Coder (Guo et al., 2024), and StarCoder (Li et al., 2023a; Lozhkov et al., 2024) represent the forefront of this field. They demonstrate excellent abilities in various code understanding and generation tasks by extensively pre-training from scratch on massive code datasets. Additionally, these base models undergo another phase of instruction tuning (Zheng et al., 2024; Wei et al., 2023; Luo et al., 2023; Royzen et al., 2023; Liu et al., 2023a; Muennighoff et al., 2024), empowering them with better instruction-following capability, leading to significant performance improvements in solving various code-related tasks.

Code Benchmark Many benchmarks are proposed to compare and evaluate code LLMs. However, these primarily focus on natural language and code generation. HumanEval (Chen et al., 2021) is one of the pioneering and most widely used benchmarks for LLM-based code synthesis, consisting of 164 pairs of Python function signature with docstring and the associated test-cases for correctness checking. Another Python-focused dataset, MBPP (Austin et al., 2021), is created by crowd-sourcing participants to write in summation 974 programming problems, each of which is comprised of the problem statement (i.e., docstring), the function signature, as well as three test-cases. Beyond Python, there are other benchmarks targeting additional languages such as Spider (SQL) (Yu et al., 2018), HumanEval-X (Zheng et al., 2023) (C++, Javascript and Go), HumanEvalPack (Muennighoff et al., 2024)(Python, JavaScript, Java, Go, C++ and Rust), CodeContests (Li et al., 2022) (C++ and Java) and MultiPL-E (Cassano et al., 2022) (extending HumanEval and MBPP to 18 programming languages). Competitive programming benchmarks include LeetCode-Hard Gym (Olausson et al., 2023), which evaluates code generation in multiple languages using LeetCode's server and the OpenAI gym framework. DebugBench (Tian et al., 2024) advances LLM evaluation by focusing on error correction across diverse programming challenges, from syntax to logical mistakes. EditEval (Hu et al., 2023) assesses LLMs' ability to understand and execute code editing instructions, measuring how accurately models can modify code based on human-written instructions.

Relatively little work addresses the objective of code editing. Previous works either focus on a subset of code editing tasks or do not give a reasonable division of code edit tasks. To fill this gap, we introduce *CodeEditorBench*, a pioneering evaluation framework designed to assess the performance of LLMs in editing code.

3 METHOD

3.1 PROBLEM DEFINITION

For any question q_i, in order to test whether the code c_i generated by LLM is correct, a reasonable approach is to construct a set R_i - a collection of test cases containing test case input-output pairs (x_i,y_i), each x_i being a program input and y_i being a corresponding desired output. Let $a_c(x) = y$ denote a program a, based on a code script c, that maps an input x to an output y. We define the answer code c_i to be correct if and only if there is not any input-output pair in R_i such that $a_{c_i}(x_i) \neq y_i$

(i) an ideal debugger D that rectifies any buggy code from c to $c*$ should satisfy that D(c) = $c*$ s.t.$\forall(x_i, y_i) \in R_i$, $a_c*(x_i) = y_i$.Debugging can be regarded as the converting process of debugger D.

(ii) an ideal translater T that translates any code from language a to language b should satisfy that T(c) = $c*$ s.t. $\forall(x_i,y_i) \in R_i$, $a_c(x_i) = y_i$ and $a_c*(x_i) = y_i$.Translating can be regarded as the converting process of translater T.

(iii) an ideal optimizer P that make any code more effective in either time or space complexity with the guarantee of accuracy should satisfy that P(c) = $c*$ s.t. $\forall(x_i,y_i) \in R_i$, $a_c(x_i) = y_i$,$a_c*(x_i) = y_i$ and $avg_{time}(c*) \leq avg_{time}(c)$ or $avg_{memory}(c*) \leq avg_{memory}(c)$.Polishing can be regarded as the converting process of optimizer P.

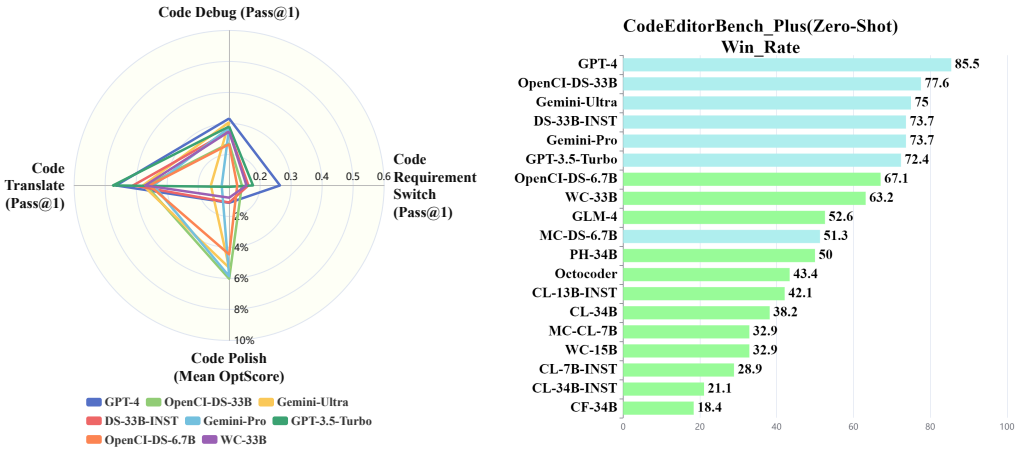


Figure 2: Based on subsection 5.2. **Left**. We propose evaluating LLMs across four scenarios capturing various code editing capabilities, namely code debug, code translate, code polish, and code requirement switch. The figure depicts various model performances across the four scenarios available in CodeEditorBench_Plus in a radial plot – highlighting how relative differences across models change across the scenarios. **Right**. Performance of open-source and closed-source models on CodeEditorBench_Plus in zero-shot evaluated through win_rate. For a comprehensive explanation of the abbreviation, refer to section 4.1.

(iv) an ideal requirement switcher S that switches a similar code that satisfies question a to a target code that satisfies question b based on the sample inputs and outputs of a and b should satisfy that S(c) = $c*$ s.t. $\forall(x_i,y_i) \in R_b$, $a_c*(x_i) = y_i$.Switching can be regarded as the converting process of switcher S.

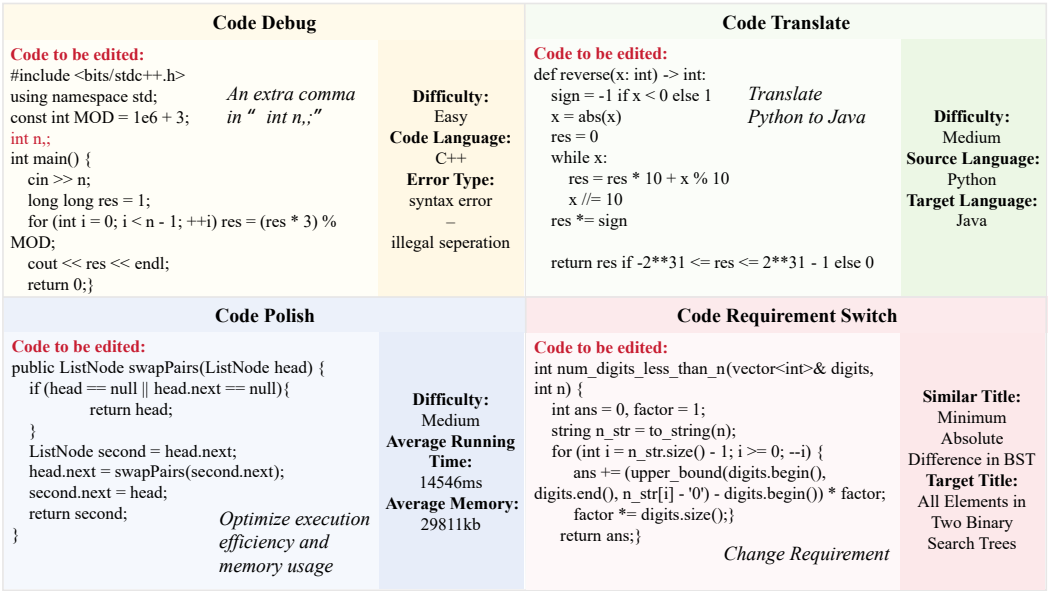


Figure 3: Data Samples of *CodeEditorBench*

3.2 Data Construction

In our meticulously crafted compilation, we gather a wide-ranging assortment of coding challenges sourced from five sources: namely leetcode[3], code_contests (Li et al., 2022), CodeXGLUE (Lu et al.,

[3] https://leetcode.com

2021), codeNet (Puri et al., 2021) and Taco (Li et al., 2023b). To align with the constraints imposed by the maximum token limit of advanced LLMs, we applied a stringent selection criterion, setting aside any question codes that surpassed 800 lines or exceeded a 1000-token threshold. This approach ensures the integrity of the code generated by LLMs.

