M^{2}RC-EVAL: Massively Multilingual Repository-level Code Completion Evaluation

Jiaheng Liu[1,2*,†], **Ken Deng**[2*], **Congnan Liu**[2], **Jian Yang**[2], **Shukai Liu**[2], **He Zhu**[2], **Peng Zhao**[2], **Linzheng Chai**[2], **Yanan Wu**[2], **Ke Jin**[2], **Ge Zhang**[3], **Zekun Wang**[2], **Guoan Zhang**[2], **Yingshui Tan**[2], **Bangyu Xiang**[2], **Zhaoxiang Zhang**[4], **Wenbo Su**[2], **Bo Zheng**[2]

[1]Nanjing University, [2]Alibaba Group, [3]University of Waterloo, [4]CASIA

liujiaheng@nju.edu.cn

Abstract

Repository-level code completion has drawn great attention in software engineering, and several benchmarks have been introduced. However, existing repository-level code completion benchmarks usually focus on a limited number of languages (<5), which cannot evaluate the general code intelligence abilities across different languages for existing code Large Language Models (LLMs). Besides, the existing benchmarks usually report overall average scores of different languages, where the fine-grained abilities in different completion scenarios are ignored. Therefore, to facilitate the research of code LLMs in multilingual scenarios, we propose a massively multilingual repository-level code completion benchmark covering 18 programming languages (called **M^{2}RC-EVAL**), and two types of fine-grained annotations (i.e., **bucket-level** and **semantic-level**) on different completion scenarios are provided, where we obtain these annotations based on the parsed abstract syntax tree. Moreover, we also curate a massively multilingual instruction corpora **M^{2}RC-INSTRUCT** dataset to improve the repository-level code completion abilities of existing code LLMs. Comprehensive experimental results demonstrate the effectiveness of our M^{2}RC-EVAL and M^{2}RC-INSTRUCT.

1 Introduction

The emergence of Large Language Models (LLMs) has marked a significant advancement in many tasks (Liu et al., 2025; Zhang et al., 2024; Wang et al., 2024; Liu et al., 2024). The code LLMs (Roziere et al., 2023; Zheng et al., 2023; Guo et al., 2024a; Hui et al., 2024; Zhang et al., 2025; He et al., 2025; Huang et al., 2024) pre-trained on extensive datasets comprising billions of code-related tokens further revolutionize the automation of software development tasks, providing contextually relevant code suggestions and facilitating the translation from natural language to code. The generation capability of code LLMs opens up diverse applications in software development, promising to enhance productivity and streamline coding processes. As the field continues to evolve, it presents exciting opportunities for future developments and innovations in automated programming and code assistance.

The code completion task is crucial in modern software development, enhancing coding efficiency and accuracy by predicting and suggesting code segments based on context. Recent advancements in code LLMs (Bavarian et al., 2022a) have introduced sophisticated completion techniques, such as prefix-suffix-middle (PSM) and suffix-prefix-middle (SPM) paradigms, which can complete middle code segments given the surrounding context. However, the current benchmark (Ding et al., 2024; Liu et al., 2023a) mainly focuses on several programming languages. For example, the CrossCodeEval (Ding et al., 2024) includes four languages (i.e., Python, Java, TypeScript, C#). Besides, existing benchmarks can only provide the average score among all samples, which cannot provide a language-specific evaluation for different programming languages based on their intrinsic structure. *Inspired by the multilingual in-file code generation benchmark MultiPL-E (Cassano et al., 2022) and McEval (Chai et al., 2024), we create a massively multilingual repository-level code completion Evaluation benchmark called* **M^{2}RC-EVAL** *to facilitate the research of the community.*

In this paper, as shown in Fig. 1, our M^{2}RC-EVAL includes 18 programming languages with two types of fine-grained annotations (i.e., **bucket-level** and **semantic-level**), where each language contains 100 validation and 500 test samples, respectively. Specifically, for the bucket-level annotations, we first generate abstract syntax tree with

* First two authors contributed equally.

† Corresponding Author: Jiaheng Liu.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15661–15684

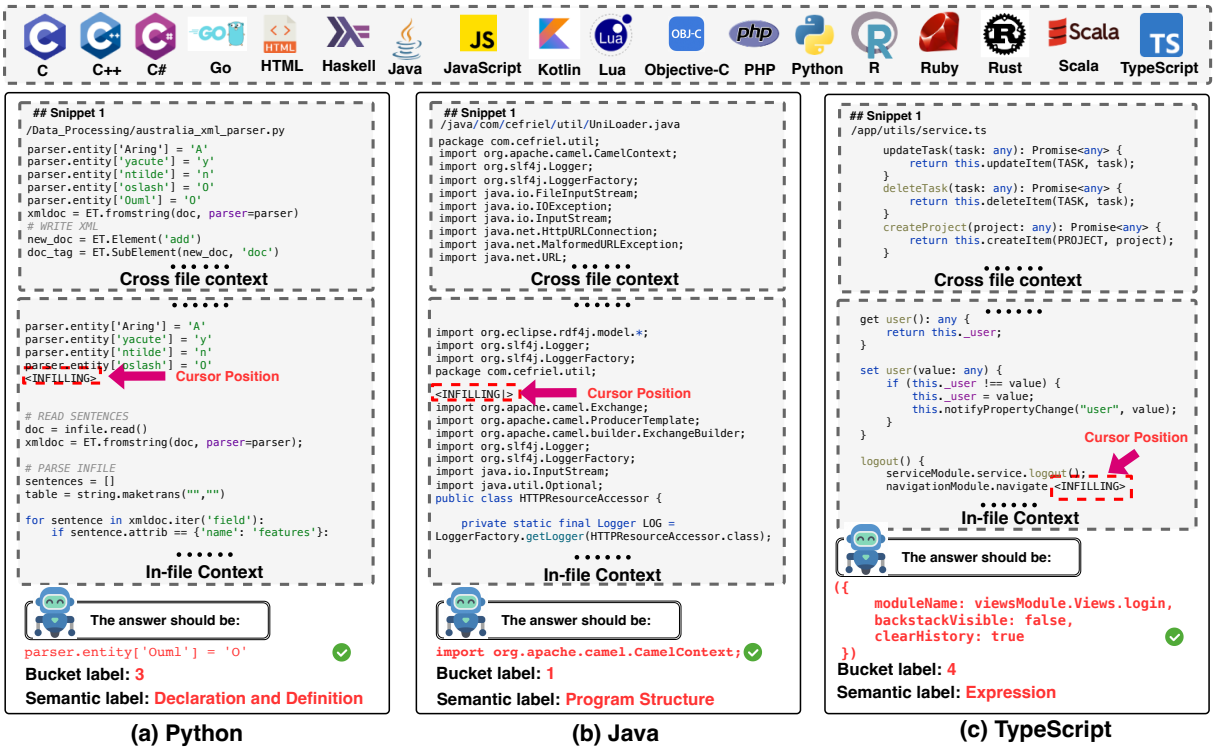


Figure 1: Overview of our M^{2}RC-EVAL with 18 languages. Specifically, first, we provide three samples from different languages (i.e., Python, Java, TypeScript) for illustration, where the bucket label and semantic label for the corresponding cursor position are provided. Second, the code LLMs need to predict the completion results given the in-file context from the current code file and the cross file context retrieved from other code files in the current repository. Note that "< INFILLING >" denotes that the current position will be triggered for code completion.

N layers using code parser (i.e., Treesitter [1]), and divide these N into fixed M buckets, Then, for each completion cursor position, we annotate the corresponding bucket-level label based on the layer to which the location belongs. In this way, we can obtain different code completion scenarios with different difficulties.

For the semantic-level annotations, inspired by (Takerngsaksiri et al., 2024), we first pre-define 11 major semantic labels (e.g., Program Structure, Statement) for each completion cursor position, which aims to analyze the fine-grained performance across different code semantics. Note that as different languages usually have specific syntax, we carefully design the subcategories under each major semantic label for different languages. Then, as the code parser usually provides syntax labels (e.g., functions, variables, classes, empty lines)[2] for each completion cursor position, we carefully define the mappings between the syntax labels to our designed semantic labels and build the semantic-level annotations for our M^{2}RC-EVAL. Finally, to enhance the performance of repository-level code completion for existing code LLMs, we also create a massively multilingual instruction corpora **M^{2}RC-INSTRUCT** of 18 languages.

The contributions are summarized as follows:

- We propose the first massively multilingual repository-level code completion benchmark M^{2}RC-EVAL covering 18 languages, where two types of annotations (bucket-level and semantic-level labels) are provided based on the parsed abstract syntax tree.
- We introduce M^{2}RC-INSTRUCT, the massively multilingual repository-level code instruction corpora covering the multilingual code snippet from 18 languages, which can greatly enhance the performance of repository-level code completion results.
- Comprehensive evaluation results and analysis demonstrate the effectiveness of our proposed M^{2}RC-EVAL and M^{2}RC-INSTRUCT.

[1] https://tree-sitter.github.io/tree-sitter/

[2] Note that the syntax label provided by code parser (e.g., tree-sitter) are highly detailed.

2 Related Works

Code Large Language Models. Code LLMs (Chen et al., 2021; Zhao et al., 2024; Black et al., 2022; Le et al., 2022; Chowdhery et al., 2023; Nijkamp et al., 2023; Fried et al., 2023; Xu et al., 2022; Jain et al., 2024b; Zhuo et al., 2025) are increasingly involved in modern programming, due to excellent capabilities of code generation (Li et al., 2022; Allal et al., 2023), code repair (Wang et al., 2021, 2023), code translation (Li et al., 2023), and other coding tasks. Recent code LLMs such as Code Llama (Roziere et al., 2023), DeepSeek-Coder (Guo et al., 2024a), and Qwen2.5-Coder (Hui et al., 2024) incorporate the fill-in-the-middle (FIM) task into their training stage for code completion. Moreover, many in-file benchmarks are proposed to evaluate capabilities of code LLMs (Zheng et al., 2023; Austin et al., 2021; Jain et al., 2024a; Chai et al., 2024).

Repository-level Code Completion. The latest repository-level code completion methods (Bairi et al., 2023; Phan et al., 2024; Liao et al., 2023; Shrivastava et al., 2023a; Agrawal et al., 2023; Shrivastava et al., 2023b; Pei et al., 2023; Zhang et al., 2023; Yu et al., 2024; Ding et al., 2022) aim to retrieve related code snippets across files within a repository, where existing datasets mainly focus on limited programming languages. For example, RepoBench (Liu et al., 2023b) and CrossCodeEval (Ding et al., 2023) only support 2 and 4 languages, respectively. To comprehensively evaluate the multilingual repository-based code completion, we propose M^{2}RC-EVAL with 18 languages.

3 M^{2}RC-EVAL

3.1 Data Collection

The Overall Data Pool. We collect the training data from The Stack v2 (Lozhkov et al., 2024) with permissively licensed repositories from GitHub. Further, we keep only repositories receiving more than 5 stars and containing $[10, 50]$ files. Lastly, we preserve files written in 18 common languages, and obtain 431,353,244 files.

Completion Cursor Position Selection. Completion cursor position selection significantly impacts the quality of a code completion benchmark. Previous studies (Ding et al., 2024; Liu et al., 2023a) randomly select a segment of consecutive characters as the completion span, which does not guarantee the integrity of identifiers and statements.

Table 1: A comparison with existing notable other datasets. "**FG**" denotes "**Fine-grained**".

Benchmark	# Lang	FG	Training	# Test Repos
RepoBench	2	✗	✓	1669
CrossCodeEval	4	✗	✗	1002
Ours	18	✓	✓	5993

Besides, recent works (e.g., Qwen2.5-Coder (Hui et al., 2024), aiXcoder (Jiang et al., 2024)) also claimed that developers often expect LLMs to complete the current code into a complete snippet, such as a completed code line or loop block, instead of suggesting an incomplete code snippet. Therefore, in M^{2}RC-EVAL, we first parse the abstract syntax tree (AST) of each source code file, and then we randomly choose a node (e.g., the node of "Function Definition" in Fig. 3) on the AST as the completion cursor position. After that, we obtain the corresponding code to obtain the ground-truth for the current completion cursor position. Finally, at inference, the code LLMs need to predict the current code span given the in-file and cross file contexts. Similarly, in training, we just use the ground-truth to tune code LLMs.

