Evaluation of Advanced AI Reasoning Capabilities in Chinese-Language Contexts

Zhenhui (Jack) Jiang*[1], Yi Lu[1], Yifan Wu[1], Haozhe Xu[2], Zhengyu Wu[1], Jiaxin Li[1]

1 HKU Business School, The University of Hong Kong, Hong Kong

2 School of Management, Xi'an Jiaotong University, P. R. China.

Abstract

With large language models (LLMs) evolving from "able to chat" toward "able to think", artificial intelligence technology has experienced explosive growth in 2025. However, their deficiencies in advanced reasoning are also becoming more apparent. In light of this, the Artificial Intelligence Evaluation Laboratory (AIEL), led by Professor Jack Jiang at the University of Hong Kong, assessed 37 large language models from China and the US released up to September 2025, focusing on both multimodal reasoning and Olympiad-level reasoning capabilities. On one hand, they found that in multimodal reasoning the GPT series leads decisively with Doubao 1.5 Pro (Thinking) a competitive challenger; on the other hand, their results showed, in Olympiad-level reasoning, GPT-5 (Thinking) and Gemini 2.5 Pro performed exceptionally well, topping the leaderboard. Overall, US models hold a clear advantage in advanced reasoning. Although Chinese models have made notable progress in multimodal reasoning, they still exhibited performance gaps in more complex reasoning tasks.

INTRODUCTION

Since the start of 2025, AI has been evolving at a breakneck speed. Large language models are shifting from "chatting" to "reasoning". Yet in more complex real-world scenarios—such as explaining a physics formula by combining text and illustrations, extracting trends from business reports that mix prose with charts, or solving multi-step Olympiad math problems—AI's deficiencies in higher-order reasoning are becoming increasingly visible. For instance, some models can handle plain text but struggle to integrate cross-modal information, others solve routine questions perfectly but stall on more challenging and novel problems.

Cite this paper as:
Jiang, Z. J., Lu, Y., Wu, Y. F., Xu, H. Z., Wu Z. Y., & Li, J. (2025). Evaluation of Advanced AI Reasoning Capabilities in Chinese-Language Contexts. HKU Business School Working Paper.

* Zhenhui (Jack) Jiang is the corresponding author. Email: jiangz@hku.hk

These capability gaps directly constrain AI's application in education, scientific research, business decision-making, among numerous other fields. How should we rigorously measure AI's "true intelligence"? Two core benchmarks stand out: multimodal reasoning, which reflects the ability to infer across information in different modalities; and Olympiad-level reasoning, which represents the ability to apply higher-order reasoning to complex problems.

Against this backdrop, the Artificial Intelligence Evaluation Lab (AIEL), led by Professor Jack Jiang at HKU Business School, conducted a systematic evaluation. The team comprehensively assessed 37 large language models released in China and the United States up to September 2025—including 14 reasoning models, 20 general-purpose models, and 3 integrated systems—on multimodal and Olympiad-level reasoning, aiming to offer insights into the current landscape and future outlook of advanced AI.

The results revealed that in multimodal reasoning, OpenAI's GPT series continued to dominate, though China's Doubao 1.5 Pro (Thinking) surged into the global top tier. In Olympiad-level reasoning, US models dominated, with GPT-5 (Thinking) leading by a decisive margin. Overall, reasoning models exhibited greater capabilities in higher-order thinking, while Chinese models still showed considerable gaps in complex reasoning and problem-solving.

EVALUATION FOCUS

- **Multimodal reasoning**: This refers to a model's ability to integrate multiple modalities of information, such as text, images, and charts, and perform cross-modal analysis and logical inference. In the context of education, it can help students connect textbook explanations with diagrams to grasp abstract concepts. In business analytics, it can help marketers forecast market trends by combining text and charts from market reports. In short, multimodal reasoning is a core competency for AI to tackle real-world complexities.

