2024 年第二届大湾区杯科技竞赛

题　　号	**CG2401**		
标　　题	大语言模型解答高中理科题目性能评估		
成员信息	姓名：谢李贝蕾	单位： 广东金融学院	邮箱： selina_xlbl@126.com
	姓名：马乔南	单位： 广东金融学院	邮箱： mjoen12138@gmail.com
	姓名：邓嘉伟	单位： 广东金融学院	邮箱： colaluckday@foxmail.com

签名：谢李贝蕾、马乔南、邓嘉伟

摘　要

近年来，大语言模型（LLMs）在自然语言处理领域取得了显著的进展，尤其在文本生成和问题回答等任务上展现出了强大的潜力。然而，关于 LLMs 在解答特定领域问题上的表现，特别是面对高中理科题目的解题能力，有待提高。因此，本研究旨在全面、系统地评估多种 LLM 在解答高中理科题目方面的能力，并分析其在教育领域中的应用前景。

我们通过自建题库并结合开源题库，设计了一个包含四轮评测的技术流程，涵盖了多个维度的性能评价，包括正确率、逻辑性、思维性、解题能力、多解能力、创造性以及响应速度。为保证评估的科学性和准确性，我们使用了多种评分方法，如精确评分、加权评分、思路多样性评分、响应时间对比评分，充分衡量各模型在多维度的实际表现，并与以往研究中采用的单一评分方法形成鲜明对比。

在实验过程中，我们设计了多类型评测任务，以全面捕捉模型在实际解题情境下的表现。每个评测任务涉及不同类型的题目和多种提示词方案，采用控制变量实验观察提示词对模型正确率的影响。同时，实验关注了模型间的横向对比、响应时间差异以及各模型在不同题目类型上的综合得分，为教育应用中的模型选择提供数据支持。

为确保实验结果的可靠性，我们对数据集误差、评分误差和应答时间误差进行了深入分析，以评估这些误差来源对结果的影响。通过这样的误差分析，我们更清晰地理解模型在解题任务中的真实表现，同时为后续的优化提供了基础数据支持。

实验结果表明，优化后的提示词显著提升了模型的正确率，尤其在逻辑性和创造性要求较高的题目上表现尤为突出。此外，不同模型在各学科领域表现差异明显，体现出各模型在知识覆盖面和推理能力方面的不同适应性（如 doubao-pro-128k 模型在数学题目中的正确率平均高出 llama-70b 模型 37%）。这些结果说明了针对不同学科内容的模型选择策略在实际应用中的必要性。

未来的研究可以进一步探索模型在多学科和多年龄层的适用性，并研究如何提升模型在逻辑推理和创新能力上的表现，最终为教育技术的智能答疑系统和辅助教学工具开发中提供理论和技术支持。

关键词：大语言模型（LLMs）、高中理科题目、性能评估、评测体系、提示词优化、响应时间分析、综合得分、教育技术

目录

一、 问题分析

大语言模型 (LLMs) 在多个领域取得显著进展，但其在解决复杂推理问题，尤其在数学等需要严谨逻辑和精确计算的学科上的能力仍待提高。近期复旦大学 NLP 实验室利用高考数学题对 LLMs 的评测结果显示模型表现不理想，这引发了我们对其在解决高中理科题目能力的关注。

本次大赛邀请参赛者对大语言模型解答高中理科题目的能力进行全面评测，我们提出了精确评分、加权评分、思路多样性评分、响应时间对比评分、elo 赛制评分等方法。我们的创新点有（多维度、多方法量化评测体系和提示词、调参优化模型表现）。

(流程图)

二、 模型分析

本文从各模型官方文档说明的信息出发，考量其训练参数大小、是否拥有视觉、api 调用方式等，选择了如下团队的具有代表性的模型。

表 1：本文所选模型及其团队介绍

团队名称	简介	模型名称	模型简介
零一万物	零一万物是李开复带队孵化的 AI2.0 公司，总部注册于北京，集中在大模型技术、人工智能算法、自然语言处理、系统架构、算力架构、数据安全、产品研发等领域。	yi-vision	复杂视觉任务模型，提供基于多张图片的高性能理解、分析能力。
		yi-large	千亿参数大尺寸模型，提供超强问答及文本生成能力，具备极强的推理能力。并且对 System Prompt 做了专属强化。
腾讯混元	腾讯混元大模型是由腾讯研发的大语言模型，具备跨领域知识和自然语言理解能力，实现基于人机自然语言对话的方式，理解用户指令并执行任务，帮助用户实现人获取信息，知	hunyuan-pro	万亿级参数规模 MOE-32K 长文模型。在各种 benchmark 上达到绝对领先的水平，复杂指令和推理，具备复杂数学能力，支持 functioncall，在多语言翻译、金融法律医疗等领域应用重点优化。

	识和灵感。	hunyuan-vision	混元最新多模态模型，支持图片+文本输入生成文本内容。 图片基础识别：对图片中主体、元素、场景等进行识别 图片内容创作：对图片进行概述、创作广告文案、朋友圈、诗词等 图片多轮对话：输出单张图片进行多轮交互问答 图片分析推理：对图片中逻辑关系、数学题、代码、图表进行统计分析 图片知识问答：对图片包含的知识点进行问答，例如历史事件、电影海报 图片 OCR：对自然生活场景、非自然场景的图片识别文字
Google Gemini	Gemini 是一款由 Google DeepMind（谷歌母公司 Alphabet 下设立的人工智能实验室）于 2023 年 12 月 6 日发布的人工智能模型，可同时识别文本、图像、音频、视频和代码五种类型信息，还可以理解并生成主流编程语言（如 Python、Java、C++）的高质量代码，并拥有全面的安全性评估。	gemini-1.5-pro	高度计算高效的多模态专家混合模型，能够从数百万个令牌的上下文 (包括多个长文档和数小时的视频和音频) 中回忆和推理细粒度信息。Gemini 1.5 Pro 在跨模态的长上下文检索任务上实现了近乎完美的召回，改进了长文档问答、长视频问答和长上下文语音识别领域的最新技术水平，并且在广泛的基准测试中与 Gemini 1.0 Ultra 的最新技术水平持平或超越。
字节跳动豆包	豆包是字节跳动公司基于云雀模型开发的 AI 工具，提供聊天机器人、写作助手以及英语学习助手等功能，它可以回答各种问题并进行对话，帮助人们获取信息，支持网页 Web 平台，iOS 以及安卓平台。	doubao-pro-128k	Doubao 效果最好的主力模型，适合处理复杂任务，在参考问答、总结摘要、创作、文本分类、角色扮演等场景都有很好的效果。支持 128k 上下文窗口的推理和精调。

智谱清言	智谱是由清华大学计算机系技术成果转化而来的公司，致力于打造新一代认知智能通用模型。公司合作研发了双语千亿级超大规模预训练模型 GLM-130B，并构建了高精度通用知识图谱，形成数据与知识双轮驱动的认知引擎，基于此模型打造了 ChatGLM。	glm-4v-plus	集图像理解与视频理解能力于一体的多模态模型。GLM-4V-Plus 在图像和视频理解领域均展现出领先水平，稳居行业前列。
阿里千问	通义千问是由阿里云自主研发的大语言模型，用于理解和分析用户输入的自然语言，在不同领域和任务为用户提供服务和帮助。	qwen-plus	Qwen-plus 是一款大型语言模型，提供稳定版和持续更新的最新快照版，以及不同时间点的快照版本供用户选择。它拥有强大的上下文理解能力，最新的快照版支持高达 131,072 个 token 的上下文长度。
		qwen-vl-plus	Qwen-VL-Plus 是一个强大的多模态大型语言模型，能够理解和处理文本和视觉信息。它支持图像、文本和检测框等多种输入模态，并能根据这些信息进行对话、生成图像描述、回答视觉问题等。相比之前的版本，Qwen-VL-Plus 拥有更强的视觉理解和推理能力，以及更优秀的跨模态生成能力。 它基于 Qwen-Plus 构建，并通过大量的图文数据进行训练，使其在处理复杂多模态任务时表现出色。
月之暗面	月之暗面是一家专注于对话式 AI 助手的科技公司，旗下产品 Kimi 智能助手支持 200 万字的无损上下文输入，能够快速学习任何领域的专业知识。	moonshot-v1-32k	一款千亿参数的语言模型，具备优秀的语义理解、指令遵循和文本生成能力。支持 32K 上下文窗口，适合长文本的理解和内容生成场景。本文详细介绍了 Moonshot-v1-32k 的 SDK 及 API 使用方法。
深度求索	深度求索（DeepSeek）是知名	deepseek-chat	一款基于专家混合（MoE）架构的 2360

	私募巨头幻方量化旗下的人工智能公司，自主研发了大语言模型开发的智能助手，可以进行自然语言处理、问答系统、智能对话、智能推荐、智能写作和智能客服等多种任务。		亿参数大语言模型，拥有强大的语言理解和生成能力。每次请求仅激活 210 亿参数，在提升性能的同时降低了训练成本、键值缓存使用，并提高了生成吞吐量。
OpenAI	OpenAI 是一个美国人工智能研究实验室。OpenAI 的目的是促进和开发友好的人工智能，使人类整体受益。OpenAI 提供了一个基于 AI 的开发和研究框架，开发、维护和训练了一批可用于通用活动的 AI 模型。	gpt-4o-mini	MMLU 评分达 82%，聊天偏好胜过 GPT-41。价格仅为每百万输入 token 15 美分，输出 60 美分，比 GPT-3.5 Turbo 便宜 60% 以上。它支持 128K 上下文窗口和文本图像输入输出，未来将支持更多模态。 GPT-4o mini 在数学 (MGSM: 87%) 和编码 (HumanEval: 87.2%) 上也表现出色，旨在更广泛地普及 AI 应用。
		gpt-4o	端到端地处理文本、图像、音频和视频输入，并能生成文本、音频和图像输出。能够更自然地理解和响应用户，例如理解语音中的语气和背景噪音，并以更接近人类的反应速度（平均 320 毫秒）进行交互。在英文文本和代码方面与 GPT-4 Turbo 持平，非英文文本处理能力显著提升，同时速度更快，API 成本更低。
		o1-preview	o1-preview 是比 o1-mini 更强大但更昂贵的 OpenAI 模型，拥有更强的推理能力和更广泛的知识库，适用于更复杂的任务，但仍在预览阶段。
Claude	Claude 是由 Anthropic 公司开发的先进人工智能模型，该公司专注于创建安全、可靠且符合	claude-3-5-sonnet-20240620	在软件工程、代理能力和计算机使用方面达到最先进水平。该模型可用于代码生成、高级聊天机器人、知识问答、视

	人类价值观的人工智能系统。Claude 以信息论之父香农的名字命名，旨在成为友善、诚实、无害的人工智能助手，协助完成从自然语言理解到复杂问题解决等一系列任务。		觉数据提取和机器人流程自动化等多种场景，并通过 API、Amazon Bedrock、Google Cloud Vertex AI 和 Claude.ai 提供给开发者和用户。
Mistral	Mistral AI 是一家法国初创公司，由 DeepMind 和 Meta Platforms 的前研究人员创立。该公司致力于构建大规模通用人工智能 （AGI） 模型。它专注于为各种任务（从聊天机器人到代码生成）构建快速、安全的大规模语言模型 （LLM）。	pixtral-12b-240910	一款开源多模态模型，可理解图像和文本。它在多模态基准测试中表现出色，超越了更大规模的模型，且不牺牲自然语言处理能力。该模型采用全新视觉编码器，可处理任意数量的自然分辨率图像，并拥有 128K 的长上下文窗口。Pixtral-12B 的性能优于同规模甚至更大规模的开源模型，并附带一个开源基准 MM-MT-Bench 用于评估。
Meta LLaMA	LLaMA （英语： Large Language Model Meta AI）是 Meta AI 公司 于 2023 年 2 月发布的 大型语言模型。 它训练了各种模型，这些模型的参数从 70 亿到 650 亿不等。	llama3-70b	提供 80 亿和 700 亿参数两种规模，均有预训练和指令微调版本。指令微调版在对话方面表现出色，胜过许多开源聊天模型。采用优化的 transformer 架构，支持 8k 上下文窗口，并使用分组查询注意力 (GQA)。

三、 技术路线

3.1 题目集

3.1.1 自建立题目集

首先将官方提供的 word 版本试卷集转化为 html 文档，使用 beautifulsoup[1]进行文档解析与读取。通过阅读各大模型的官方文档，本文决定采用 OpenAI 框架对接受测试的模型进行 api 调用。

