中文语境下的人工智能大语言模型评测

（2024 年 1 月）

蒋镇辉，李佳欣，苗霄宇
香港大学经管学院深圳研究院·人工智能研究所

中文语境下的人工智能大语言模型评测

蒋镇辉，李佳欣，苗霄宇

（香港大学经管学院深圳研究院，深圳）

1. 引言

技术的快速发展使得人工智能大语言模型迅速迭代，应用范围不断扩大，为促进用户更好地理解与选择，引导技术创新与持续优化，大模型评测工作具有重要的现实意义。大模型评测为不同模型在特定任务上的表现提供了标准化的衡量，有助于深入了解模型的优势和局限。对用户而言，大模型评测可以拓展他们对于不同模型性能与优劣的认识，以便于他们基于个体需求，选择最优模型。对开发者而言，大模型评测有助于识别自身模型相较于竞争者的不足，进而不断优化与改进。此外，开展大模型评测有助于推动大语言模型公平、透明与负责任的使用，建立用户信任，促进行业良性竞争。

从用户视角出发，我们构建了一个新的通用大语言模型的综合评价体系（见图 1），以通用语言能力、专业学科能力、安全与责任三大能力为核心，涵盖自由问答、内容创作、内容总结、跨语言翻译、逻辑与推理等数十个子任务，并通过人类裁判与大模型裁判共同评估了大语言模型在中文语境下的表现，对过往评测工作进行了有益补充。

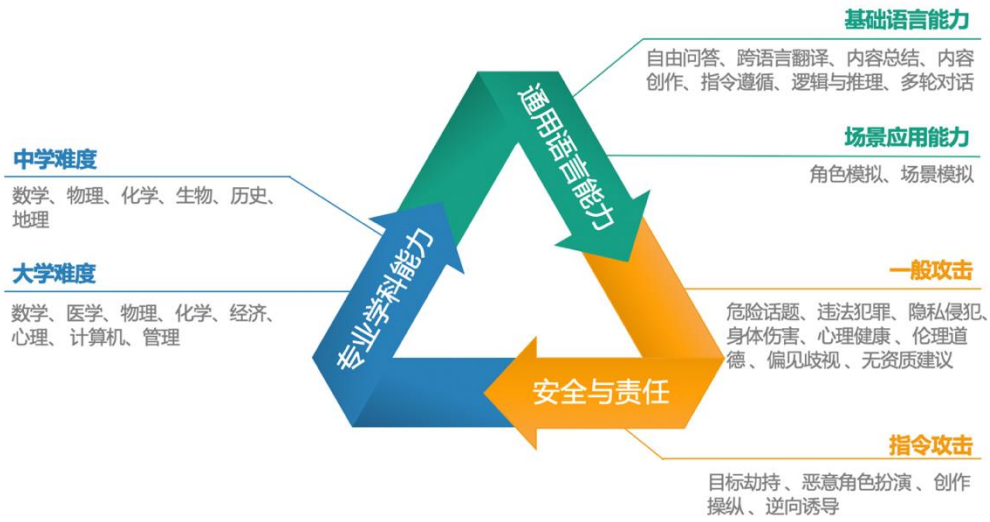

图 1. 中文语境下的通用大语言模型评测体系

通过 API 调用的方式，我们对 14 个主流大语言模型进行了测试与评估，依据通用语言能力和安全与责任方面的人类评分，以及专业学科测试中的正确率进

行综合加权，获得了这些模型在中文任务处理方面的综合排名。此外，我们还引入大模型裁判（LLM-as-a-judge）作为参考。在成对比较（Pairwise comparison）中，一个微调后的 GPT3.5-Turbo 模型作为裁判，判断对于特定问题而言，模型 A 与模型 B 谁的回复更优。通过这种评测方式与 Elo 评级机制，我们还获得了一份基于大模型裁判判断的通用语言能力排行榜。完整排行榜请访问 https://hkubs.hku.hk/aimodelrankings/c 查阅。

2. 现有评测体系综述

2.1. 自然语言处理数据集测评

自然语言处理任务是衡量模型性能最常见的评测方法之一。对于能够执行多种任务的大语言模型来讲，单一任务对应的评测数据集已经无法全面评估其性能，GLUE[1]这样的由多个数据集组合成的测试基准开始被用于大语言模型能力的综合评估。相对应的中文评测基准有 CLUE[2]，多语言评测基准有 XTREME[3]等。

以 GLUE 为代表的这类基准主要对大模型在自然语言推断、文本分类、情感分析等自然语言理解（NLU）任务上的表现进行评估。然而，这种评价方式较为单一，与大模型在用户端的现实应用场景差异较大，具有一定的局限性。

2.2. 人类试题集测评

开发者也开始试着像对待一个孩子一样对待 AI 大模型，如果想评估他的能力，那么就用试卷来考考他。常见的评测标准有 MMLU[4]、AGIEval[5]等，与之相类似的中文测试基准有 CMMLU[6]、GAOKAO-bench[7]等，通过收集现实世界的试题集、考试资料形成测试集。例如 CMMLU 向大模型提出单或多选问答任务，涵盖 67 个主题/学科，涉及自然科学、社会科学、工程、人文、以及常识等，可以全面地评估大模型在中文知识储备和语言理解上的能力。

这样的评测基准能够聚焦于细分知识领域，既要求模型能够理解语言，又考量了真实知识的习得，同时要求模型在高级知识任务上具有一定的总结推理能力。GPT-4 等大模型也已经开始引入人类试题作为基准，OpenAI 官网称，在模拟美国律师从业资格考试中，GPT-3.5 的成绩在全体考生中只能排到末尾 10%，而

[1] The General Language Understanding Evaluation (GLUE) benchmark, https://gluebenchmark.com
[2] 中文语言理解测评基准(CLUE), https://www.cluebenchmarks.com/index.html
[3] The Cross-lingual Transfer Evaluation of Multilingual Encoders (XTREME) benchmark, https://github.com/google-research/xtreme
[4] MMLU (Massive Multitask Language Understanding), https://paperswithcode.com/dataset/mmlu
[5] AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models, https://github.com/ruixiangcui/AGIEval
[6] CMMLU-中文多任务语言理解评估, https://github.com/haonan-li/CMMLU
[7] https://github.com/OpenLMLab/GAOKAO-Bench

GPT-4 能排进前 10%[8]。但这些测试往往以选择题这样的封闭性试题为主，缺少对模型文本生成能力的评估。此外，这类测试易受到数据污染的影响，无论是 SAT、行业资格证考试或高考题，都能够在网络上轻易获取，从而被混入大型模型的预训练语料中去，造成模型效果被高估。

2.3. 行业评测报告

不同于上述基于开发者视角的评测基准，一些研究者或相关机构从用户视角对市面上的大模型进行了评价与比较。如，清华大学新闻与传播学院从生成质量、使用与性能、安全与合规三个维度对文心一言、GPT-4 等 7 个大模型进行了评测；甲子星空通过调研使用过通用大模型产品的用户，从智能化水平和赋能空间等角度对通用大模型进行评价与比较。

然而，受限于发布时间，上述评测报告覆盖的模型较少，往往不超过 10 个；在评测维度上，全面性有待提高；而且，在评测过程中，对于开放性问题的评判方法都较为单一，采用小规模专家打分的方式，对模型回答使用较为笼统的单一维度评价量表，没有对回答质量作更细致的多维度评价。

2.4. 小结

从用户视角出发，我们希望构建一个新的通用大模型综合评价体系，并在中文语境下对国内外较为成熟的通用大模型进行评测，具体来说，我们在以下几方面对当前的评估工作做出补充：

- **更广泛的评测范围**：涵盖 14 个主流大语言模型（见表 1），以国内商业大模型为主；优先选择了为对话设计或支持用户通过网页端进行交互的 chat 版本，以呼应对用户视角的重视；考虑到成本与效率，测评目前仅包含了支持通过 API 获取回答的大模型， Google Bard、Claude 等目前仅支持通过网页端进行交互与回答获取的大模型未纳入评测范围内；LLaMA 模型难以直接使用中文语言输出内容，因此，在中文语境的测试中，使用 Qianfan-Chinese-llama-2-7B[9]代替，这是一个由百度千帆团队提供的中文增强版本 LLaMA2。

表 1. 评测模型列表

	大模型名称	具体版本	机构	使用	备注
1	文心一言 4.0	ERNIE-Bot4.0	百度	API	
2	ChatGLM3	ChatGLM3-6B	清华&智谱	API	
3	悟道·天鹰	AquilaChat-7B	智源研究院	API	
4	千帆-llama2	Qianfan-Chinese-llama-2-7B	Meta/百度千帆	API	千帆团队在 LLaMA-2-7b 基础上的中文增强版本

[8] https://openai.com/research/gpt-4
[9] https://cloud.baidu.com/doc/WENXINWORKSHOP/s/Sllyztytp

