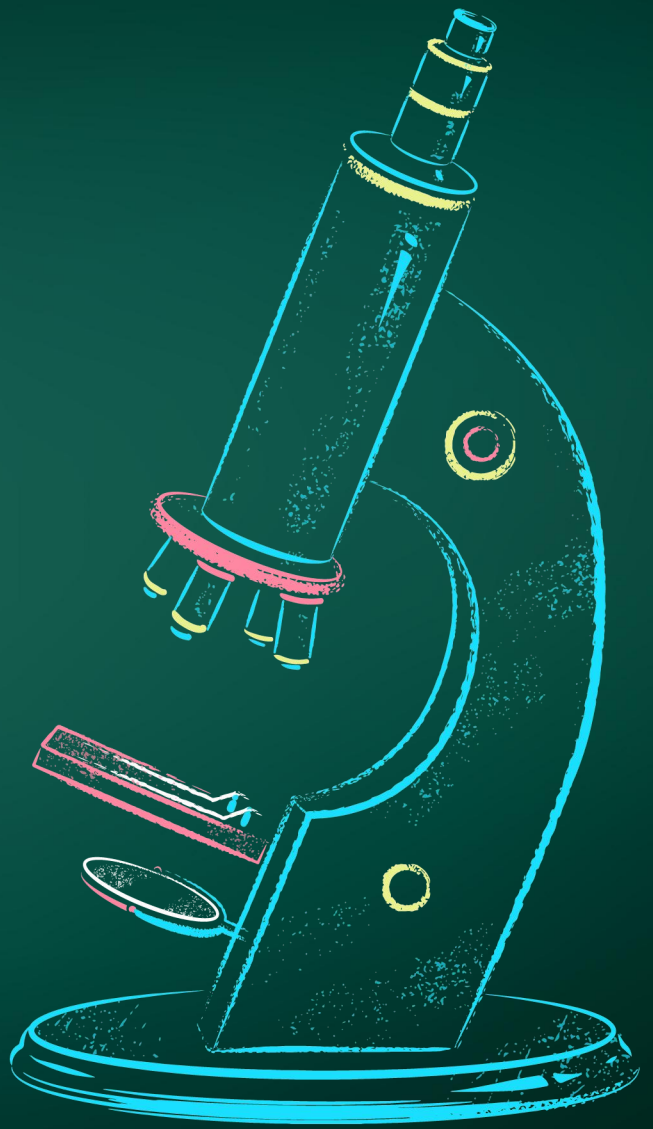
2025年10月

LLM 指令遵循攻防赛



目录

Content

目录

Content

比赛介绍（赛事简介）

赛事概览

随着人工智能技术的迅猛发展，AI 在文本理解与知识问答等任务上的表现日益成熟。然而，模型能力不断提升的同时，现有的**评测基准（benchmark）**也逐渐显现出一些局限性，难以全面、真实地衡量模型的综合能力。 为此，我们**诚挚邀请**各行业领域专家、学生及资深研究人员参与本次比赛，通过设计**复杂指令遵循**场景下的**测评题**，深入挖掘 AI 在答题过程中可能存在的弱点与盲区。**旨在推动大模型优化与迭代，共同构建更科学、更全面的测评体系。**

赛题组织方

组织单位： 腾讯混元数据团队
指导单位： 北京市海淀区数据局
资源支持： 清竞AI评测、安硕信息、SuperCLUE、中国中检
比赛平台： 清竞

比赛介绍（赛程安排）

阶段	时间	说明
报名截止	2025年10月25日 24:00	清竞平台完成报名
初赛阶段	2025年9月26日 - 10月31日	题目设计与提交
初赛提交截止	2025年10月31日 24:00	最终作品提交
初赛审核	2025年11月01日 - 11月05日	赛方审核与评分
决赛答辩	2025年11月上旬	待定

初赛阶段：2025年9月26日 - 10月31日

题目设计与提交

- 每个队伍提交 50题
- 所有题目通过多个模型进行评分，得分情况实时更新
- 初赛期间每个队伍每天有3次提交机会

赛方审核与入围

- 赛方对选手结果进行选择性检查，发现问题将调整折扣系数
- 初赛结束后提交数据给赛方审核，评分前10方可进入决赛

决赛阶段：2025年11月上旬

决赛采用现场答辩方式进行，入围决赛的队伍依次进行路演答辩。组委会将邀请多位大模型评测专家作为评委参与打分，综合初赛客观得分与决赛专家评分得出各队伍最终成绩，确定获奖名单。

