



合肥工业大学

# 大语言模型解答高中理科题目 性能评估

比赛题号: CG2401

队伍名称: 轻轻茉莉

报名编号: AFS0039

参赛队员: 游琥、杨照、王金金

# 大语言模型解答高中理科题目性能评估

**摘要：**在本研究中，我们对市面上流行的 12 个大型语言模型进行了系统评测，不仅对理科试卷进行了深入分析，还对文科试卷进行了全面评估。通过横向对比（不同厂商模型之间的比较）和纵向对比（同一厂商不同参数模型之间的比较），我们考察了各模型在理科和文科题目上的推理时间、不同提示词（prompt）对模型表现的影响、答题准确率、多解创造性和逻辑推理能力等多方面的表现。此外，我们创新性地提出了对模型回答完整性的评估方法。

本研究涵盖了 2010 年至 2024 年间的所有理科和文科试卷，结合多年的数据进行跨学科对比分析，揭示了模型在理科与文科不同学科领域的适应性。结果显示，闭源模型的推理时间在 API 请求时易受 API 服务商网络延迟的影响，但在同一模型的跨学科评估中仍具有参考意义。同时，我们为了防止数据泄露，将数据分为 2 组，从 2010 年到 2022 年高考试卷为一组，从 2023 年到 2024 年为第二组。对 2023 年和 2024 年的试卷进行了独立分析，进一步验证各模型的推理能力。

## 目录

1. 题目介绍 .....	1
1.1 题目 .....	1
1.2 要求 .....	1
1.3 目的 .....	1
2. 队伍介绍 .....	1
2.1 队伍名称 .....	1
2.2 报名信息 .....	1
2.3 学校名称 .....	1
2.4 队员信息 .....	1
3. 评测模型选择 .....	2
3.1 模型 .....	2
3.2 选择理由 .....	2
3.2.1 横向比较 .....	2
3.2.2 纵向比较 .....	3
3.3 总结 .....	4
4. 评测指标 .....	4
4.1 总体评测 .....	4
4.2 具体指标 .....	4
5. 数据集构建 .....	6
6. 评估结果 .....	7
6.1 各大模型在各科的准确率表现 .....	7
6.2 完整性、正确性、推理能力和创新性评估 .....	9
6.2.1 整体分析 .....	9
6.2.2 完整性，创新性分析 .....	10
6.2.3 推理能力分析 .....	11
6.3 模型表现水平 .....	12
6.4 推理时间 .....	13
6.5 不同 prompt 下指标表现 .....	14
6.5.1 简单的 prompt .....	14
6.5.2 精确格式化的 prompt .....	14
7. 结果分析 .....	15

7.1 关于模型一题多解创新性相关分析 .....	15
7.2 关于不同 prompt 下指标表现的分析 .....	16
8. 创新点 .....	20
8.1 对文科的题目推理 .....	20
8.2 数据泄露的分析 .....	21
8.3 对解答完整性的评估 .....	22
9. 总结与展望 .....	24

# 1. 题目介绍

## 1.1 题目

CG2401 大语言模型解答高中理科题目性能评估

## 1.2 要求

选手自主选择数学、物理、化学、生物各学科的高中题目，通过调用现有的大语言模型对题目进行解答，整理评测结果，并对结果进行分析，最终形成一份全面合理的评估报告。

## 1.3 目的

通过这些测试，我们期望深入了解大语言模型在逻辑推理和问题解答方面的优势和不足，为其未来发展提供参考。

# 2. 队伍介绍

## 2.1 队伍名称

轻轻茉莉

## 2.2 报名信息

报名渠道：赛氪

报名编号：AFS0039

## 2.3 学校名称

合肥工业大学

## 2.4 队员信息

我们的队伍成员是来自计算机与信息学院，由计算机科学与技术专业和信息与通信工程专业学子联合构成，对各方面的知识和信息都有所涉猎，因此团队也能更好地完成本次赛题。

队长：游琥

队员 1：杨照

队员 2：王金金

具体报名信息如图 2-1 所示：

姓名	来自学校	院系	手机号	邮箱
游斌	合肥工业大学	计算机学院	18280170872	1280018408@qq.com
杨照	合肥工业大学	计算机学院	15923195288	1012597462@qq.com
王基金	合肥工业大学	计算机与信息学院	13225858906	2839761781@qq.com

图 2-1 具体报名信息

## 3. 评测模型选择

### 3.1 模型

- (1) Doubao-pro-4k
- (2) Glm-4-Flash
- (3) GPT-3.5-Turbo
- (4) GPT-4o-mini
- (5) Hunyuan-lite
- (6) Llama-3-1-8B-Instruct
- (7) Mistral-7B-Instruct
- (8) Gemini-Pro
- (9) DeepSeek-V2.5
- (10) Qwen2-72B-Instruct-AWQ
- (11) Qwen2.5-7B-Instruct
- (12) Qwen2-7B-Instruct

### 3.2 选择理由

通过横向（同一厂家不同版本）和纵向（不同厂家模型）的模型对比，可以更全面地理解每个模型的表现。

#### 3.2.1 横向比较

横向比较是对 Qwen 系列不同版本进行比较。

- (1) Qwen2-72B-Instruct-AWQ

特点：这是一个大规模的模型，具备出色的推理和生成能力，能够处理复杂的逻辑问题。

适用场景：适合需要深入分析和多层次逻辑推理的高考题。

### (2) Qwen2.5-7B-Instruct

特点：该模型较小，但仍然保留了良好的逻辑推理能力，特别是在处理指令时表现突出。

适用场景：更适用于相对简单的逻辑题，资源有限时使用。

### (3) Qwen2-7B-Instruct

特点：与 Qwen2.5 相似，专注于快速响应和执行指令，具备一定的逻辑推理能力。

适用场景：适合快速反应的测试场景，可以评估其在逻辑题中的表现。

## 3.2.2 纵向比较

纵向比较是对不同厂家的模型进行比较。

### (1) Doubao-pro-4k

特点：针对逻辑推理进行了优化，处理复杂文本的能力很强。

选择理由：在高考逻辑题中表现优秀，适合细致分析。

### (2) Glm-4-Flash

特点：具备高效的推理能力，能够快速生成答案。

选择理由：适合应对时间压力大的考试环境。

### (3) Gpt-3.5-Turbo

特点：这是 OpenAI 的强大模型，逻辑推理能力很强。

选择理由：广泛应用于各种自然语言处理任务，可以提供高质量的答案。

### (4) GPT-4o-mini

特点：这是一个小型化的高效模型，减少了资源占用。

选择理由：适合需要逻辑推理但资源有限的应用场景。

### (5) Hunyuan-lite

特点：在中文处理上表现优异，优化了逻辑推理能力。

选择理由：专为中文环境设计，适合中文高考题的评估。

### (6) Llama-3-1-8B-Instruct

特点：该模型在处理复杂问题时表现出色，适应性强。

选择理由：可以处理多样化的逻辑题，适合高考题的评估。

### (7) Mistral-7B-Instruct

特点：强调指令理解，适合快速生成逻辑解答。

选择理由：适合快速反应和逻辑推理的场景。

#### (8) Gemini-Pro

特点：这是 Google 的高端大语言模型之一，具有强大的多步逻辑推理能力和理解复杂概念的优势。特别是在需要抽象思维和结构化推理的理科任务中表现出色。

选择理由：适合评估模型在多步骤逻辑题中的解答能力，尤其是在高考物理、数学等学科题目中，有助于深度分析复杂题目的解答过程。

#### (9) DeepSeek-V2.5

特点：拥有创新架构，擅长处理科学领域的内容和复杂知识点推理，尤其适合需要精确理解和深度分析的理科问题。

选择理由：在涉及专业知识和多知识点关联的题目中表现良好，适合评估模型的知识整合能力和多层次逻辑推理，能够反映在高考科学题中对专业知识的理解水平。

### 3.3 总结

通过横向比较 Qwen 系列的不同版本，我们可以观察到，随着模型规模的变化，推理能力和生成质量也会有所不同。而在纵向比较不同厂家模型时，能够展现出各自的设计哲学和技术实现的多样性。

在选择这些模型进行高考逻辑能力测评时，可以根据具体的应用场景、资源限制和期望的逻辑推理深度来决定最合适的模型。这种评估方法将有助于全面理解每个模型在处理理科高考题时的表现和优势。

## 4. 评测指标

### 4.1 总体评测

在评估不同大模型在高中数学、物理、化学、生物学科题目表现时，选取了推理速度、不同 prompt 下的表现、正确率、得分百分比、模型水平、推理结果完整性与多样性，以及拆解步骤下逻辑推理的得分等指标，以全面分析模型在各学科任务中的能力和适应性。

### 4.2 具体指标



### (1) 正确率

正确率是模型在测试题目中获得正确答案的比例。这是最直接的性能衡量标准，反映模型在回答题目时的准确性，确保其在学科知识上的基础可靠性。它直观地展现了模型对于高中数学、物理、化学、生物等学科知识的掌握程度。如果一个模型在某学科的大量测试题目中正确率较高，说明其在该学科领域内已经学习到了较为准确和丰富的知识，能够依据这些知识对各类题目做出正确的判断和解答。

