SuperCLUE

State of Chinese AI 2025

—— The SuperCLUE Report

SuperCLUE Team

2026.02.04

SuperCLUE

Independent Third-Party AGI Evaluation Institution

Accurately Quantifying the Progress of AGI,

Defining the Roadmap for Humanity's Journey towards AGI.

Contents

SuperCLUE

PART I
Key Advances in 2025

1. SuperCLUE: The Panorama of Chinese LLMs in 2025
2. SuperCLUE: The Panorama of Chinese Agent Products in 2025
3. Key Advances in Large Models: 2025
4. Top 3 in All 2025 General Evaluations (China & Overseas)

SuperCLUE: The Panorama of Chinese LLMs in 2025

SuperCLUE

Text

General Open Weights

Qwen · Z.AI · Pangu LM · LongCat · Tencent Hunyuan · Intern · MiLM · DeepSeek · MINIMAX · KIMI · Ling · StepFun · ModelBest · ERNIE-4.5

General Proprietary

Doubao-Seed · Tencent Hunyuan · ERNIE · Qwen3-Max · 日日新 sensenova · ZTE · 360智脑 · SPARK

Reasoniong

Qwen3-Max-Thinking · DeepSeek-V3.2 · K2.5 · Doubao-Seed-1.8 · ERNIE 5.0 · GLM-4.7 · Tencent HY 2.0 Think

Multimodal

Visual Understanding

Doubao-vision · Qwen3-VL · ERNIE 5.0 · K2.5 · 日日新 sensenova · GLM-4.6V · Tencent Hunyuan · StepFun

Text-to-Image

Dreamina AI · KlingAI · Wan · ERNIE 5.0 · HunyuanImage-3.0 · GLM-Image · LongCat-Image · SPARK

Image Editting

Dreamina AI · Wan · GLM-Image · Qwen-Image-Edit · HunyuanImage-3.0 · StepFun · LongCat-Image

Text-to-Video

Wan · KlingAI · Dreamina AI · Hailuo AI · PixVerse · 清影 · Tencent Hunyuan · Vidu · SkyReels · LongCat-Video

Image-to-Video

KlingAI · Hailuo AI · Dreamina AI · Wan · PixVerse · Vidu · 清影 · StepFun

Real-time Interaction

Doubao-Seed · SPARK · Qwen · 日日新 sensenova · Hailuo AI · ERNIE · ChatGLM · Kimi

Text-to-Speech

Doubao Seed TTS 2.0 · SPARK · Qwen3-TTS · Fish Audio · Speech-2.6-HD · Baidu TTS · StepFun

Industry

Healthcare Industry

Baidu Lingyi · AQ · 讯飞晓医 AI·DOC · 百川智能 BAICHUAN AI

Automotive Industry

MindGPT · ZEEKR Kr LM · 易车大模型

Industrial Sector

奇智孔明AInno-15B · HUAWEI PanguLM · 羚羊工业大模型

Education Industry

MathGPT · 子曰 · 作业帮

Financial Industry

AntFinGLM · Miaoxiang Financial LM · 度小满 XuanYuan

Legal Industry

Chat Law · CHINESE LAW 元典智库 · 得理 DLS

SuperCLUE：The Panorama of Chinese Agent Products in 2025

SuperCLUE

General Domain

coze · MINIMAX · Skywork · ChatGLM · 实在智能 INTELLIGENCE INDEED · flowith
心响 · 纳米AI · AutoGLM · ERNIE Agent · Genie · 百度文库 · GenFlow 3.0
OK Computer · iFlow · OWL · OpenManus · MasterAgent · JENIUS

Vertical Domain

Deep Research

Deep Research · Deep Research · META SOTA · Quark
Qwen Deep Research · StepFun · iFlow · SciMaster

Search

META SOTA · 纳米AI搜索 · iFlow
TizzyAI · 博查 · 开搜AI

Travel

Fliggy · iMean.AI

Coding

CODEBUDDY · TRAE · Meituan CatPaw · JoyCode · 小浣熊家族 Raccoon
Qoder · 文心快码 Baidu Comate · 通义灵码 · CodeGeeX · 扣子编程

Desktop

QoderWork · StepFun · UI-TARS-desktop
MiniMax · LOONA Deskmate · Skywork Desktop

Legal Industry

通义法睿 · 法天使 LEGAL
GptLaw法律AI · MetaLaw

Marketing

讯飞AI营销 · AI麦可
领羊 · 悠易科技 YOYI TECH · 如此AI员工

Office

Lark · WPS AI · 小浣熊家族 Raccoon
DingTalk · midu 校对通

Financial Industry

问财 · FinGenius · 金灵AI
财跃 · QUTKE

Design

星流 · Jaaz · Almake
稿定AI · 站酷AI圈 · RoboNeo

Education

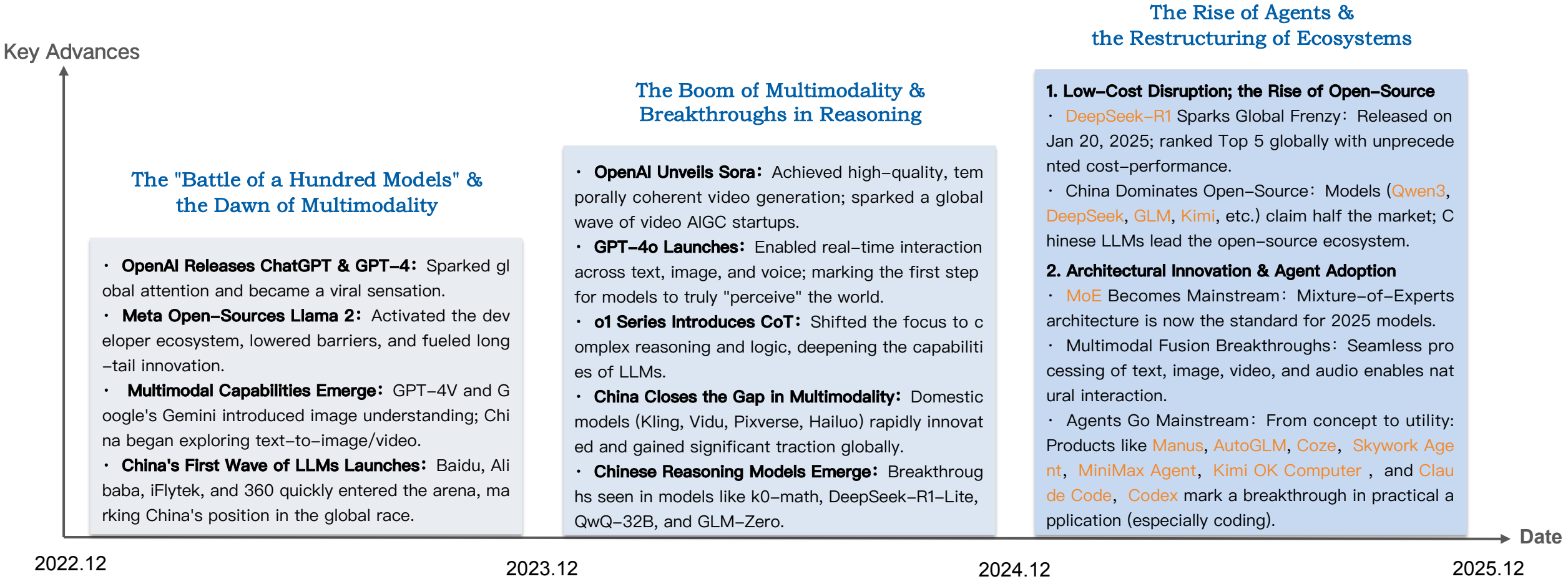
斑马 | 猿辅导在线教育
飞象老师 · 天学网·教师智能体

Key Advances in Large Models, 2025

SuperCLUE

Since the launch of ChatGPT on November 30, 2022, LLMs have triggered the largest wave of artificial intelligence in history. Over the past three years, AI institutions both domestically and internationally have achieved substantial breakthroughs, which can be specifically categorized into three phases: The "Battle of a Hundred Models" & the Dawn of Multimodality, The Boom of Multimodality & Breakthroughs in Reasoning, and The Rise of Agents & the Restructuring of Ecosystem.

SuperCLUE: Key Advances in LLMs for 2025



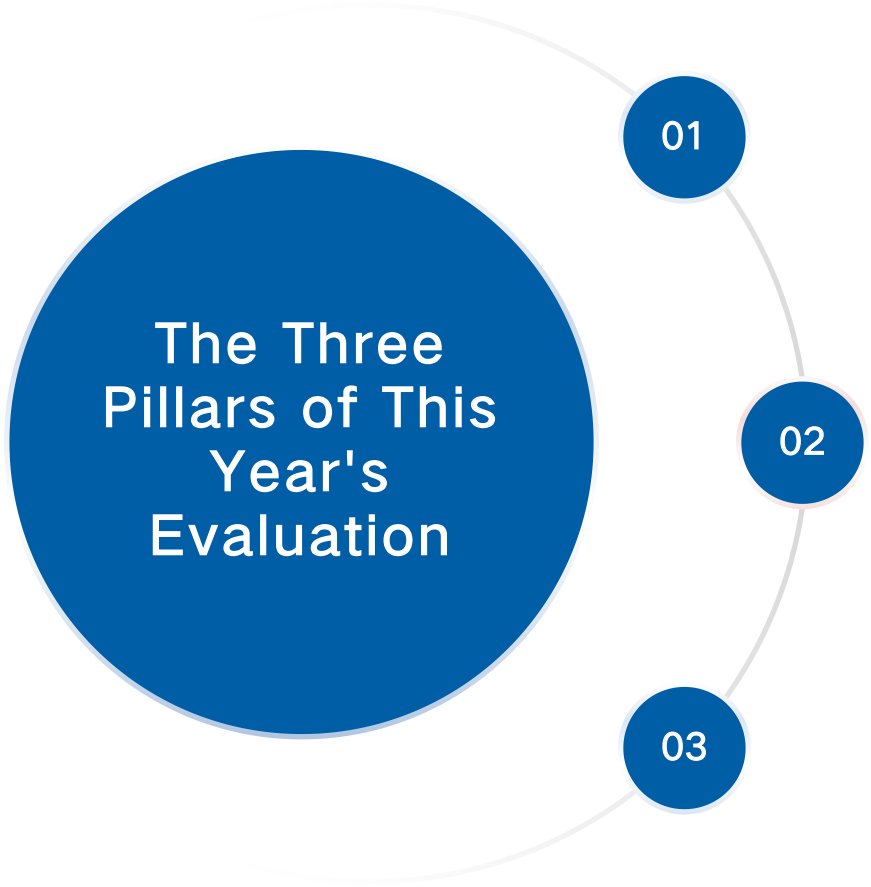
Top 3 in All 2025 General Evaluations (China & Overseas)

SuperCLUE

Date	No. 1 in China	No. 2 in China	No. 3 in China	Top 3 Overseas
January 2026	Kimi-K2.5-Thinking、Qwen3-Max-Thinking	Doubao-Seed-1.8-251228(Thinking)、DeepSeek-V3.2-Thinking	GLM-4.7、ERNIE-5.0	Claude-Opus-4.5-Reasoning、Gemini-3-Pro-Preview、GPT-5.2(high)
November 2025	DeepSeek-V3.2-Special	DeepSeek-V3.2-Thinking	ERNIE-5.0-Preview	GPT-5.2(high)、GPT-5.1(high)、Claude-Opus-4.5-Reasoning
September 2025	Kimi-K2-Thinking、DeepSeek-V3.2-Exp-Thinking	Doubao-Seed-1.6-thinking-250715、ERNIE-X1.1	Qwen3-Max、openPangu-Ultra-MoE-718B	GPT-5.1(high)、Gemini-3-Pro-Preview、GPT-5(high)
July 2025	DeepSeek-V3.1-Thinking	Doubao-Seed-1.6-thinking-250715	DeepSeek-R1-0528	GPT-5(high)、o3(high)、o4-mini(high)
May 2025	Doubao-1.5-thinking-pro-250415、SenseNova V6 Reasoner	DeepSeek-R1、NebularCoder-V6	Hunyuan-T1-20250403、DeepSeek-V3-0324	o4-mini(high)、Gemini 2.5 Pro Preview 05-06、Claude-Opus-4-Reasoning
March 2025	DeepSeek-R1	QwQ-32B	Doubao-1.5-pro-32k-250115	o3-mini(high)、Claude 3.7 Sonnet、GPT-4.5-Preview

SuperCLUE

Part II
2025 Annual Evaluation Analysis



1. Overseas proprietary models still occupy the top positions on the leaderboard.

In the 2025 Annual Chinese Large Model Benchmark, Anthropic's Claude-Opus-4.5-Reasoning ranked first with 68.25 points. Google's Gemini-3-Pro-Preview (65.59 points) and OpenAI's GPT-5.2(high) (64.32 points) followed. China's top open weights model, Kimi-K2.5-Thinking (61.50 points), and top proprietary model, Qwen3-Max-Thinking (60.61 points), placed 4th and 6th globally.

2. Chinese large models are accelerating their evolution from "playing catch-up" to "running neck and neck."

Since the release of DeepSeek-R1 in early 2025 – which matched OpenAI o1's performance and greatly narrowed the gap between Chinese and international models – Kimi-K2.5-Thinking and Qwen3-Max-Thinking have led the world in code generation and mathematical reasoning, respectively. An increasing number of Chinese large models are accelerating to catch up with the world's top ones, and even surpassing them in some fields.

3. Open-weight and proprietary models exhibit notable structural disparities between China and overseas.

The proprietary sector shows a pattern of "overseas leading, China catching up." Overseas models like Claude, Gemini, and GPT form the top tier. Chinese models such as Qwen3-Max-Thinking, Doubao-Seed-1.8-251228(Thinking), and ERNIE-5.0 trail behind but create effective competition. In contrast, the open-weight sector is characterized by "China dominance, overseas decline." Leading Chinese open-weight models including Kimi-K2.5-Thinking, DeepSeek-V3.2-Thinking, and GLM-4.7 rival top overseas proprietary models. Overseas open-weight efforts, like gpt-oss-120b and Mistral, lag significantly behind their Chinese models counterparts.

Introduction to the 2025 Chinese LLM Benchmark

CLUE (The Chinese Language Understanding Evaluation), launched in 2019, is a scientific, objective and neutral benchmark for language model evaluation. SuperCLUE is its evolution and continuation in the era of large models, focusing on comprehensive evaluation of general-purpose large models. The 2025 Annual Chinese Large Model Benchmark Evaluation focuses on general capability testing, consisting of 6 tasks with a total of 998 short-answer questions. Details of the test set are as follows:

SuperCLUE 2025 General Benchmark Dataset & Evaluation Method

1.Mathematical Reasoning

Introduction: Assesses the model's ability to perform multi-step reasoning and problem-solving using mathematical concepts and logic, covering competition-level datasets in geometry, algebra, probability, and statistics.

Evaluation Method: 0/1 scoring based on reference answers: 1 point for consistency with references, 0 otherwise. No evaluation of the reasoning process.

2.Scientific Reasoning

Introduction: Assesses the model's ability to understand and infer causality in interdisciplinary contexts, using graduate-level scientific datasets from physics, chemistry, biology, etc.

Evaluation Method: 0/1 scoring based on reference answers: 1 point for consistency with references, 0 otherwise. No evaluation of the reasoning process.

3.Code Generation

Introduction: The task has two types: (1) generating standalone functions covering data structures, algorithms, etc.; and (2) building complete interactive websites like travel booking, e-commerce, and social media platforms.

Evaluation Method: Scored 0/1 via unit tests (for standalone function generation) and functional tests simulating user interactions (for web app generation).

4.Agent (Task Planning)

Introduction: Assesses the model's ability to formulate structured action plans in complex scenarios—e.g., lifestyle services, work collaboration, learning, and healthcare—by generating logical, clear, and executable steps based on given goals and constraints.

Evaluation Method: A judge model discretely evaluates (0/1) completion of predefined checkpoints or continuously scores (0—100) overall plan quality.

5.Precise Instruction Following

Introduction: This assesses the model's ability to follow instructions—generating responses in specified formats and accurately presenting required data. Evaluated Chinese scenarios include structural, quantitative, semantic, and composite constraints (≥4 types).

Evaluation Method: Rule-based script 0/1 evaluation.

6.Hallucination Control

Introduction: Primarily evaluates the model's ability to mitigate hallucination while performing Chinese generation tasks, covering fundamental semantic understanding and generation benchmarks such as text summarization, reading comprehension, multi-document QA, and dialogue completion.

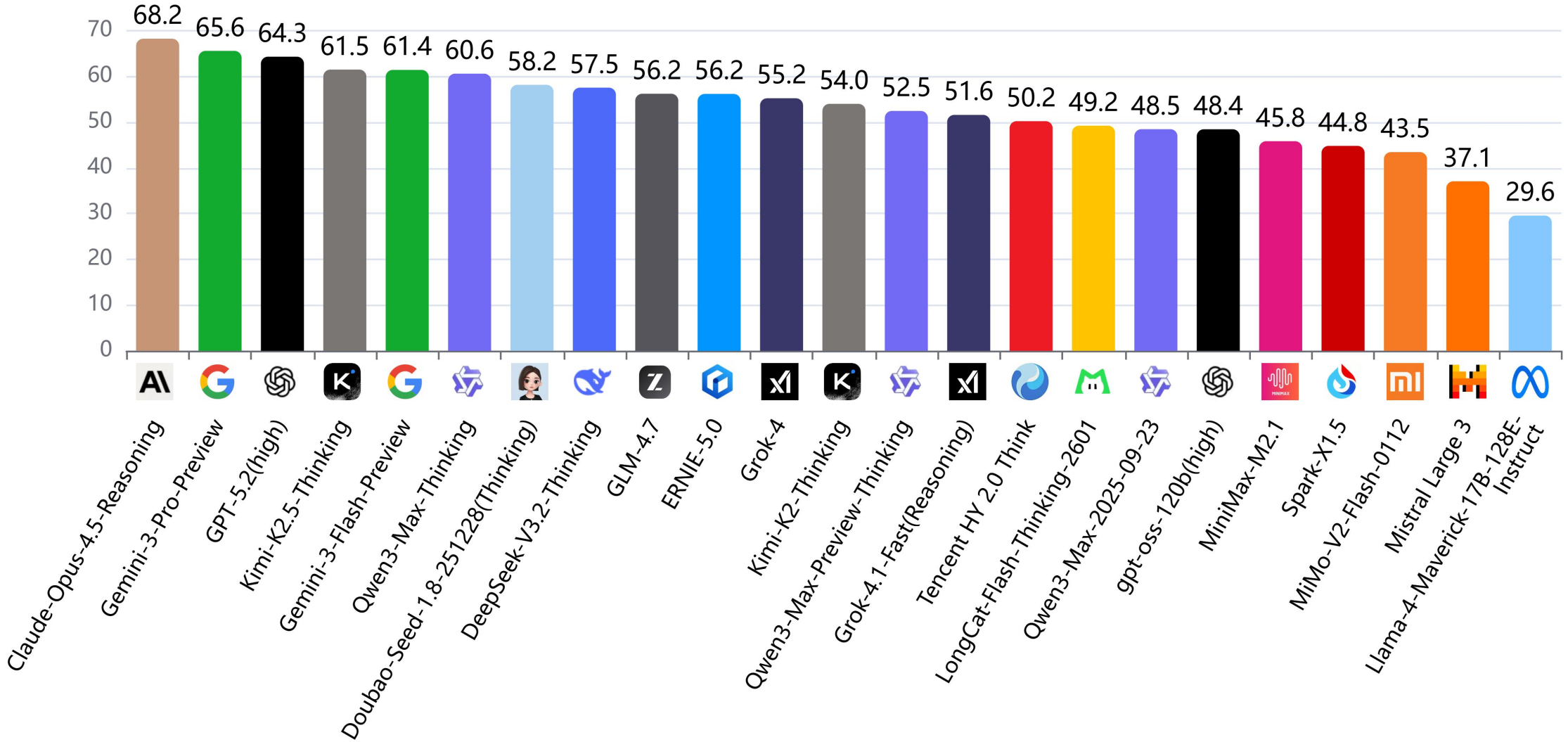
Evaluation Method: Binary (0/1) evaluation of hallucination per sentence, based on human-verified reference answers.