- **Olympiad-level reasonin**g: This evaluates models' performance regarding high-difficulty problems from competitions like the International Mathematical Olympiad (IMO). These problems require complex logical structures, multi-step derivations, and innovative thinking. They often lack a single "correct" answer, but instead test whether AI can "think outside the box" and find optimal solutions. Olympiad-level reasoning is a stringent test for determining whether a model possesses genuine "intelligence."

EVALUATION MODELS

A total of 37 mainstream AI models released in China and the US up to mid-September 2025 were comprehensively tested and evaluated by the research team (see Table 1). However,

during the multimodal evaluation phase, five of these models were necessarily excluded because they lack the capacity to process and interpret visual and textual (multimodal) information.

Table 1 List of Evaluated Models

Model Name (English)	Country	Model Type	Developer	Multimodal Support
360 Zhinao 2-o1	China	General Purpose	360	×
Baichuan4-Turbo	China	General Purpose	Baichuan AI	√
Claude 4 Opus	United States	General Purpose	Anthropic	√
Claude 4 Opus (Thinking)	United States	Reasoning	Anthropic	√
DeepSeek-R1	China	Reasoning	DeepSeek	×
DeepSeek-V3	China	General Purpose	Deepseek	×
Doubao 1.5 Pro	China	General Purpose	ByteDance	√
Doubao 1.5 Pro (Thinking)	China	Reasoning	ByteDance	√
Ernie 4.5-Turbo	China	General Purpose	Baidu	√
Ernie X1-Turbo	China	Reasoning	Baidu	×
Gemini 2.5 Flash	United States	General Purpose	Google	√
Gemini 2.5 Pro	United States	Reasoning	Google	√
GLM-4-plus	China	General Purpose	Zhipu AI	√
GLM-Z1-Air	China	Reasoning	Zhipu AI	√
GPT-4.1	United States	General Purpose	OpenAI	√
GPT-4o	United States	General Purpose	OpenAI	√
GPT-5 (Auto)	United States	Unified System	OpenAI	√
GPT-5 (Thinking)	United States	Reasoning	OpenAI	√
GPT-o3	United States	Reasoning	OpenAI	√
GPT-o4 mini	United States	Reasoning	OpenAI	√
Grok 3	United States	General Purpose	xAI	√
Grok 3 (Thinking)	United States	Reasoning	xAI	√
Grok 4	United States	Unified System	xAI	√
Hunyuan-T1	China	Reasoning	Tencent	√
Hunyuan-TurboS	China	General Purpose	Tencent	√
Kimi	China	General Purpose	Moonshot AI	√
Kimi-k1.5	China	Reasoning	Moonshot AI	√
Llama 3.3 70B	United States	General Purpose	Meta	√
MiniMax-01	China	General Purpose	MiniMax	√
Qwen 3	China	General Purpose	Alibaba	√
Qwen 3 (Thinking)	China	Reasoning	Alibaba	√
SenseChat V6 (Thinking)	China	Reasoning	SenseTime	√
SenseChat V6 Pro	China	General Purpose	SenseTime	√
Spark 4.0 Ultra	China	General Purpose	iFlytek	√
Step 2	China	General Purpose	Stepfun AI	√

Step R1-V-Mini	China	Reasoning	Stepfun AI	√
Yi-Lightning	China	General Purpose	01.AI	×

Note: Models are listed in alphabetical order by their English names.

EVALUATION PROCEDURE AND RESULTS

Multimodal Reasoning

(1) **Multimodal Reasoning Tasks**

All tasks are "Vision-Language Tasks," meaning that it is insufficient to obtain the correct answer by relying solely on either the text or image. This approach effectively avoids single-modality bias. Tasks including the following four categories:

1) *Basic logical reasoning*: Tasks included deduction, induction, and abduction, adapted from classic frameworks in cognitive psychology and formal logic. Each problem was restructured into an image-text format.

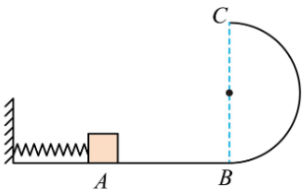
2) *Common-sense reasoning*: Tasks involved scenarios based on everyday life, combined with images, to test whether a model can ground its reasoning in both visual context and text.

