SuperCLUE
中文大模型综合性测评基准

中文大模型基准测评2025年年度报告

——2026开年特别版：含1月底重磅模型动态评测

SuperCLUE团队

2026.02.04

SuperCLUE
中文大模型综合性测评基准
独立第三方AGI测评机构

精准量化通用人工智能（AGI）进展，定义人类迈向AGI的路线图

Accurately Quantifying the Progress of AGI,
Defining the Roadmap for Humanity's Journey towards AGI.

报告目录

SuperCLUE
中文大模型综合性测评基准

第一部分
2025年关键进展

1. 2025年最值得关注的中文大模型全景图
2. 2025年最值得关注的智能体产品全景图
3. 2025年年度大模型关键进展
4. 2025年全年SuperCLUE通用基准测评海内外大模型Top3

SuperCLUE：2025年最值得关注的中文大模型全景图

SuperCLUE
中文大模型综合性测评基准

文本

通用开源	Qwen、智谱AI、盘古大模型、LongCat、腾讯混元、书生·浦语、MiLM、DeepSeek、MINIMAX、KIMI、百灵大模型、阶跃星辰、面壁小钢炮 MiniCPM、ERNIE-4.5系列
通用闭源	字节豆包、腾讯混元、Qwen3-Max、文心、日日新 sensenova、ZTE中兴、360智脑、讯飞星火
推理	Kimi-K2.5-Thinking、Qwen3-Max-Thinking、Doubao-Seed-1.8、DeepSeek-V3.2、GLM-4.7、ERNIE-5.0、Tencent HY 2.0 Think

多模态

视觉理解	Doubao-vision、Qwen3-VL、ERNIE-5.0、K2.5、日日新 sensenova、GLM-4.6V、腾讯混元、阶跃星辰
文生图	即梦AI、可灵AI、通义万相、ERNIE-5.0、腾讯混元-Image 3.0、GLM-Image、LongCat-Image、讯飞星火
图片编辑	即梦AI、通义万相、GLM-Image、Qwen-Image-Edit、腾讯混元-Image 3.0、阶跃星辰、LongCat-Image
文生视频	通义万相、可灵AI、即梦AI、海螺AI、拍我AI、清影、腾讯混元、Vidu、SkyReels、LongCat-Video
图生视频	可灵AI、海螺AI、即梦AI、通义万相、拍我AI、Vidu、清影、阶跃星辰
实时交互	字节豆包、讯飞星火、千问、日日新 sensenova、海螺AI、文心、智谱清言、Kimi
语音合成	Doubao Seed TTS 2.0、讯飞语音合成、Qwen3-TTS、Fish Audio、Speech-2.6-HD、百度TTS

行业

医疗	百度灵医、蚂蚁阿福、讯飞晓医、百川智能
汽车	理想 MindGPT、极氪Kr大模型、易车大模型
工业	奇智孔明、华为盘古工业大模型、羚羊工业大模型
教育	MathGPT、子曰、豆包爱学、作业帮
金融	蚂蚁金融大模型、妙想金融大模型、度小满 轩辕大模型
法律	Chat Law、CHINESE LAW 元典智库、得理法搜

SuperCLUE：2025年最值得关注的国内智能体产品全景图

通用领域

扣子空间 | MINIMAX | 天工Agent版 | 智谱清言 | 实在Agent | flowith
心响 | 纳米AI | AutoGLM | ERNIE Agent | Genie | 百度文库 | GenFlow 3.0
OK Computer | 心流 | OWL | OpenManus | MasterAgent | JENIUS

垂直领域

深度研究

深入研究 | Deep Research | 秘塔AI搜索 | 夸克
Qwen 深入研究 | 阶跃星辰 | 心流 | SciMaster

编程

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

法律

通义法睿 | 法天使 LEGAL
GptLaw法律AI | MetaLaw

金融

问财 | FinGenius | 金灵AI
财跃 | QUTKE

搜索

秘塔AI搜索 | 纳米AI搜索 | 心流
AI搜索 | 梯子AI | 博查 | 开搜AI

桌面

QoderWork | 阶跃AI桌面伙伴 | UI-TARS-desktop
MiniMax | LOONA Deskmate | Skywork 桌面版

营销

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

设计

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

旅行

飞猪旅行 | iMean.AI

办公

飞书 | WPS AI | 小浣熊家族 Raccoon
钉钉 AI时代的工作方式 | midu 校对通

教育

斑马 | 猿辅导在线教育
飞象老师 | 天学网·教师智能体

2025年年度大模型关键进展

◆自2022年11月30日ChatGPT发布以来，AI大模型在全球范围内掀起了有史以来规模最大的人工智能浪潮。国内外AI机构在过去3年里有了实质性的突破。具体可分为三个时期：百模大战与多模态萌芽、多模态爆发与推理突破、智能体崛起与生态重构。

