

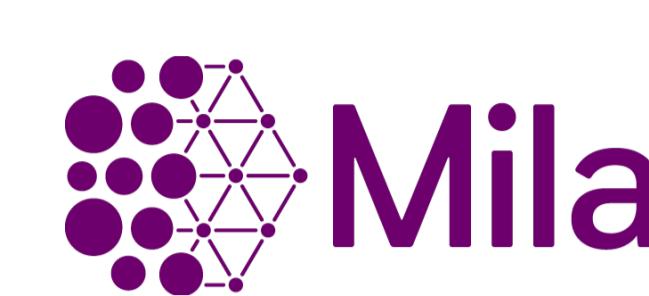
The Vault: A Comprehensive Multilingual Dataset for Advancing Code Understanding and Generation

Dung Nguyen Manh^{1*}, Nam Le Hai^{1,3*}, Anh T. V. Dau^{1,3}, Anh Minh Nguyen¹, Khanh Nghiem¹, Jin Guo^{4,5}, Nghi D. Q Bui²

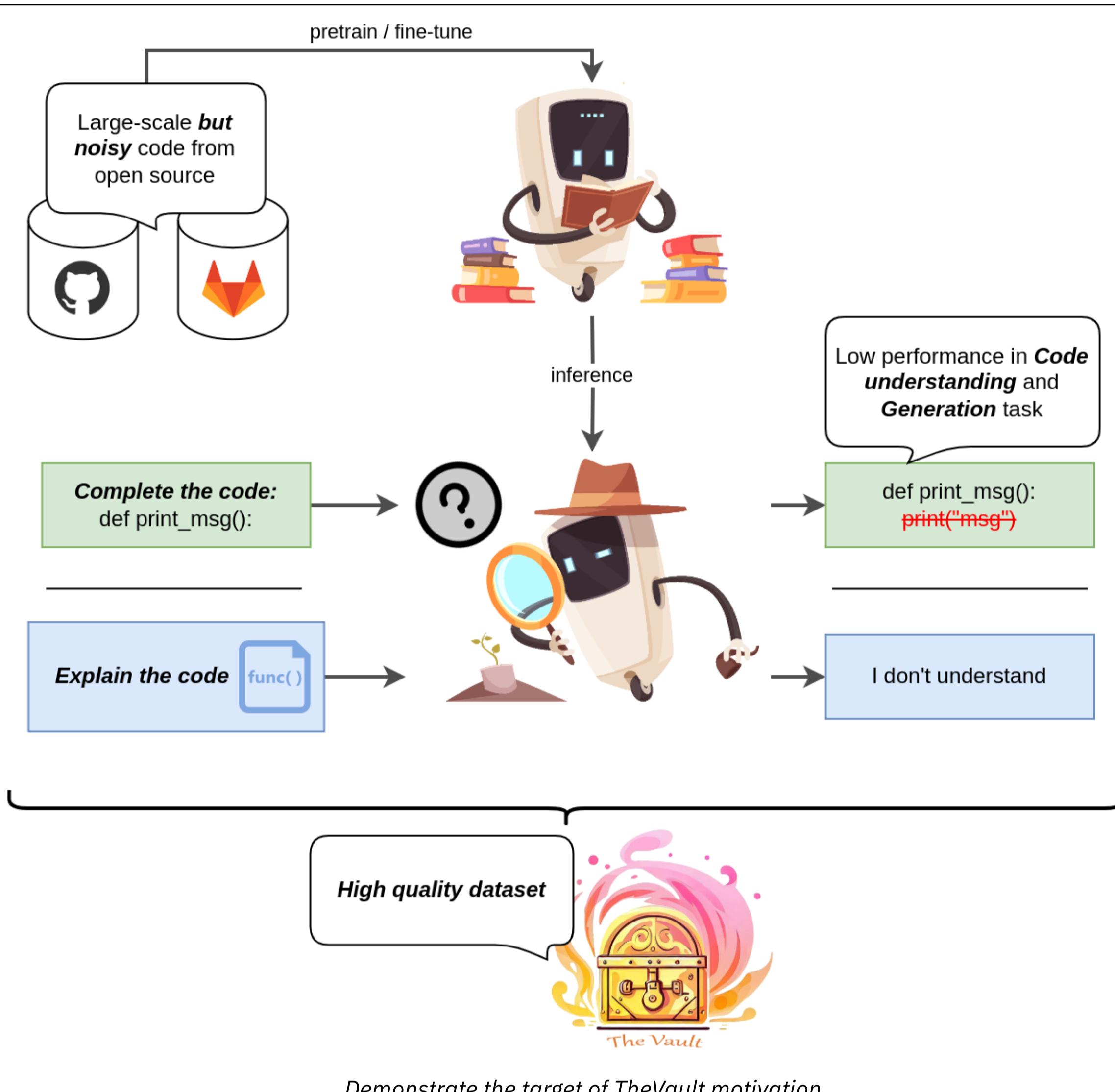
¹FPT Software AI Center {dungnm31, namlh35, anhdtv7, minhna4, khanhvn22}@fpt.com

