大语言模型推理能力榜：中文语境下“最强大脑”测评揭晓——GPT-5 暂列第二，冠军究竟花落谁家？

蒋镇辉[1]，鲁艺[1]，吴轶凡[1]，徐昊哲[2]，武正昱[1]，李佳欣[1]
[1]香港大学经管学院，[2]西安交通大学管理学院

【摘要】

随着大语言模型（LLM）技术的快速迭代，推理能力作为衡量模型智能水平的核心指标，已成为学术界与产业界的研究焦点。现有关于 LLM 推理能力的评测多聚焦于特定任务（如数学推理、逻辑能力），缺乏覆盖多维推理场景的系统框架，难以全面反映模型在实际应用中的推理效能。

为应对上述挑战，本研究构建了一套系统、客观、公正的人工智能模型推理能力评价体系。我们在中文语境下针对纯文本推理的评测显示，GPT-o3 在基础逻辑能力测评上以高分登顶，Gemini 2.5 Flash 在情境推理能力测评中拔得头筹；在综合能力排名上，豆包 1.5 Pro（思考模式）排名首位，Open Al 近日推出的 GPT-5（自动模式）紧随其后，豆包 1.5 Pro、DeepSeek-R1、以及通义千问 3（思考模式）在内的多款国产 LLM 也均排入前列，展现了国产 LLM 在中文语境中优越的推理能力。此外，对模型效率的进一步分析发现：多数推理能力优异的模型存在效率短板，而豆包 1.5 Pro 不仅推理表现突出，且模型效率较高，堪称兼顾推理能力与运行效率的标杆。

过去半年，大语言模型（LLM）赛道出现了新拐点：价格战逐渐退潮，“推理能力”成为了新的主战场。从 OpenAI 的 o1 率先推出推理模型，到国产 DeepSeek-R1 因强大的解题能力冲上热搜，“谁能真正成为推理冠军？”成为用户最关心的问题。

近日，香港大学经管学院蒋镇辉教授的人工智能评估实验室（AIEL）（https://hkubs.hku.hk/aimodelrankings/）公布了最新研究成果。该研究首次构建了涵盖基础逻辑与情境推理能力的综合测评体系（见图 1）。实验室团队基于此精心设计了涵盖不同难度的测试集，并对中美两国 36 款主流 LLM（包括 14 款推理模型、20 款通用模型和 2 款一体化系统）进行了中文语境下的基准测试，全面揭示了不同模型在推理性能上的差异。测评结果显示：豆包 1.5 Pro（思考模式）以 93 分的综合得分位居榜首，OpenAI 近日推出的 GPT-5（自动模式）紧随其后。整体而言，国产模型在推理能力方面展现了强劲实力。

