Evaluating the Reasoning Capabilities of Large Language Models 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 the rapid iteration of AI technologies, reasoning capabilities have become a core indicator for measuring the intelligence level of large language models (LLMs) and a focus of research in both academia and industry. This report aims to establish a systematic, objective, and comprehensive evaluation framework to assess AI reasoning capabilities. We compared 36 LLMs on various text-based reasoning tasks in Chinese-language contexts and found that GPT-o3 achieved the highest score in the basic logical reasoning evaluation, while Gemini 2.5 Flash led in contextual reasoning evaluation. In terms of overall ranking, Doubao 1.5 Pro (Thinking) secured the top position, closely followed by OpenAI’s recently released GPT-5 (Auto). Several Chinese-developed LLMs—including Doubao 1.5 Pro, Qwen 3 (Thinking), and DeepSeek-R1—also ranked among the leaders, demonstrating the strong reasoning performance of frontier Chinese AI technologies. Further analysis of model efficiency revealed that most models with superior reasoning capabilities often incurred higher costs in terms of token efficiency, response time, and API usage. Notably, Doubao 1.5 Pro not only achieved outstanding reasoning performance but also demonstrated high model efficiency.

Keywords: Large Language Model, LLM, Reasoning Capability, Model Efficiency, Logic Reasoning, Contextual Reasoning, Chinese-language Context

---

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

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

INTRODUCTION

Over the past few months, reasoning capabilities have emerged as the new frontier in the global race to advance Large Language Models (LLMs). Following OpenAI's launch of its reasoning models and DeepSeek-R1's rise to national prominence for its problem-solving prowess, the focus has shifted toward the central question: Which LLM performs best on reasoning tasks?

To address this issue, the Artificial Intelligence Evaluation Lab (AIEL) at HKU Business School developed a comprehensive evaluation framework that assesses basic logical inference and contextual reasoning (Figure 1). Building on this framework, the team curated a carefully designed set of questions across multiple difficulty levels to conduct a rigorous benchmark evaluation.

The study included 36 notable LLMs from China and the USA. This included 14 reasoning models, 20 general-purpose models, and two unified systems. All were tested within a Chinese-language context. The results revealed that Doubao 1.5 Pro Thinking was best, with a composite score of 93, closely followed by the recently released GPT-5 (Auto). Overall, the Chinese models demonstrated strong capabilities in reasoning tasks.

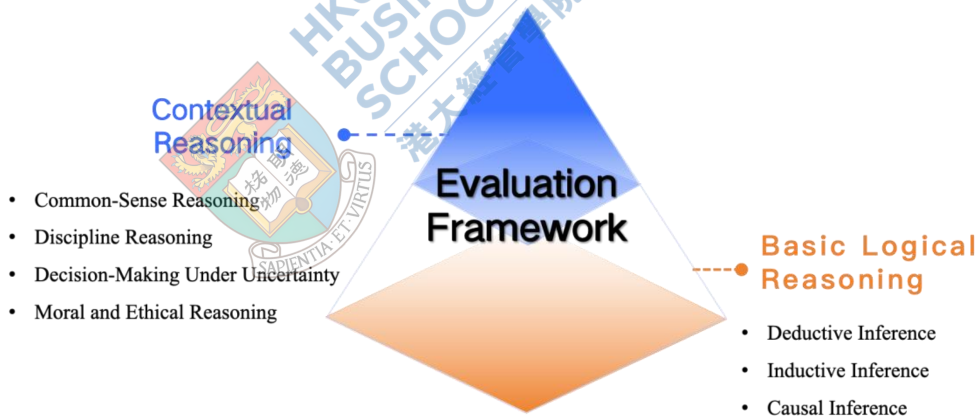

Figure 1. Reasoning Ability Assessment System

EVALUATION METHODOLOGY

(1) Models for Evaluation

The study evaluated the following LLMs from both China and the USA (Table 1). Due to local deployment constraints, Llama 4 was excluded from this round of assessment.

Table 1. Evaluated LLMs

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

Note: Models are listed in alphabetical order within each country and model type.

(2) Task Categories and Test Set

In this study, the reasoning evaluation questions were divided into two task categories: Basic Logical Reasoning and Contextual Reasoning (Table 2). Together, these categories captured a model's overall performance, spanning from fundamental reasoning skills to more advanced reasoning abilities.