²Fulbright University, Vietnam, nighi.bui@fulbright.edu.vn; ³Hanoi University of Science and Technology, Vietnam

⁴School of Computer Science, McGill University, Canada; ⁵Mila - Quebec AI Institute



MOTIVATION



MAIN CONTRIBUTIONS

- We release a dataset with **43 million samples of high-quality code-text pairs**, 243 million uni-modal samples, and 69 million pairs of line comments with context from 10 popular programming languages.
- Well-known CodeLLMs (CodeGen, CodeT5, PLBART) **fine-tuned on The Vault** **outperform their original models** significantly on wide range of tasks, including code generation, code search and code summarization.
- We propose a **novel pipeline for obtaining high-quality pairs** of code from large, noisy corpora.
- We publish an open-source toolkit used for **converting raw source code into code-text pairs** and filtering noisy samples in various programming languages.

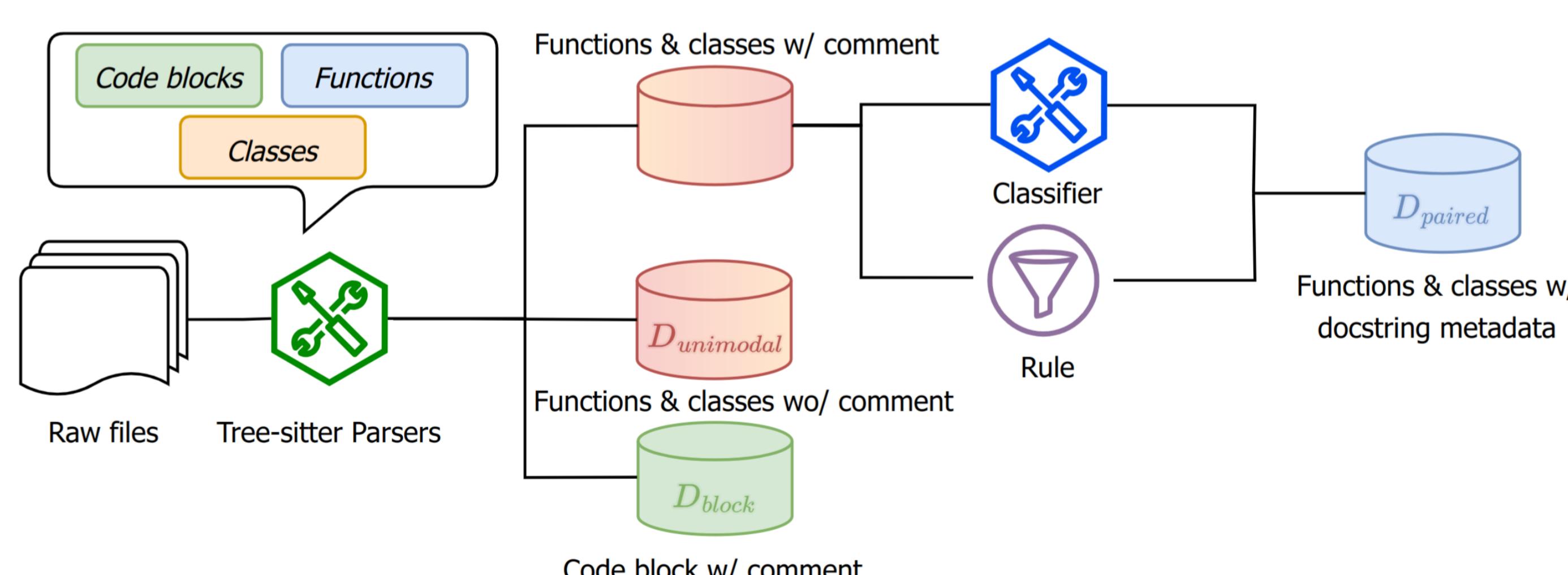