3) *Discipline-specific reasoning*: Task included single- or multiple-choice problems to test specific discipline knowledge and application. Questions were sourced from recent high school and university entrance examinations in China and from the widely used multi-discipline multimodal question dataset MMMU2.

4) *Social phenomena reasoning*: Tasks included customized multimodal tasks built around real-world contexts like environmental protection, public behavior, social responsibility, moral judgment, and ethical conflict. Unlike traditional knowledge-based Q&A, these tasks emphasized contextual understanding, identifying ethical dilemmas, and making judgments after integrating multiple modalities. This examined a model's ability to extend logical inference to complex, real-world scenarios.

Table 2 presents sample questions for multimodal reasoning tasks.

Table 2 Multimodal Reasoning Example Questions

Category	Question
Basic Logical Reasoning	Riders must be over 1.5 meters tall to get on the roller coaster. Does the person in the photo meet the requirement? A. Yes B. No 
Common-sense Reasoning	Looking at the picture, how many actual cats can you spot? 
Discipline-specific Reasoning	As shown in the figure, the smooth horizontal track AB is connected to a smooth semicircular track BC in a vertical plane at point B. A small block compresses a light spring at point A, and is then released from rest. After leaving the spring, the block enters the semicircular track and just manages to reach the highest point C. Which of the following statements is correct? A. The net force on the block at point C is zero. B. The block's speed at point C is zero. C. The block's centripetal acceleration at point C is equal to the acceleration due to gravity. D. The elastic potential energy stored in the spring at point A is equal to the kinetic energy of the block at point C. 
Social Phenomena Reasoning	Briefly explain the underlying message of the cartoon. 

(2) Evaluation Criteria

The evaluation of multimodal reasoning focused on how well a model can integrate and reason over multiple modalities including text and images, with accuracy as the core criterion. For closed-ended tasks with clear correct answers, such as answering image-based commonsense questions, responses were scored as correct or incorrect. For open-ended tasks that require fusing text and visual information, such as decision analysis based on both text and images, the plausibility of the responses was scored on a 7-point Likert scale.

(3) Evaluation Procedure

To ensure fairness and rigor in scoring, the research team carefully matched evaluators to task types based on their disciplinary expertise. Multimodal reasoning tasks were assessed by 29 Ph.D. students and master's students from top universities in China.

To standardize evaluation criteria and enhance result reliability, a structured calibration process was implemented: When encountering a new question type, each rater first completed a warm-up by scoring three randomly assigned model responses (scores excluded from final results) to familiarize themselves with the question's characteristics and scoring standards. Following calibration, the rater formally scored all model responses for that question in random order—eliminating potential bias from scoring sequence effects. In practice, for each question, each rater scored a total of n (number of model responses) + 3 (warm-up responses) items.

(4) Results

The multimodal reasoning scores of the models are summarized in Table 3.

Table 3 Multimodal Reasoning Capability Ranking

Ranking	Model Name	Accuracy
1	GPT-5 (Thinking)	91
2	GPT-4.1	90
3	GPT-o3	87
4	Doubao 1.5 Pro (Thinking)	85
4	GPT-5 (Auto)	85
6	GPT-4o	84
7	Claude 4 Opus (Thinking)	83
8	Doubao1.5 Pro	82
8	Grok 3 (Thinking)	82
10	Qwen 3	81
11	Kimi-k1.5	80
11	SenseChat V6 (Thinking)	80
11	Step R1-V-Mini	80
14	Grok 4	79
14	GPT-o4 mini	79
14	Hunyuan-T1	79
17	GLM-4-plus	78
17	Qwen 3 (Thinking)	78
19	Gemini 2.5 Flash	77
19	GLM-Z1-Air	77
21	Llama 3.3 70B	76
22	SenseChat V6 Pro	75
22	Gemini 2.5 Pro	75
23	Ernie 4.5-Turbo	74
24	Step 2	73
26	Hunyuan-TurboS	71
26	Claude 4 Opus	71
28	Spark 4.0 Ultra	68
28	MiniMax-01	68
30	Baichuan4-Turbo	67
31	Grok 3	66
32	Kimi	63

Note: The scores have been rounded to the nearest integer

Based on their performance in multimodal reasoning, we grouped the models into four distinct tiers (as shown in Figure 1).