2025 Global LLM Chinese Intelligence Index Ranking

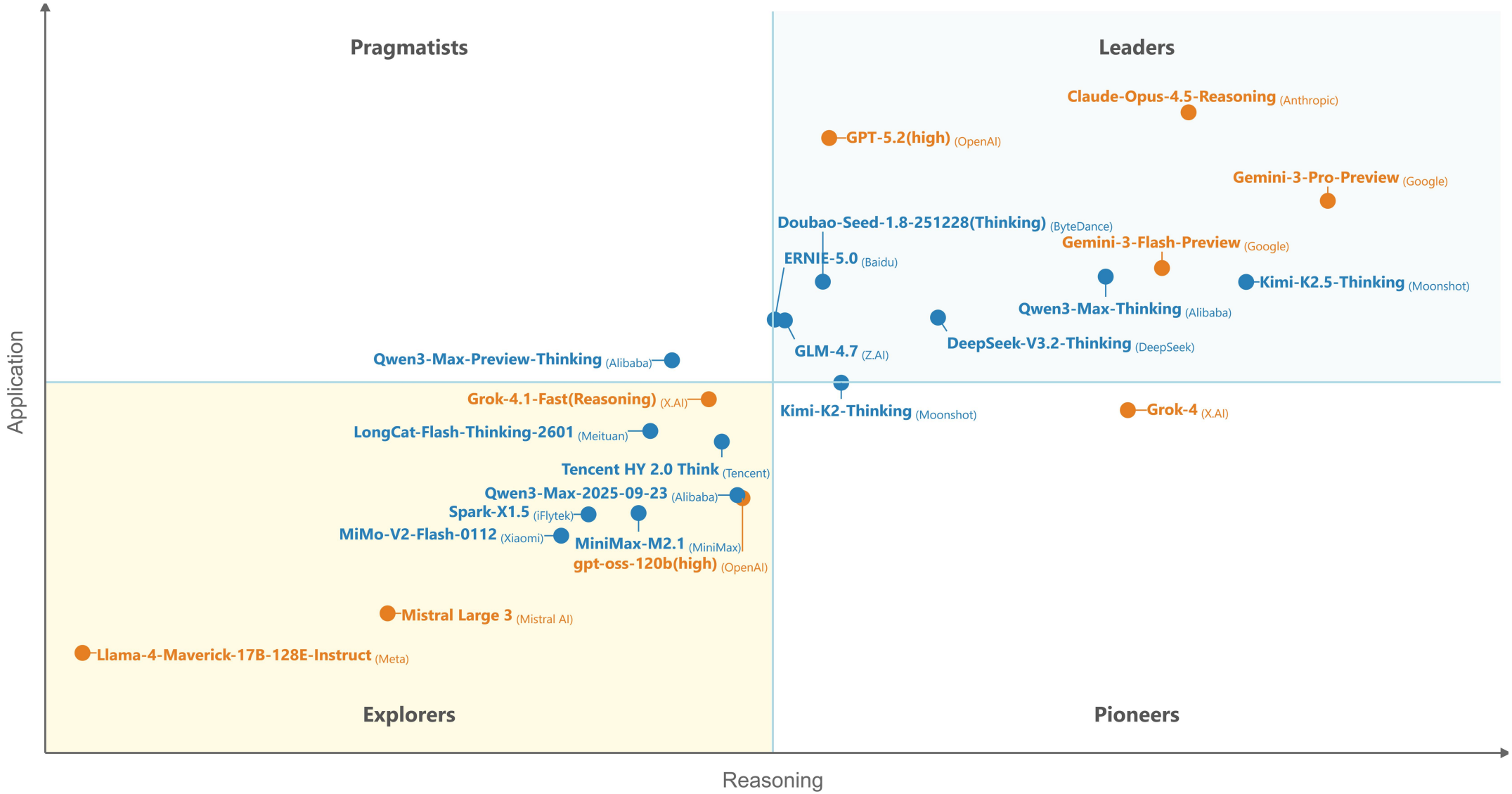
SuperCLUE

This evaluation comprises six tasks: mathematical reasoning, scientific reasoning, code generation (including web development), agent(task planning), hallucination control, and precise instruction following. It features **998 questions** and assesses **23 large language models** from China and abroad. Final scores are averaged across all tasks.

Official website: SuperCLUE.ai



SuperCLUE Model Quadrant 2025

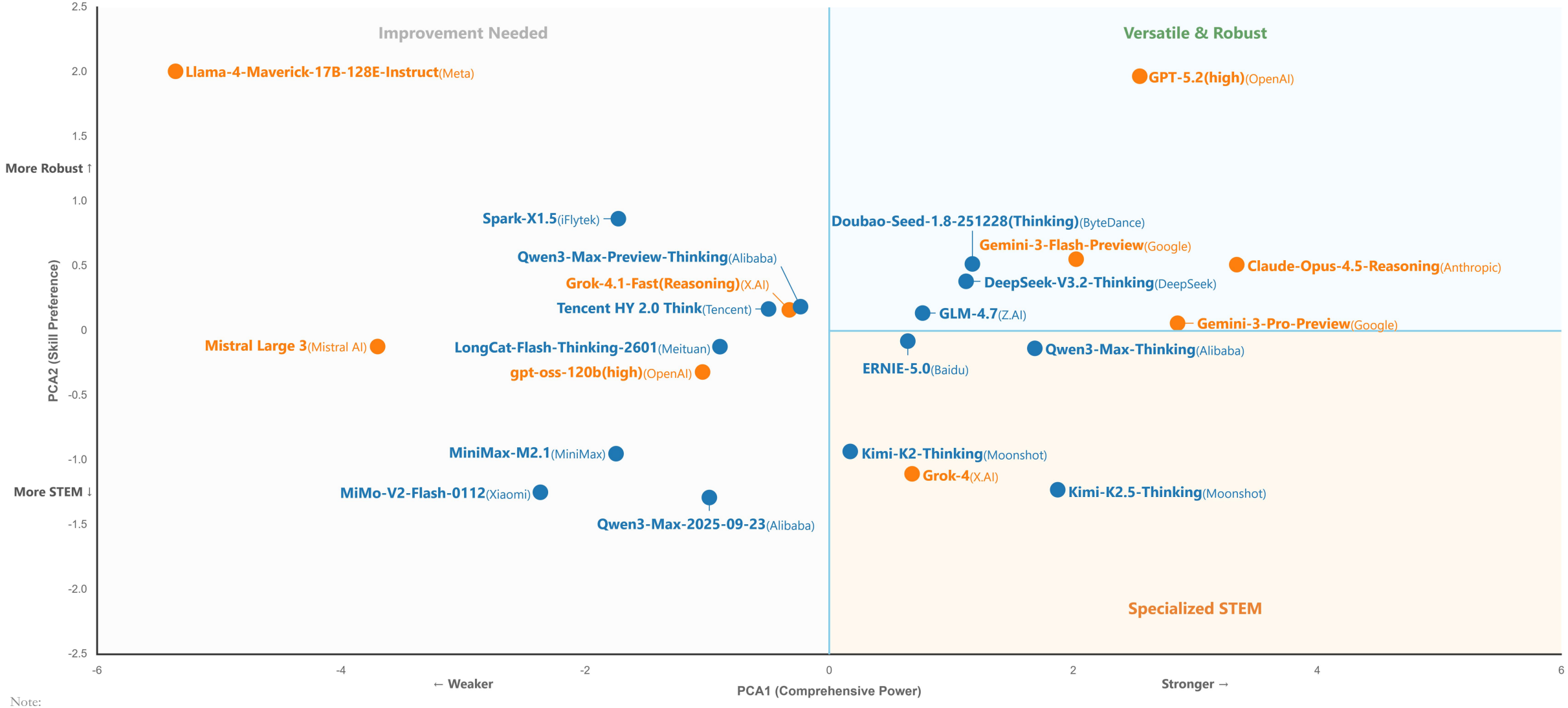


Source: SuperCLUE, January 29, 2026.

Note: 1. Composition of the two dimensions: Reasoning ability includes mathematical reasoning, scientific reasoning, and code generation; Application ability encompasses hallucination control, precise instruction following, and agent capabilities (task planning). 2. Meaning of the four quadrants: They represent different stages and positions of large models, where "Potential Explorer" indicates a model in the exploration phase with significant future potential; "Technology Leader" signifies leadership in foundational technologies; "Pragmatist" denotes superiority in depth of scenario-specific applications; and "Excellence Leader" represents a model that leads both in foundational capabilities and application scenarios, driving the advancement of domestic large models.

The 2025 SuperCLUE Model Capability Overview

SuperCLUE



Note:

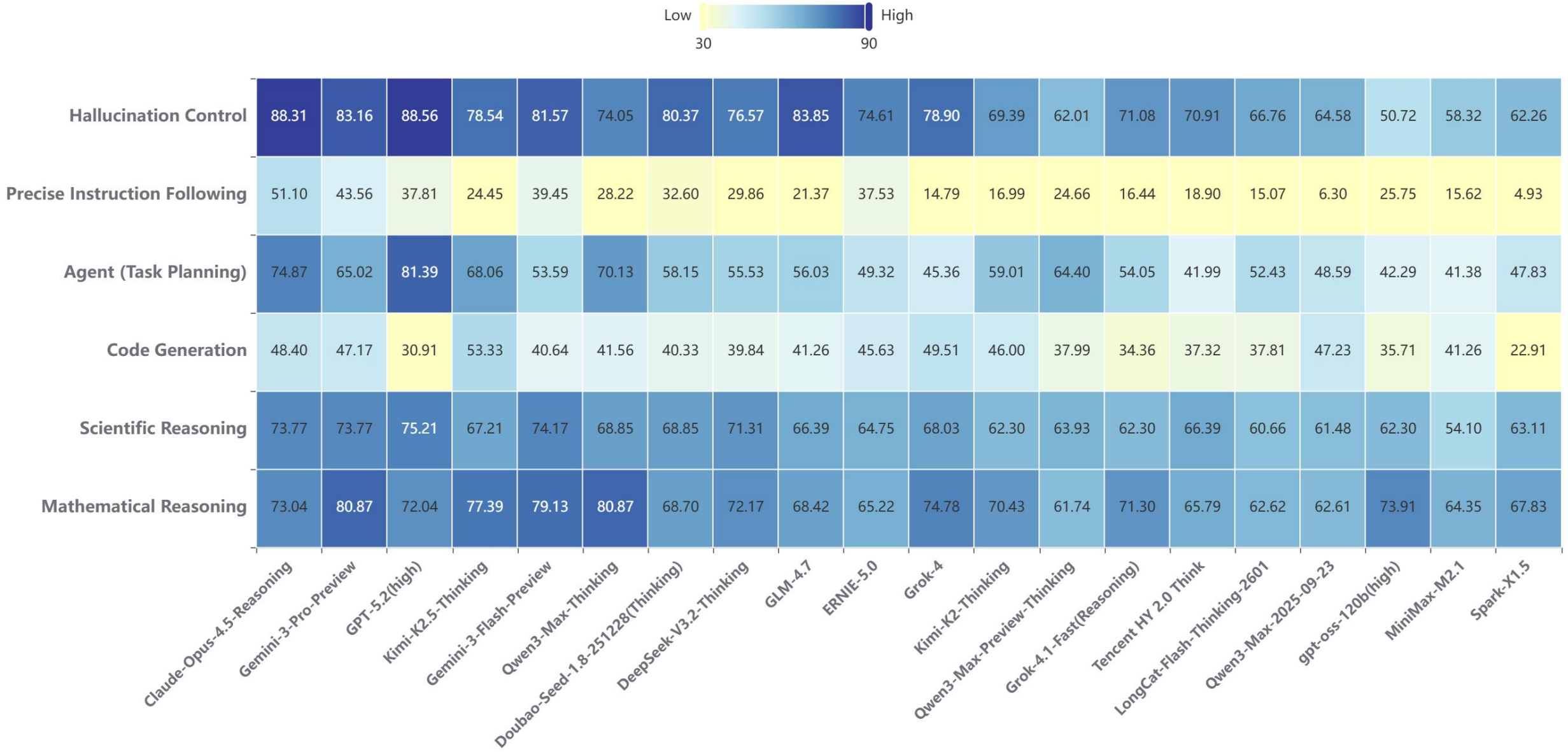
1. Source & Calculation: Based on SuperCLUE 2025 annual Chinese LLM benchmark data, generated using Principal Component Analysis (PCA). Data from six capability dimensions (Math Reasoning, Scientific Reasoning, Code Generation, Agent (Task Planning), Precise Instruction Following, Hallucination Control) are standardized (Z-Score) and projected onto a 2D plane.
2. Axis Meanings: **X-Axis (PCA1)** - Comprehensive Power (Absolute Strength): Represents the overall strength of the model. The further right, the higher the comprehensive score, meaning the model is stronger than those on the left in most dimensions. **Y-Axis (PCA2)** - Skill Preference (Relative Focus): Represents the structural difference in the model's skill tree, not absolute weaknesses. **Top:** Relative to its own level, performs better in robust tasks like "Agent-Task Planning/Hallucination Control". **Bottom:** Relative to its own level, performs better in STEM tasks like "Math Reasoning, Scientific Reasoning".
3. Region Interpretation: **Special Note**: Coordinates reflect relative preference. For example, a model in the top-right is robust-focused, but due to being on the right (strong overall), its STEM capabilities might still be stronger than a model in the bottom-left (weak overall but STEM-focused). **Versatile & Robust (Top-Right)**: Top-tier comprehensive ability, with a skill focus on long-chain planning and precise execution (Instruction Following, Agents, etc.). **Specialized STEM (Bottom-Right)**: Top-tier comprehensive ability, with a skill focus on deep thinking and logical computation (Math Reasoning, Code Generation, etc.). **Developing (Left)**: Overall capabilities have significant room for improvement.

SuperCLUE 2025: Top 3 Chinese Rankings by Dimension

SuperCLUE

Benchmark Tasks	No. 1 Overseas	No. 1 in China	No. 2 in China	No. 3 in China
Mathematical Reasoning	Gemini-3-Pro-Preview	Qwen3-Max-Thinking	Kimi-K2.5-Thinking	DeepSeek-V3.2-Thinking
Scientific Reasoning	GPT-5.2(high)	DeepSeek-V3.2-Thinking	Qwen3-Max-Thinking、Doubao-Seed-1.8-251228(Thinking)	Kimi-K2.5-Thinking、GLM-4.7、Tencent HY 2.0 Think
Code Generation	Grok-4	Kimi-K2.5-Thinking	Qwen3-Max-2025-09-23	Kimi-K2-Thinking、ERNIE-5.0
Agent (Task Planning)	GPT-5.2(high)	Qwen3-Max-Thinking	Kimi-K2.5-Thinking	Qwen3-Max-Preview-Thinking
Precise Instruction Following	Claude-Opus-4.5-Reasoning	ERNIE-5.0	Doubao-Seed-1.8-251228(Thinking)	DeepSeek-V3.2-Thinking
Hallucination Control	GPT-5.2(high)	GLM-4.7	Doubao-Seed-1.8-251228(Thinking)	Kimi-K2.5-Thinking

SuperCLUE 2025 Annual Evaluation: Top 20 Heatmap



Data Source：SuperCLUE, Jan 29, 2026.

2025 Annual Chinese LLM Benchmark – Overall Ranking

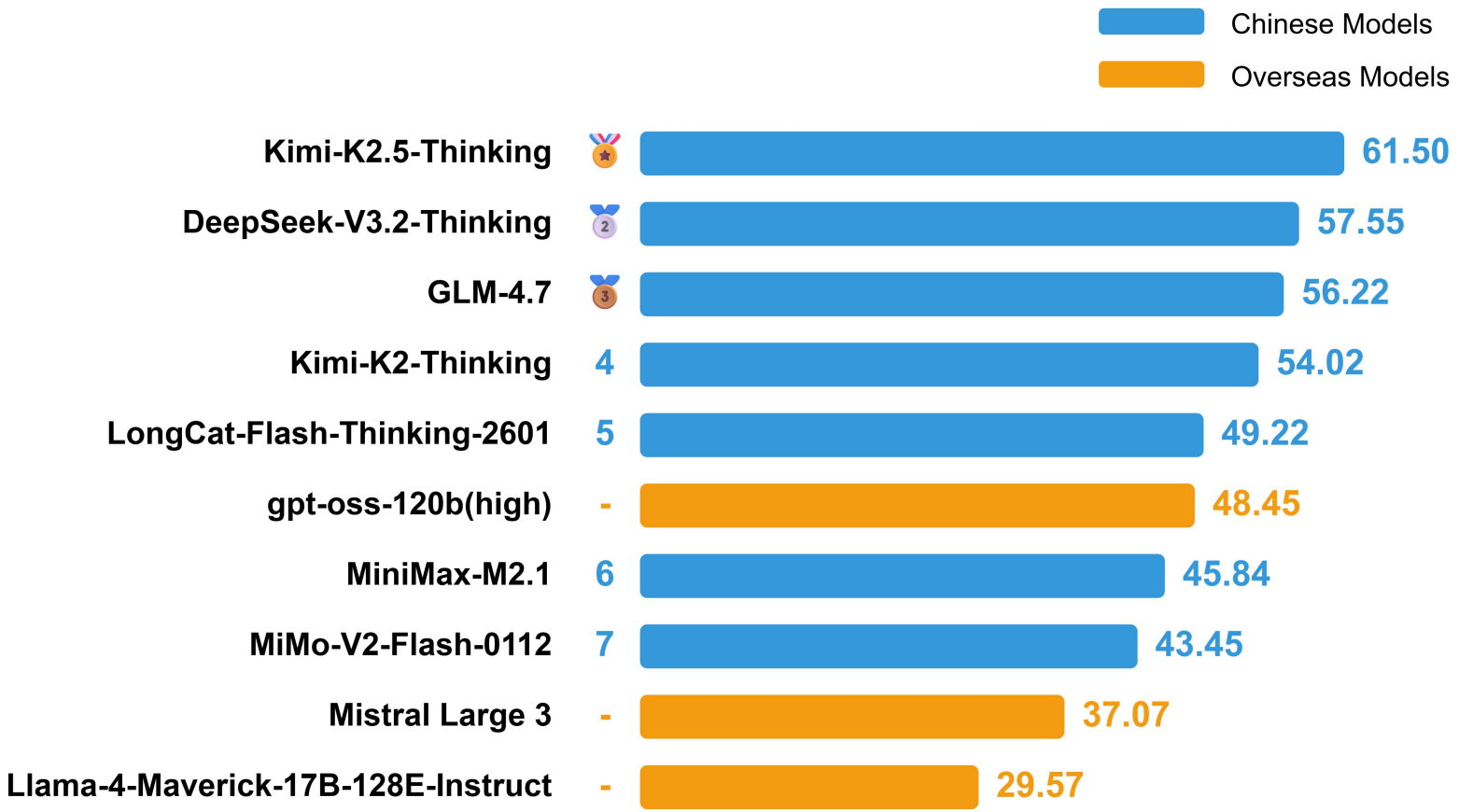
SuperCLUE

Overall Performance - SuperCLUE Benchmark 2025 Annual Report												
Rank	Model Name	Institution	License	Total Score	Mathematical Reasoning	Hallucination Control	Scientific Reasoning	Precise Instruction Following	Code Generation	Agent (Task Planning)	Region	Access Method
-	Claude-Opus-4.5-Reasoning	Anthropic	Proprietary	**68.25**	73.04	88.31	73.77	51.10	48.40	74.87	Overseas	API
-	Gemini-3-Pro-Preview	Google	Proprietary	**65.59**	80.87	83.16	73.77	43.56	47.17	65.02	Overseas	API
-	GPT-5.2(high)	OpenAI	Proprietary	**64.32**	72.04	88.56	75.21	37.81	30.91	81.39	Overseas	API
🥇	Kimi-K2.5-Thinking	Moonshot	Open Weights	**61.50**	**77.39**	**78.54**	67.21	24.45	**53.33**	**68.06**	China	API
-	Gemini-3-Flash-Preview	Google	Proprietary	**61.43**	79.13	81.57	74.17	39.45	40.64	53.59	Overseas	API
🥈	Qwen3-Max-Thinking	Alibaba	Proprietary	**60.61**	**80.87**	74.05	**68.85**	28.22	41.56	**70.13**	China	API
🥉	Doubao-Seed-1.8-251228(Thinking)	ByteDance	Proprietary	**58.17**	68.70	**80.37**	**68.85**	**32.60**	40.33	58.15	China	API
🥉	DeepSeek-V3.2-Thinking	DeepSeek	Open Weights	**57.55**	**72.17**	76.57	**71.31**	**29.86**	39.84	55.53	China	API
4	GLM-4.7	Z.AI	Open Weights	**56.22**	68.42	**83.85**	66.39	21.37	41.26	56.03	China	API
5	ERNIE-5.0	Baidu	Proprietary	**56.18**	65.22	74.61	64.75	**37.53**	45.63	49.32	China	API
-	Grok-4	X.AI	Proprietary	**55.23**	74.78	78.90	68.03	14.79	49.51	45.36	Overseas	API
4	Kimi-K2-Thinking	Moonshot	Open Weights	**54.02**	70.43	69.39	62.30	16.99	**46.00**	59.01	China	API
5	Qwen3-Max-Preview-Thinking	Alibaba	Proprietary	**52.46**	61.74	62.01	63.93	24.66	37.99	**64.40**	China	API
-	Grok-4.1-Fast(Reasoning)	X.AI	Proprietary	**51.59**	71.30	71.08	62.30	16.44	34.36	54.05	Overseas	API
6	Tencent HY 2.0 Think	Tencent	Proprietary	**50.22**	65.79	70.91	66.39	18.90	37.32	41.99	China	API
7	LongCat-Flash-Thinking-2601	Meituan	Open Weights	**49.22**	62.62	66.76	60.66	15.07	37.81	52.43	China	API
7	Qwen3-Max-2025-09-23	Alibaba	Proprietary	**48.46**	62.61	64.58	61.48	6.30	**47.23**	48.59	China	API
-	gpt-oss-120b(high)	OpenAI	Open Weights	**48.45**	73.91	50.72	62.30	25.75	35.71	42.29	Overseas	API
8	MiniMax-M2.1	MiniMax	Open Weights	**45.84**	64.35	58.32	54.10	15.62	41.26	41.38	China	API
9	Spark-X1.5	iFlytek	Proprietary	**44.81**	67.83	62.26	63.11	4.93	22.91	47.83	China	API
10	MiMo-V2-Flash-0112	Xiaomi	Open Weights	**43.45**	61.74	60.24	48.36	3.01	40.52	46.84	China	API
-	Mistral Large 3	Mistral AI	Open Weights	**37.07**	45.22	51.08	51.64	8.77	33.37	32.34	Overseas	API
-	Llama-4-Maverick-17B-128E-Instruct	Meta	Open Weights	**29.57**	32.17	57.30	48.36	5.48	13.85	20.29	Overseas	API

Data source: SuperCLUE, January 29, 2026.