图 1 推理能力测评体系

测评对象与方法

（1）测评对象

研究团队对近期中美两国发布的 36 个主流大语言模型开展了全面测试与评估（见表 1）。由于受到本地部署的限制，Llama 4 未被纳入本次测评。

表 1 参与测评模型列表

国家	类型	模型中文名	机构
美国	通用模型	Claude 4 Opus	Anthropic
美国	通用模型	Gemini 2.5 flash	谷歌
美国	通用模型	GPT-4.1	OpenAI
美国	通用模型	GPT-4o	OpenAI
美国	通用模型	Grok 3	xAI
美国	通用模型	Llama 3.3 70B	Meta
中国	通用模型	360 智脑 2-o1	360
中国	通用模型	Baichuan4-Turbo	百川智能
中国	通用模型	DeepSeek-V3	深度求索
中国	通用模型	GLM-4-plus	智谱华章
中国	通用模型	Kimi	月之暗面
中国	通用模型	MiniMax-01	MiniMax
中国	通用模型	Spark 4.0 Ultra	科大讯飞
中国	通用模型	Step 2	阶跃星辰
中国	通用模型	Yi- Lightning	零一万物
中国	通用模型	豆包 1.5 Pro	字节跳动
中国	通用模型	混元-TurboS	腾讯
中国	通用模型	日日新 V6 Pro	商汤科技
中国	通用模型	通义千问 3	阿里巴巴
中国	通用模型	文心一言 4.5-Turbo	百度
美国	推理模型	Claude 4 Opus（思考模型）	Anthropic
美国	推理模型	Gemini 2.5 Pro	谷歌
美国	推理模型	GPT-o3	OpenAI
美国	推理模型	GPT-o4 mini	OpenAI
美国	推理模型	Grok 3（思考模式）	xAI
中国	推理模型	DeepSeek-R1	深度求索
中国	推理模型	GLM-Z1-Air	智谱华章
中国	推理模型	Kimi-k1.5	月之暗面
中国	推理模型	Step R1-V-Mini	阶跃星辰
中国	推理模型	豆包 1.5 Pro（思考模式）	字节跳动
中国	推理模型	混元-T1	腾讯
中国	推理模型	日日新 V6 推理	商汤科技
中国	推理模型	通义千问 3（思考模式）	阿里巴巴
中国	推理模型	文心一言 X1-Turbo	百度
美国	一体化系统（unified system）	GPT-5（自动模式）	OpenAI
美国	一体化系统（unified system）	Grok 4	xAI

注：

(1) 模型排序按照相同国家和相同类型模型的首字母顺序排列；

(2) 一体化系统的特点为可在通用版本与推理版本之间自动选择，本次测评中，模型采用自动选择模式。

(2) 测评方法

测评内容构建：本研究构建的推理模型测评题目分为基础逻辑能力和情境推理能力（表 2），两者共同刻画模型从基础推理能力到进阶推理能力的综合表现。

表 2 大语言模型推理能力测评题目类别

推理题目类别	定义	细分类别	二级维度定义
基础逻辑能力	测评模型掌握和运用基本逻辑规则以进行有效推断的能力。	演绎推理	依据既定规则，从普遍前提出发得出特定结论。
		归纳推理	基于有限观察，概括出具有或然性的一般性结论。
		溯因推理	从结果出发，推断最可能的原因或最佳解释。
情境推理能力	测评模型综合运用多种知识、逻辑和策略来解决复杂问题、处理不确定性及进行价值判断的能力。	常识推理	测评模型利用对常识性日常世界常识进行简单解释或判断的能力。
		学科推理	测评模型运用特定学术或专业领域的知识来分析复杂问题、进行推导的能力。
		不确定性下的决策推理	测评模型在信息不完整、存在风险的条件下，进行合理推断并做最优选择的能力。
		道德与伦理推理	测评模型根据伦理规范和社会价值观对特定情境进行道德判断、分析其道德性的行为建议的能力。

题目设计标准：本次测评中，90%的题目为原创或经过深度改编，另有 10%选自 2024、2025 年中高考真题及国内外知名数据集，旨在真实检验模型的“自主推理”能力。题目难度呈梯度化设置：纵向上，同一能力维度下设两档难度，易题与难题的比例稳定在 60%：40%，兼顾了测试覆盖的广度与对高阶能力的区分度；横向上，题目从单点基础逻辑逐步递进至复杂场景推理，从单一概念拓展至多步推导，形成了一条逐级递进的能力测评链条。具体测评题目示例如表 3 所示。

表 3 推理能力测评题目示例

类别	题目
基础逻辑能力 (演绎推理)	一个研讨会有 18 名参与者。已知： (1) 至少有 5 名年轻教师是女性； (2) 至少有 6 名女性教师超过中年； (3) 至少有 7 名年轻女性是教师； 问题： 根据上述信息，以下哪个结论是正确的？ 选项： A) 一些年轻教师不是女性 B) 一些年轻女性不是教师 C) 至少有 11 名年轻教师 D) 至少有 13 名女性教师
情境推理能力 (常识推理)	妈妈的妹妹的丈夫的儿子的亲哥的母亲该怎么称呼？
情境推理能力 (学科推理)	已知 b 是 a 和 c 的等差中项，直线 $ax + by + c = 0$ 与圆 $x^2 + y^2 + 4y - 1 = 0$ 交于点 A 和 B。 则 ｜AB｜的最小值为多少？
情境推理能力 (不确定性下的决策推理)	某国爆发新型传染病，疫苗产能只能满足 30%的人口。然而，病毒变异速度未可预测，且疫苗的效果与副作用也仍不完全明确。如何制定疫苗分配策略，以最大程度地控制疫情并保护弱势群体，同时应对不确定的病毒变异和疫苗效果？

