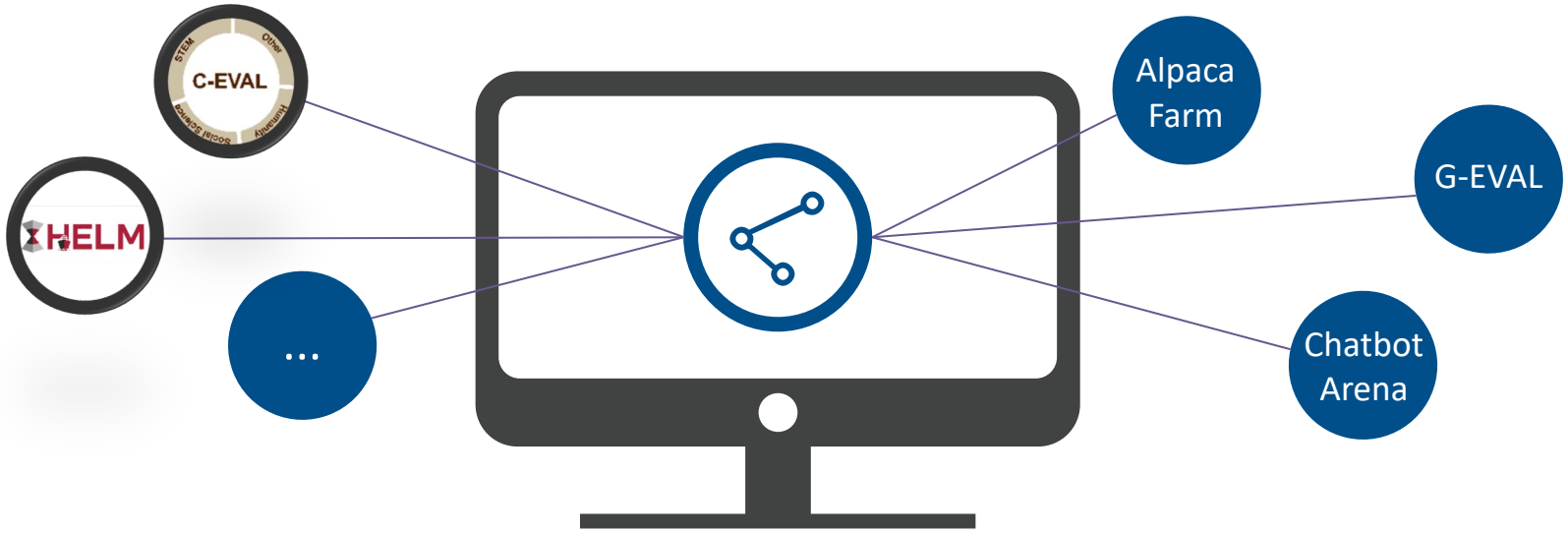
LLMEVAL-2
中文大语言模型评测第二期

复旦大学
自然语言处理实验室

前言



LLMEVAL-1已经告一段落，有大量的公众用户参与了进来，为我们的评测提供了详实的数据，我们也在数据收集阶段结束后提供了详细的评测报告（https://github.com/llmeval/llmeval-1）。

然而我们也发现由于这期评测希望更多的公众参与而在题目设计上选择了绝大多数用户都可以进行评价的问题，但是这类问题不能很好的反映模型的知识覆盖率。

因此，我们推出了LLMEVAL-2专业领域评测。LLMEVAL-2评测中以用户日常使用为主线，结合线上用户问题分布情况，重点考察不同专业本科生和研究生在日常学习和生活中，希望借助大模型得到帮助的任务。

目录

LLMEVAL-2数据集

- **测试范围：12个学科分别构造领域知识测试集**

- 对每个学科领域构造测试题集
- 题型为单项选择题与问答题
- **20个**开源及商业大模型，测试时间段为**7月5日至7月9日**。评测问题和各个参评系统的回答结果已经上传至 **https://github.com/llmeval/llmeval-2**

LLMEVAL-2评测方法

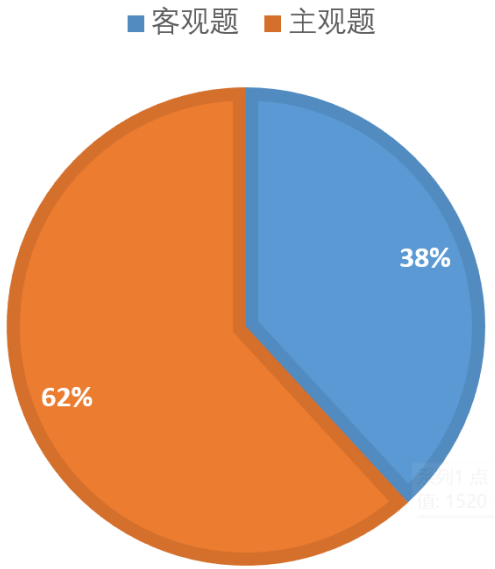
题目类型分布

每个学科设计：
- 约25-30道客观题
- 约10-15道主观题

合计480个题目

综合评价得分：
- 每个学科总分归一化为100分

评测方法：
人工评测+自动评测



评分标准

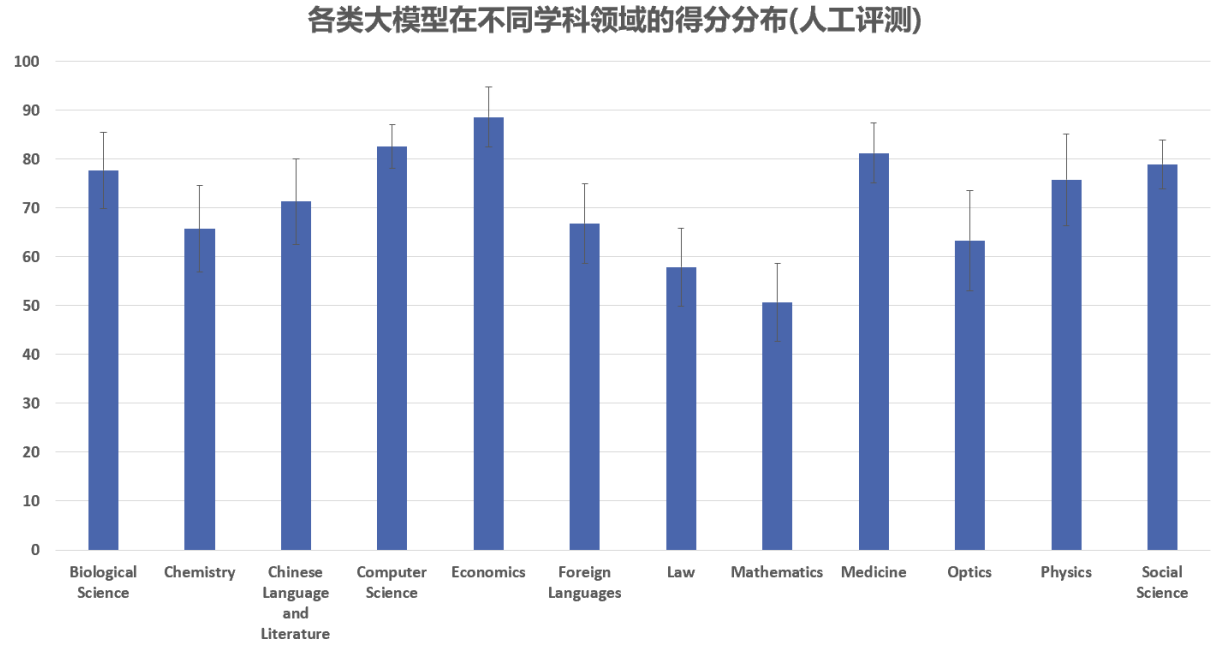
客观题：单选题或填空题
- 正确性（3分）：回答是否正确
- 解释正确性（2分）：是否生成了正确解释

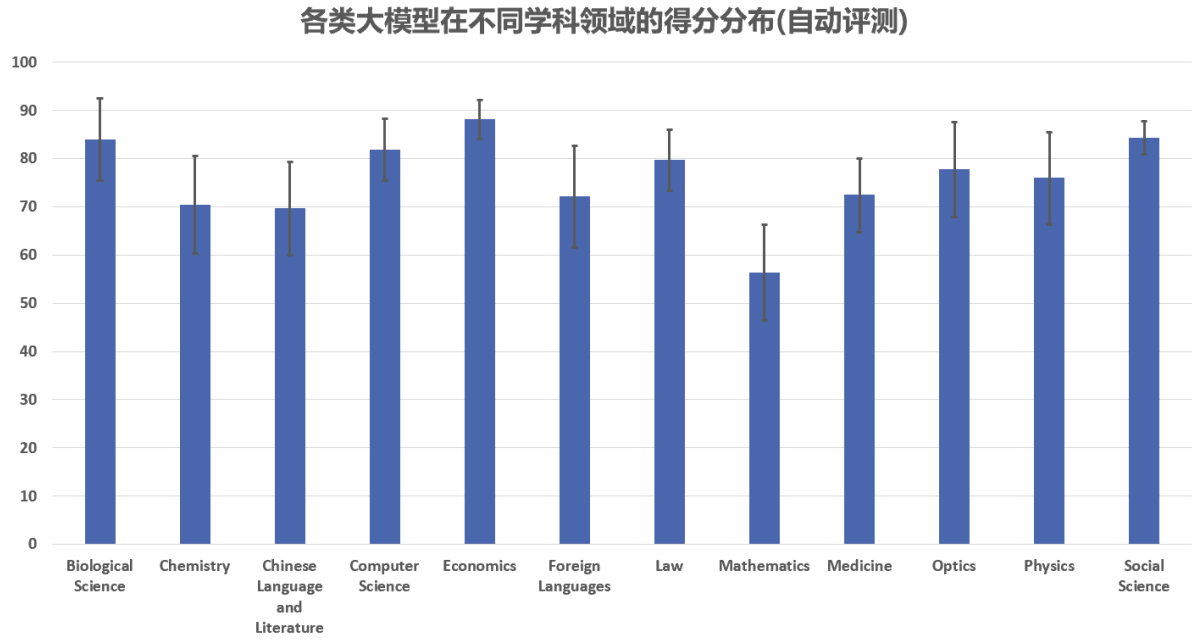
主观题：问答题（4个维度）：
- 准确性（5分）：回答内容是否有错
- 信息量（3分）：回答信息是否充足
- 流畅性（3分）：回答格式语法是否正确
- 逻辑性（3分）：回答逻辑是否严谨

评测结果

学科角度

- 大模型在不同学科问答能力表现差异较大；
- 数学学科平均得分最低，不同模型能力表现标准差较大；
- 经济学、计算机科学、药学平均得分较高；
- 人工评测和自动评测结果基本上保持一致；