### (2) 不同 prompt 下指标表现

不同 prompt 表现指模型在不同提示词 (prompt) 引导下的输出质量。由于不同的提示词可能会引导模型产生不同的回答，评估模型在各种 prompt 下的稳定性和表现有助于了解其应对复杂语言输入的灵活性和一致性。

### (3) 完整性

推理结果完整性是指模型的回答是否包括了必要的步骤和细节。完整性评估能确保模型输出的答案是清晰且有条理的，尤其在学科问题中，完整的推理步骤往往比简单的正确答案更重要，有助于学生理解。

### (4) 创新性

推理结果创新性是指模型是否可以提供多种不同的解答思路或答案视角。多样性反映了模型在处理问题上的灵活性，能够展示不同的解题思路有助于全面评估模型的智能程度，也对学习者提供更多选择。

### (5) 逻辑性

该指标是指模型在分步骤回答题目时，每一步的逻辑和准确性得分。通过细化评估模型的逻辑推理能力，可以发现模型在复杂问题拆解过程中的优势和不足，对提升模型的解释性和可靠性具有指导意义。

### (6) 模型能达到的水平

该指标评估模型在不同学科中的综合表现水平，如是否能达到高中生、大学生甚至专家水平。了解模型的表现水平有助于识别其在实际应用中的适用场景，便于用户决定是否能将其用于具体学科教育中的辅导或辅助。

### (7) 推理速度

推理速度是指模型在回答题目时的响应时间。在实际应用中，模型的响应速度直接影响用户体验和应用效率，尤其在大规模应用场景下，速度成为衡量模型是否实用的关键指标。

## 5. 数据集构建

### (1) 提取并替换图片路径

使用自动化脚本提取 Word 文档中的所有图片，并将其保存到当前目录下的 image 文件夹中。提取完成后，脚本会在文档相应位置将图片替换为其保存路径，以便后续处理与标识。这样，每张图片在文档内容中均对应具体的文件路径，便于管理和引用。

### (2) 将公式转化为 LaTeX 格式

借助 MathType 工具对 Word 文档中的公式进行转换，统一为 LaTeX 格式，以提升公式的格式一致性与兼容性。此步骤确保公式在后续处理阶段保持良好的可读性和移植性。

### (3) 文档内容人工整理

对文档内容进行人工规整和优化，确保信息以逻辑清晰、结构紧凑的方式排列，便于后续录入处理。该步骤不仅可以去除冗余信息，还能提升内容的可读性，为后续数据结构转换和输入提供便利。数据集构建工作步骤如图 5-1 所示：

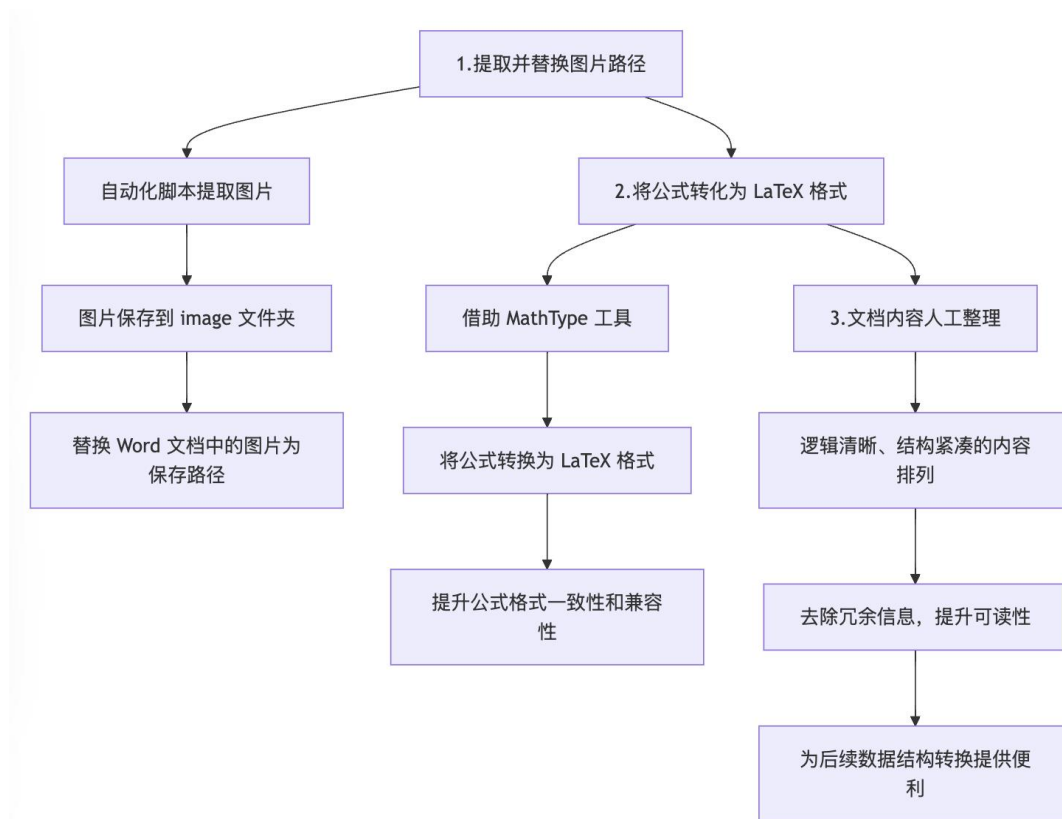


图 5-1 数据集构建工作步骤

## 6. 评估结果

### 6.1 各大模型在各科的准确率表现

实验结果热图如图 6-1 所示:

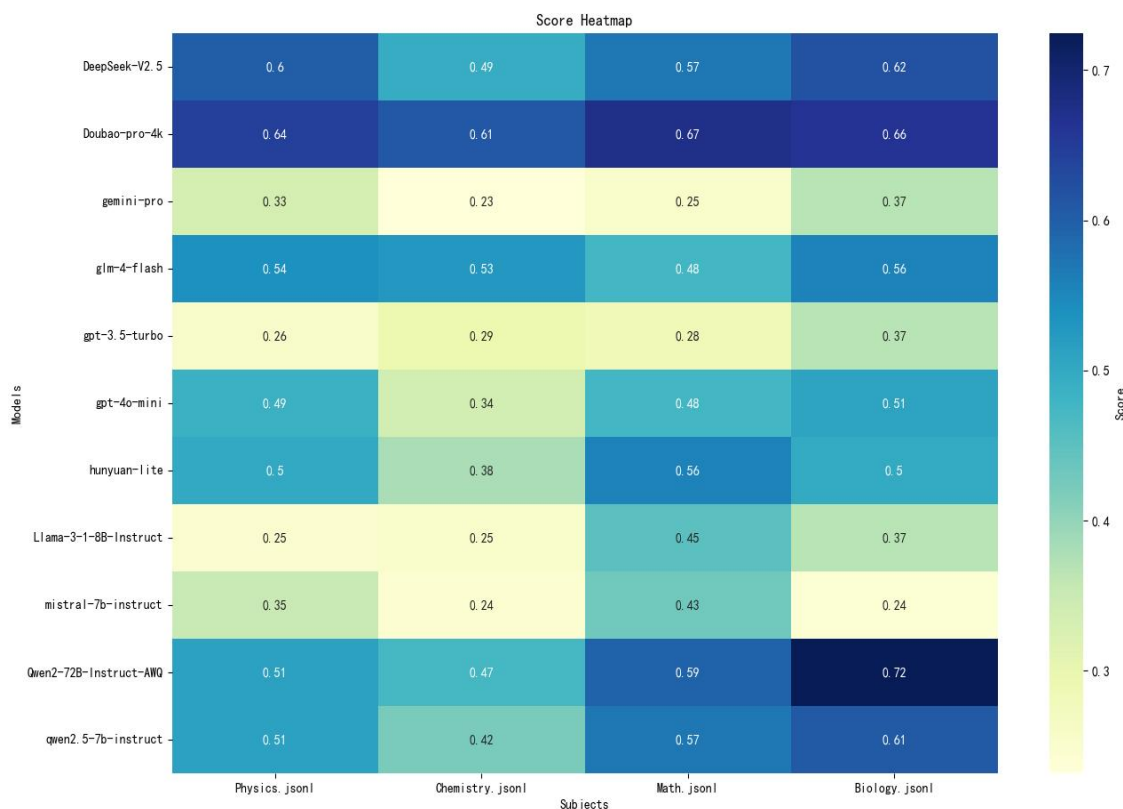


图 6-1 实验结果热图

图 6-1 展示了不同模型在物理、化学、数学和生物四个科目中的得分情况。颜色从浅绿色至深蓝色逐渐变化，其中浅绿色表示较低的正确率，深蓝色则表示较高的正确率。在 2023 至 2024 年间的全国高考模拟测试中，Doubao-pro-4k 模型表现出卓越的综合能力，Qwen2-72B-Instruct-AWQ 和 DeepSeek-V2.5 的表现也相当出色，尤其是 Qwen2-72B-Instruct-AWQ 在生物学科上获得了 0.72 的高正确率。