Table 2. Evaluation Task Categories

Category	Category Definition	Subcategory	Subcategory Definition
Basic Logical Reasoning	Ability to understand and apply fundamental logical rules to make valid inferences.	Deductive	Drawing specific conclusions from general principles or premises.
		Inductive	Drawing general conclusions from specific observations.
		Abductive	Inferring the most plausible conclusion given a set of observations.
Contextual Reasoning	Ability to integrate diverse knowledge, logic, and strategies to solve complex problems, handle uncertainty, and make evaluative judgments.	Common-sense	Interpreting or making judgments based on everyday common knowledge.
		Discipline-based	Applying knowledge from a particular discipline to solve complex questions.
		Decision-Making Under Uncertainty	Making well-reasoned and optimized decisions despite incomplete data, ambiguity, or risk.
		Moral and Ethical	Using ethical norms and social values to judge contexts, analyse dilemmas, and propose actions.

Test Set: In this evaluation, 90% of the test items were either newly created or extensively adapted, and the remaining 10% were drawn from real examination papers from the 2024 and 2025 China National College Entrance Examination (Gaokao), as well as internationally recognized benchmark datasets. Representative sample questions are provided in Table 3.

Experts: The evaluation was conducted by a team of 38 postgraduate researchers from China's leading universities. They strictly followed the standardized scoring protocol to ensure consistency and fairness.

Table 3. Representative Sample Questions

Category	Question
Basic Logical Inference (Deductive)	A seminar has 18 participants. The following information is known: (1) At least 5 young teachers are female. (2) At least 6 female teachers are middle-aged or older. (3) At least 7 young females are teachers. **Question:** Based on the above information, which of the following conclusions must be true? **Options:** A) Some young teachers are not female. B) Some young females are not teachers. C) There are at least 11 young teachers. D) There are at least 13 female teachers.
Contextual Reasoning (Common-Sense)	What do you call your mother's sister's husband's son's biological elder brother's mother?
Contextual Reasoning (Discipline-Based)	Given that **b** is the arithmetic mean of **a** and **c**, and the line $ax + by + c = 0$ intersects the circle $x^2 + y^2 + 4y - 1 = 0$ at points A and B, what is the minimum value of $\vert AB \vert$?
Contextual Reasoning (Decision-Making Under Uncertainty)	A novel infectious disease has broken out, and current vaccine production can only cover 30% of the population. However, the mutation rate of the virus is unpredictable, and the vaccine's efficacy and side effects are still unclear. What vaccine distribution strategy should be adopted to best control the outbreak, protect vulnerable populations, and address the uncertainties around mutation and vaccine effectiveness?
Contextual Reasoning (Moral and Ethical)	As a newcomer to the workplace, you face a manager who is extremely demanding and frequently pressures you. He asks you to participate in ethically questionable practices, such as concealing financial issues within the company, and he claims that doing so is necessary for you to gain his approval and secure promotion opportunities. Would you follow his instructions or adhere to ethical principles?

Evaluation Criteria: Each model's reasoning performance was assessed across three core criteria – accuracy, logical coherence and conciseness (Figure 2).

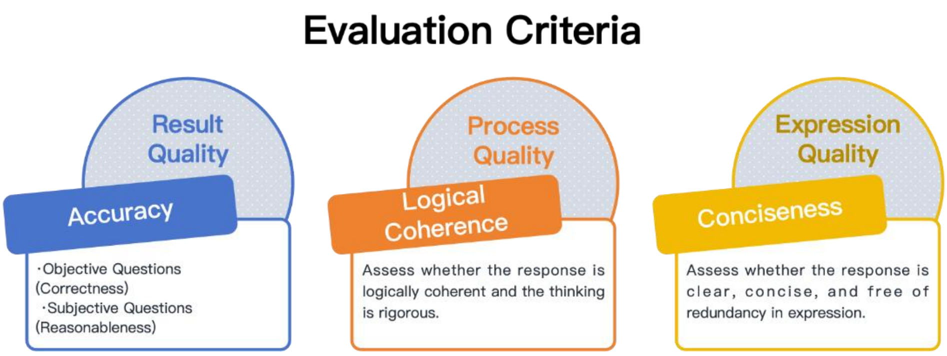

Figure 2. Evaluation Criteria for Reasoning Questions

RESULTS AND ANALYSIS

(1) Basic Logical Inference

As shown in Table 4, GPT-o3 achieved the highest score in basic logic with 97 points, closely followed by Doubao 1.5 Pro (96) and Doubao 1.5 Pro (Thinking) (95). In contrast, models like Llama 3.3 70B (64) and 360 Zhinao 2-o1 (59) displayed notable weaknesses in this category.

