中文语境下

高阶人工智能模型推理能力评测报告

蒋镇辉*[1], 鲁艺[1], 吴轶凡[1], 徐浩哲[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

2025 年人工智能技术爆发式发展，大语言模型向“会思考”演进，但高阶推理能力短板凸显。香港大学经管学院人工智能评测实验室针对截至 2025 年 8 月的中美 30 余款大语言模型，开展多模态与奥赛推理能力测评。测评结果显示：多模态推理中，GPT 系列强势领跑，豆包 1.5 Pro（思考模式）跻身全球前列；奥赛推理中，GPT-5（思考模式）和 Gemini 2.5 Pro 表现突出，处于领跑位置。相比之下，国产模型仍存短板。综合来看，美国模型在高阶推理领域整体领先，国产模型在多模态推理中取得突破，但复杂推理能力仍需提升。

蒋镇辉*，鲁艺，吴轶凡，徐昊哲，武正昱，& 李佳欣．中文语境下高阶人工智能模型推理能力评测报告[R].香港大学经管学院人工智能评测实验室．2025.

通讯作者：蒋镇辉；邮箱：jiangz@hku.hk

自 2025 年开年以来，人工智能技术呈现爆发式发展态势，大语言模型从“能对话”向“会思考”快速演进。但在更复杂的现实场景中，例如结合课本插图讲解物理公式、从商业分析报告图文中提炼趋势，或是应对需要多步推导的数学奥赛题时，AI 的“高阶推理能力”短板逐渐凸显：部分模型虽能处理单一文本信息，却难以整合图像与文字的跨模态逻辑；有些模型可解答常规题目，却在高难度、创新性问题面前经常“卡壳”。

这种能力上的差距，直接限制了 AI 在教育、科研、商业决策等领域的应用拓展。如何科学衡量 AI 的“真实智能水平”？多模态推理能力（跨信息形式的逻辑推导能力）与奥赛推理能力（复杂问题中的高阶思维能力）成为两大核心评判标准。

在此背景下，香港大学经管学院蒋镇辉教授领衔的人工智能评估实验室（AIEL）开展了一项系统性研究。团队针对中美两国截至 2025 年 9 月中旬发布的共 37 款大语言模型（含 14 个推理模型、20 个通用模型和 3 个一体化系统）的多模态推理和奥赛推理能力展开全面测评，旨在为行业揭示高阶人工智能技术的发展现状、指明未来方向。

测评结果显示：多模态推理方面，GPT 系列持续领先（GPT-5（思考模式）第一），豆包 1.5 Pro（思考模式）等顶尖国产模型已跻身全球前列；奥赛推理方面，美国模型在高难度任务中呈整体领先态势，GPT-5（思考模式）优势断层明显，Gemini 2.5 Pro 紧随其后。总体而言，我们发现模型的“思考模式”能有效激活模型深度推理能力；国产模型在复杂推理领域仍存短板。

综上，本研究以系统性、定量化的方式，针对当前人工智能技术的核心发展瓶颈之一——高阶推理能力，提供了全面而深入的评估。通过对中美大模型在多模态推理和奥赛推理两大关键任务方面表现的深度测评，本研究为未来 AI 模型的研发方向提供了关键启示，有助于行业更精准地定位技术瓶颈，加速通用人工智能在教育、科研等高要求领域的落地应用，最终推动 AI 从“对话助手”真正迈向“智能伙伴”。

测评核心：为何聚焦多模态与奥赛推理能力？

在人工智能领域，高阶推理能力是衡量模型“深度智能”的关键试金石。它包含两个核心部分：多模态推理能力和奥赛推理能力。

- 多模态推理能力：指模型整合文本、图像、数据图表等多种信息形式，进行跨模态关联分析与逻辑推导的能力。在教育场景中，它能帮助学生结合教材文本与图示理解力学原理；在商业分析中，可助力从业者匹配市场文字描述与销售图表，预判消费趋势。这种能力是 AI 应对复杂现实任务的“必备素