Our collection spans extensive data structures— from trees, stacks, and queues to arrays, hash tables, and pointers. It also covers a broad spectrum of algorithms, including but not limited to dynamic programming, various sorting techniques, depth-first search (DFS), breadth-first search (BFS), greedy algorithms, and recursion. This comprehensive compilation aims to facilitate an all-encompassing exploration and understanding of coding principles, catering to various computational problems and scenarios.

3.2.1 CONSTRUCTION METHOD

Code Debug Inspired by the DebugBench (Tian et al., 2024), we employ the insertion of a basic error to construct the data with one error and utilized a similar methodology to generate data sets with two, three, and four errors, as detailed in the basic error types in Table 3. The precise prompts used for inserting errors are detailed in subsection A.2. In addition, we manually recheck after insertion errors to ensure that they are inserted accurately.

Code Translate & Code Polish Based on a straightforward rationale, the challenge LLMs encounter in understanding code typically increases as the complexity of the code itself rises. Consequently, we stratify the dataset according to code complexity, adhering to a selection ratio of 3:4:1.

Code Requirement Switch Initially, we categorize the data into two groups. The first is designated as 'strong relation', which represents the similar questions provided by Leetcode under a certain question. These questions exhibit clear human feedback and represent the most tightly interconnected subjects. Conversely, the second category, 'weak relation', which is constructed by us. The methodology for its construction is outlined as follows:

- Collect the labels of each question.
- Cluster the questions based on the number of tags they possess, with our clustering criterion being the presence of four or more identical tags.
- Given that tags only partially convey the essence of the questions, we employ Bert (Devlin et al., 2018) to assess the semantic similarity between the descriptions of two questions within each category. Our set threshold for similarity is 0.92.

Despite the majority of the dataset being synthetically generated by us, we acknowledge the inherent risk of data pollution within the dataset. To address this concern, we implement a timestamp-based filtering process. This novel strategy enables us to methodically examine the dataset, thereby identifying and excluding outdated information, which significantly improves the overall quality of the dataset. Consequently, from CodeEditorBench_Primary, we develop CodeEditorBench_Plus.

Test cases generation During the benchmark construction, we encounter several issues: notably, some topics are deficient in test cases, failing to meet the minimum requirement of eight cases. Additionally, the comprehensiveness of some test cases is compromised due to the absence of boundary tests, thereby falling short of the standards for rigorous evaluation. To address these shortcomings, we leverage three LLMs for test case generation: ***(i)*** GPT-4: generate boundary test cases. ***(ii)*** GLM-4: generate basic test cases. ***(iii)*** Qwen-72B-Chat:check the formatting of the test cases, along with ensuring each topic has at least 8 test cases. Mindful of the potential for illusion, we opt to select only the inputs from the generated test cases. The corresponding outputs are then determined by executing the problem-solving code via OJ, which we establish to assess the correctness of code generated by LLMs. The specific prompt for generating test cases can be found in Figure 6 and Figure 7.

3.3 DATA ANALYSIS

Figure 3 presents a selection of exemplars from the *CodeEditorBench*, delineating the spectrum of code editing tasks, including Code Debugging, Code Translating, Code Polishing, and Code Requirement Switching.

The CodeEditorBench_Primary, as illustrated in Figure 8a, and CodeEditorBench_Plus, as shown in Figure 8b, establish an evaluation framework that mirrors the complexities inherent in real-world software development scenarios. This framework is meticulously designed to gauge LLMs' code editing capabilities, presenting a richer and more nuanced set of challenges than conventional code generation benchmarks. The datasets are extensive, categorizing tasks along several dimensions: programming languages (C++, Java, Python), number of errors (one, two, three, four), difficulty levels (easy, medium, hard), language transitions (e.g., C++ to Java), and relation strength (strong, weak).

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Decoding Strategy. Our dataset is significantly larger than those used in prior benchmarks, necessitating adjustments in our evaluation approach. For closed-source models, given the considerable expense associated with API calls, we opt for greedy decoding when generating code, employing the pass@1 (Kulal et al., 2019; Chen et al., 2021) metric to assess the code's pass rate. We also apply a greedy decoding strategy to open-source models to facilitate a fair comparison between them.

Hyperparameter Settings. Our dataset was filtered to exclude code exceeding 800 lines or 1024 tokens. In our benchmark's four scenarios, the generated code is typically similar in length to the source code, leading us to set the count of maximum new tokens to 1024 for closed-source models. For open-source models, we set the maximum new tokens' count to 2048. We utilize vLLM (Kwon et al., 2023) to accelerate the output process for open-source models.

The same experimental setup is used across all four scenarios.

Evaluated Models We select 19 of the most popular current code LLMs from existing leaderboards evalplus (Liu et al., 2023b), bigcode (Ben Allal et al., 2022), encompassing both open-source and closed-source models, with sizes ranging from 6.7B to 34B, including base models and instruction-tuning models. The open-source models include the GPT series (GPT-3.5-Turbo, GPT-4), the Gemini series (Gemini-Pro, Gemini-Ultra), and GLM-4. Closed-source models comprise the CodeLlama series (CL-34B-Base, CL-{7, 13, 34}B-INST), the DeepSeek series (DS-33B-INST), and the outstanding instruction-finetuned models, the WizardCoder series (WC-15B based on StarCoder, WC-33B based on DS-33B-Base), OctoCoder based on StarCoder, CF-34B based on CL-34B-Python, PH-34B based on CL-34B-Base, the Magicoder series (MC-DS-6.7B based on DS-6.7B-Base, MC-CL-7B based on CL-7B) and the OpenCodeInterpreter series(OpenCI-DS-6.7B based on DS-6.7B-Base, OpenCI-DS-33B based on DS-33B-Base). Appendix B presents more detailed information regarding our evaluated models.

4.2 PROMPT ENGINEERING

Prompt Setting. We implement various prompting techniques to evaluate our model, including zero-shot prompting and few-shot prompting methods. In selecting examples for few-shot prompts, a clustering approach is utilized to pick three fixed examples from the dataset of each scenario.

Prompt Format. In constructing prompts for open-source models, we adhere to the formats provided in their official examples. We search for a suitable prompt format in the HuggingFace model card, Github repository, and formal publications or technical reports. The variance in prompts across different open-source models primarily lies in a few special identifiers, such as for the CodeLlama-INST series of models, we add <s>[INST] and [/INST] at the beginning and end of the instruction, respectively. We consistently use the *Instruction-Question-Answer* format to construct prompts for closed-source models.