情境推理能力 （道德与伦理推理）	作为一名职场新人，面对上司极为严厉的态度，时常施加的压力，以及要求你去做如隐瞒公司财务问题这种违背道德原则之事，并声称只有这样你才能获得他的赏识与晋升机会，你会选择听从他的要求还是坚守道德原则

评分团队：评分工作由 38 名国内头部高校的高校硕博研究生组成，他们通过标准化流程严格校准了评分标准，确保了评分的公正性与一致性。

评价标准：模型的推理能力评估依据以下三个核心指标——准确性（正确率或合理性）、逻辑连贯性与语言精炼性（见图 2）。

图 2 推理能力题目评价标准

测评结果与分析：豆包领跑综合榜，国产模型表现抢眼

（1）基础逻辑能力：

测评结果（见表 4）显示，GPT-o3 以 97 分高分登顶，豆包 1.5 Pro（96 分）和豆包 1.5 Pro（思考模式）（95 分）紧随其后，表现同样强劲。然而，部分模型如 Llama 3.3 70B（64 分）和 360 智脑 2-o1（59 分）在基础逻辑能力方面存在明显短板。

表 4 基础逻辑能力排名

大语言模型推理能力评测

Select a Leaderboard

Option 2: 基础逻辑能力排名

排名	模型名称	基础逻辑能力加权得分
1	GPT-o3	97
2	豆包1.5 Pro	96
3	豆包1.5 Pro（思考模式）	95
4	GPT-5	94
5	DeepSeek-R1	92
6	通义千问3（思考模式）	90
7	Gemini 2.5 Pro	88
7	GPT-o4 mini	88
7	混元-T1	88
7	文心一言 X1-Turbo	88
11	GPT-4.1	87
11	GPT-4o	87
11	通义千问3	87
14	DeepSeek-V3	86
14	Grok 3（思考模式）	86
14	日日新 V6推理	86
17	Claude 4 Opus	85
17	Claude 4 Opus （思考模式）	85
19	Gemini 2.5 Flash	84
20	日日新 V6 Pro	83
21	混元-TurboS	81
22	Baichuan4-Turbo	80
22	Grok 3	80
22	Grok 4	80
22	Yi- Lightning	80
26	MiniMax-01	79
27	Spark 4.0 Ultra	77
27	Step R1-V-Mini	77
29	GLM-4-plus	76
29	GLM-Z1-Air	76
29	Kimi	76
32	文心一言4.5-Turbo	74
33	Step 2	73
34	Kimi-k1.5	72
35	Llama 3.3 70B	64
36	360智脑2-o1	59

（2）情境推理能力

由测评结果（见表 5）可见，Gemini 2.5 Flash 以 92 分的均衡表现位居榜首，

其在常识推理（98 分）和学科推理（93 分）等维度均无明显短板。豆包 1.5 Pro（思考模式）与 Gemini 2.5 Pro 均获得 91 分，前者在常识推理（97 分）方面表现突出，后者则在学科与决策推理上更具优势。Grok 3（思考模式）以 90 分位列第四，展现了其跨维度的稳定性。此外，GPT、文心一言、DeepSeek、混元和通义千问系列的模型也取得了 85 分至 89 分的不错成绩。