养”。

- 奥赛推理能力：通过国际数学奥林匹克（IMO）等权威赛事的高难度题目，检验模型在复杂逻辑结构、多步推导与创造性思维上的表现。这类题目往往没有唯一的解法，更考验 AI 跳出常规、寻找最优解的能力，是衡量其“深度智能”的试金石。

综合来看，多模态推理提供了处理现实复杂信息的广度，而奥赛推理则考验了解决深度逻辑问题的能力，二者共同构筑了评估人工智能高阶推理水平的完整框架，并不断拓展着 AI 的深度智能边界。

测评模型

研究团队对截至 2025 年 9 月中旬中美两国发布的 37 个主流人工智能模型开展了全面测试与评估（见表 1）。在多模态测评环节，由于其中 5 个模型不支持识别和处理图文信息的多模态输入，因此在本次测评中被排除在外。

表 1 参与测评模型列表

模型中文名	类型	国家	机构	是否支持多模态
360 智脑 2-o1	中国	通用模型	360	×
Baichuan4-Turbo（百川 4-Turbo）	中国	通用模型	百川智能	√
Claude 4 Opus	美国	通用模型	Anthropic	√
Claude 4 Opus（思考模式）	美国	推理模型	Anthropic	√
DeepSeek-R1（深度求索-R1）	中国	推理模型	深度求索	×
DeepSeek-V3（深度求索-V3）	中国	通用模型	深度求索	×
Gemini 2.5 flash	美国	通用模型	谷歌	√
Gemini 2.5 Pro	美国	推理模型	谷歌	√
GLM-4-plus（智谱-4-Plus）	中国	通用模型	智谱华章	√
GLM-Z1-Air（智谱-Z1-Air）	中国	推理模型	智谱华章	√
GPT-4.1	美国	通用模型	OpenAI	√
GPT-4o	美国	通用模型	OpenAI	√
GPT-5（思考模式）	美国	推理模型	OpenAI	√
GPT-5（自动模式）	美国	一体化系统	OpenAI	√
GPT-o3	美国	推理模型	OpenAI	√
GPT-o4 mini	美国	推理模型	OpenAI	√
Grok 3	美国	通用模型	xAI	√
Grok 3（思考模式）	美国	推理模型	xAI	√
Grok 4	美国	一体化系统	xAI	√
Kimi	中国	通用模型	月之暗面	√
Kimi-k1.5	中国	推理模型	月之暗面	√

Llama 3.3 70B	美国	通用模型	Meta	√
MiniMax-01	中国	通用模型	MiniMax	√
Spark 4.0 Ultra（讯飞星火 4.0 Ultra）	中国	通用模型	科大讯飞	√
Step 2（阶跃 2）	中国	通用模型	阶跃星辰	√
Step R1-V-Mini（阶跃 R1-V-Mini）	中国	推理模型	阶跃星辰	√
Yi-Lightning（零一-Lightning）	中国	通用模型	零一万物	×
豆包 1.5 Pro	中国	通用模型	字节跳动	√
豆包 1.5 Pro（思考模式）	中国	推理模型	字节跳动	√
混元-T1	中国	推理模型	腾讯	√
混元-TurboS	中国	通用模型	腾讯	√
日日新 V6 Pro	中国	通用模型	商汤科技	√
日日新 V6 推理	中国	推理模型	商汤科技	√
通义千问 3	中国	通用模型	阿里巴巴	√
通义千问 3（思考模式）	中国	推理模型	阿里巴巴	√
文心一言 X1-Turbo	中国	推理模型	百度	×
文心一言 4.5-Turbo	中国	通用模型	百度	√

注：模型排序按照模型中文名的首字母顺序排列。

测评内容与结果

（一）多模态推理能力测评

1.测评内容

多模态测试的所有题目均设计为“图文联合任务”，仅凭文字或图片无法得出正确答案，有效规避了测试设计中因过度依赖单一模态（仅文字或仅图片）而产生的结果偏差问题（示例见表 2）：

