大语言模型在计算机科学领域的能力评测

一、实验目的与流程

本实验旨在系统评测主流大语言模型（LLM）在计算机科学领域的知识掌握、推理、编程与指令理解等多方面能力。通过标准化问题集和自动化评测流程，定量比较不同模型的表现，分析其能力边界、缺陷与风险，并为后续模型优化和应用提供参考。

1.1 评测对象

本次评测涵盖以下主流大语言模型：

- GPT-4o-mini
- GPT-3.5-turbo
- DeepSeek-Chat
- Claude-3.5-Haiku-20241022
- Gemini-2.5-Flash-Preview-04-17

1.2 问题设计

问题集覆盖计算机科学五大核心领域：

- 计算机历史（computer_history）
- 离散数学（discrete_math）
- 程序设计（programming）
- 人工智能（artificial_intelligence）
- 计算机系统（computer_systems）

问题类型包括：

- 事实性问答（factual）
- 多项选择题（multiple_choice）
- 指令型任务（instruction，如代码实现、证明等）

每个问题包含唯一ID、类别、类型、问题内容，以及部分题目配有标准答案（reference）。

1.3 评测流程

1. **自动问答：**
   - 读取问题集（computer_science_questions.json），逐题调用各模型API，获取模型回答，保存为 result/{model}_results.json。
2. **自动化评分：**
   - 设计多套评分模板（prompts.json），针对不同题型（事实、选择、指令等）自动生成评测提示。
   - 将模型回答与评分模板结合，调用高阶LLM（如GPT-4o）进行自动评分与点评，结果保存为 evaluations/{model}_evaluation.json。
3. **结果统计与分析：**
   - 解析评分结果，统计各模型在不同领域、题型下的平均分与分布。
   - 结合自动化点评与人工复核，分析模型能力边界、缺陷与风险。

二、评测结果与定量分析

2.1 总体表现

1. **按题型统计**

题型	题数	平均分	最高分	最低分	≥8分比例
GPT-4o-mini					
事实问答	8	8.88	10	8	100%

题型	题数	平均分	最高分	最低分	≥8分比例
选择题	9	9.11	10	7	88.9%
指令题	8	9.00	10	6	100%
总计	25	8.96	10	6	96%
GPT-3.5-turbo					
事实问答	8	9.25	10	9	100%
选择题	9	9.33	10	7	100%
指令题	8	9.00	10	7	100%
总计	25	9.16	10	7	100%
DeepSeek-Chat					
事实问答	8	9.00	10	8	100%
选择题	9	9.00	10	3	88.9%
指令题	8	8.63	10	6	100%
总计	25	8.88	10	3	96%
Claude-3.5-Haiku					
事实问答	8	9.13	10	8	100%
选择题	9	9.33	10	6	88.9%
指令题	8	9.00	10	6	100%
总计	25	9.16	10	6	96%
Gemini-2.5-Flash					
事实问答	8	9.13	10	8	100%
选择题	9	9.33	10	7	100%
指令题	8	9.25	10	7	100%
总计	25	9.24	10	7	100%

2. 按领域统计

领域	题数	平均分	最高分	最低分	≥8分比例
GPT-4o-mini					
计算机史	5	8.60	10	8	100%
离散数学	5	9.60	10	9	100%
程序设计	5	8.80	10	6	80%
人工智能	5	9.20	10	8	100%
计算机系统	5	8.80	9	7	100%

领域	题数	平均分	最高分	最低分	≥8分比例
GPT-3.5-turbo					
计算机史	5	9.40	10	9	100%
离散数学	5	9.20	10	9	100%
程序设计	5	9.20	10	7	100%
人工智能	5	9.40	10	9	100%
计算机系统	5	8.80	9	7	100%
DeepSeek-Chat					
计算机史	5	9.40	10	8	100%
离散数学	5	9.40	10	9	100%
程序设计	5	8.80	10	6	80%
人工智能	5	9.20	10	9	100%
计算机系统	5	7.60	10	3	80%
Claude-3.5-Haiku					
计算机史	5	9.40	10	9	100%
离散数学	5	9.60	10	8	100%
程序设计	5	9.00	10	6	80%
人工智能	5	9.60	10	9	100%
计算机系统	5	8.60	10	6	80%
Gemini-2.5-Flash					
计算机史	5	9.20	10	9	100%
离散数学	5	9.60	10	9	100%
程序设计	5	9.40	10	7	100%
人工智能	5	9.20	10	9	100%
计算机系统	5	8.60	10	7	100%