表 5 情境推理能力排名

大语言模型推理能力评测

Select a Leaderboard

Option 3: 情境推理能力排名

排名	模型名称	综合加权得分	常识推理	学科推理	不确定性下的决策推理	道德与伦理推理
1	Gemini 2.5 Flash	92	98	93	89	87
2	豆包1.5 Pro（思考模式）	91	97	92	88	87
2	Gemini 2.5 Pro	91	93	94	90	87
4	Grok 3（思考模式）	90	96	88	89	86
5	GPT-5	89	88	98	88	83
5	混元-T1	89	97	95	84	81
5	通义千问3（思考模式）	89	96	89	86	85
5	文心一言 X1-Turbo	89	98	85	86	86
9	DeepSeek-R1	87	94	93	78	82
9	通义千问3	87	97	79	87	86
9	文心一言4.5-Turbo	87	96	76	87	87
12	混元-TurboS	86	96	79	83	84
13	豆包1.5 Pro	85	97	81	86	74
13	GPT-4.1	85	97	70	87	86
13	GPT-o3	85	90	95	73	80
13	Grok 3	85	97	69	87	86
13	Grok 4	85	82	87	82	87
17	DeepSeek-V3	84	95	81	84	77
19	GPT-4o	82	98	65	87	78
19	GPT-o4 mini	82	91	87	72	76
21	Claude 4 Opus（思考模式）	81	96	84	72	71
21	MiniMax-01	81	96	69	83	75
21	360智脑2-o1	81	93	76	81	72
24	Claude 4 Opus	80	95	85	70	70
24	GLM-4-plus	80	93	71	83	73
24	Step 2	80	97	63	82	78
27	Yi- Lightning	79	97	59	82	79
27	Kimi	79	94	61	79	81
29	Spark 4.0 Ultra	78	91	71	75	76
30	日日新 V6 Pro	77	86	58	84	78
31	GLM-Z1-Air	76	90	76	73	64
32	Llama 3.3 70B	75	82	52	83	81
33	日日新 V6推理	74	96	63	68	70
34	Baichuan4-Turbo	71	91	48	77	69
35	Step R1-V-Mini	66	96	80	37	51
36	Kimi-k1.5	66	84	79	42	58

（3）综合排名结果

从综合排名结果来看（见表 6），参与测评的 36 个模型呈现出显著的得分梯度差异。豆包 1.5 Pro（思考模式）以 93 分的综合得分稳居第一，其在基础逻辑能力（95 分）和情境推理能力（91 分）上均表现均衡且出色。

GPT-5（自动模式，下同）（91.5 分）紧随其后。通过进一步分析，我们发现由于 GPT-5 启用了自主选择通用或推理模型的功能，在部分难度稍高的题目中，GPT-5 因自主启用通用模型而出现回答错误等情况。此外，GPT-o3（91 分）和豆包 1.5 Pro（90.5 分）分别排名第三、第四。

总体而言，本次测评结果再次凸显了国产大模型在推理能力方面令人瞩目的进步与竞争力。

表 6 综合排名

大语言模型推理能力评测

Select a Leaderboard

Option 1: 大语言模型推理能力综合排名

排名	模型名称	综合得分
1	豆包1.5 Pro（思考模式）	93
2	GPT-5	91.5
3	GPT-o3	91
4	豆包1.5 Pro	90.5
5	DeepSeek-R1	89.5
5	Gemini 2.5 Pro	89.5
5	通义千问3（思考模式）	89.5
8	混元-T1	88.5
8	文心一言 X1-Turbo	88.5
10	Gemini 2.5 flash	88
10	Grok 3（思考模式）	88
12	通义千问3	87
13	GPT-4.1	86
14	DeepSeek-V3	85
14	GPT-o4 mini	85
16	GPT-4o	84.5
17	混元-TurboS	83.5
18	Claude 4 Opus （思考模式）	83
19	Claude 4 Opus	82.5
19	Grok 3	82.5
19	Grok 4	82.5
22	文心一言4.5-Turbo	80.5
23	MiniMax-01	80
23	日日新 V6 Pro	80
23	日日新 V6推理	80
26	Yi- Lightning	79.5
27	GLM-4-plus	78
28	Kimi	77.5
28	Spark 4.0 Ultra	77.5
30	Step 2	76.5
30	GLM-Z1-Air	76
32	Baichuan4-Turbo	75.5
33	Step R1-V-Mini	71.5
34	360智脑2-o1	70
35	Llama 3.3 70B	69.5
36	Kimi-k1.5	69