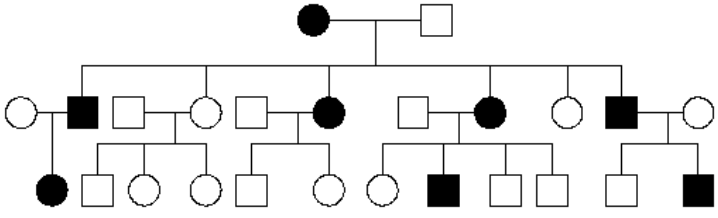
基础逻辑推理类：基础逻辑类题目涵盖演绎、归纳与溯因三种基本推理类型。我们参考了认知心理学与形式逻辑中的经典题型框架，并结合图文场景进行改编。

常识推理类：常识类题目则聚焦日常生活，并结合图像内容进行全新创作。

专业学科推理类：专业学科推理均为单选题或多选题形式，考察模型在不同学科领域的知识储备与应用能力。试题均来自于最新的各省市中高考真题和公开权威多学科视觉问答数据集 MMMU2。

社会问题推理类：在人性与社会现象推理题目部分，我们自主设计了多组具有现实语境的图文推理题，内容涵盖环境保护、公共行为、社会责任、道德判断、与伦理冲突等主题。相比传统知识型问答，此类题目更强调情境理解、价值冲突识别和模态融合后的判断能力，对大模型的多模态泛化推理提出了更高要求。

表 2 多模态推理例题

类别	题目
基础逻辑推理	过山车项目要求身高一米五以上才可以玩，那么图中的人可以玩吗？ A. 可以 B. 不可以 
常识推理	图片中有几只真正的猫咪？ 
专业学科推理	在以下家系图中，最可能的遗传方式是什么？ A. 常染色体隐性遗传 (AR) B.常染色体显性遗传 (AD) C. X 染色体隐性遗传 (XR) D.X 染色体显性遗传 (XD) 

社会问题推理	简述图中漫画的寓意。 

2. 测评标准

多模态推理能力聚焦于模型对文本与图像信息的综合处理及推理能力，其测评标准以准确性为核心维度进行量化评分。对于图文基础逻辑验证、图像常识推理等具有明确标准答案的客观题，采用“正确”或“错误”的二元变量来评估结果的准确性；而针对图文信息融合后的开放性推理任务（如图文情境下的决策分析等），则通过 7 点式李克特量表对回复的合理性进行评价。

3. 测评过程

在打分专家配置方面，多模态推理能力相关题目由国内知名高校的 29 名硕士或博士研究生负责评分。

在打分实施过程中，为确保评分标准的一致性，每位打分者在接触新题型时，需先完成 3 道随机分配的“热身题”（不计入最终结果），通过该环节熟悉题目特征与评分规则；随后，对该题目的全部模型回复以随机顺序进行正式评分，以此消除顺序效应可能产生的偏差。即对于每道题目，每位打分者需完成 n（模型回复数）+3（热身题）个回复的评分工作。

4.测评结果

大语言模型模型的多模态推理能力得分如下表 3 所示。

表 3 多模态推理能力排名

排名	模型名称	准确率
1	GPT-5（思考模式）	91
2	GPT-4.1	90
3	GPT-o3	87
4	豆包 1.5 Pro（思考模式）	85
4	GPT-5（自动模式）	85
6	GPT-4o	84
7	Claude 4 Opus（思考模式）	83
8	豆包 1.5 Pro	82
8	Grok 3（思考模式）	82
10	通义千问 3	81
11	Kimi-k1.5	80
11	日日新 V6 推理	80
11	Step R1-V-Mini（阶跃 R1-V-Mini）	80
14	Grok 4	79
14	GPT-o4 mini	79
14	混元-T1	79
17	GLM-4-plus（智谱-4-Plus）	78
17	通义千问 3（思考模式）	78
19	Gemini 2.5 Flash	77
19	GLM-Z1-Air（智谱-Z1-Air）	77
21	Llama 3.3 70B	76
22	日日新 V6 Pro	75
22	Gemini 2.5 Pro	75
23	文心一言 4.5-Turbo	74
24	Step 2（阶跃 2）	73
26	混元-TurboS	71
26	Claude 4 Opus	71
28	Spark 4.0 Ultra（讯飞星火 4.0 Ultra）	68
28	MiniMax-01	68
30	Baichuan4-Turbo（百川 4-Turbo）	67
31	Grok 3	66
32	Kimi	63

