



# 人工智能前沿探索实践

LLM实践 - 评测开源大模型

复旦大学计算机科学技术学院

2025-4-16



# 任务介绍

## ■ 任务内容

- 1.探究大模型的在具体专业领域下的能力范围，能力表现，
- 2.分析不同prompt构造方法对大模型性能的影响。
- 3.分析使用大模型评估大模型的可行性。

## ■ 任务环境

中转api ( 如果免费额度用完的话，可以自己上网找一些其他的中转api )

<https://api.deerapi.com/>

对应代码示例

<https://helpdoc.deerapi.com/code-sample>

批量化处理：可把问题和参考答案写成json的格式



# 任务介绍

## ■ 报告要求

报告标题：“学号\_姓名\_随堂练习3.pdf”

报告内容：

- 1、报告中包含定量的评测结果
- 2、分析各模型的能力边界、缺陷和风险，可尝试用LLM撰写分析报告，并结合自己的分析思考进行总结  
(这部分无字数限制)
- 3、构造的评分标准需要放在报告最后
- 4、JSON文件另外提交，与报告一起打包为学号\_姓名\_随堂练习3.zip



# 任务介绍

## ■ 任务要求

- 1.在5种任/专业领域共25个问题中评估5种LLMs，在不同prompt下的表现。
- 2.评估LLM可通过人为检查/通过参考答案交给GPT4o等高级LLM来评判
- 3.分析各模型的能力边界、缺陷和风险，可尝试用LLM撰写分析报告

```
payload = json.dumps({  
    "model": "gpt-3.5-turbo", # 这里是你需要访问的模型，改成上面你需要测试的模型名称就可以了。  
    "messages": [  
        {  
            "role": "system",  
            "content": "You are a helpful assistant."  
        },  
        {  
            "role": "user",  
            "content": "xxxxx"  
        }  
    ]  
})
```



# 任务介绍

## ■ 任务举例

### 评测LLM在计算机科学领域的表现

分类：计算机史、离散数学、程序设计、人工智能、计算机系统

Prompt方式：直接询问（是什么）、选择题（ABCD）、指令（帮我写一份代码）等

评测答案：前两者可直接提供答案+解题思路（可选），后面则只能人为或高级LLM评估

LLM-Eval / demo\_questions.json

YesianRohn upload code e864d9e · 2 months ago History

Code Blame 35 lines (35 loc) · 9.74 KB

```
1 {"question_id": 1, "category": "coding", "turns": ["实现一个Python程序，逐行读取文本文件并计算文件中特定单词的出现次数。"]
2 {"question_id": 2, "category": "coding", "turns": ["实现一个Python函数，使用动态编程找出两个输入字符串的最长公共子序列。"]
3 {"question_id": 3, "category": "coding", "turns": ["Implement a regular expression in Python to validate an em
4 {"question_id": 4, "category": "coding", "turns": ["编写一个Python程序，使用动态编程找出第n个斐波纳契数。"]}
5 {"question_id": 5, "category": "coding", "turns": ["Implement a binary search algorithm to find a specific ele
```



# 任务介绍

## ■ 任务举例

### 评测LLM在计算机科学领域的表现

分类：计算机史、离散数学、程序设计、人工智能、计算机系统

Prompt方式：直接询问（是什么）、选择题（ABCD）、指令（帮我写一份代码）等

评测答案：前两者可直接提供答案+解题思路（可选），后面则只能人为或高级LLM评估

LLM-Eval / demo\_prompts.json

YesianRohn upload code e864d9e · 2 months ago History

Code Blame 3 lines (3 loc) · 2.82 KB

Raw Copy Download Edit

```
1 {"name": "base-v1", "type": "single", "system_prompt": "You are a helpful assistant.", "prompt_template": "|
2 {"name": "math-v1", "type": "single", "system_prompt": "You are a helpful assistant.", "prompt_template": "|
3 {"name": "multi-turn", "type": "single", "system_prompt": "Please act as an impartial judge and evaluate the
```



# 任务介绍

## ■ 任务举例

### 评测LLM在计算机科学领域的表现

分类：计算机史、离散数学、程序设计、人工智能、计算机系统

Prompt方式：直接询问（是什么）、选择题（ABCD）、指令（帮我写一份代码）等

评测答案：前两者可直接提供答案+解题思路（可选），后面则只能人为或高级LLM评估

参考思路：

- 1、利用demo\_questions.json（包含问题id（question\_id），问题类别（"category"）、（多轮）对话内容（"turns"）、（可选）标准答案（"reference"）（如math题），调用LLM API获取回答
- 2、定义评价标准，可以针对每个category定义一套评价标准，建立评分模板
- 3、将回答插入自定义的评分模板，再次调用LLM，让其按评分维度自评
- 4、解析评分结果并保存

# THANKS

4/16/2025