相比之下，GPT 系列模型在该任务中并未展现出明显的优势，在各科目中的得分均处于一般水平；不过，GPT4o-mini 相比 GPT3.5 在各科目上表现显著提升。

小型模型在该任务中的整体表现较弱，但值得注意的是，Qwen2.5-7B-Instruct 模型的平均正确率超过了 0.5，在众多小型模型中表现突出。

实验结果雷达分析图如图 6-2 所示：

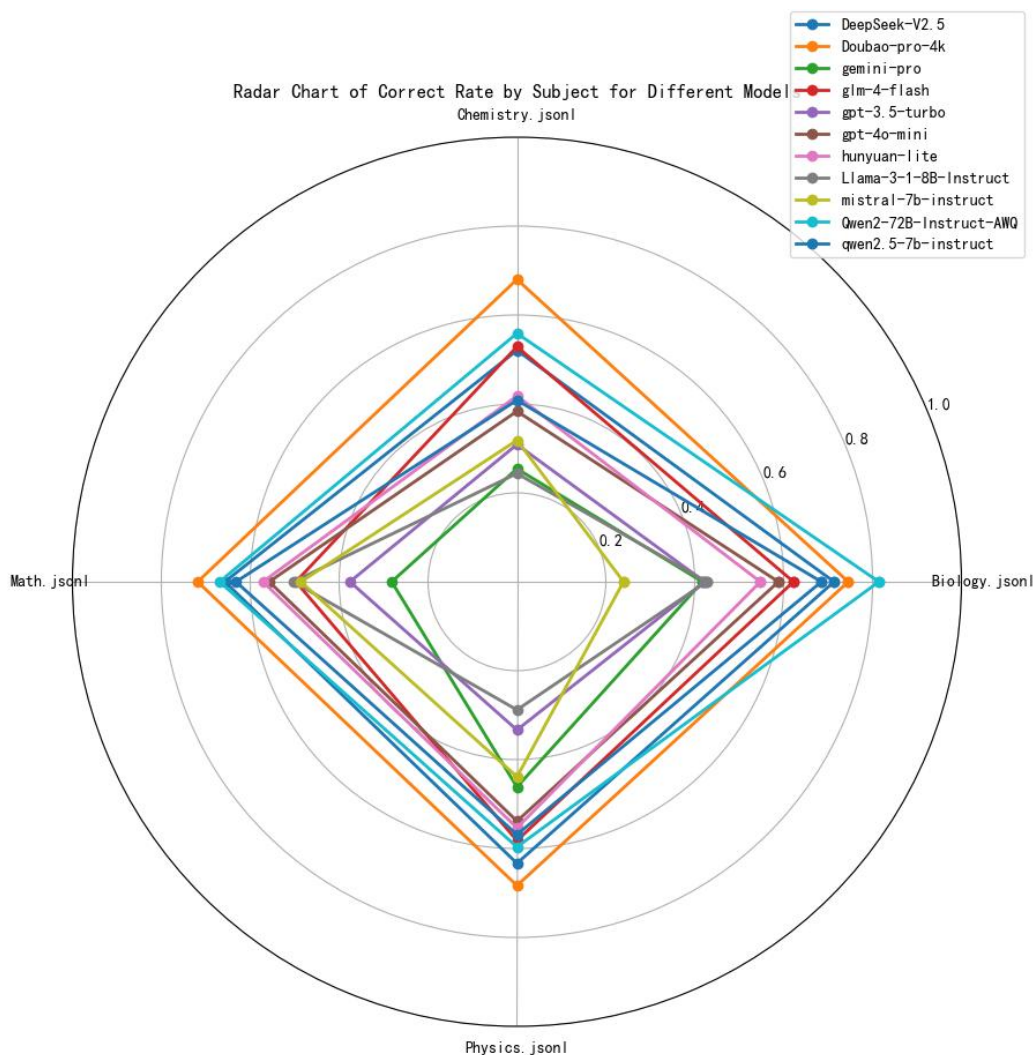


图 6-2 雷达分析图

图 6-2 展示了不同模型在化学、数学、生物和物理四个学科上的正确率情况。每种颜色代表一种模型的表现。从整体来看，各大模型在生物学科上的表现尤为突出，第一梯队的模型准确率超过 0.6，尤其是 Qwen2-72B-Instruct-AWQ 的表现最为亮眼。数学学科次之，其中表现最佳的是 Doubao-pro-4k。相较之下，各模型在化学学科上的表现则相对逊色。值得注意的是，Doubao-pro-4k 在各个学科中几乎都取得了领先的准确率，彰显了其卓越的综合性能。

## 6.2 完整性、正确性、推理能力和创新性评估

### 6.2.1 整体分析

模型整体分析如图 6-3 所示：

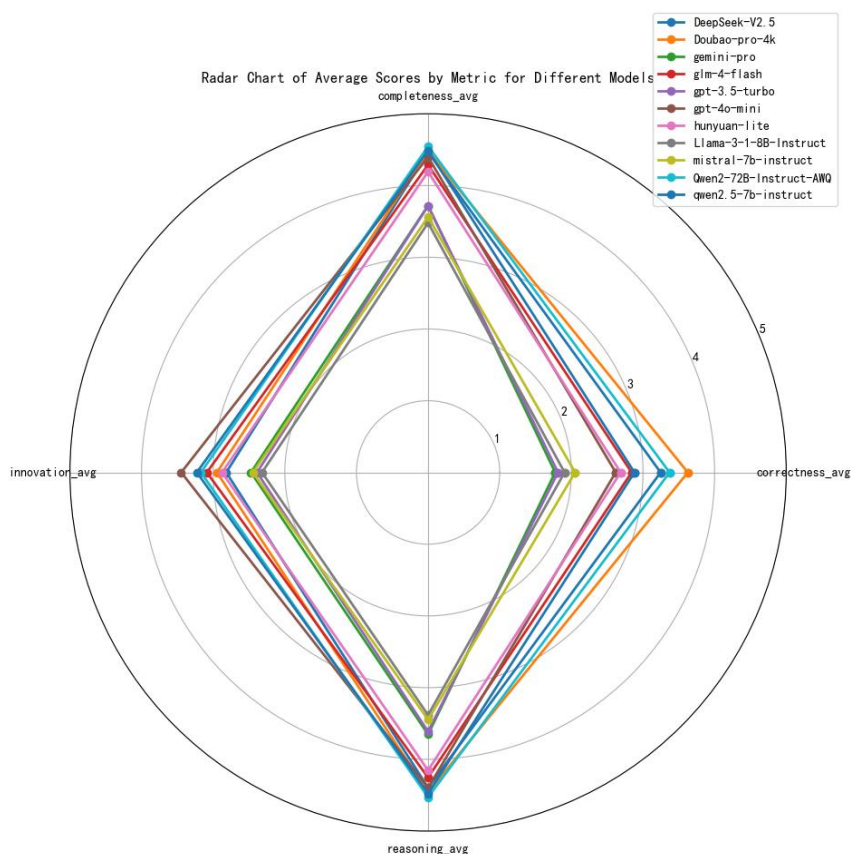


图 6-3 模型整体分析图

整体来看，各大模型在各个维度上的表现较为突出，尤其是在完整性和推理能力两个维度上，除 Mistral-7B-Instruct 和 Llama-3-1-8B-Instruct 外，大多数模型的表现接近且得分较高，表明在这些指标上模型之间具有较强的一致性。这可能得益于训练数据的充足性以及模型对回答准确性的优化，使得其在这两个维度上保持了较高的稳定性。

然而，在创新性方面，各模型的得分差异更加显著，部分模型在创新性方面表现不佳，难以实现一题多解。在此维度上，GPT-4o-mini 最佳，但考虑到其相对较低的正确率得分，这种创新能力的实际有效性仍需进一步探讨。