Table 4. Ranking for Basic Logical Inference Capability

Ranking	Model Name	Basic Logical Inference Weighted Score
1	GPT-o3	97
2	Doubao 1.5 Pro	96
3	Doubao 1.5 Pro (Thinking)	95
4	GPT-5 (Auto)	94
5	DeepSeek-R1	92
6	Qwen 3 (Thinking)	90
7	Gemini 2.5 Pro	88
7	GPT-o4 mini	88
7	Hunyuan-T1	88
7	Ernie X1-Turbo	88
11	GPT-4.1	87
11	GPT-4o	87
11	Qwen 3	87
14	DeepSeek-V3	86
14	Grok 3 (Thinking)	86
14	SenseChat V6 (Thinking)	86
17	Claude 4 Opus	85
17	Claude 4 Opus thinking	85
19	Gemini 2.5 Flash	84
20	SenseChat V6 Pro	83
21	Hunyuan-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	Ernie 4.5-Turbo	74
33	Step 2	73
34	Kimi-k1.5	72
35	Llama 3.3 70B	64
36	360 Zhinao 2-o1	59

(2) Contextual Reasoning

The ranking of contextual reasoning capability is shown in Table 5.

Table 5. Ranking for Contextual Reasoning Capability

Ranking	Model Name	Common-sense Reasoning	Discipline-Based Reasoning	Decision-Making Under Uncertainty	Moral & Ethical Reasoning	Final Weighted Score
1	Gemini 2.5 Flash	98	93	89	87	92
2	Doubao 1.5 Pro (Thinking)	97	92	88	87	91
2	Gemini 2.5 Pro	93	94	90	87	91
4	Grok 3 (Thinking)	96	88	89	86	90
5	GPT-5 (Auto)	88	98	88	83	89
5	Hunyuan-T1	97	95	84	81	89
5	Qwen 3 (Thinking)	96	89	86	85	89
5	Ernie X1-Turbo	98	85	86	86	89
9	DeepSeek-R1	94	93	78	82	87
9	Qwen 3	97	79	87	86	87
9	Ernie 4.5-Turbo	96	76	87	87	87
12	Hunyuan-TurboS	96	79	83	84	86
13	Doubao 1.5 Pro	97	81	86	74	85
13	GPT-4.1	97	70	87	86	85
13	GPT-o3	90	95	73	80	85
13	Grok 3	97	69	87	86	85
13	Grok 4	82	87	82	87	85
17	DeepSeek-V3	95	81	84	77	84
19	GPT-4o	98	65	87	78	82
19	GPT-o4 mini	91	87	72	76	82
21	Claude 4 Opus thinking	96	84	72	71	81
21	MiniMax-01	96	69	83	75	81
21	360 Zhinao 2-o1	93	76	81	72	81
24	Claude 4 Opus	95	85	70	70	80
24	GLM-4-plus	93	71	83	73	80
24	Step 2	97	63	82	78	80
27	Yi- Lightning	97	59	82	79	79
27	Kimi	94	61	79	81	79
29	Spark 4.0 Ultra	91	71	75	76	78
30	SenseChat V6 Pro	86	58	84	78	77
31	GLM-Z1-Air	90	76	73	64	76
32	Llama 3.3 70B	82	52	83	81	75
33	SenseChat V6 (Thinking)	96	63	68	70	74
34	Baichuan4-Turbo	91	48	77	69	71
35	Step R1-V-Mini	96	80	37	51	66
36	Kimi-k1.5	84	79	42	58	66

The results revealed that Gemini 2.5 Flash ranked first in contextual reasoning with an overall score of 92, demonstrating no significant weakness across any categories. It performed particularly well in common-sense reasoning (98) and discipline-based reasoning (93). Both Doubao 1.5 Pro (Thinking) and Gemini 2.5 Pro followed closely with scores of 91. The former excelled in common-sense reasoning (97), while the latter showed particular strength in discipline-based reasoning and decision-making under uncertainty.