3.2 Quality Control

We build a suite of post-processing filters to enhance the quality of M^{2}RC-INSTRUCT. We eliminate examples based on two heuristic rules: (1) The completion cursor position should be no longer than 5 lines. (2) If the completion ground truth is fewer than 20 characters, at least 20% of them should be alphabetic. To improve data independence and inference difficulty, we apply extra filters to the test cases in M^{2}RC-EVAL. (a) Repositories in M^{2}RC-EVAL should be absent from M^{2}RC-INSTRUCT. (b) We ensure that 30% of the completion ground truth is not shorter than 2 lines. (c) The completion cursor position should not be fully white-spaced. (d) We discard test cases that could be exactly predicted by `DeepSeekCoder-1.3B` (Guo et al., 2024b) without cross file contexts.

3.3 Dataset Statistics

Following the quality filters in §(3.2) from the overall data pool §(3.1). We sample 50,000 files per language to construct our M^{2}RC-INSTRUCT, and sample 100, and 500 files per language to build the validation and test sets of our M^{2}RC-EVAL, respectively. The statistics of the test set are shown in Fig. 2, and we also provide a detailed com-

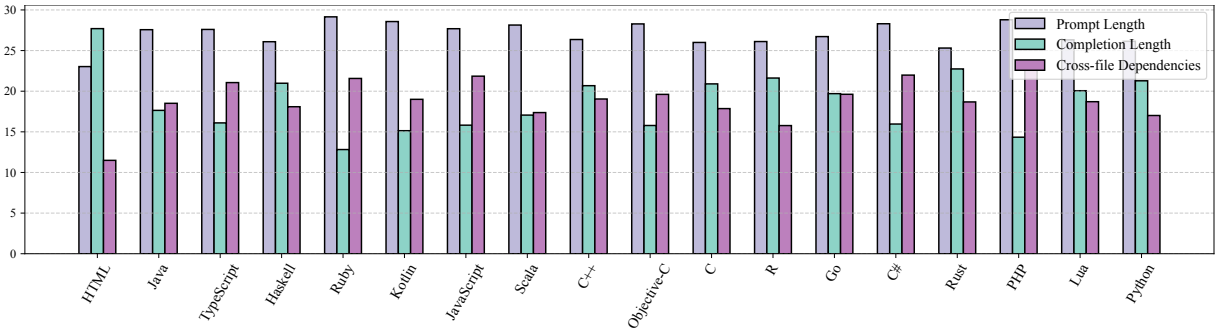


Figure 2: The average prompt length (100x tokens), completion span length (50x tokens), and cross-file dependencies (1x) in the testing set of M^{2}RC-EVAL. We define the number of other files, which are explicitly imported and implicitly referenced by the current file, as cross-file dependencies.

Table 2: Semantic-level annotations on different types of programming languages.

Major Classes	**Java**	**Go**	**Scala**
Program Structure	"Program Entry", "Namespace", "Import/Include"	"Program Entry", "Namespace", "Import/Include"	"Program Entry", "Namespace", "Import/Include"
Declaration and Definition	"Class", "Function", "Variable"	"Class", "Function", "Variable"	"Class", "Function", "Variable"
Control Flow Structure	"Conditional", "Loop", "Jump", "Exception Handling"	"Conditional", "Loop", "Jump", "Exception Handling"	"Conditional", "Loop", "Jump", "Exception Handling"
Expression	"Arithmetic Operation", "Logical Operation", "Function Call", "Object Creation", "Type Casting", "Other", "Arithmetic Operator", "Logical Operator"	"Arithmetic Operation", "Logical Operation", "Function Call", "Object Creation", "Type Casting", "Arithmetic Operator", "Logical Operator"	"Arithmetic Operation", "Function Call", "Object Creation", "Type Casting", "Tuple Expression", "Logical Operator", "Special Operator"
Data Type	"Primitive Type", "Composite Type", "Generic", "Numeric", "String", "Boolean", "Special Value"	"Primitive Type", "Composite Type", "Generic"	"Primitive Type", "Composite Type", "Generic", "Numeric", "String", "Boolean", "Special Value"
Statement	"Expression Statement", "Compound Statement", "Other Statement"	"Expression Statement", "Compound Statement"	"Compound Statement"
Modifier and Attribute	"Access Modifiers", "Other Modifiers", "Attribute Annotation"	"Access Modifiers", "Other Modifiers", "Attribute Annotation"	"Access Modifiers", "Other Modifiers", "Annotation"
Comments and Documentation	"Single-line Comment", "Multi-line Comment"	"Single-line Comment"	"Single-line Comment", "Multi-line Comment"
Preprocessing Directive	"Conditional Compilation", "Macro Definition"	"Conditional Compilation", "Macro Definition"	"Macro Definition"
Identifier and Scope	"Identifier", "Qualified Name"	"Identifier", "Qualified Name"	"Identifier", "Qualified Name", "Binding", "Delimiter"
Special Language Structure	"Lambda Expression", "Pattern Matching", "Coroutine"	"Lambda Expression", "Coroutine"	"Lambda Expression", "Pattern Matching"

parison between our M^{2}RC-EVAL with existing repository-level code completion datasets in Table 1. Note that the numbers of repositories for M^{2}RC-INSTRUCT, validation split of M^{2}RC-EVAL are 37439 and 1635, respectively.

3.4 Fine-grained Annotations

As shown in Fig. 3, to analyze the performance in a fine-grained manner, we further provide two types of fine-grained annotations (i.e., bucket-level and semantic-level) for each completion cursor. Specifically, we first generate the abstract syntax tree. For the bucket-level annotations, we first simply divide each tree into M buckets based on the depth degree of the abstract syntax tree. Note that we set M as 10 in our M^{2}RC-EVAL. For example, if the number of layers for the current abstract syntax tree is N, the i-th layer of the tree belongs to the $\lceil \frac{i}{N/M} \rceil$ bucket. Then, for each completion cursor node, we annotate the bucket label based on the layer number of each node. Similarly, for the semantic-level annotations, we directly annotate the semantic-level label for each completion cursor node. Specifically, we pre-define 11 major classes (i.e., "**Program Structure**", "**Declaration and Definition**", "**Control Flow Structure**", "**Expression**", "**Data Type**", "**Statement**", "**Modifier and Attribute**", "**Comments and Documentation**", "**Preprocessing Directive**", "**Identifier and Scope**", "**Special Language**

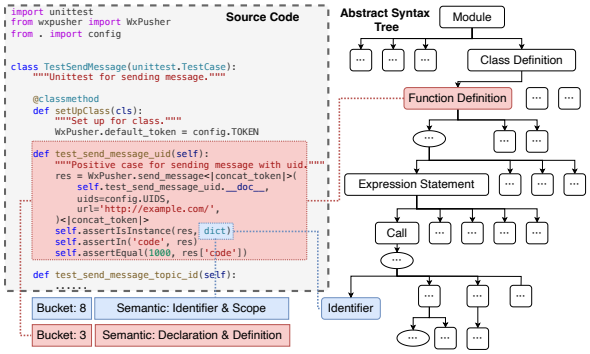


Figure 3: Illustration on generating completion cursor position and fine-grained annotations. Specifically, we first parse the source code into an abstract syntax tree (AST). Then, we choose one node as the completion cursor position and generate the bucket label based on the belonged layer number in AST, and obtain the semantic label based on the node type parsed by the Tree-sitter.

Structure"). Then, as different languages have many specific designs, the subcategories under each major class are carefully annotated for different languages. In Table 2, we provide the semantic-level annotations on three main-stream programming languages (Java, Go, Scala), where the annotations on all 18 languages are provided in Fig. 10, Fig. 11 and Fig. 12 of the Appendix.

4 Experiments

4.1 Evaluation Models and Metrics

We evaluate three Code LLMs (i.e., **StarCoder-7B** (Li et al., 2023), **DeepSeekCoder-6.7B** (Guo et al., 2024b) and **Code Llama-7B** (Roziere et al., 2023)) (See Appendix A.2 for more details) on M^{2}RC-EVAL. Following (Ding et al., 2023), we compare the generated code with the reference and compute the exact match (**EM**) and edit similarity (**ES**) metrics [3], which assess the textual similarities and ignore semantic structure similarities among predictions and ground-truth.

4.2 Experimental Setup

Baseline. Only the original code file, where the cursor position is located, is provided for the code LLMs. As no explicit inter-file context is supplied, the model must utilize its inherent knowledge-based reasoning abilities to generate code.

+ Retrieval. In line with the approach outlined in CrossCodeEval (Ding et al., 2023), the retrieval process begins by examining files within the same repository. Continuous code segments of L lines are extracted, where L matches the length of the retrieval query and is set as 10 by default. Subsequently, these extracted candidates are prioritized based on their Jaccard similarity scores. The most relevant fragments are then appended to the beginning of the in-file context in descending order of similarity. This concatenation continues until the total length, including both the added candidates and the original in-file context, reaches the predetermined maximum token limit of 4096.

+ Retrieval & Tuning. To further improve the performance of repository-level code completion, we fine-tune code LLMs on M^{2}RC-INSTRUCT mentioned in §(3). At inference, we use the same inference strategy as discussed in "+ Retrieval".

4.3 Main Results

We present the results on M^{2}RC-EVAL in Table 3. We observe that different code LLMs have different repository-level code completion abilities for different programming languages. For instance, DeepSeekCoder-6.7B demonstrates strong completion ability for Go, while its performance is weaker with HTML, a markup language, which demonstrates the necessity of evaluating code LLMs for multilingual capabilities. Besides, the results indicate that cross-file context is highly effective, resulting in a significant improvement compared to using only in-file context. In particular, the multilingual SFT on our created instruction corpora M^{2}RC-INSTRUCT also significantly enhances performance on M^{2}RC-EVAL. Notably, after SFT on M^{2}RC-INSTRUCT, Code Llama-7B, which originally ranked lowest with in-file context, outperformed the non-finetuned StarCoder-7B, demonstrating the effectiveness of M^{2}RC-INSTRUCT.

4.4 Analysis

Analysis on different model sizes. In Table 4, we provide the results on the validation set of M^{2}RC-EVAL. Notably, StarCoder-7B consistently outperforms StarCoder-3B under comparable conditions. However, after SFT on M^{2}RC-INSTRUCT, the results of StarCoder-3B exceed those of the inference-only StarCoder-7B. This finding underscores the effectiveness of our M^{2}RC-INSTRUCT in augmenting the capabilities of smaller models in repository-level code completion.

Analysis on different training data sizes. In Table 5, we evaluate the fine-tuned StarCoder-7B by employing varying sizes of M^{2}RC-INSTRUCT and report the results on the validation set of M^{2}RC-

[3] https://github.com/amazon-science/cceval

Table 3: Exact match (%) and edit similarity (%) performance on M^{2}RC-EVAL.