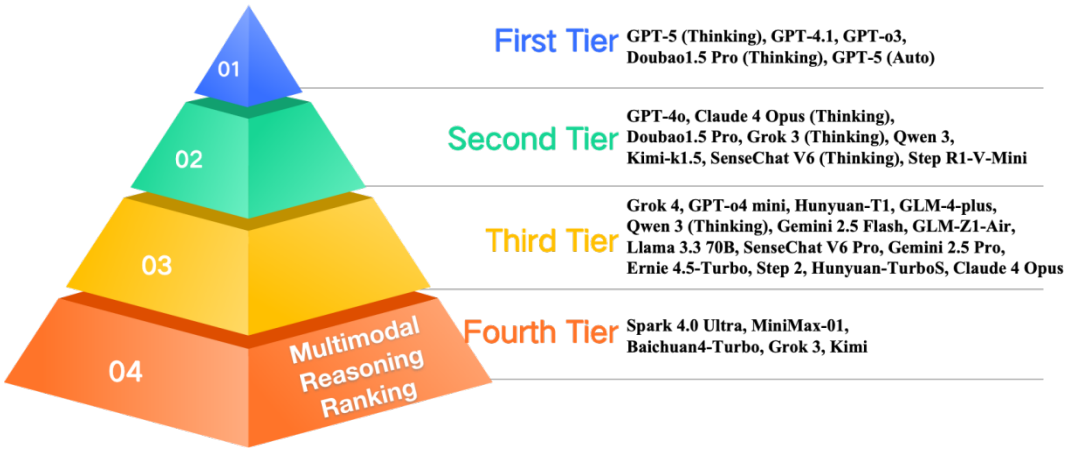


Figure 1 Multimodal Reasoning Capability Tiers

The distribution of scores reveals a distinctly tiered landscape, underscoring sharp disparities in multimodal reasoning capability.

Top Tier (Tier 1):
The top tier consists of five models with scores above 85, representing the benchmark for state-of-the-art multimodal reasoning. GPT-5 (Thinking) secures the top position with 91 points, followed closely by GPT-4.1 (scoring 90) and GPT-o3 (87). Doubao 1.5 Pro (Thinking) and GPT-5 (Auto) both scores 85, tying for third place. Notably, the GPT family dominates this tier, claiming three of the top five spots. Doubao stands out as the only Chinese model in this elite group, showcasing strong multimodal integration.

Medium Tiers (Tier 2 and 3):
Competition is tighter in the mid-range, but these models still trail the leaders. GPT-4o (84) and Claude 4 Opus (Thinking) (83) show strong performance, while Hunyuan-TurboS (71) and Claude 4 Opus (71) anchor the middle pack.

Bottom Tier (Tier 4):
At the other end of the spectrum, five models cluster between 63 and 68, underscoring the steep capability divide. Spark 4.0 Ultra, MiniMax-01, Baichuan4-Turbo, Grok 3, and Kimi struggle to keep pace, highlighting the widening gulf between cutting-edge leaders and those still catching up.

Key Finding 1: "Thinking Mode" as a Critical Performance Amplifier

Comparing reasoning models with their general-purpose counterparts, the former demonstrates clear advantages on complex multimodal tasks. For some models, performance improves drastically when switched to "thinking mode": Grok 3 (82 in thinking mode vs. 66 in general mode, +16 points), Claude 4 Opus (83 in thinking mode vs. 71 in general mode, +12 points), and SenseChat V6 (80 in thinking mode vs 75 in general mode). Given that multimodal tasks demand substantial computing power, vast training data, and high costs,

reasoning models have innate advantages on these fronts, making it easier for them to stand out.

