McEval: Massively Multilingual Code Evaluation

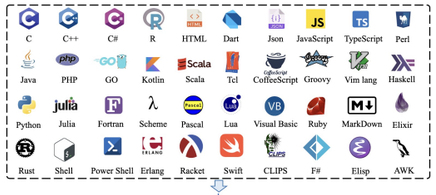
Linzheng Chai[1] *, Shukai Liu[1] *, Jian Yang[1] *†, Yuwei Yin[2], Ke Jin[1], Jiaheng Liu[1], Tao Sun[1], Ge Zhang[3], Changyu Ren[1], Hongcheng Guo[1], Zekun Wang[1], Boyang Wang[1], Xianjie Wu[1], Bing Wang[1], Tongliang Li[4], Liqun Yang[1], Sufeng Duan[5], Zhoujun Li[1]

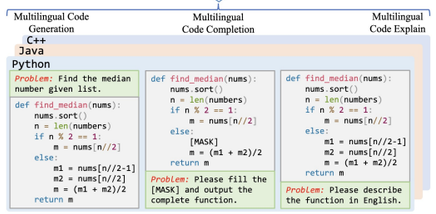
[1]CCSE, Beihang University, [2]University of British Columbia, [3]University of Waterloo
[4]Beijing Information Science and Technology University, [5]Shanghai Jiao Tong University

ICLR

Introduction

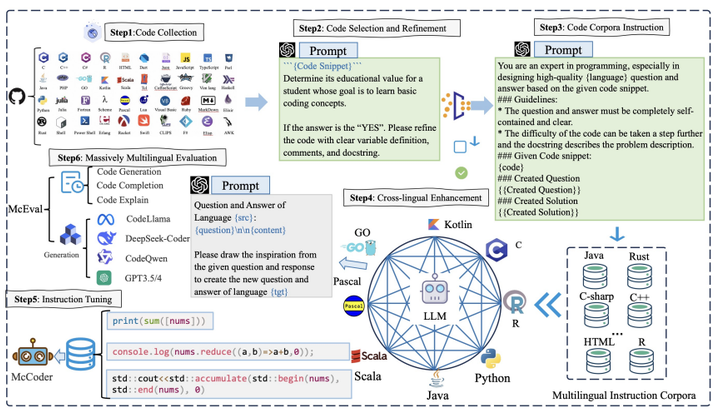
➢ **McEval:** A massively **multilingual code benchmark** covering **40** programming languages.





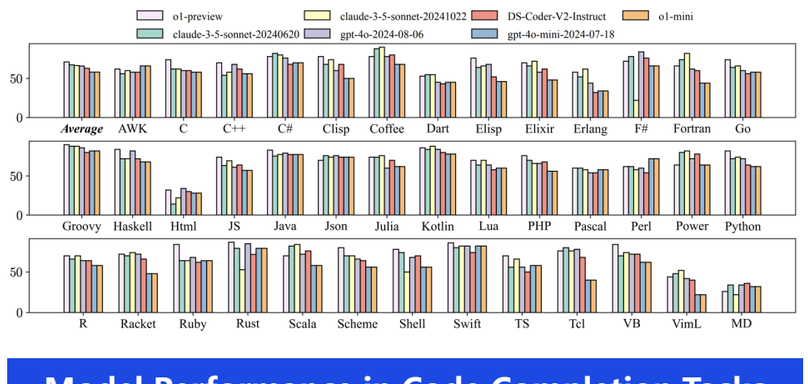
Statistics	Value
Questions	
Code Generation	2,007
Code Explanation	2,007
Code Completion	12,017
- *Single-Line*	2,998
- *Multi-Line*	2,998
- *Span*	4,014
- *Span(light)*	2,007
Total Test Cases	10,086
Difficulty Level	
- Easy	1,221
- Medium	401
- Hard	385
Length	
Prompt	
- *maximum length*	793 tokens
- *minimum length*	16 tokens
- *avg length*	173.8 tokens
Solution(Output)	
- *maximum length*	666 tokens
- *minimum length*	4 tokens
- *avg length*	120.9 tokens

➢ **McEval-Instruct:** covering 40 languages from code snippets to fine-tune mCoder.