参考各大模型的官方训练集, 本文承接训练方法, 采取 LaTex 表达[2]题目集[3]中的数学公式、物理公式与化学方程式。同时，OpenAI 官方文档对图片[4]进行 Base64 解析编码处理，因此本文采用二进制形式保存图片至数据库，接着采用 Base64 进行解析编码，使模型能够理解题目中的图片信息。

选择题（样例）：

id	origin	isMultiple	question	image	options	answer	analysis	avail
8	2022年高考数学试卷（北京）	0	若 (2x-1)^{4}=a_{4} x^{4}+a_{3} x^{3}+a_{2} x^{2}+a_{1} x+a_{0}，则 a_{0}+a_{2}+a_{4}=（ ）	PNG.bin	A. 40 B. 41 C. -A{\big{}\} D. -{\A{A}}|	B	利用赋值法可求a_{0}+a_{2}+a_{4}的值.【详解】令 \scriptstyle x\;=\;1，则a_{4}+\alpha_{3}+\alpha_{2}+ \alpha_{1}+\alpha_{0}=1，令\scriptstyle x\;=\;-1，则 \alpha_{4}-\alpha_{3}+\alpha_{2}-\alpha_{1}+ \alpha_{0}=\left(-3\right)^{4}=81，故 a_{4}+a_{2}+a_{0}={\frac{1+81}{2}}=41	1

图 1：题目集_选择题样例

表 2：题目集_选择题列信息

列信息	含义说明
id	题号，这是每道题目的唯一标识符
origin	试卷来源，存储试题来源的信息，例如试卷名称或编号
isMultiple	是否为多选题
question	问题，存储试题的具体内容
image	图片，用于存储与试题相关的图片（采用二进制编码格式）
options	选项，存储试题的所有选项，例如 A、B、C、D 等
answer	答案，存储正确答案
analysis	解析，存储答案的详细解析
avail	是否可用，标识该试题是否处于可用状态，默认值为 `1`（可用）

图 1 与表 2 展示了一道编号为 8 的来源于 2022 年高考数学北京卷的单选题样例。该题题干包含图片与文字信息，正确答案是 B 选项，同时该题为可用状态。选项、文字题面与解析中的公式使用 LaTex 表达式呈现，图片采用二进制格式存储。

填空题（样例）：

Id	origin	question	image	answer	analysis	avail
87	2024年高考数学试卷（理）（全国甲卷）	\left(\frac{1}{3}{+}x \right)^{\mathbf{10}}的展开式中，各项系数的最大值是_____.	Null	5	先设展开式中第\boldsymbol{\gamma}+1项系数最大，则根据通项公式有\left\{\ C_{\mathfrak{B}}^{\gamma}\left(\frac{1}{2}\right)^{0-\gamma}\ge\ C_{10}^{\prime+1}\left(\frac{1}{5}\right)^{9-\gamma}，进而求出\mathbf{\boldsymbol{\gamma}}即可求解.【详解】由题展开式通项公式为\widetilde{Z}_{r+1}=\mathsf{C}_{10}^{r}\left(\frac{1}{\overline{3}}\right)^{10-r}x^{\gamma}，0\le r\le10且r\in\mathbf{Z}，设展开式中第\boldsymbol{\gamma}+1项系数最大，则\left\{\ C_{\mathfrak{B}}^{\gamma}\left(\frac{1}{2}\right)^{0-\gamma}\ge\ C_{10}^{\prime+1}\left(\frac{1}{5}\right)^{9-\gamma}，\Rightarrow\left\{r\ge\frac{29}{4}\right.，即\frac{29}{4}\check{\underline{{{\Omega}}}}r\check{\underline{{{\Omega}}}}\frac{\bar{3}\bar{3}}{4}，又r\in\mathbf{Z}，故\boldsymbol{r}=\boldsymbol{\mathrm{8}}，所以展开式中系数最大的项是第9项，且该项系数为C_{10}^{\mathfrak{s}}\left(\frac{1}{3}\right)^{2}=5.	1

图 2：题目集_填空题样例

解答题（样例）：

Id	Origin	question	image	answer	analysis	avail
90	2024年高考物理试卷（新课标）	将重物从高层楼房的窗外运到地面时，为安全起见，要求下降过程中重物与楼墙保持一定的距离。如图，一种简单的操作方法是一人在高处控制一端系在重物上的绳子P，另一人在地面控制另一根一端系在重物上的绳子Q，二人配合可使重物缓慢竖直下降。若重物的质量\mathcal{m}=42\mathtt{k g}，重力加速度大小\mathcal{E}=10\mathtt{m}/\mathfrak{s}^{2}，当P绳与竖直方向的夹角\mathcal{U}{=}\bar{\mathcal{G}}\bar{\mathcal{Y}}^{\mathsf{o}}时，Q绳与竖直方向的夹角\mathcal{B}{=}5\mathfrak{Z}^{\mathtt{o}},(\mathfrak{s i n Z7}^{o})=0.6)（1）求此时P、Q绳中拉力的大小；（2）若开始竖直下降时重物距地面的高度\hbar=10\mathrm{m}，求在重物下降到地面的过程中，两根绳子拉力对重物做的总功。	PNG.bin	（1）1200\mathrm{N}，\mathtt{900N}；（2）-42005	（1）物块在薄板上做匀减速运动加速度大小为a_{\mathtt{l}}=\mathtt{A}\mathtt{g}=\mathtt{Z}\mathtt{m}/\mathtt{s}^{2}薄板做加速运动的加速度\mathcal{Q}_{2}=\frac{\mathcal{J}\mathcal{M}\mathcal{E}}{\mathcal{M}}=\mathcal{Z}\mathrm{fn}/\mathfrak{s}^{2}对物块\displaystyle\boldsymbol{\boldsymbol{\boldsymbol{\xi}}}+\boldsymbol{\Delta\boldsymbol{\boldsymbol{\xi}}}=\nu_{0}\boldsymbol{\boldsymbol{\boldsymbol{\varepsilon}}}-\frac{1}{2}\alpha_{1}\boldsymbol{\boldsymbol{\boldsymbol{\varepsilon}}}^{2}对薄板\Delta\boldsymbol{\boldsymbol{\lambda}}=\frac{1}{2}\boldsymbol{\alpha}_{\boldsymbol{\lambda}}\boldsymbol{\boldsymbol{\boldsymbol{\varepsilon}}}^{2}解得\nu_{\tt0}=4\tt{m}^{j}s\ell=\frac{1}{\bar{\jmath}}\mathfrak{s}（2）物块飞离薄板后薄板得速度\nu_{2}=a_{2}\ell=\mathtt{l m}^{j}s物块飞离薄板后薄板做匀速运动，物块做平抛运动，则当物块落到地面时运动的时间为\dot{\varepsilon}^{\cdot}=\frac{\frac{\dot{\ell}}{2}-\frac{\dot{\xi}}{\delta}}{\nu_{2}}=\frac{1}{3}s则平台距地面的高度\lambda=\frac{1}{2}\operatorname{\mathsf{g}}\ell^{\prime2}=\frac{5}{9}\operatorname{\mathsf{fn}}	1

图 3：题目集_解答题样例

表 3：题目集_填空题与解答题列信息

列信息	含义说明
id	题号，这是每道题目的唯一标识符
origin	试卷来源，存储试题来源的信息，例如试卷名称或编号
question	问题，存储试题的具体内容

image	图片，用于存储与试题相关的图片（采用二进制编码格式）
answer	答案，存储正确答案
analysis	解析，存储答案的详细解析
avail	是否可用，标识该试题是否处于可用状态，默认值为 `1`（可用）

表 3 说明题目集中的填空题与解答题所含列信息相同，结合图 2 展示了一道编号为 87 的来源于 2024 年高考数学全国甲卷的填空题样例，正确答案是 5。该题为可用状态，文字题面与解析中的公式使用 LaTex 表达式呈现，但不包含图片信息。

表 3 结合图 3 展现了一道编号为 90 的解答题样例，来自 2024 年新课标高考物理卷且状态为可用。该题包含文字与图片信息

，同样使用 LaTex 表达式呈现文字题面、答案、解析中包含的公式。

表 4：数据集情况

科目 \ 题型	选择题	填空题	解答题	总计（条）
数学	216	88	113	417
物理	308	\	144	452
生物	327	\	119	446
化学	308	\	57	365

如表 4 所示，截至目前，本文题目集拥有 1680 条数据，涵盖 2022-2024 近三年来各省份高考理科试题，并且在持续更新[5]。

数据预处理

对于收集完备的题目集，为了避免数据不合规范，我们使用人工筛选配合模型对题库进行了三次筛选，以确保题目和答案的可用性，并重点关注 LaTeX 公式的正确表达，题目和答案是否可用，如题目描述清晰完整以及答案正确无误，并且不包含任何歧义。

我们使用 GPT-4-0-mini 来判断 LaTeX 公式的语法和语义是否正确，判断置信度阈值设置为 0.9。如果置信度低于 0.9，则人工进行检查。

通过三次筛选，我们最终得到了 1201 道高质量的题目，这些题目包含清晰完整的描述、正确的答案以及语义正确的 LaTeX 公式。

3.1.2 开源题目集

GAOKAO-Bench[6]：一个以中国高考题目为数据集，测评大模型语言理解能力、逻辑推理能力的测评框架。

GSM8K[7]：为大模型数学能力评测常见的 benchmark 之一，题目类型为小学数学题，共 8.5k 样本。

MATH[8]：为大模型数学能力评测常见的 benchmark 之一，是一个包含 12500 个高中数学竞赛的问题（7500 个用于训练，5000 个用于测试）的数据集，以文本模式的 Latex 格式呈现。MATH 中的每个问题都有一个完整的逐步解决方案，有助于 CoT 训练。

3.2 评测体系

表 5：四轮评测简介[1]

评测轮次	评分方法	简介	涵盖能力
Round 1	精确评分	采用高考阅卷形式，模型需给出每道题的精确答案与必要的解题步骤。	正确率，逻辑性
Round 2	加权评分	基于 Round 1 的结果，根据题目难度系数加权求和，重新计算得分率。	思维性, 实际解题能力
Round 3	思路多样性评分	在难度系数基础上，抽取代表性题目，鼓励模型提供多种正确的解题思路。思路越多，得分越高。	一题多解能力, 创造性
Round 4	响应时间对比评分	对比模型在各轮次的平均响应时间和波动情况，评估处理速度的稳定性。	速度

如表 5 所示，我们共进行四轮正式评测，以量化对比各模型的多项能力指标：正确率、逻辑性、思维性、实际解题能力、一题多解能力、创造性和速度。

3.2.1 Round 1

本轮中，所有填空题与选择题每题总分赋 5 分，解答题总分 12 分。其中多选题答对部分赋 3 分，多选题选错一项则不得分；多空题答对第一空赋 2 分，答对第二空赋 3 分。

对于答案表述唯一的填空题与单选题，本文使用自动化程序判断模型答题的正确率。

对于答案表述不唯一的填空题与解答题, 我们使用 gpt-4o-mini 进行逻辑性与正确率的打分。

[1] 注：模型响应时长会受到网络波动影响，本文考虑最小化网络波动带来的干扰，采取多次测试取平均值。

具体操作是将结果集中的模型答案与题目集中的正确答案、解析输入给 gpt-4o-mini，让其对比模型与解析、模型与模型的答案进行打分。

同时为了增加 gpt-4o-mini 评分的准确性和稳定性，我们修改优化提示词[9]，并且不在提示词中体现被评测的模型身份。通过的多次“盲”打分，本文最终求取求平均数，得出模型们在表述不唯一的填空题与解答题中的逻辑性与正确率的综合打分。

3.2.2 Round 2

本轮在 Round 1 基础上进行题目难度量化与结果集模型得分的新一轮加权计算，旨在探究模型的实际解题能力与难题得分率中所体现的思维性。

表 6：题目难度量化标准

官方样例解析的字符串长度 L_i	难度系数 r_i	权重 w_i
$0 \le L_i \le 300$	1	0.2
$300 < L_i \le 800$	2	0.4
$800 < L_i \le 1100$	3	0.6
$1000 < L_i \le 1500$	4	0.8
$1500 < L_i$	5	1.0

如表 6 所示，设 L_i 为第 i 题官方解答字符串的长度。根据 L_i 的值，由短而长进行题目难度系数的划分，从而确定第 i 题的难度系数 r_i 和权重 w_i。

设 s_i 为模型在第 i 题的原始得分。则第 i 题的加权得分记作 S_i，计算如下：

$$S_i = s_i * w_i$$

本轮基于 Round 1 的总加权得分，是所有题目加权得分的总和：

$$S_{\text{total}} = \sum_{i=1}^{n} S_i = \sum_{i=1}^{n} s_i \cdot w_i$$

其中 n 为题目总数。此加权得分旨在反映模型的实际解题能力和思维能力，尤其是在解决难题方面的能力。

3.2.3 Round 3

本轮我们在 Round 1 的基础上，保留基础分的设定，即填空、选择 5 分，解答 12 分。

多样性加分：针对每道题，模型每提供一种**正确且不同的**解题思路，就给予额外加分。建议根据题目的难度和解法的常见程度设置不同的加分分值。例如，对于一道难题，如果模型能