Note: To minimize the impact of fluctuations, models with score differences within 1 point are considered tied in this evaluation. Overseas models are listed for reference only and are not ranked. Top three Chinese models are highlighted in red.

Open Weights Model Score Comparison



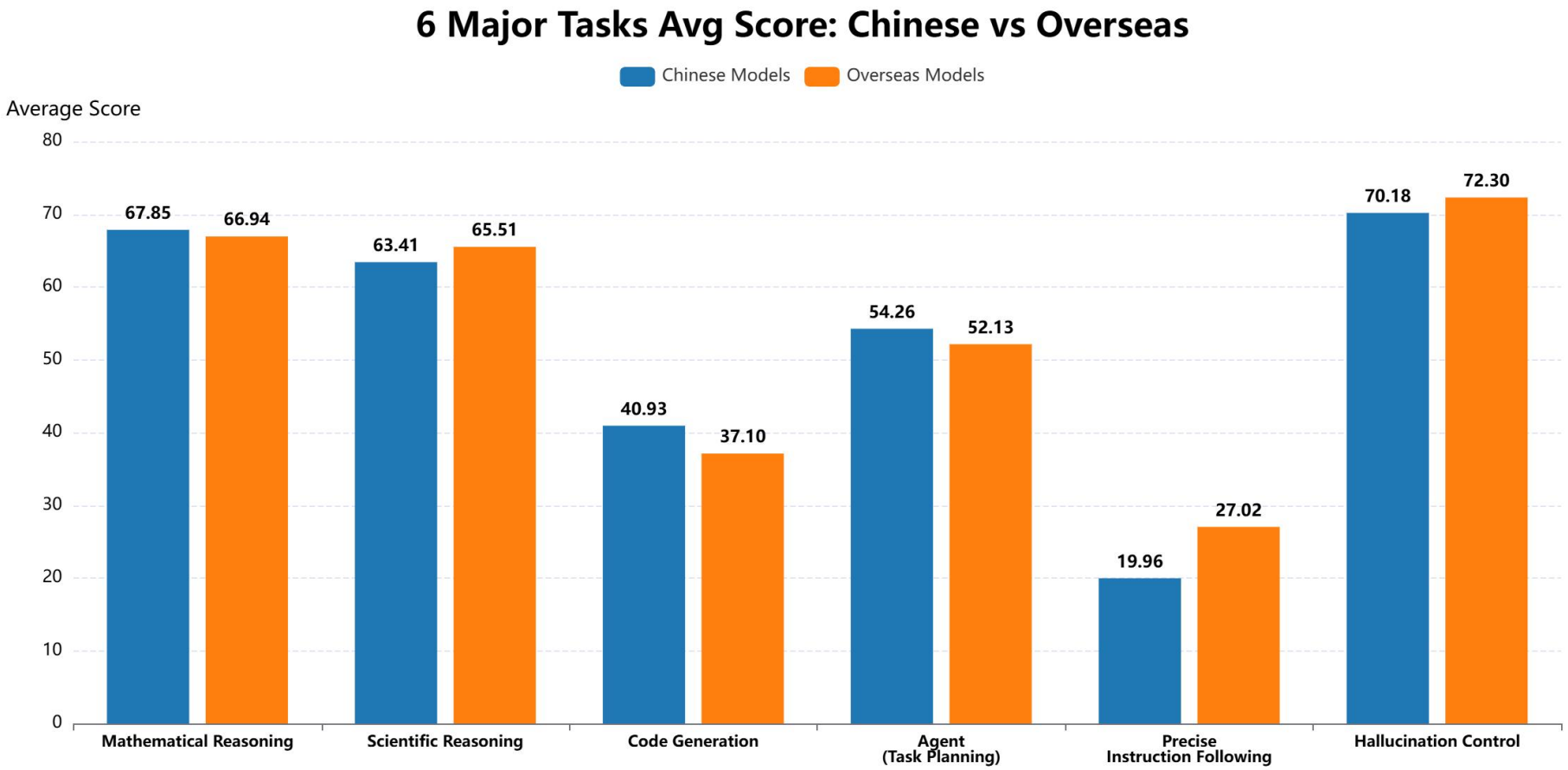
Data Source: SuperCLUE, Jan 29, 2026.

Note: Due to close scores of some models, models within a 1-point range are defined as tied to reduce the impact of fluctuations. Overseas models are for reference only and are not ranked.

Evaluation Analysis

Chinese open-weight models have comprehensively surpassed their overseas counterparts.

Chinese models occupy all top 5 spots on the open-weight leaderboard. Notably, Kimi-K2.5-Thinking claimed the No. 1 position with a score of 61.50, leading the second place by nearly 4 points. DeepSeek-V3.2-Thinking and GLM-4.7 secured the Top 3 ranking, significantly outperforming the best overseas open weights model, gpt-oss-120b(high).



Data Source: SuperCLUE, Jan 29, 2026.

Evaluation Analysis

1. Reasoning capabilities are highly aligned.

Chinese and overseas models show similar average scores in mathematical and scientific reasoning. Chinese models hold a slight edge in math, whereas overseas models (primarily top-tier) lead more notably in science.

2. Chinese models perform better in coding and agent tasks.

They outperform overseas models by over two points on average. In coding, Chinese models are generally mid-to-high tier, with the top models achieving the highest rankings. In agent tasks, overall performance is strong, though top Chinese models still lag behind top overseas models.

3. Precise instruction following and hallucination control remain weak spots.

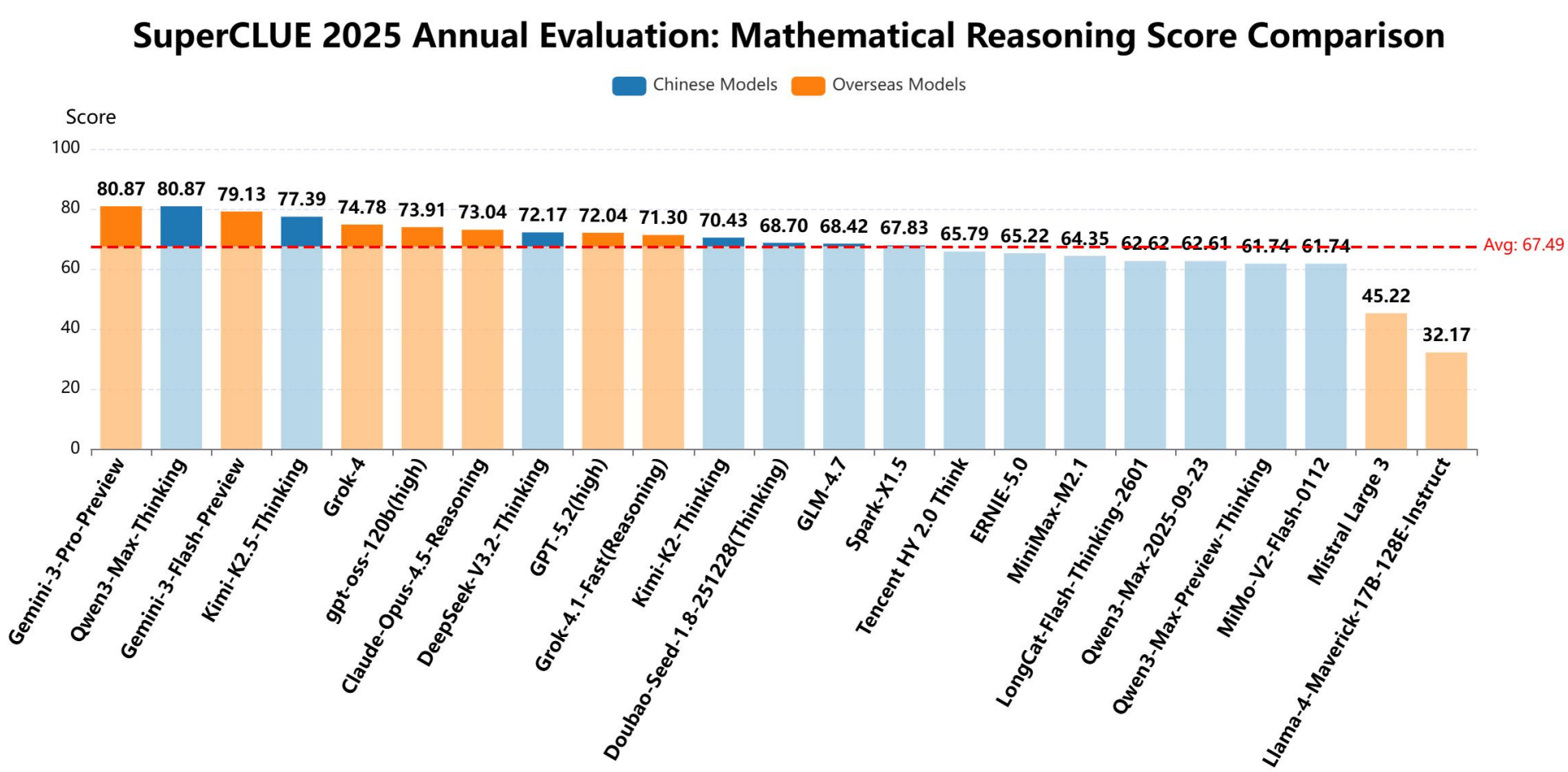
Chinese models underperform overseas models in these areas. The gap is largest in instruction following (over 7 points), followed by hallucination control (nearly 2 points).

Chinese vs. Overseas： Mathematical Reasoning

SuperCLUE

Introduction： Assesses the model's ability to perform multi-step reasoning and problem-solving using mathematical concepts and logic, covering competition-level datasets in geometry, algebra, probability, and statistics.

Evaluation Method： 0/1 scoring based on reference answers: 1 point for consistency with references, 0 otherwise. No evaluation of the reasoning process.



Data Source: SuperCLUE, Jan 29, 2026.

Evaluation Analysis

1. Top Chinese models achieve parity.

In the math reasoning task, Qwen3-Max-Thinking scored 80.87, tying with Gemini-3-Pro-Preview for the global top spot. Meanwhile, Kimi-K2.5-Thinking secured 4th place with 77.39, demonstrating a breakthrough in this domain.

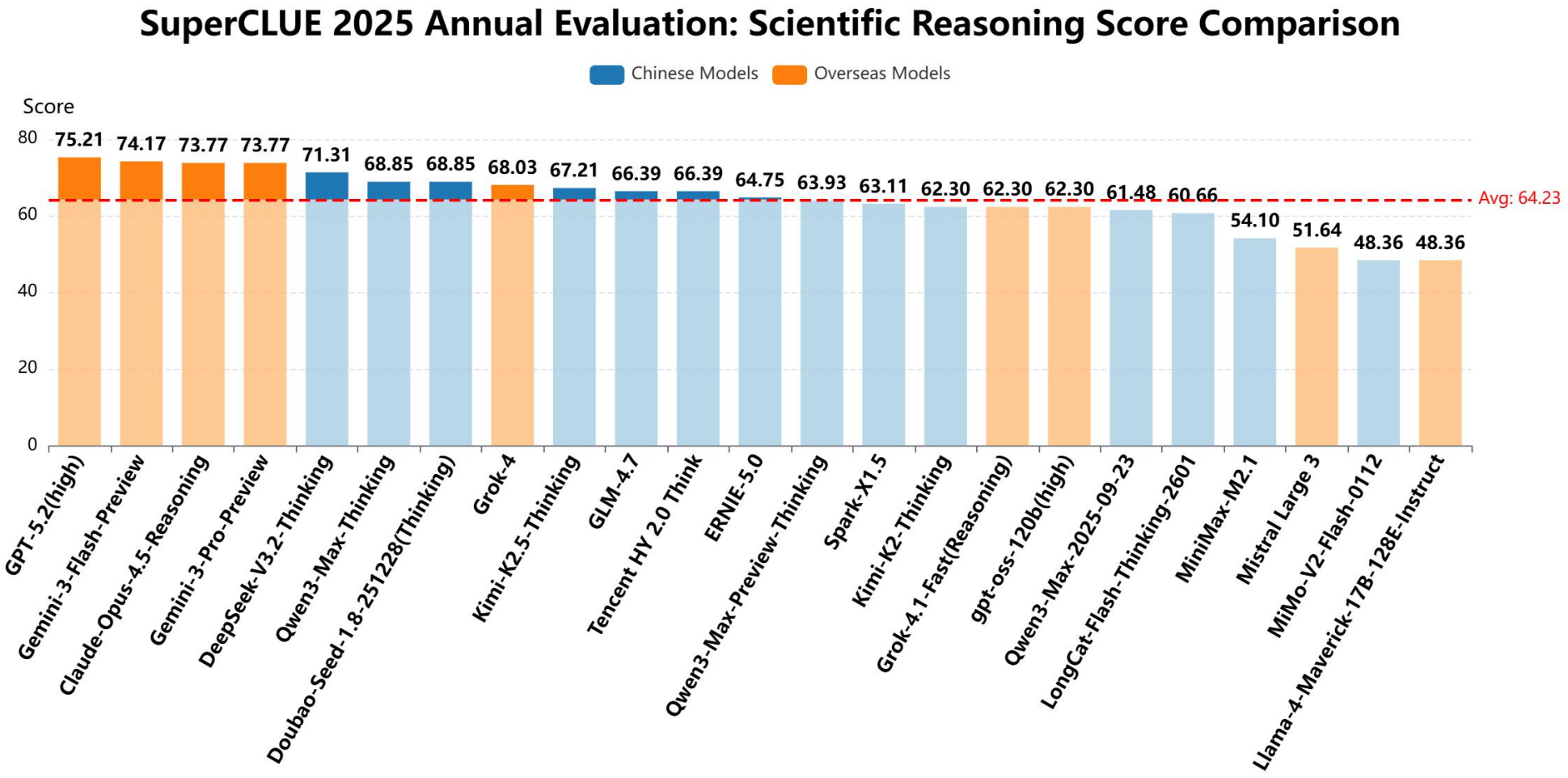
2. The overall Chinese cohort lags behind.

While top performers show promise, Chinese models are underrepresented in the Top 10 (4 out of 10). Most are clustered around the average (67.49) in the lower half, revealing a significant gap from international leaders.

Chinese vs. Overseas： Scientific Reasoning

Introduction： Assesses the model's ability to understand and infer causality in interdisciplinary contexts, using graduate–level scientific datasets from physics, chemistry, biology, etc.

Evaluation Method： 0/1 scoring based on reference answers: 1 point for consistency with references, 0 otherwise. No evaluation of the reasoning process.



Data Source: SuperCLUE, Jan 29, 2026.

Evaluation Analysis

1.Overseas models monopolize the top tier.

In scientific reasoning, overseas models swept the top four positions, led by GPT-5.2(high) (75.21). DeepSeek-V3.2-Thinking was the sole Chinese model to break into the top five, with Qwen3-Max-Thinking and Doubao-Seed-1.8-251228(Thinking) following closely behind.

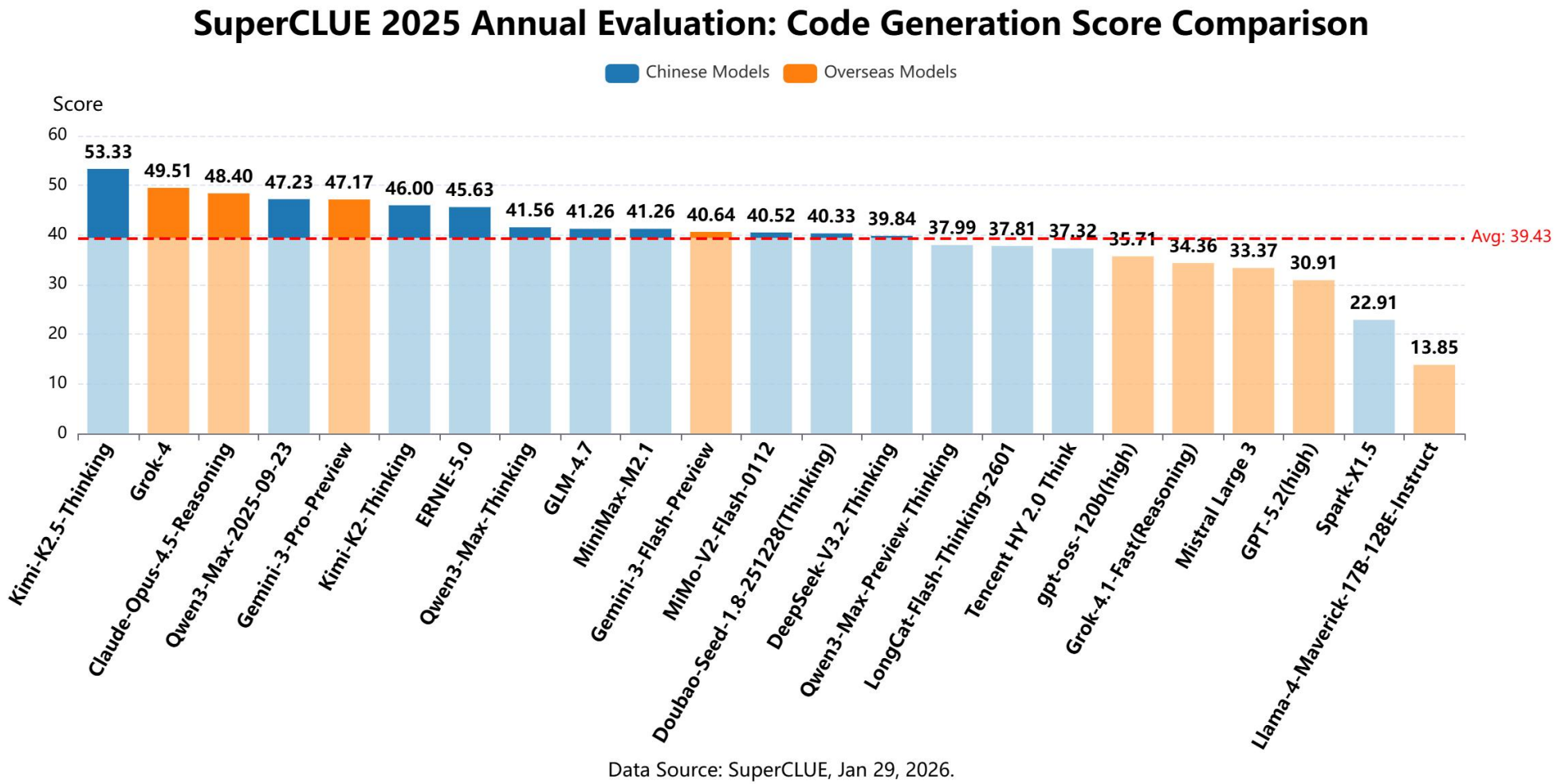
2. The Chinese models cohort shifts toward the mid-tier.