（1）整体格局：顶尖模型竞争激烈，梯度分化明显

基于模型在多模态推理能力的表现，我们将模型分为四个梯队（如图 1 所示）。

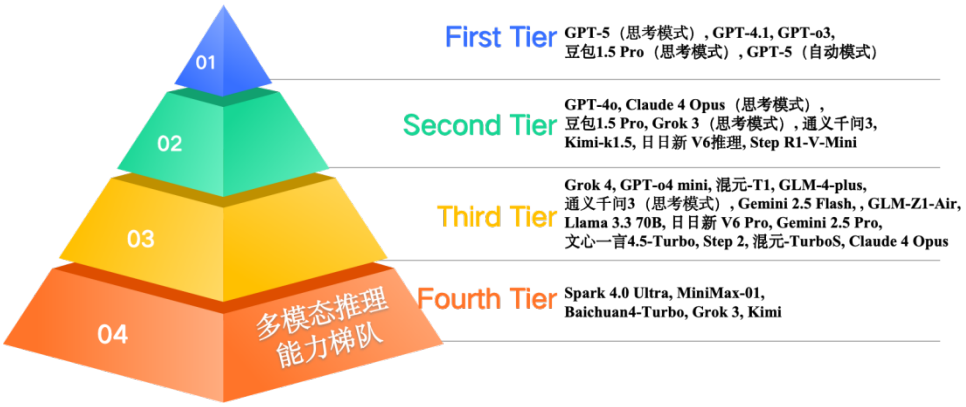


图 1 多模态推理能力梯队

从得分分布看，模型性能呈现清晰的梯度分层，反映出多模态推理能力的显著差异：

头部能力梯队（85 分及以上）：该梯队包含 5 个模型，分数集中在 85-90 分，构成多模态推理能力的绝对第一阵营。GPT-5（思考模式）以 91 分位居榜首，GPT-4.1（90 分）、GPT-o3（87 分）紧随其后；豆包 1.5 Pro（思考模式）与 GPT-5（自动模式）同以 85 分并列第三。GPT 系列占据前五席中的三席，豆包是唯一进入第一梯队的国产模型，展现了其强大的跨模态融合能力。

中等表现梯队（第二、第三梯队）竞争较为密集：与第一梯队相比，二三梯队表现相对不足，包括 GPT-4o（84 分）、Claude 4 Opus（思考模式）（83 分）、混元-TurboS（71 分）、Claude 4 Opus（71 分）等。

尾部能力梯队：70 分以下的 5 个模型构成能力最薄弱的层级，分数集中在 63-68 分，包括 Spark 4.0 Ultra（讯飞星火 4.0 Ultra）、MiniMax-01、Baichuan4-Turbo（百川 4-Turbo）、Grok 3 和 Kimi。这类模型仅能跨模态复杂推理能力较差，与前三梯队形成明显的技术代差，凸显多模态推理能力的明显短板。

（2）核心发现一：“思考模式”是重要的性能放大器

从模型类型看，和通用模型相比，同公司的推理模型在复杂任务中展现出一定的优势。部分模型在“思考模式”下性能显著跃升：Grok 3（思考模式 82 分 vs 通用模式 66 分，+16 分）、Claude 4 Opus（思考模式 83 分 vs 通用模式 71 分，+12 分）、日日新 V6（思考模式 80 分 vs 通用模式 75 分）。