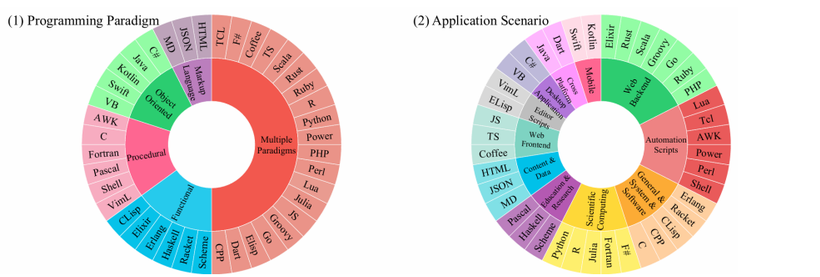
Home Page: https://mceval.github.io/

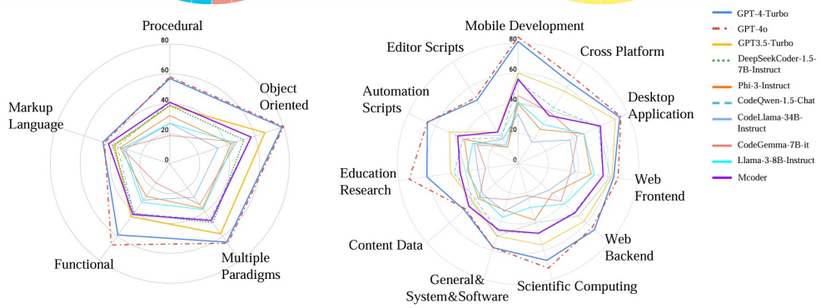
Evaluation



Model Performance in Code Completion Tasks

We categorize the McEval into 5 paradigms and 11 application scenarios and summarize the performance of code LLMs on the code generation task.



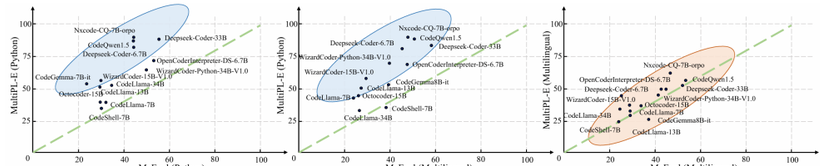


(1) Performance on Programming Paradigm Categories (2) Performance on Application Scenario Categories

Further Analysis

➢ Unbalance in Different Languages

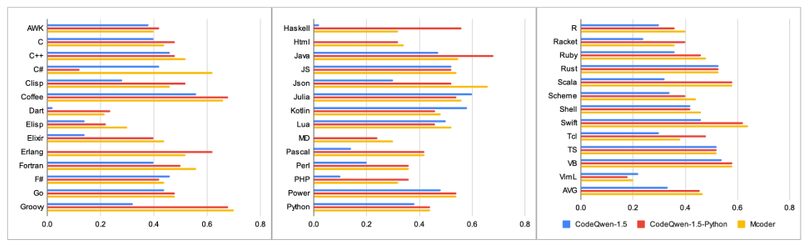
We compare the results of several open-source models on the Multipl-E multilingual benchmark with corresponding languages on McEval.



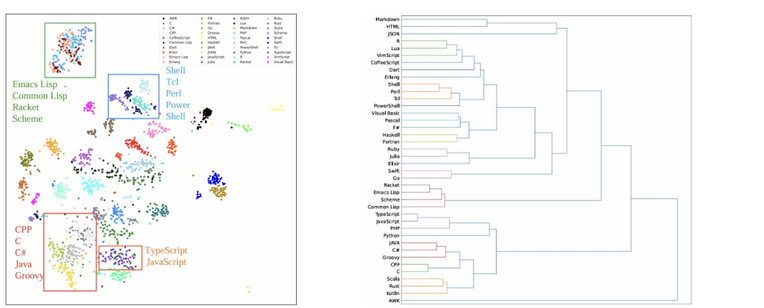
(1) Python-Python (2) Multilingual-Python (3) Multilingual-Multilingual

➢ Cross-lingual Transfer

Fine-tuning with Python-only data can still effectively transfer instruction-following abilities to other languages, resulting in superior multilingual performance.



➢ Analysis of Language Representations



(1) Representation visualization based on t-SNE (2) Representation visualization based on Hierarchical Cluster