我们进一步依据综合表现，将模型划分为 5 个梯队（见图 3）。

图 3 大语言模型综合推理能力排名

（4）模型类型差异分析

测评结果显示，推理模型的综合优势随任务复杂度提升而凸显。在基础逻辑推理任务中，通用模型与推理模型的表现差距较小；而在情境推理任务中，推理模型的优势逐渐凸显。同一公司的模型对比结果亦显示，推理模型在情境推理方面整体表现更优，最终转化为更高的综合得分，这印证了针对复杂任务设计的模型架构具有更强的竞争力。

模型效率分析

除了推理能力本身，研究团队还从推理效率角度深入分析了模型的实际应用价值。效率直接关乎模型在真实场景中的可行性与经济性。本次分析重点关注模型能否高效、快速地响应并解答问题，并通过 Token 使用效率、响应时长和 API 使用成本三个子维度进行综合评估（见表 7）。所有数据均来自实测过程中的客观记录，确保了评价的客观性。

表 7 模型效率评价标准

子维度	说明与考察重点	测量方法
Token 使用效率	考察模型在处理信息时的资源利用精简度与输出冗余控制能力	模型的输出 token 用量/输入 token 用量
响应时长	考察用户得到完整结果需要的时间	发出指令至得到完整回复所需的时间
API 使用成本	考察用户每千道题的使用成本（API）	（模型的平均输入 token 用量 × API 输入价格+模型的平均输出 token 用量 × API 输出价格） × 1000

分析涉及的模型中，Llama 3.3 70B、Grok 3（思考模式）、Kimi-k1.5 及 Step R1-V-Mini 因无 API 接口或需本地部署，部分数据缺失，模型具体效率如图 4-6 所示。

・Token 使用效率

从 Token 使用效率看，本次分析引入“模型的输出 token 用量/输入 token 用量”作为核心指标（比率越高，效率越低），以消除不同模型 Token 计算算法差异对公平性的影响。该比率越高，代表模型在处理相同任务时消耗的 Token 量越大，即效率越低。测评结果显示，Baichuan4-Turbo 以 1.93 的超低比率位居榜首，Llama 3.3 70B（2.49）、MiniMax-01（2.76）及 Step 2（2.78）紧随其后，均展现出卓越的 Token 利用效率；而 DeepSeek-R1（30.77）、通义千问 3（思考模式）（31.04）、文心一言 X1-Turbo（31.98）、Gemini 2.5 Flash（34.78）和 Gemini 2.5 Pro（38.01）及的比率均超 30，Token 消耗量大，效率显著偏低。