Compared to their performance in math reasoning, Chinese models show a clear shift toward the center in this domain. Most are now clustered around the average line, moving away from the lower ranks.

Chinese vs. Overseas： Code Generation

Introduction： The task has two types: (1) generating standalone functions covering data structures, algorithms, etc.; and (2) building complete interactive websites like travel booking, e-commerce, and social media platforms.

Evaluation Method： Scored 0/1 via unit tests (for standalone function generation) and functional tests simulating user interactions (for web app generation).



Data Source: SuperCLUE, Jan 29, 2026.

Evaluation Analysis

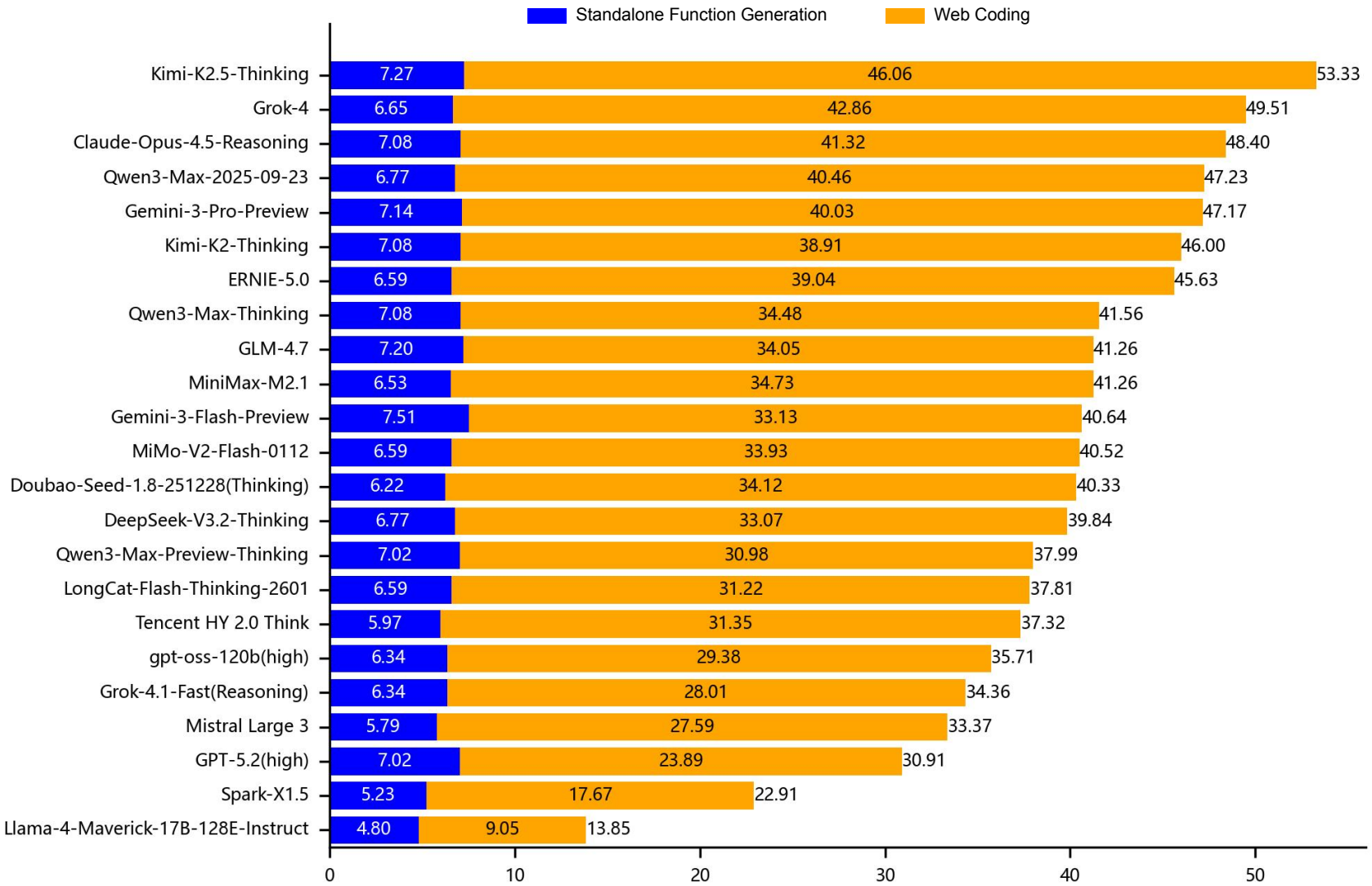
1. Chinese models shine.

The Chinese open-weight model Kimi-K2.5-Thinking topped the global ranking with 53.33, outperforming leading overseas models like Grok-4 and Claude-Opus-4.5-Reasoning. Qwen3-Max-2025-09-23 also entered the Top 5 with 47.23, indicating Chinese models leaders have achieved parity or a slight edge in code generation.

2. GPT-5.2(high)'s setback.

GPT-5.2(high) scored only 30.91, ranking third from bottom. This was due to our timeout mechanism: with a 30-minute limit per question (and two retries), the model incurred heavy penalties as many attempts timed out.

Chinese vs. Overseas: Code Generation



Data Source: SuperCLUE, January 29, 2026.

Notes:1.The final score for the Code Generation task is calculated by weighting the Standalone Function Generation task and the Web Coding task. The weight for each subtask is determined as follows: Weight = (Number of test cases in the subtask) / (Total number of test cases)

2.The number in the center of the blue bar represents the weighted score for the Standalone Function Generation task, while the number in the center of the orange bar represents the weighted score for the Web Coding task. The number on the right side of the bar indicates the total score, which is the sum of the two subtasks. (The subtask scores displayed in the chart are rounded to two decimal places, which may result in a cumulative rounding error when summed. However, the total score is calculated directly by dividing the total number of passed test cases by the total number of test cases, ensuring no cumulative error.)

Evaluation Analysis

1. Web Coding: The overtaking zone for Chinese models.

Kimi-K2.5-Thinking took the lead in the Web Coding subtask with a high score of 46.06, securing the top spot and leading the second place by 3.2 points. This was the key factor propelling its overall ranking, closely tied to its native multimodal architecture. Other leading Chinese models, such as Qwen3-Max-2025-09-23 and ERNIE-5.0, also showed a narrow gap (around 3 points) compared to top international models (Grok-4, Claude-Opus-4.5-Reasoning) in this subtask.

2. Web Coding serves as a key differentiator.

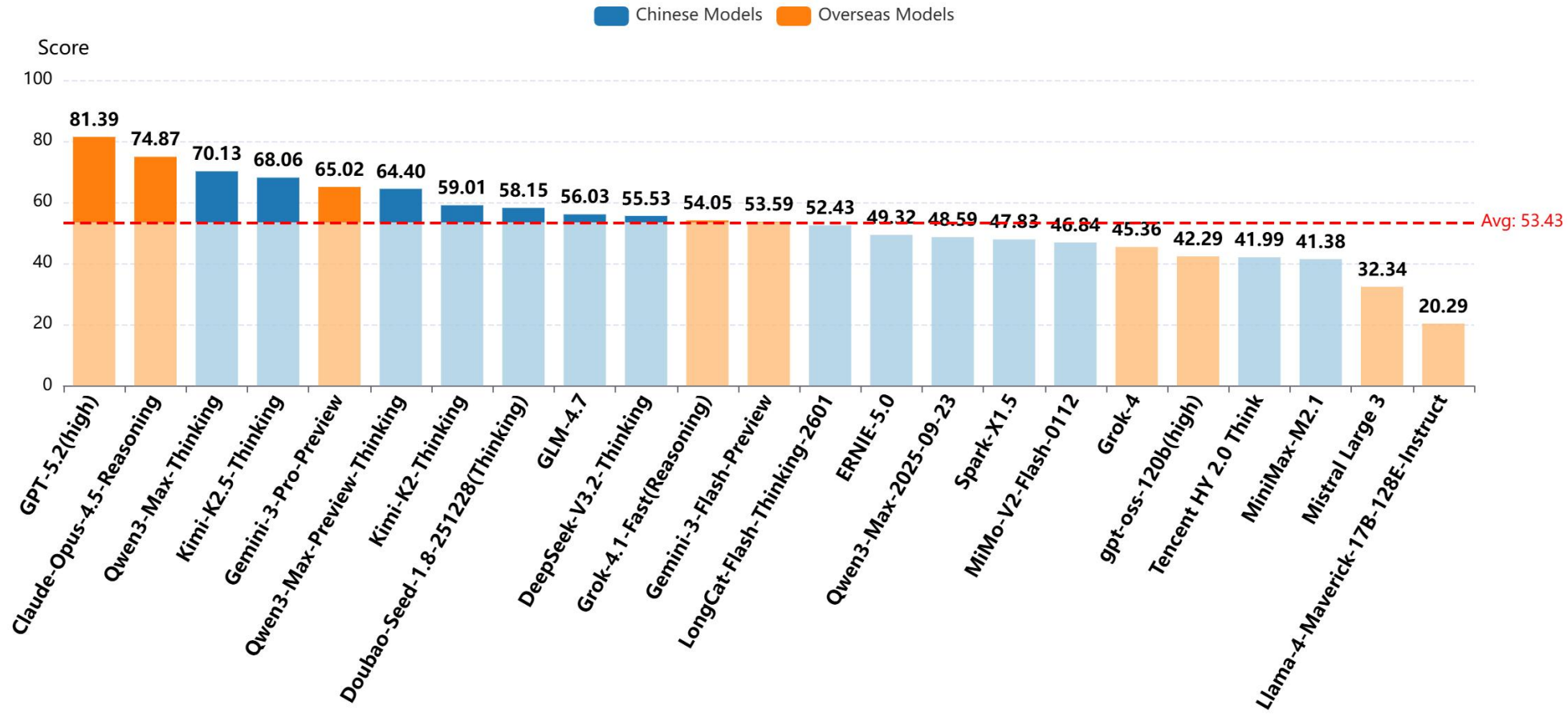
The performance gap among models in the standalone function generation subtask was insignificant (standard deviation: 0.66). In contrast, the standard deviation in Web Coding reached 8.23, highlighting its role as the primary driver of separation in overall code generation performance.

Chinese vs. Overseas： Agent(Task Planning)

Introduction： Assesses the model's ability to formulate structured action plans in complex scenarios—e.g., lifestyle services, work collaboration, learning, and healthcare—by generating logical, clear, and executable steps based on given goals and constraints.

Evaluation Method： A judge model discretely evaluates (0/1) completion of predefined checkpoints or continuously scores (0—100) overall plan quality.

SuperCLUE 2025 Annual Evaluation: Agent (Task Planning) Score Comparison



Data Source: SuperCLUE, Jan 29, 2026.

Evaluation Analysis

1. Overseas leaders dominate.

GPT-5.2(high) leads with 81.39, followed by Claude-Opus-4.5-Reasoning (74.87). Chinese models Qwen3-Max-Thinking (70.13) and Kimi-K2.5-Thinking (68.06) rank third and fourth. The >10 point gap highlights ample room for improvement in task planning.

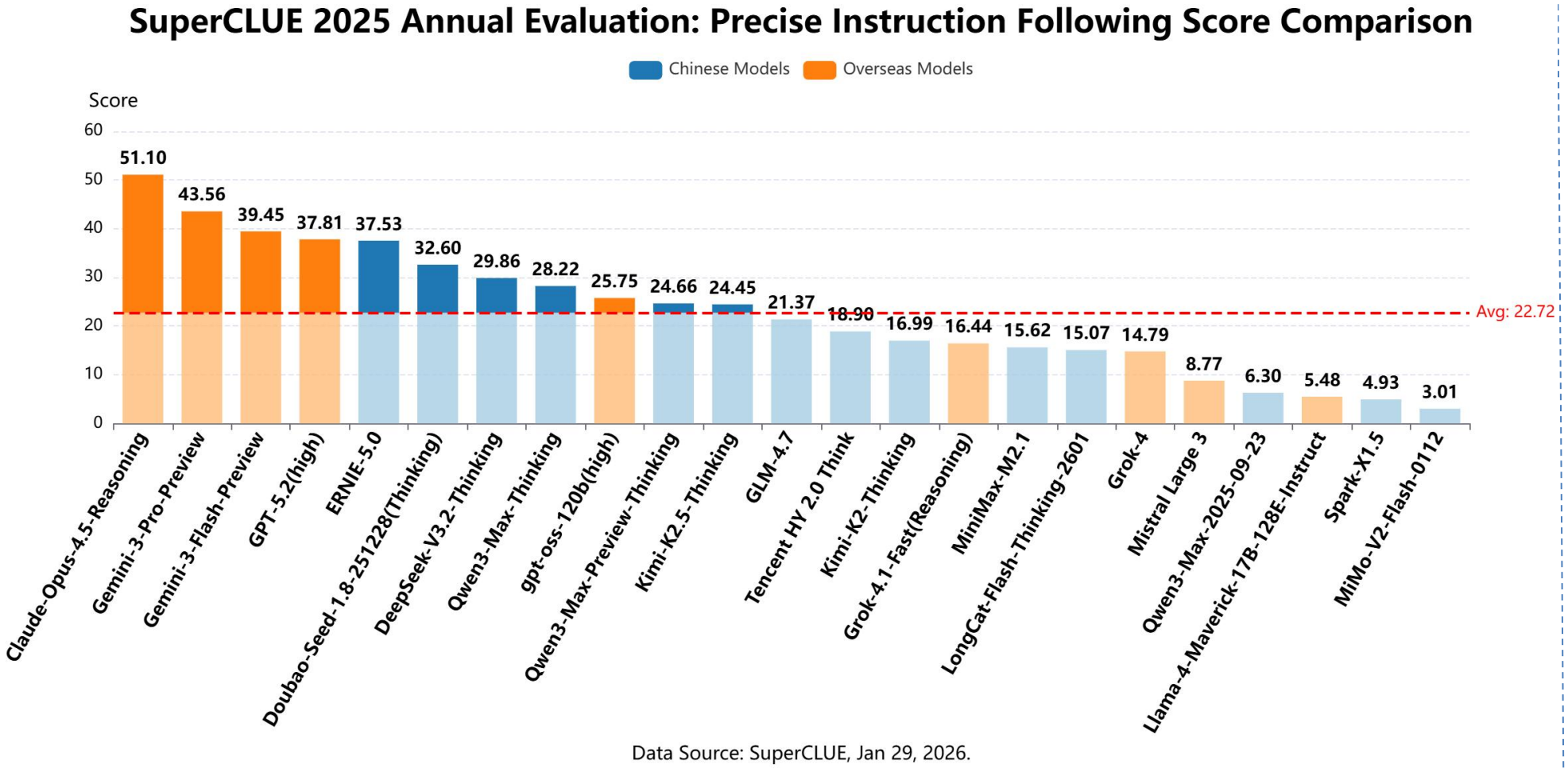
2. Wide gap and long tail.

The industry shows extreme polarization: the top score is 4x the bottom. With the highest standard deviation (13.78), this task remains a key differentiator and major challenge for LLMs.

Chinese vs. Overseas： Precise Instruction Following

Introduction： This assesses the model's ability to follow instructions—generating responses in specified formats and accurately presenting required data. Evaluated Chinese scenarios include structural, quantitative, semantic, and composite constraints (≥4 types).

Evaluation Method： Rule-based script 0/1 evaluation.



Data Source: SuperCLUE, Jan 29, 2026.

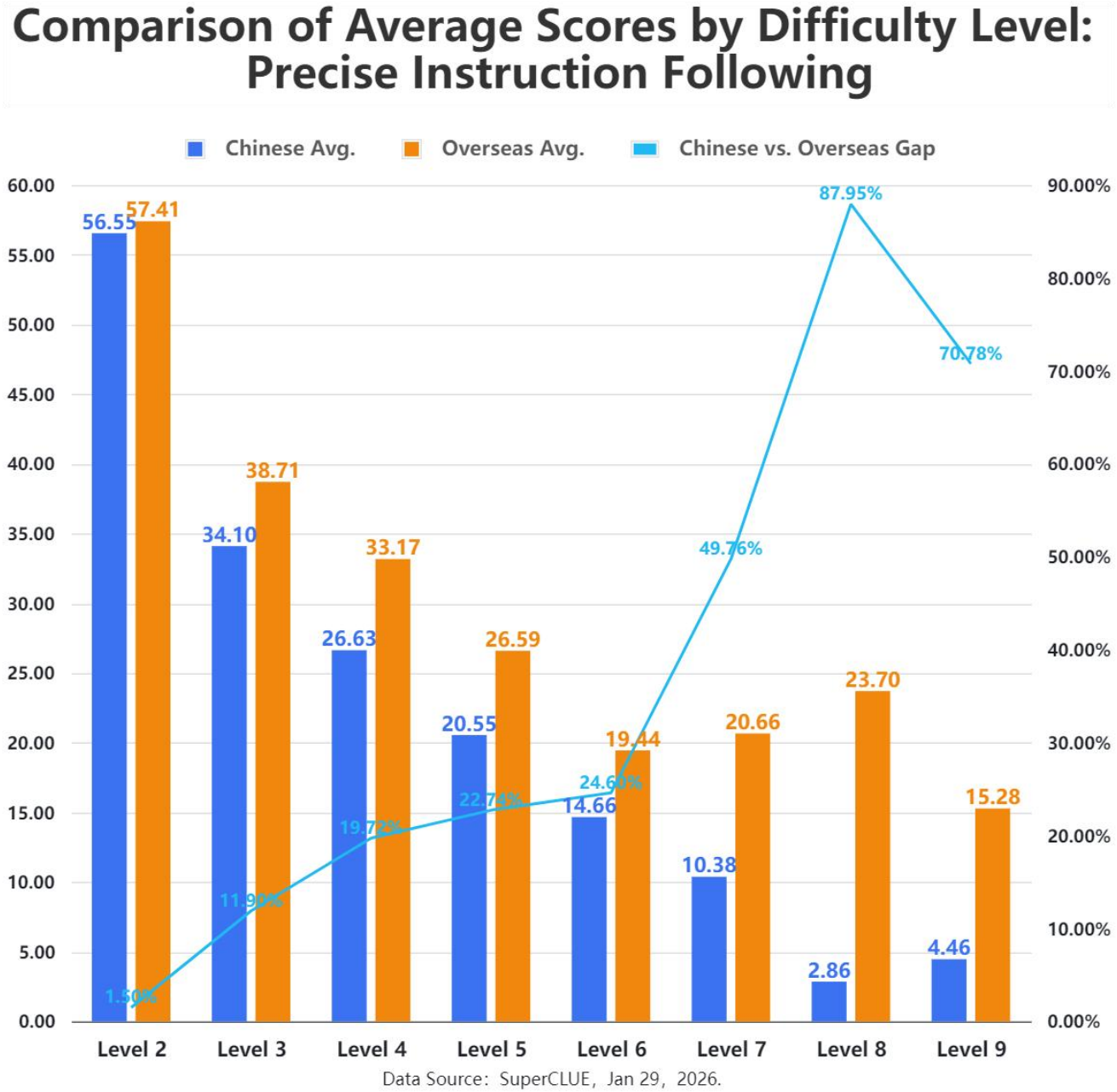
Evaluation Analysis

1. Overseas lead with a clear gradient.

The top 4 are all overseas models. Claude-Opus-4.5-Reasoning leads with 51.10, holding an 8-point lead over second place. Chinese models like ERNIE-5.0 trail by over 13 points. The Chinese models average (19.97) lags the overseas average (27.02) by 7 points, indicating room for growth.

2. Low overall scores and severe polarization.

This task spans 8 difficulty levels (from Level 2 to 9), presenting a high challenge. The overall average score was merely 22.72. More than half of the models scored below this average, highlighting the severity of the polarization.



Evaluation Analysis

1. Strong inverse correlation between difficulty and scores.

From Level 2 to Level 9, scores for both Chinese and overseas models show an exponential decline, dropping into single digits at higher difficulties (L7-L9). This indicates current models struggle to fully satisfy users when facing extremely complex tasks with multi-nested constraints.

