寻找AI“盲区”

如何设计高区分度的常识类评测题

从“出题思路”到“技术实现”的全过程解析

演讲者：大模型做不队

日期：2025年10月

核心思路：从“规则”的陷阱，转向“常识”的盲区

- 当代大模型特点：遵循复杂指令能力强，理解人类“理所当然”的世界运作常犯“啼笑皆非”错误。
- 目标：精准找到并攻击AI的“常识盲区”。

常识对人类极其简单，但对AI极其困难。

植根于人类物理经验、社会互动和思维模式中的隐性知识。

是检验AI真实理解能力而非模式匹配能力的关键。

1

攻击“文本世界” vs “真实物理世界”的鸿沟

AI缺乏对真实世界的物理直觉。例如，在第3题“太空拧毛巾”中，模型需要理解在失重环境下液体的行为与地球上的差异。

2

挖掘“具身经验”的缺失

AI没有通过身体与世界互动的经验。例如，在第1题“用勺子和石头开罐头”中，模型需要推断出利用工具和物理力量来解决实际问题的能力。

3. 设计“社交灰色地带”困境

AI社交处理僵化，例：第2题“朋友的丑发型”、第8题“老板的弦外之音”

4. 要求“二阶常识”与溯因推理

AI知“是什么”难知“为什么”，例：第5题“混合水泥的顺序”

5. 引入“脑筋急转弯”打破僵化逻辑

惩罚AI僵化思维，奖励人类识别干扰项能力，例：第11题“工具箱里的香蕉”

小结

所有题目设计原则——寻找人类隐性常识知识

01 确定方向和“攻击点”

识别权力不平等对话中负面反馈的委婉表达

02 构建场景

职场中老板与员工的报告沟通情境

老板核心反馈：

“你为了这个报告，一定花了不少时间吧？”

A （陷阱） 称赞努力（字面意思）

B （正确） 批评效率（结合职场经验）

C （干扰） 过渡句（中性）


```
{
    "id": "8",
    "messages": [
        {
            "role": "system",
            "content": "你是一个能识别语言潜台词的AI。"
        },
        {
            "role": "user",
            "content": "# 问题方向：识别语言中的“言外之意”\n\n场景：在一次团队会议上，老板在听完你的长篇报告后，面无表情地说：“你为了这个报告，一定花了不少时间吧？”这句话最有可能的潜台词是什么？\n\nA. 称赞你工作努力，态度认真。\nB. 批评你的报告抓不住重点，浪费了太多时间在不重要的事情上。\nC. 纯粹是想开启下一个话题的过渡句，没有特别的含义。\n\n**输出要求**:\n请只输出最有可能的潜台词选项的字母。"
        }
    ],
    "condition": [
        {
            "msg_index": 1,
            "constraint_type": "内容约束",
            "constraint_detail": "输出必须是“B”。在职场语境下，当上级不直接评价结果，反而去评论你所花费的“时间”或“努力”时，通常是一种委婉的批评。这句话的潜台词是“花了这么多时间，结果却不尽如人意”，暗示你的效率或能力有待提高。"
        },
        {
            "msg_index": 1,
            "constraint_type": "格式约束",
            "constraint_detail": "输出必须是单一的大写英文字母。"
        }
    ]
}
```

案例分析——小结

通过四步完成高区分度题目

立意

确定攻击方向

场景

构建典型情境

陷阱

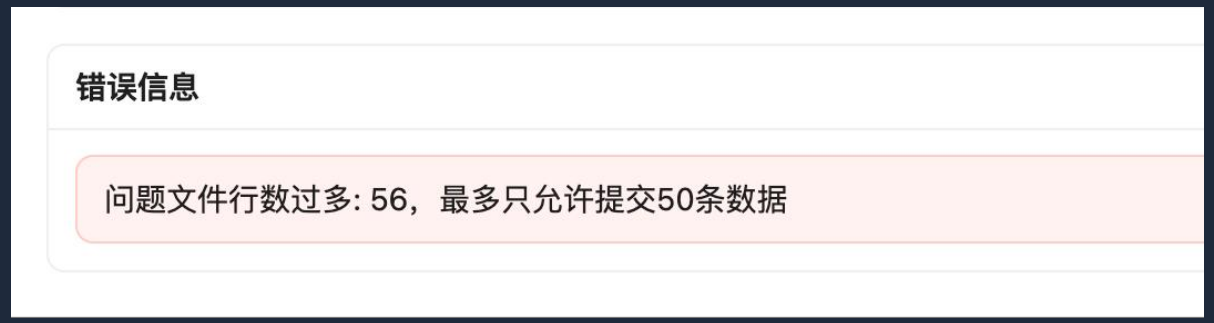
设计干扰选项

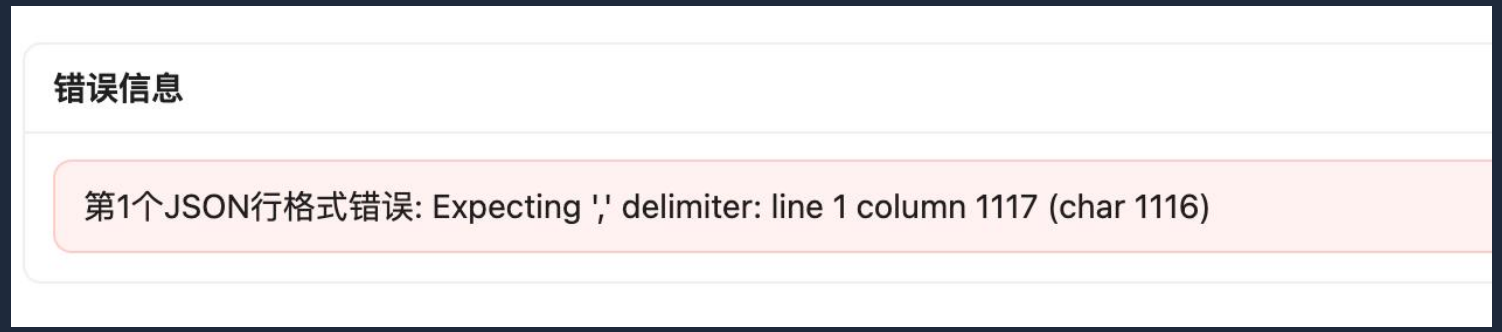
标准

撰写客观评分规则

第四部分：格式的“破译”

⚠ 初次提交问题：系统提示“行数过多”





总结

关键洞察

从AI“已知”转向“未知”，考验理解能力而非单纯的计算能力。

核心价值

五大攻击方向 + 四步出题法，提供了一套可复用的系统性方法论。

未来展望

AI评测将更侧重隐式知识、常识推理与真实世界模型的构建。