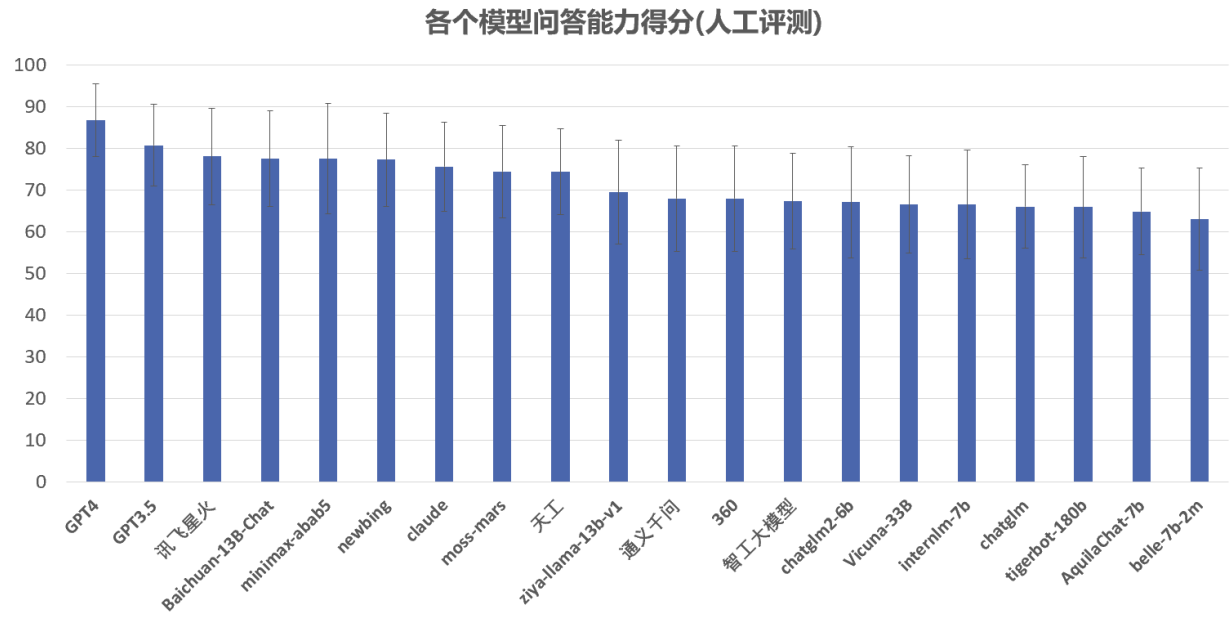


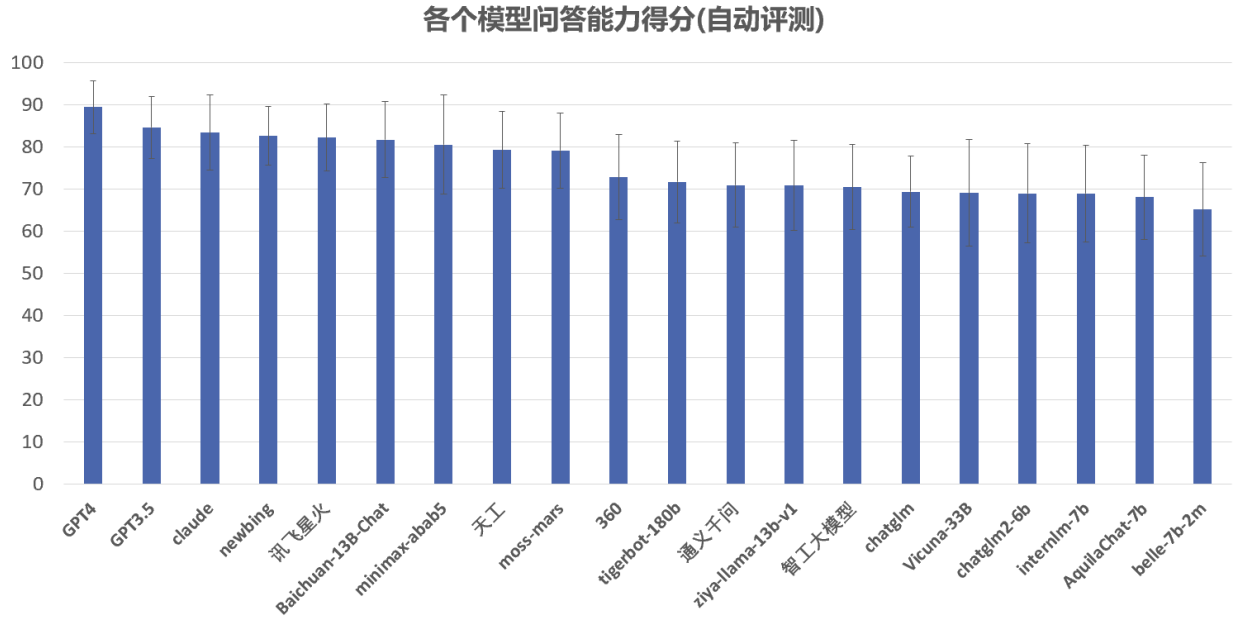


评测结果

模型角度

- GPT-4在主观题和客观题都具有明显优势；
- 很多模型距离GPT-3.5差距已经很少；
- 人工评测和自动评测基本保持一致，但是模型之间微小的分差两者之间存在差异；



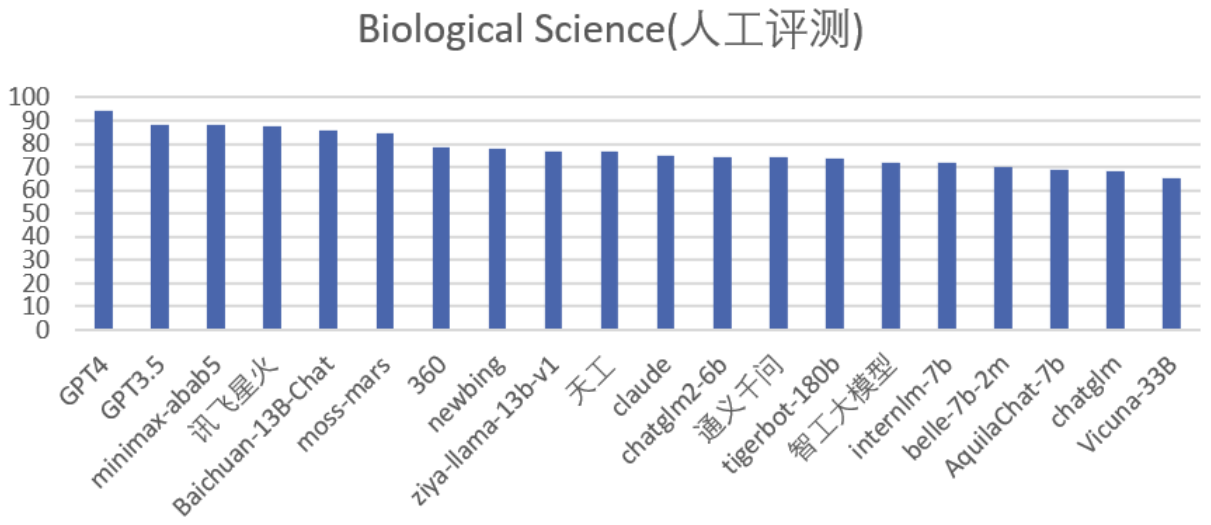


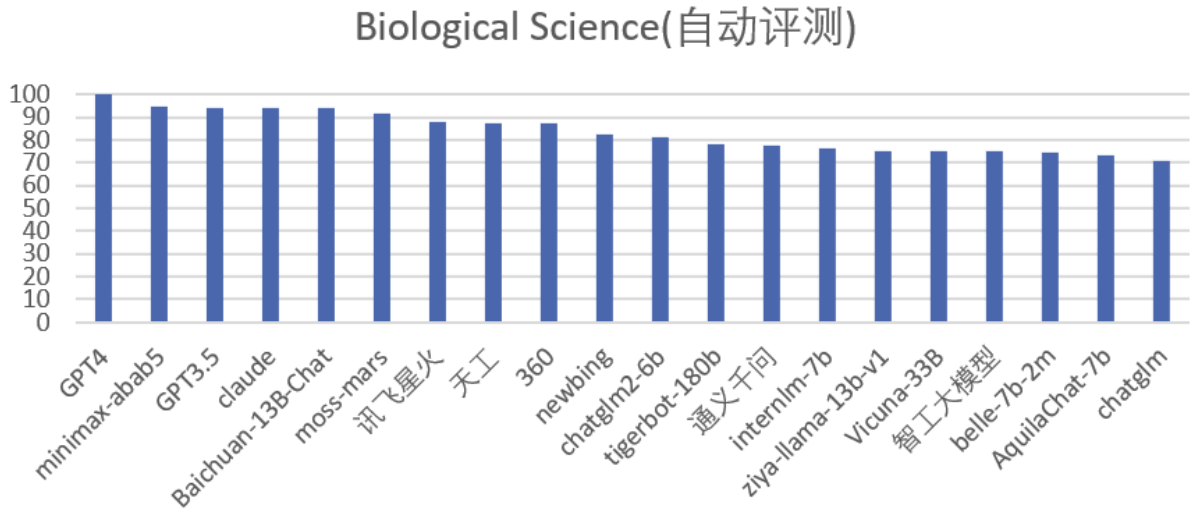
注: 图中提及大模型测试版本号为GPT4(gpt-4-0314), GPT3.5(gpt-3.5-turbo-0301), 讯飞星火(v1.5), Baichuan-13B-Chat, minimax-abab5(chat v1), newbing(Bing Chat), Claude(Claude-2-100k),moss-mars(v0.0.3), 天工(天工大模型v3.5.20230705.a), ziya-llama-13b(v1), 通义千问(1.0.3), 360(360智脑beta-2.00)), 智工大模型, ChatGLM2-6b(v1.1.0), Vicuna-33b(v1.3), ChatGLM(ChatGLM-130B-v0.8), TigerBot-180B (research version) , AquilaChat-7B(v0.6), belle-7b-2(v0.95)