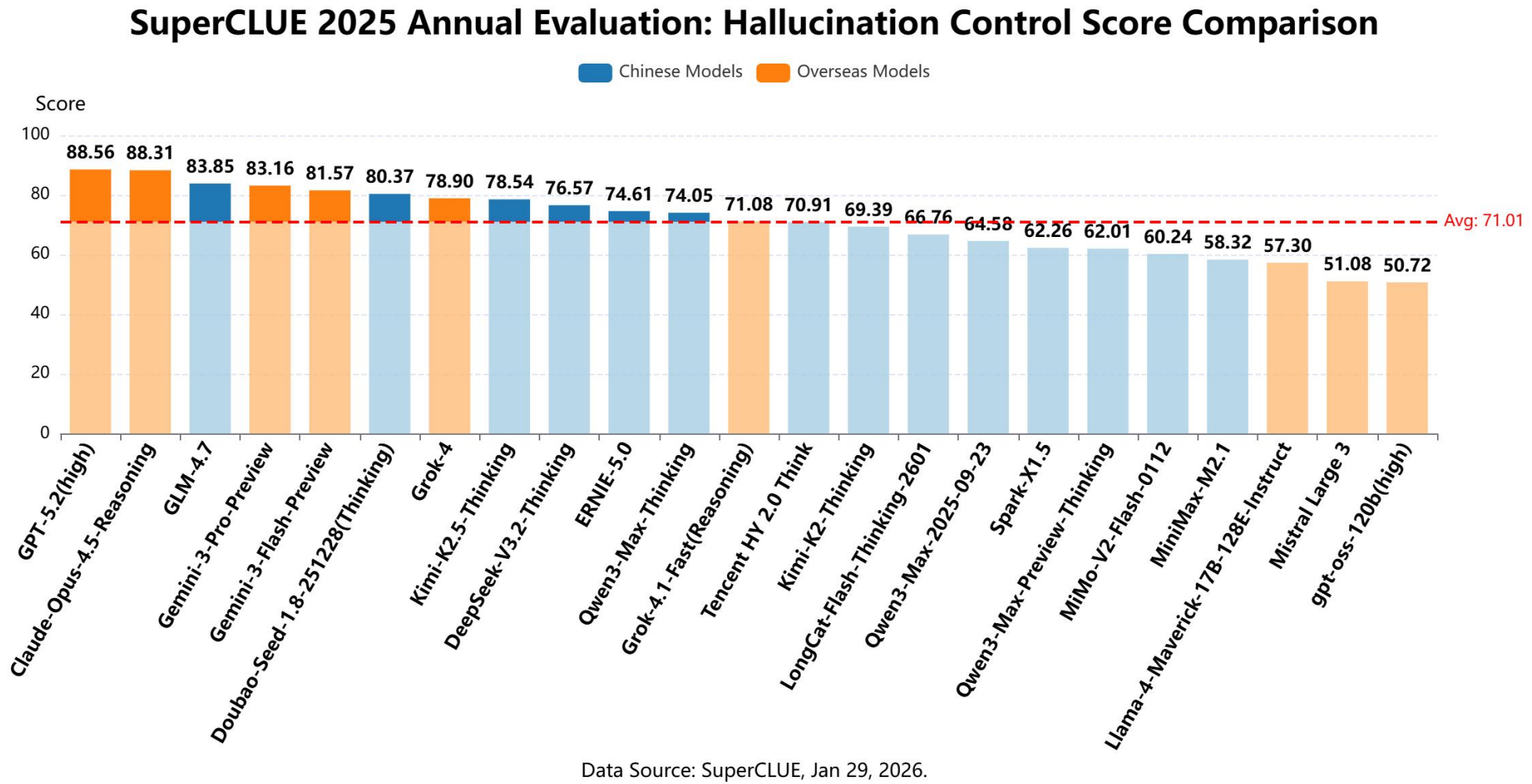
2. Overseas models demonstrate superior robustness.

Overall, the gap between Chinese and overseas models widens as instruction count increases. While the gap remains relatively stable (under 25%) at lower difficulties (L2-L6), it diverges significantly from Level 7 onwards. As complexity increases, overseas models show greater robustness—even scoring higher at L7 and L8—whereas Chinese models exhibit a near-total monotonic decrease.

Chinese vs. Overseas： Hallucination Control

Introduction： Primarily evaluates the model's ability to mitigate hallucination while performing Chinese generation tasks, covering fundamental semantic understanding and generation benchmarks such as text summarization, reading comprehension, multi-document QA, and dialogue completion.

Evaluation Method： Binary (0/1) evaluation of hallucination per sentence, based on human-verified reference answers.



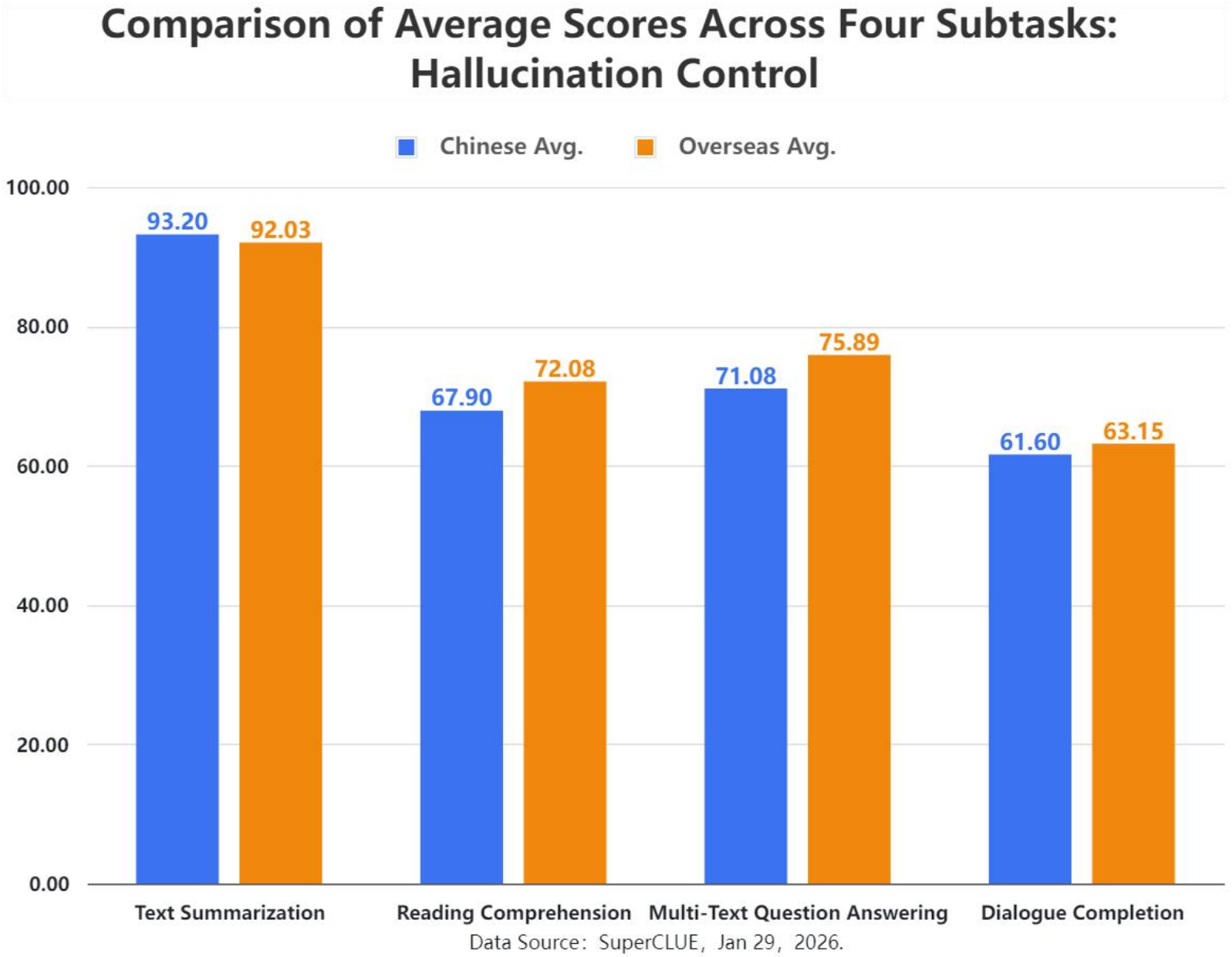
Data Source: SuperCLUE, Jan 29, 2026.

Evaluation Analysis

Overseas leaders hold a dominant edge, while Chinese leaders show breakthroughs.

GPT-5.2(high) (88.56) and Claude-Opus-4.5-Reasoning (88.31) lead the pack, outperforming the average by over 17 points and demonstrating the first-tier overseas models' dominance in hallucination control . Notably, GLM-4.7 broke into the Top 3 with 83.85 points, narrowing the gap with the overseas leaders to within 5 points. Additionally, Doubao-Seed-1.8-251228(Thinking) delivered a strong performance exceeding 80 points, surpassing Grok-4.

Chinese vs. Overseas：Hallucination Control



Evaluation Analysis

As tasks evolve from information integration to open-ended generation, both Chinese and overseas large models show a distinct decline in hallucination control.

Text Summarization is the easiest task (yielding the highest scores) because it strictly relies on the source text for compression and paraphrasing, leaving little room for creation.

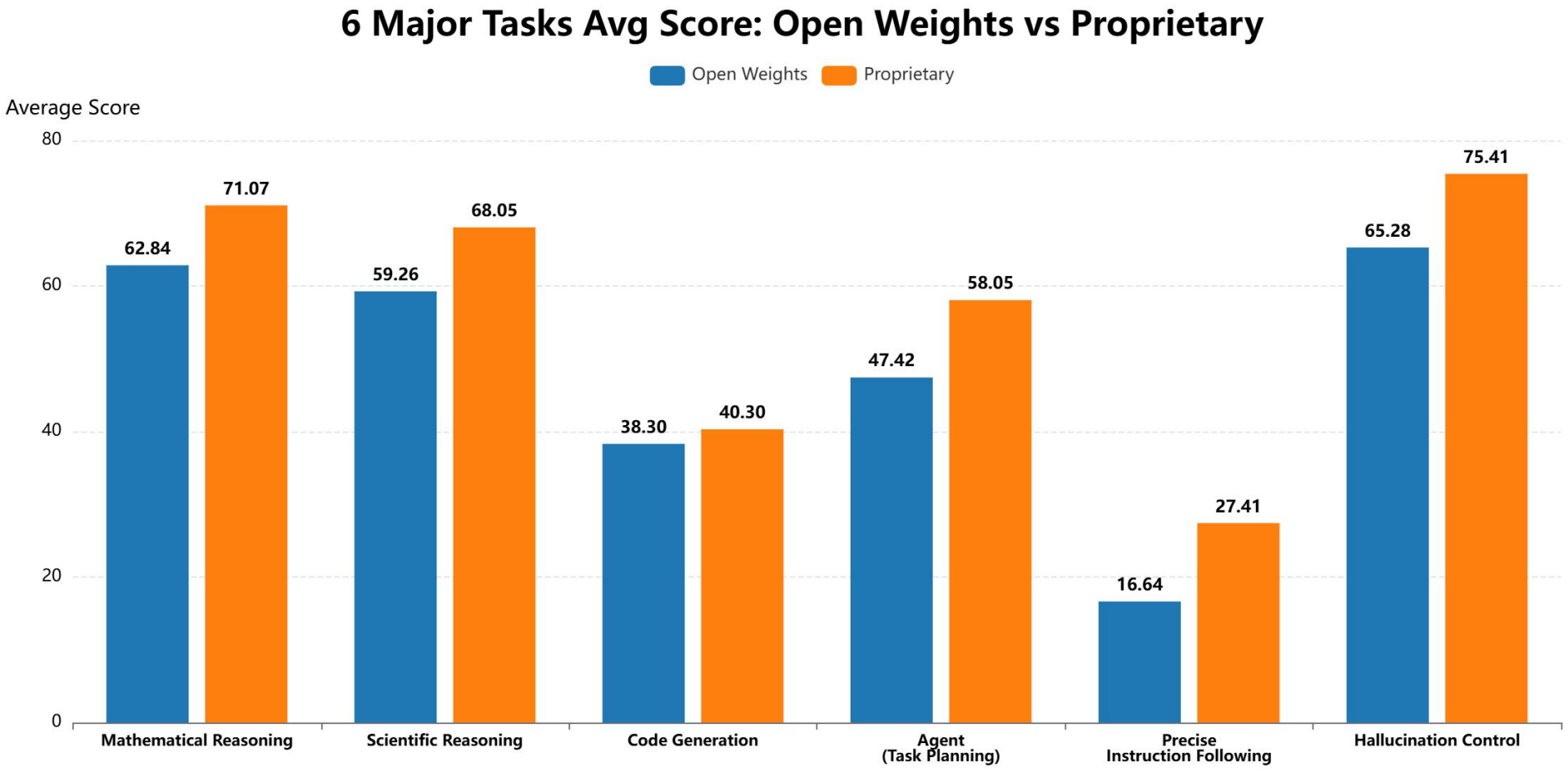
Reading Comprehension introduces more risk, as it demands reasoning and judgment rather than simple repetition, creating space for hallucinations.

Multi-Text Question Answering is more challenging; integrating and comparing multiple sources increases the likelihood of information confusion and misattribution.

Finally, Dialogue Completion is the most prone to hallucinations. Its open-ended nature requires the model to invent information to ensure fluency, inevitably leading to factual errors and confabulation .

In summary, the more open-ended and creative the task, the higher the risk of hallucination.

Open Weights vs. Proprietary



Data Source: SuperCLUE, Jan 29, 2026.

Evaluation Analysis

1. Proprietary models lead across the board.

They outperform open-weight models in all six tasks. Despite rapid open weights progress, proprietary models maintain a clear edge in top-tier performance, showing a double-digit advantage in Agent, Precise Instructions, and Hallucination Control.

2. Open-weight models narrows the gap in reasoning and breaks through in coding.

Open-weight models continue catching up in Math and Science. In Code Generation, the gap is minimal (~2 points), likely due to focused optimization in this domain.

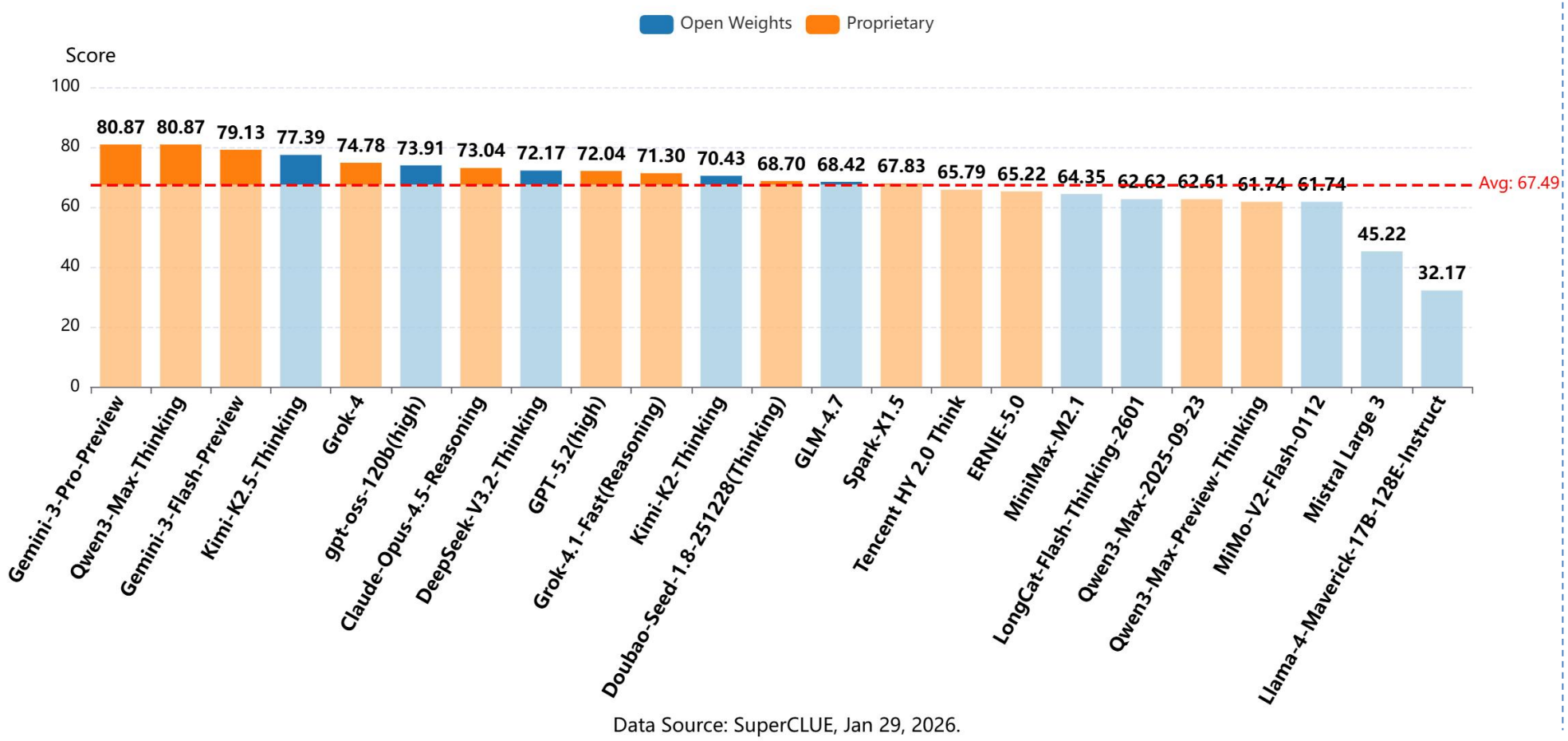
Open Weights vs. Proprietary：Mathematical Reasoning

SuperCLUE

Introduction： Assesses the model's ability to perform multi-step reasoning and problem-solving using mathematical concepts and logic, covering competition-level datasets in geometry, algebra, probability, and statistics.

Evaluation Method： 0/1 scoring based on reference answers: 1 point for consistency with references, 0 otherwise. No evaluation of the reasoning process.

SuperCLUE 2025 Annual Evaluation: Mathematical Reasoning Score Comparison



Data Source: SuperCLUE, Jan 29, 2026.

Evaluation Analysis

1. Proprietary models still hold the lead.

Proprietary models take the top 3 spots in math reasoning, with only 3 open-weight models in the top 10. The average gap is nearly 8 points (71.07 vs. 62.84).

2. Top-tier open-weight models show strong momentum, nearing the leaders.

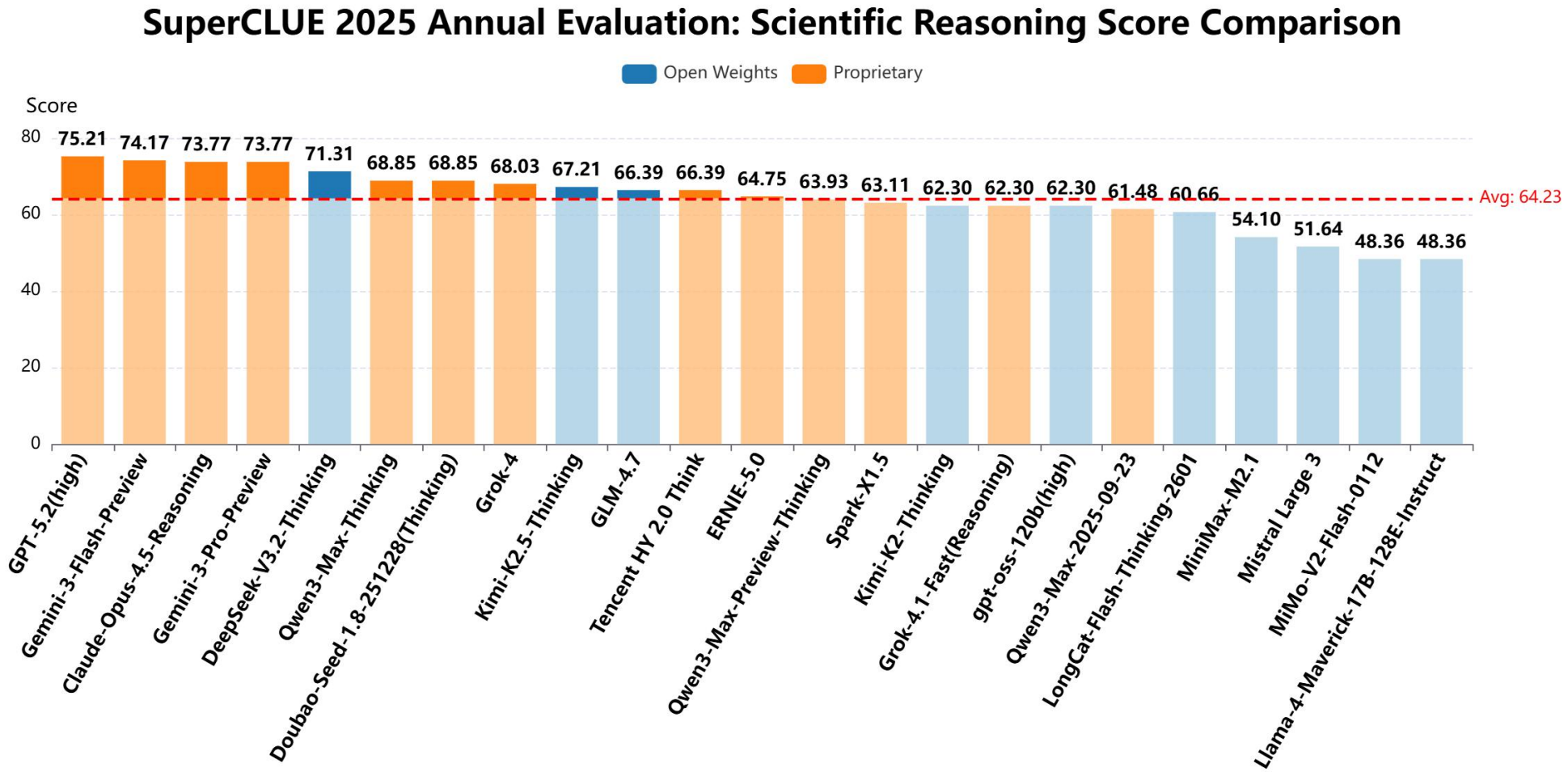
The Chinese open-weight model Kimi-K2.5-Thinking ranked fourth globally with a score of 77.39, surpassing top-tier proprietary models including Grok-4, Claude-Opus-4.5-Reasoning, and GPT-5.2 (high).