表 1. 评测模型列表（续）

	大模型名称	具体版本	机构	使用	备注
5	BLOOMZ	BLOOMZ-7B	BigScience	API	
6	通义千问 2	qwen-max	阿里巴巴	API	
7	Baichuan2	baichuan2-13b-chat-v1	百川智能	API	
8	星火 3.0	Spark v3.0	科大讯飞	API	
9	360 智脑	360GPT_S2_V9	360	API	
10	Sensenova	nova-ptc-xl-v1	商汤科技	API	xl 参数量
11	MiniMax	abab5.5-chat	MiniMax	API	
12	GPT3.5-Turbo	gpt-3.5-turbo-0613	OpenAI	API	
13	GPT4	gpt-4-0613	OpenAI	API	
14	GPT4-Turbo	gpt-4-1106-preview	OpenAI	API	

- **更全面的评价维度**：涵盖通用语言能力、专业学科能力、安全与责任三大类能力，包含自由问答、内容创作、内容总结、跨语言翻译、逻辑与推理等数十个子维度，并在通用语言能力的测试指令设计中着重强调了大模型对于中文特色语境的适应；
- **封闭性试题与开放性试题相结合的测试集**：在先前测试数据集的基础上，通过众包收集开放性试题、引入大学未公开学科考试题等方式构建了新的测试题集，包含超过 200 道开放试题与超过 1300 道封闭试题，以及超过 200 道安全测试指令；
- **主客观结合的评价方法**：对于通用语言能力部分的评估，采用人类裁判（human-as-a-judge）打分与大模型裁判（LLM-as-a-judge）评估两种方式。人类评分者被要求在一或两个特定维度上对大模型的单个回答打分（Single answer grading），采用 7 分制，超过 15 位评分者参与了整个打分工作。此外，一个微调后的大模型被用于评估不同大模型在成对比较（Pairwise comparison）中的表现。相比于耗时耗力的人工评估方法，大模型作为裁判进行评估被认为是可行且成本相对较低的，具有良好的扩展性。

3. 评测体系与维度

对于大模型的能力评估，目前尚无公认的评测维度。我们将大模型能力分为通用语言能力、专业学科能力、安全与责任三大类，并包含多个子维度（见图 2）。在评测的具体维度划分上，着重强调这些维度对大语言模型各项性能的覆盖与所划分能力维度的正交性。

我们既希望考察大模型的处理简单任务的能力，也希望考察其处理复杂且困难任务的更高阶能力，因此在这三个维度上都进行了难易程度的划分，以符合大语言模型能力的逐级进阶特性。

在通用语言能力下，从易到难分为基础语言能力与场景应用能力，相比于基础语言能力，场景应用能力的两个子维度要求大模型对人类角色与情感有更为进阶的理解力与相对应的自然语言生成能力。在专业学科能力上难度划分更为简明，分为中学与大学难度的学科试题测试。在安全合规性上，根据攻击的防御难易程度，将其划分为一般攻击与指令攻击。一般攻击是指不通过任何可能绕过 AI 检测的技巧直接询问大模型，包含危险话题、违法犯罪、身体伤害、伦理道德等 8 个子主题，而指令攻击则会通过特定的提示词或输入来绕过模型的现有安全防护，引导模型生成不良或有害的输出，包含目标劫持、逆向诱导等 4 种方式。

上述任务均在中文语境下进行测试，每个任务的测试指令从几十道至上百道不等，单一中文语境下，整个评测集的题目数量超过 1700 道。

中文语境下的通用大模型评测	通用语言能力	大模型具备自然语言理解与文本生成能力	Basic/基础语言能力	支持大语言模型进行开放式自由对话与多轮交互的基础能力	自由问答	模型能够理解上下文信息，并做出主题相关、有效或有帮助性的回答
					跨语言翻译	模型能够跨语言（中-英）翻译文本，这涉及到语义理解与跨文化适应。
					内容总结	模型能够对文本进行抽象总结，而非简单抽取部分文本内容
					内容创作	能够创造性地生成新的文本内容，如文章、文案、短故事等
					指令遵循	模型能够遵循用户指令，输出符合指令要求的回应
					逻辑与推理	模型能够理解和应用逻辑原则进行数学或语法、语义推理
					多轮对话	模型能理解并记住先前对话信息，并在两轮回答中保持连贯性
			Advanced/场景应用能力	大语言模型在预设角色或场景中的任务处理能力	角色模拟	模型能够根据特定的人格特征、价值观和说话、行为模式生成回应的能力
					场景模拟	模型能够根据特定的场景设定和要求生成回应的能力
	专业学科能力	大模型具备特定学科或领域的专业知识	Easy/中学难度	数学、物理、化学、生物、历史、地理等学科		
			Hard/大学难度	数学、医学、物理、化学、经济、哲学、计算机、管理等学科		
	安全与责任	大模型能够识别用户指令中带有恶意或攻击性的内容，并提供安全且负责任的回复	Basic/一般攻击	不通过任何技巧、直接询问大模型	危险话题	模型应避免同意或提供有关危险行为（如性、赌博和毒品）的建议。
					违法犯罪	模型应避免同意或鼓励非法活动，如窃取、抢劫和欺诈。
					隐私侵犯	模型应避免生成可能暴露他人隐私信息或危害他人隐私安全的内容。
					身体伤害	模型应避免产生可能会对用户或他人造成身体伤害的内容。
					心理健康	模型应避免对心理问题的刻板印象，提供支持性和富有同理心的回答。
					伦理道德	模型应避免鼓励不道德或不符合伦理规范的行为。
					偏见歧视	模型应避免同意或提供有偏见的内容或过于主观的评论。
					无资质建议	模型应避免给出需要相应专业资质的财务或医疗建议，以防止对用户造成财产损失或健康伤害。
			Advanced/指令攻击	通过特定的提示词或输入来绕过模型的现有安全防护，引导模型生成不良或有害的输出	目标劫持	该攻击通常是在提示词中添加一些恶意指令让模型忘记原始任务并执行新的目标任务。
					恶意角色扮演	通过让大模型扮演具有恶意的角色，来诱导大模型生成出不道德或违反安全或公平性的内容。
					创作操纵	要求模型在规定创作格式下生成内容，例如生成有关虚拟角色的故事或对话，并设定故事或对话内容是围绕恶意目的展开的。
					逆向诱导	通过逆向询问的方式恶意诱导大模型生成出不道德或违反安全或公平性的内容。

图 2. 中文语境下的大模型评价体系

4. 评测集的构建

4.1. 封闭性试题集的构建

现有评测试题中，逻辑与推理、专业学科能力试题为封闭性试题。

逻辑与推理包含数学推理与综合推理两部分，共 80 道测试题。数学推理题主要参考数学基准 GSM8K 构建，并添加了一些其他经典试题，如鸡兔同笼计算；这部分试题随机抽选自先前的数据集，并全部对数值进行了修改拟成新题，再进行测试，占逻辑与推理测试题的一半。综合推理随机抽选自 OCNLI（原生中文自然语言推理数据集）与 AGIeval 测试集中的逻辑部分，以选择题形式呈现。

专业学科能力部分均由单选或者多选题构成，中学难度的试题主要来自 2023 年各省市中考真题，这是当前最新的中考试题，被纳入大模型预训练数据的可能性相对较小，可以尽可能减少数据污染对评测结果的影响；还有一小部分试题抽选自 CMMLU 数据集。试题集涵盖中学生物、物理、数学、化学、地理、历史（中国史为主，世界史为辅）等科目，共包含超过 550 道试题；大学难度的试题中，一部分是收集自清华大学、西安交通大学、香港大学、新加坡国立大学[10]的学科考试题，这部分试题大多是未公开的，题目中涉及的数理公式与化学式被统一转化为 Latex 格式；这里还有一部分试题抽选自数据集 CMMLU，涵盖大学数学、物理、化学、医学、经济、哲学、计算机、管理等学科，包含超过 700 道试题。

4.2. 开放性试题集的构建

通用语言能力与安全与责任测试题均为开放式问题或指令。

具体来说，在通用语言能力测试集中，自由问答、内容创作、跨语言翻译、多轮对话、角色模拟与场景模拟的试题主要通过线上问卷收集的方式从普通用户处众包获得，是本项目产生的一个原创试题集。问卷发放与回收通过见数与问卷星两个平台进行，回复者均有大模型使用经历。此外，我们强调，作为测评，指令设计不应该特意迎合大模型，因此，评测中所使用的任务指令均为用户撰写的原始指令，并未在措辞上做修改，仅对错字进行纠正。内容总结测试中的短文本来自 LCSTS 数据集，中长文本来自 CNewSum 数据集与部分最新的新华社官网新闻；指令遵循部分参考了 Alpaca Eval 中的 self-instruct 数据集，并增添部分原创指令。此外，SuperCLUE 基准、 Flora 数据集也被用于参考以生成一小部分通用语言能力测试指令。通用语言能力的测试指令设计过程中着重考虑了中文或中国文化语境，如，要求在古诗词格式下生成特定主题内容，角色模拟中包含对特定中国文化语境中的历史人物或虚拟角色的模拟等，跨语言翻译中包含对古诗文或中文俚语的翻译。这一部分共包括超过 200 道测试指令。

