算法类题目的代码自动生成能力评估

* 提供benchmark Java/C++/Python
* 生成代码的质量评估（功能性指标/非功能性指标）
* 生成代码的理解和修复

利用LLM（例如ChatGPT 3.5）的方法

（in-context learning、思维链、检索增强RAG）

提供解决思路/提供相似题目

关键词：

Novice programs

相关论文

[TSE 2024] Liu Z, Tang Y, Luo X, et al. No Need to Lift a Finger Anymore? Assessing the Quality of Code Generation by ChatGPT[J]. arXiv preprint arXiv:2308.04838, 2023.

评估ChatGPT生成的代码质量

[Arxiv 2023] Khan M A M, Bari M S, Do X L, et al. xcodeeval: A large scale multilingual multitask benchmark for code understanding, generation, translation and retrieval[J]. arXiv preprint arXiv:2303.03004, 2023.