给出 3 种不同的正确解法，每种解法可以加 2-3 分；而对于一道简单的题目，即使模型给出了多种解法，每种解法也只加 1 分。

创造性加分：对于特别具有创造性或新颖性的解法，给予更高的额外加分。

在题目选择上，我们优先选择里具有多种解题思路的题目，避免选择那些答案唯一或解法单一的题目。同时为了保证评估深度，我们分批次选取 5 道具有代表性题目。

对于模型调用参数设置，我们将温度（temperature）参数设置得稍高一些（例如 0.7 - 1.0），可以鼓励模型生成更多样化的答案。最大回复长度（max_tokens)设置为 2000，确保模型有足够的空间完整地表达多种解题思路。

同时我们采用人工与 o1-preview 模型对模型的回复进行打分，从思维性，合理性，正确性，有效性等角度，对模型的回复进行打分。

3.2.4 Round 4

本轮我们将重点评估模型在处理任务时的响应时间。响应时间是指从请求发送到模型接收到完整响应之间的时间间隔。为了确保评分的公平性和准确性，我们采用以下步骤来计算每个模型的平均时间得分。

数据清洗：去除响应时间过长的数据，例如超过 99%分位数的数据。

通过以下公式计算得分：时间得分=(最小平均响应时间\模型的平均响应时间)×10。

该公式旨在对应答速度快的模型进行奖赏，对应答速度过慢的模型进行有效惩罚。

3.3 工具平台

SDK 调用统一：

由于各家大模型 API 调用所采用的 SDK 不同, 为了便于测试与统计, 我们采用了 OneAPI[10] 项目进行对 API Key 的统一管理，并统一使用 OpenAI-SDK[11]进行 API 调用。

对于部分服务限制，我们采用了分渠道、分线路、分节点的设计，构建了一套完整、可靠的 API 调用平台，以避免因网络原因导致对模型的调用异常。

OCR 工具识别：

在试题读取中，对于数学公式，物理公式，化学方程式等数据，原卷采用了大量的图片表达相关数据，对此，我们采用了 OCR 工具[12][13]识别成标准的 latex 表达式，使模型能正确理解

相关数据的含义。

四、 实验过程

在本实验中，我们挑选主流的纯文本大模型与多模态大模型，分别用开源数据集与自建数据集对所选模型们进行评估。本文将在 4.1 中详细介绍实验设置，在 4.2 中介绍纯文本大模型在开源数据集上的定量评估，在 4.3 中介绍多模态大模型在自建数据集上的定量评估，在 4.4 中进行实验误差分析。

4.1 实验建立

4.1.1 数据集的划分

GAOKAO-Bench 包含 1354 条数据，自建数据集中包含 1680 条数据。为了实现更快的评估和模型验证，我们进行了小批量测试，在 GAOKAO-Bench 中随机抽取选择题、填空题、解答题各为 12 道、6 道、5 道（仅数学存在填空题）构成一个最小批量 batch。在自建数据集中随机抽取选择题、填空题、解答题各为 20 道、4 道、6 道（仅数学），随机抽取选择题、解答题各为 12 道、6 道构成一个最小批量。

4.1.2 模型划分

我们将模型划分为两大类:

(a) 专注文本的大模型

qwen-plus,hunyuan-pro,moonshot,doubao,deepseek,llama3-70b,gpt-4o,yi-large-turbo, claude-3-5-sonnet-20240620,gemini-1.5-pro。

(b) 多模态大模型

qwen-vl-plus,hunyuan-vision,glm-4v-plus,yi-vision,gpt-4o,claude-3-5-sonnet-20240620, pixtral-12b-2409,gemini-1.5-pro。

4.2 LLMs 评估分析

4.2.1 提示词对 LLMs 正确率的影响

我们提供了两个版本的提示词,

(a) **直接输出答案**。例如，对于数学题“计算：2 + 3 * 4”，提示词为“计算以下表达式的值：2 + 3 * 4”。

(b) **逐步 CoT 推理。**对于同样的数学题，提示词为“请逐步解释你的推理过程，并最终给出答案：2 + 3 * 4”。 我们希望通过这样的提示词引导模型展现类似人类的思考逻辑，例如：先进行乘法运算，再进行加法运算。

表 7：两个版本的提示词对比

prompt(a)	prompt(b)

prompt = r"""
现在你正在参加高考，我将给你一些题目,你需要根据固定的 json 格式输出答案，请直接给出你的答案。
请严格按照以下 JSON 格式回答:
{
"answer": "你的答案"
}
对于选择题，请将选项（ABCD）放置在 `answer` 字段中,不要包含选项的内容,并且用""包裹你的答案。
样例:
{
"answer": "AB"
}
对于填空题, 请将你的答案放置在 `answer` 字段中,不能使用（\\u7ea6\\u90e8）表示内容。
样例
{
	prompt = r"""
	现在你正在参加高考，我将给你一些题目,你需要根据固定的 json 格式输出答案，你可以先解析题目然后分析得出你的准确答案。
	请严格按照以下 JSON 格式回答:
	{
	"analysis": "你的解析",
	"answer": "你的答案"
	}
	对于选择题，请将选项（ABCDE）放置在 `answer` 字段中,不要包含选项的内容,并且用""包裹你的答案。
	样例:
	{
	"analysis": "你的解析",
	"answer": "AB
	"}
	对于填空题，请将你的答案放置在 `answer` 字段中,不能使用（\\u7ea6\\u90e8）表示内容。
	样例

"answer": "光合作用是植物利用太阳能将二氧化碳和水转化为有机物的过程"
}
对于解答题, 请将你的答案放置在 `answer` 字段中, 并且你的回答要包含具体的解题步骤。
{
"answer": "你的答案"
}
注意:
你的回答不能为空或者略, 否则将无法计算分数! ! !
回答仅限于上述 JSON 格式且单独一条 Json 内容! ! !
你的回答只能包含 JSON 格式的内容! !
不能包含任何解释性文字! ! !
answer 字段中只能为一串字符串（UTF-8）! ! !
你需要确保回答的格式正确并且能正确解析成 Json 对象，否则将无法计算分数。
"""
	{
	"analysis": "你的解析",
	"answer": "光合作用是植物利用太阳能将二氧化碳和水转化为有机物的过程"
	}
	对于解答题, 请将你的解析放置在 `analysis` 字段中，并且你的回答要包含具体的解题步骤，我们会按照你的解题步骤进行给分。
	{
	"analysis": "你的解析"
	}
	注意:
	你的回答不能为空或者略，否则将无法计算分数! ! !
	回答仅限于上述 JSON 格式且单独一条 Json 内容! ! !
	你的回答只能包含 JSON 格式的内容! !
	不能包含任何解释性文字! ! !
	answer 字段中只能为一串字符串（UTF-8）! ! !
	你需要确保回答的格式正确并且能正确解析成 Json 对象，否则将无法计算分数。
	"""

"让模型先思考" 的含义:

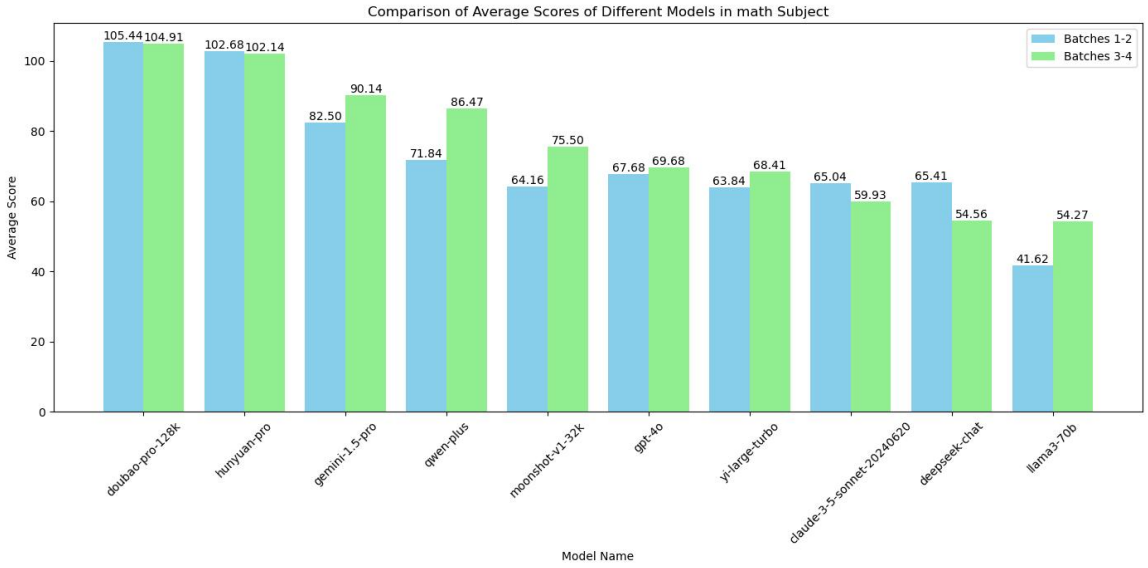
模拟人类认知过程: 人类在解决问题时，通常会先理解问题，然后进行分析、推理，最后得出结论。 “让模型先思考”的目标就是模拟这种认知过程，使模型的输出更符合人类的思维方式，更容易被理解和接受。

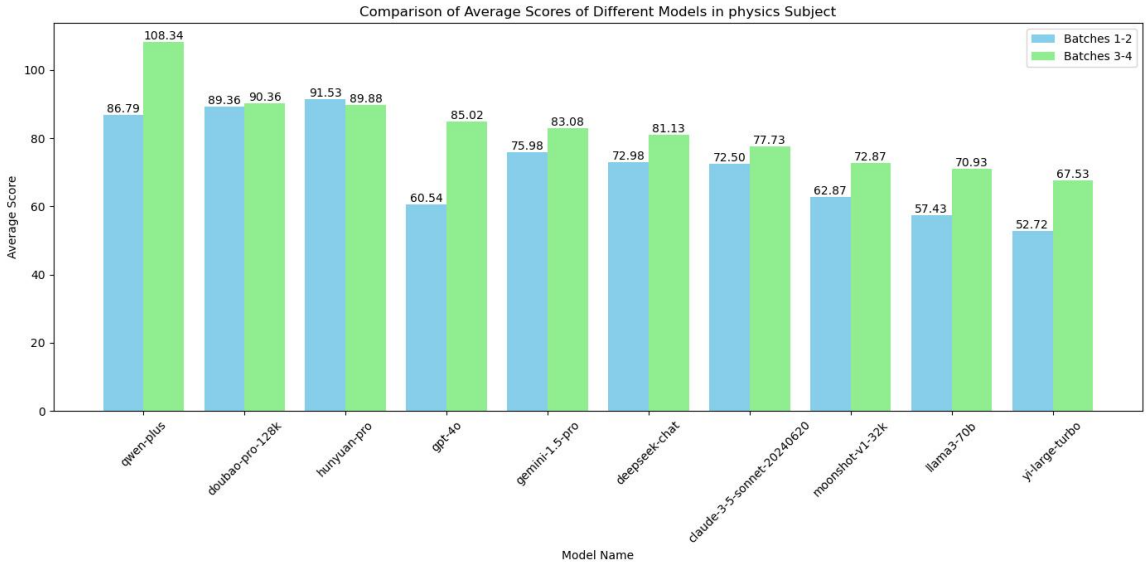
避免直接猜测: 如果直接要求模型输出答案，它可能会尝试直接猜测或基于统计规律进行推断，而没有进行深入的推理。 这在处理复杂问题时容易导致错误或不合理的答案。 “让模型先思考” 可以促使模型进行更严谨的逻辑推理，减少猜测的成分。

提高输出的可解释性: 当模型直接输出答案时，我们很难理解其背后的推理过程。 “让模型先思考” 可以引导模型输出中间步骤和推理过程，提高输出的可解释性，使我们更容易理解模型是如何得出最终答案的，从而更好地评估答案的可靠性。

