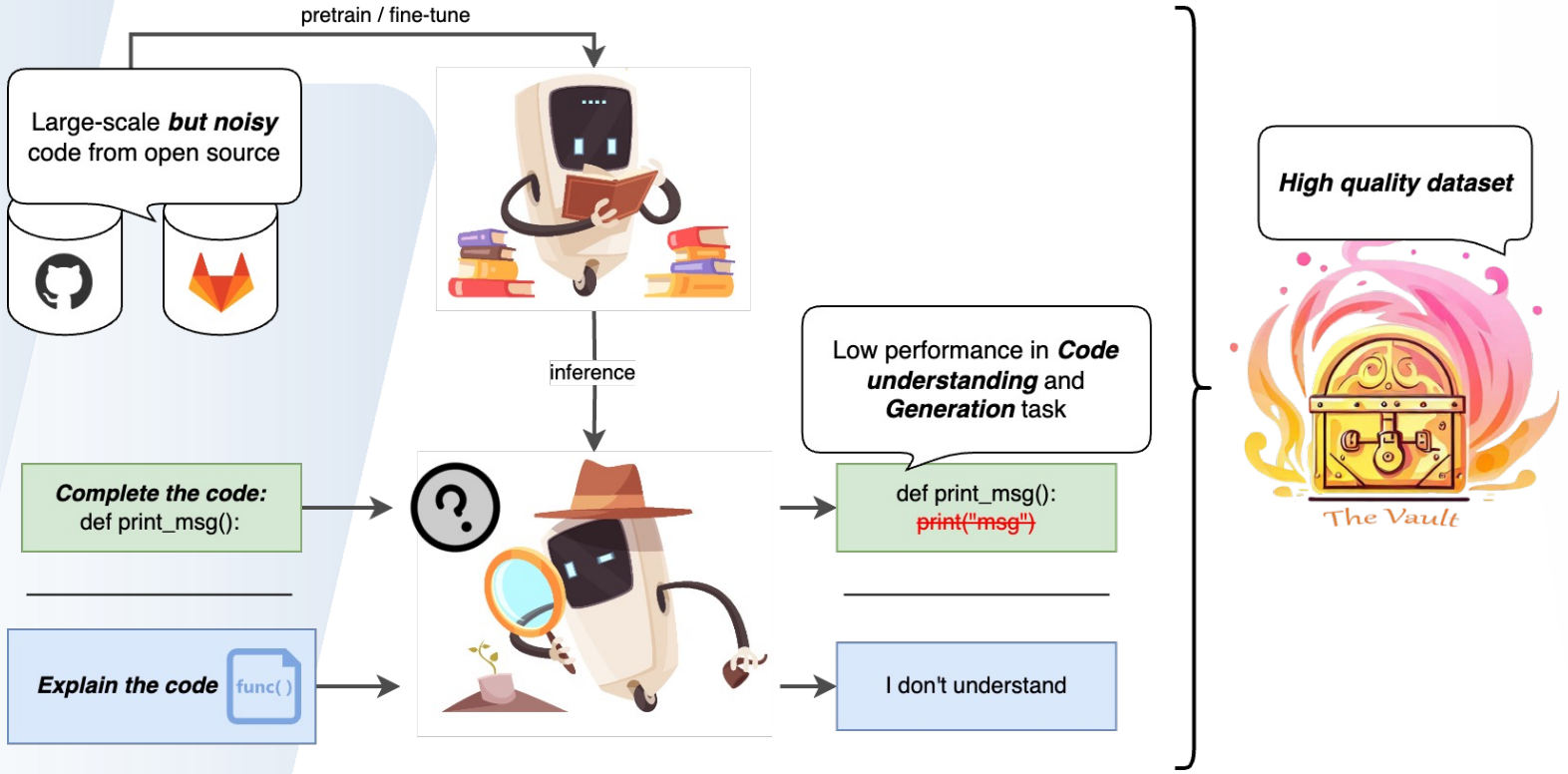
The Vault: A Comprehensive Multilingual Dataset for Advancing Code Understanding and Generation

EMNLP 2023

FPT Software AICenter

TL;DR

- TheVault is the ***largest corpus*** containing a ***multilingual code-text dataset.***
- CodeLLMs show a ***superior in performance*** when fine-tuned on The Vault for a wide range of tasks.



Check our repo