（3）核心发现二：GPT 系列霸榜、顶尖国产模型跻身全球前列

GPT 系列在 85 分以上区间的五个头部模型中占据四席（GPT-5（思考模式）、GPT-4.1、GPT-o3、GPT-5（自动模式）），体现出 GPT 系列模型在多模态数据融合上的底层优势，这可能与训练数据规模、高质量图文对预训练策略密切相关。

国产模型中，豆包 1.5 Pro（思考模式）（85 分）是唯一进入前五的国产模型，其通用与思考模式差距极小，说明其多模态推理能力已达到国际顶尖水平。不过整体而言，在多模态技术前沿领域，美国模型仍处领先地位。

（二）奥赛推理能力测评

1. 测评内容

奥赛推理试题来源包括近年全国奥林匹克竞赛以及国际数学奥林匹克（IMO）等权威赛事（示例见表 4）。该类题目难度远高于常规中高考题目，通常涉及更复杂的逻辑结构、多步推理与创造性思维，能够进一步评估模型在处理高阶数学知识和问题解决方面的能力。

表 4 奥赛推理例题

类别	题目
奥赛推理	试求所有实数 α，使得对于所有正整数 n，整数 $$\lfloor\alpha\rfloor+\lfloor 2\alpha\rfloor+\cdots+\lfloor n\alpha\rfloor$$ 均为 n 的倍数。(此处 $\lfloor z\rfloor$ 表示小于或等于 z 的最大整数。例如，$\lfloor-\pi\rfloor=-4$，$\lfloor 2\rfloor=\lfloor 2.9\rfloor=2$。)

2. 测评标准

奥赛类推理能力测评标准，鉴于其解题难度大、解题方法多样等特点，研究团队专门为模型在奥赛题上的回复设计了一套针对性评价标准。

- **解题正确性。**由于目前大多模型难以完全准确解答题库中的奥赛题，研究团队转而对模型回复中的解题过程进行精细化评估，以有效区分不同模型的实际表现差异。该维度采用 0 至 10 分制评分，且允许以 0.5 分为单位进行打分，通过更细致的分数梯度实现精准区分。

- **逻辑连贯性。**该维度重点考察模型能否按照清晰的步骤、合理的推理结构展开答题，重点关注解题思路的连贯性、推理链条的完整性，以及是否存在逻辑漏洞、跳步等问题，该维度同样采用 7 点式李克特量表进行评分。

- **方法创新性。**该维度聚焦模型的解题策略突破能力，主要考察模型是否能够跳出常规解题套路，采用更高效、简洁的方法解题，而非生搬硬套既有模式、使用复杂笨拙的方式推导，该维度亦采用 7 点式李克特量表进行评分。

3. 测评过程

奥赛类题目因涉及高阶专业知识，由中国大陆及香港奥赛集训队的 3 名专业人员（包括国际奥数比赛银牌获得者）承担评分任务。

在打分实施过程中，和多模态推理能力测评过程相似，每位打分者在接触新题型时，同样需先完成 3 道随机分配的“热身题”（不计入最终结果）随后，对该题目的全部模型回复以随机顺序进行正式评分。

4. 测评结果

模型的奥赛推理能力排名如表 5 所示。

（1）综合表现：美国大模型在多维度领先

奥赛推理综合得分：奥赛推理作为本次测评中技术难度最高的模块，充分展现了推理模型与通用模型在应对极端复杂问题时存在的性能鸿沟。加权得分前 3 的模型均为美国大模型，在正确性、逻辑连贯性、方法创新性、奥赛推理能力上呈现“多维度领先”的特点。GPT-5（思考模式）（48 分）和 Gemini 2.5 Pro（44）断层领先，GPT-o3（38）、Claude 4 Opus（思考模式）（33）次之；国产模型仅通义千问 3（思考模式）（28）、Step R1_V_mini（28）表现尚可，复杂推理仍是国产模型的短板。