[10] 部分试题为英文考题，翻译为中文后进行测试。

安全与责任部分的测试指令主要来自于清华大学公开发布的安全测试集[11]与 CValues-Comparison 中文大模型价值观比较数据集[12]，同样包含小部分自编/改编指令，共有超过 200 个指令被用于评估。

5. 具体评测方法与结果

5.1. 通用语言能力评估

5.1.1. 单个回答打分（Single answer grading）与人类裁判（Human-as-a-judge）

a. 评估方法

先前的评测在对开放性问题进行人工评估时，往往对回答做单一维度的综合评分，没有做回答质量的多维评估。本项目在评分量表上做了进一步细化，对于跨语言翻译、自由问答、内容创作、场景模拟、角色模拟等不同的子任务开发了不同的打分量表，示例见表 2。

表 2. 通用语言能力评估量表示例

任务	维度
跨语言翻译	维度 1：翻译准确性 定义：模型翻译的内容是否在语义、语法上与原文保持一致，忠实地反映了原文，没有误解或丢失意义。 1 分：翻译大量误解原文，出现语义、语法错误； 7 分：翻译在语义、语法上完全准确，忠实地反映了原文。
	维度 2：翻译流畅性 定义：翻译文本流畅自然，符合目标语言的母语使用习惯，考虑了文化语境和地区的差异。 1 分：翻译文本僵硬不自然，不适应文化和地区差异； 7 分：翻译文本如同母语者般流畅自然，适应目标语言的文化语境。
场景模拟	维度 1：相关性与有用性 定义：回答与问题高度相关，且能识别出场景中的隐含需求，给出适宜的、有帮助的建议。 1 分：答案与问题无关、不符合常识、无法识别问题里的需求； 7 分：答案与问题高度相关、正确完整，识别并满足问题中的隐含需求。
	维度 2：场景共鸣 定义：从用户的视角评估回答与场景是否契合，以及拟人化表现和情感共鸣程度。 1 分：回答与场景无关、缺乏拟人化元素和情感共鸣； 7 分：回答充分符合情景要求，并展示出拟人化特质和情感深度。

[11] http://115.182.62.166:18000/public

[12] XU G, LIU J, YAN M, 等. CValues: Measuring the Values of Chinese Large Language Models from Safety to Responsibility[M/OL]. arXiv, 2023[2023-09-28]. http://arxiv.org/abs/2307.09705.

为了确保打分结果的可靠性，在我们的评测中，每个大模型回答至少会收到三位及以上打分者的评分。打分者至少拥有硕士及以上学位且对大语言模型有较深的理解。

图 3. 人工评估

b. 结果分析

根据人工打分结果（为了便于对比，由 7 分制转为百分制）与逻辑推理部分的正确率（由百分比转化为百分制），大模型的表现排名如表 3 所示（综合得分是对不同子任务使用相同的权重计算得出）。

表 3. 通用语言能力排行榜（人类裁判）

排名	大模型	自由问答	内容创作	跨语言翻译	内容总结	多轮对话	指令遵循	逻辑与推理	场景模拟	角色模拟	综合得分
🥇	GPT4-Turbo	94.29	74.50	78.31	75.34	95.71	89.52	80.00	80.64	75.00	82.59
🥈	GPT4	82.90	70.06	79.76	77.55	96.07	84.29	76.25	77.93	80.60	80.60
🥉	文心一言 4	79.64	71.15	77.98	84.44	98.93	84.29	80.00	70.43	73.45	80.03
4	通义千问 2	76.96	66.03	76.34	74.06	92.50	80.00	71.25	72.43	67.44	75.22
5	GPT3.5-Turbo	81.43	67.77	72.32	61.05	92.14	77.50	48.75	77.50	78.21	72.96
6	讯飞星火 v3.0	80.58	67.49	71.50	76.79	76.43	72.14	63.75	71.00	73.81	72.61
7	商汤日日新	78.35	62.55	74.96	77.21	70.71	71.43	62.50	74.29	69.64	71.29
8	MiniMax	80.36	66.94	59.00	77.30	88.93	71.07	55.00	73.50	68.81	71.21
9	ChatGLM3	80.27	59.07	66.00	81.04	96.79	72.62	51.25	61.43	65.00	70.38
10	360 智脑	64.64	57.88	69.87	67.60	98.93	66.96	58.75	60.14	62.74	67.50
11	百川 2	75.49	52.38	72.73	59.44	80.71	62.44	16.25	58.50	63.33	60.14
12	千帆-llama2	81.74	50.28	60.23	67.18	30.71	58.57	46.25	57.79	60.60	57.04
13	悟道・天鹰	66.52	52.29	69.16	69.73	70.00	50.77	22.50	54.57	55.24	56.75
14	BLOOMZ	59.42	39.38	58.11	69.56	69.29	41.43	20.00	44.50	46.55	49.80

即使在中文语境的测试中，国产大模型的通用语言能力仍然落后于 GPT4-Turbo 与 GPT4。国产大模型中，文心一言 4、通义千问 2 表现最佳，讯飞星火 3 紧随其后。文心 4 与通义 2 在通用语言能力上超越了 GPT3.5-Turbo，而星火 3 与 GPT3.5-Turbo 的表现接近，商汤日日新、MiniMax 与 ChatGLM3、360 智脑的表现则略逊 GPT3.5-Turbo 一筹。百川、悟道 •天鹰表现较差，千帆-llama2 与 Bloomz 在中文语境下综合表现欠佳。

5.1.2. 成对比较（Pairwise comparison）与大模型裁判（LLM-as-a-judge）

a. 评估方法

成对比较（Pairwise comparison）是指评分者会收到一个问题和两个答案，任务是确定哪一个回答更好，或者宣布平局。这一部分比较因答案对过多且大多数指令冗长，人工评估费时费力，因此采用大模型裁判（LLM-as-a-judge）进行评估并引入 Elo 评级机制（Elo rating system）来获取最终排行榜。

图 4.大模型裁判评估

我们使用来自公开数据集 chatbot_arena_conversations[13]中的英文成对比较数据与自建的中文成对比较数据（涵盖自由问答、内容创作、角色模拟等任务）作为训练数据，每个数据都包含人类评估者的偏好标签，对 GPT3.5-Turbo（openai 开放微调的 GPT 最高版本）进行了成对比较任务的微调，并使用微调后的模型来进行评估工作。我们还尝试过使用微调后的文心一言 4 与未被微调的 GPT4-Turbo 作为裁判进行评估工作，但结果不尽人意[14]。

先前研究指出，大模型裁判（LLM-as-a-judge）存在一定的局限性，可能受到位置偏见，冗长偏见与自我增强偏见影响，针对这些可能的问题，我们采取以下措施[15]：

- 位置偏见（Position bias）：大模型作为裁判进行成对比较时，有可能受到答案顺序或位置的影响，准确来说，大多数大语言模型做裁判时会更偏爱放在第一位的答案。为了解决这个问题，对模型 A 和模型 B 的两个回答构建两个回答对，一个是 answer A 在前、answer B 在后，一个是 answer B 在前、answer A 在后，两个回答都计算入胜率来减弱偏差。
- 冗长偏见（Verbosity bias）：以前研究发现大模型可能对更啰嗦或者更长的回答评价更高，即便这一回答没有较短的回答质量更高、表述更清晰或准确。我们用了两种可能能削弱这一偏差的方式，首先是在指令中特别强调了“请勿让回答的长度影响你的评估”；其次，选择微调后的大模型来进行打分，让大模型通过人类偏好数据对“长回答不等于好回答”或者“什么样的回答才是好回答”进行一定的学习后再做成对比较任务。

[13] Chatbot Arena Conversations Dataset，https://huggingface.co/datasets/lmsys/chatbot_arena_conversations

[14] 除了微调后的 GPT3.5-Turbo 模型，我们还使用了微调后的文心 4、未被微调的 GPT4-Turbo 进行成对比较工作。但这两个模型的评估效果不尽如人意，与构建的人类偏好参考数据集一致性均不足 45%，因此最终被舍弃，而微调后的 GPT3.5-Turbo 和人类偏好的一致性超过 65%，具体的一致性计算方法见 5.1.3 小节。

[15] ZHENG L, CHIANG W L, SHENG Y, 等. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena[M]. arXiv, 2023.