Model	Size	Open	Debug	Translate	Switch	Polish	Win Rate
				Zero-shot			
GPT-4	-	✗	**0.316**(0.493)	0.465(**0.503**)	**0.264**	1.12%(1.33%)	**0.855**(**0.868**)
OpenCI-DS-33B	33B	✔	0.236(0.429)	0.368(0.428)	0.141	**6.02%**(6.49%)	0.776(0.816)
Gemini-Ultra	-	✗	0.304(0.459)	0.378(0.278)	0.041	5.31%(3.77%)	0.750(0.579)
DS-33B-INST	33B	✔	0.275(0.487)	0.410(0.451)	0.162	1.10%(1.14%)	0.737(0.757)
Gemini-Pro	-	✗	0.286(0.423)	0.344(0.344)	0.076	5.86%(**6.65%**)	0.737(0.711)
GPT-3.5-Turbo	-	✗	0.290(**0.494**)	**0.475**(0.480)	0.177	0.09%(0.84%)	0.724(0.776)
OpenCI-DS-6.7B	6.7B	✔	0.233(0.402)	0.357(0.384)	0.126	4.45%(4.28%)	0.671(0.697)
WC-33B	33B	✔	0.274(0.487)	0.371(0.438)	0.156	0.79%(0.90%)	0.632(0.704)
GLM-4	-	✗	0.220(0.271)	0.278(0.365)	0.085	5.17%(6.46%)	0.526(0.592)
MC-DS-6.7B	6.7B	✔	0.242(0.406)	0.343(0.401)	0.130	0.21%(1.99%)	0.513(0.697)
PH-34B	34B	✔	0.230(0.369)	0.279(0.331)	0.074	2.84%(1.78%)	0.500(0.539)
Octocoder	15.5B	✔	0.042(0.145)	0.392(0.223)	0.030	1.39%(2.70%)	0.434(0.289)
CL-13B-INST	13B	✔	0.176(0.368)	0.333(0.275)	0.021	2.31%(1.82%)	0.421(0.368)
CL-34B	34B	✔	0.163(0.250)	0.310(0.240)	0.052	1.10%(0.84%)	0.382(0.171)
MC-CL-7B	7B	✔	0.174(0.317)	0.272(0.276)	0.039	1.31%(1 31%)	0.329(0.342)
WC-15B	15B	✔	0.159(0.354)	0.309(0.278)	0.067	0.91%(0.96%)	0.329(0.408)
CL-7b-INST	7B	✔	0.155(0.336)	0.289(0.231)	0.017	1.47%(1.17%)	0.289(0.250)
CL-34B-INST	34B	✔	0.131(0.250)	0.287(0.240)	0.027	1.02%(0.84%)	0.211(0.171)
CF-34B	34B	✔	0.166(0.223)	0.218(0.177)	0.028	0.33%(0.70%)	0.184(0.105)
				Few-shot			
Gemini-Ultra	-	✗	0.286(0.448)	0.443(0.307)	0.152	5.62%(4.55%)	**0.855**(0.632)
GPT-4	-	✗	**0.345**(**0.523**)	**0.517**(**0.514**)	**0.303**	1.13%(1.14%)	0.816(**0.882**)
OpenCI-DS-6.7B	6.7B	✔	0.233(0.440)	0.372(0.399)	0.165	**6.47%**(**8.59%**)	0.770(0.750)
OpenCI-DS-33B	33B	✔	0.230(0.463)	0.371(0.437)	0.229	5.75%(4.82%)	0.763(0.803)
DS-33B-INST	33B	✔	0.272(0.489)	0.417(0.465)	0.235	1.18%(0.93%)	0.737(0.763)
GPT-3.5-Turbo	-	✗	0.270(0.511)	0.364(0.431)	0.201	1.54%(1.70%)	0.684(0.803)
Gemini-Pro	-	✗	0.229(0.386)	0.392(0.356)	0.139	5.23%(5.64%)	0.671(0.645)
WC-33B	33B	✔	0.279(0.515)	0.362(0.447)	0.243	0.65%(0.63%)	0.645(0.711)
MC-DS-6.7B	6.7B	✔	0.262(0.478)	0.321(0.381)	0.192	1.44%(0.89%)	0.605(0.632)
GLM-4	-	✗	0.233(0.341)	0.299(0.360)	0.100	5.30%(6.41%)	0.572(0.592)
CL-34B	34B	✔	0.133(0.367)	0.307(0.252)	0.113	1.75%(1.11%)	0.474(0.447)
PH-34B	34B	✔	0.239(0.468)	0.275(0.326)	0.092	1.20%(0.75%)	0.421(0.461)
CL-13B-INST	13B	✔	0.160(0.330)	0.327(0.284)	0.028	1.75%(1.25%)	0.414(0.322)
MC-CL-7B	7B	✔	0.157(0.355)	0.245(0.230)	0.075	1.70%(1.18%)	0.329(0.382)
WC-15B	15B	✔	0.114(0.332)	0.271(0.224)	0.099	1.65%(1.11%)	0.322(0.329)
CF-34B	34B	✔	0.166(0.262)	0.240(0.158)	0.050	1.61%(1.39%)	0.289(0.250)
Octocoder	15.5B	✔	0.050(0.263)	0.290(0.206)	0.054	1.09%(0.85%)	0.211(0.184)
CL-7B-INST	7B	✔	0.167(0.362)	0.271(0.224)	0.028	1.00%(0.71%)	0.211(0.204)
CL-34B-INST	34B	✔	0.143(0.330)	0.303(0.264)	0.032	0.32%(0.67%)	0.211(0.211)

Table 1: Evaluating LLMs on CodeEditorBench. All results of models are generated by greedy decoding. Code Debug, Code Translate and Code Requirement Switch are evaluated with pass@1, while Code Polish is evaluated with Mean OptScore. Values outside parentheses denote Plus results and inside denote Primary results. For the Switch class, Primary and Plus results are identical, and only one score is displayed.

5 RESULT ANALYSIS

5.1 EVALUATION ONLINE SYSTEM

To comprehensively assess LLM's code editing performance across the four scenarios, we construct OJ based on the hustoj (Hao-bin, 2012). The system processes LLM-generated code to ensure adherence to operational requirements, tailors a set of focused test problems, and passes criteria for each scenario, facilitating a comprehensive evaluation.

Pass Criteria. For Code Debug, Code Translate, and Code Requirement Switch, we verify whether the code passes all test cases within the time and memory constraints of our OJ. For Code Translate, we specifically run the code in the target language environment to ensure an accurate evaluation of translation performance. To meet the pass criteria for Code Polish, the code must pass all test cases and demonstrate improved efficiency by reducing execution time or memory usage.

	Pass	Wrong Answer	Runtime Error	Compile Error	Other
Debug	21.57%	53.41%	13.40%	7.51%	4.11%
Polish	19.23%	53.15%	5.46%	3.01%	19.15%
Switch	11.18%	64.74%	10.94%	8.09%	5.05%
Translate	33.15%	45.67%	3.69%	6.06%	11.43%
ALL	20.34%	55.26%	8.53%	6.34%	9.53%

Table 2: Judgment results across different problem types in CodeEditorBench_Plus

For more detailed information on the pass criteria and evaluation configuration, please refer to Appendix E and Appendix F respectively.

5.2 Performance Metrics

We evaluate the 19 models described in section 4.1 using a zero-shot and few-shot approach on *CodeEditorBench*. For Code Debug, Code Translate, and Code Requirement Switch, we employ pass@1 evaluation criteria. For Code Polish, we use the Mean OptScore as a ranking metric. We measure the average runtime $\bar{T}$ and average memory usage $\bar{M}$ over 20 executions of the original code. For each polish problem, we conduct two measurements of the model-generated code and calculate the average values $\bar{T}_{\text{avg}}$ and $\bar{M}_{\text{avg}}$. For each model-generated code pass all test cases, if $\bar{T} > \bar{T}_{\text{avg}}$ or $\bar{M} > \bar{M}_{\text{avg}}$, the code is considered to pass, and the score is calculated as:

$$OptScoreTime = \frac{(\bar{T} - \bar{T}_{\text{avg}})}{\bar{T}}$$

$$OptScoreMem = \frac{(\bar{M} - \bar{M}_{\text{avg}})}{\bar{M}}$$

$$OptScore = \max\left[\frac{(OptScoreTime + OptScoreMem)}{2}, 0\right]$$

Otherwise, the score is set to 0. We calculate the average OptScore across questions and obtain the mean OptScore.

Finally, we calculate each model's rank based on its performance on CodeEditorBench_PLUS for each problem category. Inspired by (Ben Allal et al., 2022), we utilize the win rate to evaluate the model's overall performance across various types of problems. We compute the win rate using $1 - (rank - 1)/num_models$ for each problem category, and average them across all categories as the win rate. The results are summarized in Table 1.

5.3 Model Performance Comparison

Analysis of Table 1 reveals that, in Debug and Translate data types, some LLMs exhibit a significant difference in pass@1 between Primary and Plus datasets, with values exceeding 0.2. This discrepancy suggests the potential for data leakage within the Primary dataset. Therefore, we focus on analyzing LLMs' performance on the Plus dataset.

Situation. The comparative analysis of model efficacy on the Plus dataset detailed in Table 1 reveals that closed-source LLMs like Gemini-Ultra and GPT-4 generally outperform their open-source analogs. GPT-4 excels in zero-shot tasks across Debug, Translate, and Switch categories, whereas Gemini-Ultra leads in few-shot settings, particularly in Polish tasks, but struggles with Switch tasks. Among open-source models, OpenCI-DS-33B stands out in zero-shot scenarios, surpassing Gemini-Ultra, though it shares similar weaknesses in Switch tasks. Notably, OpenCI-DS-6.7B outperforms its open-source peers and several closed-source models, including Gemini-Pro and GPT-3.5, demonstrating significant efficacy despite its smaller size.