评测结果

各个学科领域评测结果

- 生命科学
- 化学
- 汉语言文学
- 物理学
- 外语
- 药学
- 社会科学
- 光学
- 经济学
- 法学
- 数学
- 计算机科学

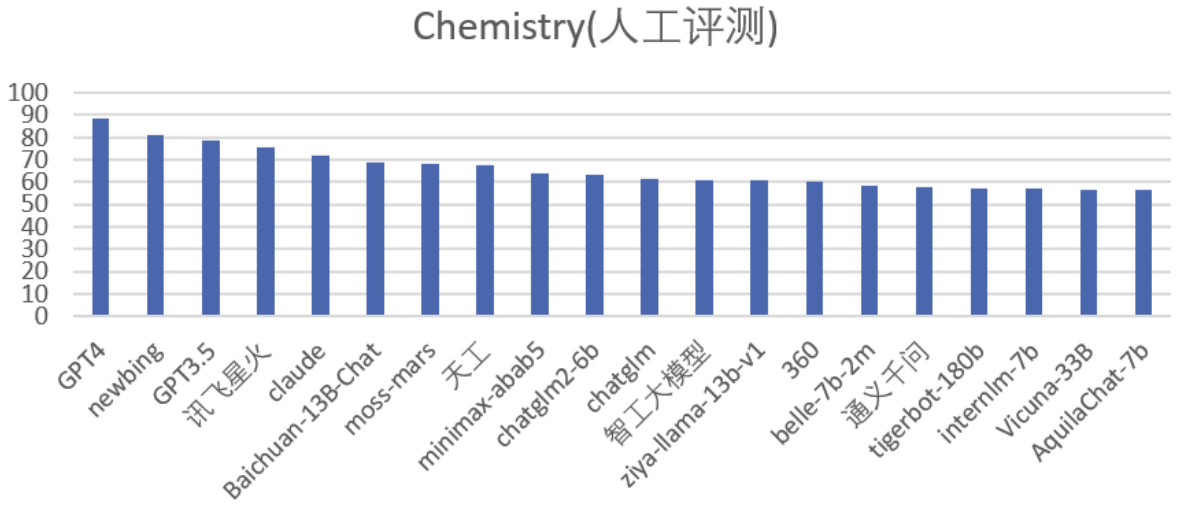


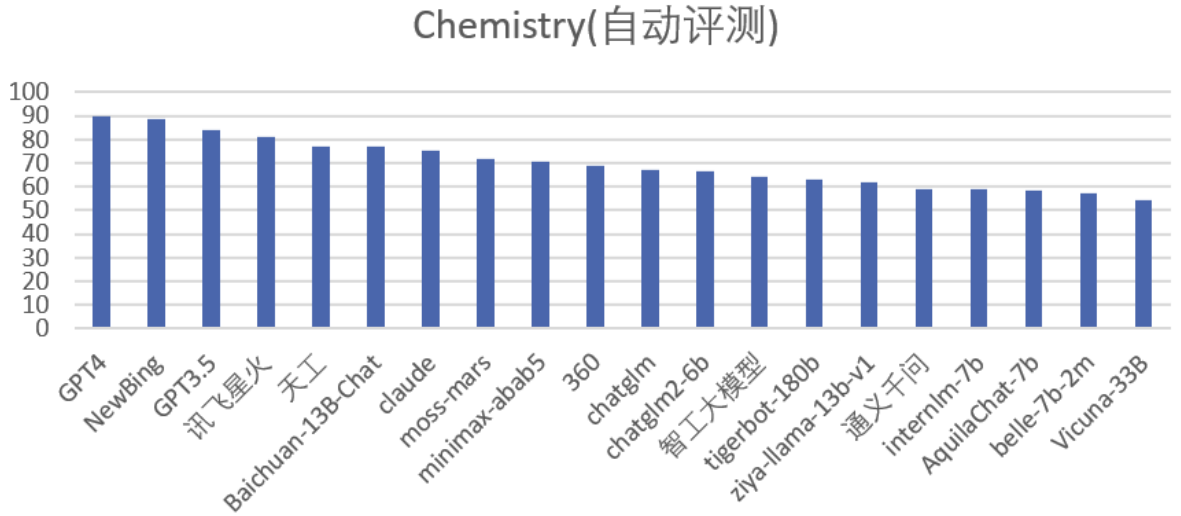


评测结果

各个学科领域评测结果

- 生命科学
- 化学
- 汉语言文学
- 物理学
- 外语
- 药学
- 社会科学
- 光学
- 经济学
- 法学
- 数学
- 计算机科学

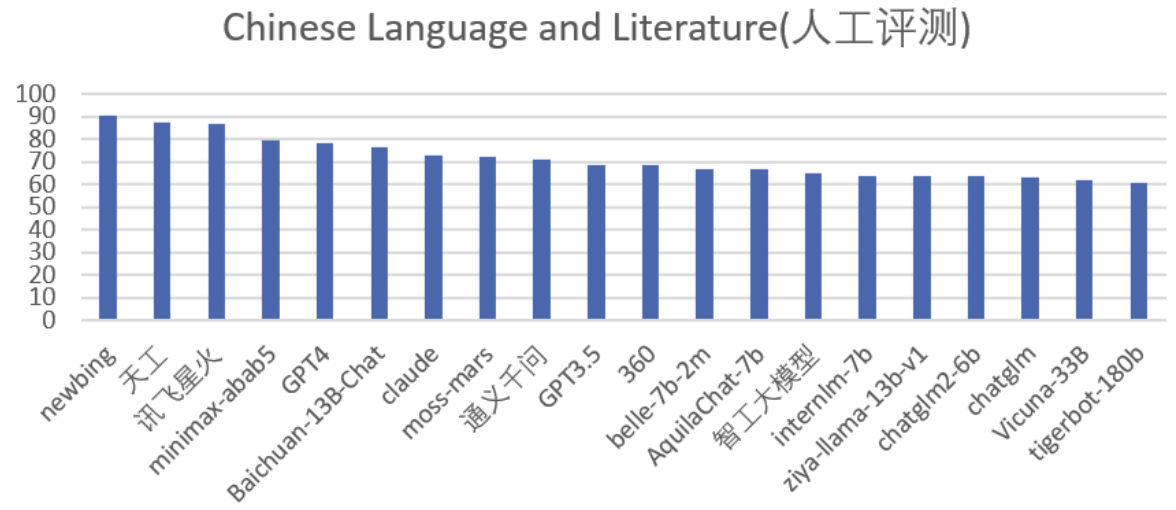


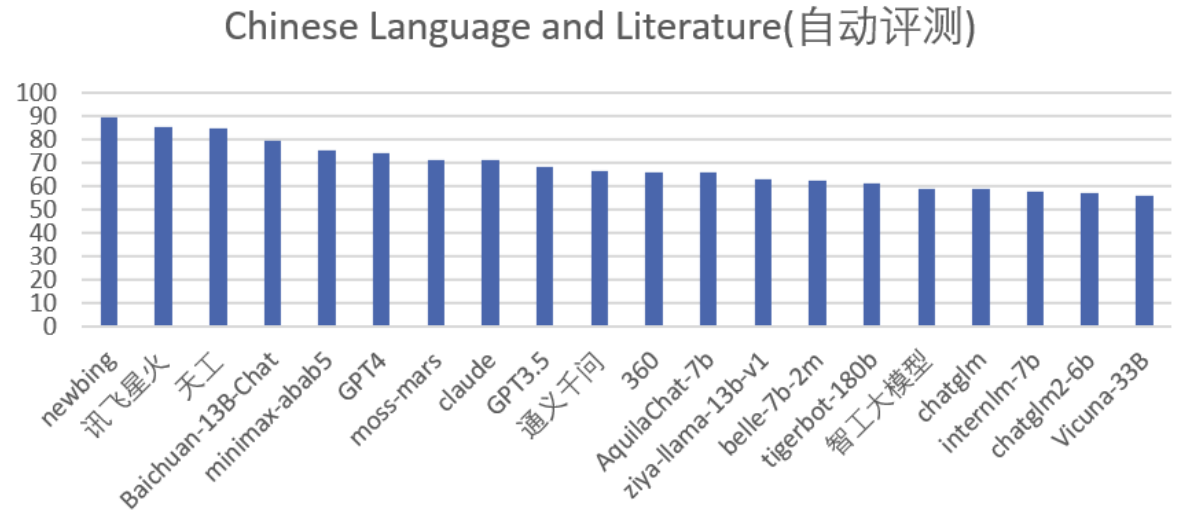


评测结果

各个学科领域评测结果

- 生命科学
- 化学
- 汉语言文学
- 物理学
- 外语
- 药学
- 社会科学
- 光学
- 经济学
- 法学
- 数学
- 计算机科学

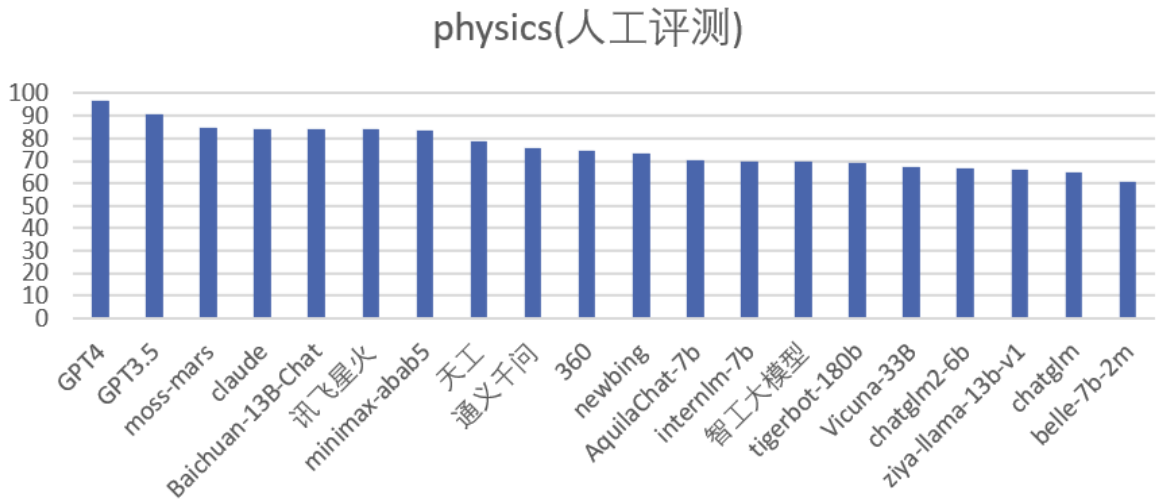


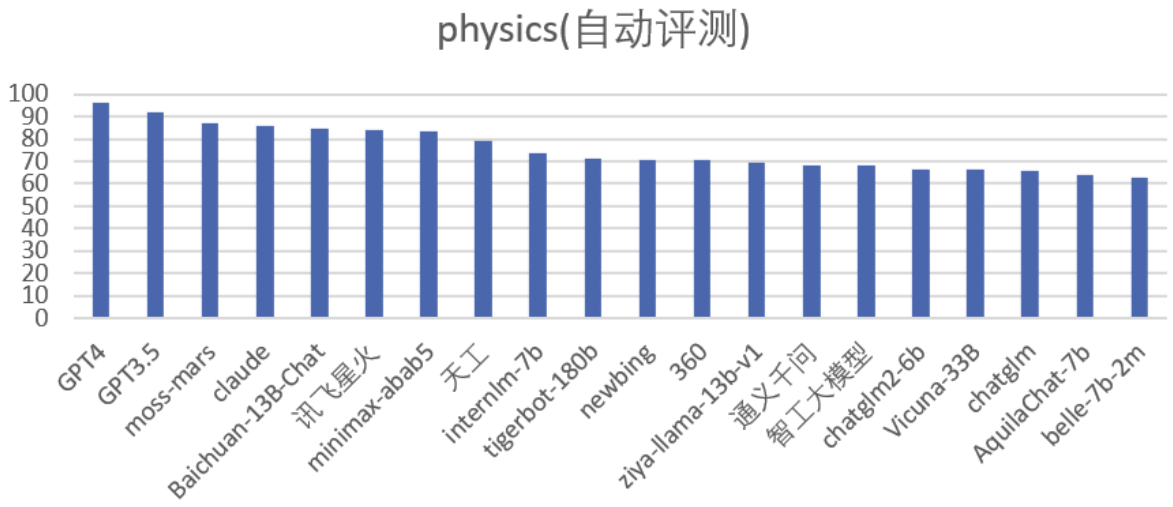


评测结果

各个学科领域评测结果

- 生命科学
- 化学
- 汉语言文学
- 物理学
- 外语
- 药学
- 社会科学
- 光学
- 经济学
- 法学
- 数学
- 计算机科学





评测结果

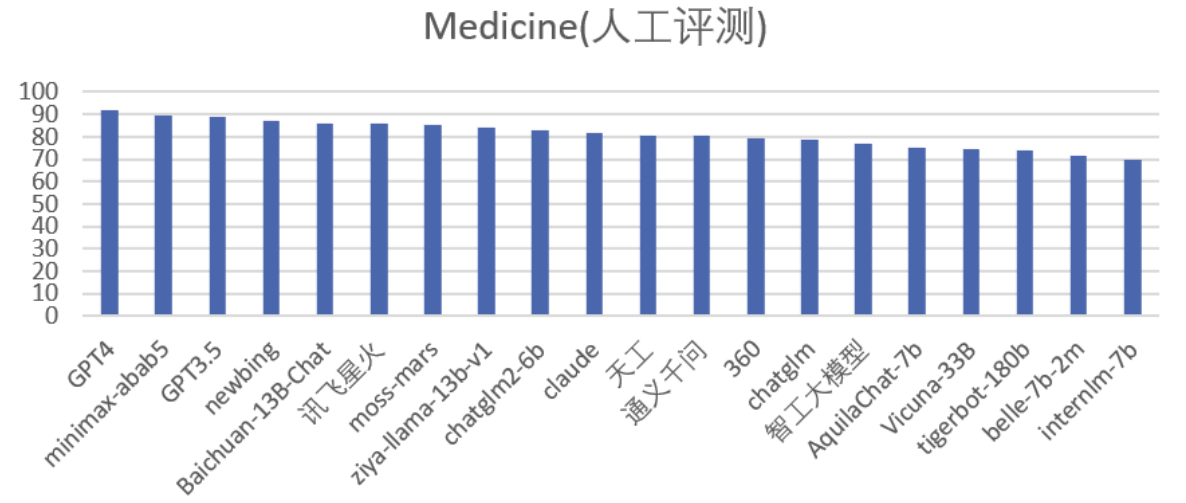
各个学科领域评测结果

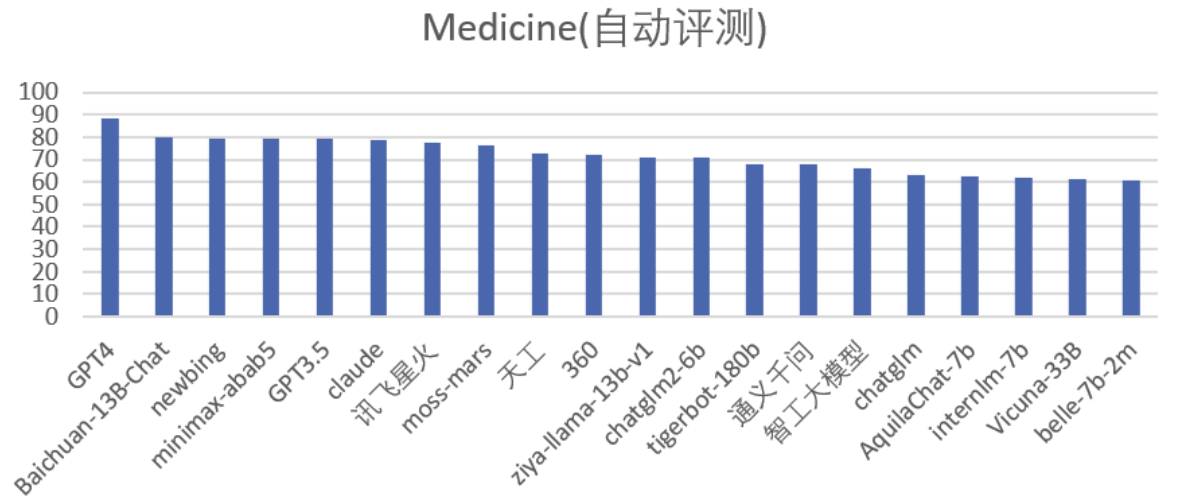
- 生命科学
- 化学
- 汉语言文学
- 物理学
- 外语
- 药学
- 社会科学
- 光学
- 经济学
- 法学
- 数学
- 计算机科学