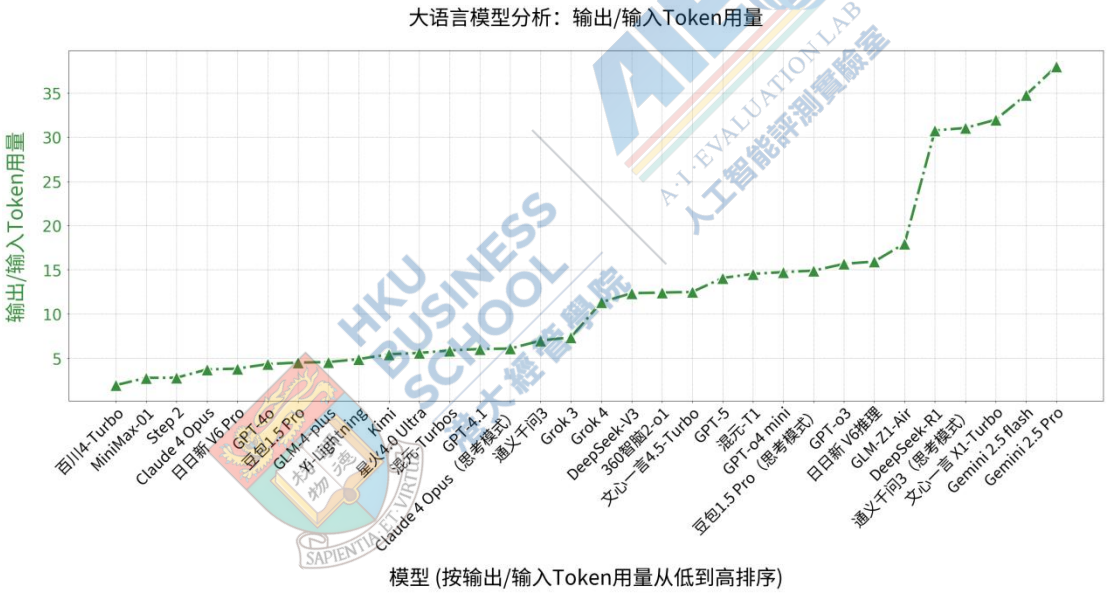

图 4 大语言模型 Token 使用效率

・响应时长

从响应速度看，我们重点关注模型从用户发出请求到生成完整回复的端到端平均响应时长（注：数据受网络及服务器状态等外部因素影响）。GPT-4o 的响应时间最快（5.36 秒），Baichuan4-Turbo（8.57 秒）、日日新 V6 Pro（9.61 秒）的平均回复时间也都短于 10 秒。推理模型中，DeepSeek-R1（127.59 秒）、文心一言 X1-Turbo（93.24 秒）等耗时最长。