### 6.2.2 完整性，创新性分析

模型整体分析如图 6-4 所示：

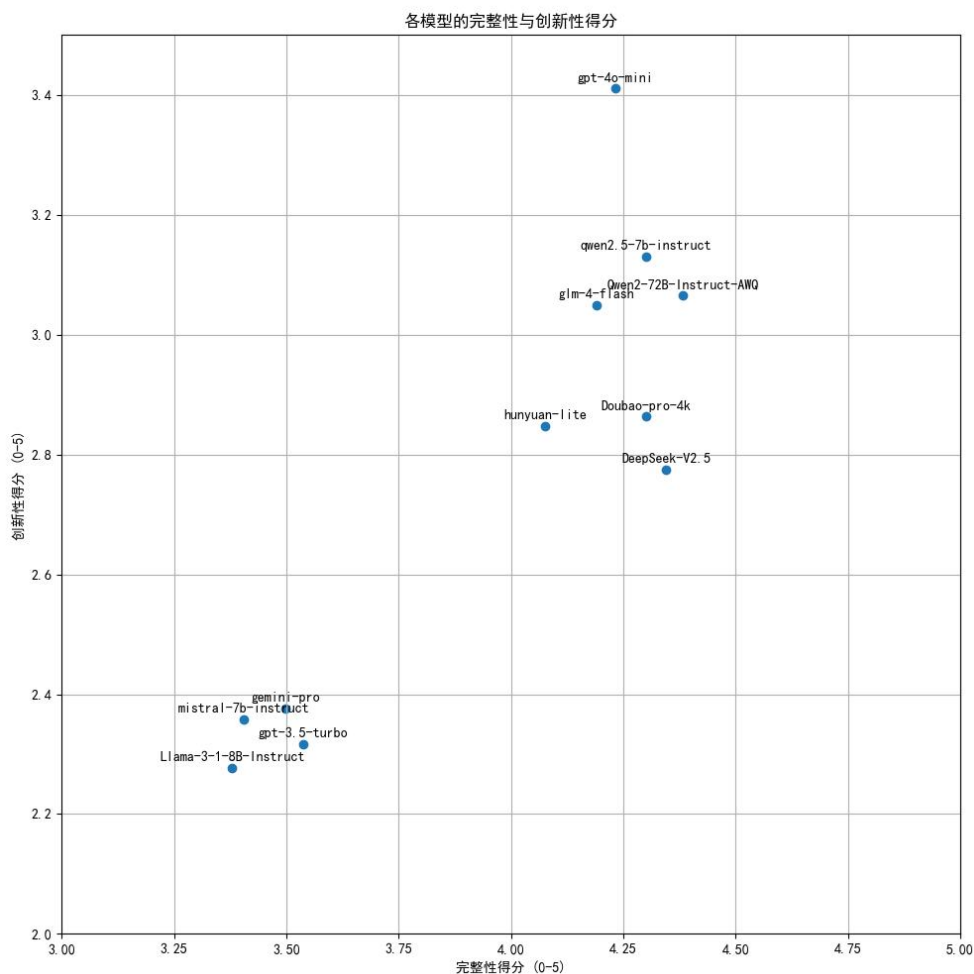


图 6-4 模型整体分析图

图 6-4 展示了多个大型模型在完整性（x 轴）和创新性（y 轴）两个维度的评分表现。评测中，完整性用于衡量模型回答的逻辑性和细节程度，而创新性则评估模型是否具备多样化的解题视角。在这两个维度上，GPT-4-mini 表现优异，尤其在处理高考推理题时，展现出较强的解答全面性和独特的解题方式，是整体表现最为突出的模型。

相比之下，Doubao-pro-4k 和 DeepSeek-V2-5 在完整性评分上达到 4.3 至 4.4 之间，表明它们在需要严谨且详尽解答的任务中具有优势。然而，这两款模型的创新性评分稍显不足，解题方法相对单一，适合应用于对内容完整性要求较高的场景。

位于图中左下角的 Gpt-3.5-Turbo、Mistral-7B-Instruct 和 Llama-3-13B-Instruct 在这两个维度上的评分均较低，显示它们在处理复杂推理任务时表现一般，更适用于较为基础的问题解答。

### 6.2.3 推理能力分析

模型推理能力分析如图 6-5 所示:

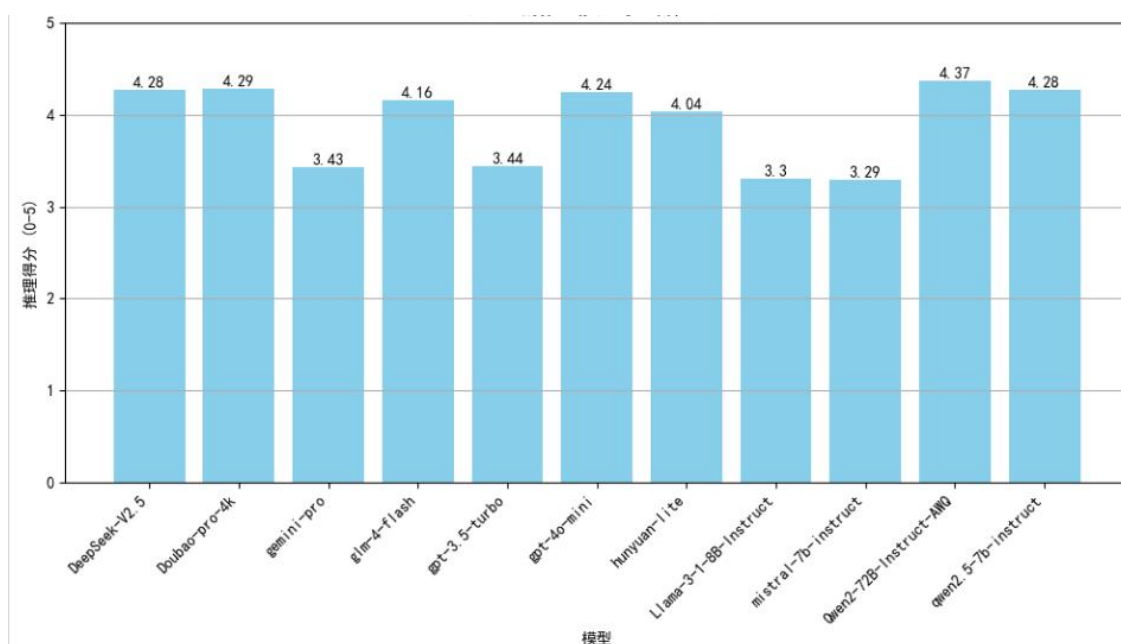


图 6-5 模型推理能力分析图

图表显示了多种模型在特定任务中的评分情况，评分范围从 3.29 到 4.37，揭示了各模型在推理能力上的显著差异。整体而言，“Qwen2-72B-Instruct-AWQ”模型以 4.37 的高分位居榜首，显示其在推理能力方面表现最为卓越。紧随其后的是“DeepSeek-V2.5”、“Qwen2.5-7B-Instruct”和“Doubao-pro-4k”，其评分分别为 4.28、4.28 和 4.29，表明这些模型在特定任务或数据处理上具有独特的优势。

相较之下，得分较低的模型如“Llama-3-1-8B-Instruct”和“Mistral-7B-Instruct”，分别得分 3.3 和 3.29，表明它们在推理任务中存在较大劣势。此外，“Gpt-3.5-Turbo”和“GPT-4-Turbo”分别获得 4.16 和 4.24 的评分，展现出较为稳定的推理能力，与高分模型较为接近。

总体趋势显示，模型的推理能力呈现出一定的两极分化现象：部分模型的得分显著高于 4.0，具备较强的解题能力；而一些模型则徘徊在 3.3 至 3.44 之间，可能存在“猜对答案”的情况，反映出其在复杂推理任务中的不足。

## 6.3 模型表现水平

经过各项指标分析，并参考 ChatGPT-4o 的建议，我们认为 Doubao-pro-4k 和 Qwen2-72B-Instruct-AWQ 的表现已达到较为良好的高中生水平；DeepSeek-



V2.5 和 Qwen2.5-7B-Instruct 则接近一般高中生水平。而其余模型的表现尚未达到高中生水平。

### 6.4 推理时间

推理时间如图 6-6 所示：



图 6-6 推理时间分析图

结果显示，随着模型参数的增加，推理速度显著下降。同时，我们观察到 Glm-4-Flash 模型的推理速度特别慢。经过验证，推测这一现象可能与 API 提供商的网络响应延迟有关，因此在不同模型之间的对比上可参考性较低，更适合在同一模型的不同学科间进行分析。

此外，我们发现模型在数学和物理科目的推理时间整体较长，而在化学和生物科目上推理时间相对较短。这一现象可能反映了不同学科问题的复杂性对推理效率的影响。

## 6.5 不同 prompt 下指标表现

### 6.5.1 简单的 prompt

简单的 prompt 示例如图 6-7 所示:

```
你是一个Biology专家，解决所给题目
1. （6分）下列有关细胞的叙述，正确的是（    ）
A. 病毒是一类具有细胞结构的生物
B. 蓝藻细胞具有细胞核且 DNA分子呈环状
C. 人体所有细胞的细胞周期持续时间相同
D. 内质网膜和高尔基体膜都具有流动性
```

图 6-7 简单的 prompt

### 6.5.2 精确格式化的 prompt

精确格式化的 prompt 示例如图 6-8 所示:

```
###身份###
你是一名参加高考的考生，拥有强大的推理和分析能力。
###科目###
生物
###指令###
你将对以下题目进行解答，在阅读题目后，你会根据问题判断所属的类型
|（概念题，分析题，计算题，公式应用题），然后进行作答
###挑战###
如果可以，请尽可能一题多解
### 输出格式###
{
  "type": "题目所属的类型",
  "analysis": [["解法一"], ["解法二"], .....]],
  #写尽可能多的写法，而非我限定的两种解法
  #如果解法有多个步骤，按1，2，3，4这样分点分析
  #每一个[]对应一种解法
  "answer": "你的答案"
}
***注意，仅仅按***### 输出格式###里的格式输出，不要输出其它内容
###题目###
1. （6分）下列有关细胞的叙述，正确的是（    ）
A. 病毒是一类具有细胞结构的生物
B. 蓝藻细胞具有细胞核且 DNA分子呈环状
C. 人体所有细胞的细胞周期持续时间相同
D. 内质网膜和高尔基体膜都具有流动性
```