评测结果

各个学科领域评测结果

- 生命科学
- 化学
- 汉语言文学
- 物理学
- 外语
- **药学**
- 社会科学
- 光学
- 经济学
- 法学
- 数学
- 计算机科学

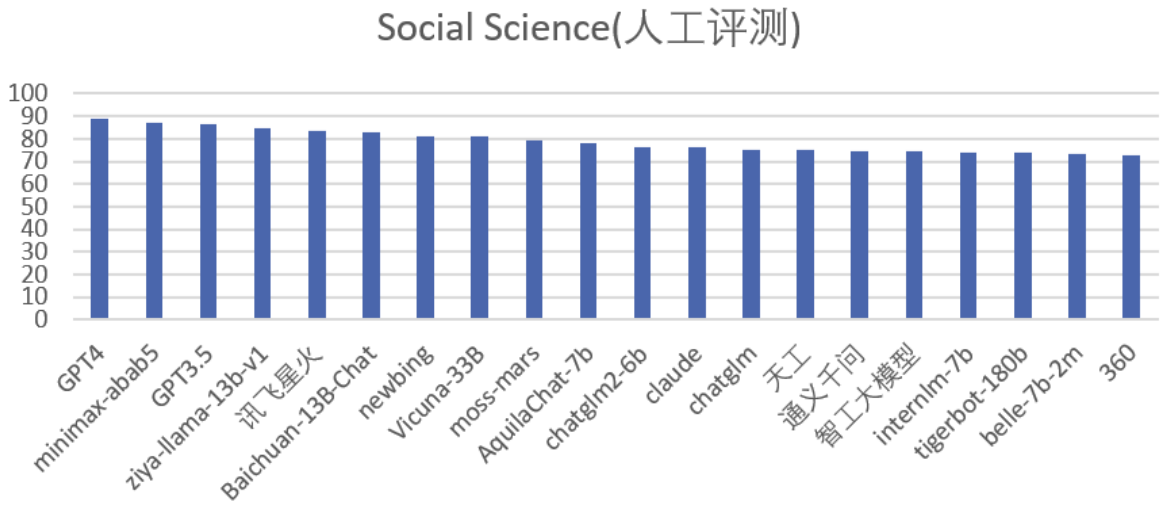


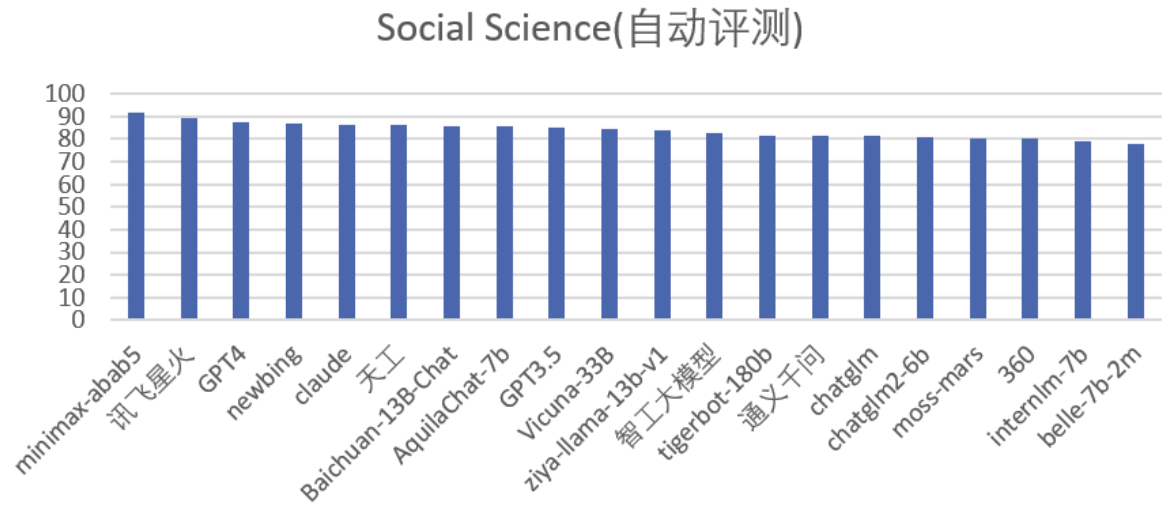


评测结果

各个学科领域评测结果

- 生命科学
- 化学
- 汉语言文学
- 物理学
- 外语
- 药学
- 社会科学
- 光学
- 经济学
- 法学
- 数学
- 计算机科学

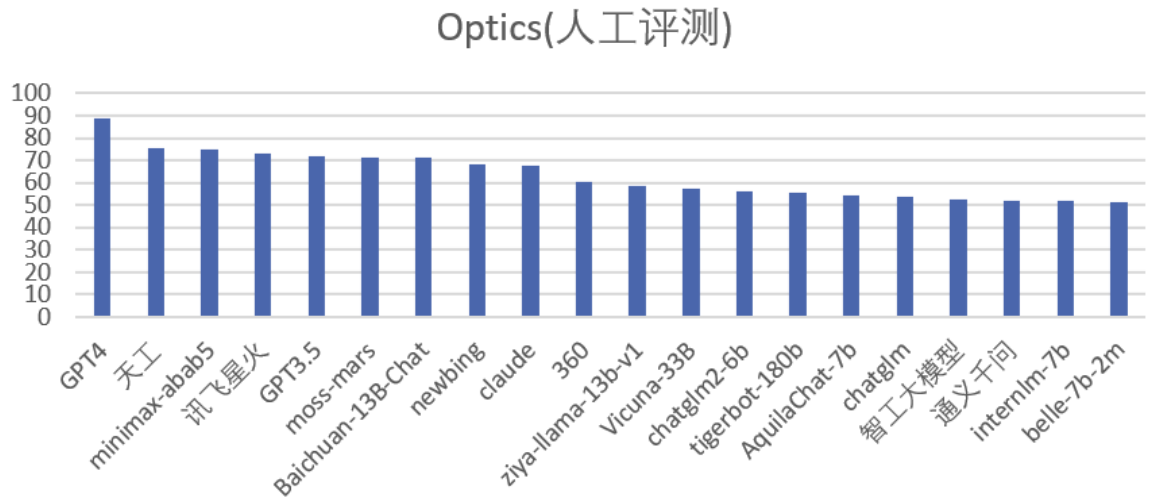


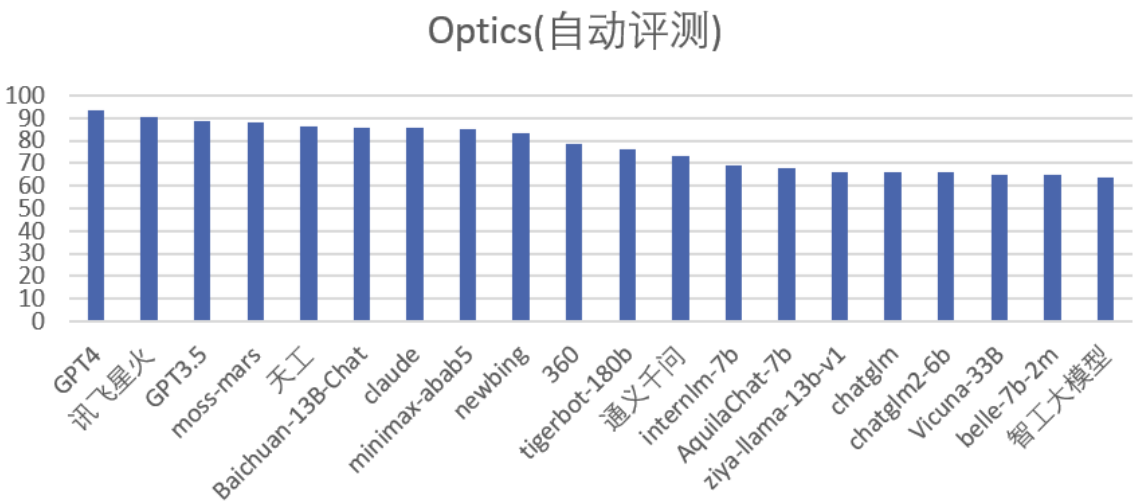


评测结果

各个学科领域评测结果

- 生命科学
- 化学
- 汉语言文学
- 物理学
- 外语
- 药学
- 社会科学
- 光学
- 经济学
- 法学
- 数学
- 计算机科学

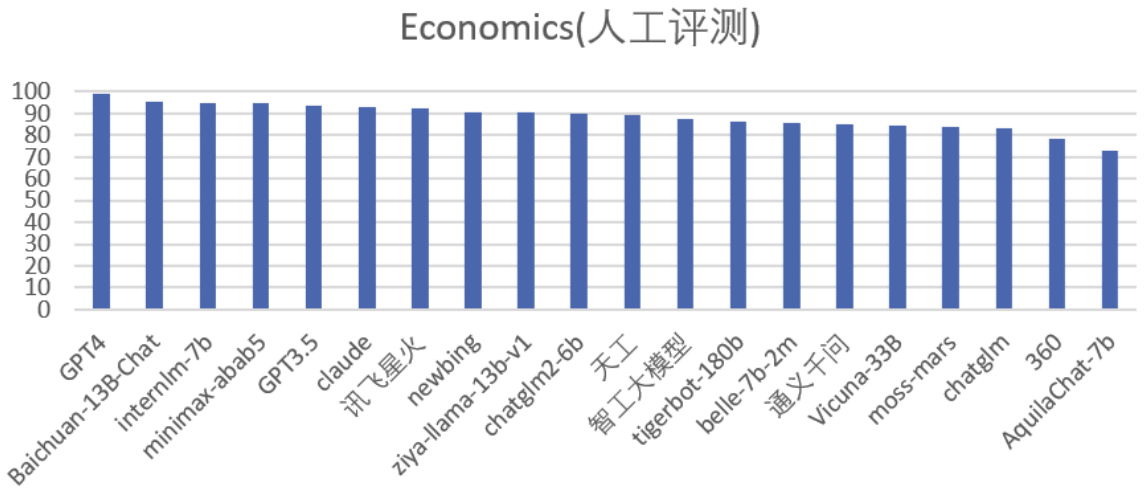


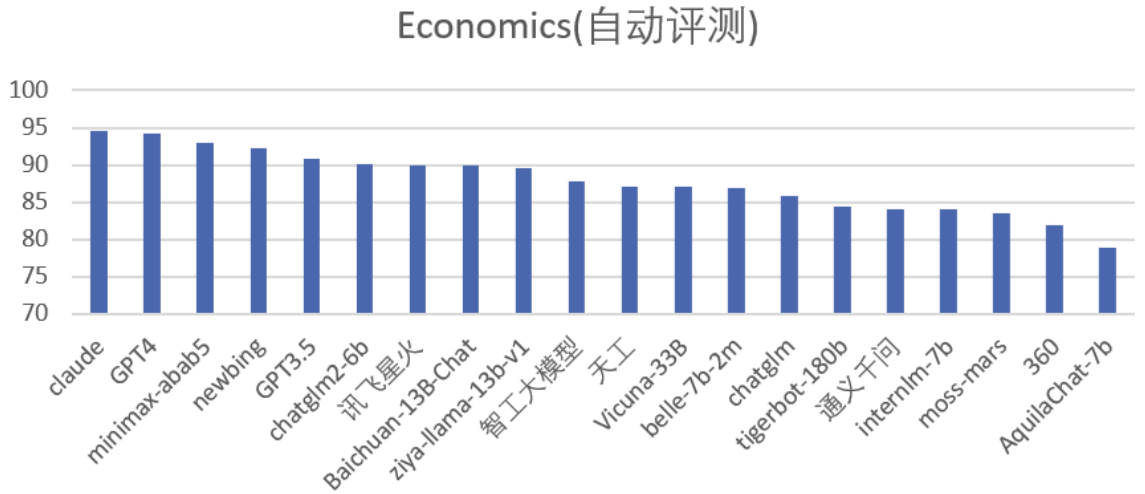


评测结果

各个学科领域评测结果

- 生命科学
- 化学
- 汉语言文学
- 物理学
- 外语
- 药学
- 社会科学
- 光学
