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.

基于算法代码的数据集

[TOSEM 2023]Liu Y, Le-Cong T, Widyasari R, et al. Refining ChatGPT-generated code: Characterizing and mitigating code quality issues[J]. ACM Transactions on Software Engineering and Methodology, 2023.

生成代码的重新润色？

Pornprasit C, Tantithamthavorn C. GPT-3.5 for Code Review Automation: How Do Few-Shot Learning, Prompt Design, and Model Fine-Tuning Impact Their Performance?[J]. arXiv preprint arXiv:2402.00905, 2024.

给出了提示的改进路线

Mu F, Shi L, Wang S, et al. ClarifyGPT: Empowering LLM-based Code Generation with Intention Clarification[J]. arXiv preprint arXiv:2310.10996, 2023.