SuperCLUE：AI大模型2025年关键进展

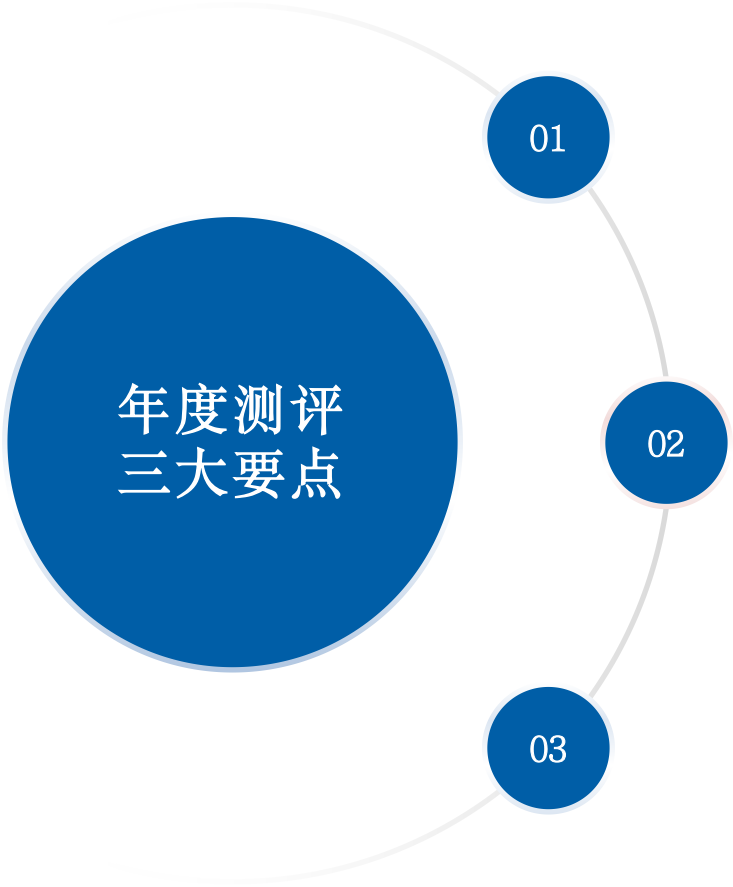
2025年全年SuperCLUE通用基准测评海内外大模型Top3

测评时间	国内第一	国内第二	国内第三	海外Top3
2026年1月	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)
2025年11月	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
2025年9月	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)
2025年7月	DeepSeek-V3.1-Thinking	Doubao-Seed-1.6-thinking-250715	DeepSeek-R1-0528	GPT-5(high)、o3(high)、o4-mini(high)
2025年5月	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
2025年3月	DeepSeek-R1	QwQ-32B	Doubao-1.5-pro-32k-250115	o3-mini(high)、Claude 3.7 Sonnet、GPT-4.5-Preview

SuperCLUE
中文大模型综合性测评基准

第二部分
2025年年度测评结果与分析

1. 2025年年度中文大模型基准测评介绍
2. 2025年全球大模型中文智能指数排行榜
3. 2025年SuperCLUE模型象限
4. 2025年SuperCLUE模型能力格局
5. SuperCLUE2025年年度测评六大任务国内Top3
6. SuperCLUE2025年年度测评六大任务国内外Top20热力图
7. 2025年年度中文大模型基准测评——总榜
8. 2025年年度中文大模型基准测评——开源模型
9. 海内外大模型对比分析
10. 开闭源大模型对比分析
11. 大模型性价比区间分布
12. 大模型推理效能区间分布
13. 代表性模型分析：Kimi-K2.5-Thinking&Qwen3-Max-Thinking
14. 评测与人类一致性验证：对比LMArena