- 经济学
- 法学
- 数学
- 计算机科学

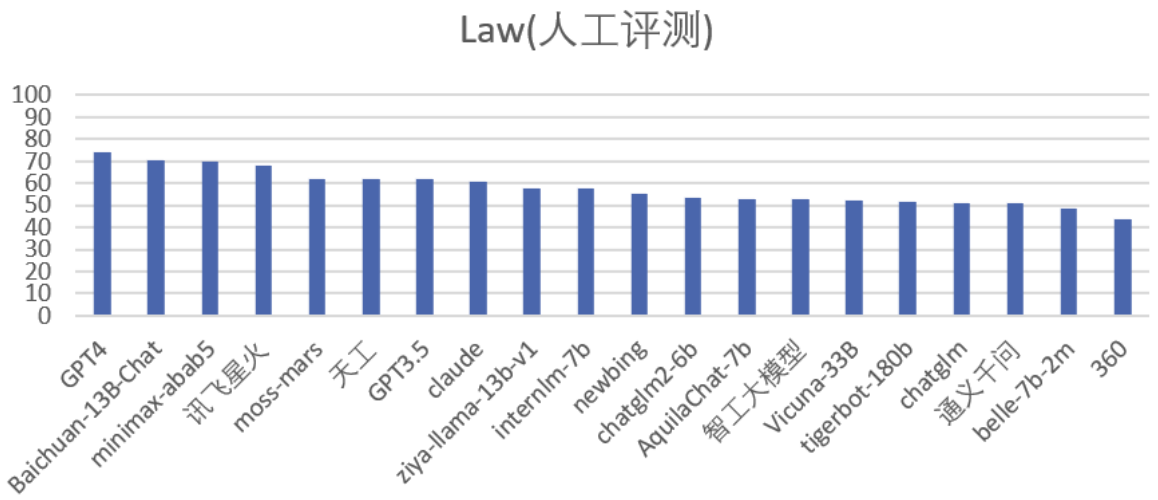


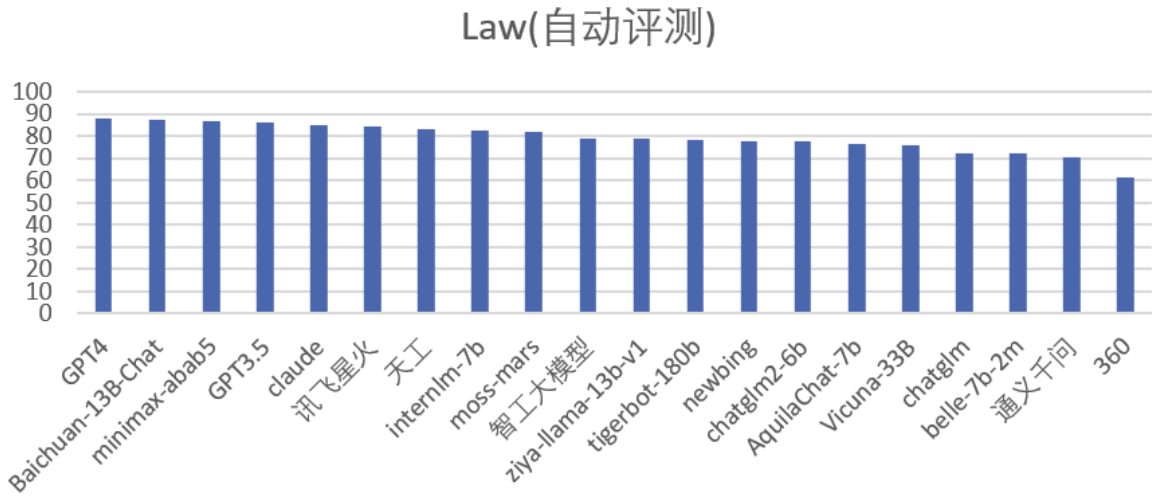


评测结果

各个学科领域评测结果

- 生命科学
- 化学
- 汉语言文学
- 物理学
- 外语
- 药学
- 社会科学
- 光学
- 经济学
- 法学
- 数学
- 计算机科学

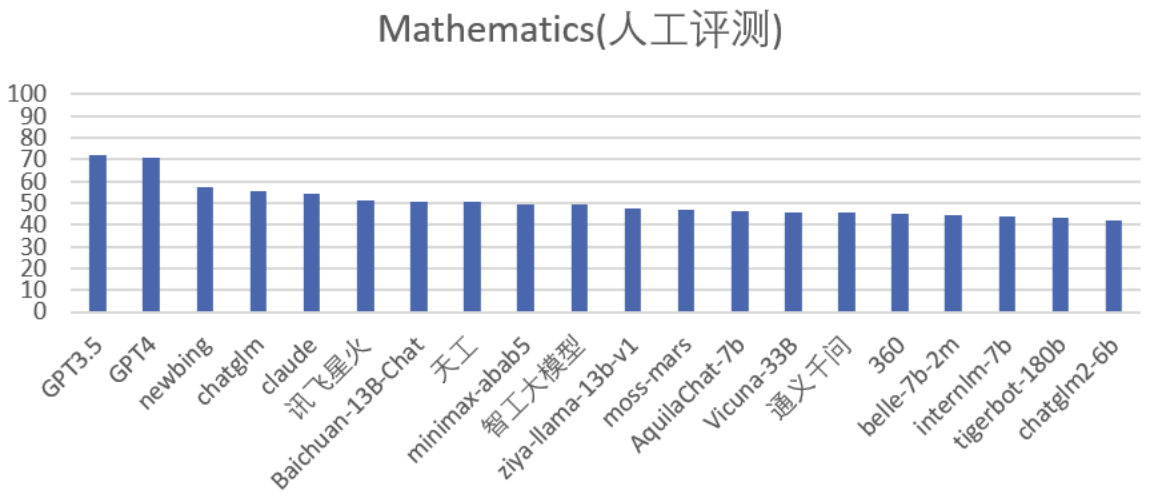


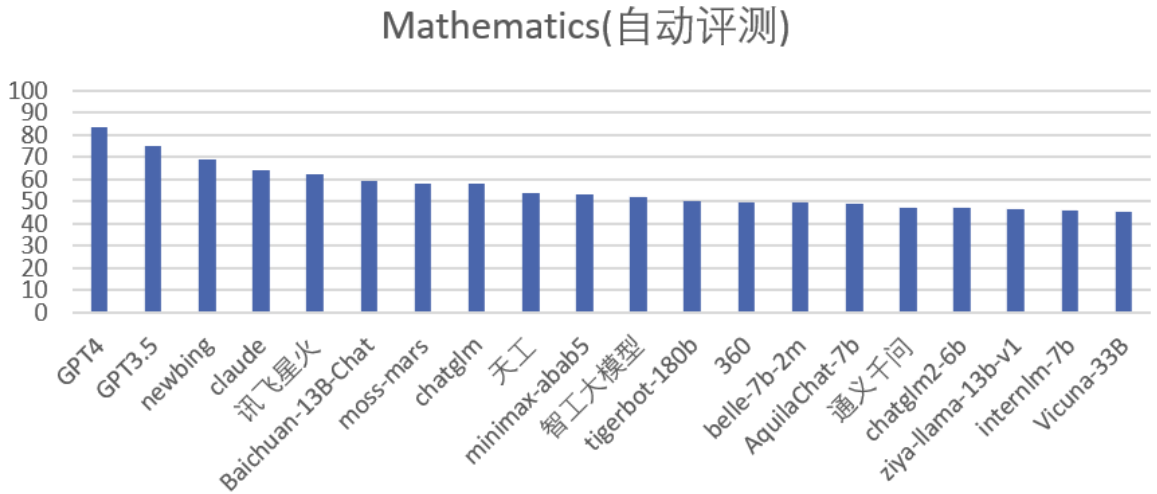


评测结果

各个学科领域评测结果

- 生命科学
- 化学
- 汉语言文学
- 物理学
- 外语
- 药学
- 社会科学
- 光学
- 经济学
- 法学
- 数学
- 计算机科学

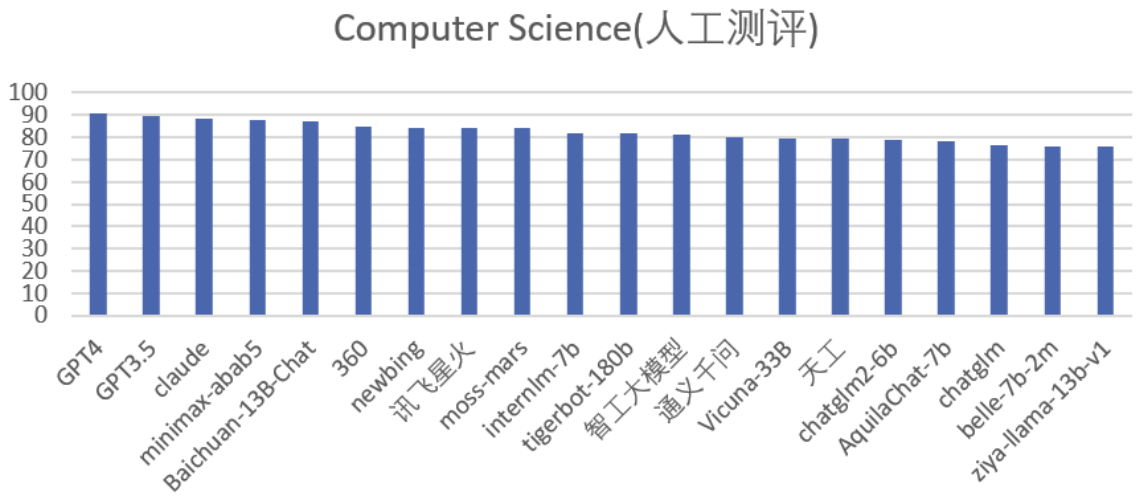


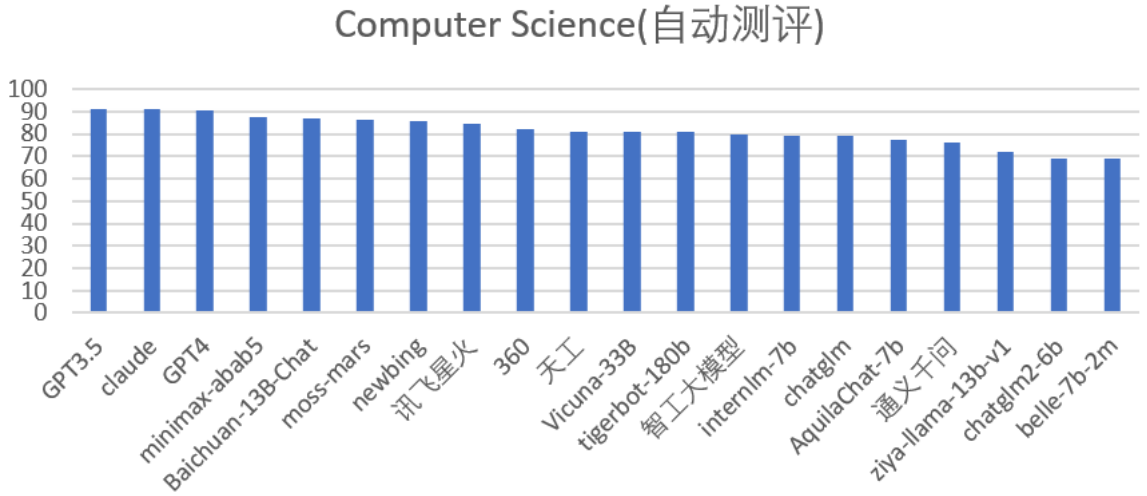


评测结果

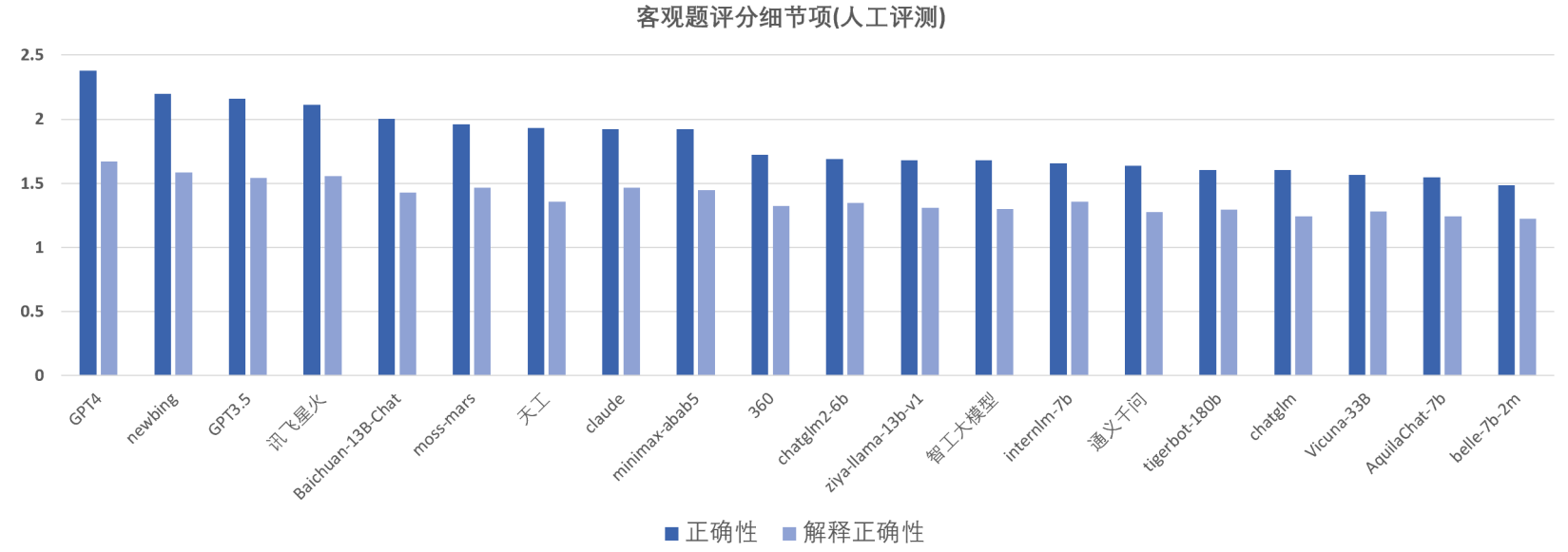
各个学科领域评测结果

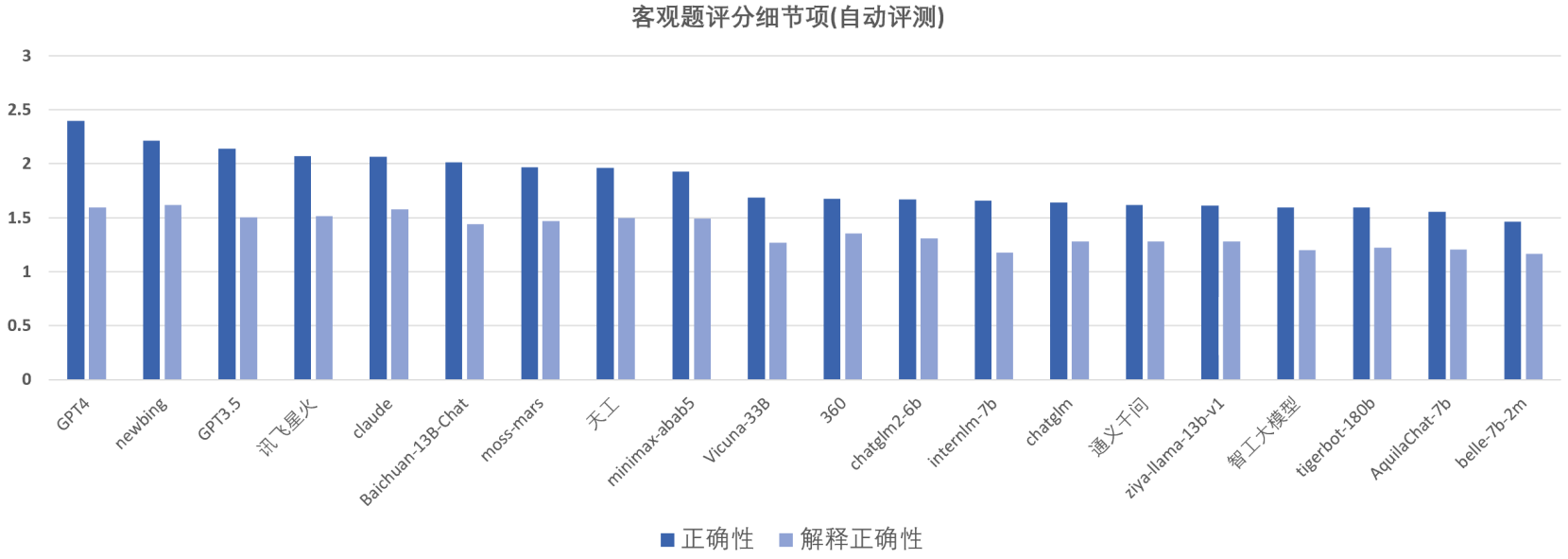
- 生命科学
- 化学
- 汉语言文学
- 物理学
- 外语
- 药学
- 社会科学