Model	C		C#		C++		Go		HTML		Haskell		-	
	EM	ES	EM	ES	EM	ES	EM	ES	EM	ES	EM	ES	EM	ES
Code Llama-7B	18.6	47.2	19.6	52.6	21.8	51.1	26.0	53.6	20.6	40.4	22.6	48.9	-	-
+ Retrieval	21.8	47.2	22.9	48.9	23.2	46.6	23.8	52.4	12.6	35.6	22.6	48.9	-	-
+ Retrieval & Tuning	45.4	72.0	43.5	72.3	50.8	74.9	43.4	72.9	41.8	63.6	39.8	66.3	-	-
StarCoder-7B	20.0	50.4	20.0	53.3	22.4	51.8	25.4	58.2	17.4	40.7	25.0	51.1	-	-
+ Retrieval	23.8	47.8	27.1	53.2	24.6	48.0	26.0	53.6	20.6	40.4	25.0	47.7	-	-
+ Retrieval & Tuning	47.0	72.7	45.1	74.8	52.4	76.3	43.2	73.7	45.8	67.1	44.8	70.2	-	-
DeepSeekCoder-6.7B	22.4	53.7	21.4	56.2	23.2	54.2	29.4	61.4	17.6	43.4	25.2	51.3	-	-
+ Retrieval	28.2	52.6	25.3	52.6	27.6	52.2	29.4	61.4	17.6	43.4	25.8	51.0	-	-
+ Retrieval & Tuning	48.6	75.2	47.9	76.9	54.4	78.2	48.8	78.4	45.0	66.3	45.8	72.0	-	-
Model	**Java**		**JavaScript**		**Kotlin**		**Lua**		**Objective-C**		**PHP**		-	
Code Llama-7B	23.4	58.5	17.2	52.0	23.6	57.0	20.0	45.7	17.8	49.5	19.2	54.9	-	-
+ Retrieval	23.4	57.5	19.6	48.0	20.8	50.0	19.6	42.2	21.4	46.6	21.2	49.0	-	-
+ Retrieval & Tuning	41.8	74.1	38.8	70.1	45.0	75.6	43.8	70.5	49.8	75.9	45.6	76.7	-	-
StarCoder-7B	24.0	59.2	16.6	52.0	24.4	59.3	21.4	48.6	17.6	49.6	18.6	54.4	-	-
+ Retrieval	25.0	53.1	22.0	50.8	22.8	52.6	26.4	48.5	23.6	48.0	18.6	54.4	-	-
+ Retrieval & Tuning	47.4	76.9	38.8	70.1	45.0	75.6	43.8	70.5	50.8	75.9	45.6	76.7	-	-
DeepSeekCoder-6.7B	22.2	61.0	20.4	56.5	26.0	61.0	22.0	48.8	21.0	55.6	24.2	58.6	-	-
+ Retrieval	21.6	51.4	24.4	53.6	26.0	61.0	22.0	49.9	27.6	53.5	28.6	56.9	-	-
+ Retrieval & Tuning	48.2	79.1	43.6	73.5	46.0	75.7	44.6	70.6	52.2	77.6	49.8	78.8	-	-
Model	**Python**		**R**		**Ruby**		**Rust**		**Scala**		**TypeScript**		**Avg.**	
Code Llama-7B	24.6	54.2	15.2	41.2	17.2	45.8	26.2	56.0	22.8	48.5	23.4	52.3	19.4	50.3
+ Retrieval	17.4	46.4	15.2	39.8	17.2	42.3	26.0	51.3	22.8	48.5	19.4	48.6	20.2	46.1
+ Retrieval & Tuning	39.2	69.9	38.6	65.5	43.0	68.5	42.0	69.2	41.0	70.1	37.0	68.2	41.9	70.0
StarCoder-7B	19.4	52.9	16.4	43.7	19.4	47.4	26.2	56.0	23.6	53.4	19.8	53.3	21.0	52.0
+ Retrieval	24.6	54.2	22.6	47.2	23.6	47.4	26.4	53.5	22.8	48.5	23.4	52.3	24.1	50.0
+ Retrieval & Tuning	39.2	69.9	41.0	66.6	43.0	68.5	45.8	72.6	43.6	71.5	39.2	69.7	44.5	72.2
DeepSeekCoder-6.7B	21.8	55.1	19.4	48.5	23.6	52.2	23.8	54.3	24.6	56.7	19.4	55.4	22.6	54.7
+ Retrieval	21.8	55.1	19.4	48.5	23.6	52.2	23.8	54.3	22.4	50.4	26.0	54.5	25.1	51.7
+ Retrieval & Tuning	41.6	71.3	45.4	69.4	45.6	70.3	47.6	73.4	44.8	73.7	43.2	73.4	**46.8**	**74.1**

Table 4: Performance on M^{2}RC-EVAL.

Model	EM	ES
StarCoder-3B	14.9	43.5
+ Retrieval	14.6	38.4
+ Retrieval & Tuning	41.7	69.1
StarCoder-7B	20.6	49.9
+ Retrieval	23.6	49.3
+ Retrieval & Tuning	44.4	71.4

Table 5: Performance under different training data sizes.

Data Size (Per lang.)	100	1k	5k	10k	50k
EM (Avg.)	23.4	35.7	40.5	42.4	44.4
ES (Avg.)	49.1	62.9	68.2	69.4	71.4

EVAL. Our observations indicate that increasing the dataset from 0.1k to 50k samples per language yields improved results. This suggests that more training data can help boost the model's performance. Therefore, we select 50k samples per language as the default training set size.

Analysis on the granularity of different bucket levels. As mentioned in §(3.4), we categorize M^{2}RC-EVAL into ten bucket levels based on the positions of the code requiring completion within the abstract syntax tree. As shown in Fig. 4, we presents the performance of StarCoder-7B on the test set of M^{2}RC-EVAL across these different bucket levels, and we observe that as the bucket level decreases, the performance of StarCoder-7B correspondingly declines, which means that the

Table 6: CodeBLEU results on ten representative programming languages.

Model	C	C#	C++	Go	Java	JavaScript	PHP	Python	Ruby	Rust	Avg.
StarCoder-7B	48.3	48.9	50.4	51.5	50.6	46.4	48.2	46.4	46.1	50.4	48.7
+ Retrieval	50.1	52.3	51.1	52.5	51.4	49.3	52.2	49.3	49.1	51.4	50.9
+ Retrieval & Tuning	56.0	57.4	57.6	57.0	57.6	54.8	57.8	52.0	52.9	55.5	**55.9**

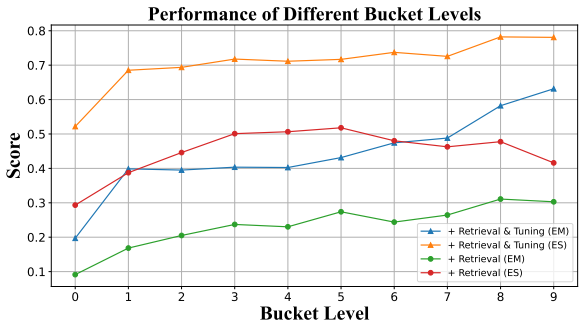


Figure 4: Effectiveness of different bucket levels.

Table 7: Performance on M^{2}RC-EVAL.

Model	Syntax Accuracy
StarCoder-7B	82.8
+ Retrieval	83.9
+ Retrieval & Tuning	96.9

code completion on the shallow layer is usually more challenging than on the deep layer. For more experimental data on single-language completion performance and its relation to bucket levels, please refer to Fig.7 and Fig.8 in the Appendix. These findings suggest that the code LLMs encounter challenges when addressing shallow nodes within the syntax tree during the code completion process.

Analysis on the granularity of different semantic levels. Similarly, in §(3.4), we also categorize the nodes within the abstract syntax tree into eleven primary semantic levels based on their semantic characteristics, and we provide the performance of StarCoder-7B for these various semantic levels across multilingual languages on the test set of the M^{2}RC-EVAL, as illustrated in Fig. 5. Notably, we observe significant performance disparities across different semantic levels. Specifically, StarCoder-7B shows superior performance on "Identifier and Scope", while it exhibits lower efficacy on "Special Language Structure", This suggests that current code LLMs are proficient at completing tasks related to variable definitions and references, yet their capacity to handle characteristics of different languages requires further enhancement. For single-language completion performance across various node types, please refer to Fig. 9 in the Appendix.

Analysis on completion on different lines. As shown in Fig.6, StarCoder-7B can effectively complete tasks involving a small number of lines. However, as the number of lines to be completed increases, the scores of the generated code gradually decline. This indicates that completing multi-line code remains a challenge for code LLMs.

Analysis on various input lengths. In Fig. 6, we report the results produced by StarCoder-7B ("Retrieval & Tuning") on our M^{2}RC-EVAL when the input lengths of range in {512, 1024, 2048, 4096} tokens. In Fig. 6, we observe that a scaling law exists, where better performance is achieved when the input length is larger. Thus, we set the default input length as 4096 tokens.

Analysis on CodeBLEU metric. In Table 3, we mainly report the EM and ES metrics based on the textual similarity, which neglects important syntactic and semantic features of codes and underestimates different outputs with the same semantic logic. Thus, the CodeBLEU (Ren et al., 2020) [4] is proposed, which considers information from not only the shallow match, but also the syntactic match and the semantic match. In Table 6, we report the results of 10 popular programming languages using the test split of M^{2}RC-EVAL based on the StarCoder-7B model and observe that we can still achieve better performance by fine-tuning on our constructed M^{2}RC-INSTRUCT, which further demonstrates the effectiveness of our M^{2}RC-INSTRUCT on repository-level code completion.

Analysis on syntax accuracy. In Table 7, for syntax static analysis, to further verify the syntax correctness of the predicted code snippets, we use the code static checking tools (Tree-Sitter) for all predicted code snippets of M^{2}RC-EVAL. Specifically, we parse the code snippet into the abstract syntax tree and filter out the code snippet, where the parsed nodes in the code snippet have parsing errors. For

[4]We test the CodeBLEU metric based on https://github.com/k4black/codebleu.

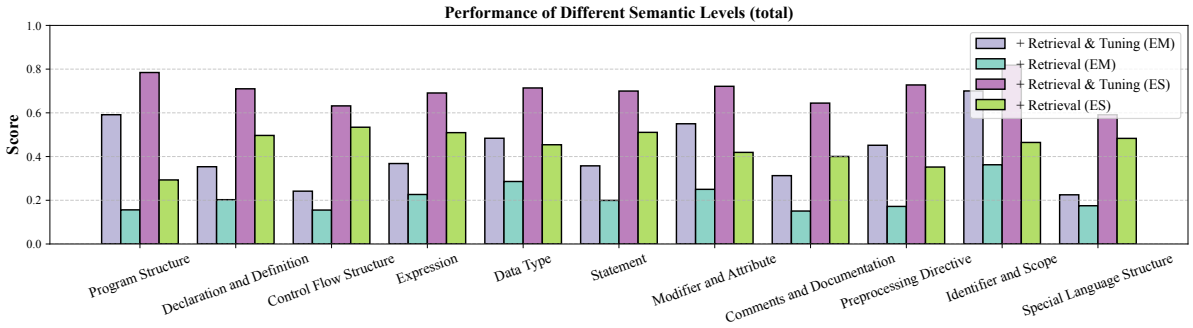


Figure 5: Effectiveness of different semantic levels based on StarCoder-7B.

Table 8: Results of M^{2}RC-EVAL based on paradigm types.

Paradigm Types	StarCoder	+ Retrieval	+ Retrieval & Tuning
Procedural	19.2	23.7	47.6
Object Oriented	21.3	24.2	47.4
Multiple Paradigms	21.1	24.3	43.4
Functional	25.1	24.8	44.6
Markup Language	17.2	21.3	46.7

Table 9: Results of M^{2}RC-EVAL based on application scenarios.

Application Scenarios	StarCoder	+ Retrieval	+ Retrieval & Tuning
Mobile	21.2	22.9	48.0
Cross Platform	24.2	25.3	48.1
Desktop Application	18.7	26.1	45.3
Web Frontend	17.9	22.2	41.5
Web Backend	22.8	24.8	44.6
Scientific Computing	18.2	23.5	40.3
System & Software	21.3	24.0	50.1
Education & Research	25.1	24.8	44.6
Automation Scripts	21.7	26.3	43.7

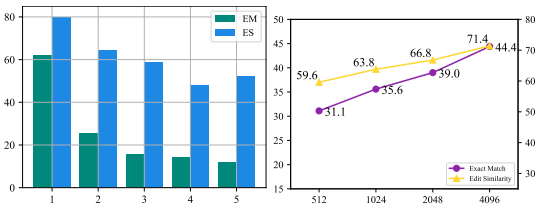


Figure 6: Left figure: Results of different lines. Right figure: Results of various input lengths.

execution analysis, generating unit test cases and providing execution sandboxes for repository-level code completion are very challenging. In Table 7, we observe that tuning leads to better performance in both syntax and execution analysis. Notably, we observe that syntax accuracy improves a lot after tuning, which means that code LLMs can easily learn the basic syntax rules for programming languages.

