课程作业：逻辑谜题推理

自然语言处理2025

助教　陈宇飞

2025年5月6日

背景

自 OpenAI 推出 GPT 系列以来，研究者们愈发关注 Large Language Model（LLM）能否通过仅在自然语言上的学习就能实现真正的智能 。

其中，逻辑推理能力作为智能的核心能力，被广泛视为衡量模型理解与思考水平的关键指标，几乎在所有模型评测中均出现并占据重要位置，特别是随着近期推理模型的出现更受关注。

尽管语言模型的结构并非适合逻辑推理任务，其信息筛选、多步推理、一致性保持等方面仍存在挑战，但近年来该方向仍展现出令人振奋的发展态势。

包括 DeepSeek R1、GPT-o3 和 Claude 3.7 等在内的前沿模型，不仅在多个逻辑推理基准测试中实现突破性进展，其在用户的实际使用中也得到广泛认可。

本次课程作业将围绕“LLM + 逻辑推理”展开。

数据集类型

本次课程作业的数据集类型是谜题（Puzzle）。

在推理任务中，与常见的数学题或代码题相比，谜题通常有如下的特点：

1. 结构简洁：问题形式标准
2. 推理导向：聚集逻辑推理能力，强调因果关系
3. 可验证性强：具有清晰可验证的解决方案
4. 可扩展复杂性：谜题通常由程序生成，难度可调，便于逐步研究
5. 有趣味性

数据集

Temporal Clue：灵感来源于经典谋杀推理游戏 Clue （Cluedo），玩家在游戏中需要找出是谁在他富丽堂皇的庄园中杀死了博迪先生。

Temporal Clue 将这一游戏设定转化为形式化的逻辑谜题，扩展了传统游戏中的“谁是凶手、用什么凶器、在哪杀死” ，进一步引入“时间”和“动机”两个维度，构成五元组的完整推理任务。

所有谜题由程序随机生成，并通过 OR-Tools 的 CP-SAT 求解器选择最小但足够的线索。

由于作业涉及使用LLM API，考虑成本和速率的限制，将原先的数据集按难度均匀随机抽样200个问题，共四种难度，每种50个问题，并保存在data/tc_200_en.json。

数据集

由于 tc_200_en 具有如下特点：

1. 英文形式，可能不容易观察、理解和分析模型的推理回答

2. 由程序生成，语言机械，缺乏自然语言的灵活性，无法观察模型在真实语言环境中的泛化与鲁棒性。

进一步设计了多风格中文数据集：将每种难度级别的问题，均匀地转换为五种不同的中文风格：小说、新闻、审讯记录、广播剧脚本、档案报告。每种风格下，每个难度包含10个问题，保存在data/tc_200_zh.json 。