- 光学
- 经济学
- 法学
- 数学
- 计算机科学





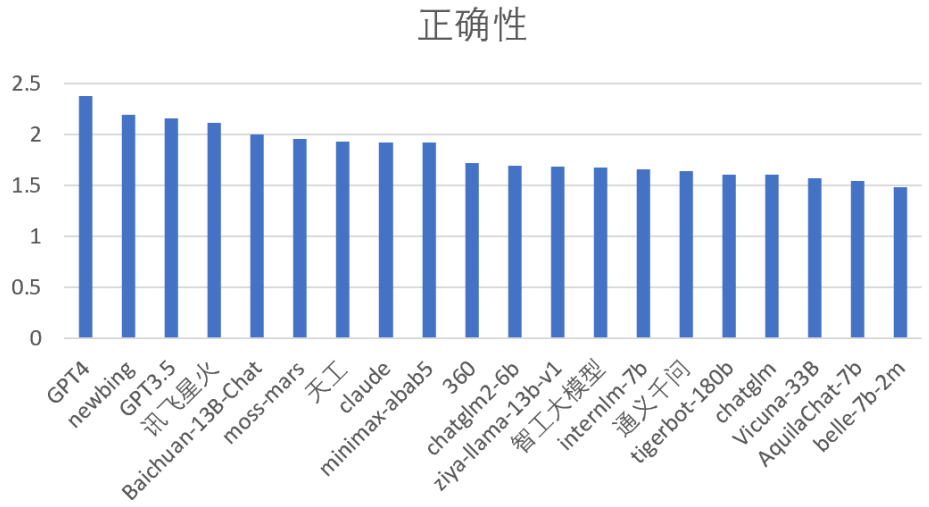
评测结果

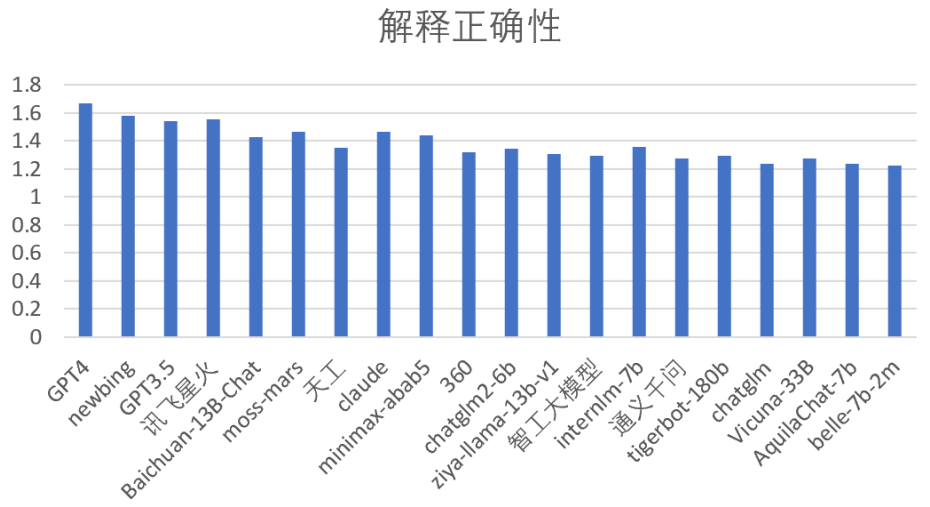

客观题评分细节项(人工评测)
2.5
2
1.5
1
0.5
0
GPT4
newbing
GPT3.5
讯飞星火
Baichuan-13B-Chat
moss-mars
天工
claude
minimax-abab5
360
chatglm2-6b
ziya-llama-13b-v1
智工大模型
internlm-7b
通义千问
tigerbot-180b
chatglm
Vicuna-33B
AquilaChat-7b
belle-7b-2m
正确性
解释正确性


客观题评分细节项(自动评测)
3
2.5
2
1.5
1
0.5
0
GPT4
newbing
GPT3.5
讯飞星火
claude
Baichuan-13B-Chat
moss-mars
天工
minimax-abab5
Vicuna-33B
360
chatglm2-6b
internlm-7b
chatglm
通义千问
ziya-llama-13b-v1
智工大模型
tigerbot-180b
AquilaChat-7b
belle-7b-2m
正确性
解释正确性