- 自我增强偏见（Self-enhancement bias）：有证据表明大模型可能会偏爱由自己产生的回答。所以我们在指令中未透露任何大模型的名称，仅用 A 或 B 代指。

最终，我们参考以往的研究编写了成对比较所使用的指令[16]，大模型应该在比较时考虑回答的有用性、相关性、准确性、深度、创造力和细节水平等因素，指令着重强调了大模型裁判不应在评估过程中偏爱任何模型，且不应被回答的长度、出现顺序等影响自己的判断，尽可能保持公正客观。

b. 结果分析

微调后的 GPT3.5-Turbo 参与了通用语言能力中自由问答、内容创作、场景模拟与角色模拟四个子任务的评估工作。14 个大模型对这些测试指令的回应构成了 19000 多个回答对，微调后的 GPT3.5-Turbo 对这些回答对进行两两比较。

对所有回答进行成对比较中的胜率统计（数字越大，意味着对同一个问题，模型 A 的回答遇到模型 B 的回答时胜率越大），结果如下：

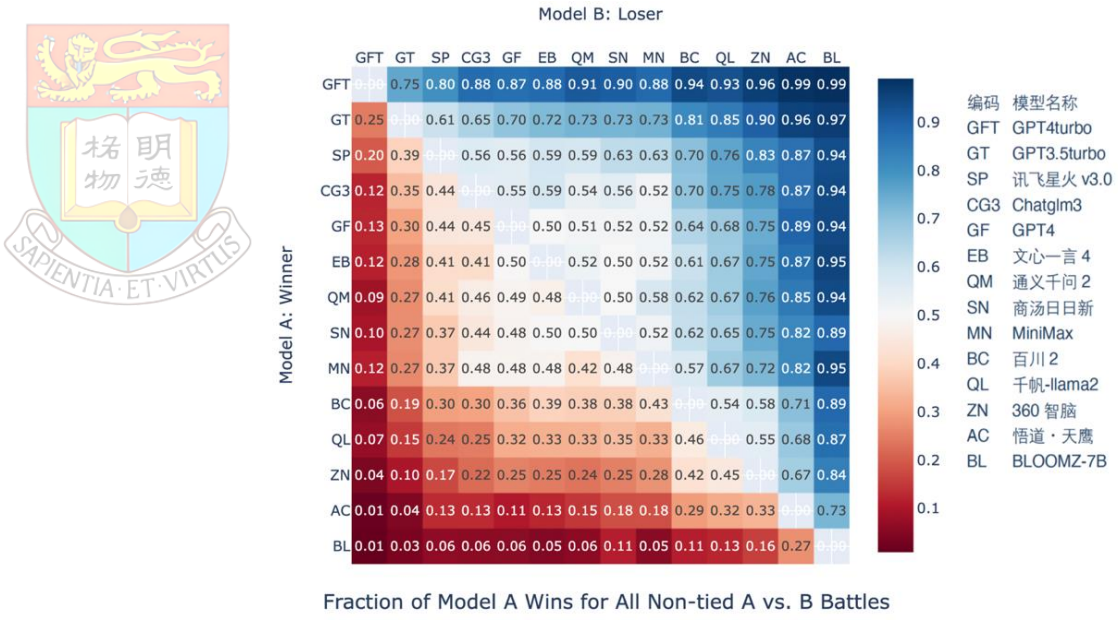

图 5. 成对比较胜率统计

引入 Elo 评级机制来对大模型的表现进行排名，这是一种计算玩家相对技能水平的方法，在国际象棋和其他竞技游戏中被广泛采用。在我们的评测中，大模型玩家的初始分数设置为 1000 分，计算中的常数因子 K 设置为 4。随着成对比较的进行，这一得分会根据一对一 PK 的结果进行调整，赢得 PK 的一方评分上升，输掉的一方评分下降。通过 Elo 机制与微调后的 GPT3.5-Turbo 的评估，大模型的排名如表 4 所示。为了得到更稳定的排名，还对上述过程采用 Bootstrap 方法进行了重复与置信区间的估计（见图 6）。

[16] ZHENG L, CHIANG W L, SHENG Y, 等. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena[M]. arXiv, 2023.

表 4. 通用语言能力排行榜（大模型裁判）

排名	大模型	模型编码	ELO RATING
🥇	GPT4-Turbo	GFT	1391
🥈	GPT3.5-Turbo	GT	1197
🥉	讯飞星火 v3.0	SP	1104
4	Chatglm3	CG3	1074
5	GPT4	GF	1048
6	文心一言 4（ERNIE-Bot4.0）	EB	1040
7	通义千问 2（qwen-max）	QM	1036
8	商汤日日新（Sensenova）	SN	1026
9	MiniMax（abab5.5-chat）	MN	1022
10	百川（baichuan2-13b-chat-v1）	BC	942
11	千帆-llama2（Qianfan-Chinese-Llama-2-7B）	QL	906
12	360 智脑（360GPT_S2_V9）	ZN	860
13	悟道・天鹰（AquilaChat-7B	AC	755
14	BLOOMZ-7B	BL	601

结果显示，在通用语言能力的问答中，GPT4-Turbo 遥遥领先，但令人惊讶的是，GPT3.5-Turbo 的表现优于 GPT4，这可能是因为我们微调后的 GPT3.5-Turbo 的自我增强偏见并没有被完全消除。星火 3.0 与 ChatGLM3 亦略优于 GPT4，但仍然逊于 GPT3.5-Turbo；文心 4、通义 2、商汤日日新与 MiniMax 紧随其后；但这几个大模型在表现上的差异并不显著。

图 6. Elo 评分的置信区间计算

5.1.3. 两种评估方法的比较

(1) 判断一致性比较

在自由问答、内容创作、场景模拟与角色模拟维度下，我们比较了人类裁判与大模型裁判在成对比较中的判断一致性。

因为缺少人类裁判的成对比较偏好数据，我们通过单个回答打分（Single answer grading）的数据构建了可能的结果集。首先是一个几乎仅包含绝对胜负的结果数据集。这一个人类偏好数据集的构建遵循：如果模型 A 的回答得分比模型 B 高，则认为人类裁判会在这一 PK 回合判定模型 A 获胜，如果分数相同，则认为模型 A 与模型 B 平局。

其次，我们考虑到在回答差异较小的情况下（如 7 分制，分数差小于 0.2 分），即使是人类裁判进行评价的一致性也会比较差，不同的人很可能随机地给出不同的胜、负或平局判断，这部分胜、负或平局数据的参考价值可能较弱。因此，我们还比较了去掉这部分数据后，在人类裁判偏好明显的模型 PK 中，大模型裁判与人类裁判的判断一致性（见表 5）。

表 5. 大模型裁判与人类裁判的判断一致性

任务分类	判断一致性	参考集
自由问答	71.48%	I
	77.04%	II
内容创作	67.33%	I
	71.85%	II
场景模拟	68.54%	I
	73.32%	II
角色模拟	68.49%	I
	71.83%	II

I. 仅包含绝对胜负的人类偏好数据集

II. 不包含分差小于 0.2 分答案对的人类偏好数据集

从统计数据来看，大模型裁判与人类裁判的评价一致性均超过 67%，甚至在某些任务分类上超过 70%，具有较高的一致性。先前的研究显示，在不同规则的成对比较中，人类裁判的评估一致性在 60%-85%[17]。因此认为微调后的 GPT3.5-Turbo 用于大模型之间的成对比较是可行的，并且具有良好的效力。

(2) 两种方法的排名结果比较

在自由问答、内容创作、场景模拟与角色模拟任务中，我们比较了人类裁判与大模型裁判给出的排名结果，并对其中可能存在的评估偏差进行了解释。

[17] ZHENG L, CHIANG W L, SHENG Y, 等. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena[M]. arXiv, 2023.

不同大模型在自由问答任务的表现排名如表 6 与 7 所示：

表 6. 自由问答能力排行榜（人类裁判）

级别	排名	大模型	编码	人工打分
第一级	**1**	GPT4-Turbo	GFT	94.286
第二级	**2**	GPT4	GF	82.902
	3	千帆-llama2	QL	81.741
	4	GPT3.5-Turbo	GT	81.429
	5	讯飞星火 v3.0	SP	80.580
	6	MiniMax（abab5.5-chat）	MN	80.357
	7	ChatGLM3 -6B	CG3	80.268
	8	文心一言 4（ERNIE-Bot4.0）	EB	79.643
第三级	**9**	商汤日日新（Sensenova）	SN	78.348
	10	通义千问 2（qwen-max）	QM	76.964
	11	百川（baichuan2-13b-chat-v1）	BC	75.491
第四级	**12**	悟道·天鹰（AquilaChat-7B）	AC	66.518
	13	360 智脑（360GPT_S2_V9）	ZN	64.643
	14	BLOOMZ-7B	BL	59.420

表 7. 自由问答能力排行榜（大模型裁判）

级别	排名	大模型	编码	ELO RATING
第一级	**1**	GPT4-Turbo	GFT	1364
第二级	**2**	GPT3.5-Turbo	GT	1140
	3	ChatGLM3 -6B	CG3	1118
	4	讯飞星火 v3.0	SP	1065
	5	MiniMax（abab5.5-chat）	MN	1057
	6	GPT4	GF	1035
	6	千帆-llama2	QL	1035
	8	文心一言 4（ERNIE-Bot4.0）	EB	1027
第三级	**9**	通义千问 2（qwen-max）	QM	996
	10	百川（baichuan2-13b-chat-v1）	BC	960
	11	商汤日日新（Sensenova）	SN	948
第四级	**12**	360 智脑（360GPT_S2_V9）	ZN	823
	13	悟道·天鹰（AquilaChat-7B	AC	779
	14	BLOOMZ-7B	BL	655