Grok 3 (Think) ranked fourth with 90, reflecting consistent performance across all evaluated categories. In addition, the series of GPT, Ernie, DeepSeek, Hunyuan, and Qwen also performed well, with scores between 85 to 89.

(3) Composite Ranking Results

As shown in Table 6, the 36 models assessed exhibited a clear performance gradient in the composite rankings. Doubao 1.5 Pro (Thinking) ranked first with a top composite score of 93, demonstrating consistently strong and balanced performance across both basic logical inference and contextual reasoning.

GPT-5 (Auto) (91.5 points) followed closely behind. Further analysis revealed that because GPT-5 (Auto) is enabled with the function to automatically select between the general-purpose mode and the reasoning mode, it sometimes defaulted to the general-purpose version on more difficult questions, leading to errors. In addition, GPT-o3 (91 points) and Doubao 1.5 Pro (90.5 points) ranked third and fourth, respectively.

In general, these results highlight the significant progress and growing competitiveness of China-developed LLMs in reasoning-intensive tasks.

Table 6. Composite Ranking

Ranking	Model Name	Score
1	Doubao 1.5 Pro (Thinking)	93
2	GPT-5 (Auto)	91.5
3	GPT-o3	91
4	Doubao 1.5 Pro	90.5
5	DeepSeek-R1	89.5
5	Gemini 2.5 Pro	89.5
5	Qwen 3 (Thinking)	89.5
8	Hunyuan-T1	88.5
8	Ernie X1-Turbo	88.5
10	Gemini 2.5 flash	88
10	Grok 3 (Thinking)	88
12	Qwen 3	87
13	GPT-4.1	86
14	DeepSeek-V3	85
14	GPT-o4 mini	85
16	GPT-4o	84.5
17	Hunyuan-TurboS	83.5
18	Claude 4 Opus (Thinking)	83
19	Claude 4 Opus	82.5
19	Grok 3	82.5
19	Grok 4	82.5
22	Ernie 4.5-Turbo	80.5
23	MiniMax-01	80
23	SenseChat V6 Pro	80
23	SenseChat V6 (Thinking)	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 Zhina o2-o1	70
35	Llama 3.3 70B	69.5
36	Kimi-k1.5	69

To better illustrate relative performance, the models were organized into a five-tier pyramid based on their composite scores, with higher tiers representing stronger composite reasoning ability (Figure 3).

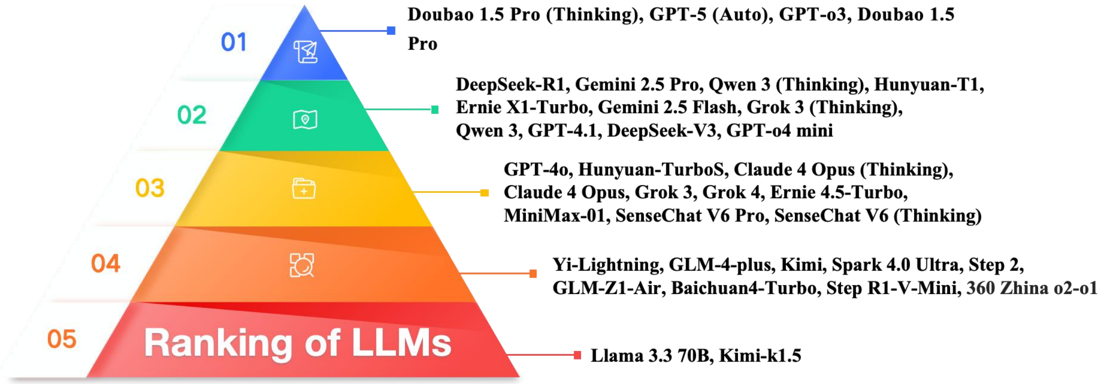

Figure 3. Ranking LLMs Based on Their Composite Scores

(4) Analysis of Performance by Model Type

The evaluation shows that the comparative advantage of reasoning models grows with task complexity. For basic logical inference, their performance is only marginally better than that of general-purpose models. However, for contextual reasoning, the gap widens significantly in favour of the specialized models.

This trend is also evident when comparing models from the same developer. Reasoning models consistently outperform their general-purpose counterparts in areas such as contextual reasoning and hallucination control, leading to higher overall composite scores. These findings highlight the competitive edge of LLMs explicitly optimized for complex reasoning tasks.