Analysis on language-specific insights. We have classified 18 programming languages in M^{2}RC-EVAL into 5 programming paradigms and 9 application scenarios as shown in Table 11 and Table 12: Based on the above programming classification structure, we also report the EM results based on StarCoder-7B as shown in Table 8 and Table 9, and have the following observations: (1). For different programming paradigms, we observe that the markup language paradigm has the lowest performance, and the functional paradigm has the best performance. Besides, after tuning, the performance of the markup language paradigm improves greatly. We assume that the syntax rules for markup language are easy, and these code LLMs can quickly obtain these rules after tuning. (2). For different application scenarios, the performance varies significantly. Specifically, the Web Frontend and Scientific Computing have relatively low performance, which needs to be improved for existing code LLMs. (3). We observe these LLMs share some common strengths and weaknesses and we

will continue to investigate more language-specific insights to better improve the code completion abilities of existing code LLMs.

5 Conclusion

In this paper, we propose the first massively multilingual repository-level code completion benchmark (M^{2}RC-EVAL) with 18 popular programming languages, where two types of fine-grained annotations (bucket-level and semantic-level) are provided to comprehensively analyze the effectiveness of different code LLMs. Besides, we also curate a high-quality instruction corpus M^{2}RC-INSTRUCT to enhance the performance of existing models on repository-level code completion. Finally, we hope M^{2}RC-EVAL could facilitate the growth of code intelligence and software engineering.

6 Limitations

First, there are several hyperparameters (e.g., training sizes, input length) to tune, which is laborious and expensive. Second, the current work only focuses on the repository-level code completion task, where other repository-level code intelligence tasks are not considered. Third, while M^{2}RC-EVAL aims to simulate realistic debugging scenarios, the tasks and data may not fully capture the complexity and diversity of real-world software development. We will continue include challenging and complex real-world repos into the benchmark.

7 Acknowledgement

This work was supported by the Jiangsu Science and Technology Major Project (BG2024031) and Nanjing University AI & AI for Science Funding (2024300540).

References

Lakshya A Agrawal, Aditya Kanade, Navin Goyal, Shuvendu K Lahiri, and Sriram K Rajamani. 2023. Guiding language models of code with global context using monitors.

Loubna Ben Allal, Raymond Li, Denis Kocetkov, Chenghao Mou, Christopher Akiki, Carlos Munoz Ferrandis, Niklas Muennighoff, Mayank Mishra, Alex Gu, Manan Dey, et al. 2023. Santacoder: don't reach for the stars! *arXiv preprint arXiv:2301.03988*.

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

Ramakrishna Bairi, Atharv Sonwane, Aditya Kanade, Arun Iyer, Suresh Parthasarathy, Sriram Rajamani, B Ashok, Shashank Shet, et al. 2023. Codeplan: Repository-level coding using llms and planning. *arXiv preprint arXiv:2309.12499*.

Mohammad Bavarian, Heewoo Jun, Nikolas Tezak, John Schulman, Christine McLeavey, Jerry Tworek, and Mark Chen. 2022a. Efficient training of language models to fill in the middle. *arXiv preprint arXiv:2207.14255*.

Mohammad Bavarian, Heewoo Jun, Nikolas Tezak, John Schulman, Christine McLeavey, Jerry Tworek, and Mark Chen. 2022b. Efficient training of language models to fill in the middle. *arXiv preprint arXiv:2207.14255*.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin. Association for Computational Linguistics.

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. 2022. Multipl-e: A scalable and extensible approach to benchmarking neural code generation. *arXiv preprint arXiv:2208.08227*.

Linzheng Chai, Shukai Liu, Jian Yang, Yuwei Yin, Ke Jin, Jiaheng Liu, Tao Sun, Ge Zhang, Changyu Ren, Hongcheng Guo, et al. 2024. Mceval: Massively multilingual code evaluation. *arXiv preprint arXiv:2406.07436*.

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*, abs/2107.03374.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. 24(240):1–113.

Yangruibo Ding, Zijian Wang, Wasi Ahmad, Hantian Ding, Ming Tan, Nihal Jain, Murali Krishna Ramanathan, Ramesh Nallapati, Parminder Bhatia, Dan Roth, et al. 2024. Crosscodeeval: A diverse and multilingual benchmark for cross-file code completion. *Advances in Neural Information Processing Systems*, 36.

Yangruibo Ding, Zijian Wang, Wasi Uddin Ahmad, Hantian Ding, Ming Tan, Nihal Jain, Murali Krishna Ramanathan, Ramesh Nallapati, Parminder Bhatia, Dan Roth, and Bing Xiang. 2023. Crosscodeeval: A diverse and multilingual benchmark for cross-file code completion. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yangruibo Ding, Zijian Wang, Wasi Uddin Ahmad, Murali Krishna Ramanathan, Ramesh Nallapati, Parminder Bhatia, Dan Roth, and Bing Xiang. 2022. Cocomic: Code completion by jointly modeling in-file and cross-file context. *arXiv preprint arXiv:2212.10007*.

Daniel Fried, Armen Aghajanyan, Jessy Lin, Sida Wang, Eric Wallace, Freda Shi, Ruiqi Zhong, Scott Yih, Luke Zettlemoyer, and Mike Lewis. 2023. Incoder: A generative model for code infilling and synthesis. In *The Eleventh International Conference on Learning Representations*.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y Wu, YK Li, et al. 2024a. Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y Wu, YK Li, et al. 2024b. Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*.

Yancheng He, Shilong Li, Jiaheng Liu, Weixun Wang, Xingyuan Bu, Ge Zhang, Zhongyuan Peng, Zhaoxiang Zhang, Zhicheng Zheng, Wenbo Su, and Bo Zheng. 2025. Can large language models detect errors in long chain-of-thought reasoning? *Preprint*, arXiv:2502.19361.

Siming Huang, Tianhao Cheng, Jason Klein Liu, Jiaran Hao, Liuyihan Song, Yang Xu, J. Yang, J. H. Liu, Chenchen Zhang, Linzheng Chai, Ruifeng Yuan, Zhaoxiang Zhang, Jie Fu, Qian Liu, Ge Zhang, Zili Wang, Yuan Qi, Yinghui Xu, and Wei Chu. 2024. Opencoder: The open cookbook for top-tier code large language models.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Kai Dang, et al. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

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

Nihal Jain, Robert Kwiatkowski, Baishakhi Ray, Murali Krishna Ramanathan, and Varun Kumar. 2024b. On mitigating code llm hallucinations with api documentation. *ArXiv*, abs/2407.09726.

Siyuan Jiang, Jia Li, He Zong, Huanyu Liu, Hao Zhu, Shukai Hu, Erlu Li, Jiazheng Ding, Yu Han, Wei Ning, Gen Wang, Yihong Dong, Kechi Zhang, and Ge Li. 2024. aixcoder-7b: A lightweight and effective large language model for code completion.

Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Carlos Muñoz Ferrandis, Yacine Jernite, Margaret Mitchell, Sean Hughes, Thomas Wolf, et al. 2022. The stack: 3 tb of permissively licensed source code. *arXiv preprint arXiv:2211.15533*.

Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven C. H. Hoi. 2022. Coderl: Mastering code generation through pretrained models and deep reinforcement learning. *ArXiv*, abs/2207.01780.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, et al. 2023. Starcoder: may the source be with you! *arXiv preprint arXiv:2305.06161*.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. 2022. Competition-level code generation with alphacode. *ArXiv preprint*, abs/2203.07814.

Dianshu Liao, Shidong Pan, Qing Huang, Xiaoxue Ren, Zhenchang Xing, Huan Jin, and Qinying Li. 2023. Context-aware code generation framework for code repositories: Local, global, and third-party library awareness.

Jiaheng Liu, Chenchen Zhang, Jinyang Guo, Yuanxing Zhang, Haoran Que, Ken Deng, Jie Liu, Ge Zhang, Yanan Wu, Congnan Liu, et al. 2024. Ddk: Distilling domain knowledge for efficient large language models. *Advances in Neural Information Processing Systems*, 37:98297–98319.

Jiaheng Liu, Dawei Zhu, Zhiqi Bai, Yancheng He, Huanxuan Liao, Haoran Que, Zekun Wang, Chenchen Zhang, Ge Zhang, Jiebin Zhang, et al. 2025. A comprehensive survey on long context language modeling. *arXiv preprint arXiv:2503.17407*.

Tianyang Liu, Canwen Xu, and Julian McAuley. 2023a. Repobench: Benchmarking repository-level code auto-completion systems. *arXiv preprint arXiv:2306.03091*.

Tianyang Liu, Canwen Xu, and Julian J. McAuley. 2023b. Repobench: Benchmarking repository-level code auto-completion systems. abs/2306.03091.

Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, et al. 2024. Starcoder 2 and the stack v2: The next generation.

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

Hengzhi Pei, Jinman Zhao, Leonard Lausen, Sheng Zha, and George Karypis. 2023. Better context makes better code language models: A case study on function call argument completion. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Huy Nhat Phan, Hoang N. Phan, Tien N. Nguyen, and Nghi D. Q. Bui. 2024. Repohyper: Search-expand-refine on semantic graphs for repository-level code completion.

Shuo Ren, Daya Guo, Shuai Lu, Long Zhou, Shujie Liu, Duyu Tang, Neel Sundaresan, Ming Zhou, Ambrosio Blanco, and Shuai Ma. 2020. Codebleu: a method for automatic evaluation of code synthesis. *arXiv preprint arXiv:2009.10297*.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code.

Disha Shrivastava, Denis Kocetkov, Harm de Vries, Dzmitry Bahdanau, and Torsten Scholak. 2023a. Repofusion: Training code models to understand your repository. *arXiv preprint arXiv:2306.10998*.

Disha Shrivastava, Hugo Larochelle, and Daniel Tarlow. 2023b. Repository-level prompt generation for large language models of code. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31693–31715. PMLR.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. 568:127063.

Wannita Takerngsaksiri, Chakkrit Tantithamthavorn, and Yuan-Fang Li. 2024. Syntax-aware on-the-fly code completion. *Inf. Softw. Technol.*, 165(C).

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.

Pei Wang, Yanan Wu, Noah Wang, Jiaheng Liu, Xiaoshuai Song, Z.Y. Peng, Ken Deng, Chenchen Zhang, JiakaiWang, Junran Peng, Ge Zhang, Hangyu Guo, Zhaoxiang Zhang, Wenbo Su, and Bo Zheng. 2024. Mtu-bench: A multi-granularity tool-use benchmark for large language models.

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

Yue Wang, Weishi Wang, Shafiq Joty, and Steven C.H. Hoi. 2021. CodeT5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8696–8708, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Frank F Xu, Uri Alon, Graham Neubig, and Vincent Josua Hellendoorn. 2022. A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, pages 1–10.

Hao Yu, Bo Shen, Dezhi Ran, Jiaxin Zhang, Qi Zhang, Yuchi Ma, Guangtai Liang, Ying Li, Qianxiang Wang, and Tao Xie. 2024. Codereval: A benchmark of pragmatic code generation with generative pre-trained models. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*, pages 1–12.

Alexander Zhang, Marcus Dong, Jiaheng Liu, Wei Zhang, Yejie Wang, Jian Yang, Ge Zhang, Tianyu Liu, Zhongyuan Peng, Yingshui Tan, Yuanxing Zhang, Zhexu Wang, Weixun Wang, Yancheng He, Ken Deng, Wangchunshu Zhou, Wenhao Huang, and Zhaoxiang Zhang. 2025. Codecriticbench: A holistic code critique benchmark for large language models. *Preprint*, arXiv:2502.16614.