增强模型的泛化能力: 通过 “让模型先思考”，可以帮助模型学习更通用的问题解决策略，而不是仅仅记住特定的输入输出模式。 这可以提高模型的泛化能力，使其能够更好地应对未见过的问题。

batchs1-2 为(a),batchs3-4 为（b）





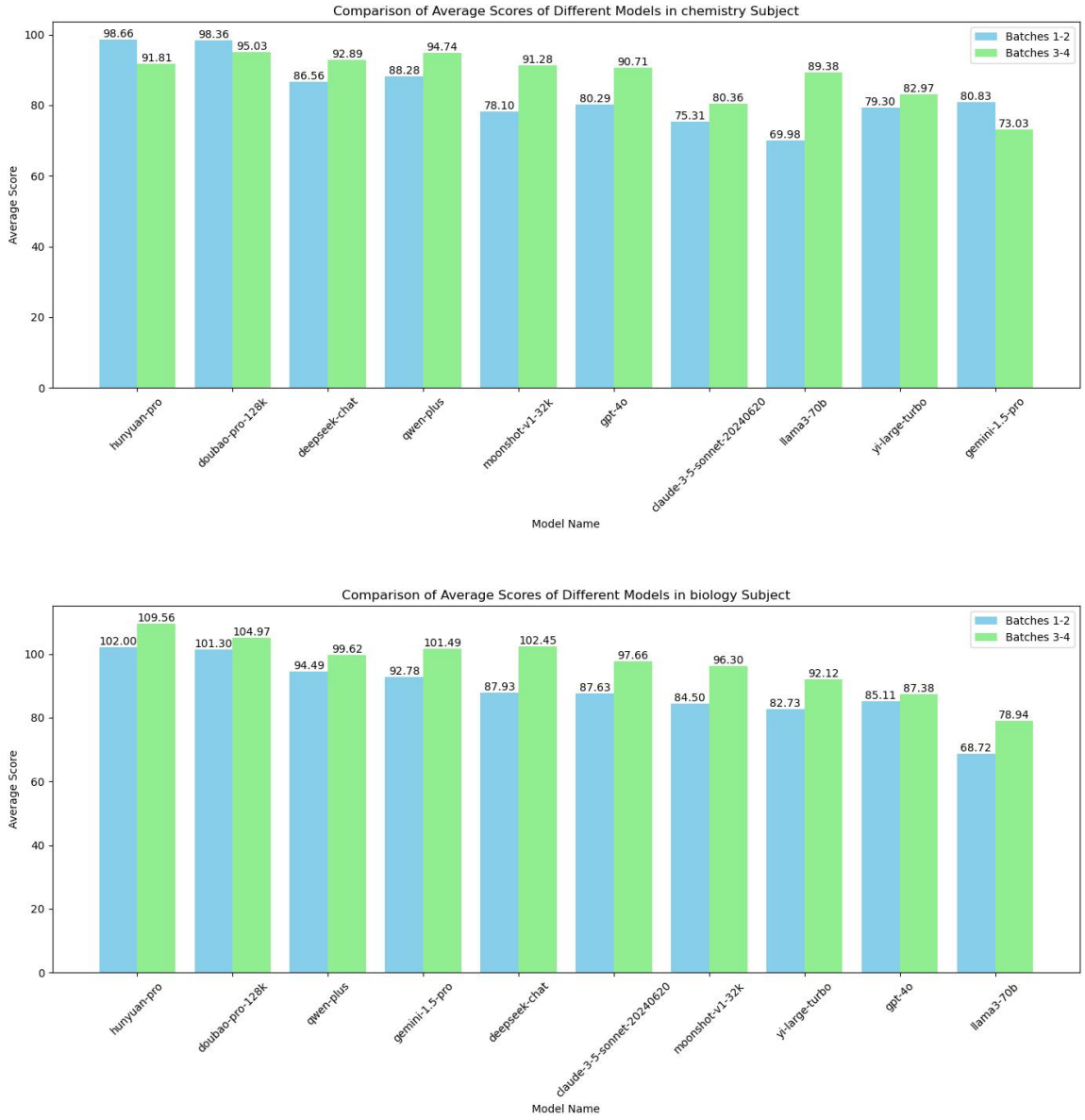


图 4：各学科在提示词优化前后模型正确率对比

总体观察:

在进行图表分析后，我们可以发现在优化提示词之后，几乎所有模型在使用优化后的提示词 (Batches 3-4) 后，平均得分都有所提高其中，gpt-4o 的提升效果最好，在化学数据集中提升达到了 40.43%。这表明提示词优化策略对提高模型在数学任务上的性能是有效的。

4.2.2 模型正确率横向对比

在该项中我们对两个版本提示词进行合并，得到各模型加权得分：

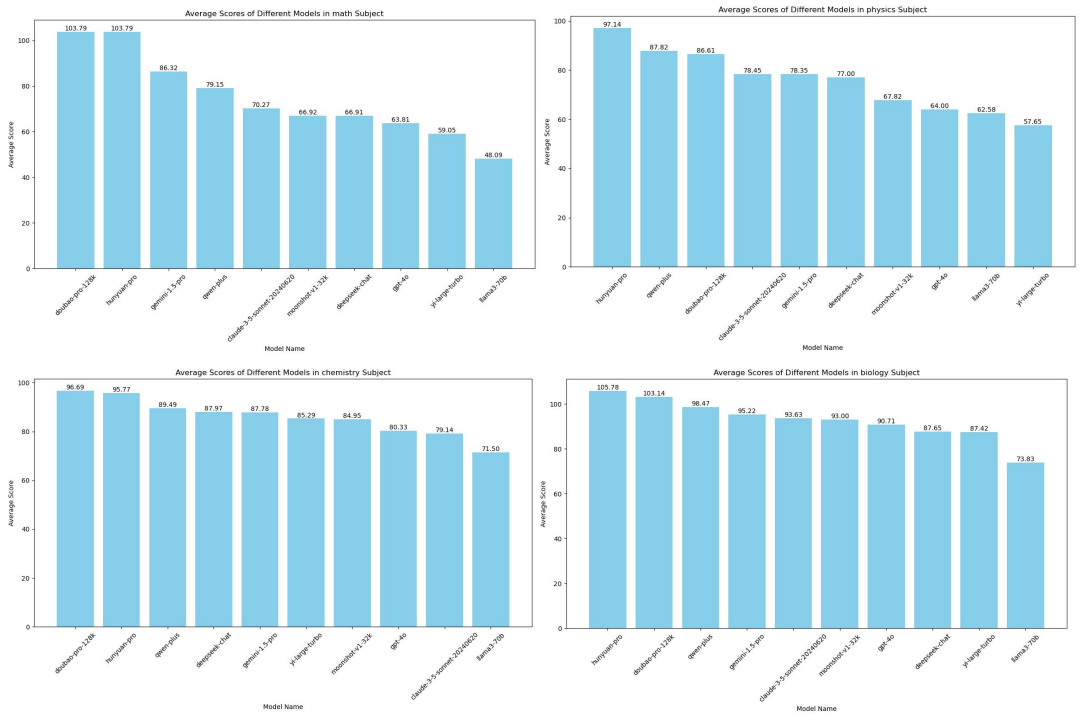


图 5：各学科模型得分对比图

总体观察:

hunyuan-pro 在所有四个学科中都表现出色，尤其是在物理和生物学科中名列前茅。doubao-pro-128k 也表现强劲，在数学和生物学科中得分最高。

4.2.3 响应时间分析

在本实验中，我们使用方差作为指标来衡量各模型的应答波动情况，旨在分析模型的稳定性（即应答时间对题目复杂程度的敏感度）以及适应性（即模型对不同类型任务的适应能力）。

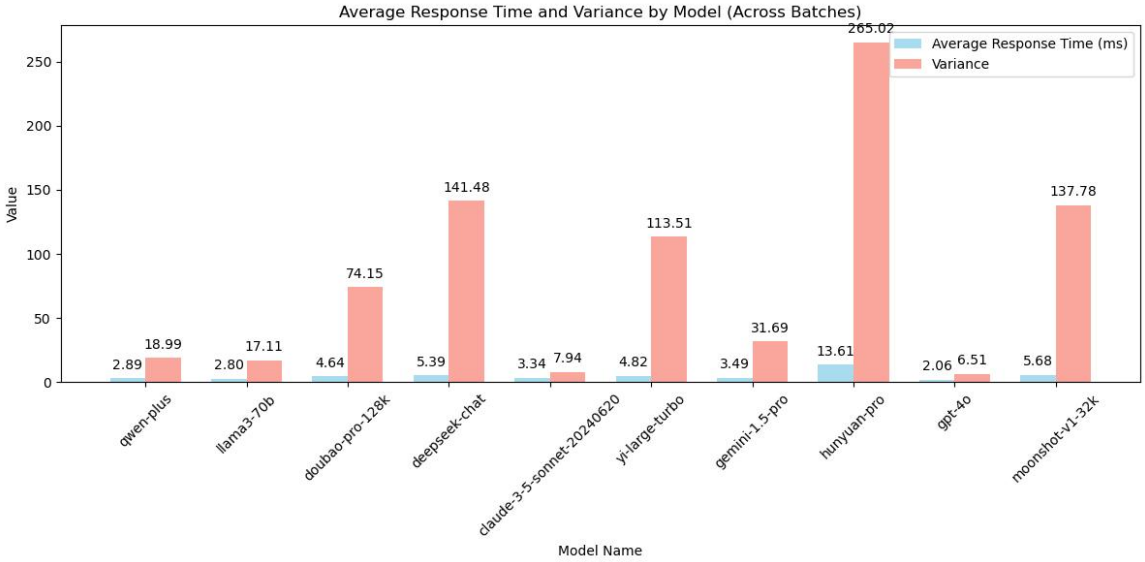


图 6：提示词优化前各模型平均响应时间和响应方差横向对比图

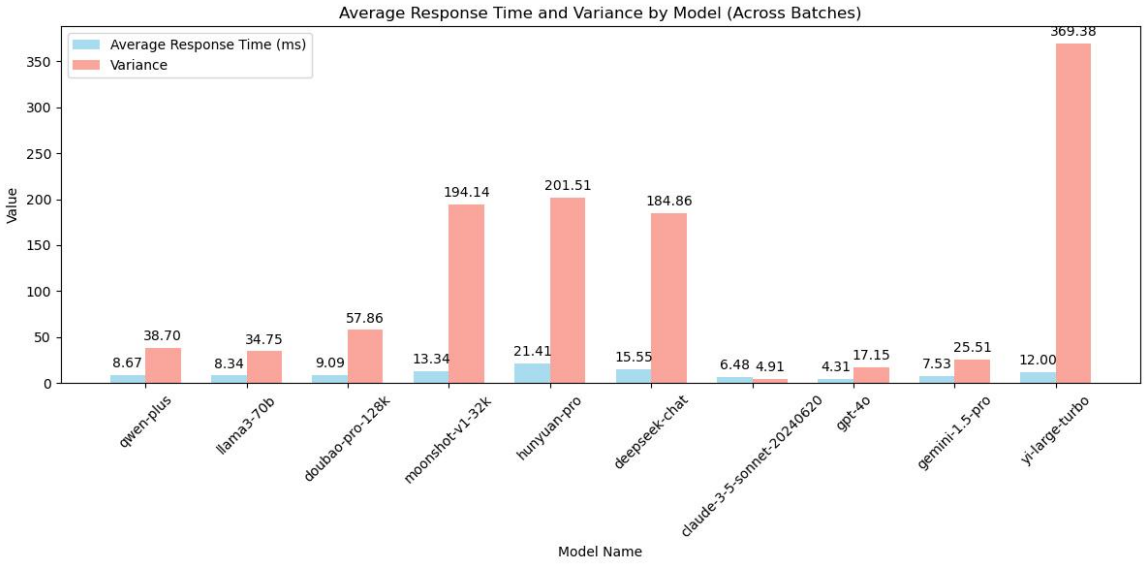


图 7：各模型在提示词优化后的平均响应时间与响应方差对比

在调整提示词前，各模型中平均应答时间和应答方差表现最佳的是 gpt-4o，表现最差的是 hunyuan-pro。具体而言，gpt-4o 的平均应答时间比 hunyuan-pro 快 6.6 倍，应答方差低 8.4 倍。这表明 gpt-4o 在稳定性和适应性方面具有显著优势，而 hunyuan-pro 在这两方面的表现相对较差。

在调整提示词后，平均应答时间表现最佳的模型为 gpt-4o，最差的为 hunyuan-pro，两者的平均应答时间相差 4.9 倍；应答方差表现最佳的是 claude-3-5-sonnet-20240620，最差的是 yi-large-turbo，二者的应答方差相差高达 75.2 倍。这表明 yi-large-turbo 在应对复杂逻辑推理任