ADDITIONAL ANALYSIS: MODEL EFFICIENCY

In addition to evaluating reasoning performance, the research team conducted an in-depth analysis of model efficiency to assess their practical utility in real-world applications. Specifically, the analysis examined how quickly and cost-effectively a model can generate high-quality responses. Efficiency was assessed across three dimensions—token consumption, response time, and API usage cost (Table 7). All metrics were derived from empirical logs captured during live testing, thereby ensuring an objective evaluation.

Table 7. Evaluation Criteria for Model Efficiency

Dimension	Definition and Evaluation Focus	Measurement Method
Token Efficiency	Measures how efficiently a model processes information, minimizing redundant output	Output token count/ Input token count
Response Time	Measures how quickly the model returns a complete result to the user	Time from prompt issuance to full response
API Usage Cost	Measures user-facing cost per thousand questions based on token usage	(Average input token usage × API input price + Average output token usage × API output price) × 1000

Due to local deployment constraints or a lack of public API access, Llama 3.3 70B**,** Grok 3 (Think), Kimi-k1.5, and Step R1-V-Mini were excluded from the analysis due to missing data. Efficiency results for the remaining models are presented in Figures 4-6.

Token Efficiency

To benchmark token efficiency, we employed the output–input token ratio as a core metric, where a higher ratio indicates lower efficiency. This metric helps normalize differences in token accounting across models and ensures comparability (Figure 4).

Results show that Baichuan4-Turbo leads with an exceptionally low ratio of 1.86, followed by Llama 3.3 70B (2.49), MiniMax-01 (2.76)**,** and Step 2 (2.78)**,** all of which demonstrate excellent token efficiency.

In contrast, DeepSeek-R1 (30.77), Qwen 3 (Thinking) (31.04), and Ernie X1-Turbo (31.98), Gemini 2.5 Flash (34.78), and Gemini 2.5 Pro (38.01) exceeded a ratio of 30, indicating high token consumption and significantly lower efficiency.

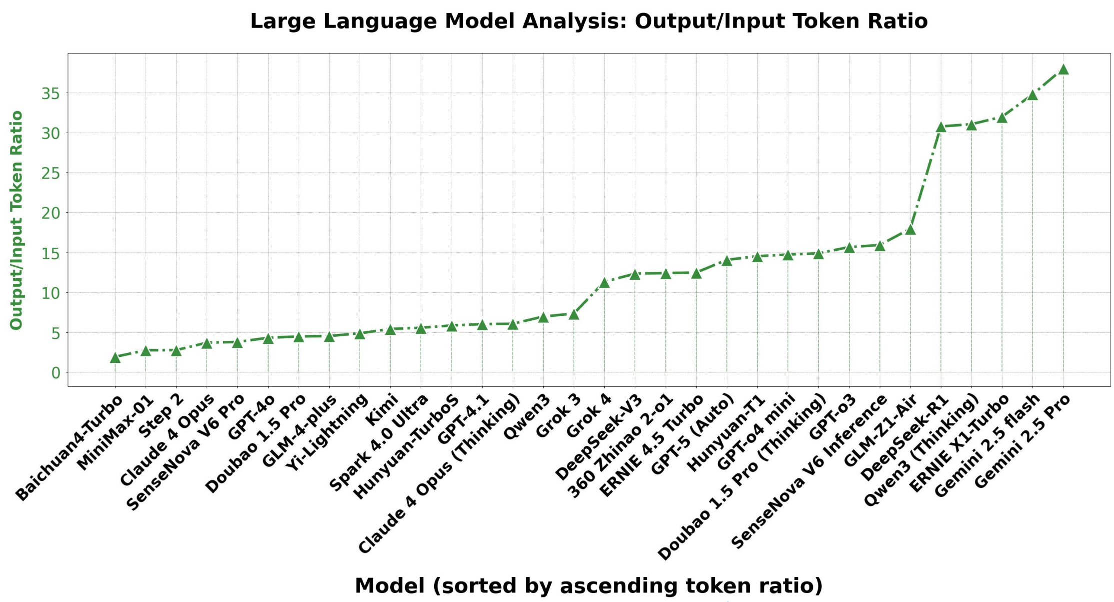

Figure 4. Output/Input Token Ratio

Response Time

In terms of speed, we measured the end-to-end average response time, defined as the duration from when a user sends a prompt to when they receive the model's complete response, noting that results may be affected by network and server conditions (Figure 5).