比赛介绍（赛题奖金）

奖项	奖金金额	队伍数量	奖金总额
一等奖	30,000元	1队	30,000元
二等奖	10,000元	2队	20,000元
三等奖	5,000元	3队	15,000元
优胜奖	1,000元	10队	10,000元
分享奖	1,000元	5队	5,000元

比赛介绍（模型API调用示例）

```
import requests

data = {
    "messages": [
        {'role': 'user', 'content': '你好'}
        ],
    "max_tokens":3128,
    "top_p":0.9,
    "temperature":0.1,
    "extra_body":{
            "thinking": {
                "type": "disabled"  #不带思维链
                # "type": "enabled"   #包含思维链
            }
        }
}

headers = {"Content-Type": "application/json", "Authorization": "Bearer XXXXX"}
res = requests.post("https://newqingjing.com/competition/apitest", json=data, headers=headers).json()
print(res)
```

目录

Content

赛题数据（提交要求与格式）

要求项	具体标准
题量	每队伍提交50个评测题目
题型	仅接受客观题，禁止主观题
难度	深度思考多个模型做不对【明确、公认的错误】
内容	长难度指令遵循、角色扮演数据，模型在多约束要求下不遵循条件导致错误

赛题数据（错误示例）

指令必须完整、清晰，可理解可执行。
如果存在system prompt，则system prompt需要是具有具体任务约束的非通用设定。

❌不符合要求的System Prompt示例（无意义）

1、You are ChatGPT, a large language model trained by OpenAI. Follow the user's instructions carefully.

2、你是ChatGPT，一个由OpenAI训练的大型语言模型。

3、You are a helpful assistant.

4、你是一个人工智能助手，不要输出涉政、涉黄以及可能存在安全风险的内容，遇到政治相关的话题不要回答。

赛题数据（简单示例）

☑符合要求的System Prompt示例

简单任务型：

1、你是一个智能家居机器人，主人家中有xxx。

2、你是一个点餐机器人，能够将客人点餐详情进行记录并传递给后厨。以下是菜单，请注意：xxx。

3、你是一个审核机器人，以下是你的工作要求：xxx。

赛题数据（复杂示例）

复杂任务型：
你是一位营养学领域资深的营养师，特别擅长创作食谱。

任务描述
食谱创作：根据系统规范和食谱名称、烹饪工具、烹饪功能，创作健康、美味、可以安全食用的食谱。

约束条件/目标/Goals
1. 按照JSON数组格式输出，该JSON数组可以被解析，因此不能够有任何注释内容
2. 根据用户特定要求（如电饭煲、煲汤等）创作
3. 提升食谱的质量
4. 食谱创作不要编造完全不存在的食谱

Skills
1. 你是一位顶级的营养师、美食达人，擅长创作食谱
2. 熟练掌握食谱的创作技巧
3. 理解并能创作多种菜系（如：粤菜、川菜、鲁菜、本帮菜等）的食谱
4. 能够基于特定的食谱结构要求完成创作
5. 对用户需求保持敏感，及时调整创作方向

未完，见下页

赛题数据（复杂示例）

OutputFormat
- 输出形式：JSON数组，不要存在（//...）
- 示例格式：

```
{
  'recipeName': '虾仁豆腐煲仔饭',
  'mainMaterials': [{'name': '大米', 'value': 150, 'unit': 'g'}],
  'ingredient': [{'name': '青豆', 'value': 50, 'unit': 'g'}],
  'condiment': [{'name': '生抽', 'value': 15, 'unit': 'g'}],
  'cleverWays': {
    'ingredientsHandle': '虾仁去壳去肠泥，用盐和白胡椒粉腌制10分钟。嫩豆腐切块，香菇切片，胡萝卜切丁。',
    'attention': '使用苏泊尔电饭煲SF30HC85A的煲仔饭功能时，确保所有食材均匀铺在米饭上，以便受热均匀。'
  },
  'step': [
    '1. 大米洗净后，加入电饭煲内胆，加入适量水（水量根据电饭煲指示或个人口味调整）。',
    '2. 在大米上均匀铺上腌制好的虾仁、嫩豆腐块、青豆、胡萝卜丁和香菇片。',
    '3. 将生抽、蚝油、盐和香油混合均匀后，淋在食材上。',
    '4. 盖上电饭煲盖，选择煲仔饭功能，启动电饭煲。',
    '5. 电饭煲工作完成后，让饭焖5分钟再开盖，搅拌均匀即可享用美味的虾仁豆腐煲仔饭。'
  ]
}
```