务时的稳定性较弱，而 gpt-4o 在整体表现中处于领先地位。

4.2.4 综合加权得分分析

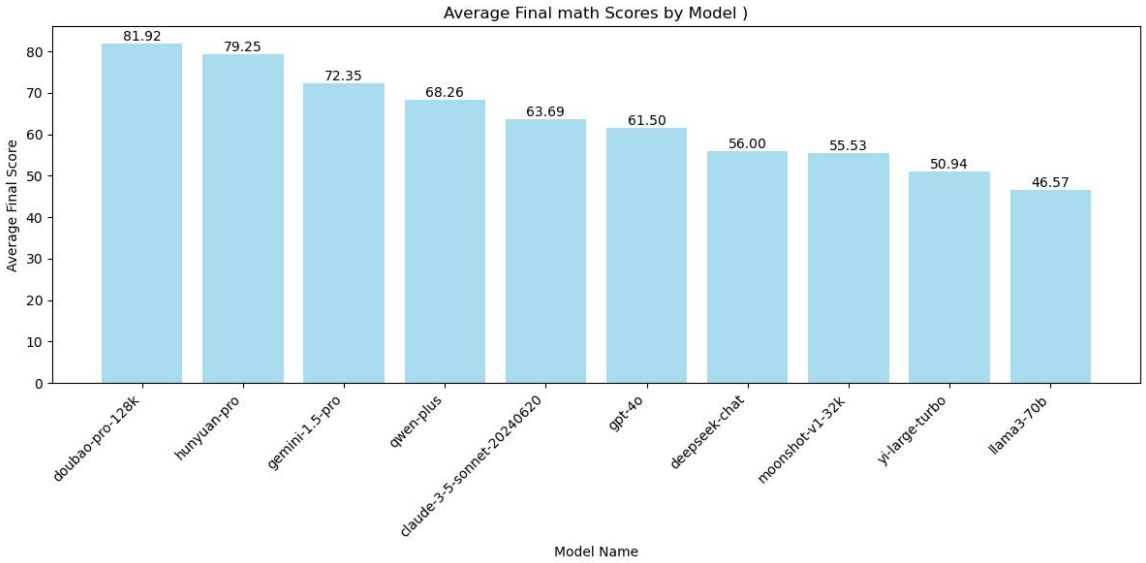


图 8：数学综合得分对比

数学排名：

doubao-pro-128k>hunyuan-pro>gemini-1.5-pro>qwen-plus>claude-3-5-sonnet-20240620>gpt-4o>deepseek-chat>moonshot-32k>yi-large-turbo>llama3-70b

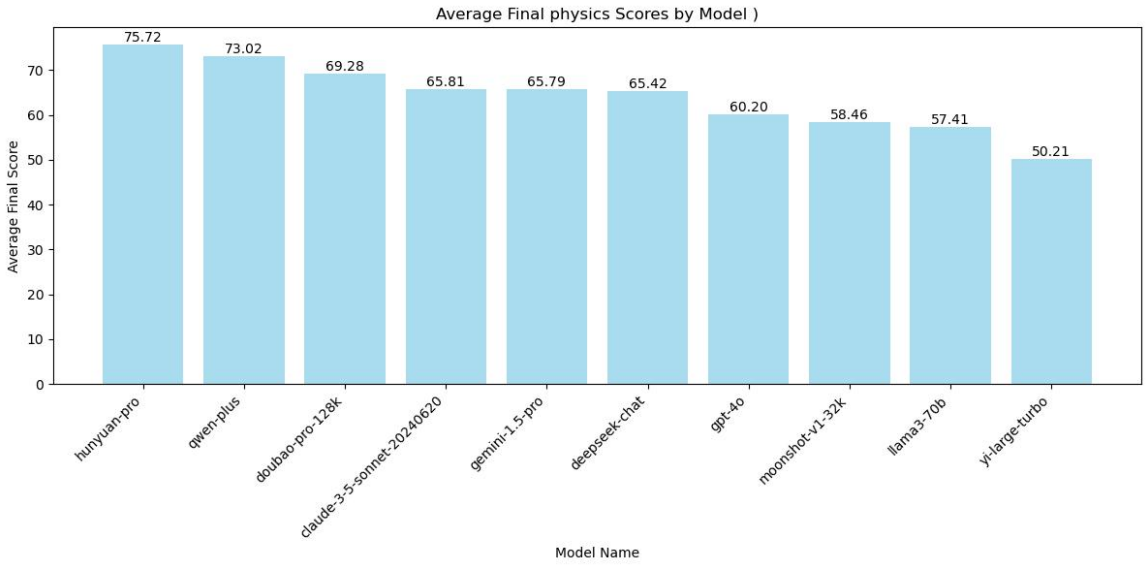


图 9：物理综合得分对比

物理排名：

hunyuan-pro>qwen-plus>doubao-pro-128k>claude-3-5-sonnet-20240620>gemini-1.5-pro>deepseek-chat>gpt-4o>deepseek-chat>moonshot-32k>llama3-70b>yi-large-turbo

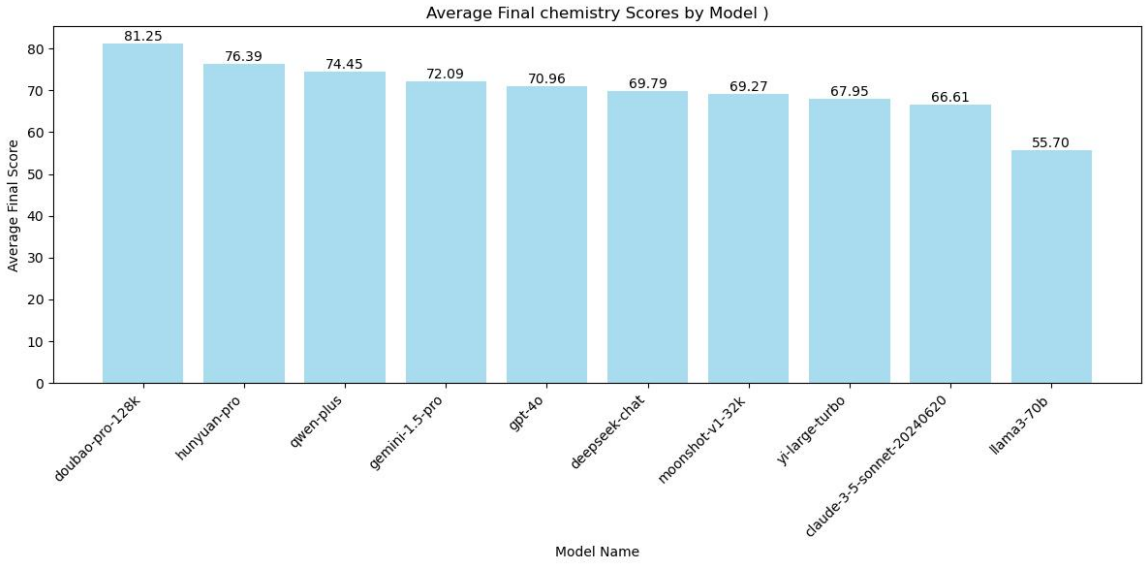


图 10：化学综合得分对比

化学排名：

doubao-pro-128k>hunyuan-pro>qwen-plus>gemini-1.5-pro>gpt-4o>deepseek-chat>moonshot-32k>yi-large-turbo>claude-3-5-sonnet-20240620>llama3-70b

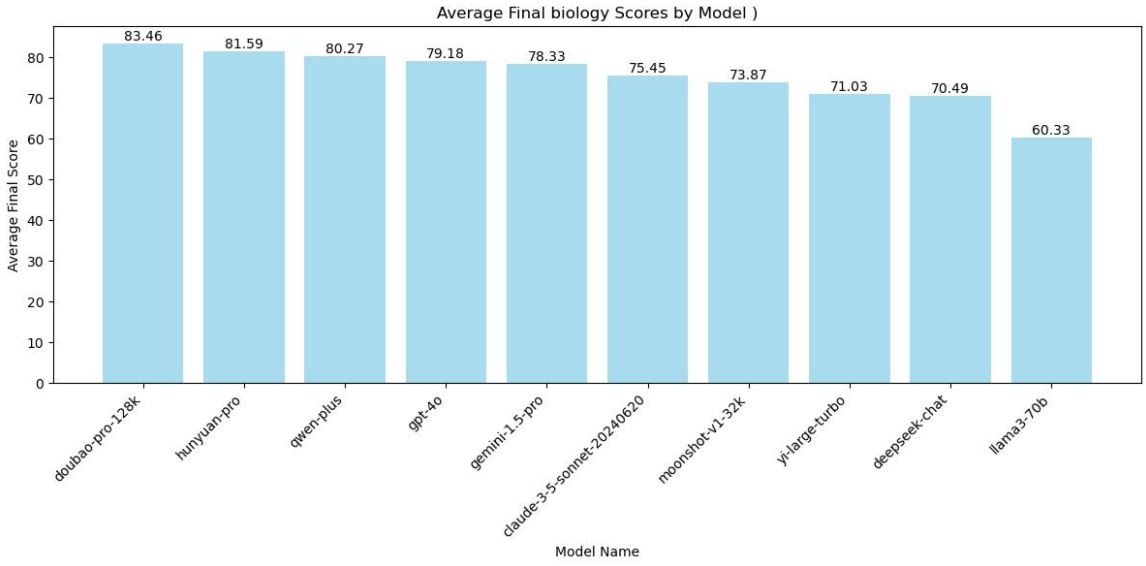


图 11：生物综合得分对比

生物排名：

doubao-pro-128k>hunyuan-pro>qwen-plus>gpt-4o>gemini-1.5-pro>claude-3-5-sonnet-20240620>moonshot-32k>yi-large-turbo>deepseek-chat>llama3-70b