本次作业的所有实验应在 data/tc_200_zh.json 上进行，而 data/tc_200_en.json 仅用于参考或者在实验中可能用作辅助测试。

数据集

tc_200_zh.json 中的每条属性包括：

1. prompt：问题
2. solution：标准答案
3. num_clues：线索数量
4. evaluation：问题难度评定

示例代码中的评测方法为所有问题都正确才正确。

Deepseek V3 参考测试结果：

zh acc：11%，en acc：22%


```
{
  "prompt": "冬夜深沉，雪花无声地落在都铎庄园的尖顶上。富可敌国却行踪诡秘的约翰·Q·博迪先生，正在宅邸里举办一场小型而奢华的晚宴，邀请了几位最亲密的伙伴。然而凌晨时分，博迪先生被发现陈尸于宅中某处。此刻，壁炉里的余烬忽明忽暗，照出两位嫌疑人的轮廓：\n\n• 芥末上校\n• 桃子小姐\n\n现场遗留的凶器包括：\n\n• 扳手\n• 毒药\n\n命案可能发生的房间只有两处：\n\n1. 画室\n2. 图书室\n\n宅邸布局如下：\n\n 北 北  \n西 1|2 东\n 南 南\n\n每位嫌疑人都怀揣着独特的杀人动机：\n\n• 恐惧\n• 仇恨\n\n要完成谋杀，凶手必须与博迪先生独处一室，且现场至少存在一件凶器。\n\n关键线索在煤油灯下若隐若现：\n\n- 博迪先生生前所在的房间，位于扳手所在房间的正东方\n- 桃子小姐的动机是恐惧\n- 心怀仇恨的嫌疑人当时在画室\n\n请回答以下问题：\n\nA. 真凶是谁？\nB. 凶器为何？\nC. 命案发生在何处？\nD. 杀人动机是什么？\n\n以及附加问题：\n\nE. 桃子小姐当时身在何处？\nF. 毒药最初存放在哪个房间？\n\n请按以下格式提交答案：\n\nA. 嫌疑人\nB. 凶器\nC. 房间\nD. 动机\nE. 房间\nF. 房间\n\n祝您好运，侦探。煤油灯快要熄灭了.....",
  "solution": {
    "A": "桃子小姐",
    "B": "毒药",
    "C": "图书室",
    "D": "恐惧",
    "E": "图书室",
    "F": "图书室"
  },
  "num_clues": 3,
  "evaluation": {
    "difficulty_score": 0.11,
    "difficulty_level": "Easy",
    "metrics": {
      "clue_complexity": 0.2,
      "temporal_complexity": 0.0,
      "spatial_complexity": 0.0,
      "motive_complexity": 0.2
    },
    "statistics": {
      "num_clues": 3,
      "num_suspects": 2,
      "num_weapons": 2,
      "num_rooms": 2,
      "num_times": 0,
      "num_motives": 2,
      "has_unique_motives": true
    },
    "style": "novel"
  }
},
```

数据集例子：Prompt

冬夜深沉，雪花无声地落在都铎庄园的尖顶上。富可敌国却行踪诡秘的约翰·Q·博迪先生，正在宅邸里举办一场小型而奢华的晚宴，邀请了几位最亲密的伙伴。然而凌晨时分，博迪先生被发现陈尸于宅中某处。此刻，壁炉里的余烬忽明忽暗，照出两位嫌疑人的轮廓：

- 芥末上校
- 桃子小姐

现场遗留的凶器包括：

- 扳手
- 毒药

命案可能发生的房间只有两处：

- 画室
- 图书室

宅邸布局如下：

北 北
西 1|2 东
南 南

画室	图书馆

数据集例子：Prompt

每位嫌疑人都怀揣着独特的杀人动机：

- 恐惧
- 仇恨

要完成谋杀，凶手必须与博迪先生独处一室，且现场至少存在一件凶器。

关键线索在煤油灯下若隐若现：

- 博迪先生生前所在的房间，位于扳手所在房间的正东方
- 桃子小姐的动机是恐惧
- 心怀仇恨的嫌疑人当时在画室

画室	图书馆

请回答以下问题：

A.真凶是谁?

B.凶器为何?

C.命案发生在何处?

D.杀人动机是什么?

以及附加问题：

E.桃子小姐当时身在何处?

F.毒药最初存放在哪个房间?

请按以下格式提交答案：A. 嫌疑人\nB. 凶器\nC. 房间\nD. 动机\nE. 房间\nF. 房间\n

祝您好运，侦探。煤油灯快要熄灭了……

数据集例子：Prompt

每位嫌疑人都怀揣着独特的杀人动机：

- 恐惧
- 仇恨

要完成谋杀，凶手必须与博迪先生独处一室，且现场至少存在一件凶器。

关键线索在煤油灯下若隐若现：

- 博迪先生生前所在的房间，位于扳手所在房间的正东方
- 桃子小姐的动机是恐惧
- 心怀仇恨的嫌疑人当时在画室

画室 扳手	图书馆 博迪先生

请回答以下问题：

A.真凶是谁?