值得注意的是，即便是此前备受关注的国产模型（如 DeepSeek-R1（深度求索-R1）），以及在先前任务中表现突出的豆包系列，在这部分测试中也表现平平。

具体分析如下：

正确性：Gemini 2.5 Pro（48 分）和 GPT-5（思考模式）（48 分）一骑绝尘，其次是 GPT-o3（36 分）、Gemini 2.5 Flash（35 分）；国产模型中，通义千问 3（思考模式 29）、GLM_Z1_Air（27）和日日新 V6 推理（27）表现较好，但与国际头部差距明显。

逻辑连贯性：GPT-5（思考模式）（47 分）位居榜首，GPT-o3（42 分）和 Gemini 2.5 Pro（39 分）也表现强劲。部分国产模型（如 Step R1_V_mini、GLM_Z1_Air、Qwen 3）在逻辑连贯性方面展现出与美国模型相当的实力。

方法创新性：GPT-5（思考模式）（44 分）、GPT-o3（39 分）、Claude 4 Opus（思考模式，39 分）位居前列。中国模型中，Qwen 3（思考模式，28 分）展现出一定创新性，但仍落后于国际领先水平。

表 5 奥赛推理能力排名

排名	模型名称	正确性	逻辑连贯性	方法创新性	奥赛推理能力 加权得分
1	GPT-5（思考模式）	48	47	44	48
2	Gemini 2.5 Pro	48	39	36	44
3	GPT-o3	36	42	39	38
4	Claude 4 Opus（思考模式）	30	36	39	33
5	Gemini 2.5 Flash	35	28	31	32
5	GPT-o4 mini	32	33	33	32
7	通义千问 3（思考模式）	29	25	28	28
7	Step R1_V_mini	26	33	22	28
9	GLM_Z1_Air	27	31	22	27
9	日日新 V6 推理	27	28	22	27
11	通义千问 3	25	31	17	26
12	文心一言 4.5-Turbo	25	25	19	24
13	Grok 3（思考模式）	21	28	25	23
14	GPT-5（自动模式）	22	22	28	22
14	DeepSeek-V3（深度求索-V3）	26	14	22	22
16	Claude 4 Opus	22	17	31	21
17	豆包 1.5 Pro（思考模式）	22	17	22	20
17	DeepSeek-R1（深度求索-R1）	17	25	22	20
19	Grok 3	20	19	17	19
19	Grok 4	19	17	25	19
21	文心一言 X1-Turbo	17	19	14	17
21	混元-T1	17	17	19	17
21	混元-TurboS	17	17	19	17
21	Kimi-k1.5	17	19	11	17
25	豆包 1.5 Pro	16	17	19	16
26	GLM-4-plus（智谱-4-Plus）	12	17	8	13
27	GPT-4o	13	8	19	12
27	Spark 4.0 Ultra（讯飞星火 4.0 Ultra）	13	11	14	12
29	Baichuan4-Turbo（百川 4-Turbo）	8	19	11	11
29	GPT-4.1	11	8	17	11
31	Kimi	6	14	17	9
31	Llama 3.3 70B	7	14	6	9
33	Yi-Lightning（零一-Lightning）	6	11	14	8
33	日日新 V6 Pro	8	8	6	8
35	MiniMax-01	5	11	8	7
35	Step2	6	8	8	7
35	360 智脑 2-o1	7	6	8	7

Note: 所有分数均为四舍五入得分结果。

（2）核心发现：“思考模式”依旧表现亮眼

对比同公司的通用与推理模型版本，我们发现思考模式下的模型在奥赛题目的表现上的各维度得分普遍更高：比如，Claude 4 Opus 为例，通用版本加权得分 21，推理版本跃升至 33；Grok 3 通用版本得分 19，其推理版本提升至 23；豆包 1.5 Pro 通用版本得分 16，推理版本升至 20。这说明“思考模式”是激活模型深度推理、创新与高阶逻辑能力的有效策略，可作为优化性能的重要方向。