Open Weights vs. Proprietary：Scientific Reasoning

SuperCLUE

Introduction： Assesses the model's ability to understand and infer causality in interdisciplinary contexts, using graduate-level scientific datasets from physics, chemistry, biology, etc.

Evaluation Method： 0/1 scoring based on reference answers: 1 point for consistency with references, 0 otherwise. No evaluation of the reasoning process.



Data Source: SuperCLUE, Jan 29, 2026.

Evaluation Analysis

Proprietary models hold a significant advantage.

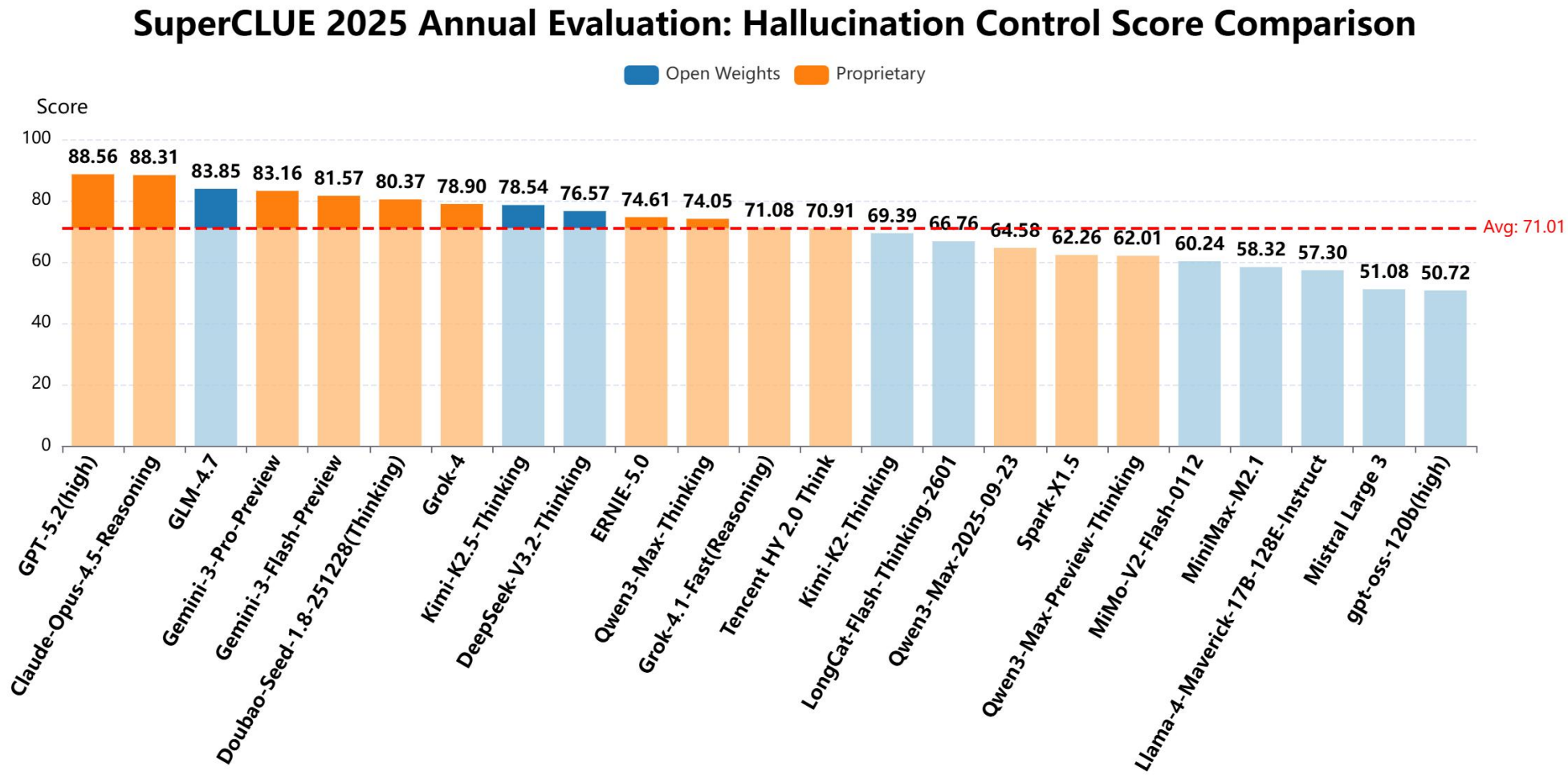
The top 4 spots—GPT-5.2 (high), Gemini-3-Flash-Preview, Claude-Opus-4.5-Reasoning, and Gemini-3-Pro-Preview—are all occupied by proprietary models. Only the Chinese open-weight model DeepSeek-V3.2-Thinking made it into the top 5.

Open-weight models scored an average of 59.26, trailing closed-source models by nearly 9 points (68.05). This indicates a clear disparity between the two. Most open-weight models are clustered in the lower half of the leaderboard, with the majority failing to reach the average benchmark.

Open Weights vs. Proprietary：Hallucination Control

Introduction： Primarily evaluates the model's ability to mitigate hallucination while performing Chinese generation tasks, covering fundamental semantic understanding and generation benchmarks such as text summarization, reading comprehension, multi-document QA, and dialogue completion.

Evaluation Method： Binary (0/1) evaluation of hallucination per sentence, based on human-verified reference answers.



Data Source: SuperCLUE, Jan 29, 2026.

Evaluation Analysis

Proprietary models demonstrate superior reliability.

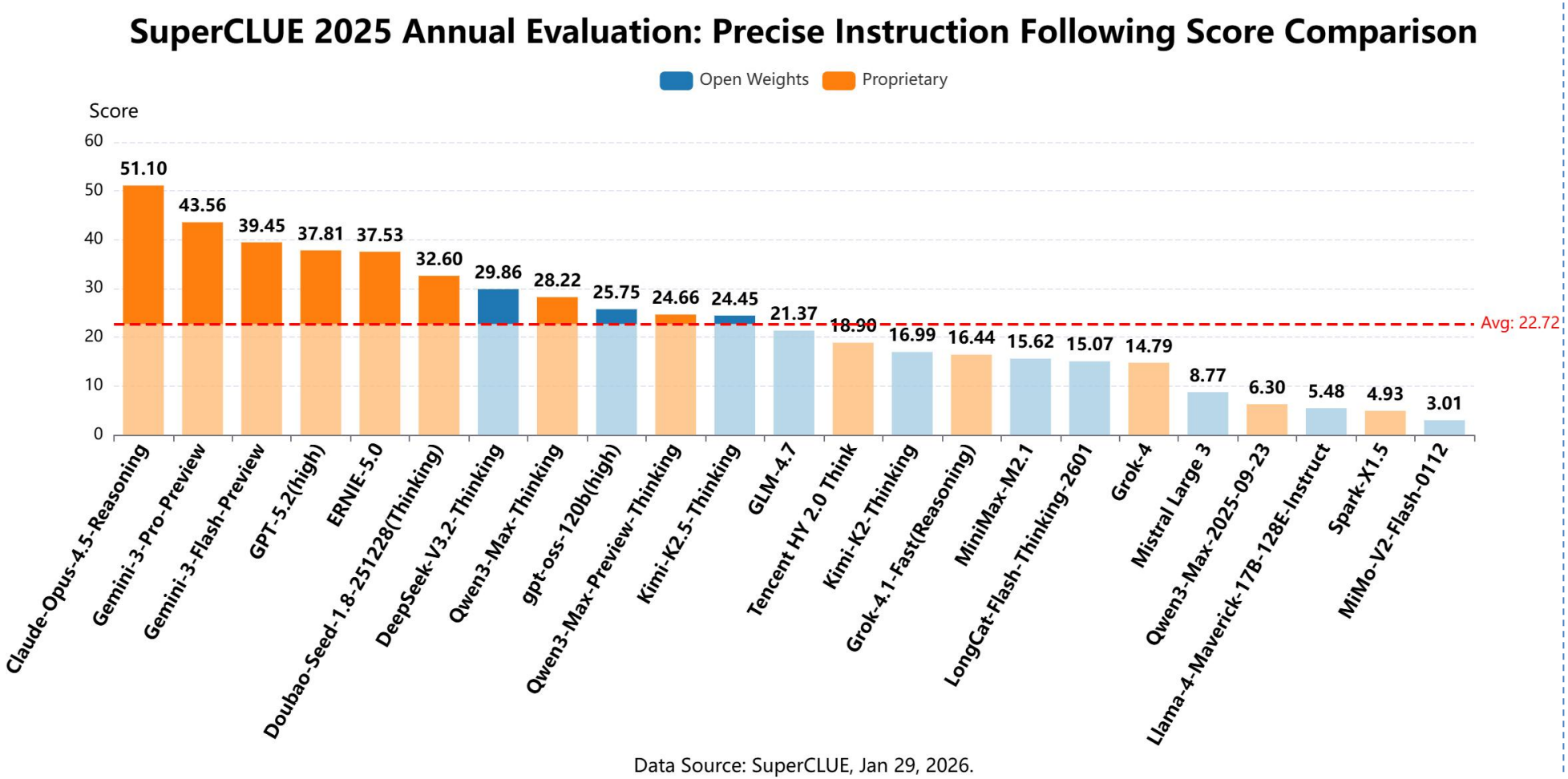
The top 2 spots are secured by proprietary models (GPT-5.2 (high) and Claude-Opus-4.5-Reasoning), both surpassing the 88-point mark.

Proprietary models occupy nearly the entire leaderboard summit, with only one open-weight model (GLM-4.7) breaking into the top 3. This highlights their distinct advantage in factual accuracy and contextual consistency.

Open Weights vs. Proprietary：Precise Instruction Following

Introduction： This assesses the model's ability to follow instructions—generating responses in specified formats and accurately presenting required data. Evaluated Chinese scenarios include structural, quantitative, semantic, and composite constraints (≥4 types).

Evaluation Method： Rule-based script 0/1 evaluation.



Data Source: SuperCLUE, Jan 29, 2026.

Evaluation Analysis

Proprietary models lead by an absolute margin.

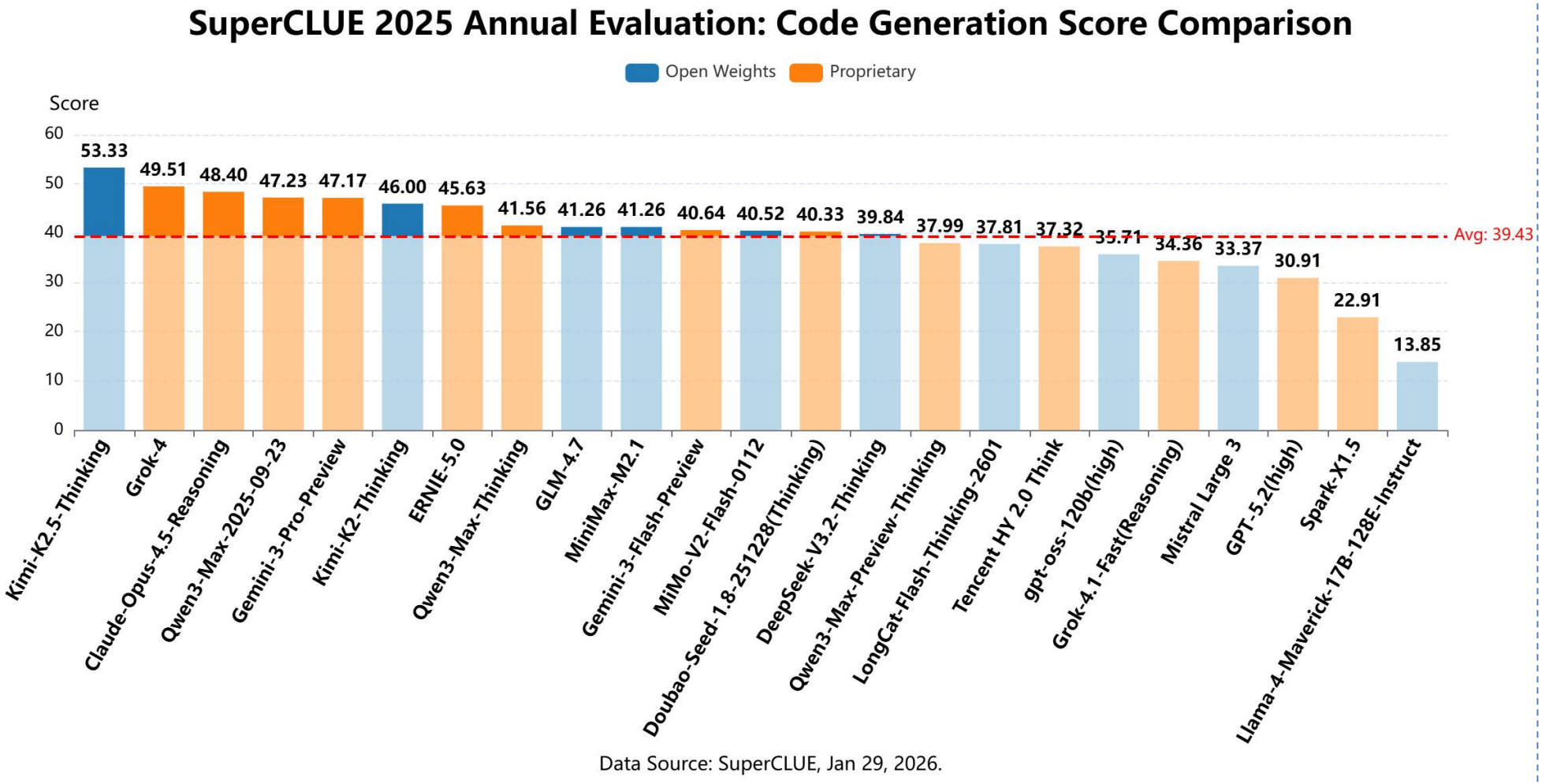
Precise instruction following is the area where the generational gap between open-weight and proprietary models is most evident. The proprietary camp demonstrates a near-crushing dominance, occupying the top 6 spots.

Although the open-weight model DeepSeek-V3.2-Thinking ranks 7th, it trails the leader by over 21 points. The runner-up among open-weight models, gpt-oss-120b (high), scores nearly half of the top model's points, highlighting the huge disparity between the two.

Open Weights vs. Proprietary：Code Generation

Introduction： The task has two types: (1) generating standalone functions covering data structures, algorithms, etc.; and (2) building complete interactive websites like travel booking, e-commerce, and social media platforms.

Evaluation Method： Scored 0/1 via unit tests (for standalone function generation) and functional tests simulating user interactions (for web app generation).



Data Source: SuperCLUE, Jan 29, 2026.

Evaluation Analysis

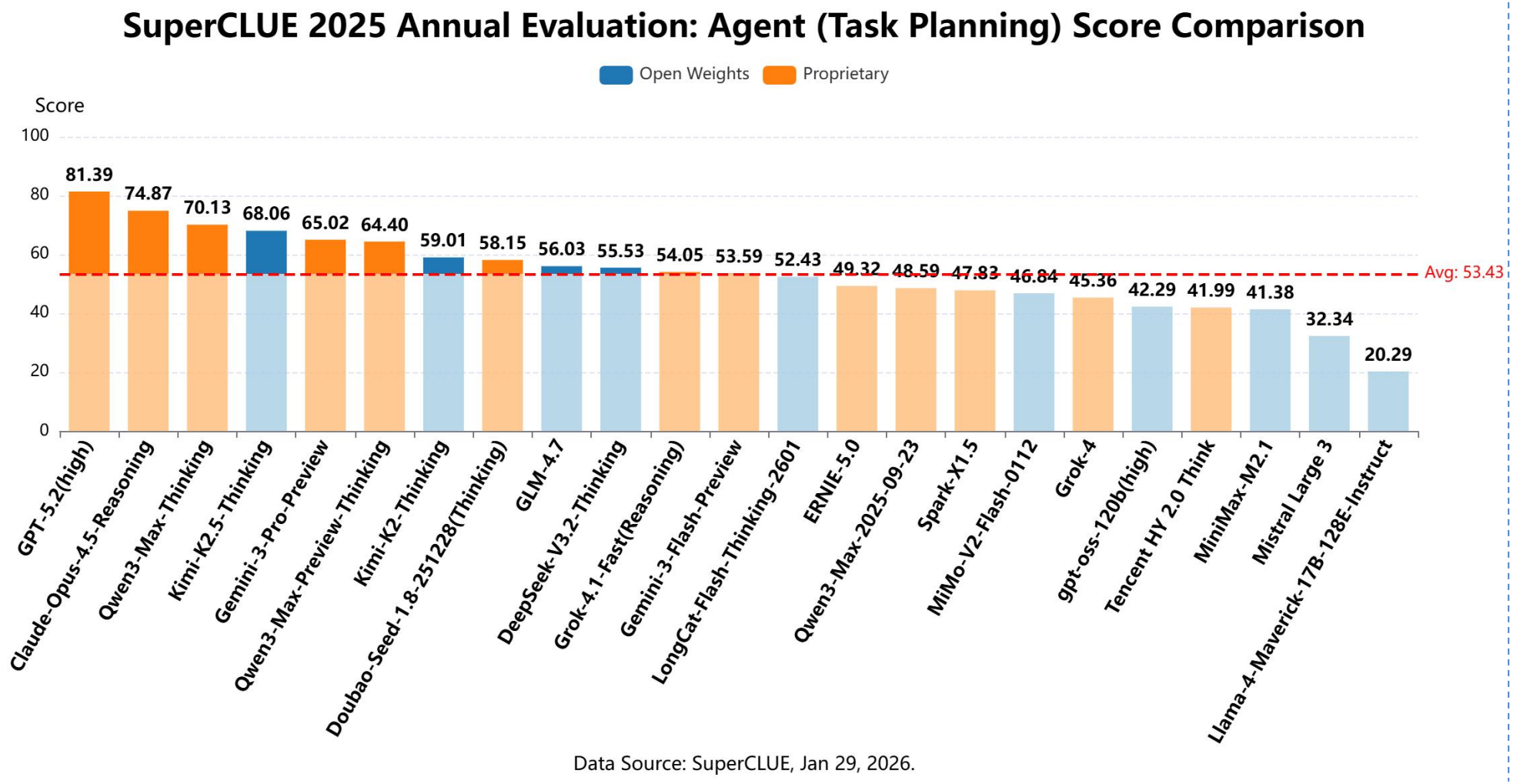
Open-weight models achieve a breakthrough at the top.