4.3 MM-LLMs 评估分析

4.3.1 提示词对 MM-LLMs 准确率的影响

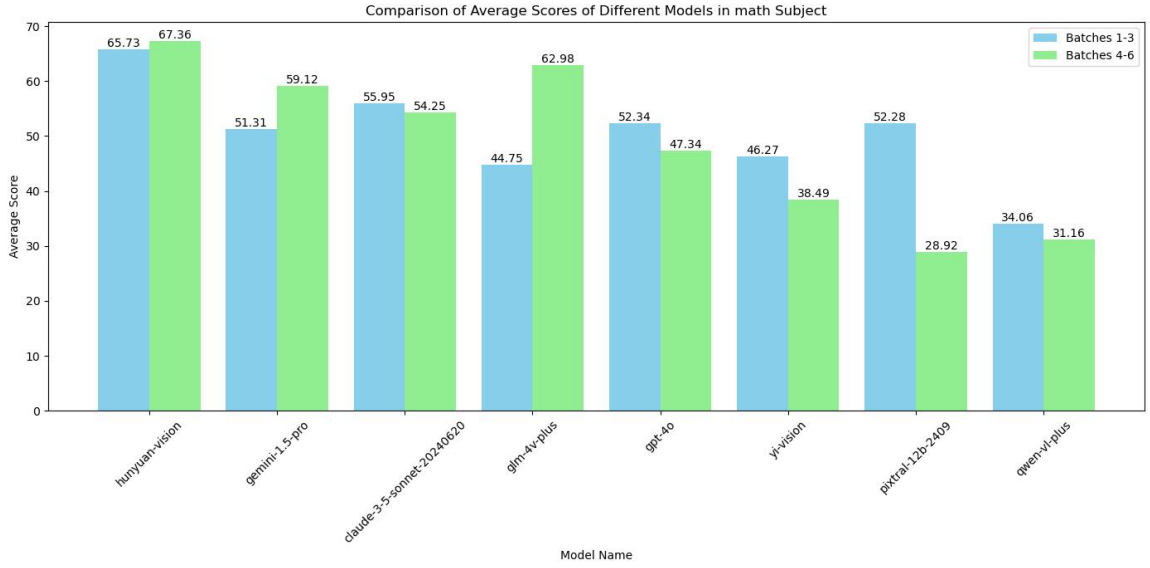
在多模态大语言模型的测试上，我们同样提供了两个版本的提示词。

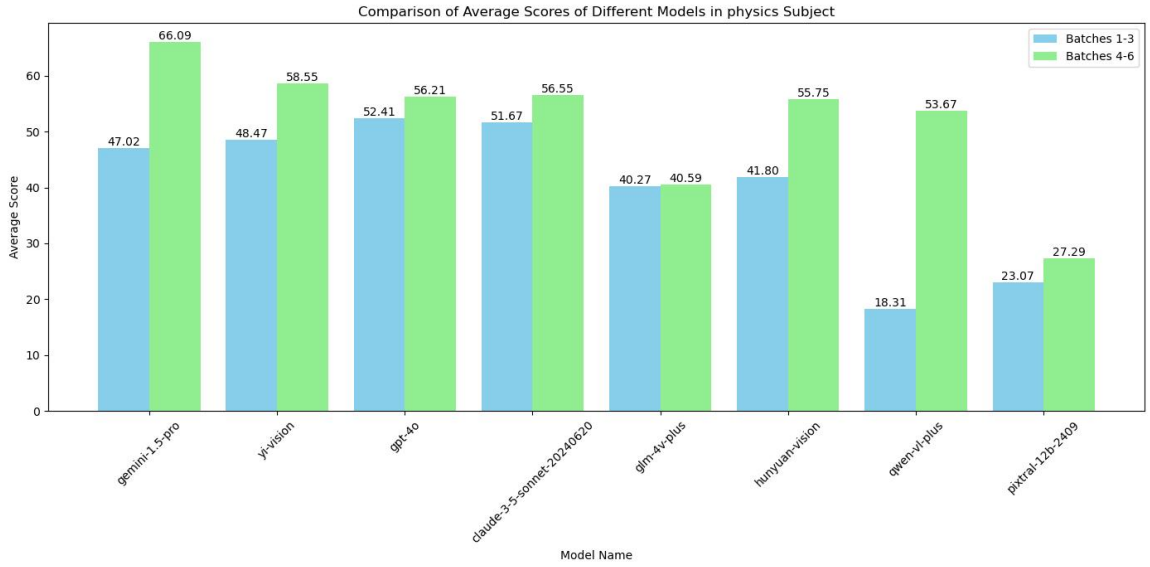
(a) 要求模型直接输出答案。

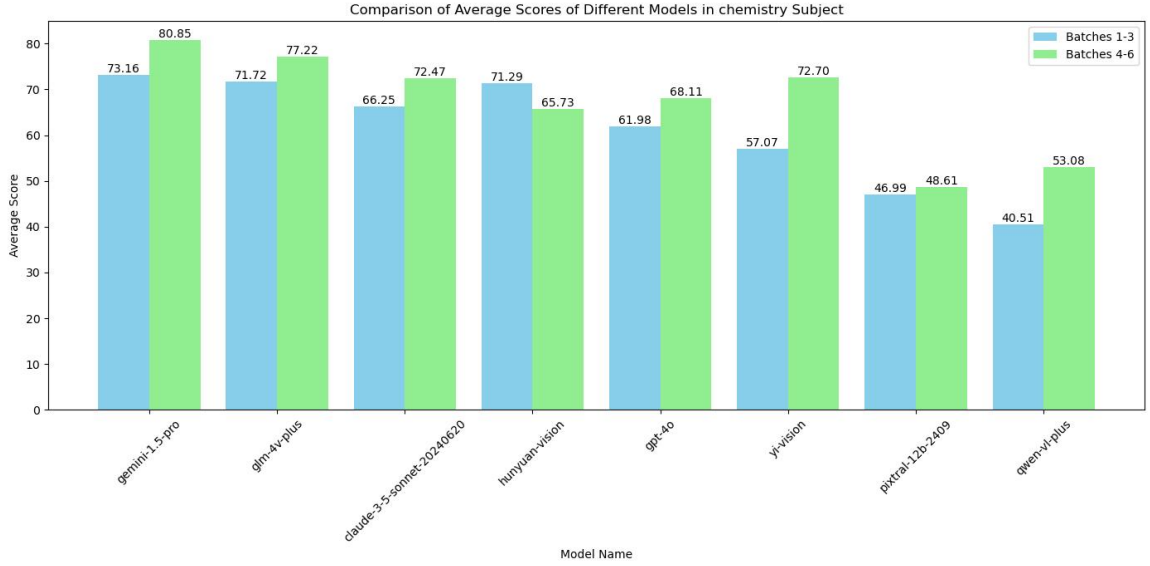
(b) 要求模型细微的、逐步的 CoT 推理，先给出解析，再给出答案。

batchs1-3 为(a)

batchs4-6 为(b)




Comparison of Average Scores of Different Models in physics Subject
Batches 1-3
Batches 4-6
Average Score
Model Name
gemini-1.5-pro
yi-vision
gpt-4o
claude-3-5-sonnet-20240620
glm-4v-plus
hunyuan-vision
qwen-vl-plus
pixtral-12b-2409
47.02
66.09
48.47
58.55
52.41
56.21
51.67
56.55
40.27
40.59
41.80
55.75
18.31
53.67
23.07
27.29


Comparison of Average Scores of Different Models in chemistry Subject
Batches 1-3
Batches 4-6
Average Score
Model Name
gemini-1.5-pro
glm-4v-plus
claude-3-5-sonnet-20240620
hunyuan-vision
gpt-4o
yi-vision
pixtral-12b-2409
qwen-vl-plus
73.16
80.85
71.72
77.22
66.25
72.47
71.29
65.73
61.98
68.11
57.07
72.70
46.99
48.61
40.51
53.08

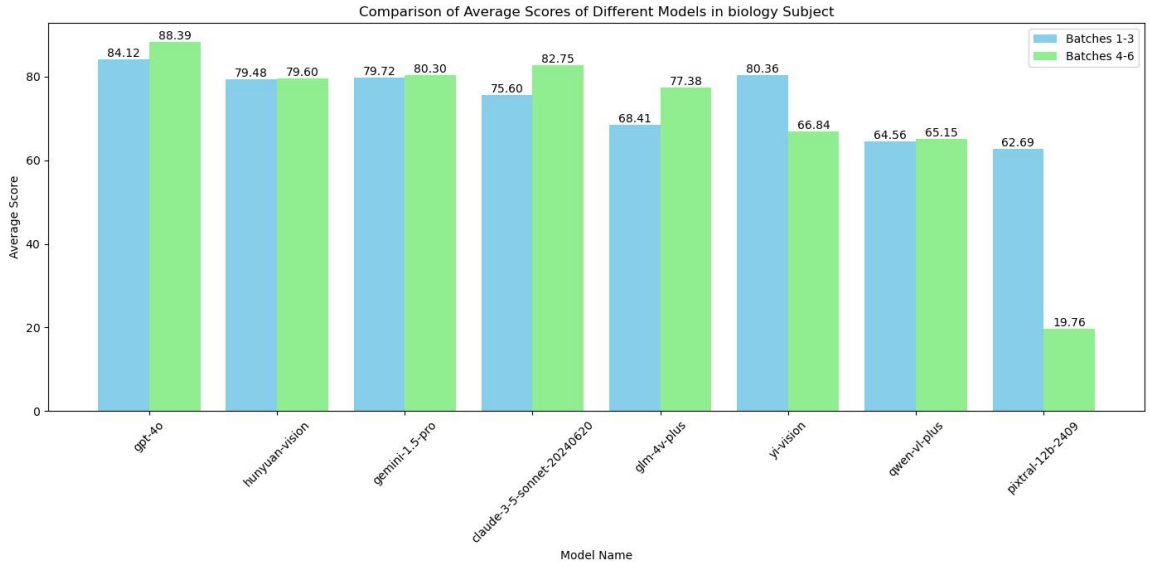
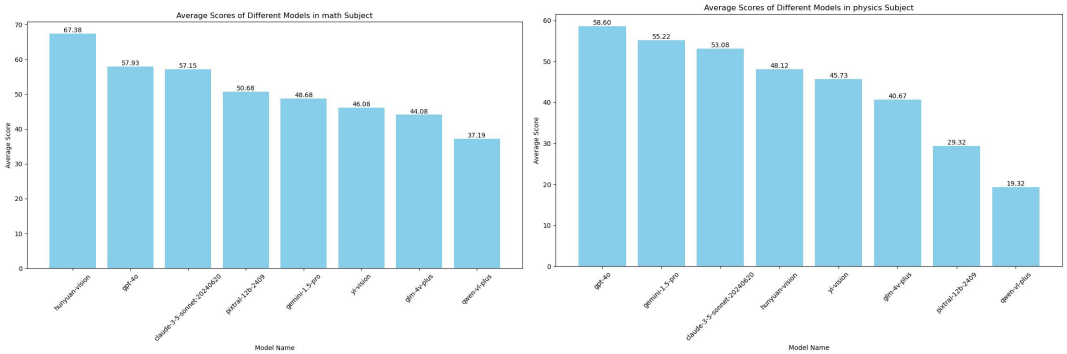


图 12：各学科在提示词优化前后模型准确率对比

我们通过图 12 分析可以发现，提示词优化策略对提升大多数大型语言模型在数学、物理、化学和生物学科上的准确率有积极作用。然而，pixtral-12b-2409 模型在数学、物理和生物三个学科上的表现出现了显著下降，尤其在生物学科中下降幅度高达 68.47%，这可能与其处理复杂问题的能力不足，以及在思维链推理过程中误差积累有关。其他模型在使用优化后的提示词后，准确率普遍有所提高，但提升幅度因模型而异。

4.3.2 模型准确率横向对比

我们同样对两个版本提示词进行合并，得到各模型加权得分



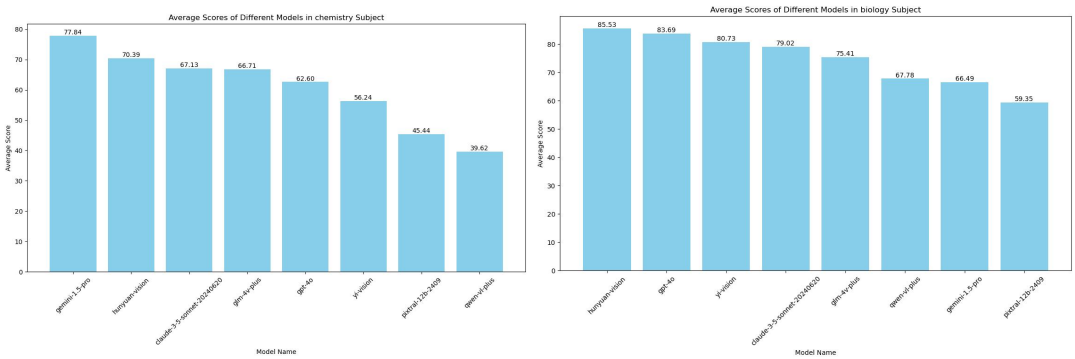


图 13：各学科模型得分对比图

由图 13 可知，综合加权之下，**hunyuan-vision 和 gpt-4o 是综合表现最好的两个模型。**hunyuan-vision 更擅长生物和数学，而 gpt-40 在物理和化学上更胜一筹。我们在分析数据后发现，qwen-vl-plus 存在严重的幻觉现象。