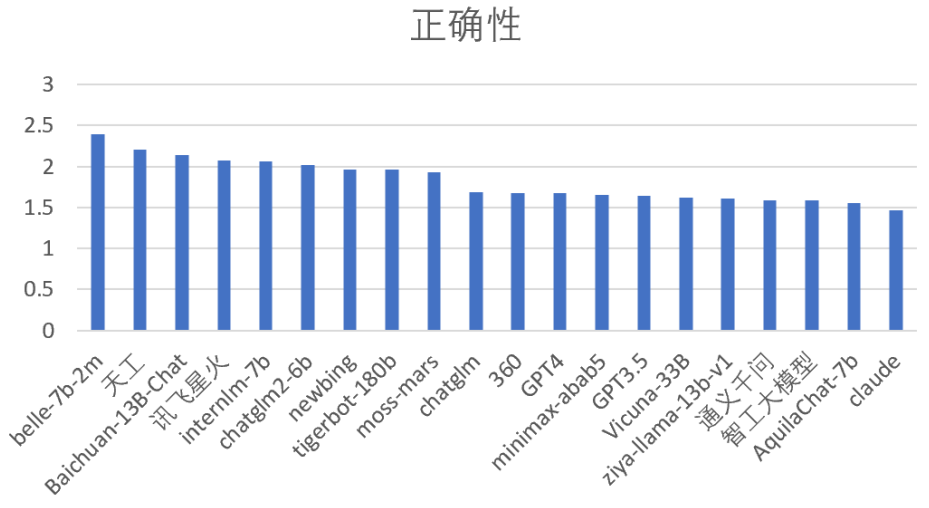
评测结果

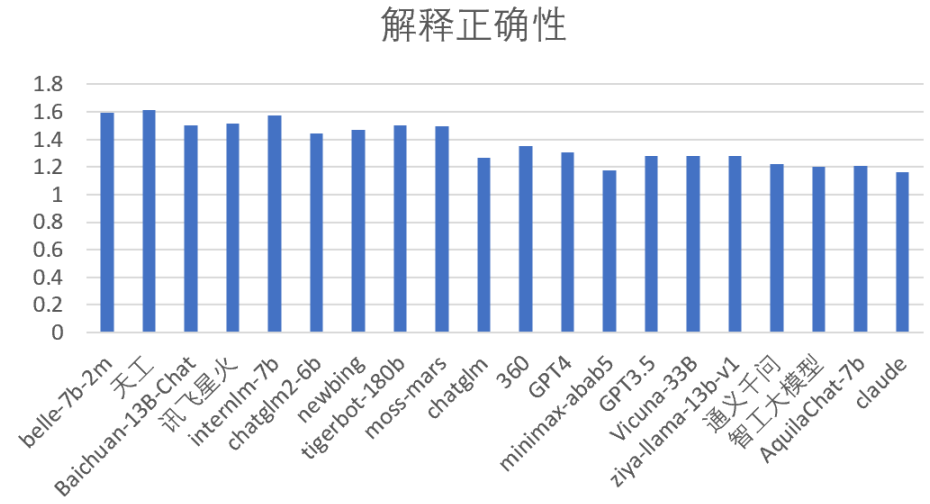
人工评测



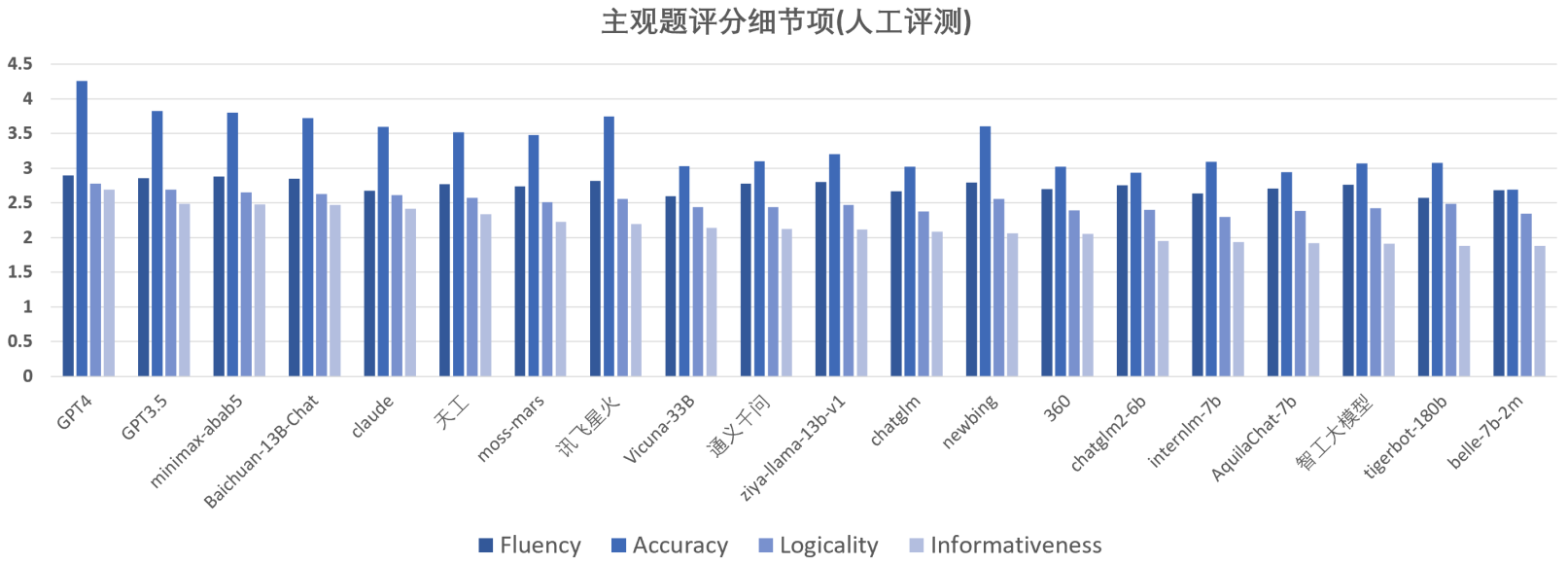


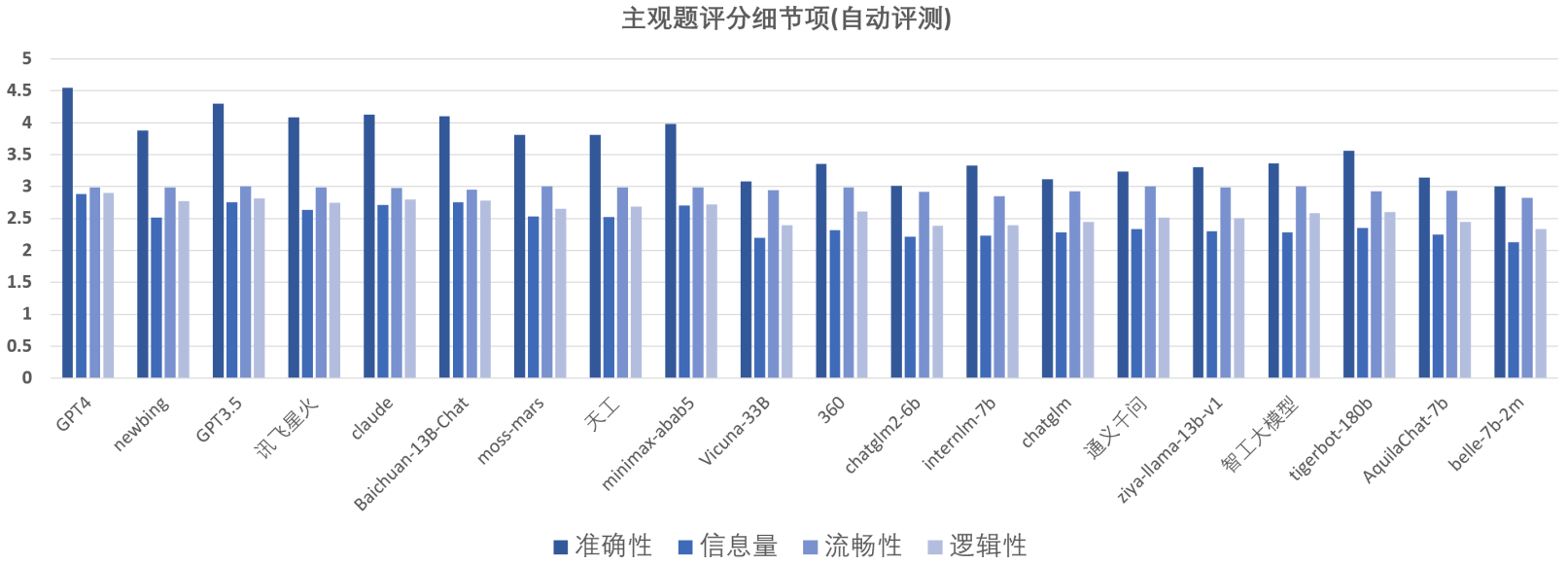
自动评测





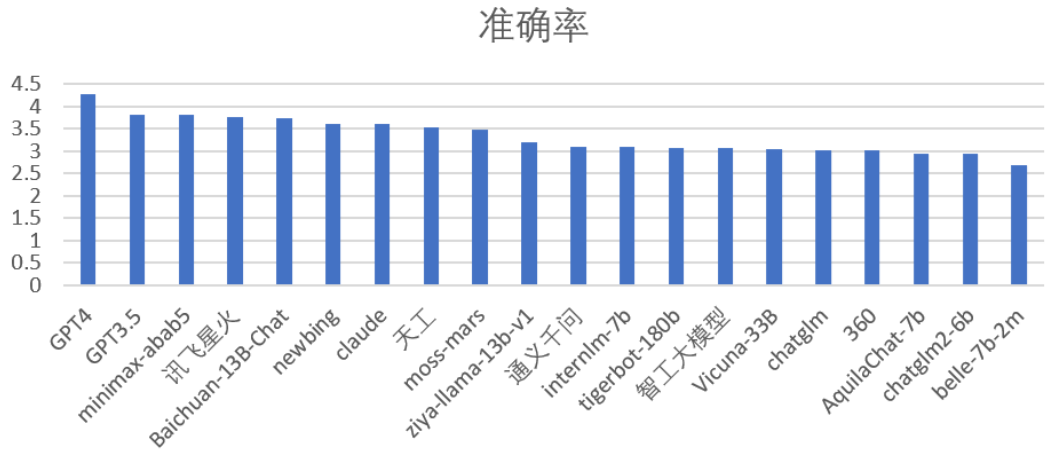
评测结果

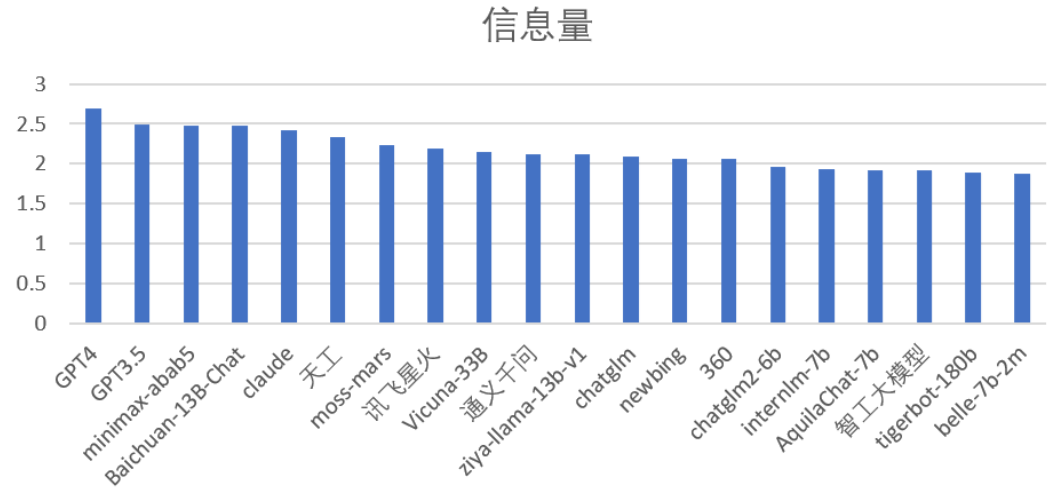

主观题评分细节项(人工评测)
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
GPT4
GPT3.5
minimax-abab5
Baichuan-13B-Chat
claude
天工
moss-mars
讯飞星火
Vicuna-33B
通义千问
ziya-llama-13b-v1
chatglm
newbing
360
chatglm2-6b
internlm-7b
AquilaChat-7b
智工大模型
tigerbot-180b
belle-7b-2m
Fluency
Accuracy
Logicality
Informativeness

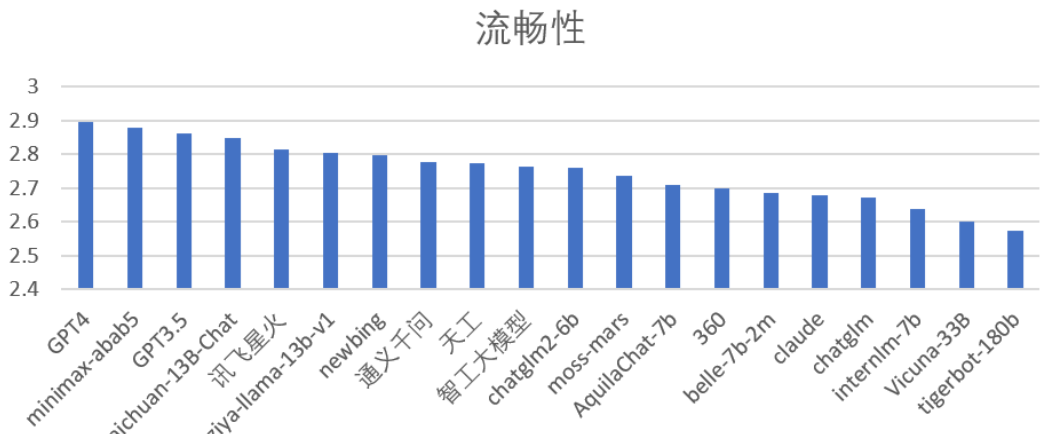

主观题评分细节项(自动评测)
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
GPT4
newbing
GPT3.5
讯飞星火
claude
Baichuan-13B-Chat
moss-mars
天工
minimax-abab5
Vicuna-33B
360
chatglm2-6b
internlm-7b
chatglm
通义千问
ziya-llama-13b-v1
智工大模型
tigerbot-180b
AquilaChat-7b
belle-7b-2m
准确性
信息量
流畅性
逻辑性