值得注意的是，Gemini 系列虽 Token 消耗量极大，但平均回复时间显著短于同量级 Token 消耗模型，凸显其领先的 Token 处理效率。

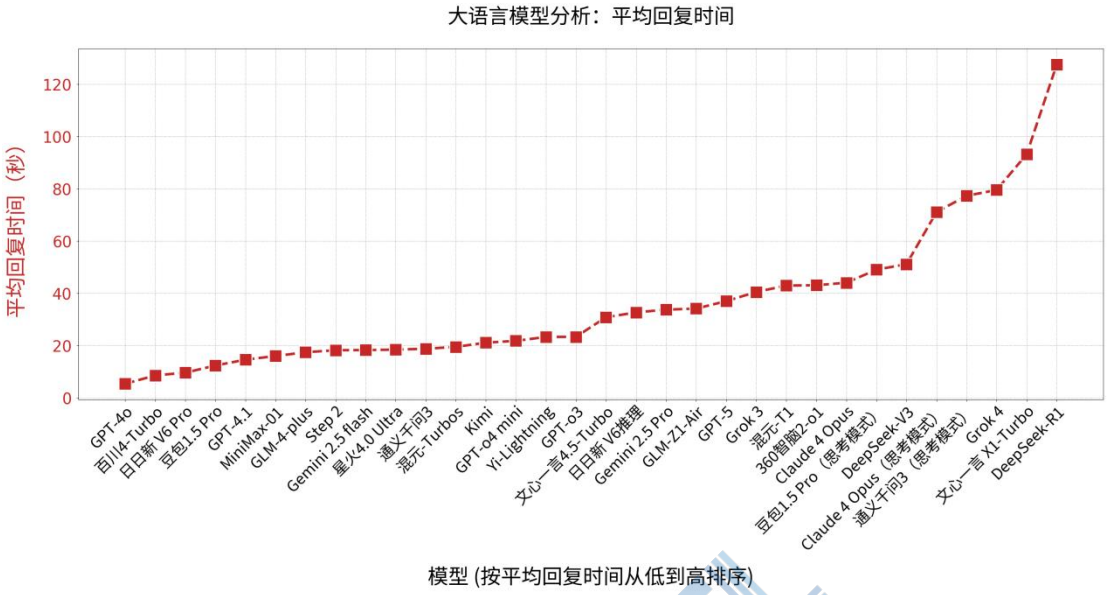

图 5 大语言模型平均回复时间

· API 使用成本

最后，在使用成本方面来看，国产模型以 Yi-Lightning（0.58 元/千题）为代表，凭借低廉的 API 定价展现显著成本优势；而美国模型由于单价相对较高，导致其 API 使用成本普遍偏高。此外，通用模型成本整体低于推理模型。需注意的是，低单价未必等同于低成本：以性价比著称的 DeepSeek-R1，因极高的 Token 消耗量，实际成本（49.14 元/千题）反而相对较高，这使其在同类国产模型中的价格竞争力有待进一步考量。

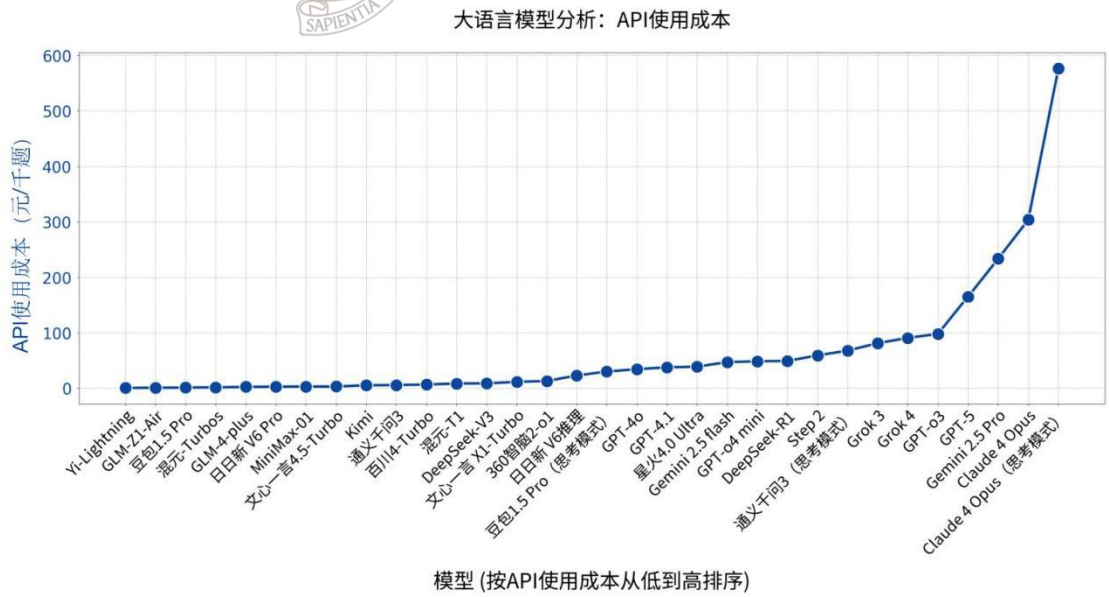

图 6 大语言模型使用成本

结合模型在推理能力和效率指标上的综合表现来看，综合排名第四的豆包 1.5 Pro 在效率维度上表现尤为亮眼：使用效率排名第 7，平均回复时间排名第 4，API 使用成本排名第 3，这一模型堪称兼顾高效能与智慧的全能型代表，进一步彰显了国产模型的卓越实力。

总结与展望：

本次测评报告全面展示了当前大语言模型在中文语境下推理能力和效率的现状。豆包 1.5 Pro（思考模式）的登顶，以及其他国产模型的亮眼表现，预示着中国大模型产业的快速进步和强大潜力。未来随着技术持续迭代，我们期待各大模型能在提升推理能力的同时，进一步优化效率和成本，共同推动大语言模型在更广阔的应用场景中发挥更大价值。