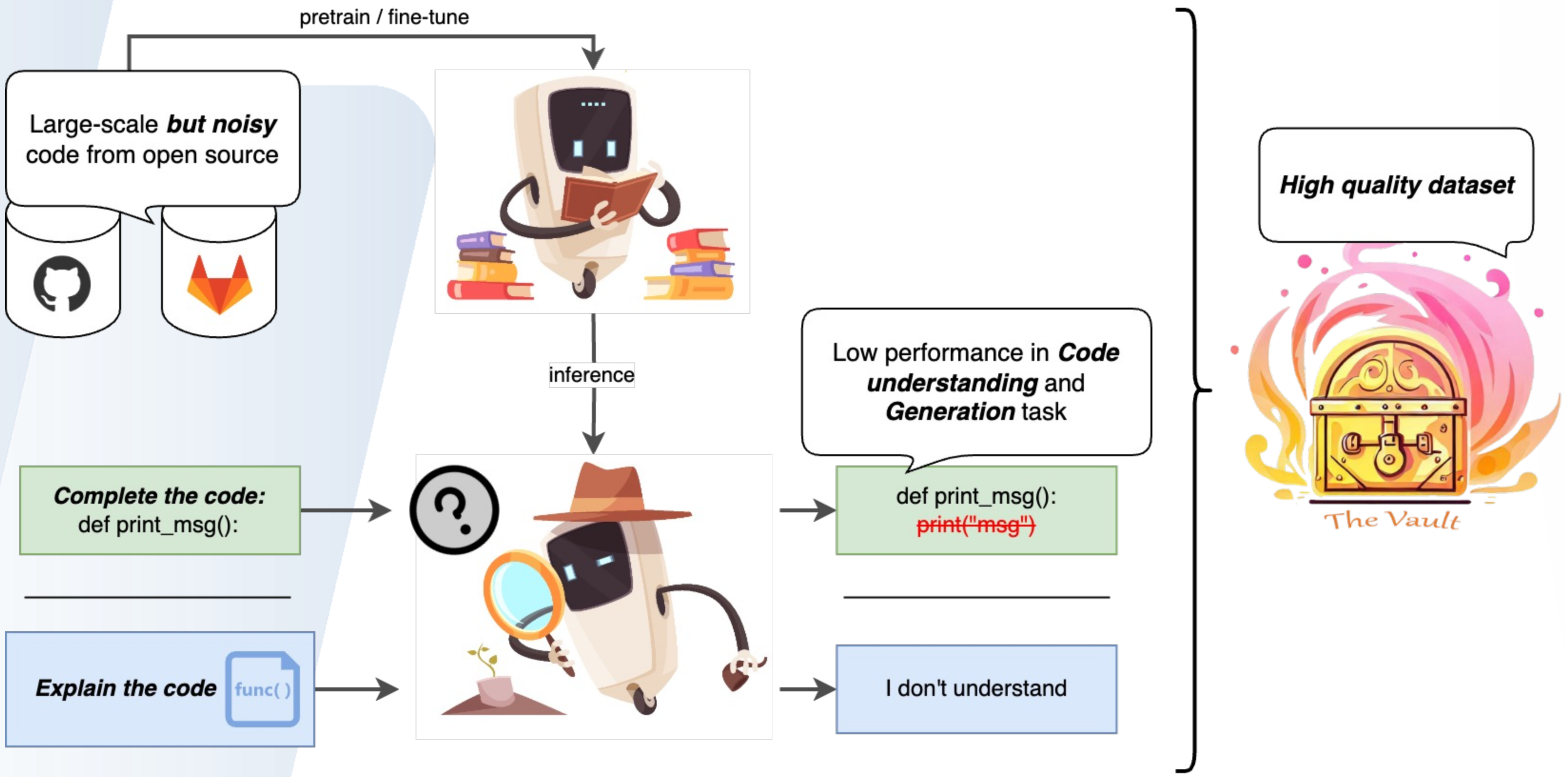
The Vault: A Comprehensive Multilingual Dataset for Advancing Code Understanding and Generation

EMNLP 2023

FPT Software AICenter

TL;DR

- TheVault is the ***largest corpus*** containing a ***multilingual code-text dataset*** with over 40 million pairs covering 10 popular programming languages.
- CodeLLMs (CodeGen, CodeT5, PLBART) show ***superior performance*** when fine-tuned on TheVault for a wide range of tasks, including code generation, code search, and code summarization.



Check our repo