B.凶器为何?

C.命案发生在何处?

D.杀人动机是什么?

以及附加问题：

E.桃子小姐当时身在何处?

F.毒药最初存放在哪个房间?

请按以下格式提交答案：A. 嫌疑人\nB. 凶器\nC. 房间\nD. 动机\nE. 房间\nF. 房间\n

祝您好运，侦探。煤油灯快要熄灭了……

数据集例子：Prompt

每位嫌疑人都怀揣着独特的杀人动机：

- 恐惧
- 仇恨

要完成谋杀，凶手必须与博迪先生独处一室，且现场至少存在一件凶器。

关键线索在煤油灯下若隐若现：

- 博迪先生生前所在的房间，位于扳手所在房间的正东方
- 桃子小姐的动机是恐惧
- 心怀仇恨的嫌疑人当时在画室

画室 扳手 芥末上校 （仇恨）	图书馆 博迪先生 桃子小姐 （恐惧）

请回答以下问题：

A.真凶是谁?

B.凶器为何?

C.命案发生在何处?

D.杀人动机是什么?

以及附加问题：

E.桃子小姐当时身在何处?

F.毒药最初存放在哪个房间?

请按以下格式提交答案：A. 嫌疑人\nB. 凶器\nC. 房间\nD. 动机\nE. 房间\nF. 房间\n

祝您好运，侦探。煤油灯快要熄灭了……

数据集例子：Solution

问题：
A.真凶是谁？ B.凶器为何？
C.命案发生在何处？ D.杀人动机是什么？
E.桃子小姐当时身在何处？
F.毒药最初存放在哪个房间？

画室 扳手 芥末上校 （仇恨）	图书馆 博迪先生 桃子小姐 （恐惧） 毒药

根据线索间的逻辑关系，我们可以得出以下结论：

芥末上校因仇恨滞留画室时，扳手作为工具也存放在同一位置。由于博迪所在房间位于扳手的正东方，他必然身处东侧的图书室。

桃子小姐的恐惧动机促使她必须出现在被害人所在处。当图书室成为唯一符合方位条件的案发现场时，留在该空间的毒药自然成为作案工具。

毒药的原始存放位置无需移动即可完成犯罪，说明其本就放置在实施犯罪的图书室。这种空间与凶器的对应关系，恰好解释了为何持有恐惧动机者能在特定地点完成投毒。

数据集例子：Solution

数据集中保存的答案如下，顺序和问题一一对应：

```
"solution": {
        "A": "桃子小姐",
        "B": "毒药",
        "C": "图书室",
        "D": "恐惧",
        "E": "图书室",
        "F": "图书室"
    },
```

数据集例子：Evaluation

如下复杂度分数的加权求和

1. 线索复杂度（35%）： 0.5×0.4 + 0×0.3 + 0×0.3=0.2
    1. 线索数量与总元素数量（嫌疑人+武器+房间）的比值（40%）： 3÷ (2+2+2)=0.5
    2. 间接逻辑线索（含"if and only if"）占比（30%）
    3. 包含时间信息的线索占比（30%）
2. 时间复杂度（25%）：时间点数量减1后除以5，最高取1.0。
3. 空间复杂度（20%）：房间数量减2后除以14，最高取1.0。
4. 动机复杂度（20%）：动机数量乘以系数（唯一动机乘0.1，非唯一即多人共用乘0.15），最高取1.0。