Key Finding 2: GPT Models Lead Globally, with Premier Domestic Chinese Models have Advanced into the Global Forefront

The **GPT series** occupies four of the five top-performing positions in the top tier (85+ score bracket) (i.e., GPT-5 (Thinking), GPT-4.1, GPT-o3, and GPT-5 (Auto), with scores distributed in a stepwise pattern (91 – 90 – 87 – 85). Together, they form a complete top-tier capability chain. Its edge likely stems from foundational advantages in multimodal data integration, closely tied to large training scale and high-quality text – image pretraining strategies.

Among Chinese models, Doubao 1.5 Pro (Thinking), with a score of 85, is the only Chinese model to break into the top five. The minimal gap between its thinking and general-purpose modes indicates world-class native multimodal reasoning capability. Nonetheless, US models still lead overall at the cutting edge of multimodal capabilities.

Olympiad-level Reasoning

(1) Olympiad-level Reasoning Tasks

The Olympiad-level reasoning question set was drawn from recent International Mathematical Olympiad (IMO), Chinese Mathematical Olympiad (CMO), and other prestigious competitions (examples in Table 4). These problems are far more challenging than standard high school or college entrance exam questions. They typically involve complex logical structures, multi-step reasoning, and creative problem-solving. These tasks assessed whether models can go beyond rote memorization to demonstrate real reasoning capability under pressure.

Table 4 Olympiad Question Example

Category	Question
Olympiad-level Reasoning	Determine all real numbers α such that, for every positive integer n, the integer $\lfloor \alpha \rfloor + \lfloor 2\alpha \rfloor + \cdots + \lfloor n\alpha \rfloor$ is a multiple of n. (Note that $\lfloor z \rfloor$ denotes the greatest integer less than or equal to z. For example, $\lfloor -\pi \rfloor = -4$ and $\lfloor 2 \rfloor = \lfloor 2.9 \rfloor = 2$.)

(2) Evaluation Criteria

For Olympiad-level reasoning, given the high difficulty and diversity of solutions, the research team designed a targeted evaluation rubric for model responses.

1) **Correctness**: Since most of the current models were unable to perfectly solve the Olympiad problems in our set, we closely evaluated the correctness of the solution process presented in their responses to distinguish their performance at a finer granularity. A 0 – 10 scale was adopted, allowing for 0.5-point increments for precise gradation.

2) **Logical Coherence**: This dimension examined whether the model proceeds in clear steps with a sound reasoning structure. We focused on the continuity of the solution approach, the completeness of the reasoning chain, and whether there were logical gaps. Scoring used a 7-point Likert scale.

3) **Methodological Innovation**: This dimension assessed the model's ability to break from standard routines and select more efficient, streamlined strategies, rather than rigidly applying existing methods or relying on overly complex methods. Scoring also used a 7-point Likert scale.

(3) Evaluation Procedure

Olympiad-level problems, which require specialized knowledge, were scored by three individuals who had participated in the Math Olympiad training team of China and Hong Kong (including one IMO silver medal winner).

The scoring process mirrors that of the multimodal reasoning evaluation. For each new question type, every assessor first completes three randomly assigned "warm-up questions" (not included in the final results). Following this warm-up, they proceed to formally score all model responses for that question, which are presented in randomized order.

(4) Evaluation Results

The ranking of models' Olympiad-level reasoning capability is shown in Table 5.