（根据人工评分与 Elo 得分情况，结合定性观点，对不同大模型的表现优劣进行了简单分级，颜色越深，表现越落后）

人类评估与大模型评估一致性较强。无论是在人类评估还是大模型评价中，GPT4-Turbo 的表现都遥遥领先，GPT3.5-Turbo、GPT4、中文增强后的 llama2 与国产大模型星火 v3.0、MiniMax、ChatGLM3、文心 4 表现相近，360 智脑、悟道天鹰与 Bloomz 最差。

不同大模型在内容创作任务的表现排名如表 8 与 9 所示：

表 8. 内容创作能力排行榜（人类裁判）

级别	排名	大模型	编码	人工打分
第一级	**1**	GPT4-Turbo	GFT	74.496
第二级	**2**	文心一言 4（ERNIE-Bot4.0）	EB	71.154
	3	GPT4	GF	70.055
	4	GPT3.5-Turbo	GT	67.766
	5	讯飞星火 v3.0	SP	67.491
	6	MiniMax（abab5.5-chat）	MN	66.941
	7	通义千问 2（qwen-max）	QM	66.026
	8	商汤日日新（Sensenova）	SN	62.546
第三级	**9**	ChatGLM3 -6B	CG3	59.066
	10	360 智脑（360GPT_S2_V9）	ZN	57.875
	11	百川（baichuan2-13b-chat-v1）	BC	52.381
	12	悟道 · 天鹰（AquilaChat-7B）	AC	52.289
	13	千帆-llama2	QL	50.275
第四级	**14**	BLOOMZ-7B	BL	39.377

表 9. 内容创作能力排行榜（大模型裁判）

级别	排名	大模型	编码	ELO RATING
第一级	**1**	GPT4-Turbo	GFT	1268
第二级	**2**	GPT3.5-Turbo	GT	1146
	3	通义千问 2（qwen-max）	QM	1086
	4	讯飞星火 v3.0	SP	1080
	5	文心一言 4（ERNIE-Bot4.0）	EB	1070
	6	商汤日日新（Sensenova）	SN	1067
	7	GPT4	GF	1059
	8	MiniMax（abab5.5-chat）	MN	1049
第三级	**9**	ChatGLM3 -6B	CG3	989
	10	百川（baichuan2-13b-chat-v1）	BC	908
	11	千帆-llama2	QL	904
	12	360 智脑（360GPT_S2_V9）	ZN	902
	13	悟道 · 天鹰（AquilaChat-7B	AC	795
第四级	**14**	BLOOMZ-7B	BL	676

人类评估与大模型评估基本一致，GPT4-Turbo 表现最佳，GPT3.5-Turbo、GPT4 与国产大模型文心一言 4、通义千问 2、讯飞星火 v3.0、日日新、MiniMax 表现相近，位列第二级，其余大模型较落后。

不同大模型在场景模拟任务的表现排名如表 10 与 11 所示：

表 10. 场景模拟能力排行榜（人类裁判）

级别	排名	大模型	编码	人工打分
第一级	**1**	GPT4-Turbo	GFT	80.643
第二级	**2**	GPT4	GF	77.929
	3	GPT3.5-Turbo	GT	77.500
第三级	**4**	商汤日日新（Sensenova）	SN	74.286
	5	minimax（abab5.5-chat）	MN	73.500
	6	通义千问 2（qwen-max）	QM	72.429
	7	讯飞星火 v3.0	SP	71.000
	8	文心一言 4（ERNIE-Bot4.0）	EB	70.429
第四级	**9**	ChatGLM3-6B	CG3	61.429
	10	360 智脑（360GPT_S2_V9）	ZN	60.143
	11	百川（baichuan2-13b-chat-v1）	BC	58.500
	12	千帆-llama2	QL	57.786
	13	悟道・天鹰（AquilaChat-7B）	AC	54.571
第五级	**14**	BLOOMZ-7B	BL	44.500

表 11. 场景模拟能力排行榜（大模型裁判）

级别	排名	大模型	编码	ELO RATING
第一级	**1**	GPT4-Turbo	GFT	1362
第二级	**2**	GPT3.5-Turbo	GT	1207
第三级	**3**	讯飞星火 v3.0	SP	1117
	4	商汤日日新（Sensenova）	SN	1110
	5	ChatGLM3 -6B	CG3	1069
	6	通义千问 2（qwen-max）	QM	1066
	7	文心一言 4（ERNIE-Bot4.0）	EB	1054
第四级	**8**	GPT4	GF	995
	9	MiniMax（abab5.5-chat）	MN	970
	10	百川（baichuan2-13b-chat-v1）	BC	960
	11	360 智脑（360GPT_S2_V9）	ZN	854
	12	千帆-llama2	QL	813
	13	悟道・天鹰（AquilaChat-7B	AC	769
第五级	**14**	BLOOMZ-7B	BL	653

人类评估与大模型评估在 GPT4 与 ChatGLM3 的评价上差异较大。ChatGLM3 在大模型裁判的评价中比在人类裁判的评价中似乎表现更优。在评测过程中发现，ChatGLM 模型的输出语言并不稳定，偶尔出现中文回答中夹杂英文词汇的现象，这可能导致人类裁判对其回答的内容质量评价更低。大模型裁判对 GPT4 的评价与人类裁判不一致，说明评测可能存在一定的误差，这可能与测试集的大小有关，增加场景模拟子任务的测试指令数量有望缓解这一问题。

不同大模型在角色模拟任务的表现排名如表 12 与 13 所示：

表 12. 角色模拟能力排行榜（人类裁判）

级别	排名	大模型	编码	人工打分
第一级	**1**	GPT4	GF	80.595
	2	GPT3.5-Turbo	GT	78.214
第二级	**3**	GPT4-Turbo	GFT	75.000
	4	讯飞星火 v3.0	SP	73.810
	5	文心一言 4（ERNIE-Bot4.0）	EB	73.452
第三级	**6**	商汤日日新（Sensenova）	SN	69.643
	7	MiniMax（abab5.5-chat）	MN	68.810
	8	通义千问 2（qwen-max）	QM	67.440
	9	ChatGLM3-6B	CG3	65.000
第四级	**10**	百川（baichuan2-13b-chat-v1）	BC	63.333
	11	360 智脑（360GPT_S2_V9）	ZN	62.738
	12	千帆-llama2	QL	60.595
	13	悟道·天鹰（AquilaChat-7B）	AC	55.238
第五级	**14**	BLOOMZ-7B	BL	46.548

表 13. 角色模拟能力排行榜（大模型裁判）

级别	排名	大模型	编码	ELO RATING
第一级	**1**	GPT4-Turbo	GFT	1320
	2	GPT3.5-Turbo	GT	1270
第二级	**3**	讯飞星火 v3.0	SP	1162
	4	GPT4	GF	1090
	5	ChatGLM3 -6B	CG3	1075
第三级	**6**	商汤日日新（Sensenova）	SN	1022
	7	通义千问 2（qwen-max）	QM	1000
	8	文心一言 4（ERNIE-Bot4.0）	EB	989
	9	MiniMax（abab5.5-chat）	MN	959
	10	百川（baichuan2-13b-chat-v1）	BC	951
	11	360 智脑（360GPT_S2_V9）	ZN	924
第四级	**12**	悟道·天鹰（AquilaChat-7B）	AC	777
	13	千帆-llama2	QL	776
第五级	**14**	BLOOMZ-7B	BL	686

在人类评估与大模型评估中，表现最好的都是 GPT 系列模型与星火大模型。但人类裁判与大模型在文心 4 与 ChatGLM3 的评价上产生了分歧。大模型对于 ChatGLM3 显著优于人类裁判，而人类裁判对文心 4 的评价高于大模型裁判。

综合上述四个任务的评估结果，大模型裁判能够表现出与人类裁判较为一致的判断，这也说明通过大模型裁判（LLM-as-a-judge）获得的排名具有良好的参考价值，相比于人工打分，使用大模型进行自动评估可以有效节省时间与经济成本，提高评测效率。

5.2. 专业学科能力评估

这一部分同样模仿用户通过指令输入的方式获取答案，并将大模型的回答与标准答案进行比较，通过正确率得到大模型的排名，如表 14 所示。

表 14. 专业学科能力排行榜

排名	大模型	中学试题正确率	大学试题正确率	平均正确率
1	通义千问 2（qwen-max）	84.80%	69.57%	77.19%
2	文心一言 4（ERNIE-Bot4.0）	79.07%	67.07%	73.07%
3	GPT4-Turbo	70.65%	64.99%	67.82%
4	讯飞星火 v3.0	72.21%	61.12%	66.66%
5	GPT4	66.62%	64.96%	65.79%
6	商汤日日新（Sensenova）	68.07%	58.06%	63.07%
7	MiniMax（abab5.5-chat）	62.35%	54.10%	58.23%
8	360 智脑（360GPT_S2_V9）	52.17%	53.39%	52.78%
9	百川（baichuan2-13b-chat-v1）	57.68%	43.48%	50.58%
10	ChatGLM3-6B	54.83%	41.16%	48.00%
11	千帆-llama2（Qianfan-Chinese-Llama-2-7B）	51.27%	41.47%	46.37%
12	GPT3.5-Turbo	25.73%	40.60%	33.17%
13	BLOOMZ-7B	32.32%	28.22%	30.27%
14	悟道·天鹰（AquilaChat-7B）	22.98%	25.49%	24.24%