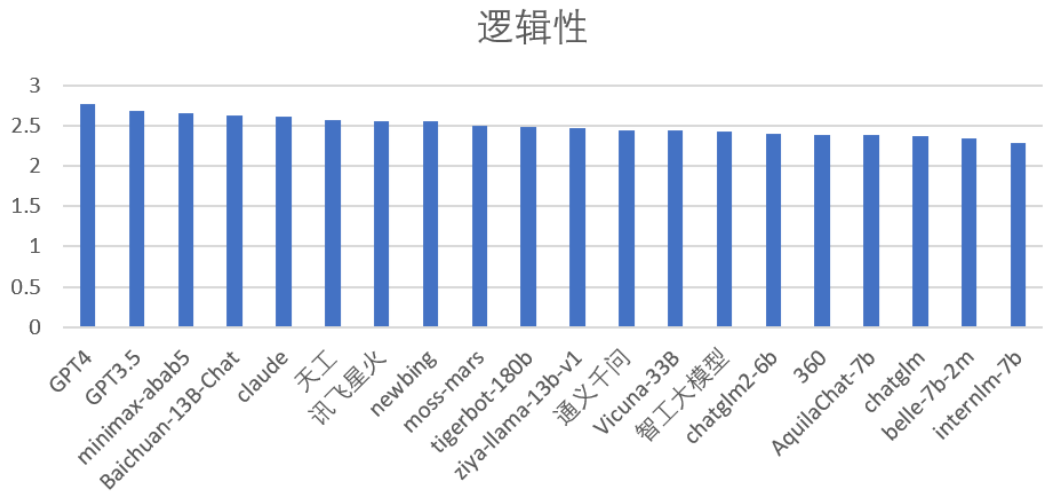
评测结果

主观题人工评分细节





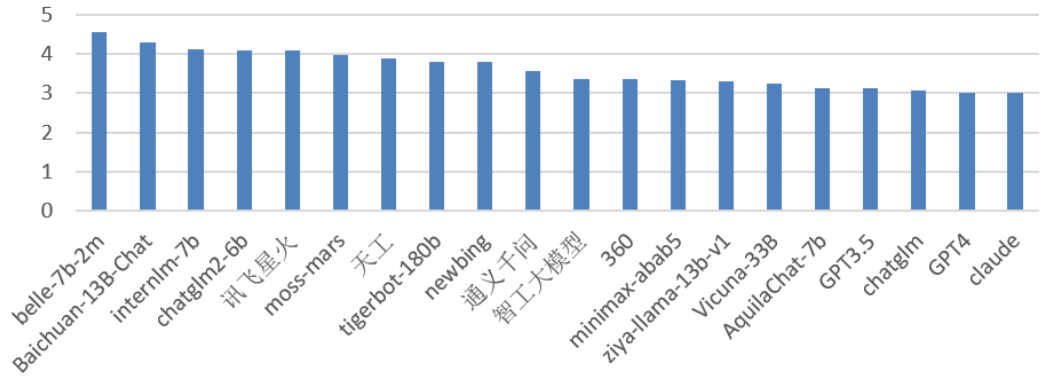




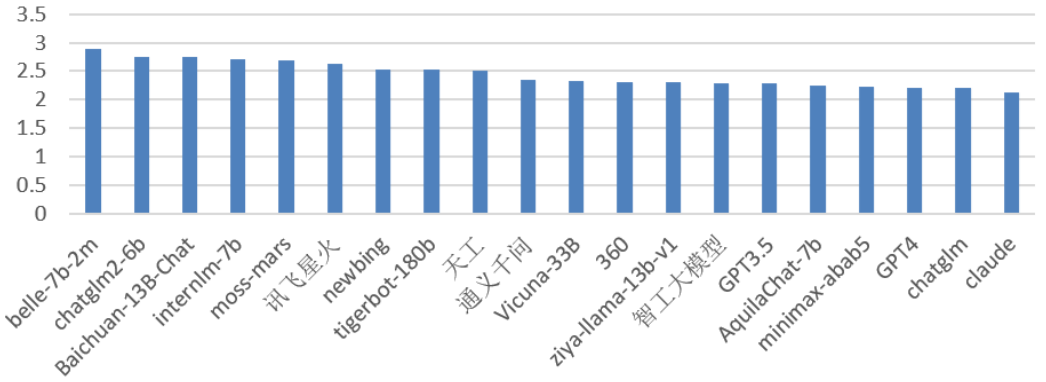
评测结果

主观题自动评分细节

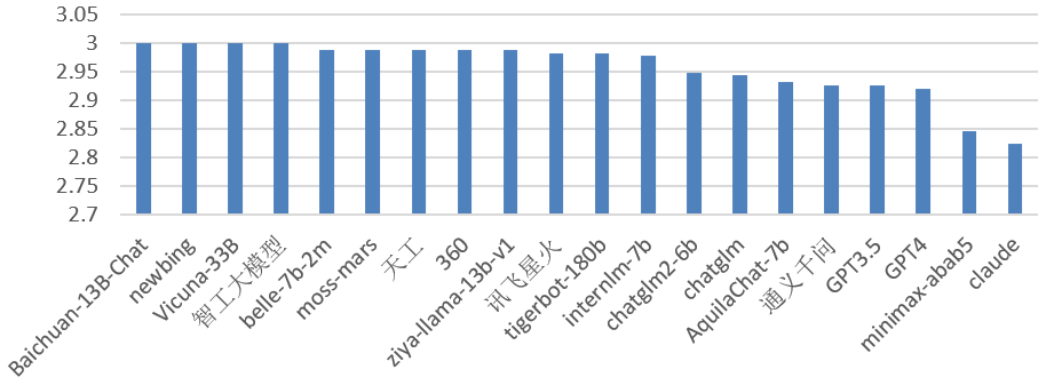
准确度



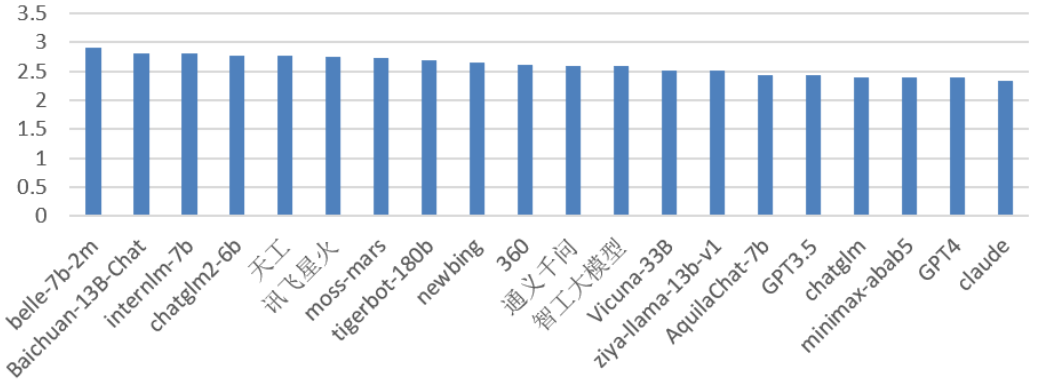
信息量



流畅度



逻辑性



附录 1 模型排名（人工/GPT4）

	客观题		主观题				排名	总分
模型名称	答案准确性	解释准确性	流畅性	准确率	逻辑性	信息量		
GPT4	2.378 (2.395)	1.670 (1.595)	2.895 (2.989)	4.260 (4.545)	2.779 (2.903)	2.691 (2.886)	1(1)	86.72 (89.54)
GPT3.5	2.160 (2.138)	1.542 (1.503)	2.861 (3.000)	3.822 (4.295)	2.694 (2.818)	2.489 (2.750)	2(2)	80.71 (84.69)
讯飞星火	2.114 (2.243)	1.557 (1.632)	2.815 (2.977)	3.750 (4.193)	2.560 (2.739)	2.196 (2.716)	3(5)	78.05 (82.26)
Baichuan-13B-Chat	2.003 (2.013)	1.428 (1.441)	2.847 (2.949)	3.727 (4.102)	2.631 (2.778)	2.472 (2.756)	4(6)	77.51 (81.82)
minimax-abab5	1.922 (1.928)	1.443 (1.493)	2.878 (2.989)	3.800 (3.977)	2.656 (2.722)	2.478 (2.699)	5(7)	77.47 (80.64)
newbing	2.197 (2.211)	1.583 (1.615)	2.796 (2.989)	3.608 (3.875)	2.558 (2.773)	2.061 (2.511)	6(4)	77.28 (82.63)
claude	1.923 (2.066)	1.463 (1.576)	2.680 (2.977)	3.597 (4.125)	2.613 (2.801)	2.414 (2.710)	7(3)	75.57 (83.49)
moss-mars	1.961 (1.967)	1.465 (1.470)	2.737 (3.000)	3.480 (3.807)	2.508 (2.648)	2.229 (2.534)	8(9)	74.41 (79.21)
天工	1.933 (1.961)	1.354 (1.500)	2.774 (2.983)	3.520 (3.807)	2.576 (2.682)	2.339 (2.523)	9(8)	74.36 (79.31)
ziya-llama-13b-v1	1.681 (1.592)	1.306 (1.201)	2.804 (3.000)	3.207 (3.364)	2.473 (2.585)	2.120 (2.278)	10(13)	69.48 (70.92)
通义千问	1.638 (1.618)	1.275 (1.280)	2.776 (3.000)	3.098 (3.239)	2.443 (2.511)	2.126 (2.335)	11(12)	68.01 (71.02)
360	1.720 (1.678)	1.322 (1.352)	2.700 (2.989)	3.022 (3.352)	2.394 (2.608)	2.056 (2.313)	12(10)	67.97 (72.86)
智工大模型	1.680 (2.072)	1.297 (1.516)	2.764 (2.983)	3.067 (4.080)	2.427 (2.744)	1.916 (2.631)	13(14)	67.27 (70.53)
chatglm2-6b	1.690 (1.671)	1.345 (1.306)	2.758 (2.920)	2.934 (3.011)	2.401 (2.386)	1.956 (2.210)	14(17)	67.07 (69.06)
Vicuna-33B	1.567 (1.684)	1.277 (1.270)	2.599 (2.943)	3.033 (3.080)	2.440 (2.398)	2.143 (2.199)	15(16)	66.53 (69.16)
internlm-7b	1.655 (1.658)	1.355 (1.174)	2.636 (2.847)	3.091 (3.330)	2.295 (2.392)	1.938 (2.233)	16(18)	66.52 (69.00)
ChatGLM	1.602 (1.638)	1.239 (1.280)	2.670 (2.926)	3.022 (3.114)	2.374 (2.443)	2.084 (2.278)	17(15)	66.05 (69.48)
Tigerbot-180b	1.604 (1.592)	1.294 (1.220)	2.573 (2.926)	3.079 (3.557)	2.489 (2.602)	1.882 (2.352)	18(11)	65.90 (71.77)
AquilaChat-7b	1.548 (1.553)	1.239 (1.207)	2.710 (2.932)	2.945 (3.136)	2.383 (2.443)	1.918 (2.244)	19(19)	64.82 (68.19)
belle-7b-2m	1.484 (1.461)	1.224 (1.164)	2.685 (2.824)	2.695 (3.000)	2.347 (2.335)	1.880 (2.131)	20(20)	62.98 (65.27)

$$\text{总分} = \frac{1}{n}\sum_{i=1}^{n} score_i$$

总分为所有科目归一化分数的平均值，括号内数值为GPT-4自动评测的打分及排名。

附录 2　主观题打分界面

在线评测界面

- 客观题
- 主观题

附录 3　客观题测评界面

在线评测界面

- 客观题
- 主观题

附录 4　GPT-4自动测评Prompt – 客观题

Score the following LLM output on a question related to a specific discipline with respect to the following aspects, using a 1 to 2 star scale.

Correctness:
1 star means wrong
2 stars means correct

Explanation Correctness:
1 star means Incorrect explanation
2 stars means correct explanation

User: [question]

LLM: [answer from llm]

The correct answer to user's question is: correct answer
Please give me the answers like this:

{"Correctness":numbers of its stars(int),"Explanation Correctness":numbers of its stars(int)}

附录 5　GPT-4自动测评Prompt – 主观题

Score the following LLM output on a question related to a specific discipline with respect to the following aspects, using a 1 to 3 star scale.

Accuracy:
1 star means Completely wrong
2 stars means Partially correct
3 stars means Completely correct

Informativeness:
1 star means Lack of necessary information or off-topic response
2 stars means Insufficient information to provide a complete answer
3 stars means Information is correct and sufficient.

Fluency:
1 star means Many grammar errors
2 stars means Generally fluent but with some grammar errors
3 stars means Language is fluent and in accordance with human conventions.

Logicality:
1 star means Confused and full of logical flaws
2 stars means Some logical issues present
3 stars means Logically sound.

User: [question]

LLM: [answer from llm]

The correct answer to user's question is: correct answerPlease give me the answers like this:

{"Accuracy":numbers of its stars(int),"Informativeness":numbers of its stars(int),"Fluency":numbers of its stars(int),"Logicality":numbers of its stars(int)}

谢谢！

Email: cs_nlp@fudan.edu.cn