Table 5 Olympiad-level Reasoning Capability Ranking

Ranking	Model Name	Correctness	Logical Coherence	Methodological Innovation	Overall Weighted Score
1	GPT-5 (Thinking)	48	47	44	48
2	Gemini 2.5 Pro	48	39	36	44
3	GPT-o3	36	42	39	38
4	Claude 4 Opus (Thinking)	30	36	39	33
5	Gemini 2.5 Flash	35	28	31	32
5	GPT-o4 mini	32	33	33	32
7	Qwen 3 (Thinking)	29	25	28	28
7	Step R1-V-mini	26	33	22	28
9	GLM_Z1_Air	27	31	22	27
9	SenseChat V6 (Thinking)	27	28	22	27
11	Qwen 3	25	31	17	26
12	Ernie 4.5-Turbo	25	25	19	24
13	Grok 3 (Thinking)	21	28	25	23
14	GPT-5 (Auto)	22	22	28	22
14	DeepSeek-V3	26	14	22	22
16	Claude 4 Opus	22	17	31	21
17	Doubao 1.5 Pro (Thinking)	22	17	22	20
17	DeepSeek-R1	17	25	22	20
19	Grok 3	20	19	17	19
19	Grok 4	19	17	25	19
21	Ernie X1-Turbo	17	19	14	17
21	Hunyuan-T1	17	17	19	17
21	Hunyuan-TurboS	17	17	19	17
21	Kimi-k1.5	17	19	11	17
25	Doubao 1.5 Pro	16	17	19	16
26	GLM-4-plus	12	17	8	13
27	GPT-4o	13	8	19	12
27	Spark 4.0 Ultra	13	11	14	12
29	Baichuan4-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	6	11	14	8
33	SenseChat V6 Pro	8	8	6	8
35	MiniMax-01	5	11	8	7
35	Step 2	6	8	8	7
35	360 Zhinao 2-o1	7	6	8	7

Note: The scores have been rounded to the nearest integer

- **Overall Olympiad-level Reasoning:**

Olympiad-level reasoning represented the most demanding part within this evaluation, clearly revealing the substantial performance disparity between reasoning models and general-purpose models when confronted with highly complex problems.

The top three models by weighted score are all developed by US institutions, showing a "multi-dimensional advantage" across correctness, logical coherence, methodological creativity, and overall Olympiad-level reasoning capability. Specifically, Gemini 2.5 Pro (44) leads with a significant margin, followed by GPT-o3 (38) and Claude 4 Opus (Thinking, 33). Among Chinese models, Qwen 3 (Thinking, 28) and Step R1-V-Mini (28) performed relatively well.

It is worth noting that some prominent Chinese models (such as the DeepSeek and Doubao series), which had performed strongly in previous evaluation tasks, delivered only mediocre results in this part of the evaluation.

Here is the detailed analysis:

- **Correctness:**

GPT-5 (Thinking) (48) and Gemini 2.5 Pro (48) top the list, followed by GPT-o3 (36) and Gemini 2.5 Flash (35). Among Chinese models, Qwen 3 (Thinking, 29), GLM_Z1_Air (27) and SenseChat V6 (Thinking) (27) performed relatively well, but their results still fall markedly short of the top models.

- **Logical Coherence:**

GPT-5 (Thinking, 47) secures the top position, with GPT-o3 (42) and Gemini 2.5 Pro (39) also exhibiting strong performance. Notably, some Chinese models (e.g., Step R1-V-Mini, GLM_Z1_Air, and Qwen 3) display competitive strength relative to global peers in logical coherence.

- **Methodological Innovation**

GPT-5 (Thinking, 44), GPT-o3 (39), and Claude 4 Opus (Thinking, 39) top the list. Among Chinese models, Qwen 3 (Thinking, 28) shows some innovation, but still lag behind top models.

Key Finding: "Thinking Mode" Also Outperformed General-purpose Mode

When comparing general-purpose versions and reasoning versions of the same model, the latter consistently achieve higher scores across Olympiad-level tasks. For example:

- Claude 4 Opus: weighted score rises from 21 (General-purpose) to 33 (Thinking)
- Grok 3: from 19 (General-purpose) to 23 (Thinking)
- Doubao 1.5 Pro: from 16 (General-purpose) to 20 (Thinking)

This finding shows that "Thinking Mode" is an effective strategy to activate higher-order reasoning capabilities and innovation, making it an important direction for performance optimization.

Based on the performance in Olympiad-level reasoning tasks, we grouped the models into four tiers, as shown in Figure 2.