3. 各模型-题型均分汇总表

模型名称	事实问答均分	选择题均分	指令题均分	总平均分
GPT-4o-mini	8.88	9.11	9.00	8.96
GPT-3.5-turbo	9.25	9.33	9.00	9.16
DeepSeek-Chat	9.00	9.00	8.63	8.88
Claude-3.5-Haiku	9.13	9.33	9.00	9.16
Gemini-2.5-Flash	9.13	9.33	9.25	9.24

三、能力边界、缺陷与风险分析

3.1 能力边界

- **事实性知识**：主流LLM在计算机科学基础知识、历史、常见理论等方面表现优异，绝大多数问题能给出准确、结构化的答案。
- **选择题推理**：对于标准选择题，模型不仅能选出正确答案，还能给出较为合理的解释和排除理由，表现接近人类本科生水平。
- **指令与代码实现**：在常见算法、数据结构实现等任务上，模型能生成结构合理、可运行的代码，并能解释设计思路。但在复杂算法（如红黑树删除）、边界情况处理等方面，部分模型存在实现不完整或遗漏的情况。

3.2 主要缺陷

- **复杂算法实现不完整**：如红黑树的删除操作，部分模型仅给出框架或注释，未能完整实现，导致评分明显下降。
- **专业细节遗漏**：在涉及底层原理、协议细节（如Belady异常、缓存一致性协议）时，部分模型存在知识点混淆或解释不够严谨的现象。
- **多义性与歧义处理**：对于题目描述不够精确或存在多种解释的情况，模型有时会给出模棱两可的答案，缺乏主动澄清和假设说明。
- **自信错误**：个别模型在选择题中会自信地给出错误答案（如Belady异常相关问题），并给出看似合理但实际错误的解释，存在误导风险。

3.3 风险与应用建议

- **自动化评测风险**：完全依赖LLM自评可能掩盖模型的知识盲区和推理漏洞，建议结合人工抽查和多模型交叉验证。
- **代码生成安全性**：模型生成的代码需经过严格测试和审查，避免因边界条件遗漏或实现不完整导致潜在安全隐患。
- **专业应用限制**：在高可靠性、强一致性要求的场景（如分布式系统、金融系统）中，LLM的建议和代码应仅作为辅助，最终决策需由专业人员把关。

四、结论与建议

本次实验表明，主流大语言模型在计算机科学领域已具备较强的知识掌握和推理能力，能够胜任大部分基础教学、答疑和代码生成任务。但在复杂算法实现、底层原理解释和专业细节把控方面仍有提升空间。未来应加强模型在专业领域的持续训练，完善自动化评测与人工复核结合的评测体系，提升模型的可靠性和实用性。

Copilot的上下文窗口（以Claude 3.5 Haiku为例）大约支持 200k tokens，这远超一般AI模型的4k/8k/32k token限制。可以一次性把所有大模型的评测结果全部附入，从而分析各自优劣。

五、评分标准（附录）

满分10分，采用多套评分模板，针对不同题型设计如下（详细英文评分模板见prompts.json）：

5.1 事实性知识题（factual）

- **准确性**：所有关键事实是否正确呈现
- **完整性**：是否覆盖问题所有方面
- **清晰度**：解释是否清楚、结构是否合理

5.2 多项选择题（multiple_choice）

- **选项判断**：是否选对正确答案
- **解释质量**：是否清楚说明正确选项理由，并简要说明其他选项为何不正确

5.3 指令型任务（instruction）

- **指令遵循度**：是否完成所有要求
- **工作质量与正确性**：产出是否正确、完整
- **专业性**：是否体现应有的领域知识