DATASET COMPARISON

Dataset	#PL	#Function	
		w/ docstring	w/o docstring
PyMT5 [Clement et al., 2020]	1	~ 7,700,000	-
CoDesc [Hasan et al., 2021]	1	4,211,516	-
CodeSearchNet [Husain et al., 2019]	6	2,326,976	4,125,470
CodeXGLUE CSN [Lu et al., 2021]	6	1,005,474	-
Deepcom [Hu et al., 2020]	1	424,028	-
CONCODE [Iyer et al., 2018b]	1	2,184,310	-
Funcom [LeClair et al., 2019]	1	2,149,121	-
CodeT5 [Wang et al., 2021]	8	3,158,313	5,189,321
THEVAULT	10	34,098,775	205,151,985

A comparison between programming language of current available parallel dataset on function level

DATA CREATION

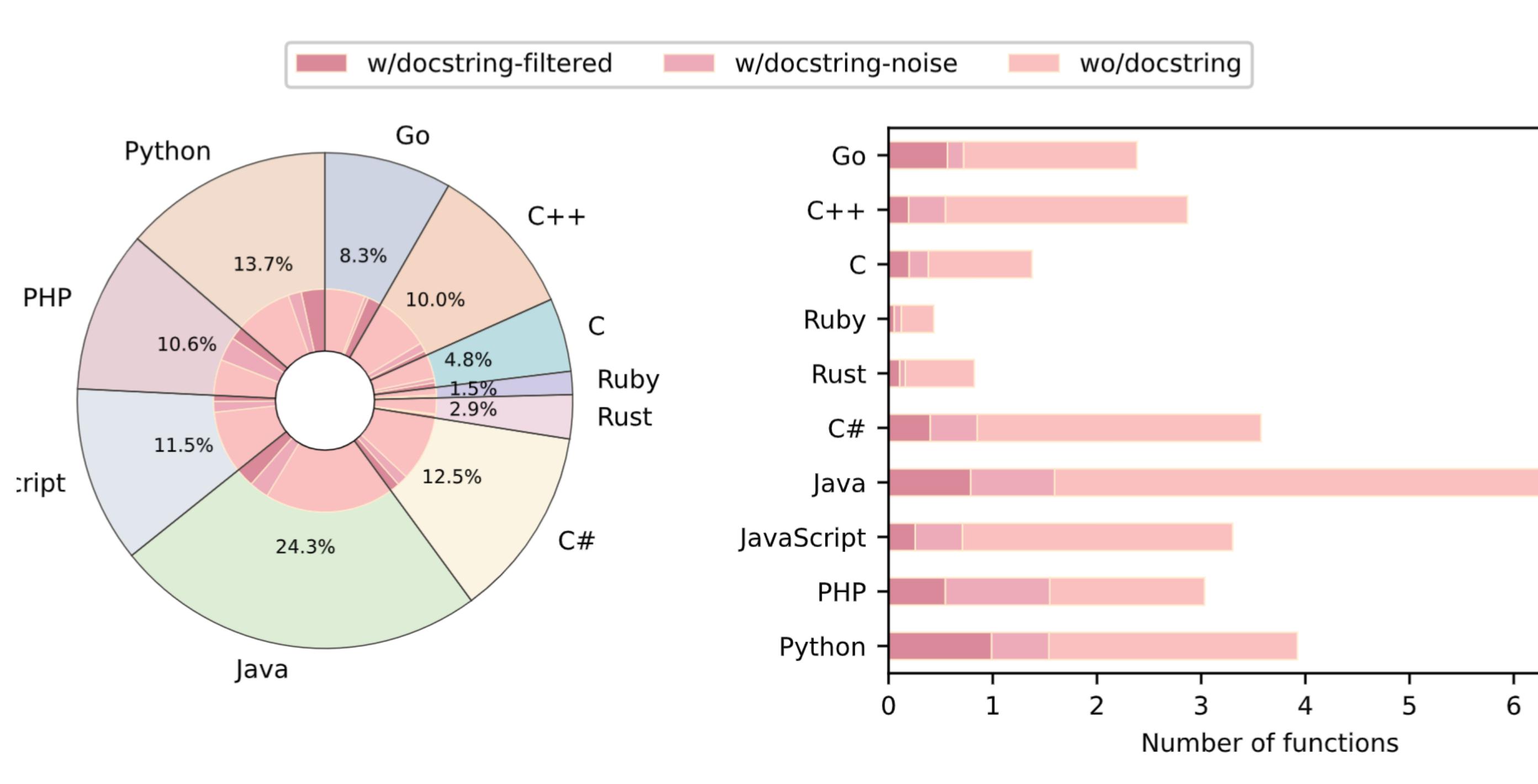
In The Vault, we leverage a subset of The Stack, recognized as the largest publicly accessible, multilingual, permissive licensed source code dataset. We comprise code in 10 prevalent programming languages and use a customized 'tree-sitter', an abstract syntax tree parser, to extract function, method, class, etc and their docstring.