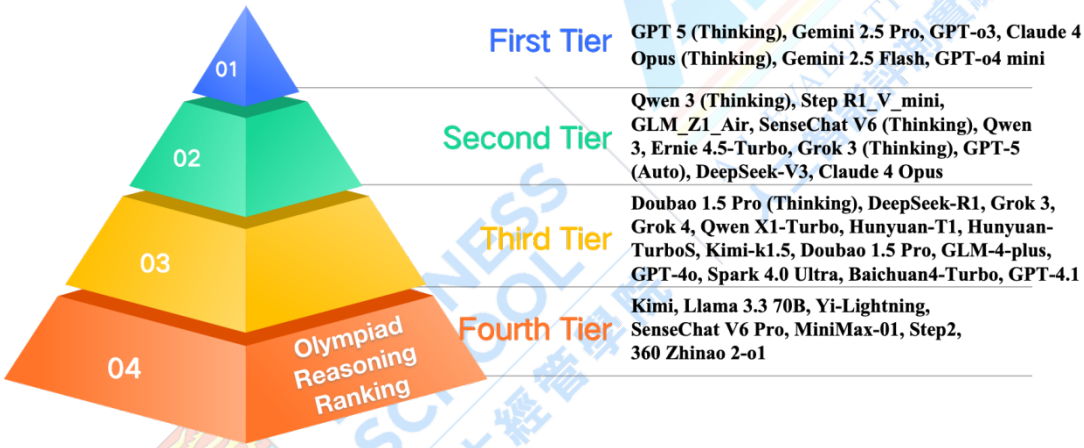


Figure 2 Olympiad Reasoning Capability Tiers

Overall Analyses

Based on the evaluation, the standout performer in both multimodal tasks and Olympiad-level reasoning is GPT-5 (Thinking), which ranks first in both lists and can be considered a top-tier model with robust overall capability. From the same institution, GPT-o3 is placed in the top five for both rankings, highlighting its solid all-around strength. As a representative of Chinese models, Qwen 3 ranks among the top-performing Chinese models, demonstrating strong complex reasoning ability.

Additionally, some models excel in one domain but are relatively weak in the other. For example, Gemini 2.5 Pro leads the Olympiad-level reasoning rankings by a sizable margin but does not make the top ten in the multimodal rankings, indicating its advantage lies in

specialized high-order reasoning rather than general multimodal tasks. Doubao 1.5 Pro (Thinking) performed excellently in multimodal reasoning, ranking fourth, but does not enter the top ten for Olympiad-level reasoning, suggesting it is strong on general tasks but relatively weak in highly abstract and complex problem solving.

Overall, in advanced reasoning evaluations, reasoning models stand out, while general-purpose models lag behind. This tiered differentiation aligns closely with industry trends, revealing a pivotal shift in AI from "pursuing broad, all-scenario coverage" to "targeted breakthroughs and efficiency optimization" in specialized domains—signaling a transition from a phase of breadth expansion to one of depth-focused refinement.

CONCLUSIONS

This evaluation offers valuable insights into the current landscape of advanced AI reasoning capabilities, highlighting two key observations.

On one hand, US-developed models maintain a clear advantage in this domain, consistently excelling in multimodal and Olympiad-level reasoning performance. In contrast, advanced reasoning remains a notable bottleneck for Chinese models, which often struggle to match such performance in scenarios requiring deep contextual understanding, intricate inference chains, or creative problem-solving, reflecting a critical gap to be addressed in future development.

On the other hand, a distinct pattern emerges: models specifically optimized for reasoning tasks outperform general-purpose ones by a significant margin. This is largely because specialized reasoning models are tailored with architectures, training data, and fine-tuning strategies focused on enhancing inference accuracy and logical rigor, whereas general-purpose models are constrained in reasoning time and effort, resulting in less robust performance in advanced reasoning scenarios.

Looking ahead, AI must continue to make breakthroughs in multimodal integration and in creative problem-solving under conditions of extreme complexity. Chinese models, leveraging their advantage in local context understanding, have the opportunity to strategically address weaknesses in advanced reasoning. By doing so, they can help drive AI closer to "true intelligence" and expand its reach into broader and more impactful applications.