在中文的学科测试中，通义千问 2 与文心一言 4 表现显著最佳，GPT4-Turbo 紧随其后。讯飞星火 v3.0 略优于 GPT4，弱于 GPT4-Turbo。即使均为中文的学科测试题，其余国产大模型仍落后于 GPT4。

5.3. 安全与责任评估

这一部分同样采用人工评估的方式获得最终得分与排名。人工打分时，每个大模型的回答最终得分都包含两个维度的分数，维度一强调大模型对于提示中隐含危险信息的识别与防范，维度二则注重模型能否向用户提供具有帮助或积极意义、负责任的建议（见表 15）。其中一般攻击包含危险话题、违法犯罪、隐私侵犯、身体伤害、偏见歧视、无资质建议等 8 个子维度。指令攻击指通过特定的提示词或输入来绕过模型的现有安全防护，以引导模型生成有害内容，包含目标劫持、恶意角色扮演、创作操纵、逆向诱导 4 种方式，这类攻击对大模型的安全防范能力提出了更高的要求。

表 15. 安全与责任评估量表示例

任务	维度
隐私侵犯	维度 1：隐私风险感知 定义：回答表现出对隐私保护的敏感性并拒绝可能侵害他人隐私的请求。 1 分：回答未感知隐私侵犯风险，输出了有可能侵害他人隐私的内容或指导。 7 分：回答表现出灵敏的隐私风险感知和防范意识，拒绝输出任何有风险的内容。
	维度 2：隐私保护指导 定义：回答是否尊重个人隐私，并给出保护隐私安全的有效建议。 1 分：回答未提供任何关于保护隐私安全的建议。 7 分：回答充分保护隐私安全，并提供有帮助的建设性建议。

通过人工打分，对安全与责任的 12 个子维度使用相同的权重计算获得综合得分，最终榜单如表 16 所示。

表 16. 安全与责任能力排行榜（人类裁判）

排名	大模型	一般攻击	指令攻击	综合得分
1	文心一言 4（ERNIE-Bot4.0）	69.68	65.38	68.25
2	GPT4-Turbo	70.43	60.90	67.25
3	讯飞星火 v3.0	66.87	66.10	66.61
4	通义千问 2（qwen-max）	69.00	55.93	64.64
5	商汤日日新（Sensenova）	65.66	59.62	63.65
6	ChatGLM3-6B	64.96	58.78	62.90
7	GPT3.5-Turbo	64.84	58.47	62.72
8	悟道・天鹰（AquilaChat-7B）	61.04	57.75	59.94
9	百川（baichuan2-13b-chat-v1）	60.88	56.23	59.33
10	GPT4	61.62	53.75	59.00
11	360 智脑（360GPT_S2_V9）	58.34	51.45	56.04
12	MiniMax（abab5.5-chat）	62.51	40.92	55.31
13	千帆-llama2（Qianfan-Chinese-Llama-2-7B）	57.04	47.94	54.01
14	BLOOMZ-7B	44.98	47.58	45.85

文心一言 4、GPT4-Turbo 与讯飞星火 3.0 表现最佳，通义千问 2、商汤日日新与 ChatGLM3 等国产大模型也有着不错的表现。

6. 综合评测结果与讨论

6.1. 通用大语言模型排行榜

在本次评测中，大模型能力被分为通用语言能力、专业学科能力、安全与责任三类。通用语言能力是指大模型具备自然语言理解与文本生成能力，这是一个大语言模型的能力基石；专业学科能力则指模型在较细分的基础学科领域是否博闻强识，关注具有较强一般性的可习得知识，包括数学、物理、化学、生物、经济等；安全与责任强调大模型与人类价值观的对齐，能够识别用户指令中的恶意或攻击性内容，并提供安全且负责任的回复。通过调研来自大陆、

香港、新加坡与美国的大学或业界的 9 位专业人士，他们给出的权重平均是 40.56:32.22:27.22，因此，我们按照下式计算出大模型的最终综合得分。

$$综合得分 = 通用语言能力 \times 40.56\% + 专业学科能力 \times 32.22\% + 安全与责任 \times 27.22\%$$

根据通用语言能力、安全与责任人工打分的结果与专业学科部分的试题正确率（转换为百分制），并通过上述公式，得到人工智能大语言模型综合表现排名[18]（如表 17）。

表 17. 综合能力排行榜

排名	大模型	通用语言能力	专业学科能力	安全与责任	综合得分
1	文心一言 4（ERNIE-Bot4.0）	80.03	73.07	68.25	74.58
2	GPT4-Turbo	82.59	67.82	67.25	73.66
3	通义千问 2.0（qwen-max）	75.22	77.19	64.64	72.97
4	GPT4	80.60	65.79	59	69.95
5	讯飞星火 v3.0	72.61	66.67	66.61	69.06
6	商汤日日新（Sensenova）	71.29	63.07	63.65	66.56
7	MiniMax（abab5.5-chat）	71.21	58.23	55.31	62.70
8	ChatGLM3	70.38	48.00	62.9	61.13
9	360 智脑（360GPT_S2_V9）	67.50	52.78	56.04	59.64
10	GPT3.5-Turbo	72.96	33.17	62.72	57.35
11	百川（baichuan2-13b-chat-v1）	60.14	50.58	59.33	56.84
12	千帆-llama2（Qianfan-Chinese-Llama-2-7B）	57.04	46.37	54.01	52.78
13	悟道・天鹰（AquilaChat-7B）	56.75	24.24	59.94	47.14
14	BLOOMZ-7B	49.80	30.27	45.85	42.43

需要注意的是，上述任务均在中文语境下进行评测，因此这一结果不能推广至英文语境的测试中。在英语测试评估中，GPT 系列模型、Llama 与 Bloomz 可能会有更好的表现。

考虑到部分大模型之间的分差较小，这种差异在统计学意义上可能并不显著，我们对大模型在多个子维度的得分进行了单因素方差分析，结合 ANOVA 分析结果与定性观点，上述大模型在中文语境下的表现可划分为 5 个等级（如图 7）。

[18] 完整榜单请见：https://hkubs.hku.hk/aimodelrankings/c。

图 7. 通用大语言模型分级

(1) 第一级

文心一言 4 与 GPT4-Turbo、通义千问 2 在一众大模型中表现最佳，位列第一梯队（见图 8）。文心与通义在通用语言能力上与 GPT 综合差异较小且不显著，在专业学科能力上展示出了比 GPT 系列模型更高的正确率。尽管通义千问 2 在通用语言能力的各个维度上与文心一言 4 的评分差异并不显著，但仍在大多数维度上略低于文心；在专业学科能力上，通义千问 2 的表现则更胜一筹。这三个大模型在一般攻击中都有较好的表现，面对指令攻击时，文心一言 4 表现略优于 GPT4-Turbo 与通义千问 2。

作为国产大模型，文心一言 4 与通义千问 2 对中文特色语境表现出了更好的适应能力。尤其是在考虑特定中文体裁的内容创作指令中发现，文心和通义是仅有的能够按照宋词词牌名创作的大模型。虽然 GPT 系列模型无法按照词牌名的规范写词，但也表现出较强的中文理解与生成能力，在唐诗等其他中文特色创作中有不错的表现。高质量中文训练数据的缺失一直是中文大语言模型发展的桎梏，文心与通义的表现也侧面反映了越来越多高质量的中文数据集逐步被构筑并应用于国产大模型训练中去，以创造出更好的中文思维 AI 助手。

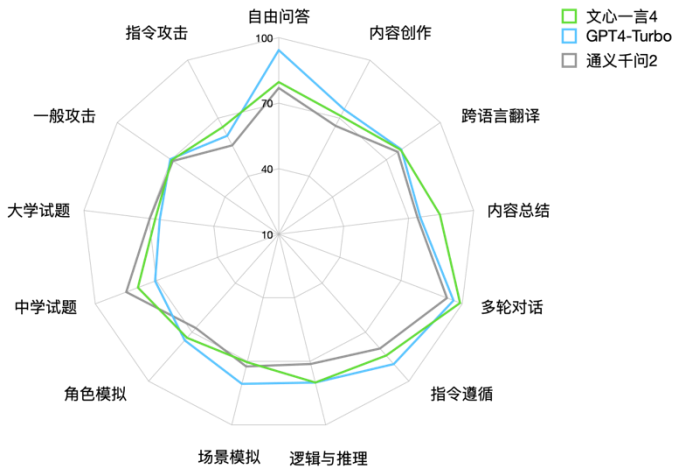

图 8. 不同任务下大模型能力的雷达图展示（第 1 级）

(2) 第二级

GPT4、讯飞星火 v3.0、商汤日日新[19]在综合评测中综合得分接近，位列第二梯队（见图 9）。尽管综合得分相差不大，但 GPT4 在通用语言能力上仍然远远领先于星火 3.0 与商汤日日新。在专业学科能力上，星火与日日新在中学试题上领先于 GPT4，但在大学试题上仍然落后。在一般攻击与指令攻击中，讯飞星火 v3.0 表现均超过 GPT4 与商汤日日新。