Fengji Zhang, Bei Chen, Yue Zhang, 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*.

Ge Zhang, Scott Qu, Jiaheng Liu, Chenchen Zhang, Chenghua Lin, Chou Leuang Yu, Danny Pan, Esther Cheng, Jie Liu, Qunshu Lin, Raven Yuan, Tuney Zheng, Wei Pang, Xinrun Du, Yiming Liang, Yinghao Ma, Yizhi Li, Ziyang Ma, Bill Lin, Emmanouil Benetos, Huan Yang, Junting Zhou, Kaijing Ma, Minghao Liu, Morry Niu, Noah Wang, Quehry Que, Ruibo Liu, Sine Liu, Shawn Guo, Soren Gao, Wangchunshu Zhou, Xinyue Zhang, Yizhi Zhou, Yubo Wang, Yuelin Bai, Yuhan Zhang, Yuxiang Zhang, Zenith Wang, Zhenzhu Yang, Zijian Zhao, Jiajun Zhang, Wanli Ouyang, Wenhao Huang, and Wenhu Chen. 2024. Map-neo: Highly capable and transparent bilingual large language model series. *arXiv preprint arXiv: 2405.19327*.

CodeGemma Team Heri Zhao, Jeffrey Hui, Joshua Howland, Nam Nguyen, Siqi Zuo, Andrea Hu, Christopher A. Choquette-Choo, Jingyue Shen, Joe Kelley, Kshi tij Bansal, Luke Vilnis, Mateo Wirth, Paul Michel, Peter Choy, Pratik Joshi, Ravin Kumar, Sarmad Hashmi, Shubham Agrawal, Zhitao Gong, Jane

Fine, Tris Brian Warkentin, Ale Jakse Hartman, Bin Ni, Kathy Korevec, Kelly Schaefer, and Scott Huffman. 2024. Codegemma: Open code models based on gemma. *ArXiv*, abs/2406.11409.

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. 2023. Codegeex: A pre-trained model for code generation with multilingual evaluations on humaneval-x. *arXiv preprint arXiv:2303.17568*, abs/2303.17568.

Terry Yue Zhuo, Vu Minh Chien, Jenny Chim, Han Hu, Wenhao Yu, Ratnadira Widyasari, Imam Nur Bani Yusuf, Haolan Zhan, Junda He, Indraneil Paul, Simon Brunner, Chen GONG, James Hoang, Armel Randy Zebaze, Xiaoheng Hong, Wen-Ding Li, Jean Kaddour, Ming Xu, Zhihan Zhang, Prateek Yadav, Naman Jain, Alex Gu, Zhoujun Cheng, Jiawei Liu, Qian Liu, Zijian Wang, David Lo, Binyuan Hui, Niklas Muennighoff, Daniel Fried, Xiaoning Du, Harm de Vries, and Leandro Von Werra. 2025. Bigcodebench: Benchmarking code generation with diverse function calls and complex instructions. In *The Thirteenth International Conference on Learning Representations*.

A Appendix

A.1 Broader Impacts & Potential Risks

In this paper, we propose a repository-level code completion benchmark with 18 programming languages. Therefore, we hope our work can enhance the improvements on the multilingual repository-level code completion task. For potential risks, we have not seen any risks in our M^{2}RC-EVAL.

A.2 Details of the Baseline Models

StarCoder (Li et al., 2023) is a series of generative language models (e.g., 7B, 15.5B). These decoder-only models are trained on the Stack dataset (Kocetkov et al., 2022) and can support 8K tokens in context.
DeepSeekCoder (Guo et al., 2024b) is a collection of code-oriented models with capacities from 1.3B to 33B parameters. Trained on a manually curated 2-trillion-token corpus, these models leverage Fill-in-the-Middle (FIM) (Bavarian et al., 2022b) and Rotary Position Embedding (RoPE) (Su et al., 2024) techniques, which enables efficient code generation and infilling within a 16K token window.
Code Llama (Roziere et al., 2023) is a family of code large language models based on Llama 2 (Touvron et al., 2023) with 7B, 13B, 34B, and 70B parameters. While trained on 16K token sequences, these models can handle inputs up to 100K tokens during inference.

Note that we just use the base model versions of these three models.

A.3 Analysis on the Quality Control

In §(3.2), we discard test samples that could be exactly predicted by DeepSeekCoder-1.3B without cross-file contexts. Meanwhile, to discuss more clearly, we also use the DeepSeekCoder-6.7B, StarCoder-7B, and DeepSeekCoder-33B to analyze the ratios of evaluation cases with or without using repository-level contexts. Specifically, we prompt DeepSeekCoder-6.7B, StarCoder-7B, and DeepSeekCoder-33B using the in-file contexts of each sample and obtain three predictions. If one prediction is exactly matched ground-truth, this sample is considered to be predicted without requiring repository-level contexts. Finally, we observe that 71% samples cannot be well predicted only using in-file contexts, which indicates that it is necessary to use the cross-file contexts to achieve better performance in our M^{2}RC-EVAL.

A.4 Analysis on Inference Cases in Different Languages

We manually inspect the behavior of StarCoder-7B (Li et al., 2023) on completion cases in different programming languages. As shown in Fig. 13, the model successively predicts the attribute `position.y`, which is an easy pattern that could be inferred from `position.x` in the prefix. Besides, the `(x, y)` pattern that occurs multiple times in the cross-file context. On the contrary, the model seems to struggle with complex expressions and statements. In Fig. 14, the model should complete the function with a combined condition and return statement. However, the retriever could not provide useful references and the model only predicts half of the condition correctly. Fig. 15 illustrates a Python script to execute memory calculations. Although some calculations appear in the cross-file context, there are no precisely matched calculation procedures. The recurrent conditions in the ground truth require calculations on the data shape, but the model clumsily guesses the data shape. Moreover, we observe that the model prediction is usually affected by frequent identifiers in the retrieved contents. In Fig. 16, the model repeats the `gc.Client` and results in a hallucination for object `PullRequests`, where the ground truth is `gc`. Similarly, in Fig. 17, the model blindly catches the "err" with error level. Yet the correct log level is "warn", which could be judged from `c.on('error', console.warn)` in the Cross file Context 1. Further, in Fig. 18, the ground truth and the model prediction differ by only two characters "()" from a textual perspective, but the ground truth passes the method reference while the model prediction passes the method return.

A.5 Analysis on the Number of Bucket Granularity

In Table 10, we have analyzed the minimum, maximum, and average depths of these 18 programming languages, and observed that depths of AST for different languages are different greatly, where the minimum depth is from 1 to 7, and the maximum depth is from 23 to 51. Therefore, we cannot use

the node's layer number in the AST as the fine-grained annotation well. Besides, we observe that the average depth of different languages is from 9.7 to 20.2, and the overall average depth is 14.6. In our implementation, to explain clearly, we just chose 10 as the bucket number. Note that we will provide the tree depth number for each completion sample, and users can select the corresponding bucket number freely.

Table 10: Depth statistics of different languages.

Language	Minimum	Maximum	Average
Objective-C	3	32	11.3
C++	3	32	11.3
C	4	41	13.8
Haskell	1	28	10.9
Ruby	2	45	13.7
Kotlin	1	34	14.9
Rust	5	51	20.2
C#	3	41	16.4
PHP	7	51	17.7
Go	1	40	13.6
Java	1	46	14.3
HTML	2	46	14.0
R	1	23	9.7
JavaScript	1	31	13.4
TypeScript	3	51	18.3
Scala	3	51	17.5
Lua	2	51	16.1
Python	1	38	15.7

Table 11: Classification of M^{2}RC-EVAL based on paradigm types.

Paradigm Types	Languages
Procedural	C
Object Oriented	C#, Java, Kotlin, Objective-C
Multiple Paradigms	C++, Go, JavaScript, Lua, PHP, Python, R, Ruby, Rust, Scala, TypeScript
Functional	Haskell
Markup Language	HTML

Table 12: Classification of M^{2}RC-EVAL based on application scenarios.

Application Scenarios	Languages
Mobile	Kotlin, Objective-C
Cross Platform	Java
Desktop Application	C#
Web Frontend	JavaScript, TypeScript, HTML
Web Backend	Go, PHP, Ruby, Rust, Scala
Scientific Computing	Python, R
System & Software	C, C++
Education & Research	Haskell
Automation Scripts	Lua

A.6 More Experiments

- We provide the analysis on the bucket levels in Fig. 7 and Fig. 8, respectively.
- We analyze the effect of different semantic levels on Rust, Objective-C, and Haskell in Fig. 9. respectively.
- We provide the semantic-level annotations on 18 languages in Fig. 10, Fig. 11 and Fig. 12.
- We provide the results of problems from different difficulty levels in Fig. 19, where we define completion on 1 line, completion on 2-3 lines and completion on 4-5 lines as easy, middle, and hard settings, respectively.

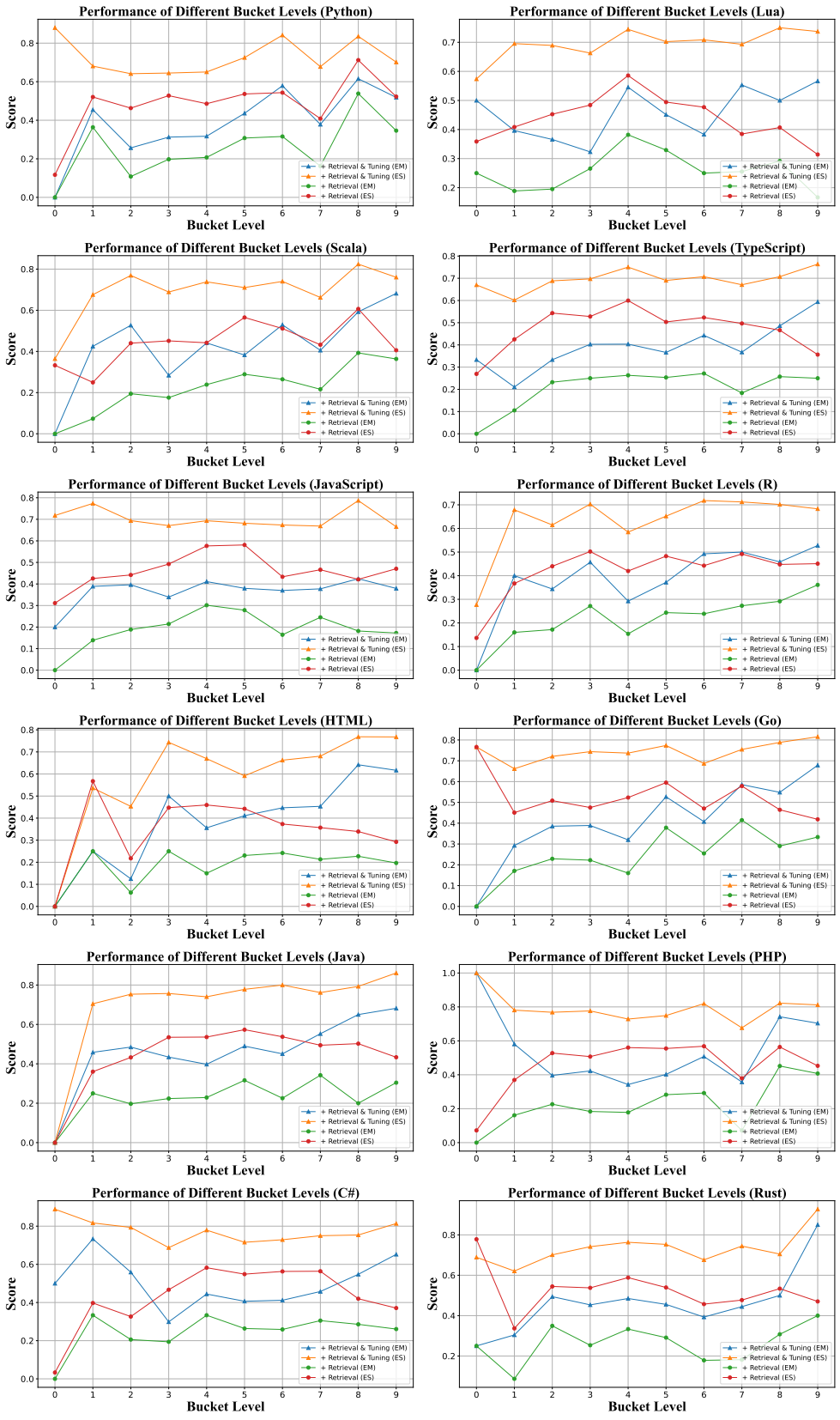


Figure 7: Effectiveness of different bucket levels based on StarCoder-7B for different languages.

