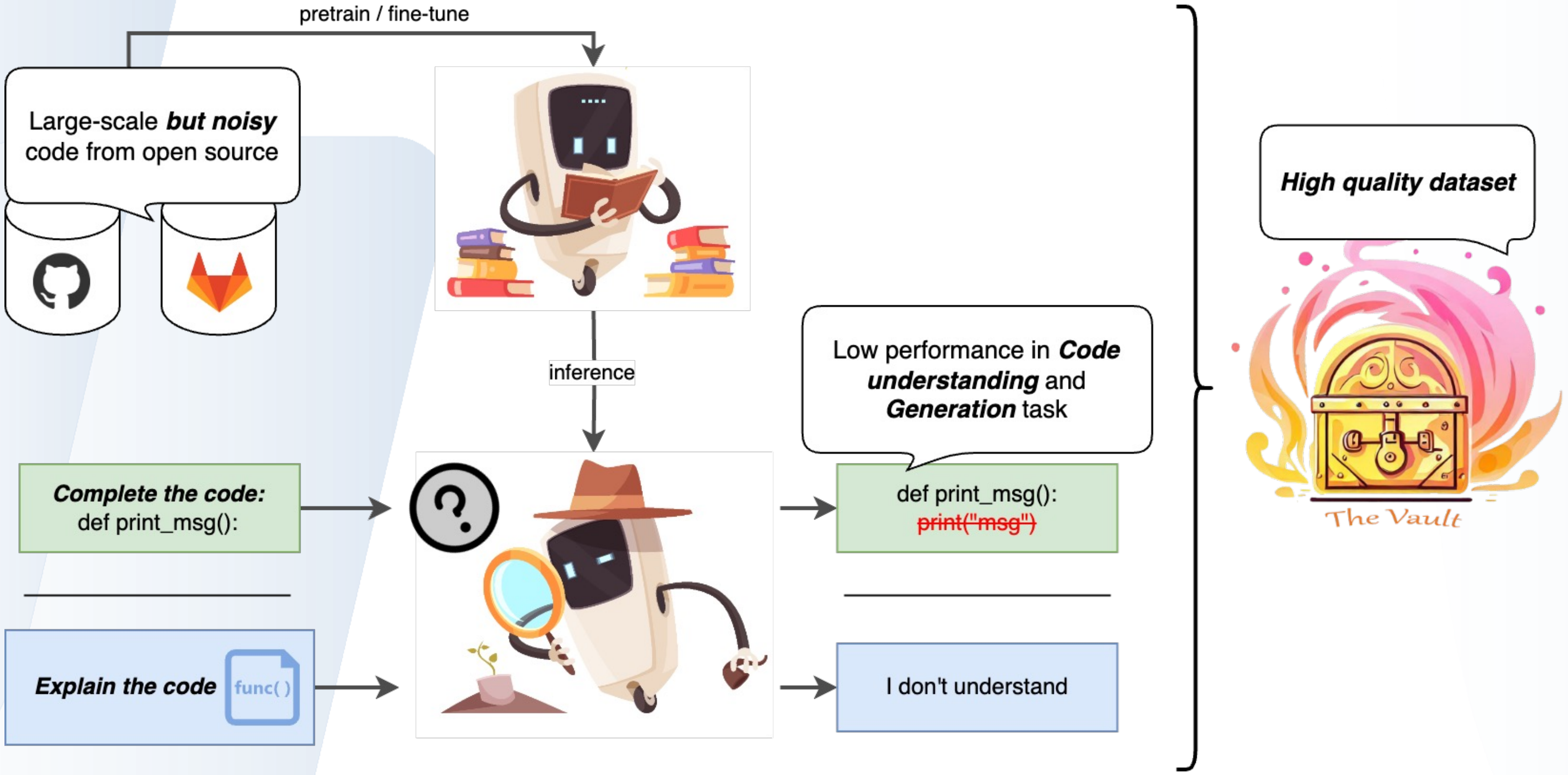
The Vault: A Comprehensive Multilingual Dataset for Advancing Code Understanding and Generation

EMNLP 2023

Software AICenter

TL;DR TheVault is the ***largest corpus*** containing a ***multilingual code-text dataset*** with over 40 million pairs covering 10 popular programming languages to leverage CodeLLMs understanding and generation abilities.



pretrain / fine-tune

Large-scale ***but noisy*** code from open source

inference

Complete the code:
def print_msg():

Explain the code func()

Low performance in ***Code understanding*** and ***Generation*** task

def print_msg():
~~print("msg")~~

I don't understand

High quality dataset

Check our repo