GPT-4o was the fastest model, averaging 5.36 seconds, followed by Baichuan4-Turbo (8.57s) and SenseChat V6 Pro (9.61s), all of which responded in under 10 seconds. Among reasoning models, DeepSeek-R1 (127.59s) and Ernie X1-Turbo (93.24s) had the slowest response times. Notably, despite its high token usage, the Gemini series delivered significantly faster responses than other models with similar token consumption, suggesting high token processing efficiency.

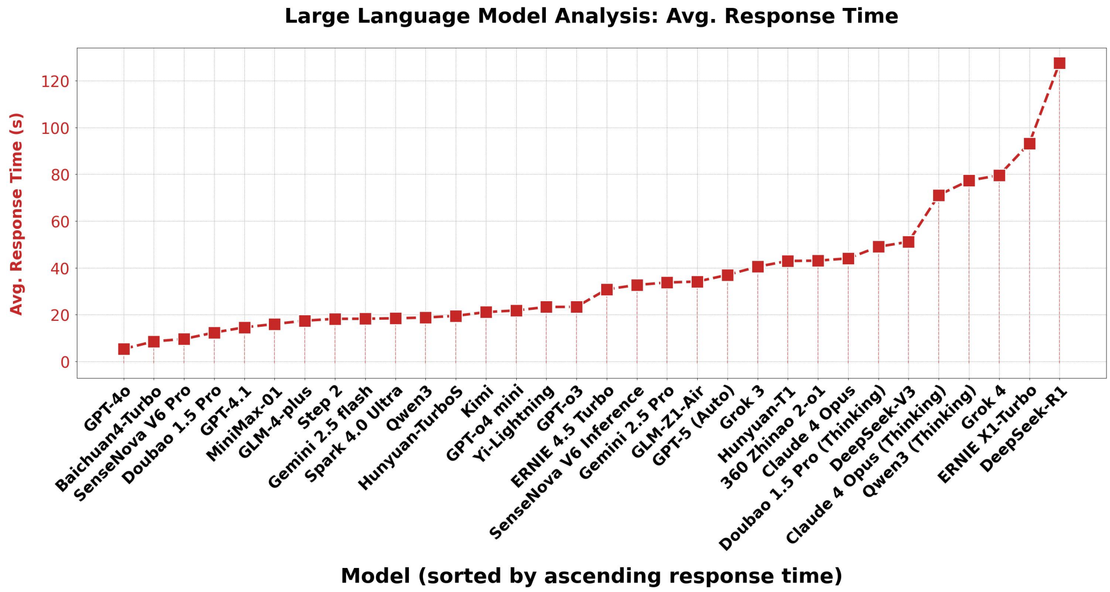

Figure 5. Response Time

API Cost

In terms of API pricing, Chinese models, as exemplified by Yi-Lightning ($0.08 per thousand questions), offer clear cost advantages due to low API rates, whereas the USA-based models are relatively expensive due to higher unit prices. Overall, general-purpose models were less costly to run than reasoning ones. It is worth noting that a low unit price does not always translate to lower total cost. For example, even though DeepSeek-R1 is marketed for value, its excessive token usage makes its actual cost ($6.77 per thousand questions) higher than that of GPT-o4 mini, decreasing its price competitiveness within the domestic arena (Figure 6).