图 6-8 简单的 prompt

不同 prompt 下指标表现结果分析详见 7.2。

## 7. 结果分析

### 7.1 关于模型一题多解创新性相关分析

我们以表现最好的 GPT-4o-mini 和表现最差的 Mistral-7B-Instruct 某一道数学题上的表现为例：

题目："已知集合 $M=\{x|x+2\geq 0\}$ , $N=\{x|x-1<0\}$ ，则  $M\cap N=()$  A. $\{x|-2\leq x<1\}$  B. $\{x|-2<x\leq 1\}$  C. $\{x|x\geq -2\}$  D. $\{x|x<1\}$ "

(1) GPT-4o-mini 的解答：

解法一：

1. 解集合 M:  $x+2\geq 0\Rightarrow x\geq -2\Rightarrow M=\{x|x\geq -2\}$ ,"
2. 解集合 N:  $x-1<0\Rightarrow x<1\Rightarrow N=\{x|x<1\}$ ,"
3. 计算交集  $M\cap N$ :  $M\cap N=\{x|x\geq -2\}\cap\{x|x<1\}\Rightarrow -2\leq x<1$ ,"
4. 得出结果:  $M\cap N=\{-2\leq x<1\}$ 。

解法二：

1. 将 M 与 N 的方法用数轴表示: M 自-2 向右延伸, N 自左向 1 延伸。","
2. 从数轴中找出重合部分: 即 M 与 N 交集的部分为 $[-2, 1)$ 。","
3. 形成的集合用符号表示:  $\Rightarrow M\cap N=\{-2\leq x<1\}=\{x|-2\leq x<1\}$ 。

解法三：

1. 先求 M 的解集:  $M=\{x|x\geq -2\}$ ,"
2. 再求 N 的解集:  $N=\{x|x<1\}$ ,"
3. 利用逻辑运算法, 结合两个不等式: 需要满足  $-2\leq x$  和  $x<1$ ,  $\Rightarrow -2\leq x<1$ 。","
4. 所以, 交集表示为  $\{x|-2\leq x<1\}$ 。

可以看到, 对于这个问题, GPT-4o-mini 给出了 3 种不同的解答。解法一展示了直接分析集合的求解过程, 逐步通过计算和符号逻辑推导得出答案。解法二使用数轴来直观地说明集合的交集, 提供了一个图形化理解的角度。解法三利用逻辑运算法来推导解集, 展示了一个结合逻辑分析的方式。

GPT-4o-mini 在解答时展现了很好的解题创新性、逻辑严密性和教学适应性。其内容详实、富有层次, 既满足了解题的严谨性, 也符合教学导向的多样化需

求，是一个非常优秀的解题辅助模型。

(2) Mistral-7B-Instruct 的解答：

解法一：首先，我们需要找到  $M$  和  $N$  的交集，即找到那些  $x$  使得  $x+2 \geq 0$  且  $x-1 < 0$ 。这意味着  $x \geq -2$  且  $x < 1$ 。因此， $M$  和  $N$  的交集是  $(-2, 1)$ 。因为这是一个开区间，所以我们需要将其闭合，得到  $[-2, 1)$ 。

解法二：我们可以直接使用集合的定义，找到  $M$  和  $N$  的交集，即找到那些  $x$  同时满足  $x+2 \geq 0$  和  $x-1 < 0$ 。这意味着  $x \geq -2$  且  $x < 1$ 。因此， $M$  和  $N$  的交集是  $(-2, 1)$ 。因为这是一个开区间，所以我们需要将其闭合，得到  $[-2, 1)$ 。