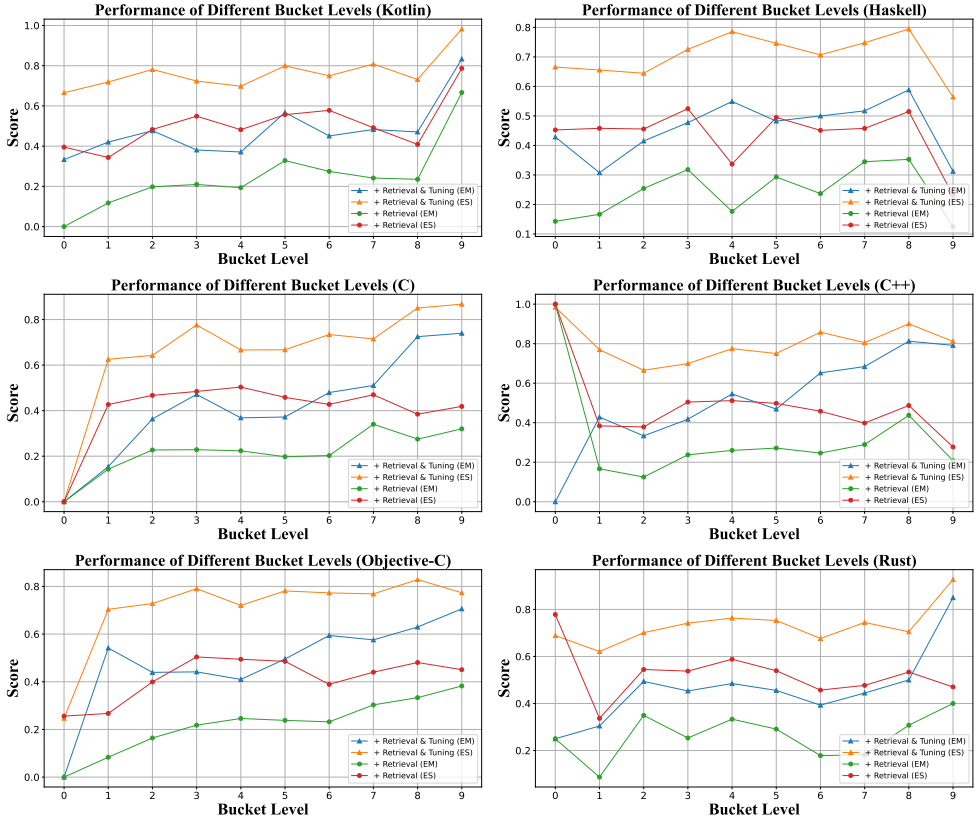


Figure 8: Effectiveness of different bucket levels based on StarCoder-7B for different languages.

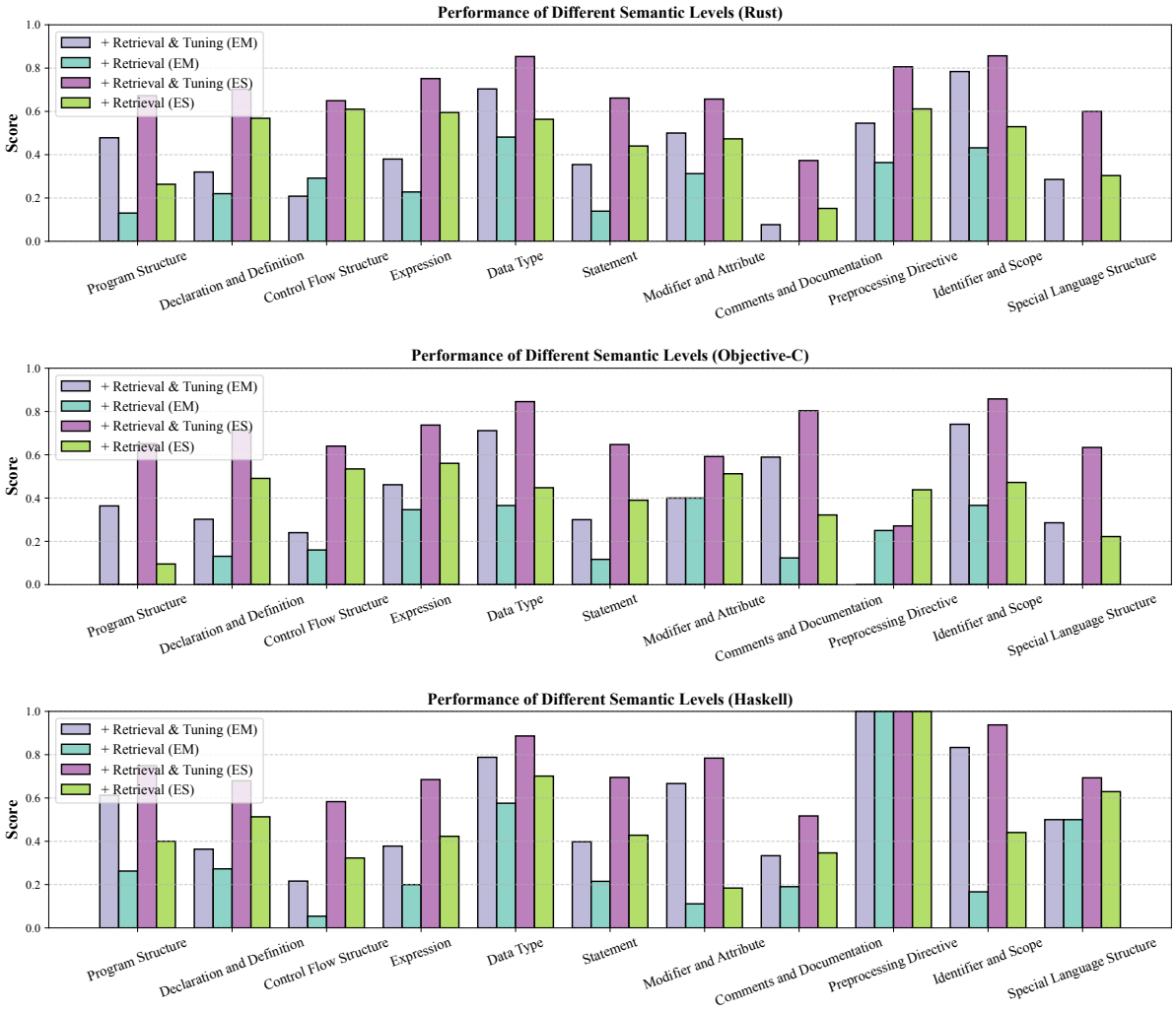
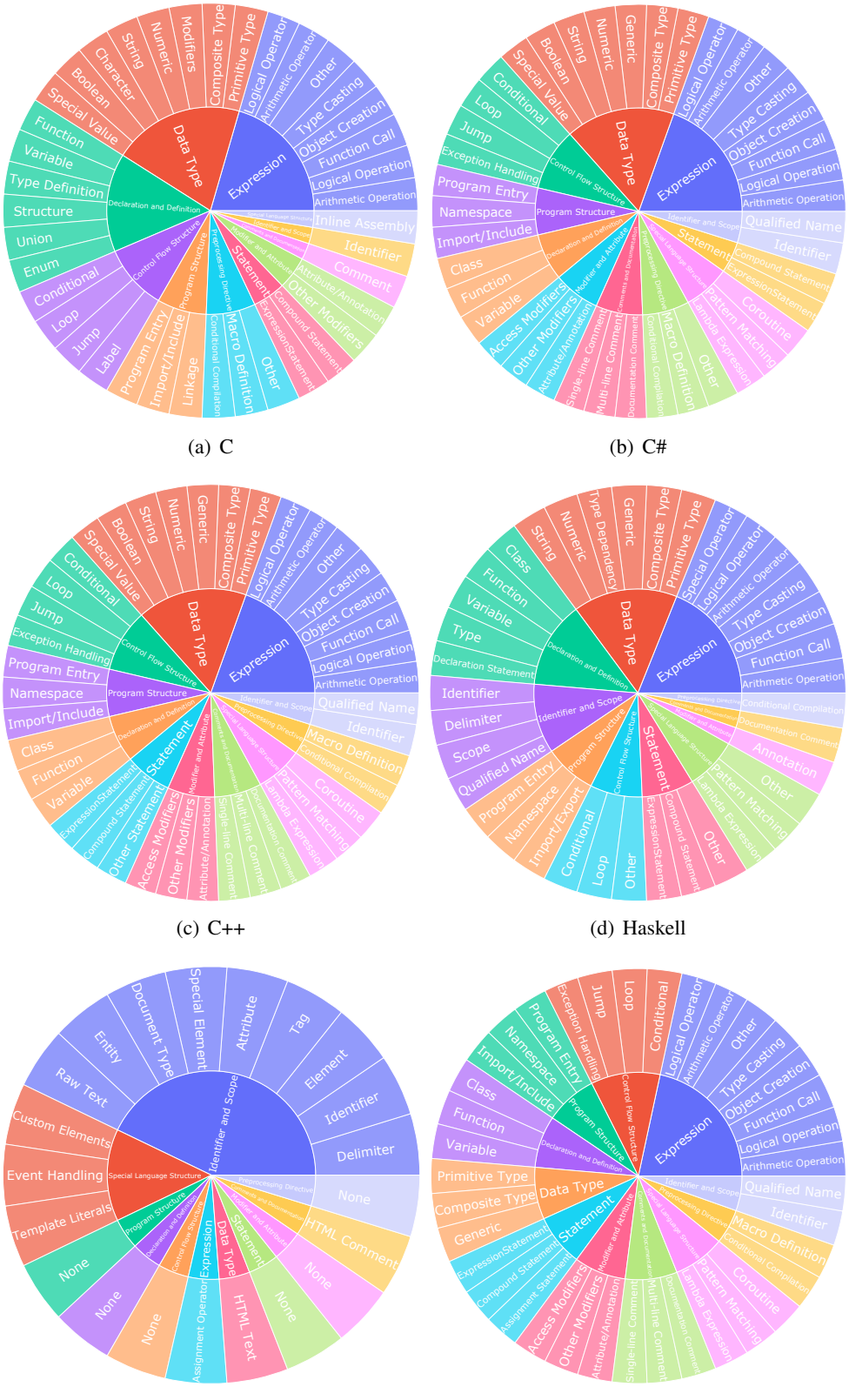


Figure 9: Effectiveness of different semantic levels based on StarCoder-7B.