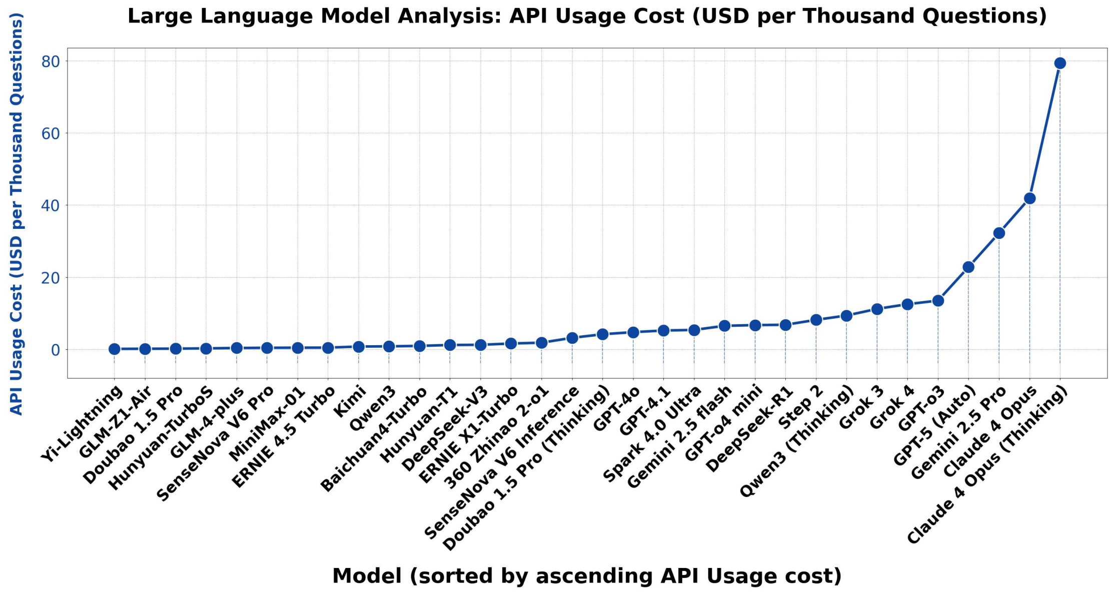

Figure 6. Usage Cost

Considering the models' overall performance in both reasoning capability and efficiency metrics, Doubao 1.5 Pro—ranked 4th in overall reasoning performance—stands out particularly for its efficiency: it ranked 7th in token efficiency, 4th in average response time, and 3rd in API usage cost. This model can be regarded as a well-rounded representative that balances high efficiency with strong intelligence, further highlighting the outstanding performance of Chinese-developed models.

GENERAL DISCUSSION

This benchmark report provides a comprehensive evaluation of the reasoning capabilities and efficiency of LLMs in Chinese-language contexts. The strong performance of Doubao 1.5 Pro (Thinking), along with impressive showings from other Chinese models, signals rapid progress and significant potential within China's LLM ecosystem.

Looking ahead, continued model iteration is expected to enhance reasoning quality further, while also optimizing latency and cost-efficiency. These improvements will be key to unlocking broader real-world adoption of LLMs across a variety of use cases.

REFERENCES

Bondarenko, A., Wolska, M., Heindorf, S., Blübaum, L., Ngomo, A. C. N., Stein, B., ... & Potthast, M. (2022, October). CausalQA: A benchmark for causal question answering. In Proceedings of the 29th International Conference on Computational Linguistics (pp. 3296-3308).

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., ... & Schulman, J. (2021). Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168.

Gao, B., Song, F., Yang, Z., Cai, Z., Miao, Y., Dong, Q., ... & Chang, B. (2024). Omni-math: A universal olympiad level mathematic benchmark for large language models. arXiv preprint arXiv:2410.07985.

He, C., Luo, R., Bai, Y., Hu, S., Thai, Z. L., Shen, J., ... & Sun, M. (2024). Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. arXiv preprint arXiv:2402.14008.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., ... & Steinhardt, J. (2021). Measuring mathematical problem solving with the math dataset. arXiv preprint arXiv:2103.03874.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. ACM computing surveys, 55(12), 1-38.

Lin, Z., Gou, Z., Liang, T., Luo, R., Liu, H., & Yang, Y. (2024). Criticbench: Benchmarking llms for critique-correct reasoning. arXiv preprint arXiv:2402.14809.

Liu, J., Cui, L., Liu, H., Huang, D., Wang, Y., & Zhang, Y. (2021, January). LogiQA: a challenge dataset for machine reading comprehension with logical reasoning. In Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence (pp. 3622-3628).

Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., ... & Gao, J. (2023). Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255.

Roemmele, M., Bejan, C. A., & Gordon, A. S. (2011, March). Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In AAAI spring symposium: logical formalizations of commonsense reasoning (pp. 90-95).

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., & Manning, C. D. (2018). HotpotQA: A dataset for diverse, explainable multi-hop question answering. arXiv preprint arXiv:1809.09600.

Yu, W., Jiang, Z., Dong, Y., & Feng, J. (2020). ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning. In International Conference on Learning Representations.