可以看到，Mistral-7B-Instruct 的两种解法实质上是同一方法的不同表述，且未引入新的解题思路或图示化的视觉化手段，缺乏多角度的分析。

## 7.2 关于不同 prompt 下指标表现的分析

化学、物理和生物三个科目上，规整后指令的表现分别如图 7-1、7-2、7-3 所示：

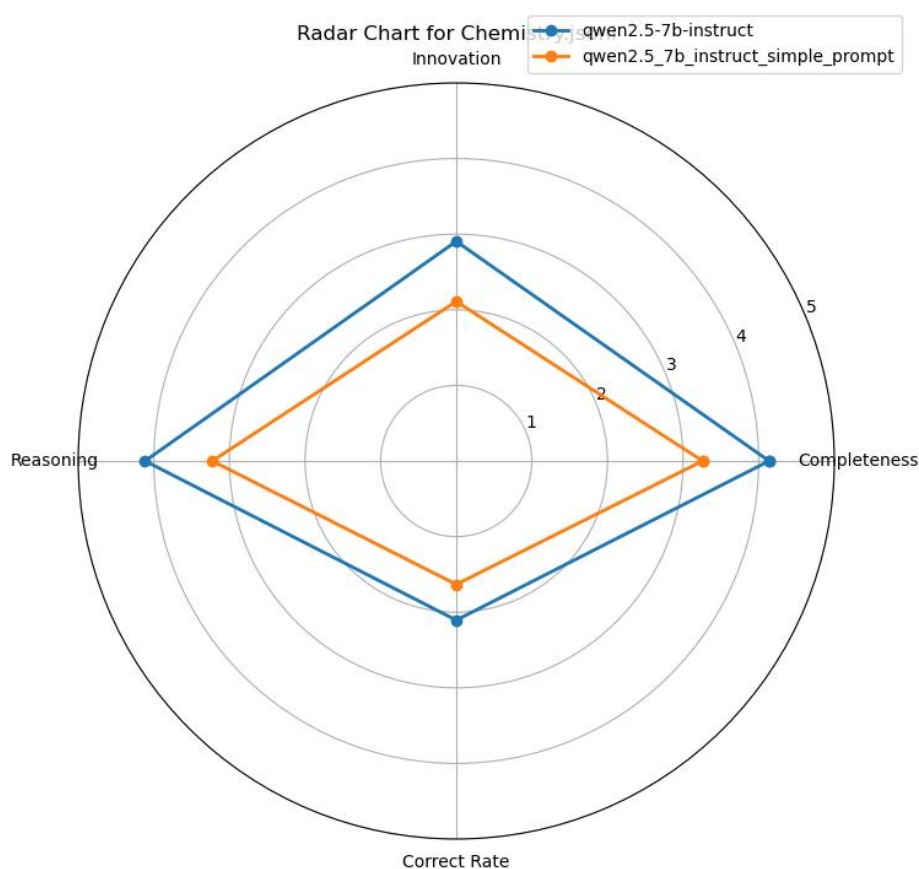


图 7-1 化学科目规整后指令表现

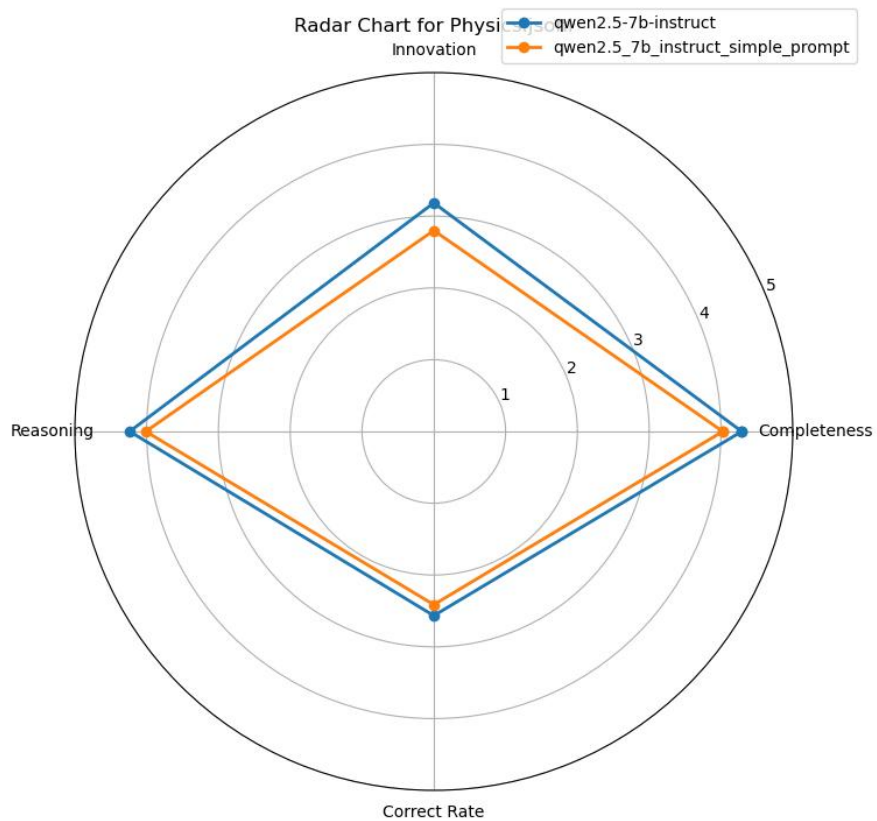


图 7-2 物理科目规整后指令表现

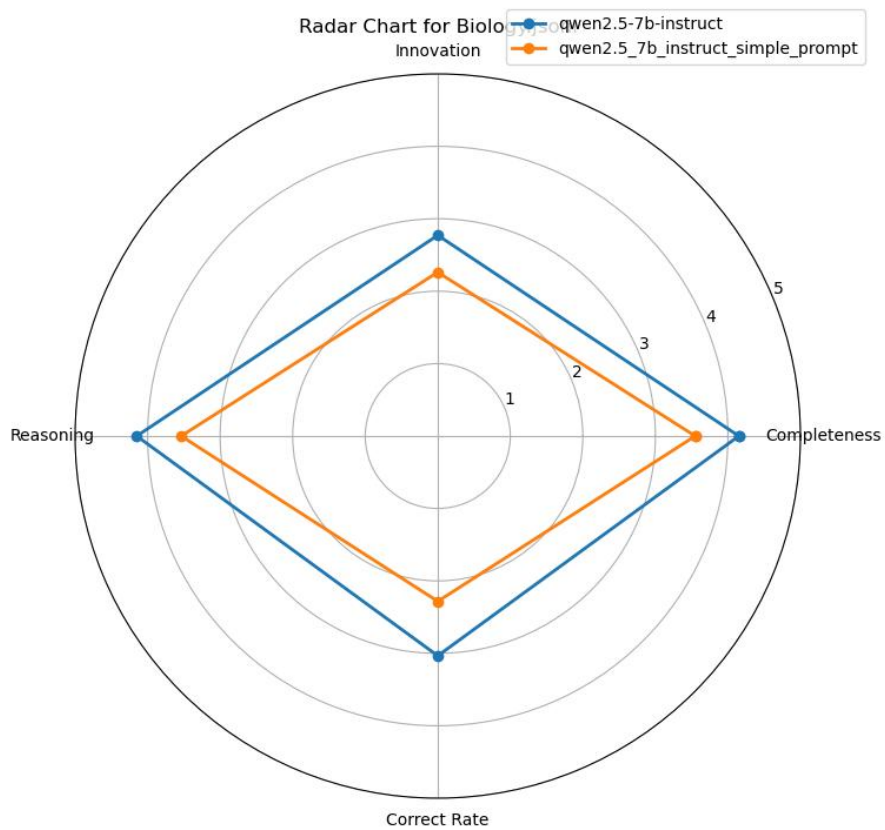


图 7-3 生物科目规整后指令表现

整体来看，在化学、物理和生物三个科目上，规整后指令的表现明显优于简单指令，尤其在创新性和完整性方面，前者展现出更为显著的优势，表明其能够生成更具创造性且内容更全面的回答。在正确率上，规整后指令也略胜一筹，进一步凸显了其在回答质量上的优越性。此外，规整后指令下的模型在推理能力上的表现也优于简单指令，表明其在逻辑推理和综合能力上更具优势。综合而言，规整后指令使模型在各项指标上均展现出更强的能力，表明其指令处理机制可能更为复杂或高效，从而显著提升了回答的质量与准确性。

数学科目规整后指令的表现分别如图 7-4 所示：

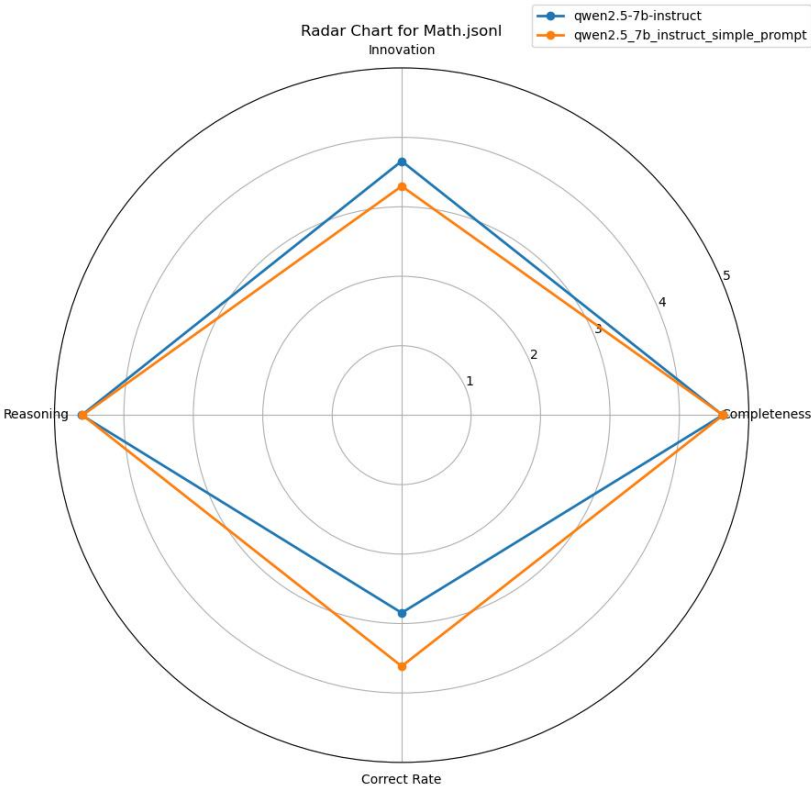


图 7-4 生物科目规整后指令表现

令人意外的是，在数学科目上，反而是简单指令下的模型表现更为出色。通过对推理结果的分析可以看出，简单指令由于未规定输出格式，为模型提供了更大的推理空间。例如，当推理过程出现错误时，模型可以及时发现并进行自我纠正，因此在步骤严密、推理复杂的数学领域表现得更为出色。而规整后的指令往往只包含解题步骤的标准化输出，模型缺乏自我检查和纠正错误的机会，这可能导致了准确率的下降。如何在输出格式的标准化与大模型的准确率之间找到平衡，是一个值得深思的问题。

同一模型在两种 prompt 设置下的表现如图 7-5 所示：

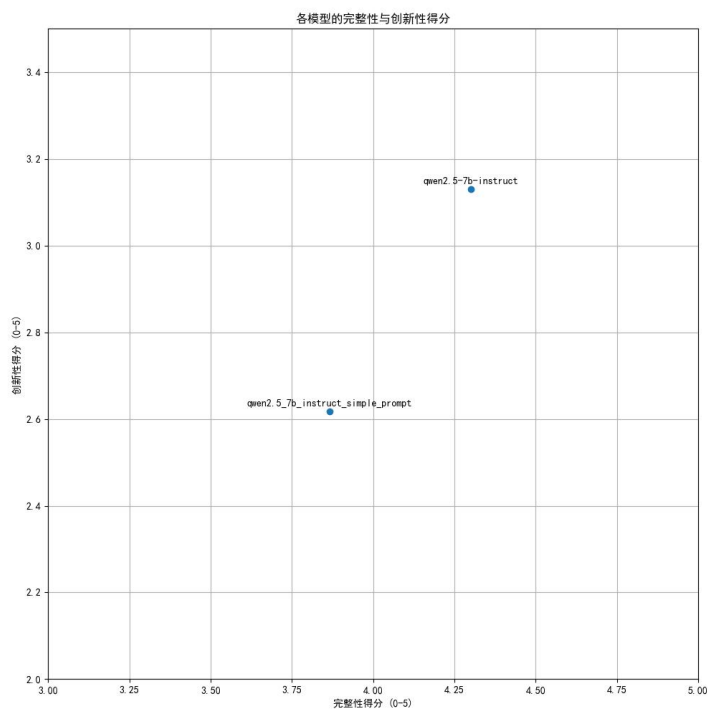


图 7-5 同一模型在两种 prompt 设置下的表现

从图中可以看出，同一模型在两种 prompt 设置下的表现差异。规整化的 prompt（在完整性和创新性方面得分更高，显示出其生成内容更全面且具有更高的创造力。相比之下，简单 prompt 的得分稍低，说明它在生成的内容丰富度和创新性上有所欠缺。

这表明规整化的 prompt 有助于提升模型的生成效果，尤其是在复杂任务中，它能引导模型提供更为详细和富有创造性的回答。这可能是因为规整化的 prompt 为模型提供了更清晰的上下文或结构，从而提升了内容质量。