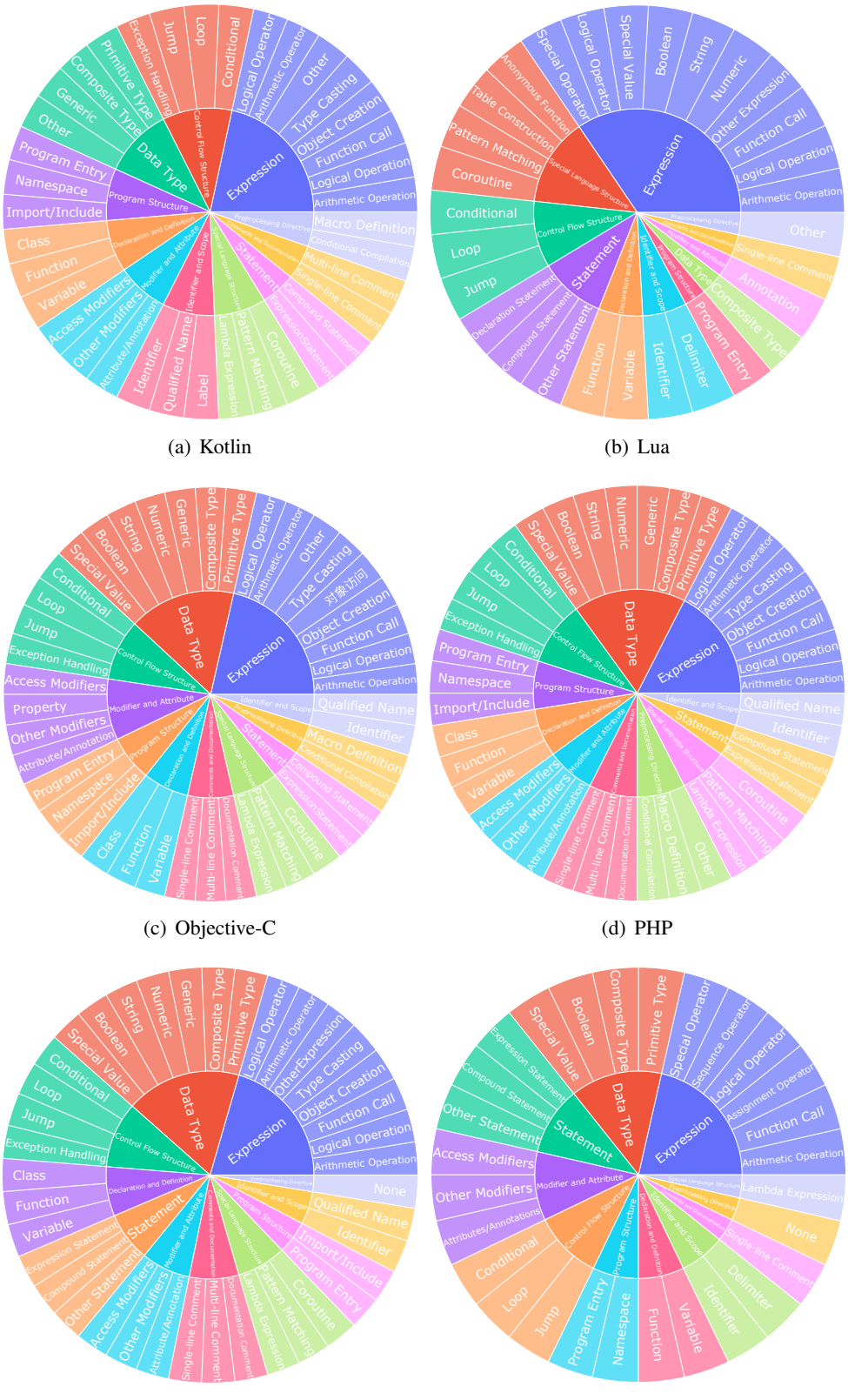


(a) C (b) C#

(c) C++ (d) Haskell

(e) HTML (f) JavaScript

Figure 10: Semantic-level annotations on different types of programming languages. “none” is used if this language does not have corresponding subcategories.

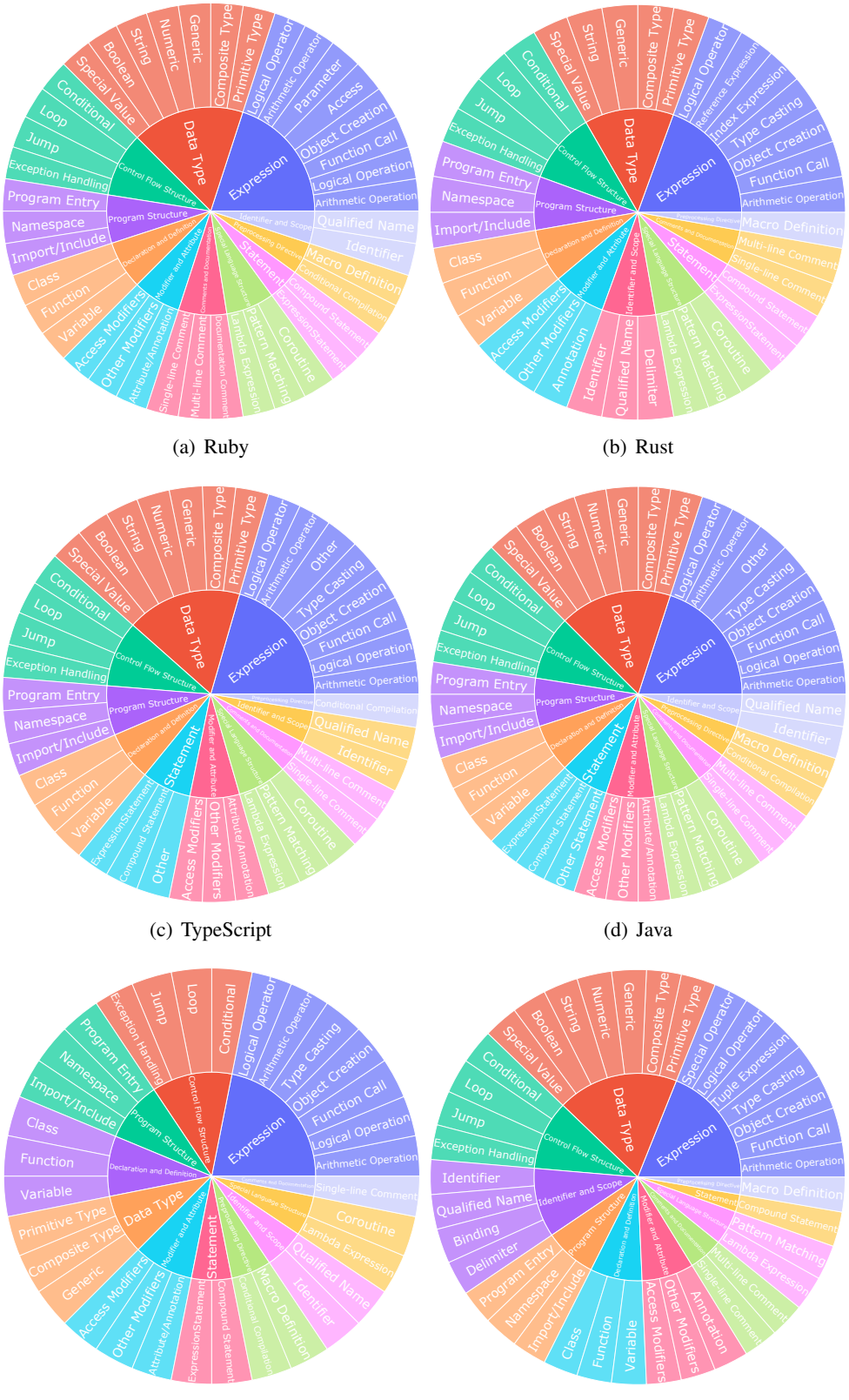


(a) Kotlin (b) Lua

(c) Objective-C (d) PHP

(e) Python (f) R

Figure 11: Semantic-level annotations on different types of programming languages. “none” is used if this language does not have corresponding subcategories.



(a) Ruby (b) Rust

(c) TypeScript (d) Java

(e) Go (f) Scala

Figure 12: Semantic-level annotations on different types of programming languages. “none” is used if this language does not have corresponding subcategories.

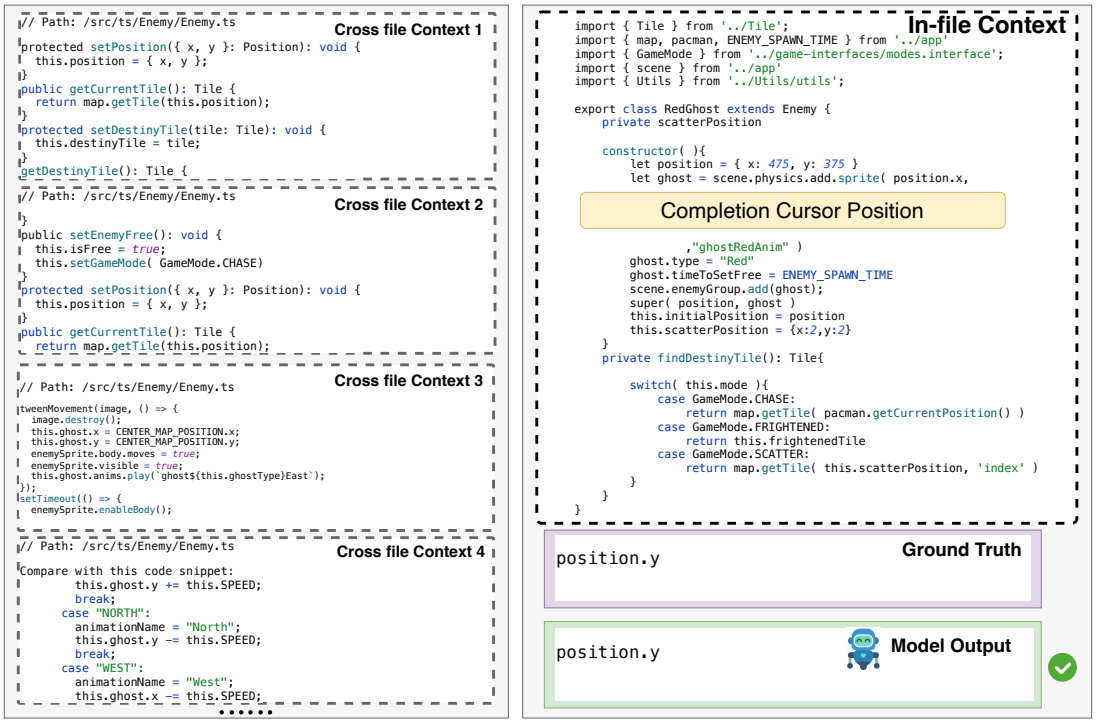


Figure 13: Visualization on success case for TypeScript. (Semantic label: *Modifier and Attribute*)

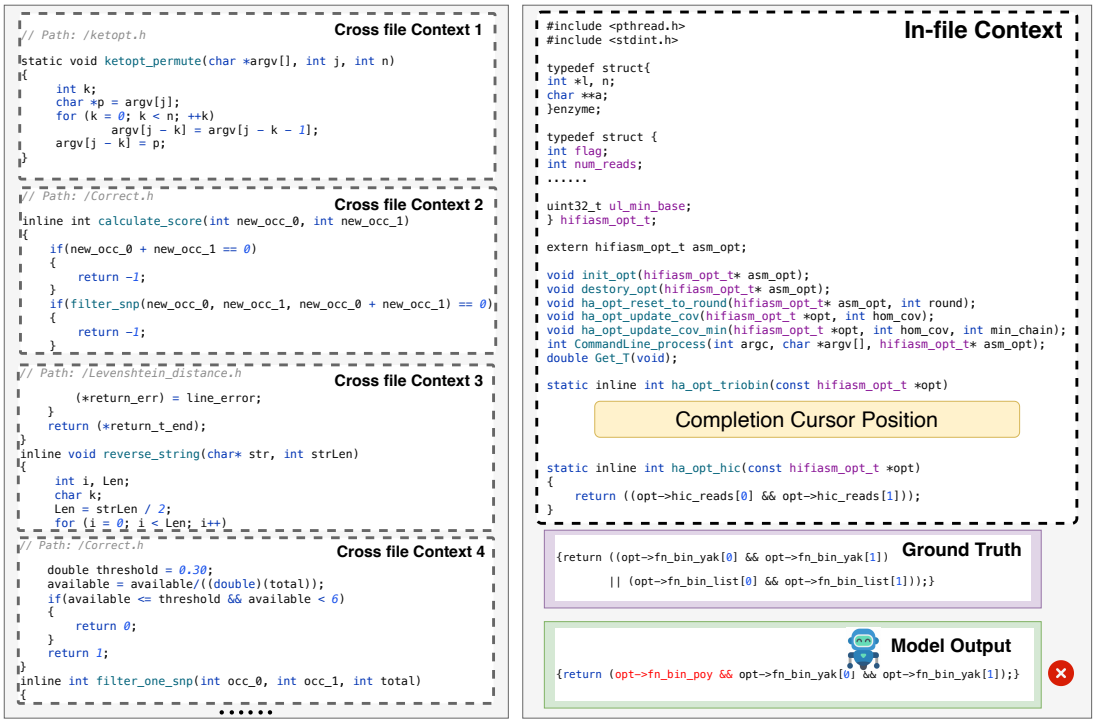


Figure 14: Visualization on failure case for the C language. (Semantic label: *Statement*)