Pass Rate Distribution. The pass rates exhibit significant variation across different problem types, as illustrated in Table 2. The PLUS dataset identifies Switch problems as the most challenging, with a mere 11.18% pass rate. Debug and Translate problems exhibit pass rates of approximately 20% and 30%, respectively. For Polish problems, even with the correct original code provided, only 37.47% of

the solutions meet all testing criteria. Additionally, only a limited 19.23% of the dataset passes all tests and demonstrates superior average runtime or memory efficiency relative to the original code. It is also significant to note that a notable fraction of solutions simply replicate the original code with no alterations.

Reasons for Not Passing. We analyze the aggregated solutions from all models on CodeEditorBench_Plus, as detailed in Table 2, and discovered that only 20.34% of solutions successfully solve the problem, with a significant 55.26% failing due to incorrect answers. Other prevalent causes of failure include compilation and runtime errors, while instances of timeouts or exceeding memory limits are comparatively rare. Specifically, 6.34% of the dataset experiences compilation errors, a phenomenon that may partly stem from post-processing losses incurred during the extraction of code blocks from solutions that include textual explanations. Models producing poorly formatted output, such as OctoCoder, are notably more susceptible to compilation errors. Interestingly, Polish problems demonstrate the lowest frequencies of both runtime and compilation errors, likely attributable to the minimal alterations made to the original code by the models. Conversely, Translate problems are characterized by a lowest rate of incorrect answers (45.67%), yet suffer the highest rate of timeout errors (10.21%).

6 CONCLUSION

In this study, we introduce *CodeEditorBench*, a pioneering benchmark created to evaluate Large Language Models (LLMs) in code editing tasks. *CodeEditorBench* is envisioned as a dynamic and scalable framework that will be periodically updated to incorporate new problems, scenarios, and models. Our findings indicate that despite the relative consistency observed between the models' code editing and code generation abilities, notable differences persist.The results highlight the models' limitations in code polishing and code rewriting as required and also indicate that models specifically tailored for code feedback capabilities show significant improvements in code editing tasks. The analysis also underscores the variability in model performance based on problem category and scenario, revealing trends in model sensitivity to prompt formulation and highlighting instances where smaller models surpass their larger counterparts in efficiency. Through establishing a holistic evaluation platform, *CodeEditorBench* aims to foster advancements in LLMs for code editing and serve as a valuable resource for researchers and practitioners.

7 LIMITATIONS

While our study on the *CodeEditorBench* introduces a novel and rigorous framework for assessing the code editing capabilities of LLMs, several limitations accompany our research. These limitations are integral to understanding our benchmark's scope, applicability, and areas for future improvement. **Model Coverage**: The evaluation of 19 LLMs may not fully represent the extensive diversity of models available, indicating a need for a more inclusive approach in subsequent studies. **Bias in Task Selection**: Despite efforts to increase diversity, our array of coding challenges may continue to exhibit a preference for specific languages or tasks, potentially compromising the benchmark's impartiality. **Evaluation Metrics**: The utilized metrics may not comprehensively encompass the intricacies of code editing tasks, pointing towards a necessity for more refined assessment techniques. **Real-world Relevance**: The benchmark simulation may not fully capture the complexity of real-world software development projects, highlighting a gap in applicability. **Dynamic LLM Landscape**: The rapid advancement in Large Language Model (LLM) technologies may render our findings obsolete, necessitating continuous updates to the benchmark. Addressing these limitations is crucial for refining the benchmark and its enhanced utility in evaluating LLMs for code editing.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Wasi Uddin Ahmad, Md Golam Rahman Tushar, Saikat Chakraborty, and Kai-Wei Chang. Avatar: A parallel corpus for java-python program translation. *arXiv preprint arXiv:2108.11590*, 2021.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Loubna Ben Allal, Niklas Muennighoff, Logesh Kumar Umapathi, Ben Lipkin, and Leandro von Werra. A framework for the evaluation of code generation models. https://github.com/bigcode-project/bigcode-evaluation-harness, 2022.

Tianle Cai, Xuezhi Wang, Tengyu Ma, Xinyun Chen, and Denny Zhou. Large language models as tool makers. *arXiv preprint arXiv:2305.17126*, 2023.

Federico Cassano, John Gouwar, Daniel Nguyen, Sydney Nguyen, Luna Phipps-Costin, Donald Pinckney, Ming-Ho Yee, Yangtian Zi, Carolyn Jane Anderson, Molly Q Feldman, et al. Multipl-e: A scalable and extensible approach to benchmarking neural code generation. *arXiv preprint arXiv:2208.08227*, 2022.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Daniel Fried, Armen Aghajanyan, Jessy Lin, Sida Wang, Eric Wallace, Freda Shi, Ruiqi Zhong, Wen tau Yih, Luke Zettlemoyer, and Mike Lewis. Incoder: A generative model for code infilling and synthesis, 2023.

Google. Gemini: A family of highly capable multimodal models. https://arxiv.org/abs/2312.11805v2, 2023.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y. Wu, Y. K. Li, Fuli Luo, Yingfei Xiong, and Wenfeng Liang. Deepseek-coder: When the large language model meets programming – the rise of code intelligence, 2024.

Zhang Hao-bin. Design and implementation of the open cloud platform based open source online judge system. *Computer Science*, 2012. URL https://api.semanticscholar.org/CorpusID:57976285.

Md Mahim Anjum Haque, Wasi Uddin Ahmad, Ismini Lourentzou, and Chris Brown. Fixeval: Execution-based evaluation of program fixes for programming problems. In *2023 IEEE/ACM International Workshop on Automated Program Repair (APR)*, pp. 11–18. IEEE, 2023.

Qisheng Hu, Kaixin Li, Xu Zhao, Yuxi Xie, Tiedong Liu, Hui Chen, Qizhe Xie, and Junxian He. Instructcoder: Empowering language models for code editing, 2023.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024.

Sumith Kulal, Panupong Pasupat, Kartik Chandra, Mina Lee, Oded Padon, Alex Aiken, and Percy Liang. Spoc: Search-based pseudocode to code, 2019.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention, 2023.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier, Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muhtasham Oblokulov, Zhiruo Wang, Rudra Murthy, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nour Fahmy, Urvashi Bhattacharyya, Wenhao Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. Starcoder: may the source be with you!, 2023a.

Rongao Li, Jie Fu, Bo-Wen Zhang, Tao Huang, Zhihong Sun, Chen Lyu, Guang Liu, Zhi Jin, and Ge Li. Taco: Topics in algorithmic code generation dataset. *arXiv preprint arXiv:2312.14852*, 2023b.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. Competition-level code generation with alphacode. *arXiv preprint arXiv:2203.07814*, 2022.

Bingchang Liu, Chaoyu Chen, Cong Liao, Zi Gong, Huan Wang, Zhichao Lei, Ming Liang, Dajun Chen, Min Shen, Hailian Zhou, Hang Yu, and Jianguo Li. Mftcoder: Boosting code llms with multitask fine-tuning. *arXiv preprint arXiv*, 2023a.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation, 2023b.

Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, Tianyang Liu, Max Tian, Denis Kocetkov, Arthur Zucker, Younes Belkada, Zijian Wang, Qian Liu, Dmitry Abulkhanov, Indraneil Paul, Zhuang Li, Wen-Ding Li, Megan Risdal, Jia Li, Jian Zhu, Terry Yue Zhuo, Evgenii Zheltonozhskii, Nii Osae Osae Dade, Wenhao Yu, Lucas Krauß, Naman Jain, Yixuan Su, Xuanli He, Manan Dey, Edoardo Abati, Yekun Chai, Niklas Muennighoff, Xiangru Tang, Muhtasham Oblokulov, Christopher Akiki, Marc Marone, Chenghao Mou, Mayank Mishra, Alex Gu, Binyuan Hui, Tri Dao, Armel Zebaze, Olivier Dehaene, Nicolas Patry, Canwen Xu, Julian McAuley, Han Hu, Torsten Scholak, Sebastien Paquet, Jennifer Robinson, Carolyn Jane Anderson, Nicolas Chapados, Mostofa Patwary, Nima Tajbakhsh, Yacine Jernite, Carlos Muñoz Ferrandis, Lingming Zhang, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. Starcoder 2 and the stack v2: The next generation, 2024.