The open-weight model Kimi-K2.5-Thinking claimed the global #1 spot with a score of 53.33. It not only significantly surpassed the average but also outperformed the second-place proprietary model Grok-4 by 3.82 points. Notably, it is the only model to break the 50-point barrier in this coding task.Furthermore, Kimi-K2-Thinking (46.00), GLM-4.7 (41.26), and MiniMax-M2.1 (41.26) all ranked within the top 10. This demonstrates that the open-weight camp has developed strong competitiveness in specific vertical domains, such as programming.

Open Weights vs. Proprietary：Agent(Task Planning)

Introduction： Assesses the model's ability to formulate structured action plans in complex scenarios—e.g., lifestyle services, work collaboration, learning, and healthcare—by generating logical, clear, and executable steps based on given goals and constraints.

Evaluation Method： A judge model discretely evaluates (0/1) completion of predefined checkpoints or continuously scores (0—100) overall plan quality.



Data Source: SuperCLUE, Jan 29, 2026.

Evaluation Analysis

1. Proprietary models dominate the top tier in task planning.

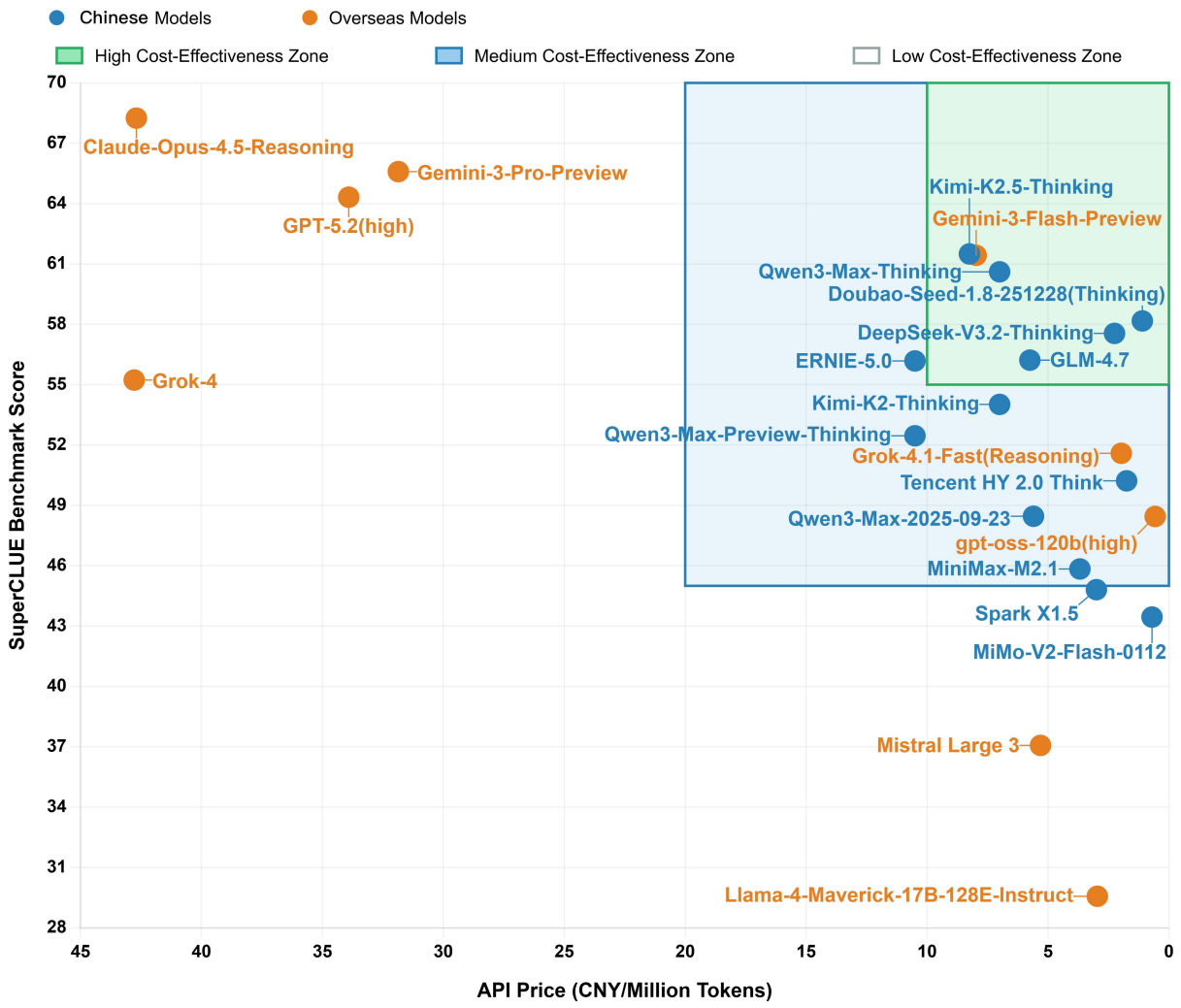
Leading proprietary models—GPT-5.2 (high), Claude-Opus-4.5-Reasoning, and Qwen3-Max-Thinking—all scored above 70, significantly outperforming open-source alternatives.

2. Open-weight models are catching up, nearing upper-mid level performance.

Open-weight leaders Kimi-K2.5-Thinking and Kimi-K2-Thinking scored over 59, nearing proprietary models like Qwen-3-Max-Thinking and Gemini-3-Pro-Preview. While they show potential in task planning, a gap remains.

Cost-Effectiveness Distribution

SuperCLUE 2025 Annual Evaluation: Cost-Effectiveness Distribution



Source: SuperCLUE, January 29, 2026.

Note: Open weights models such as DeepSeek-V3.2-Thinking are priced based on API usage, with official data as the reference. For models where API pricing is determined separately for input and output tokens, the overall cost is estimated here based on a 3:1 ratio of input to output tokens. Prices reflect the official standard rates (excluding promotional discounts) as of January 2026. Re-evaluated models use real-time pricing.

Evaluation Analysis

1. Chinese models offer higher cost-effectiveness compared to overseas models.

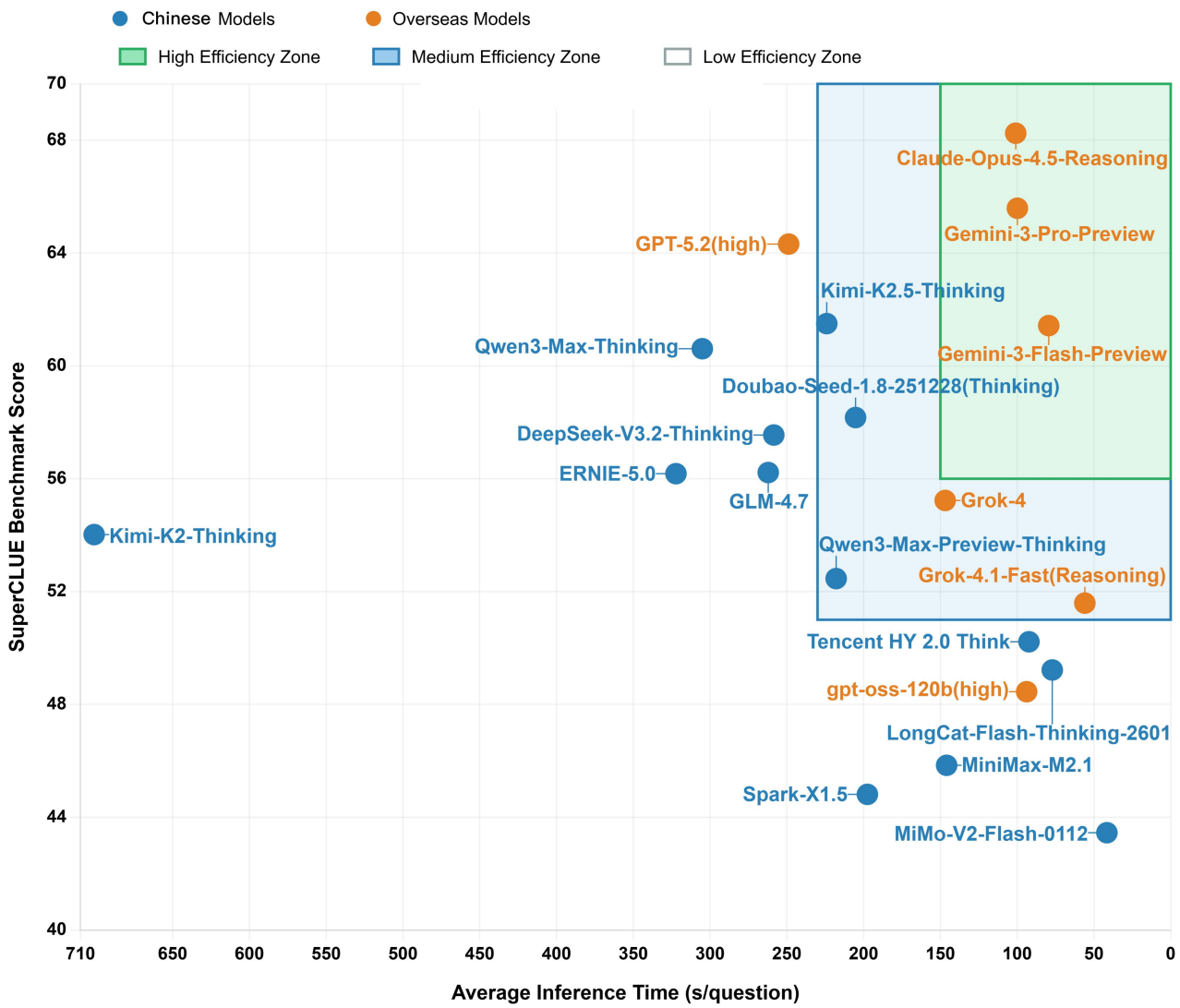
Chinese models are primarily positioned in the medium-to-high cost-effectiveness range, whereas overseas models are mostly in the medium-to-low range. Specifically, leading domestic models (such as Kimi-K2.5-Thinking and Qwen3-Max-Thinking) offer performance comparable to top international models at prices below 10 RMB per million tokens. In contrast, overseas models with similar performance are generally priced over three times higher, highlighting the cost advantage of Chinese models.

2. Overseas models generally follow a "High Quality, High Price; Low Quality, Low Price" trend.

Top overseas models (e.g., Claude-Opus-4.5-Reasoning, GPT-5.2) in the top-left quadrant show strong performance but carry high API costs (30 – 45 RMB/million tokens). Conversely, lower-priced models (e.g., Llama-4-Maverick, Mistral Large 3) in the bottom-right quadrant perform poorly (scores <40), indicating that low prices without adequate capability do not constitute true cost-effectiveness.

Inference Efficiency Distribution

SuperCLUE 2025 Annual Evaluation: Inference Efficiency Distribution



Source: SuperCLUE, January 29, 2026.

Note: The model inference speed data is based on selected models with publicly available APIs from the January 2026 evaluation. The average inference time represents the mean duration (seconds) across all task evaluation datasets.

Evaluation Analysis

1. Reasoning Performance: Overseas models significantly outperform Chinese models counterparts.

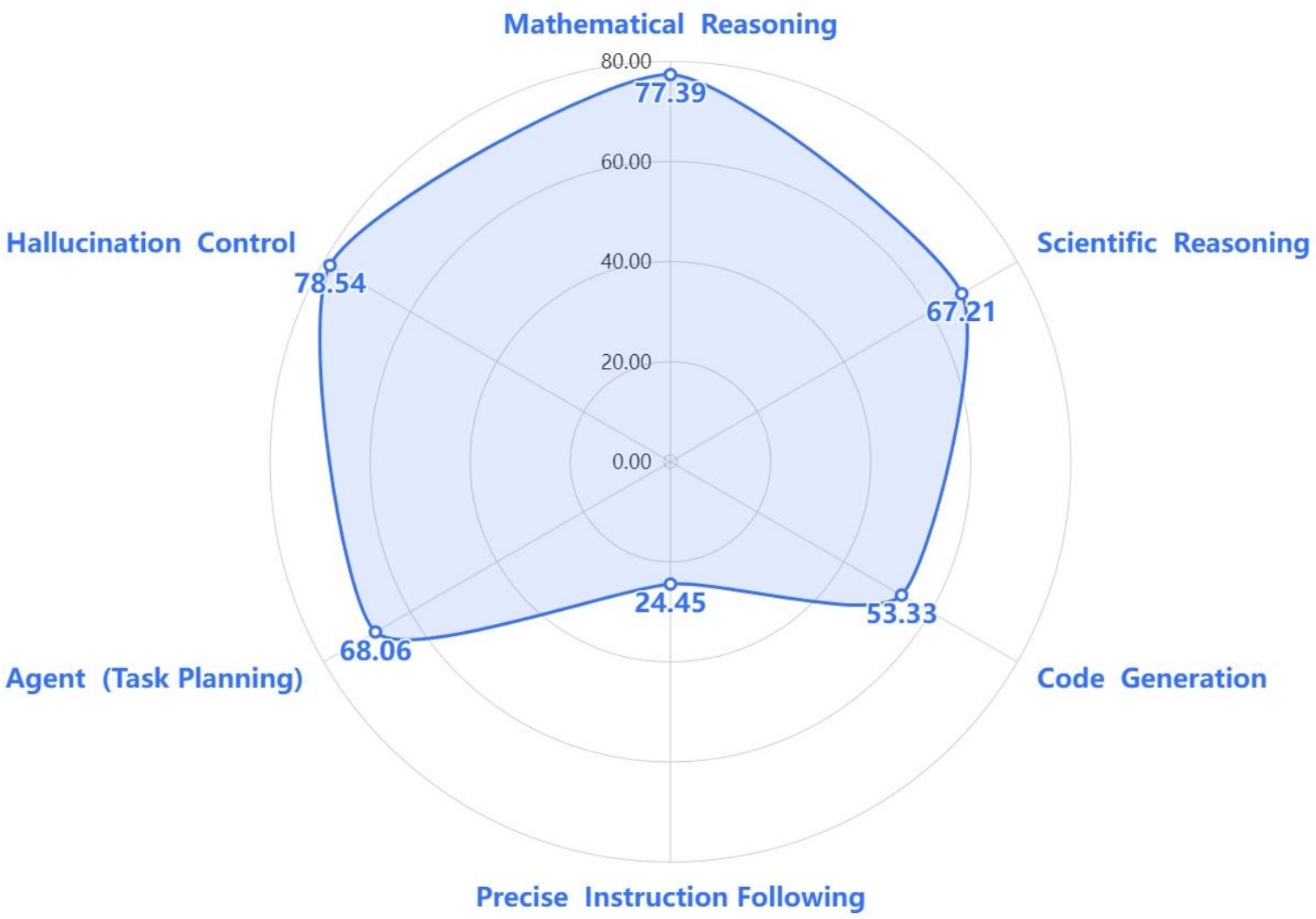
All models in the high-performance tier are overseas ones (Claude-Opus-4.5-Reasoning, the Gemini-3 series), with no Chinese models included. These three overseas models can balance reasoning efficiency while maintaining top-tier reasoning quality, achieving dual-dimensional optimization of both quality and speed. In the medium-performance tier, there are only three Chinese models: Kimi-K2.5-Thinking, Doubao-Seed-1.8-251228(Thinking) and Qwen3-Max-Preview-Thinking. All other Chinese models fall into the low-performance tier. This reflects that Chinese models still lag behind the world's top models in the collaborative optimization of reasoning quality and efficiency, leaving considerable room for improvement.

2. Chinese models: beginning to deliver high performance and high efficiency.

Taking the Kimi series models as an example, in the iterative evolution from Kimi-K2-Thinking (scoring 54.02 points with an average of 701.09 seconds per question) to Kimi-K2.5-Thinking (scoring 61.50 points with an average of 224 seconds per question), the reasoning capability has increased by nearly 14% and the reasoning speed has nearly tripled. This fully demonstrates that Chinese large models are shifting from standalone performance optimization to the collaborative optimization of performance and efficiency, with remarkable results achieved.

SuperCLUE 2025 Annual Benchmark Evaluation: Kimi-K2.5-Thinking

Scores Across Six Tasks



Evaluation Analysis

1. Overview

Kimi-K2.5-Thinking, released by Moonshot AI on Jan 27, 2026, is a natively multimodal model. It sets new open-weight SoTA records in Agents, coding, and visual understanding, with a major leap in front-end coding.

2. Strengths

(1) Code: True to its promotion, the model excels in code generation (53.33), leading global rankings. It ranks 1st in Web Coding and 2nd in standalone function generation, demonstrating world-class front-end proficiency.

(2) Agent - Task Planning: It scores 68.06 on Agent tasks, performing on par with top-tier international models such as GPT-5.2(high) and Claude-Opus-4.5-Reasoning.

(3) Complex Reasoning: The model scores 77.39 in mathematical reasoning (4th globally), trailing Gemini-3-Pro-Preview by only 3 points. In scientific reasoning (67.21), it ranks within the domestic Top 5, placing its overall reasoning ability among the global leaders.

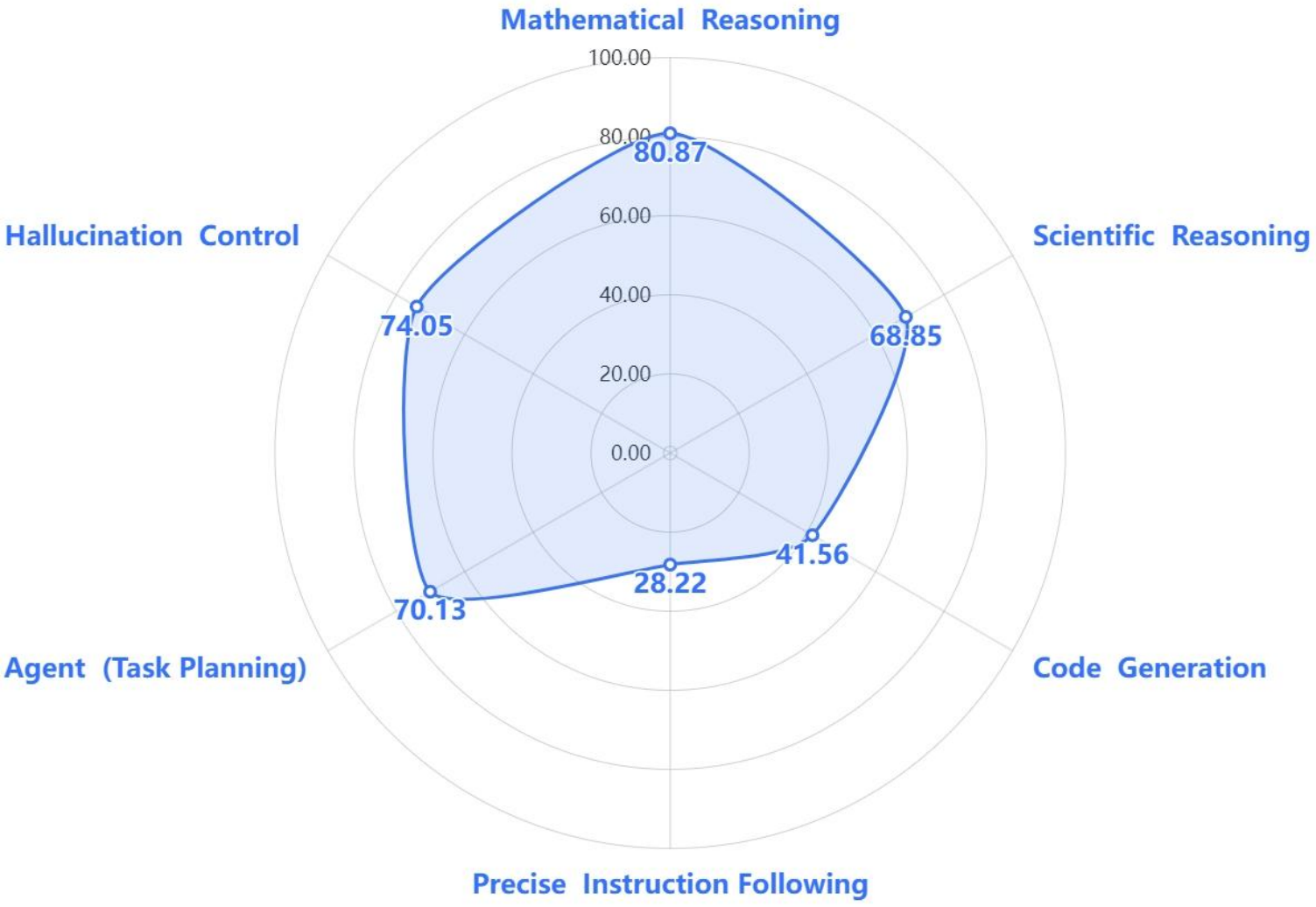
3. Areas for Improvement

(1) Precise Instruction Following: The model scored only 24.45 on this task, ranking in the middle. It lags behind the best overseas models by over 26 points and Chinese leaders by over 13 points.