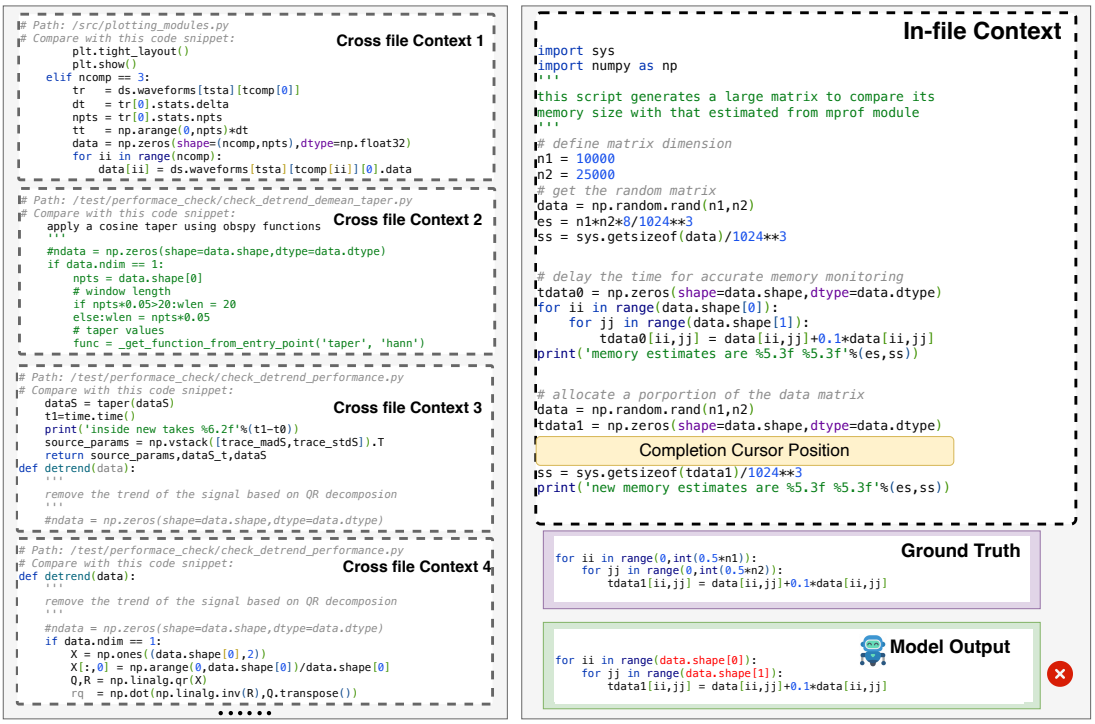


Figure 15: Visualization on failure case for Python. (Semantic label: *Expression*)

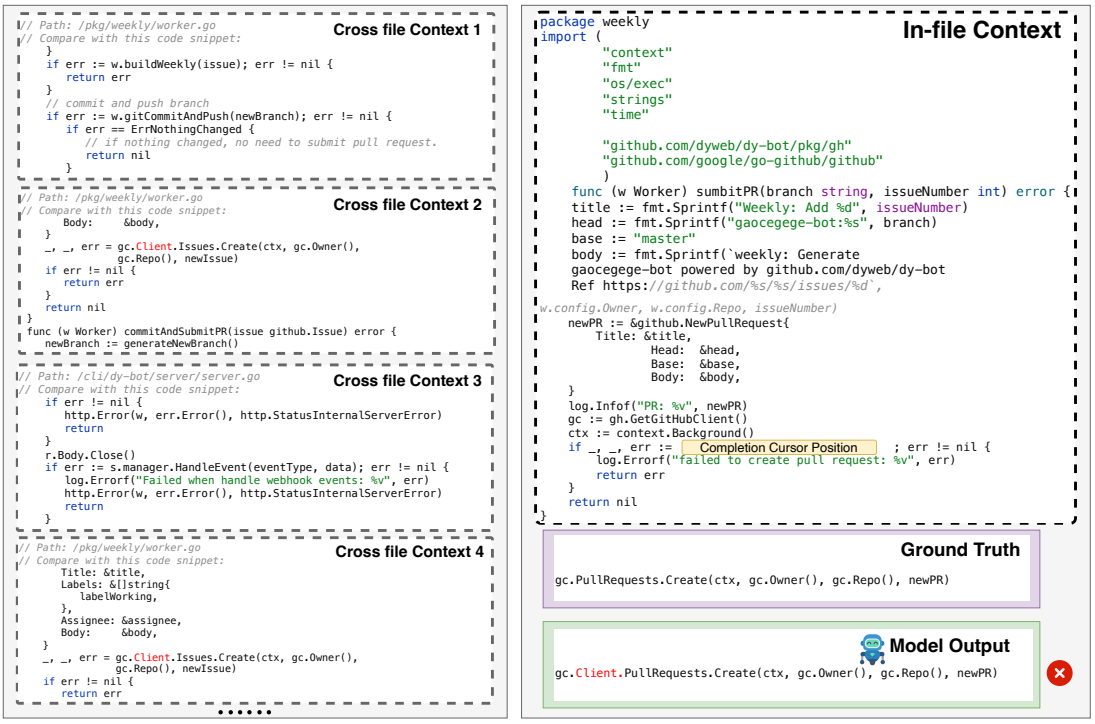


Figure 16: Visualization on failure case for Go. (Semantic label: *Expression*)

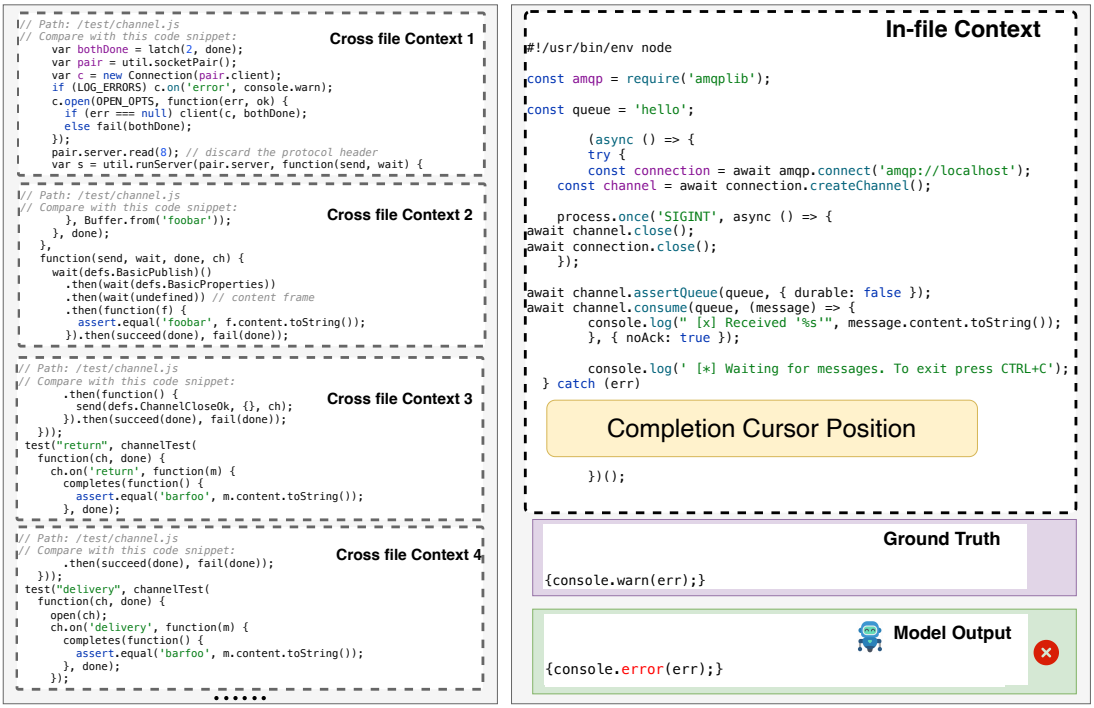


Figure 17: Visualization on failure case for Javascript. (Semantic label: *Statement*)

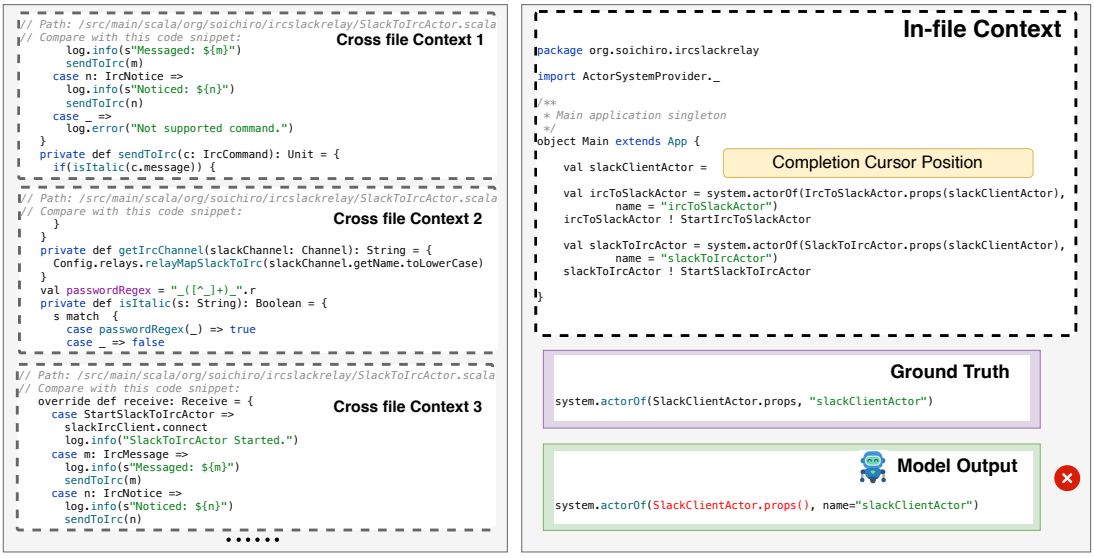


Figure 18: Visualization on failure case for Scala. (Semantic label: *Expression*)

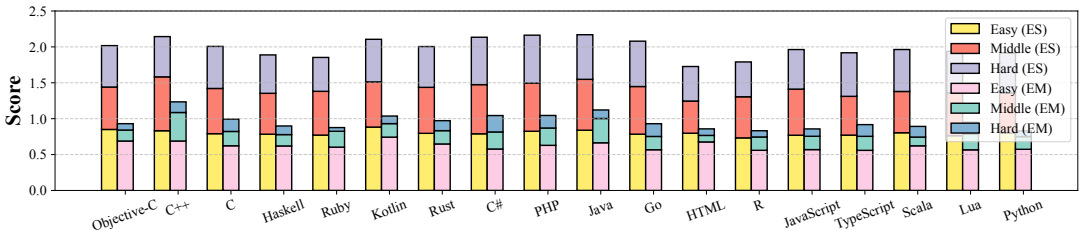


Figure 19: Performance on M^{2}RC-EVAL for problems of different difficulty levels.