Shuai Lu, Daya Guo, Shuo Ren, Junjie Huang, Alexey Svyatkovskiy, Ambrosio Blanco, Colin B. Clement, Dawn Drain, Daxin Jiang, Duyu Tang, Ge Li, Lidong Zhou, Linjun Shou, Long Zhou, Michele Tufano, Ming Gong, Ming Zhou, Nan Duan, Neel Sundaresan, Shao Kun Deng, Shengyu Fu, and Shujie Liu. Codexglue: A machine learning benchmark dataset for code understanding and generation. *CoRR*, abs/2102.04664, 2021.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. Wizardcoder: Empowering code large language models with evol-instruct, 2023.

Niklas Muennighoff, Qian Liu, Armel Zebaze, Qinkai Zheng, Binyuan Hui, Terry Yue Zhuo, Swayam Singh, Xiangru Tang, Leandro von Werra, and Shayne Longpre. Octopack: Instruction tuning code large language models, 2024.

Erik Nijkamp, Hiroaki Hayashi, Caiming Xiong, Silvio Savarese, and Yingbo Zhou. Codegen2: Lessons for training llms on programming and natural languages. *arXiv preprint arXiv:2305.02309*, 2023a.

Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. Codegen: An open large language model for code with multi-turn program synthesis, 2023b.

Theo X Olausson, Jeevana Priya Inala, Chenglong Wang, Jianfeng Gao, and Armando Solar-Lezama. Demystifying gpt self-repair for code generation. *arXiv preprint arXiv:2306.09896*, 2023.

OpenAI. Chatgpt: Optimizing language models for dialogue. `https://openai.com/blog/chatgpt`, 2022.

OpenAI. Gpt-4 technical report. `https://arxiv.org/abs/2303.08774`, 2023.

Ruchir Puri, David S. Kung, Geert Janssen, Wei Zhang, Giacomo Domeniconi, Vladimir Zolotov, Julian Dolby, Jie Chen, Mihir Choudhury, Lindsey Decker, Veronika Thost, Luca Buratti, Saurabh Pujar, Shyam Ramji, Ulrich Finkler, Susan Malaika, and Frederick Reiss. Codenet: A large-scale ai for code dataset for learning a diversity of coding tasks, 2021.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*, 2023.

Michael Royzen, Justin Wei, and Russell Coleman. Phind. `https://www.phind.com`, 2023.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. Code llama: Open foundation models for code, 2024.

Alexander Shypula, Aman Madaan, Yimeng Zeng, Uri Alon, Jacob Gardner, Milad Hashemi, Graham Neubig, Parthasarathy Ranganathan, Osbert Bastani, and Amir Yazdanbakhsh. Learning performance-improving code edits. *arXiv preprint arXiv:2302.07867*, 2023.

Haoye Tian, Weiqi Lu, Tsz On Li, Xunzhu Tang, Shing-Chi Cheung, Jacques Klein, and Tegawendé F Bissyandé. Is chatgpt the ultimate programming assistant–how far is it? *arXiv preprint arXiv:2304.11938*, 2023.

Runchu Tian, Yining Ye, Yujia Qin, Xin Cong, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Debugbench: Evaluating debugging capability of large language models, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Yue Wang, Weishi Wang, Shafiq Joty, and Steven C. H. Hoi. Codet5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation, 2021.

Yue Wang, Hung Le, Akhilesh Deepak Gotmare, Nghi D. Q. Bui, Junnan Li, and Steven C. H. Hoi. Codet5+: Open code large language models for code understanding and generation, 2023.

Yuxiang Wei, Zhe Wang, Jiawei Liu, Yifeng Ding, and Lingming Zhang. Magicoder: Source code is all you need, 2023.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, et al. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. *arXiv preprint arXiv:1809.08887*, 2018.

Qinkai Zheng, Xiao Xia, Xu Zou, Yuxiao Dong, Shan Wang, Yufei Xue, Zihan Wang, Lei Shen, Andi Wang, Yang Li, Teng Su, Zhilin Yang, and Jie Tang. Codegeex: A pre-trained model for code generation with multilingual evaluations on humaneval-x, 2023.

Tianyu Zheng, Ge Zhang, Tianhao Shen, Xueling Liu, Bill Yuchen Lin, Jie Fu, Wenhu Chen, and Xiang Yue. Opencodeinterpreter: Integrating code generation with execution and refinement. *arXiv preprint arXiv:2402.14658*, 2024.

zhipuai. Glm-4. `https://open.bigmodel.cn/`, 2024.

A CONSTRUCT DATASET

A.1 CLASSIFICATION BASIS

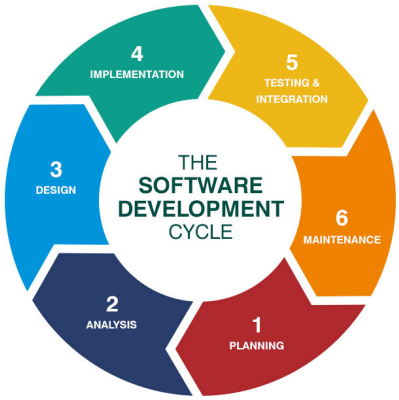


Figure 4: Software Development LifeCycle(Source:`https://bigwater.consulting/2019/04/08/software-development-life-cycle-sdlc/`)

The software development lifecycle (SDLC) is the cost-effective and time-efficient process that development teams use to design and build high-quality software. The goal of SDLC is to minimize project risks through forward planning so that software meets customer expectations during production and beyond. This methodology outlines a series of steps that divide the software development process into tasks you can assign, complete, and measure.

Categorizing code editing tasks based on SDLC in Figure 4 provides a way to understand and organize the different scenarios of code editing from the perspective of the overall flow of a software project, which typically includes phases such as planning, analysis, design, implementation, testing & integration, and maintenance. Here is how to explain and categorize debug, translate, polish, and requirement switch:

1. **Planning.**
   Code Requirement Switch: The planning phase is the period during which the objectives, scope, schedule, and resources of the project are defined. During this phase, initial changes or adjustments in requirements may be identified, and the concept of requirements switching is initially developed here to meet project goals or user expectations.
2. **Analysis.**
   Code Requirement Switch: In the Requirements Analysis phase, the requirements of the project are analyzed and defined in detail. In this process, the requirements may be further adjusted or refined, in order to adapt to these changes, the code requirements switch in this phase becomes particularly important.
3. **Design.**
   Code Translate: The design phase is responsible for translating requirements into system architecture and detailed design. In this phase, the code may need to be translated to fit the design requirements, such as migrating certain components from one technology stack to another, or adapting to different platforms and frameworks.
4. **Implementation.**
   Code Polish: The implementation phase mainly involves coding. Code optimization and polishing are especially important in this phase to ensure the maintainability and performance of the software through refactoring, improving code structure and code quality.
   Code Translate: In addition to code optimization, the implementation phase may also involve code translation, especially in multi-language programming environments or when existing code needs to be adapted to a new framework.

5. **Testing & Integration.**
   Code Debug: The goal of the Testing & Integration phase is to ensure the quality of the software and to find and fix any defects. Code debugging is a core activity in this phase to identify and resolve errors in the code to ensure that the software works as expected.
6. **Maintenance.**
   All Categories: The Maintenance phase covers all the activities that take place after the software is deployed, including fixing defects, updating the software to accommodate new requirements or changes in the environment, improving performance, and so on. In this phase:
   - Code Debug continues to play a role in dealing with user feedback and defects found in the software.
   - Code Translate may involve code migration or rewriting efforts for compatibility or technology upgrades.
   - Code Polish focuses on improving the quality and performance of code through refactoring and optimization.
   - Code Requirement Switch reflects the need to adjust functionality and performance at any point in the software lifecycle in response to changing business requirements or user feedback.

Through the SDLC-based categorization, we can see that 'debug', 'translate', 'polish', and 'requirement switch' are not only different aspects of code editing, but they also reflect the key tasks and challenges faced at each stage of the software development process. This basis for categorization emphasizes the fact that software development is a continuous, iterative process in which the activities in each phase are interdependent and work together to drive the success of a software project.

This basis for categorization emphasizes the fact that software development is a continuous, iterative process in which the activities in each phase are interdependent and work together to drive the success of a software project.