示例难度得分为 0.2×35% + 0 + 0 + 2×0.1×20%=0.11

数据集例子：Evaluation

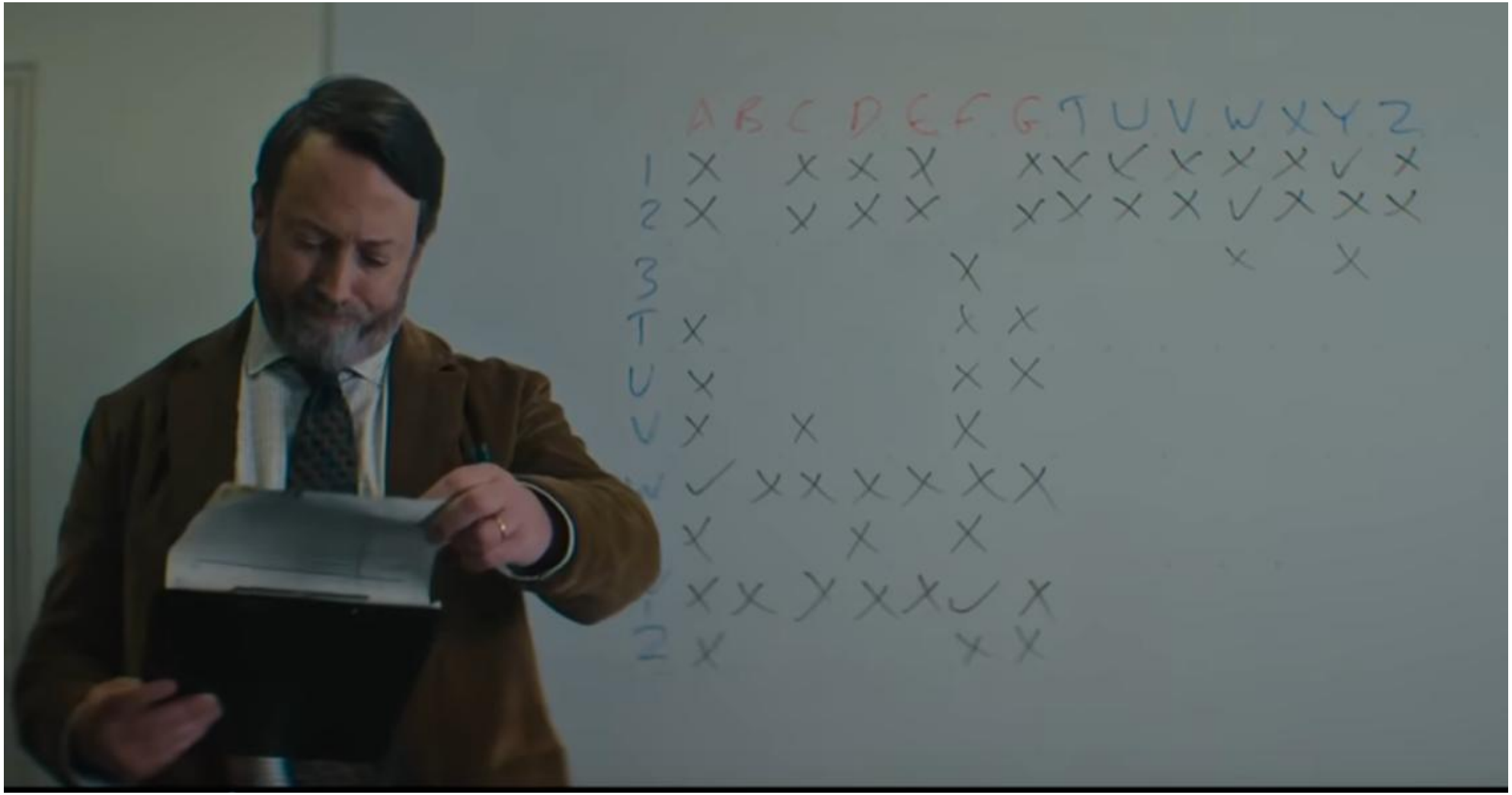
难度评级：

1. Easy（<0.3）
2. Medium（0.3-0.6）
3. Hard（0.6-0.8）
4. Expert（≥0.8）

示例难度为 Easy


```
"evaluation": {
    "difficulty_score": 0.11,
    "difficulty_level": "Easy",
    "metrics": {
      "clue_complexity": 0.2,
      "temporal_complexity": 0.0,
      "spatial_complexity": 0.0,
      "motive_complexity": 0.2
    },
    "statistics": {
      "num_clues": 3,
      "num_suspects": 2,
      "num_weapons": 2,
      "num_rooms": 2,
      "num_times": 0,
      "num_motives": 2,
      "has_unique_motives": true
    },
    "style": "novel"
  }
```