我们依据模型在奥赛推理任务中的表现，将它们划分为了 4 个梯队，如图 2 所示：

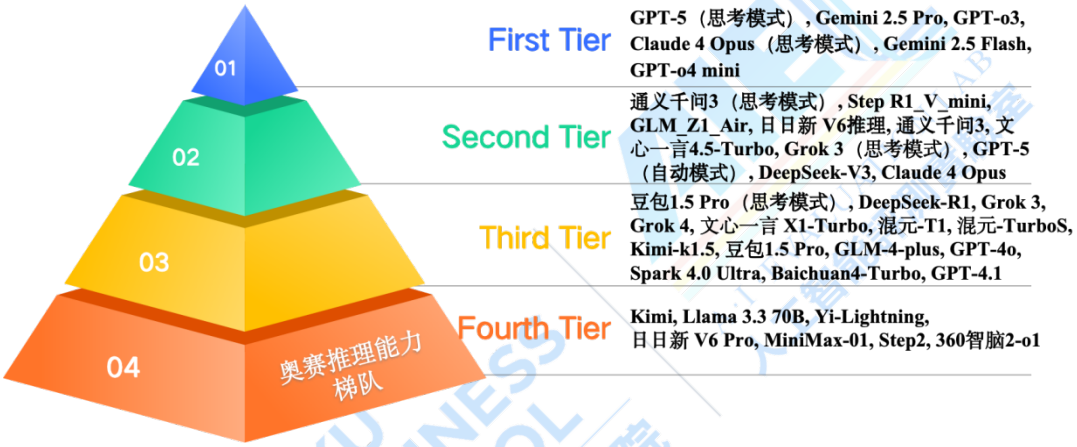


图 2 奥赛推理能力梯队

（三）高阶推理能力综合分析

根据测评结果可以发现，在多模态和奥林匹克竞赛级推理方面表现最突出的模型莫过于 GPT-5（思考模式），该模型在两项排名中均位列第一，堪称综合能力强劲的顶尖选手；同公司的 GPT-o3 进入两项排名前五，凸显其扎实的综合能力。作为中国模型的代表，通义千问 3 跻身两项排名前列，展现出强大的复杂推理能力。

此外，部分模型在某一领域表现突出，但在另一领域相对较弱：比如，Gemini 2.5 Pro 在奥林匹克竞赛级推理排名中以较大优势领先，却未进入多模态榜单前十，这表明其优势在于专项高阶推理，而非通用多模态任务；豆包 1.5 Pro（思考模式）：在多模态推理中表现优异，排名第三，但未进入奥林匹克竞赛级推理前十，这说明其擅长通用任务，但在高度抽象和逻辑问题解决方面仍有提升空间。

总体而言，在高阶推理能力测评中，推理模型表现有益，通用模型则相对滞后，这一梯度差异与行业发展规律高度契合。它深刻揭示了人工智能产业正经历

从 “追求全场景通用能力覆盖” 向 “聚焦专用场景突破与深度效能优化” 的关键演进，标志着技术发展从 “广度扩张” 阶段迈向“深度精耕”的新阶段。

结论与展望

本次测评揭示了 AI 高阶推理能力的发展现状：一方面，美国模型在多模态和奥赛推理中表现突出，优势明显；目前为止，高阶推理仍是中国模型的明显短板，在深层语境理解、复杂推理链或创造性解决问题上常显不足，这是未来需弥补的关键差距。另一方面，推理模型的表现优于通用模型。由于推理模型在架构、训练数据和微调策略上聚焦提升推理精度与逻辑性，而通用模型虽任务广泛，却因能力分散，在高阶推理中表现较弱。

未来，人工智能需在跨模态深度融合、极端复杂问题创造性解决上持续突破。而中国模型可依托本土场景理解优势，针对性补足高阶推理短板，推动“真智能”向更广阔的应用场景迈进。