A.2 INSERT ERROR

Error Name	Definition
misused ==/=	Operator misuse: equality (==) vs. assignment (=)
missing colons	Omit colons in control structures and function definitions
unclosed parentheses	Unclosed parentheses cause syntax errors
illegal separation	Syntax errors due to improper separator usage
illegal indentation	Incorrect indentation violates syntax rules (only for Python)
unclosed string	Unclosed string literals: mismatched quotation marks
illegal comment	Incorrect comment syntax or placement
faulty indexing	Incorrect indexing in collections leads to runtime errors
undefined objects	Reference to undefined object: missing definition or import
undefined methods	Calling non-existent method on object/class
illegal keywords	Reserved words misused in programming
condition error	Logical errors in control structure conditions
operation error	Arithmetic errors, like division by zero
variable error	Variable misuse errors: uninitialized variables

Table 3: Basic Error Types

PS: There is an blemish in DebugBench. Illegal separation is a basic error defined in the debugbench.But the error exists only in java and c++ cases, not in python. However,this error also exists in Python in reality. For example, 'print(a, b)' and 'print(a; b)'.So we fill that tiny gap.

Instruction:
Given code: *code*
Please add error: *error name* to the above code
Please output the modified code directly

Answer:

Figure 5: Prompt for inserting basic error

<|im_start|>user You're an experienced programmer.
input data format: *input*
description of the problem: *content*
Please give me as much sample inputs as possible based on the description of the problem in the format "input: xxx". Each sample input must conform to the input data format, and the length of each sample input must not exceed 40 tokens.
Please give the inputs directly without any explanation.

<|im_start|>assistant

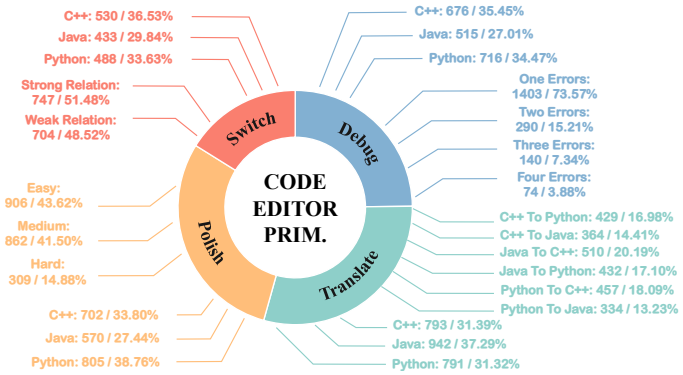
Figure 6: Prompt for generating test cases

<|im_start|>user You're an experienced programmer.
input data format: *input*
description of the problem: *content*
Please give me as much boundry tests as possible based on the description of the problem in the format "input: xxx". Each sample input must conform to the input data format, and the length of each sample input must not exceed 40 tokens.
Please give the inputs directly without any explanation.

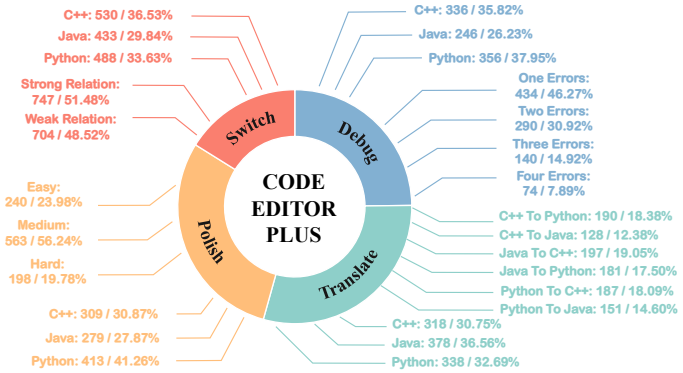
<|im_start|>assistant

Figure 7: Prompt for generating boundary test cases

A.3 Data Analysis

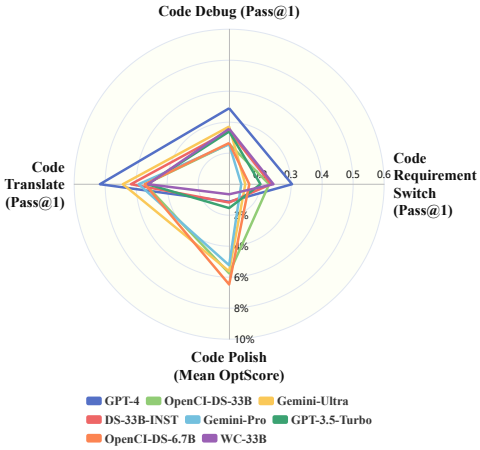


(a) Primary Dataset Analysis



(b) Plus Dataset Analysis

Figure 8: Comparison of Primary and Plus Dataset Analyses



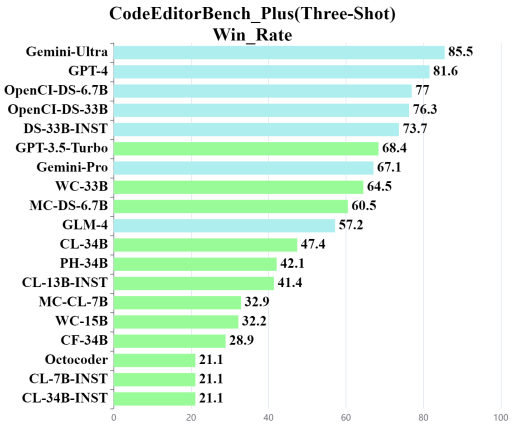


Figure 9: Based on subsection 5.2. **Left**. We propose evaluating LLMs across four scenarios capturing various code editing capabilities,namely code debug, code translate, code polish, and code requirement switch. The figure depicts various model performances across the four scenarios available in CodeEditorBench_Primary in a radial plot – highlighting how relative differences across models change across the scenarios. **Right**. Performance of open-source and closed-source models on CodeEditorBench_Primary in three-shot. For a comprehensive explanation of the abbreviation, refer to section 4.1.

B EVALUATED MODELS

We describe the details of models evaluated in our study in Table 4. To determine the model release date, we search the model's GitHub repository first. If there is no relevant information, the publication date of the paper is used as the release date. It is worth noting that for models such as GPT-3.5-Turbo and GPT-4, the referenced date is actually the cut-off date for the model's training data.

Model	Size	Release Date	Open	Link
ise-uiuc/Magicoder-S-DS-6.7B	6.7B	2023-12-04	✔	Magicoder-S-DS-6.7B
ise-uiuc/Magicoder-S-CL-7B	7B	2023-12-04	✔	Magicoder-S-CL-7B
bigcode/octocoder	15.5B	2023-08-14	✔	octocoder
WizardLM/WizardCoder-15B-V1.0	15B	2023-06-16	✔	WizardCoder-15B-V1.0
WizardLM/WizardCoder-33B-V1.1	33B	2024-01-04	✔	WizardCoder-33B-V1.1
deepseek-ai/deepseek-coder-33b-instruct	33B	2023-02-01	✔	deepseek-coder-33b-instruct
codefuse-ai/CodeFuse-CodeLlama-34B	34B	2023-09-11	✔	CodeFuse-CodeLlama-34B
Phind/Phind-CodeLlama-34B-v2	34B	2023-09-01	✔	Phind-CodeLlama-34B-v2
m-a-p/OpenCodeInterpreter-DS-6.7B	6.7B	2024-02-22	✔	OpenCodeInterpreter-DS-6.7B
m-a-p/OpenCodeInterpreter-DS-33B	33B	2024-02-22	✔	OpenCodeInterpreter-DS-33B
codellama/CodeLlama-34b-hf	34B	2023-08-24	✔	CodeLlama-34b-hf
codellama/CodeLlama-7b-Instruct-hf	7B	2023-08-24	✔	CodeLlama-7b-Instruct-hf
codellama/CodeLlama-13b-Instruct-hf	13B	2023-08-24	✔	CodeLlama-13b-Instruct-hf
codellama/CodeLlama-34b-Instruct-hf	34B	2023-08-24	✔	CodeLlama-34b-Instruct-hf
gpt-3.5-turbo-1106 (OpenAI, 2022)	-	2021-10-01	✗	gpt-3.5-turbo-1106
gpt-4-0613 (OpenAI, 2023)	-	2021-10-01	✗	gpt-4-0613
glm-4 (zhipuai, 2024)	-	2024-01-16	✗	glm-4
gemini-pro (Google, 2023)	-	2023-12-06	✗	gemini-pro
gemini-ultra (Google, 2023)	-	2023-12-06	✗	-

Table 4: Overview of evaluated models.