各模型的推理能力与逻辑性得分如图 7-6 所示：

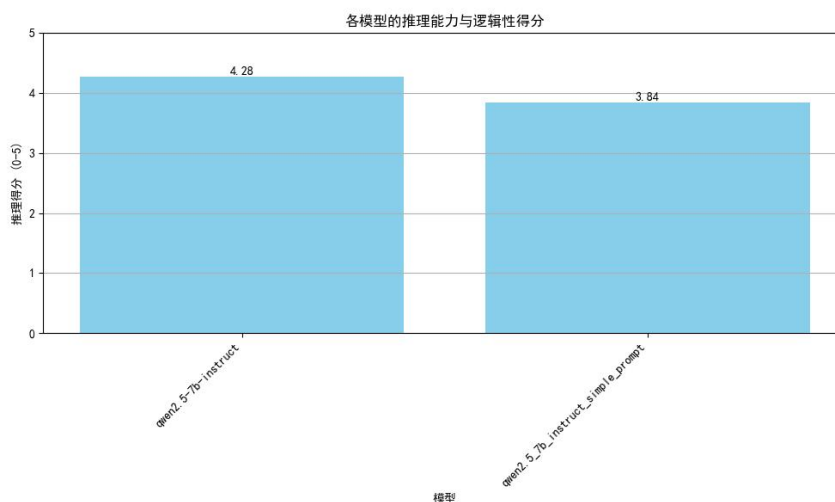


图 7-6 各模型的推理能力与逻辑性得分



规整化的 prompt 的推理得分为 4.28，显著高于简单 prompt 的 3.84。这表明，规整化的 prompt 有助于模型在推理任务中表现得更好，可能是因为规整化的 prompt 提供了更明确的结构和信息，使模型能够更准确地理解和处理信息，从而在推理时表现出更高的准确性和深度，使生成的结果更符合预期、逻辑性更强。

## 8. 创新点

### 8.1 对文科的题目推理

对文科的题目推理如图 8-1 所示：

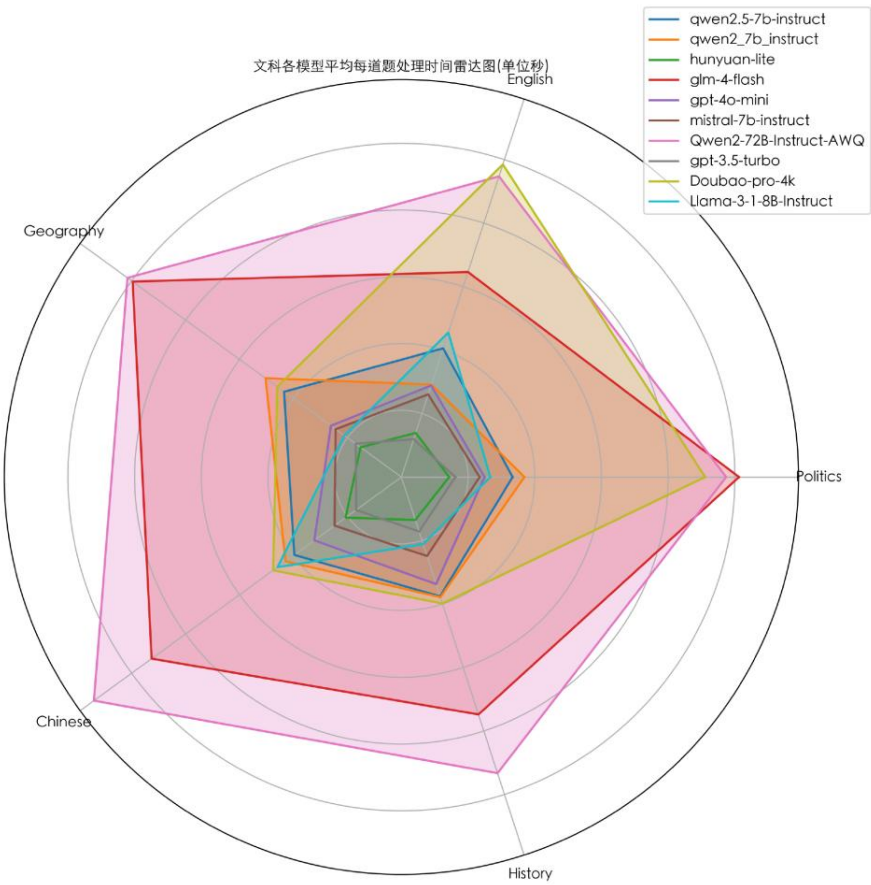


图 8-1 各模型的推理能力与逻辑性得分

从图 8-1 可以看出，大部分模型在语文和英语题目上需要更长的推理时间，且通常对这些题目需输入更多的 tokens。这表明，大模型的推理速度对输入 token 数量具有较高的敏感性。

各大模型在文科题目上的正确率表现如图 8-2 所示：



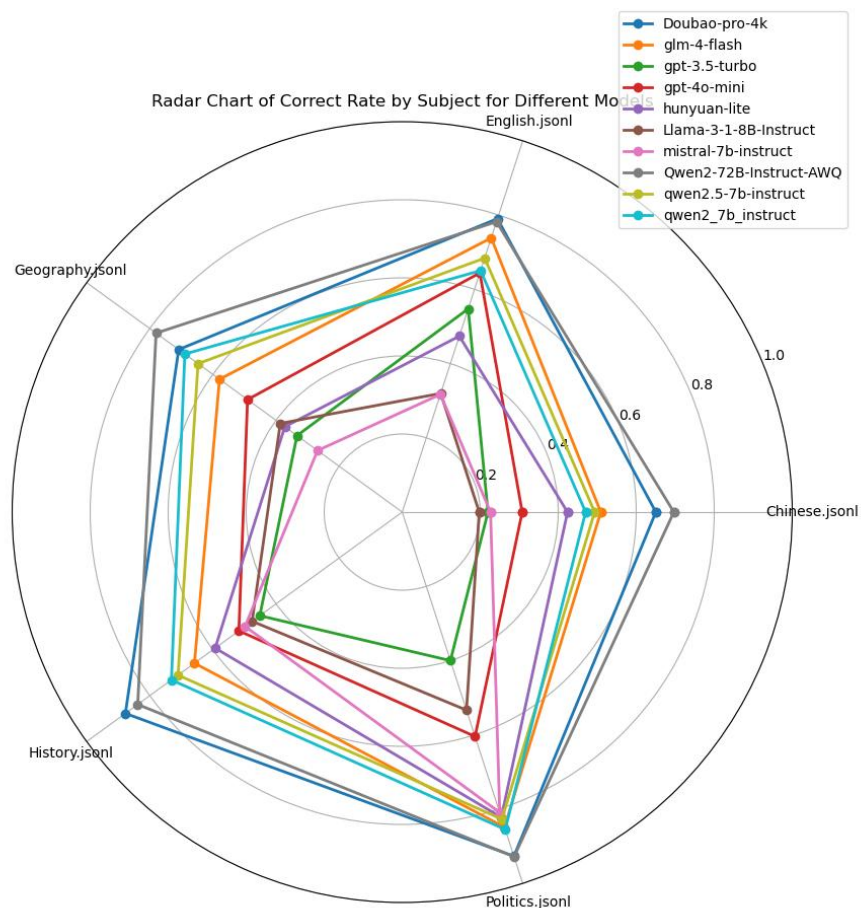


图 8-2 各模型在文科题目上的正确率表现

图 8-2 展示了各大模型在文科题目上的正确率表现。可以看出，模型在政治学科的准确率尤其高，这可能是由于大模型在记忆类问题上具有较好的优势，但在推理类问题上的表现相对较弱。同时，大多数国外模型在语文学科上的表现较差，表明其在中文数据训练上的投入仍显不足。

## 8.2 数据泄露的分析

对 22 年以前的大模型推理结果打分如图 8-3 所示：

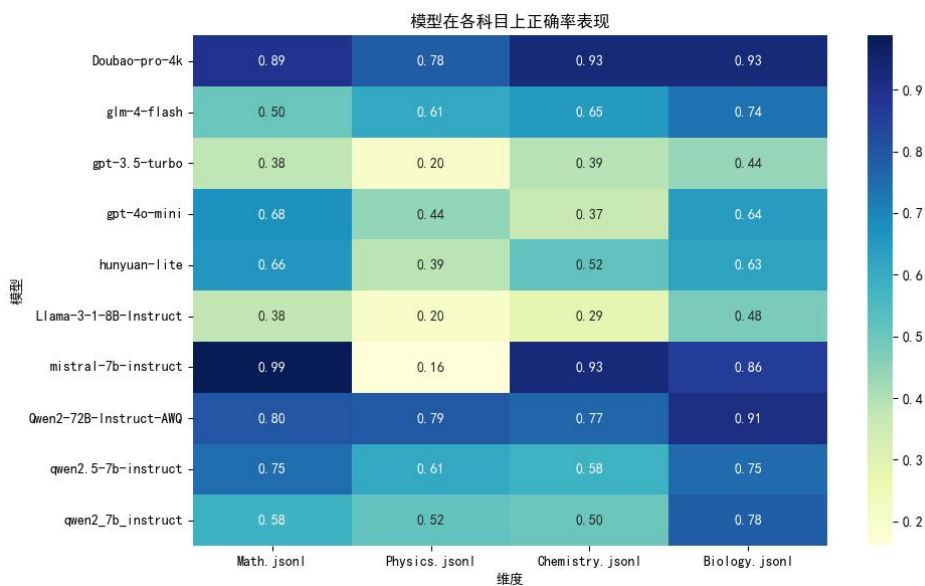


图 8-3 对 22 年以前的大模型推理结果打分情况

图 8-3 是对 22 年以前的大模型推理结果打分情况，可以看到 Mistral-7B-Instruct 在数学上的正确率达到了 99%，我们怀疑发生了数据泄露，于是重新对 23-24 的数据进行处理如图 8-4 所示：