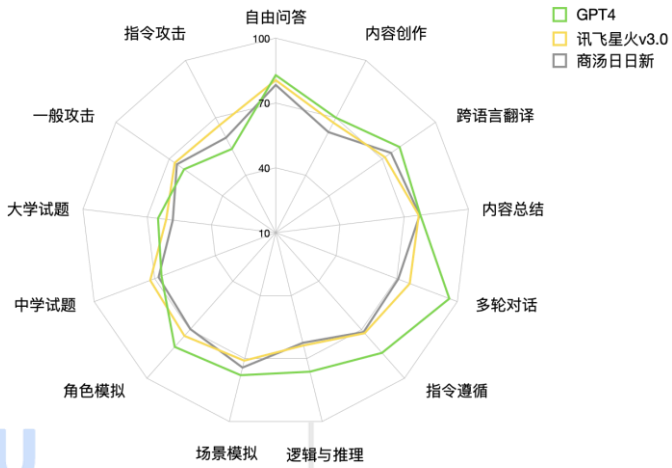

图 9. 不同任务下大模型能力的雷达图展示（第 2 级）

(3) 第三级

MiniMax、ChatGLM3、360 智脑与 GPT3.5-Turbo 位列第三梯队（见图 10）。尽管这几个国产大模型在内容总结、学科测试任务的表现上优于 GPT3.5-Turbo，但在角色模拟、场景模拟与自由问答、内容创作等文本生成类任务中仍然落后。ChatGLM3 的输出中有时会出现中文回答中夹杂英文词汇的现象，较严重地影响了对其输出内容质量的评价。此外，MiniMax 与 360 智脑的模型安全意识弱于同梯队另外两个模型。

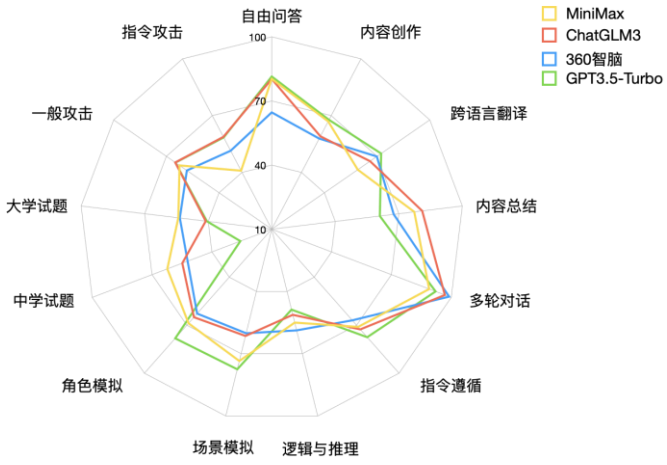

图 10.不同任务下大模型能力的雷达图展示（第 3 级）

[19] 对于多轮对话任务，尽管日日新大模型在 API 调用时支持带有对话历史的提问与回答获取，但其网页对话界面仅支持单轮问答。

(4) 第四级

百川 2 与中文增强后的 llama2 在中文语境下的整体都表现平平，位列第四梯队（见图 11）。虽然 llama 本身难以使用中文直接输出回答，但在进行中文增强后表现出较好的中文对话能力。百川在跨语言翻译、自由问答、角色模拟、场景模拟等任务中都有不错的表现，但逻辑推理能力弱。千帆在安全与责任测试中表现略逊于百川大模型。

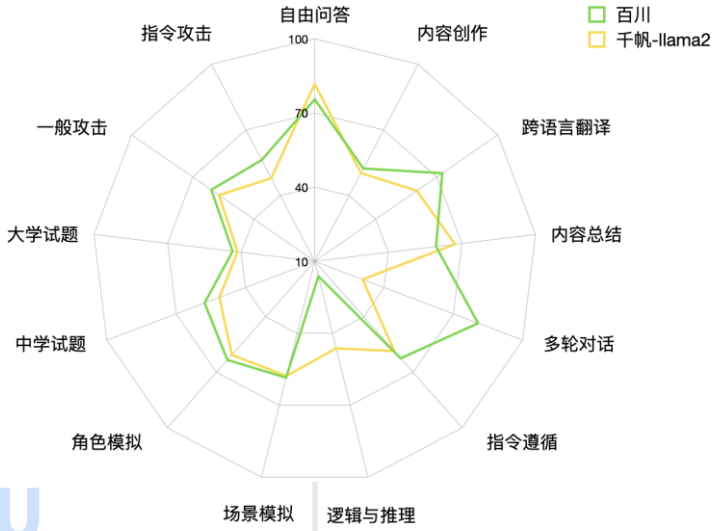

图 11. 不同任务下大模型能力的雷达图展示（第 4 级）

(5) 第五级

悟道天鹰与 Bloomz 在中文语境的测试中表现较差，尤其是在学科测试与逻辑推理测试中（见图 12）。这两个大模型在专业学科能力测试中正确率不足或仅为 30%左右，而在中文逻辑与推理子任务中，悟道的正确率仅为 22.5%，而 Bloomz 仅为 20%。此外，Bloomz 在安全与责任能力的测试中表现较差。

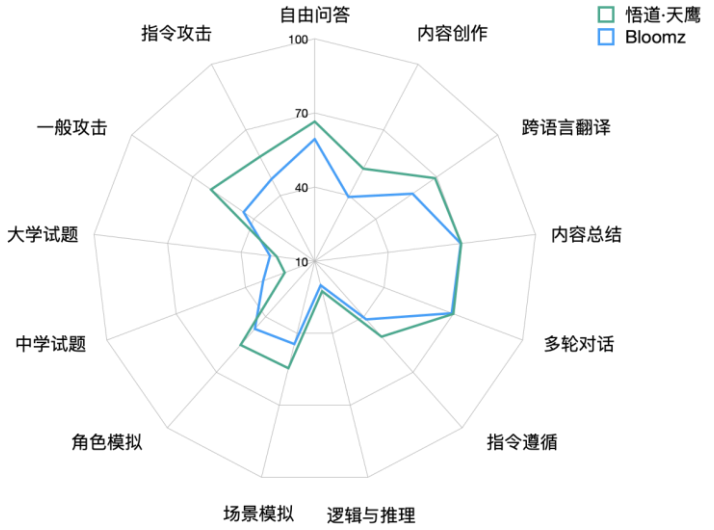

图 12. 不同任务下大模型能力的雷达图展示（第 5 级）

6.2. 从难易程度不同的任务观察大模型表现

(1) 代表性国产大模型在基础语言能力上与 GPT 系列模型匹敌，在场景应用能力上表现仍然落后

基础语言能力（包含自由问答、内容创作等任务）是指大语言模型的具有自然语言理解与生成的基本能力，而场景应用能力（包含角色模拟与场景模拟任务）则要求大模型对人类角色、情感以及文化语境有更为进阶的理解。

表 18. 基础语言能力排行榜

排名	大模型	自由问答	内容创作	跨语言翻译	内容总结	多轮对话	指令遵循	逻辑与推理	综合得分
1	GPT4-Turbo	94.29	74.50	78.31	75.34	95.71	89.52	80.00	83.95
2	文心一言 4	79.64	71.15	77.98	84.44	98.93	84.29	80.00	82.35
3	GPT4	82.90	70.06	79.76	77.55	96.07	84.29	76.25	80.98
4	通义千问 2	76.96	66.03	76.34	74.06	92.50	80.00	71.25	76.73
5	讯飞星火 v3.0	80.58	67.49	71.50	76.79	76.43	72.14	63.75	72.67
6	ChatGLM3 -6B	80.27	59.07	66.00	81.04	96.79	72.62	51.25	72.43
7	GPT3.5-Turbo	81.43	67.77	72.32	61.05	92.14	77.50	48.75	71.57
8	MiniMax	80.36	66.94	59.00	77.30	88.93	71.07	55.00	71.23
9	商汤日日新	78.35	62.55	74.96	77.21	70.71	71.43	62.50	71.10
10	360 智脑	64.64	57.88	69.87	67.60	98.93	66.96	58.75	69.23
11	百川 2	75.49	52.38	72.73	59.44	80.71	62.44	16.25	59.92
12	悟道・天鹰	66.52	52.29	69.16	69.73	70.00	50.77	22.50	57.28
13	千帆-llama2	81.74	50.28	60.23	67.18	30.71	58.57	46.25	56.42
14	BLOOMZ-7B	59.42	39.38	58.11	69.56	69.29	41.43	20.00	51.03