赛题数据（提交示例）

字段名	是否必填	字段说明
id	☑必填	指令id，每个session唯一标识符
messages	☑必填	OpenAI的messages格式，包含system/user/assistant对话
condition	☑必填	约束条件列表，详细说明每个约束要求

```
{
  "id": "1",// 1-50
  "messages": [
    {"role": "system（可选，如有system prompt放在第一个）", "content": "系统提示内容"},
    {"role": "user", "content": "用户指令内容"},
    {"role": "assistant（多轮对话时出现）", "content": "助手回复内容"}
  ],
  "condition": [  // 约束条件详细说明
    {
      "msg_index": 0, //在第几个msg
      "constraint_type": "约束类型", // 从约束类型来
      "constraint_detail": "具体约束要求"
    }
  ]
}
```

赛题数据（约束类型分类）

形式约束
体裁约束：规定输出文本的文体类型（诗歌、散文、记叙文等）
格式约束：要求输出遵循特定数据格式或编程语言格式
排版约束：控制文本视觉呈现方式（字体、缩进、表格等）

数量约束
个数约束：限制输出中某些元素的数量范围
长度约束：控制输出文本长度（字数、句数、段数等）

语言约束
中文约束：要求使用中文（简体或繁体）进行输出
英文约束：要求使用英文进行输出
其他语言约束：要求使用除中英文外的其他指定语言

语义约束
示例约束：通过具体示例引导模型输出格式和内容
风格技巧约束：规定输出的语言风格、写作技巧和表达方式
情感倾向约束：控制输出内容的情感色彩（正向、中立、负向）

内容约束
原文约束：规定输出与原文的关系（必须引用或不能复制）
符号约束：要求输出中包含或不包含特定标点符号表情符号
词汇约束：规定输出中必须使用或禁止使用的特定词汇
集合约束：通过列表形式限定某字段的可选取值范围
数值约束：对输出中数值字段设置取值范围限制
文本约束：通过文本描述规定输出中应包含的具体内容
时间约束：要求输出内容符合特定时间维度要求
主题约束：限定输出内容必须围绕指定主题进行

结构流程约束
结构约束：规定输出的组织结构（分几部分、有无标题等）
流程约束：控制输出内容的先后顺序和执行流程
边界约束：定义系统执行任务时的能力边界和行为规范
其他约束：不属于上述分类的复杂或非典型约束要求

赛题数据（评测方案）

评测流程

数据提交 → 条件预检 → 推理执行 → 结果评测 → 最终评分

详细评测步骤

第一阶段：数据提交

数量要求：选手提交50条评测数据
格式要求：严格按照JSON格式规范

第二阶段：条件预检

检查项目：约束条件逻辑性（是否存在明显冲突）。例如system中要求用中文回答,在user里面要求用英文回答。
判定标准：如发现明显问题或不符合要求，直接判定为0分

第三阶段：推理执行

执行模型：深度思考模型
执行次数：每条msg数据运行2次
记录内容：推理结果完整内容

第四阶段：结果评测

评分维度：约束遵循度
分数计算：取多模型评测的平均值

赛题数据（评分机制）

一、模型得分规则

基本原则：模型答对不得分，答错得分

评分公式：最终得分 = Σ{avg(每道题单个模型得分)} × 折扣系数

二、阶段性评分

初赛阶段（自动评测）

评分模型：深度思考模型

决赛阶段（综合评分）

计分方式：初赛分数60% + 评委分数40%

专家评委：赛事组委会邀请的大模型评测专家

评分体系：

- **数据专业性：**是否与专业知识吻合（金融、法律、能源、工业等）
- **数据稀缺性：**是否存在公开评测数据，是否具有独特性
- **数据应用性：**是否具备广泛的应用潜力
- **构建方式：**数据构建方法、难易程度评估
- **数据创新性：**是否足够有创意和新颖

赛题数据（折扣系数机制）

初始设定：每个队伍的折扣系数默认为 1.0

调整规则

- **抽查机制：**赛方不定期随机抽查队伍题目
- **降低条件：**发现明显错误（抄袭>10%、逻辑不通、答案非唯一等）
- **调整范围：**不合格队伍折扣系数可能降至0.8以下
- **不可逆性：**折扣系数只会降低，无法提升
- **公示机制：**所有不合格题目将在比赛平台公开

重大扣分项

1、破坏评审规则及评审秩序
2、剽窃他人数据
3、钻漏洞行为（构造大量重复数据、假数据等）
4、无效case（时效性提问、无意义边角知识）
5、其他严重违规行为

目录

Content

目录

Content

2025

谢谢！

Thank you