4.3.3 响应时间分析

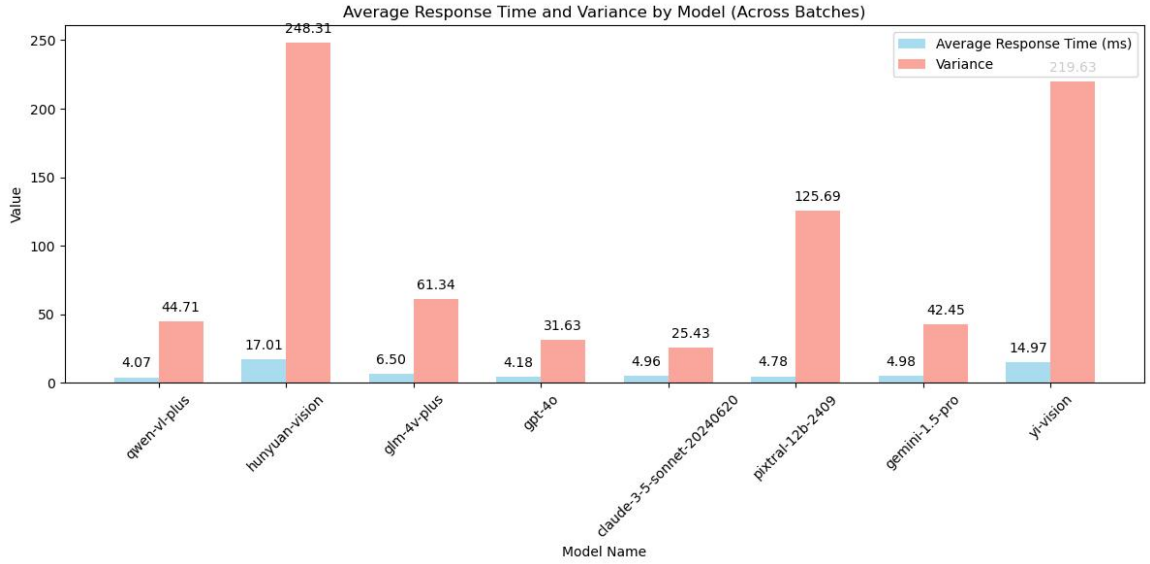


图 14：提示词优化前多模态模型平均响应时间和响应方差横向对比图

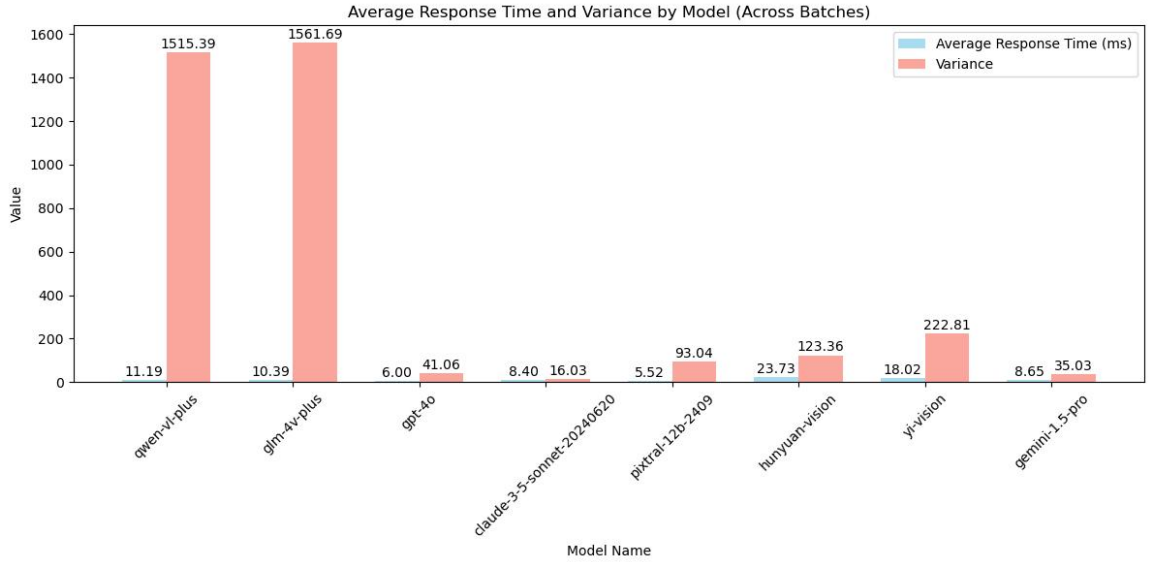


图 15：提示词优化后多模态模型平均响应时间和响应方差横向对比图

在调整提示词前，平均应答时间表现最佳的是 gpt-4o，表现最差的是 hunyuan-pro。具体而言，gpt-4o 的平均应答时间比 hunyuan-pro 快 4.1 倍。应答方差表现最佳的是 claude-3-5-sonnet-20240620，最差的是 hunyuan-pro，应答方差低 9.8 倍。这表明 gpt-4o 与 claude-3-5-sonnet-20240620 在可视化中的稳定性和适应性方面具有显著优势，而 hunyuan-pro 在这两方面的表现相对较差。

在调整提示词后，平均应答时间表现最佳的模型为 claude-3-5-sonnet-20240620，两者的平均应答时间相差 4.9 倍；应答方差表现最佳的是 claude-3-5-sonnet-20240620，qwen-vl-plus 与 glm-4v-plus 出现了巨大的波动。这表明 qwen-vl-plus 与 glm-4v-plus 在应对复杂逻辑推理任务时的稳定性较弱，而 claude-3-5-sonnet-20240620，gpt-4o 在整体表现中处于领先地位。

4.3.4 综合加权得分分析

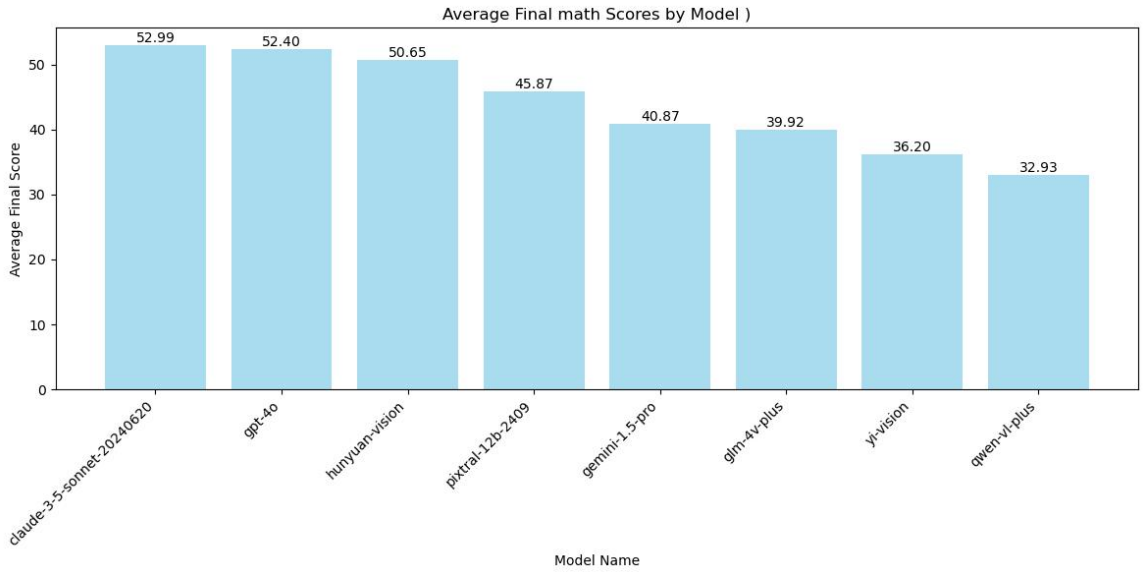


图 16：数学综合得分对比

数学排名：

claude-3-5-sonnet-20240620>gpt-4o>hunyuan-vision>pixtral-12b-2409>gemini-1.5-pro>glm-4v-plus>yi-vision>qwen-vl-plus

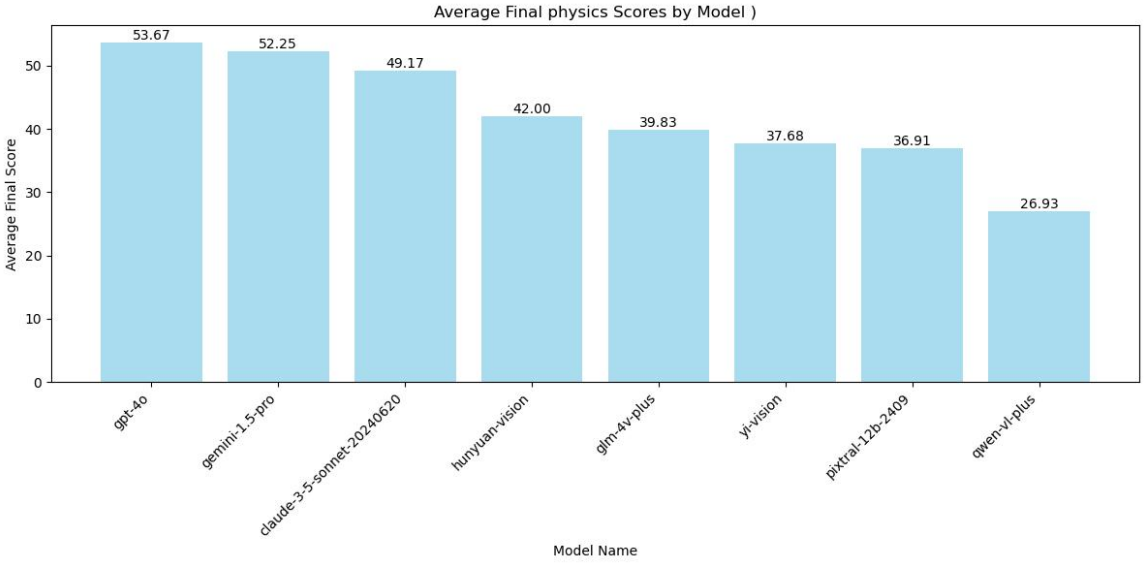


图 17 物理综合得分对比

物理排名：

gpt-4o>gemini-1.5-pro>claude-3-5-sonnet-20240620>hunyuan-vision>glm-4v-plus>yi-vision>pi

xtral-12b-2409>qwen-vl-plus

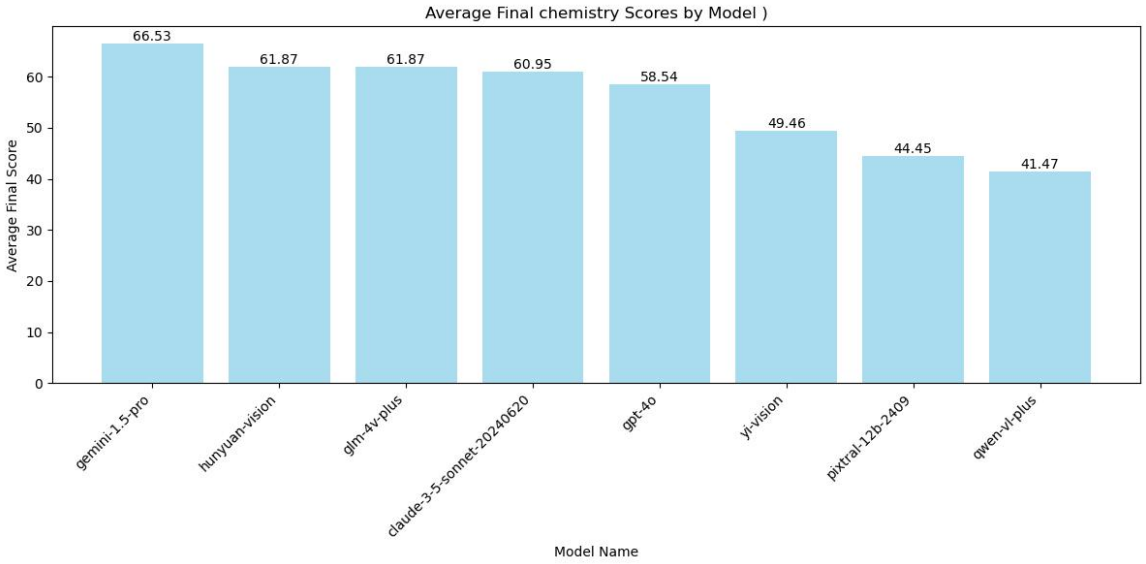


图 18：化学综合得分对比

化学排名：

gemini-1.5-pro>hunyuan-vision>glm-4v-plus>claude-3-5-sonnet-20240620>gpt-4o>yi-vision>pixtral-12b-2409>qwen-vl-plus

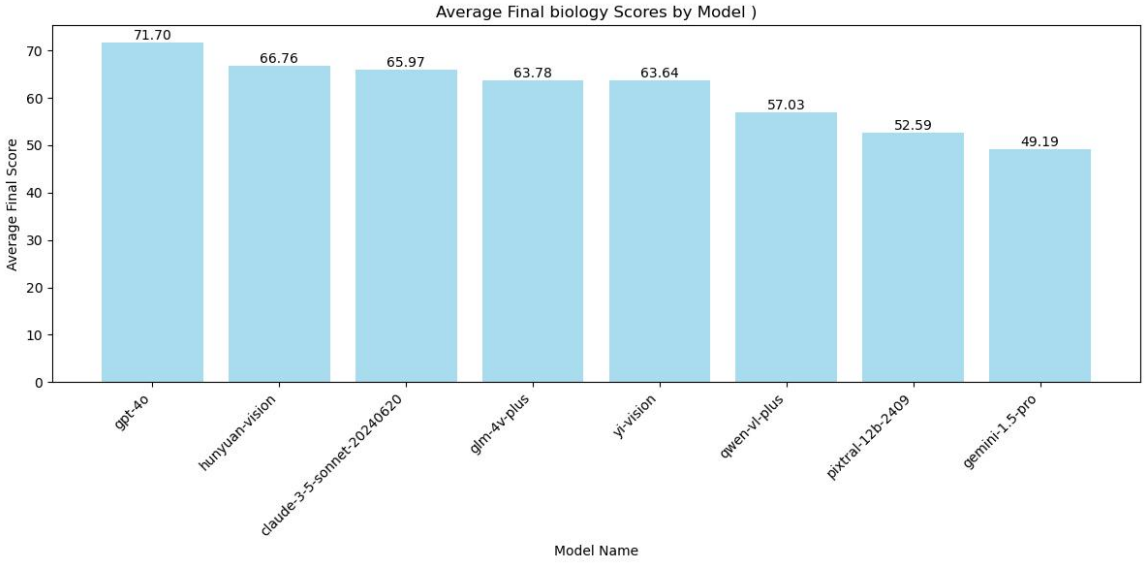


图 19：生物综合得分对比

生物排名：

gpt-4o>hunyuan-vision>claude-3-5-sonnet-20240620>glm-4v-plus>yi-vision>qwen-vl-plus>pixt