C DETAILED PROMPTS

The prompt formats demonstrated here are utilized by closed-source models. The instructions used by open-source models are similar to those of closed-source models, with the main differences being as follows:

- Given the limited ability of open-source models to generate code in standard format, we explicitly specify that open-source models generate code enclosed in "```", facilitating post processing.
- Open-source models typically adhere to a fixed prompt format during the instruction fine-tuning phase, requiring the addition of special identifiers before and after the instruction.

C.1 CODE DEBUG

Below we present the prompt formats used by closed-source models in the code debug scenario, under zero-shot and few-shot. See details in Figure 10, Figure 11.

C.2 CODE TRANSLATE

Below we present the prompt formats used by closed-source models in the code translate scenario, under zero-shot and few-shot. See details in Figure 12,Figure 13.

C.3 CODE POLISH

Below we present the prompt formats used by closed-source models in the code polish scenario, under zero-shot and few-shot. See details in Figure 14,Figure 15.

Instruction:
Please correct the errors in the buggy code snippet below, ensuring that your corrected code adheres to the specified programming language syntax and logic requirements. Validate your solution against the provided test cases to ensure its accuracy. Note that your solution should strictly consist of the corrected code only.

Question:
Below is the *code_language* buggy code:
error_code

Correct the code and ensure it passes the following test case:
Input: *public_tests_input*
Output: *public_tests_output*

Answer:

Figure 10: Zero-shot Prompt for Closed Models in Code Debug Dataset.

Instruction:
Please correct the errors in the buggy code snippet below, ensuring that your corrected code adheres to the specified programming language syntax and logic requirements. Validate your solution against the provided test cases to ensure its accuracy. Note that your solution should strictly consist of the corrected code only.

Question:
Below is the ***cpp*** buggy code:
```cpp\nclass Solution {\npublic:\n unordered_map<long long int,long long int>mp;\n int trailingZeroes(int n) {\n for(int i=n;i>=1;i--)\n fun(i);\n \n\t\t Solution s;\n\t\t return s.min(mp[2],mp[5]); \n }\n void fun(int n)\n {\n for(int i=2;i*i<=n;i++)\n {\n \twhile((n%i) == 0)\n \t{\n \t mp[i]++;\n\t\t\t\tn = n / i;\t\n\t\t\t}\n\t\t }\n\t\t if(n > 1)\n\t\t mp[n]++;\n }\n};```

Correct the code and ensure it passes the following test case:
Input: ***n = 3***
Output: ***0***

Answer:
```cpp\nclass Solution {\npublic:\n unordered_map<long long int,long long int>mp;\n int trailingZeroes(int n) {\n for(int i=n;i>=1;i--)\n fun(i);\n \n\t\t return min(mp[2],mp[5]); \n }\n void fun(int n)\n {\n for(int i=2;i*i<=n;i++)\n {\n \twhile((n%i) == 0)\n \t{\n \t mp[i]++;\n\t\t\t\tn = n / i;\t\n\t\t\t}\n\t\t }\n\t\t if(n > 1)\n\t\t mp[n]++;\n }\n};```

Two Other Examples.

Question:
[The real question that requires reasoning.]

Answer:

Figure 11: Few-shot Prompt for Closed Models in Code Debug Dataset.

C.4 CODE REQUIREMENT SWITCH

Below we present the prompt formats used by closed-source models in the code requirment switch scenario, under zero-shot and few-shot. See details in Figure 16,Figure 17.

Instruction:
Please translate the following code snippet to another programming language, ensuring that your translated code meets the syntax and logic requirements of the target programming language. Validate your solution against the provided test cases to confirm its accuracy. Note that your submission should strictly contain the translated code only.

Question:
Below is the source code snippet in *source_lang*:
source_code

Translate this code to *target_lang*. Ensure your translated code works correctly with the test case provided:
Input: *public_tests_input*
Output: *public_tests_output*

Answer:

Figure 12: Zero-shot Prompt for Closed Models in Code Translate Dataset.

Instruction:
Please translate the following code snippet to another programming language, ensuring that your translated code meets the syntax and logic requirements of the target programming language. Validate your solution against the provided test cases to confirm its accuracy. Note that your submission should strictly contain the translated code only.

Question:
Below is the source code snippet in ***java***:
```java\npublic String longestCommonPrefix(String[] strs) {\n   if (strs.length == 0) return "";\n\n   for (int i = 0; i < strs[0].length(); ++i) {\n   char c = strs[0].charAt(i);\n   for (int j = 1; j < strs.length; ++j) {\n   if (i == strs[j].length() || strs[j].charAt(i) != c) {\n   return strs[0].substring(0, i);\n   }\n   }\n   }\n   return strs[0];\n}```

Translate this code to ***python***. Ensure your translated code works correctly with the test case provided:
Input: ***strs = \["flower ", "flow ", "flight "\]***
Output: ***"fl "***

Answer:
```python\ndef longest_common_prefix(strs):\n   if not strs:\n   return ""\n\n   for i, c in enumerate(strs[0]):\n   for j in range(1, len(strs)):\n   if i == len(strs[j]) or strs[j][i] != c:\n   return strs[0][:i]\n   return strs[0]```

Two Other Examples.

Question:
[The real question that requires reasoning.]

Answer:

Figure 13: Few-shot Prompt for Closed Models in Code Translate Dataset.

Instruction:
Please optimize the given code snippet to enhance its execution efficiency and reduce memory usage, ensuring the accuracy of the code remains unaffected. Validate your solution against the provided test cases to ensure its accuracy. Note that your submission should strictly consist of the optimized code only.

Question:
Below is the source code snippet that needs optimization:
source_code

Optimize the code and ensure it passes the following test case:
Input: *public_tests_input*
Output: *public_tests_output*

Answer:

Figure 14: Zero-shot Prompt for Closed Models in Code Polish Dataset.

Instruction:
Please optimize the given code snippet to enhance its execution efficiency and reduce memory usage, ensuring the accuracy of the code remains unaffected. Validate your solution against the provided test cases to ensure its accuracy. Note that your submission should strictly consist of the optimized code only.

Question:
Below is the source code snippet that needs optimization:
```java\nclass Solution {\n public String convert(String s, int numRows) {\n int n = s.length(), r = numRows;\n if (r == 1 || r >= n) {\n return s;\n }\n int t = r * 2 - 2;\n int c = (n + t - 1) / t * (r - 1);\n char[][] mat = new char[r][c];\n for (int i = 0, x = 0, y = 0; i < n; ++i) {\n mat[x][y] = s.charAt(i);\n if (i % t < r - 1) {\n ++x;\n } else {\n --x;\n ++y;\n }\n }\n StringBuffer ans = new StringBuffer();\n for (char[] row : mat) {\n for (char ch : row) {\n if (ch != 0) {\n ans.append(ch);\n }\n }\n }\n return ans.toString();\n }\n}```

Optimize the code and ensure it passes the following test case:
Input: ***s = "PAYPALISHIRING ", numRows = 3***
Output: ***"PAHNAPLSIIGYIR "***

Answer:
```java\nclass Solution {\n public String convert(String s, int numRows) {\n int n = s.length(), r = numRows;\n if (r == 1 || r >= n) {\n return s;\n }\n StringBuffer[] mat = new StringBuffer[r];\n for (int i = 0; i < r; ++i) {\n mat[i] = new StringBuffer();\n }\n for (int i = 0, x = 0, t = r * 2 - 2; i < n; ++i) {\n mat[x].append(s.charAt(i));\n if (i % t < r - 1) {\n ++x;\n } else {\n --x;\n }\n }\n StringBuffer ans = new StringBuffer();\n for (StringBuffer row : mat) {\n ans.append(row);\n }\n return ans.toString();\n }\n}```

Two Other Examples.

Question:
[The real question that requires reasoning.]

Answer:

Figure 15: Few-shot Prompt for Closed Models in Code Polish Dataset.

Instruction:
Please modify the given code snippet to implement a new function that is related to the original function implemented by the code, ensuring your modified code adheres to the programming language's syntax and logic requirements. Validate your solution against the provided test cases to ensure its accuracy. Your submission should strictly consist of the target code only.