(2) Hallucination Control: Scoring 78.54 (up 9 points from the previous version), it sits in the upper-middle range but still trails the leading models by approximately 10 points.

SuperCLUE 2025 Annual Benchmark Evaluation: Qwen3–Max–Thinking

Scores Across Six Tasks



Evaluation Analysis

1. Overview

Qwen3-Max-Thinking, Alibaba's latest flagship reasoning model released on January 26, 2026, rivals top international models such as GPT-5.2(high), Claude-Opus-4.5-Reasoning, and Gemini-3-Pro-Preview in factual knowledge, complex reasoning, and agent capabilities.

2. Strengths

（1）Complex Reasoning: The model achieved outstanding results in the general evaluation. In mathematical reasoning, it tied for first place globally with Gemini-3-Pro-Preview at 80.87, surpassing GPT-5.2(high) and Claude-Opus-4.5-Reasoning. It also ranked sixth globally in scientific reasoning with a score of 68.85, demonstrating formidable overall reasoning power.

（2）Agent - Task Planning: Scoring 70.13, the model entered the global Top 3, surpassing Gemini-3-Pro-Preview and matching the performance of Claude-Opus-4.5-Reasoning.

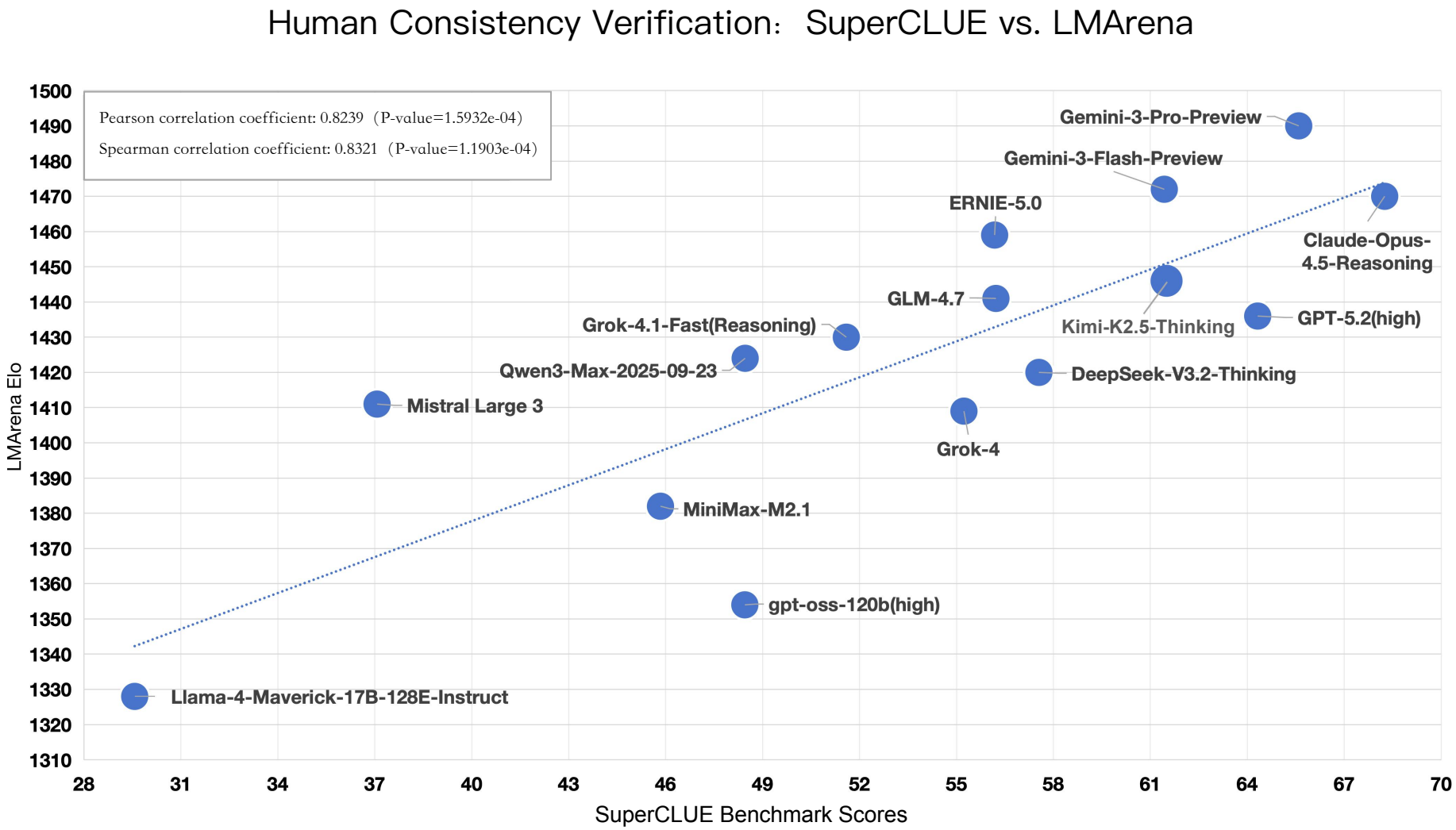
3. Areas for Improvement

（1）Hallucination Control: The model scored 74.05, a roughly 12-point improvement over the Preview version. However, it remains mid-tier, trailing leading models by approximately 14 points.

（2）Precise Instruction Following: With a score of 28.22, it sits in the middle range, nearly 23 points behind the best overseas models and over 9 points behind the domestic leaders.

（3）Code Generation: Scoring 41.56, it surpasses Gemini-3-Flash-Preview but still lags behind the best models by about 12 points.

Human Consistency Verification: SuperCLUE vs. LMArena



LMArena is a leading English-language leaderboard for large language models, conducting pairwise evaluations based on anonymous public voting. Correlating SuperCLUE and LMArena scores yields:

Pearson correlation coefficient: 0.8239 (P-value=1.5932e-04)

Spearman correlation coefficient: 0.8321 (P-value=1.1903e-04)

SuperCLUE benchmark scores show high consistency with human evaluations, exemplified by the crowdsourced anonymous voting on LMArena.

Source: SuperCLUE, January 29, 2026.

Spearman correlation coefficient: Measures monotonic association between two variables (range: [−1, 1]); higher absolute values indicate stronger correlation.

Pearson correlation coefficient: Measures linear correlation between two continuous variables (range: [−1, 1]); higher absolute values indicate stronger correlation.

Appendix I: Introduction to SuperCLUE

SuperCLUE

The SuperCLUE is the evolution and continuation of the CLUE (The Chinese Language Understanding Evaluation) benchmark in the era of large models. It is an independent and leading comprehensive evaluation benchmark for general-purpose large models. The CLUE benchmark for Chinese language understanding was launched in 2019 and has successively introduced widely cited evaluation benchmarks such as CLUE, FewCLUE, and ZeroCLUE.

2019 — CLUE Benchmark Release

2020 — Released multiple papers in top-tier conferences and organized the open evaluation task at NLPCC 2020.

2021 — Released FewCLUE、DataCLUE、ZeroCLUE and so on.

2022 — Collaborated with Westlake University to release DataCentricAI. Released SimCLUE, a large-scale Chinese semantic understanding dataset.

2023 — SuperCLUE officially launched; Released the first Chinese multi-turn open-domain evaluation benchmark, SuperCLUE-Open; Released industry-specific evaluation benchmarks.

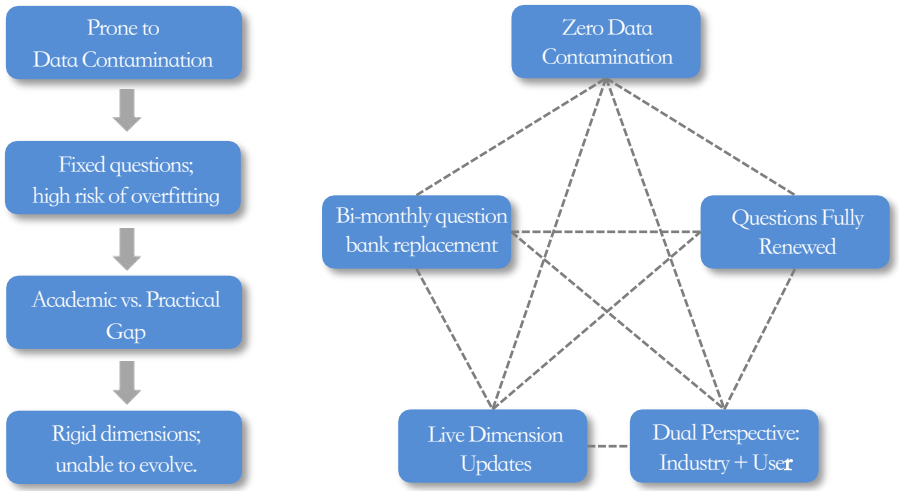
2024 — Released Math, Code, Long-text, and RAG benchmarks. Released multimodal benchmarks for Video, Image, and Vision. Released reasoning benchmarks. Released the AgentCLUE benchmark.

2025.01-2025.06 — Released AgentCLUE-General (General Agent). Released Multimodal (Text/Image-to-Video). Released Core Ability (Hallucination, Deep Research).

2025.07-2025.12 — Released new Benchmarks: Precise Instruction, Deep Search, SWE, Embodied AI, WebCoding. Released Agent Benchmarks: Mobile GUI Agent, CUA. Released a new Interactive Evaluation Mode: SuperCLUE Chinese Arena.

The Difference between SuperCLUE and Traditional Evaluation

Traditional Evaluation VS **SuperCLUE**



Three Key Features of SuperCLUE

01 Live Updates, Zero Data Contamination

100% Refresh / Bi-Monthly — No Overfitting; Live Framework — Evolves with LLMs

02 Consistent with User Interaction

Mimicking real user interaction and scenarios to ensure an authentic evaluation perspective.

03 Independent 3rd-Party

Fully Independent 3rd-Party: No proprietary models; Commitment to Objectivity: Delivering unbiased, neutral evaluations.

Appendix II: Full View of Rankings

SuperCLUE

Leveraging expertise in LLM development trends, SuperCLUE has established a multi-dimensional evaluation system. It covers the spectrum from foundational capabilities to advanced applications, including General, Text, Multimodal, Reasoning, Agent, AI Application, and Performance benchmarks—providing critical insights for industry and academic research.

The Complete Landscape of SuperCLUE Leaderboard

General Benchmark

- Math Reasoning
- Scientific Reasoning
- Code Generation
- Agent (Task Planning)
- Precise Instruction Following
- Hallucination Control

Text Benchmark

- Precise Instruction Following SuperCLUE-CPIF
- Factual Hallucination SuperCLUE-Fact
- Faithfulness Hallucination SuperCLUE-Faith
- Long-form Text SuperCLUE-Long
- Ultra-long Context SuperCLUE-200K
- Role-playing SuperCLUE-Role
- RAG SuperCLUE-RAG
- Adversarial Safety SuperCLUE-Safety

Mutimodal Benchmark

Video

- World Model SuperCLUE-World
- Text-to-Video SuperCLUE-T2V
- Image-to-Video SuperCLUE-I2V

Vision

- Image-Editting SuperCLUE-Edit
- Real-time-Interaction SuperCLUE-Live
- Text-to-Image SuperCLUE-Image
- Mutimodal Understanding SuperCLUE-VLM

Voice

- Real-time Voice SuperCLUE-Voice
- Text-to-Speech SuperCLUE-TTS
- Voice Cloning SuperCLUE-Cloning

Reasoning Benchmark

- Web Coding SuperCLUE-WebCoding
- Software Engineering SuperCLUE-SWE
- Visual Reasoning SuperCLUE-VLR
- Scientific Reasoning SuperCLUE-Science
- Math Competition MathCLUE
- Project-level Code SuperCLUE-Project
- Chain-of-Thought SuperCLUE-COT
- Primary Olympic Math SuperCLUE-Math6o
- Code SuperCLUE-Code3
- Multi-step Reasoning SuperCLUE-Math6
- Graduate-Level Mathematics

Agent Benchmark

General

- Agentic Coding AgentCLUE-SWE
- Embodied AI EmbodiedCLUE-VLA
- Deep Search SuperCLUE-DeepSearch
- Deep Research SuperCLUE-DeepResearch
- General AI Agent AgentCLUE-General AgentCLUE-tGeneral
- Executable Agent AgentCLUE
- Non-executable Agent SuperCLUE-Agent

Endpoint

- Computer Use Agent(offline) AgentCLUE-Mobile
- Mobile GUI-Agent（offline） AgentCLUE-Mobile

Industry

- Cockpit Intelligent Agent AgentCLUE-ICabin

AI Application Benchmark

AI Products

- AI Coding Kit AgentCLUE-AICodeKit
- AI Search SuperCLUE-AISearch
- Code Assistant SuperCLUE-Coder

Industry

- Advertising & Marketing SuperCLUE—Mkt
- Automotive Knowledge SuperCLUE—AutoQA
- Finance SuperCLUE-Fin
- Industry SuperCLUE-Industry
- Automobile SuperCLUE-Auto
- Intelligent Cockpit SuperCLUE-Icabin

Performance Benchmark

DeepSeek-R1 Series

- DeepSeek-R1 Third-Party Web Search Capability Test (Web Version)
- DeepSeek-R1 Third-Party Stability Test (App)
- DeepSeek-R1 Third-Party Stability Test (API)
- DeepSeek-R1 Third-Party Stability Test (Web)

- LLM Inference Speed Benchmark
- Domestic Chip Benchmark

Others

- Open-source Dataset Math24o

Released　Coming Soon

Note: General benchmarks are detailed in the report; click links to access the latest release articles for other benchmarks.

Appendix III: List of 2025 Annual Assessment Models

SuperCLUE

The SuperCLUE 2025 Annual Benchmark evaluated 23 representative large models from China and abroad; the following table provides detailed information on these models.

Number	Model Name	Institution	Introduction	Number	Model Name	Institution	Introduction
1	Qwen3-Max-2025-09-23	Alibaba	The foundational model officially released, utilizing the publicly available API from Alibaba Cloud: qwen3-max-2025-09-23.	13	Mistral Large 3	Mistral AI	The latest open-source model officially released, utilizing the official API: mistral-large-2512.
2	Qwen3-Max-Preview-Thinking	Alibaba	The official reasoning model, using the Alibaba Cloud public API: qwen3-max-preview-thinking.	14	Grok-4	X.AI	The reasoning model officially released, utilizing the official API: grok-4-0709.
3	DeepSeek-V3.2-Thinking	DeepSeek	The latest open-source reasoning model officially released, utilizing the official API: deepseek-reasoner.	15	Grok-4.1-Fast(Reasoning)	X.AI	The reasoning model officially released, utilizing the official API: grok-4-1-fast-reasoning.
4	GLM-4.7	Z.AI	The latest open-source reasoning model officially released, utilizing the official API: glm-4.7.	16	Llama-4-Maverick-17B-128E-Instruct	Meta	Utilizing the Together.ai interface: meta-llama/Llama-4-Maverick-17B-128E-Instruct-FP8.
5	Spark X1.5	iFlytek	The latest reasoning model officially released, utilizing the official API: Spark X1.5.	17	LongCat-Flash-Thinking-2601	Meituan	The latest open-source reasoning model officially released, utilizing the official API: LongCat-Flash-Thinking-2601.
6	Kimi-K2-Thinking	Moonshot	The open-source reasoning model officially released, utilizing the official API: kimi-k2-thinking.	18	Doubao-Seed-1.8-251228(Thinking)	ByteDance	The latest model officially released, utilizing the official API: doubao-seed-1-8-251228.
7	MiniMax-M2.1	MiniMax	The latest open-source reasoning model officially released, utilizing the official API: MiniMax-M2.1.	19	MiMo-V2-Flash-0112	Xiaomi	The latest reasoning model officially released and open-sourced, utilizing the official API: mimo-v2-flash.
8	Claude-Opus-4.5-Reasoning	Anthropic	The mixed-reasoning model officially released, utilizing the official API: Claude-Opus-4.5-Reasoning.	20	ERNIE-5.0	Baidu	The latest official full-modal model, using the official API: ERNIE-5.0-Thinking-Preview.
9	Gemini-3-Flash-Preview	Google	The latest model officially released by Google, utilizing the official API: gemini-3-flash-preview.	21	Qwen3-Max-Thinking	Alibaba	This is the latest reasoning model released officially, using the official API: qwen3-max-2026-01-23, with enable_thinking=True.
10	Gemini-3-Pro-Preview	Google	The latest model officially released by Google, utilizing the official API: gemini-3-pro-preview.	22	Kimi-K2.5-Thinking	Moonshot	The latest open-source multimodal model released by the official team, accessible via the official API: kimi-k2.5.
11	GPT-5.2(high)	OpenAI	The latest reasoning model officially released, utilizing the official API: gpt-5.2-2025-12-11.	23	Tencent HY 2.0 Think	Tencent	The latest reasoning model released officially, using the official API: hunyuan-2.0-thinking-20251109.
12	gpt-oss-120b(high)	OpenAI	The open-source reasoning model officially released, utilizing OpenRouter's API: gpt-oss-120b.	/	/	/	/

SuperCLUE

Independent Third-Party AGI Evaluation Institution

“

Provide professional evaluation services and independent analysis for AI applications and R&D teams to assist in technology selection and performance optimization

”

Contact Us

——Provide professional evaluation services and independent analysis for AI applications and R&D teams to assist in technology selection and performance optimization

——As a leading third-party large model evaluation institution, we are committed to delivering professional assessment services to the industry.

General LLMs Evaluation

Offer comprehensive large-model evaluation services, delivering holistic assessment reports—including multi-dimensional results, benchmark comparisons, representative examples, and optimization recommendations.

Industry and Specialized LLMs Evaluation

Focus on evaluating the real-world application effectiveness of large models across industries—including automotive, mobile, finance, industrial, education, and healthcare—with assessments covering Chinese-agent capabilities, model safety, multimodal performance, and personalized role-playing abilities.

Multimodal LLMs Evaluation

Comprehensively evaluate multimodal large models across multiple dimensions, covering core and applied capabilities—including real-time multimodal interaction, video generation, text-to-image synthesis, multimodal understanding, image editing, speech synthesis, and world modeling.

Agent Evaluation

Deliver evaluations of AI large model applications and tools, including general-purpose agents (e.g., AgentCLUE, AgentCLUE-General), coding assistants, AI search, and on-device implementations such as AI PCs, AI smartphones, XR devices, and embodied intelligence systems.

In-Depth Report on LLMs

Delivers in-depth research reports on domestic and international large models, offering comprehensive analysis of their technological advancements and real-world applications to provide timely, expert third-party insights for enterprises and institutions.

Business Collaboration: Please briefly describe your requirements and send them to contact@superclue.ai

SuperCLUE

Exchange Collaboration

Scan to follow

- **Official Leaderboard URL：https://www.superclueai.com**
- **Official Website：www.CLUEbenchmarks.com**
- **Github：https://github.com/CLUEbenchmark**
- **Contact Person：Professor Xu 18806712650（Wechat）**
  Professor Zhu 18621237819（Wechat）

Legal Notice

• Copyright Notice

This report is produced by the SuperCLUE team and copyrighted by SuperCLUE. Any reproduction or citation must credit SuperCLUE and shall not alter, abridge, or misrepresent the original content. Unauthorized use, unattributed reproduction, or commercial exploitation violates the PRC Copyright Law, other applicable laws, and relevant international conventions. SuperCLUE disclaims all liability for any consequences arising from misinterpretation, malicious distortion, abridgement, or modification of this report and reserves the right to pursue legal remedies.

• Disclaimer

This report is based on the automated evaluation results from the 2025 annual Chinese Large Model Benchmark (SuperCLUE) and publicly available information, aiming for accuracy and objectivity. All data and analyses reflect conditions as of the report's issuance date; no assurance is provided regarding their continued validity or future changes. The opinions, assessments, and forecasts herein represent views as of the issuance date and are subject to change without notice. Different conclusions may be drawn based on varying assumptions, methodologies, real-time information, or market performance, and there is no obligation to update recipients accordingly.

The team strives for objectivity and fairness in this report. However, the views, conclusions, and recommendations herein are for reference only and do not constitute investment advice. Neither the company nor the authors assume any legal liability for any consequences arising from reliance on or use of this report or any other related research reports issued by the company.