ral-12b-2409>gemini-1.5-pro

4.3.5 模型一题多解能力

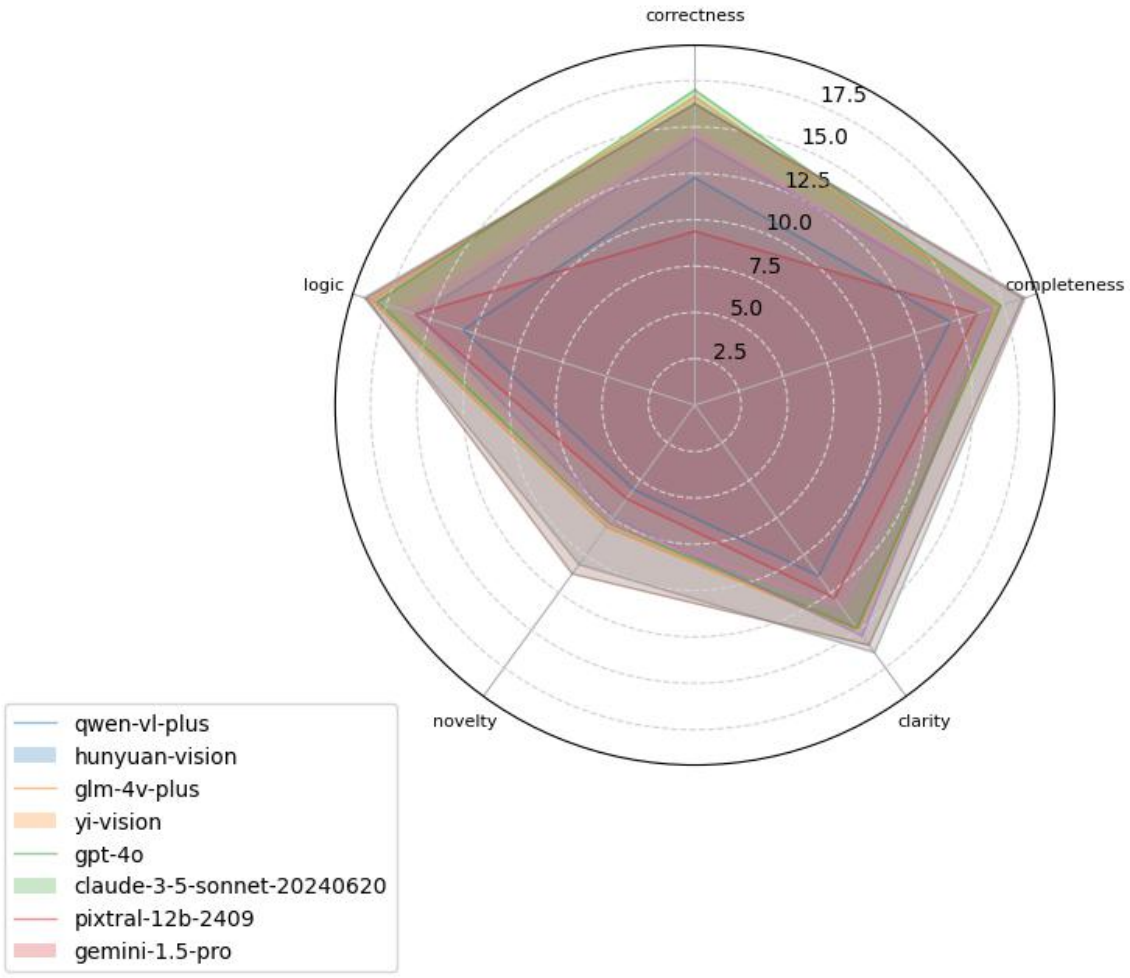


图 20：模型一题多解能力多维度分析

由图 20 雷达图可知，可以看出 Gemini-1.5-pro 模型在五个维度上都显著领先于其他模型，尤其是在正确性和完整性方面。其他模型的性能各有优劣，例如 gpt-40 在逻辑性方面表现较好，而 pixtral-12b-2409 在清晰度上略胜一筹。总体而言，Gemini-1.5-pro 在此项评估中展现出最佳的综合性能。

4.4 误差分析

4.4.1 数据集误差

本实验中数据集的公式部分由 OCR 光学识别得到，虽然通过 gpt-4o-mini 进行多次可用性判断，但数据集中仍然不可避免地会出现不可用的数据，可能降低大模型的正确率。

4.4.2 评分误差

本实验评分为人工校验配合 gpt-4o-mini 对模型作答进行评分，采用多次求取平均值作为最终得分，但是仍然不可避免地会出现一定偏差，从而造成得分的波动。

4.4.3 应答时间误差

本实验中虽通过统一项目 OneAPI 统一管理各模型调用，但是因不同时间段造成的网络波动不同，仍可能影响最终得到的应答时间。

五、 总结与展望

本研究系统评估了大语言模型 (LLMs) 在解答高中理科题目中的表现，并对模型的准确率、逻辑性、思维性、多解能力、创造性和响应速度等多维度进行了量化分析。通过构建多轮评测流程，我们揭示了不同 LLMs 在高中理科解题任务中的表现差异，并提出了有针对性的改进策略。

研究结果表明，提示词优化在提高模型正确率上具有显著成效，不同模型在不同学科题目上的表现也存在明显差异。这些发现不仅有助于模型在教育领域的实际应用，也为 LLM 在解题能力上的提升提供了可靠的数据支撑。

5.1 关键发现

1. 模型优化潜力：

提示词优化对于模型的准确率提升效果显著，特别在逻辑性和创造性要求较高的题目中尤为明显。本研究的多维度评分法有效揭示了模型在思维性和多解能力方面的表现差异，为进一步优化提示词设计和模型微调提供了方向。

2. 学科表现差异：

各模型在不同学科领域的表现存在显著差异，尤其在物理和化学等推理性较强的题目中，模型的表现有所不足。通过对各学科题目特点的深入分析，本研究为针对不同学科题目的模型选择和优化提供了量化依据。

3. 响应时间与评分方法：

在响应时间上，不同模型之间的差异也为教育技术应用中的模型部署提供了参考。在评估方法上，通过引入加权评分、思路多样性评分，本研究建立了更适合评测多解能力与实际解题能力的评分方式，使得评测结果更具说服力和实用性。

5.2 未来展望

1. 多学科与多层级适用性研究：

随着 LLM 在教育领域的应用逐渐拓展，未来可以进一步探索模型在不同学科和多年龄层教育内容上的适用性。特别是在高等教育、职业教育中的应用，将为大规模个性化学习提供可能性。

2. 逻辑推理与创造性提升：

当前 LLMs 在一些高要求逻辑推理和创造性任务上表现有限，未来可以尝试结合提示词优化和微调策略，增强模型的思维深度和创新能力。同时，可以探讨融合多模态信息（如图表和公式）来提升模型的综合解题能力。

3. 教育技术与智能学习系统的构建：

本研究成果为智能答疑系统和个性化学习系统的开发提供了重要参考。未来可以考虑将模型应用于课堂辅助、在线作业批改、以及智能辅导系统等实际场景，为教师和学生提供更智能化的学习支持。此类应用不仅有助于提升学生的学习效果，也有助于教育资源的公平分配和普及。

4. 数据集和评测方法的优化：

本研究还发现，数据集和评分误差对实验结果具有一定影响，因此未来可以尝试通过扩展题库和优化评分体系，进一步提高评测结果的可靠性。更完善的数据集和更加标准化的评分方法将为模型性能的客观评估和持续改进奠定基础。

六、 创新点介绍

6.1 多维度、多方法量化评测体系

本研究设计的多维度评测方法涵盖了正确率、逻辑性、思维性、多解能力、创造性和响应速度等方面，超越了以往研究中的单一指标评估。尤其是加入多解能力评分，创新性地量化了 LLM 在一题多解和创造性表现上的能力，为模型的综合评估设立了新标准。

6.2 提示词、调参优化模型表现

在提示词设计上进行了优化实验，探索了不同提示词策略对模型解题表现的影响。本研究不仅量化了提示词在不同题型中的效果差异，还对优化策略进行了归纳总结，提出了可推广的提示词优化准则。为后续研究中提示词设计提供了数据支持，推动 LLM 在教育、答疑系统中的有效应用。提示词优化也为模型调优提供了一种轻量级的实现方式。

研究成果不仅评估了模型当前的表现，还从个性化学习的角度提出了系统设计的建议，包括智能答疑系统、题目批改、和学习路径推荐等应用。

附录

[1] Beautiful Soup is a library that makes it easy to scrape information from web pages. It sits atop an HTML or XML parser, providing Pythonic idioms for iterating, searching, and modifying the parse tree. https://pypi.org/project/beautifulsoup4/#description

[2] 是一种基于 TeX 的排版系统，由美国计算机科学家莱斯利·兰伯特在 20 世纪 80 年代初期开发，利用这种格式系统的处理，即使用户没有排版和程序设计的知识也可以充分发挥由 TeX 所提供的强大功能，不必一一亲自去设计或校对，能在几天，甚至几小时内生成很多具有书籍质量的印刷品生成复杂表格和数学公式，这一点表现得尤为突出。因此它非常适用于生成高印刷质量的科技和数学、物理文档。引用自 https://zh.wikipedia.org/zh-cn/LaTeX

[3] latex 数据训练集
GSM8K
访问网址：https://huggingface.co/datasets/openai/gsm8k
GSM8K 数据集为大模型数学能力评测常见的 benchmark 之一，题目类型为小学数学题，共 8.5k 样本。
GSM8K 是大小为 8.5K 的小学数学数据集，涉及基本算术运算，需要 2-8 个步骤才能解决，包含 7.5K 训练集和 1K 测试集。GSM8K 每道题的答案包含完整的解题过程，有助于 CoT 训练。在 GSM8K 数据集中，测试数据的正确答案在 #### 后面，而且一般为整数。GSM8K 数据集的测试集共 1319 条。
MATH
访问网址：https://huggingface.co/datasets/lighteval/MATH/viewer
MATH 数据集为大模型数学能力评测常见的 benchmark 之一，是一个包含 12500 个高中数学竞赛的问题（7500 个用于训练，5000 个用于测试）的数据集，以文本模式的 Latex 格式呈现。
MATH 中的每个问题都有一个完整的逐步解决方案，有助于 CoT 训练。
MATH 数据集的最终答案用包围起来，答案表述一般比较多变，形式多样，有分数，根式，多项式，整数，小数等等。在其官网 Github 项目 hendrycks/math 中，提供了用于判断两个最终答案是否相等的代码脚本 math_equivalence.py ，网址为：
https://github.com/hendrycks/math/blob/main/modeling/math_equivalence.py.
Ape210K
访问网址：https://github.com/Chenny0808/ape210k
Ape210K 数据集由猿辅导开源共享，它是一个新的大规模和模板丰富的 Math Word Problem（MWP）数据集，包含 210,488 个（约 210K）中国小学水平的数学问题。每个问题都包含最佳答案和得出答案所需的方程式。

[4] https://platform.openai.com/docs/guides/vision#uploading-base64-encoded-images

[5] https://huggingface.co/datasets/MoryForCola/CG2401-Dataset/tree/main 本文题目集开源仓库地址

[6] https://github.com/OpenLMLab/GAOKAO-Bench GAOKAO-Bench 是一个以中国高考题目为数据集，测评大模型语言理解能力、逻辑推理能力的测评框架。

[7] https://huggingface.co/datasets/openai/gsm8k

[8] https://huggingface.co/datasets/lighteval/MATH/viewer

[9] prompt = '''
你正在批改高考试卷，我将给你提供学生答案和正确答案，请你根据答案和题目类型，给出分数，答案中连续的##代表或者，即两个答案都正确。
如果是填空题，请直接对比答案，给出分数。形式不同答案相同仍然给分。满分 5 分，如有两个空，第一空正确给 2 分，第二空正确给 3 分。
"如果是解答题，请根据答案和解析按解题步骤给出分数。满分 12 分。对应步骤过程及该步骤答案正确得分，错误不得分。"
注意回答你的输出只需给出分数即可，格式为：{"score":xx}，xx 为分数，注意不要回答其他内容。
'''

[10] *通过标准的 OpenAI API 格式访问所有的大模型，访问网址：*
https://github.com/songquanpeng/one-api

[11] OpenAI 官方文档，访问网址：
https://platform.openai.com/docs/quickstart

[12] Using a ViT to convert images of equations into LaTeX code.
引用至 https://github.com/lukas-blecher/LaTeX-OCR

[13] Large scale training of Latex formula recognition model, currently being organized and open source. 引用至： https://github.com/chaodreaming/Simple-LaTeX-OCR