Pipeline to create datasets of code blocks with comments D_{block} , unimodal code $D_{unimodal}$, and code-textpairs D_{paired} from raw source code

The Vault's construction process consist of 5 steps:

- Step 1:** Sampling a raw dataset of code snippets written in 10 programming languages, including Python, Java, Rust, JavaScript, Ruby, PHP, C, C#, C++.
- Step 2:** Use customized tree-sitter parsers to parse the code snippets into AST.
- Step 3:** Extract code features from the AST, such as class, function, codeblock, and corresponding docstring.
- Step 4:** Use the filtering process, which includes rule-based filters and a neural-based filter, to extract docstrings that match the semantic of the code snippet.
- Step 5:** Keep the samples that match and discard the one that do not, returning TheVault



Distribution and the number of functions by the presence of docstrings.

EXPERIMENTS

Model	Fine-tune dataset	pass@1	pass@10	pass@100
HUMAN EVAL				
CodeGen 350M	-	6.67	10.61	16.84
	Py/CodeSearchNet (250K)	2.76	8.76	14.72
	Py/TheVault raw/PyTheStack	3.74	10.57	16.26
	Py/TheVault	6.64	15.42	24.80
CodeGen 2B	-	8.14	18.12	30.07
	Py/TheVault	14.51	24.67	38.56
MBPP				
CodeGen 350M	-	7.46	24.18	46.37
CodeGen 2B	Py/TheVault	10.13	33.96	53.20
	-	18.06	45.80	65.34
CodeGen 2B	Py/TheVault	27.82	50.06	65.06

Code generation benchmarks running on CodeGen Multi model.

- CodeT5 and PLBART fine-tuned on (similar size) The Vault significantly better performance (~17.6% average) than on CodeSearchNet on task **Code Summarization**.
- On **Code Search**, superior results (~28% average) are observed in most languages when using (similar size) The Vault than on CodeSearchNet when fine-tuning CodeBERT, RoBERTa, UniXcoder.



Check our Github repository