Question:
Below is the code snippet that implements a specific function:
similar_source_code

It currently performs the operation:
Input: *public_similar_tests_input*
Output: *public_similar_tests_output*

You are required to modify this code to implement a new function that is related to the original one, as detailed below:
Input: *public_target_tests_input*
Output: *public_target_tests_output*

Ensure your modified code passes the provided test case.

Answer:

Figure 16: Zero-shot Prompt for Closed Models in Code Switch Dataset.

Instruction:
Please modify the given code snippet to implement a new function that is related to the original function implemented by the code, ensuring your modified code adheres to the programming language's syntax and logic requirements. Validate your solution against the provided test cases to ensure its accuracy. Your submission should strictly consist of the target code only.

Question:
Below is the code snippet that implements a specific function:
```cpp\nint maxBooks(vector<int>& books) {\n   int n = books.size();\n   int max_books = 0;\n   \n   for (int l = 0; l < n; ++l) {\n   int r = l + 1, books_taken = books[l], min_taken = books[l];\n   max_books = max(max_books, books_taken);\n   \n   while (r < n) {\n   min_taken = min(min_taken, books[r]);\n   books_taken += min_taken;\n   max_books = max(max_books, books_taken);\n   \n   if (min_taken <= 1) break; // Can't decrease further\n   ++r;\n   }\n   }\n   \n   return max_books;\n}```

It currently performs the operation:
Input: ***books = \[8,5,2,7,9\]***
Output: ***19***

You are required to modify this code to implement a new function that is related to the original one, as detailed below:
Input: ***nums = \[1,4,3,7,4,5\], k = 3***
Output: ***15***

Ensure your modified code passes the provided test case.

Answer:
```cpp\nint minMoves(vector<int>& nums, int limit) {\n   int n = nums.size();\n   vector<int> delta(2 * limit + 2, 0);\n\n   for (int i = 0; i < n / 2; i++) {\n   int lo = min(nums[i], nums[n - 1 - i]) + 1;\n   int hi = max(nums[i], nums[n - 1 - i]) + limit;\n   delta[lo]--;\n   delta[nums[i] + nums[n - 1 - i]]--;\n   delta[hi + 1]++;\n   }\n   int moves = n, current = n;\n   for (int i = 2; i <= 2 * limit; i++) {\n   current += delta[i];\n   moves = min(moves, current);\n   }\n   return moves;\n}```

Two Other Examples.

Question:
[The real question that requires reasoning.]

Answer:

Figure 17: Few-shot Prompt for Closed Models in Code Switch Dataset.

D CODE PROCESSING AND TEMPLATE INTEGRATION

We developed executable templates for C++, Python, and Java, as shown in Figure 18,Figure 19 and Figure 20. Submitted code only needs to contain core functions and the high-level function name used for main function calls, aligning with the submission requirements of LeetCode. For code parsing and function name extraction, all code uses tree-sitter[4] - an incremental parsing system for programming tools to retrieve function calls to ensure that the correct high-level function name is obtained.

```
### Include Other Necessary Headers
#include <bits/stdc++.h>
#include "../leetcode_template/cpp/LeetcodeIO.h"
using namespace std;

class Solution {
public:
        ### Function bodies to be tested
};

int main() {
    REGISTER_CONSTRUCTOR_SOLUTION;
    REGISTER_MEMBERFUNCTION_SOLUTION(### High-level Function name to be called);
    while (true) {
        executor.constructSolution();
        executor.executeSolution();
    }
}
```

Figure 18: C++ Template.

[4] https://github.com/tree-sitter/tree-sitter

```python
### Include Other Necessary Headers
import json
import sys
from parse_input import *
from leetcode_class import ListNode, Node, TreeNode
from typing import List

parse_function_map = {
    '"Node"': parse_node,
    '"Optional[Node]"': parse_node,
    '"TreeNode"': parse_treeNode,
    "ListNode": parse_listNode,
    "List['Node']": parse_list_node,
    "List[List[int]]": parse_list_list_int,
    "List[List[str]]": parse_list_list_str,
    "List[Optional[ListNode]]": parse_list_listNode,
    "List[TreeNode]": parse_list_treeNode,
    "List[bool]": parse_list_bool,
    "List[float]": parse_list_float,
    "List[int]": parse_list_int,
    "List[str]": parse_list_str,
    "Optional['Node']": parse_node,
    "Optional[ListNode]": parse_listNode,
    "Optional[TreeNode]": parse_treeNode,
    "TreeNode": parse_treeNode,
    "bool": parse_bool,
    "float": parse_float,
    "int": parse_int,
    "str": parse_str,
    "treeNode": parse_treeNode,
}

class Solution:
        ### Function bodies to be tested

if __name__ == '__main__':

    object_func_name = ### High-level Function name to be called

    func_input_type_list = ### Input Type of the calling Function

    while True:
        try:
            input_data = []
            for _ in range(len(func_input_type_list)):
                input_data.append(input())
            input_argus = []
            for input_data_item, input_data_type in zip(input_data, func_input_type_list):
                input_argus.append(parse_function_map[input_data_type](input_data_item))
            s = Solution()
            func = getattr(s, object_func_name)
            output = func(*input_argus)
            print(output)
        except EOFError:
            break
```

Figure 19: Python Template.

```
. ### Include Other Necessary Headers
import com.template.Node;
import com.template.ListNode;
import com.template.TreeNode;
import com.template.ParseInputUtil;
import java.lang.reflect.Method;
import java.lang.reflect.Type;
import java.util.*;
class Solution {
        . ### Function bodies to be tested
};
public class Main
{
    public static void main( String[] args )
    {
        String methodName = . ### High-level Function name to be called;

        Scanner scanner = new Scanner(System.in);
        Method[] methods = Solution.class.getMethods();
        Method method = null;
        int numberOfParams = 0;
        for (Method method_check : methods) {
            if (method_check.getName().equals(methodName)) {
                method = method_check;
                Type[] genericParameterTypes = method.getGenericParameterTypes();
                numberOfParams = genericParameterTypes.length;
                break;
            }
        }
        while(scanner.hasNext()) {
            List<String> stringParams = new ArrayList<String>();
            for (int i = 0; i < numberOfParams; i++) {
                String line = scanner.nextLine();
                stringParams.add(line);
            }
            Solution s = new Solution();
            try {
                Type[] genericParameterTypes = method.getGenericParameterTypes();
                List<Object> parsedParams = new ArrayList<>();
                for (int i = 0; i < genericParameterTypes.length; i++) {
                    Type paramType = genericParameterTypes[i];
                    String stringParam = stringParams.get(i);
                    if (stringParam.contains("None")) {
                        stringParam = stringParam.replace("None", "null");
                    }
                    Object parsedParam = ParseInputUtil.parseStringToType(stringParam, paramType);
                    parsedParams.add(parsedParam);
                }
                Object result = method.invoke(s, parsedParams.toArray());
                System.out.println(result);
            } catch (Exception e) {
                e.printStackTrace();
            }
        }
        scanner.close();
    }
}
```

Figure 20: Java Template.

E PASS CRITERIA

Table 5 shows our detailed pass criteria.

Scenario	Time limit	Memory limit	Pass Criteria
Debug	300s for all test cases. 30s for single test case.	512MB	Pass all test cases.
Translate			Pass all test cases in target language.
Switch			Pass all test cases.
Polish	$\bar{T}$ms	$\bar{M}$MB	Pass all test cases and $\bar{T}_{\text{avg}} < \bar{T}$ or $\bar{M}_{\text{avg}} < \bar{M}$

Table 5: Pass criteria in four scenarios. $\bar{T}$ and $\bar{M}$ are obtained by averaging 20 runs of the standard code. During the judging process, the LLM-generated code is run twice to obtain $\bar{T}_{\text{avg}}$ and $\bar{M}_{\text{avg}}$.

F EVALUATION CONFIGURATION

Our OJ is built on a server equipped with a high-performance Intel(R) Xeon(R) Platinum 8480C processor boasting 224 cores. It supports up to 4 petabytes (PB) of physical memory and has a virtual memory space of up to 128 terabytes (TB). The judging environment for C++ is based on the g++ compiler, version 9.4.0, utilizing the C++17 standard. For Python, the judging environment is based on Python 3.8.10, and for Java, it relies on OpenJDK version 11.0.22. The previously mentioned template utilizes the gson-2.9.1.jar library to process the input.