南開大學
Nankai University

计算机学院

大语言模型代码书写能力测试

基于多模型自动化测试框架的实验评估

姓名：陈宇昕

学号：2310675

专业：计算机科学与技术

2025 年 3 月 28 日

目录

1. 引言

随着大语言模型技术发展，其代码生成能力日益增强，但不同模型之间存在显著差异。本次创新作业构建自动化测试框架，针对代码生成能力进行系统评估，特别关注幻觉现象的检测与分析。

2. 测试方法

我们构建了基于 Python 的自动化测试框架，集成了 API 接口封装、沙箱执行环境、动态测试验证、结果分析和幻觉检测模块。核心流程包括：配置题目需求、统一 API 调用、代码生成获取、沙箱执行测试、结果比对和幻觉分析。

幻觉检测模块针对以下几类情况进行监测：

- 引用不存在的库或模块
- 使用未定义的函数或变量
- 错误的 API 调用方式
- 不符合语法的代码结构

测试实现通过 subprocess 模块在隔离环境执行代码，确保测试安全性和结果可靠性，同时记录执行时间和内存使用情况。

3. 测试对象与内容

本研究选取四个代表性大语言模型：DeepSeek-R1、GeneralV3.5（星火）、Doubao-1.5-Pro 和 Moonshot-V1（Kimi）。测试内容包括三道算法题目，涵盖二叉树、最短路径和 KMP 算法等不同类型编程挑战，考察模型在基础数据结构、图论和字符串算法方面的能力。

4. 测试结果

通过系统测试，我们获得以下关键发现：

题目	DeepSeek-R1	GeneralV3.5	Doubao-1.5-pro	Moonshot-V1
第一题通过率	23%	17%	9%	12%
第二题通过率	92%	54%	38%	61%
第三题通过率	88%	32%	87%	76%
幻觉出现率	7%	21%	18%	13%

表 1. 各模型测试结果

各模型在不同问题类型上表现差异明显：DeepSeek-R1 整体表现最佳，尤其在图论问题上；所有模型在二叉树等复杂结构问题上表现不佳；在幻觉现象方面，GeneralV3.5 出现率最高 (21%)，DeepSeek-R1 最低 (7%)。

5. 大语言模型推理成本评估

根据 Brown 等人 [1] 和 Chowdhery 等人 [2] 的研究，推理成本受模型规模、上下文长度、推理策略、硬件加速和批处理优化等因素影响。Wang 等人 [3] 指出，代码生成任务中，推理成本与代码质量非简单线性关系。

针对 DeepSeek-R1 模型的分析显示，其在代码生成任务中具备高效的输入处理能力和上下文利用率，但在处理复杂算法（特别是树结构操作）时，推理延迟增加约 40%。

6. 推理成本改进建议

基于研究分析，我们提出以下改进建议：

1. **专业知识蒸馏**：将大模型编程能力转移到更小的专用模型
2. **检索增强生成**：结合代码库检索，减少” 从头生成” 的内容

3. **任务分解策略**：将复杂任务分解，降低单次推理复杂度
4. **自适应批处理**：根据问题复杂度动态调整批处理策略
5. **硬件感知优化**：根据部署硬件特性调整模型量化策略

7. 结论

本研究通过自动化测试框架，全面评估了四种大语言模型的代码生成能力。研究发现，当前模型在经典算法实现方面表现较好，但在处理复杂数据结构和创新性问题时仍有不足。幻觉现象普遍存在但程度不同，DeepSeek-R1 整体表现最佳但推理成本较高。我们提出的优化建议有望在保持代码质量的同时提高模型效率。作为辅助编程工具，建议开发者保持批判思维，特别是对于复杂或创新性问题。

8. 项目资源与代码库

本研究的完整代码、报告和相关资源已托管在 GitHub 上，欢迎访问和参考：

https://github.com/daybreak159/homework1

8.1. 代码实现

仓库中包含三个核心 Python 脚本，分别对应本文中讨论的测试框架核心组件：

- `1.py` - 二叉树动态规划问题测试模块，实现了分数最大化和前序遍历算法验证
- `2.py` - 图论最短路径问题测试模块，验证满足约束条件的图计数功能
- `3.py` - KMP 算法测试模块，包含幻觉检测和性能分析功能

这些代码展示了本文所述测试框架的实际实现，包括 API 调用封装、沙箱执行环境、动态测试验证和结果分析模块。

8.2. 报告文档

仓库中提供了本研究报告的两个版本：

- 完整版报告 - 详细阐述了测试方法、实验设计和研究结果
- 作业要求版报告 - 提供研究的核心发现和关键要点概述（约 1000 字）

8.3. 宣传海报

为了更好地展示研究成果，仓库中还包含两张宣传海报：

- 易拉宝尺寸海报 (80cm×200cm) - 适合会议和展览展示
- 墙贴海报 (A0 尺寸: 84.1cm×118.9cm) - 适合学术海报展示

这些海报采用了本文介绍的设计方案，直观展示了研究的核心内容、测试框架和实验结果，可作为研究成果的直观参考。

参考文献

[1] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020. 2

[2] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 2

[3] Wei Wang, Zejun Zhao, Xuxi Chu, Jiazhan Jiao, Binyuan Xie, Zhe Liu, Guosun Wang, Yaoyuan Wang, Yi Ding, Jimmy Lin, et al. Code generation and understanding: Dimensions of machine learning models. *arXiv preprint arXiv:2301.08485*, 2023. 2