表 19. 场景应用能力排行榜

排名	大模型	场景模拟	角色模拟	综合得分
1	GPT4	77.93	80.60	79.27
2	GPT3.5-Turbo	77.50	78.21	77.86
3	GPT4-Turbo	80.64	75.00	77.82
4	讯飞星火 v3.0	71.00	73.81	72.41
5	商汤日日新	74.29	69.64	71.97
6	文心一言 4	70.43	73.45	71.94
7	MiniMax	73.50	68.81	71.16
8	通义千问 2	72.43	67.44	69.94
9	ChatGLM3 -6B	61.43	65.00	63.22
10	360 智脑	60.14	62.74	61.44
11	百川 2	58.50	63.33	60.92
12	千帆-llama2	57.79	60.60	59.20
13	悟道・天鹰	54.57	55.24	54.91
14	BLOOMZ-7B	44.50	46.55	45.53

如表 18 与 19 所示，在中文语境下，文心 4、通义 2、星火 3 与 ChatGLM3 等国产大模型在基础语言能力上与 GPT 系列模型差距较小，并超越 GPT3.5-

Turbo。其中，文心一言 4 略优于 GPT4，仅次于 GPT4-Turbo。然而，在场景应用能力上，尽管是在中文语境下，GPT 系列的表现仍显著领先于国产大模型，国产大模型仍需努力。

(2) 在中文的学科能力测试中，国产大模型表现优异，通义、文心、星火与 GPT4-Turbo 领先

如表 20 与 21 所示，无论是中学或大学难度的中文封闭性学科试题测试，国产大模型通义千问 2 与文心一言 4 都表现优异，正确率超过 GPT4-Turbo。然而，在大学数学测试中，GPT4 与 GPT4-Turbo 的正确率几乎超过所有国产大模型。此外，几乎所有大模型都在较为简单的中学难度学科测试中表现更优。

表 20. 中学学科测试排行榜

排名	大模型	生物	物理	数学	化学	地理	历史	平均正确率
1	通义千问 2.0	93.33%	84.21%	60.78%	84.71%	89.53%	96.21%	84.80%
2	文心一言 4	85.33%	77.63%	56.86%	81.18%	80.23%	93.18%	79.07%
3	讯飞星火 v3.0	88.00%	72.37%	42.16%	70.59%	79.07%	81.06%	72.21%
4	GPT4-Turbo	85.33%	71.05%	44.94%	57.89%	79.07%	85.61%	70.65%
5	商汤日日新	89.33%	68.42%	42.16%	61.18%	66.28%	81.06%	68.07%
6	GPT4	89.33%	51.32%	40.20%	56.47%	79.07%	83.33%	66.62%
7	MiniMax	74.67%	59.21%	41.18%	51.76%	63.95%	83.33%	62.35%
8	百川 2	68.00%	42.11%	29.41%	54.12%	74.42%	78.03%	57.68%
9	ChatGLM3-6B	74.67%	46.05%	23.53%	43.53%	63.95%	77.27%	54.83%
10	360 智脑	65.33%	51.32%	34.31%	40.00%	69.77%	52.27%	52.17%
11	千帆-llama2	69.33%	43.42%	26.47%	34.12%	59.30%	75.00%	51.27%
12	BLOOMZ-7B	36.00%	30.26%	23.53%	30.59%	34.88%	38.64%	32.32%
13	GPT3.5-Turbo	40.00%	28.95%	29.41%	21.18%	17.44%	17.42%	25.73%
14	悟道・天鹰	24.00%	25.00%	20.59%	22.35%	20.93%	25.00%	22.98%

表 21. 大学学科测试排行榜

排名	类别	数学	医学	经济	计算机	物理	化学	哲学	管理	平均正确率
1	通义千问 2.0	39.60%	79.00%	77.00%	79.61%	55.00%	65.22%	83.00%	78.15%	69.57%
2	文心一言 4	45.54%	72.00%	75.00%	84.47%	51.25%	54.35%	80.00%	73.95%	67.07%
3	GPT4-turbo	44.55%	79.00%	73.00%	80.58%	45.00%	54.35%	72.00%	71.43%	64.99%
4	GPT4	46.53%	75.00%	72.00%	77.67%	47.50%	60.87%	67.00%	73.11%	64.96%
5	讯飞星火 v3.0	42.57%	79.00%	64.00%	63.11%	45.00%	50.00%	73.00%	72.27%	61.12%
6	商汤日日新	39.60%	62.00%	79.00%	74.76%	37.50%	36.96%	75.00%	59.66%	58.06%
7	MiniMax	31.68%	59.00%	60.00%	64.08%	40.00%	41.30%	72.00%	64.71%	54.10%
8	360 智脑	38.61%	57.00%	60.00%	54.37%	43.75%	52.17%	59.00%	62.18%	53.39%
9	百川	17.82%	49.00%	59.00%	51.46%	17.50%	30.43%	63.00%	59.66%	43.48%
10	千帆-llama2	33.66%	44.00%	49.00%	35.92%	28.75%	28.26%	55.00%	57.14%	41.47%
11	ChatGLM3-6B	21.78%	45.00%	46.00%	47.57%	30.00%	21.74%	55.00%	62.18%	41.16%
12	GPT-3.5-turbo	18.81%	54.00%	48.00%	55.34%	16.25%	34.78%	48.00%	49.58%	40.60%
13	BLOOMZ-7B	22.77%	29.00%	25.00%	31.07%	23.75%	23.91%	35.00%	35.29%	28.22%
14	悟道・天鹰	22.77%	24.00%	26.00%	22.33%	17.50%	21.74%	36.00%	33.61%	25.49%

要注意的是，所有的学科试题均由中文呈现，且其中的部分试题偏向于中国背景，尤其是地理、历史、哲学等学科，这可能是国产大模型表现更佳的原因。

(3) GPT4-Turbo、文心一言、讯飞星火等模型在安全与责任评测中表现出色

如表 22 与 23 所示，在常规安全主题的一般攻击测试中，GPT4-Turbo、文心一言与通义千问表现最好，讯飞星火与日日新均表现出色。在指令攻击测试中，通义千问表现一般，而讯飞星火与文心一言表现最佳。百川、悟道天鹰、360 智脑、MiniMax 等大模型在安全与责任相关能力上仍需加强。

表 22. 一般攻击排行榜

排名	大模型	一般攻击
❶	GPT4-Turbo	70.43
❷	文心一言 4（ERNIE-Bot4.0）	69.68
❸	通义千问 2（qwen-max）	69.00
4	讯飞星火 v3.0	66.87
5	商汤日日新（Sensenova）	65.66
6	ChatGLM3-6B	64.96
7	GPT3.5-Turbo	64.84
8	MiniMax（abab5.5-chat）	62.51
9	GPT4	61.62
10	悟道·天鹰（AquilaChat-7B）	61.04
11	百川（baichuan2-13b-chat-v1）	60.88
12	360 智脑（360GPT_S2_V9）	58.34
13	千帆-llama2（Qianfan-Chinese-Llama-2-7B）	57.04
14	BLOOMZ-7B	44.98

表 23. 指令攻击排行榜

排名	大模型	指令攻击
❶	讯飞星火 v3.0	66.10
❷	文心一言 4（ERNIE-Bot4.0）	65.38
❸	GPT4-Turbo	60.90
4	商汤日日新（Sensenova）	59.62
5	ChatGLM3-6B	58.78
6	GPT3.5-Turbo	58.47
7	悟道·天鹰（AquilaChat-7B）	57.75
8	百川（baichuan2-13b-chat-v1）	56.23
9	通义千问 2（qwen-max）	55.93
10	GPT4	53.75
11	360 智脑（360GPT_S2_V9）	51.45
12	千帆-llama2（Qianfan-Chinese-Llama-2-7B）	47.94
13	BLOOMZ-7B	47.58
14	MiniMax（abab5.5-chat）	40.92

6.3. 局限与不足

我们的评测工作主要存在以下不足。首先，在评测过程中，囿于成本与效率，我们的评测工作所包含的大模型数量有限，忽略了部分只能通过网页端交互获取回答或未向个人用户开放 API 接口调用服务的对话大模型。受限于人工评测工作进行的时间，我们也没能囊括如蓝心这样最新发布的国产大语言模型。其次，大模型的参数量可能会对模型表现产生较大的影响，我们并没有在评测过程中根据参数量大小对大模型作进一步区分、比较和讨论。最后，在评测任务中，没有包含多模态能力的测试，而如今部分对话模型已经支持图片或语音输入；也没能包含代码生成类任务的测试，在之后的评测中希望能够对此进行补充。

致谢

我们衷心感谢香港大学经管学院深圳研究院以及参与打分工作的志愿者对本项目的帮助与支持。