谜探路德维希



任务一

任务一：调用 API 生成

选择一个 LLM API 对数据集中的所有问题生成回答并评测结果，挑选一个 case 观察模型的推理是否准确，在哪个步骤如何出错，如何修正。

对模型不做限定，但模型本身的推理能力不要太弱。如果能使用不同模型的API，可以比较不同模型的效果，例如比较不同系列模型、同系列不同大小模型、通用模型和推理模型。考虑到 API 限制，这并非强制要求。

已提供了一个异步处理的示例代码，可能需要根据具体使用的API做调整，也可以自己写调用 API 的代码。可以自行设计其他合理的评测指标用于分析，例如对不同难度问题的准确率分别报告，或者计算排除 bonus 问题的准确率，但原本评测指标的结果仍需要报告。

免费 API 参考网站（可能限时免费或者不再免费）：OpenRouter，Chutes，Cloudflare Workers AI，Free-QWQ，字节扣子，智谱AI 等等。

任务二

任务二：Prompt Engineer

探索和对比不同 Prompt 策略的效果，需要设计不同的Prompt模板，来提升模型在逻辑推理中的表现。

尝试多种 Prompt 策略，可以包括角色扮演、引入与目标问题不同的示例、强化或者引导模型生成推理链、基于模型初始回复进行二次提问。

例子：“你是福尔摩斯，接到一个案子：{Q}”。

如果探索的 Prompt 策略比较多，可以选择部分数据来实验，最后在最好的策略上测试全部数据。

如有参考资料或相关工作，需在实验报告中注明引用。

任务三

任务三：工具使用（Tools）

编写代码，使 LLM 能结合外部工具完成任务，并在数据集的所有问题上评测结果。

由于谜题是通过程序生成的，所以可以通过 LLM 提取相关信息，使用一个逻辑判断程序来获得结果。也鼓励使用其他形式的工具（程序、小型模型、网页接口等等）。

可参考 Temporal Clue 中的谜题生成方式，反向设计工具调用流程。

可使用 MCP 协议实现工具接口，但并非强制要求。

任务四

任务四：多智能体对话（Agents）

编写代码，模拟现实场景中的多智能体交流，让多个 LLM 以不同角色协同解决问题。

可选场景包括侦探破案、法庭辩论、苏格拉底式提问等。

仅需选择数据集中的一个合适的问题即可，通过设定不同角色（如侦探、证人、法官等）实现多轮互动。

设计目标可侧重于高效性或者趣味性，并以此作为任务四的作业评分标准：要么多智能体协作能提升推理效率与准确性；要么使多轮对话更具故事性，表达更加生动，如剧本或小说形式，可以自行添加或者让模型生成更加具体的信息来帮助展现一个完整的故事。

可使用 A2A 协议进行多智能体管理，但并非强制要求。

作业要求

DDL：5.20

提交格式要求：压缩包“学号_姓名. zip” ，内含代码、实验报告、使用 Prompt 模板、实验中涉及的完整模型生成结果，报告命名为“学号_姓名.pdf”。

实验报告需包含以下内容：所有任务的设计思路、测试效果和思考；Prompt 策略涉及的参考资料；遇到的具体问题，如何解决等。

主要依据任务完成度、报告质量、不同方法探索程度、基于实验结果的分析和思考质量来综合评分。

鼓励在实验中对比不同难度和不同语言风格上的表现。

严禁结果造假和代码抄袭。

Thanks!