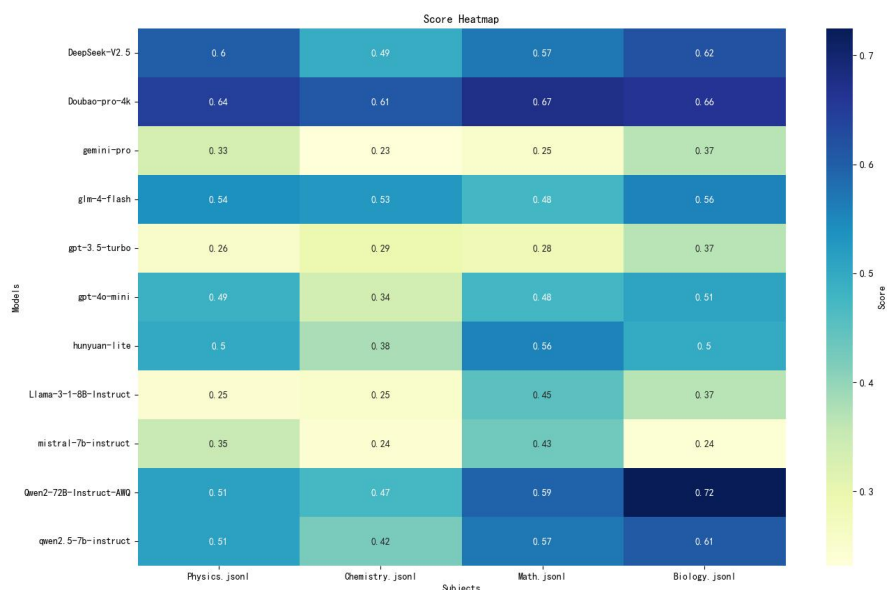


图 8-4 对 23-24 年的大模型推理结果打分情况

基于 23-24 年的打分情况，我们可以得到更加真实的大模型能力情况。

### 8.3 对解答完整性的评估

完整性推理结果完整性是指模型的回答是否包括了必要的步骤和细节。完整性评估能确保模型输出的答案是清晰且有条理的，尤其在学科问题中，完整的推理步骤往往比简单的正确答案更重要，有助于学生理解。

以之前提到的数学题答案为例：

解法一：

- 1.解集合  $M$ :  $x + 2 \geq 0 \Rightarrow x \geq -2 \Rightarrow M = \{x \mid x \geq -2\}$ ,"
- 2.解集合  $N$ :  $x - 1 < 0 \Rightarrow x < 1 \Rightarrow N = \{x \mid x < 1\}$ ,"
- 3.计算交集  $M \cap N$ :  $M \cap N = \{x \mid x \geq -2\} \cap \{x \mid x < 1\} \Rightarrow -2 \leq x < 1$ ,"
- 4.得出结果:  $M \cap N = \{-2 \leq x < 1\}$ 。

解法二：

- 1.将  $M$  与  $N$  的方法用数轴表示:  $M$  自-2 向右延伸,  $N$  自左向 1 延伸。","
- 2.从数轴中找出重合部分: 即  $M$  与  $N$  交集的部分为 $[-2, 1)$ 。","
- 3.形成的集合用符号表示:  $\Rightarrow M \cap N = \{-2 \leq x < 1\} = \{x \mid -2 \leq x < 1\}$ 。

解法三：

- 1.先求  $M$  的解集:  $M = \{x \mid x \geq -2\}$ ,"
- 2.再求  $N$  的解集:  $N = \{x \mid x < 1\}$ ,"
- 3.利用逻辑运算法, 结合两个不等式: 需要满足  $-2 \leq x$  和  $x < 1$ ,  $\Rightarrow -2 \leq x < 1$ 。","
- 4.所以, 交集表示为  $\{x \mid -2 \leq x < 1\}$ 。

对于这三种解答方式, qwen72b 给出的评价如图 8-5 所示:

分析

解法一

- 步骤1: 明确地指出了如何从给定的不等式  $x + 2 \geq 0$  得到集合  $M$  的定义。
- 步骤2: 同样清晰地展示了如何从不等式  $x - 1 < 0$  获得集合  $N$  的定义。
- 步骤3: 通过逻辑运算解释了如何计算  $M$  与  $N$  的交集, 并给出了具体的结果。
- 步骤4: 最终给出了  $M \cap N$  的解集形式。

此解法的每一步都提供了足够的信息来理解解题过程, 因此具有较高的完整性。

解法二

- 步骤1: 使用数轴的方式直观地表示了  $M$  和  $N$  两个集合。
- 步骤2: 通过数轴上的重合部分直观地找到了  $M$  与  $N$  的交集。
- 步骤3: 将数轴上的观察结果转化为数学符号表达。

这种方法通过视觉辅助工具帮助理解, 对于某些学习者来说可能更容易掌握。同时, 它也保持了一定程度的完整性, 因为每个步骤都是连贯的, 能够清楚地展示出解题思路。

解法三

- 步骤1: 直接给出了  $M$  的解集。
- 步骤2: 同样直接给出了  $N$  的解集。
- 步骤3: 利用逻辑运算符将两个条件结合起来, 得出了交集的条件。
- 步骤4: 最后以数学符号的形式表示了交集。

解法三同样遵循了清晰的逻辑顺序, 逐步解析了问题, 确保了解题过程的透明度和逻辑性, 体现了良好的完整性。

图 8-5 qwen72b 给出的评价情况

三种解法都展现了从问题描述到最终答案的完整路径，其中包含了必要的步骤和细节。这些解法不仅有助于验证答案的正确性，更重要的是，它展示了大模型理解数学概念的过程。

而对于某道数学题 mistral7b 给出的解题过程：

“解法三：利用极限的概念来检验每个函数在区间 $[(0,+\infty)]$ 上的行为。对于选项 C，当  $x$  从 0 趋向于正无穷大时， $f(x)=-\frac{1}{x}$  的值从负无穷大逐渐增加到 0，这表明函数在这个区间上是单调递增的。”对于这个回答，模型的回答并没有展现出很好的完整性，" $f(x)=-\frac{1}{x}$  的值从负无穷大逐渐增加到 0"，和"函数在这个区间上是单调递增的"只是对单调递增的两种表述，大模型的回答缺少了证明函数单调递增（比如说用导数大于零）的步骤。

## 9. 总结与展望

在本研究中，我们对 12 个市面上流行的大型语言模型进行了系统评测，重点考察了这些模型在理科和文科试卷上的表现。评测采用横向对比（不同厂商的模型之间）和纵向对比（同厂商不同参数配置的模型之间），分析了以下几个关键方面：

- (1) 推理时间：记录模型在不同题型上的推理速度，以衡量其效率。
- (2) 提示词影响：观察不同提示词（prompt）对模型答题表现的影响，了解提示词设计的优化空间。
- (3) 答题准确率：统计各模型在不同题型上的正确率，以此评价模型的准确度。
- (4) 创造性：特别针对主观题，分析模型是否能够提供多解以及多解的多样性，评价其在回答上的创新能力。
- (5) 逻辑推理能力：重点关注模型在逻辑推理过程中是否能够符合题目要求，逻辑推理链是否清晰且合理。
- (6) 回答完整性：观察模型的回答是否具备充分的信息覆盖，评估其回答的全面性。

基于以上评测，研究为全面了解主流大语言模型在应对高考题目时的性能提供了重要参考，有助于未来在提升模型推理和理解能力方面做出优化。

尽管已经对大模型在高考题目上进行了细粒度的分析，但未来的评估工作仍有以下潜在方向：

(1) 多模态模型的评估：虽然目前已对数据集进行文本和图片多模态的划分，但现阶段仅完成了文本数据的处理。后续可以考虑加入多模态模型的评估，从而了解多模态大语言模型在视觉文本多模态信息上解题的表现。

(2) 细化题目标签：研究中已将题目划分为主观题和客观题，未来可以在此基础上进一步细分不同题目子类型，探讨在各细分类别下模型的表现，以便更好地识别模型在不同题目类型上的推理瓶颈，优化模型推理性能。

(3) 标准化答案处理：在选择题中，可以尝试使用正则化处理对答案进行提取，使评分更加一致。在主观题评估方面，可以适当放宽内容多样性限定，提升大模型评测的准确性，并发掘模型的创造性潜力。