1. 海外闭源模型仍占据榜单头部位置。

在本次2025年年度中文大模型基准测评中，Anthropic旗下的Claude-Opus-4.5-Reasoning以68.25分的总分位居榜首，Google的Gemini-3-Pro-Preview（65.59分）和OpenAI的GPT-5.2(high)（64.32分）紧随其后。国内开源最佳模型Kimi-K2.5-Thinking（61.50分）和闭源最佳模型Qwen3-Max-Thinking（60.61分）分列全球第四和第六。

2. 国产大模型正从"跟跑"向"并跑"阶段加速演进。

从2025年年初DeepSeek-R1发布，以对标OpenAI o1的性能极大地缩小了海内外模型的差距，到Kimi-K2.5-Thinking和Qwen3-Max-Thinking的发布分别在代码生成任务和数学推理任务上领跑全球，越来越多的国产大模型开始加速追赶上国际顶尖大模型的步伐，甚至在部分领域有所超越。

3. 海内外开闭源模型结构性差异显著。

闭源阵营呈现出"**海外领先、国产追赶**"的格局。海外闭源模型以Claude、Gemini、GPT为代表，构成了海外闭源大模型的第一梯队。国产闭源模型以Qwen3-Max-Thinking、Doubao-Seed-1.8-251228(Thinking)、ERNIE-5.0为代表，虽然与海外头部仍有差距，但已形成有效的竞争态势。开源阵营则呈现出"**国产主导、海外式微**"的格局。国产开源模型以Kimi-K2.5-Thinking、DeepSeek-V3.2-Thinking、GLM-4.7为代表，构成了国产开源模型的第一梯队，媲美海外顶尖闭源模型。海外开源模型的表现相对平淡，gpt-oss-120b、Mistral等被国产开源模型大幅超越。

2025年年度中文大模型基准测评介绍

中文语言理解测评基准CLUE（The Chinese Language Understanding Evaluation）是致力于科学、客观、中立的语言模型评测基准，发起于2019年。SuperCLUE是大模型时代CLUE基准的发展和延续，聚焦于通用大模型的综合性测评。本次2025年年度中文大模型基准测评聚焦通用能力测评，测评集由六大任务构成，总量为998道简答题，测评集的介绍如下：

SuperCLUE-2025年年度通用基准数据集及评价方式

1.数学推理

介绍：主要考察模型运用数学概念和逻辑进行多步推理和问题解答的能力。包括但不限于几何学、代数学、概率论与数理统计等竞赛级别数据集。

评价方式：基于参考答案的0/1评估，模型答案与参考答案一致得1分，反之得0分，不对回答过程进行评价。

2.科学推理

介绍：主要考察模型在跨学科背景下理解和推导因果关系的能力。包括物理、化学、生物等在内的研究生级别科学数据集。

评价方式：基于参考答案的0/1评估，模型答案与参考答案一致得1分，反之得0分，不对回答过程进行评价。

3.代码生成

介绍：该任务分为两大类型：一是独立功能函数生成，生成覆盖数据结构、算法等领域的独立函数。二是Web应用生成，要求模型构建旅游订票、电商、社交媒体等完整的交互式网站。

评价方式：通过单元测试进行0/1评分（独立功能函数生成）；通过模拟用户交互的功能测试进行0/1评分（Web应用生成）。

4.智能体(任务规划)

介绍：主要考察模型在复杂任务场景中制定结构化行动方案的能力，包括且不限于生活服务、工作协作、学习成长、健康医疗等。要求模型基于给定目标和约束条件，生成逻辑连贯、步骤清晰、可执行的行动计划。

评价方式：利用裁判模型根据行动方案对预设检查点的完成情况进行离散判定（0/1），或对方案整体质量进行连续评分（0–100）。

5.精确指令遵循

介绍：主要考察模型的指令遵循能力，包括但不限于定义的输出格式或标准来生成响应，精确地呈现要求的数据和信息。涉及的中文场景包括但不限于结构约束、量化约束、语义约束、复合约束等不少于4个场景。

评价方式：基于规则脚本的0/1评估。

6.幻觉控制

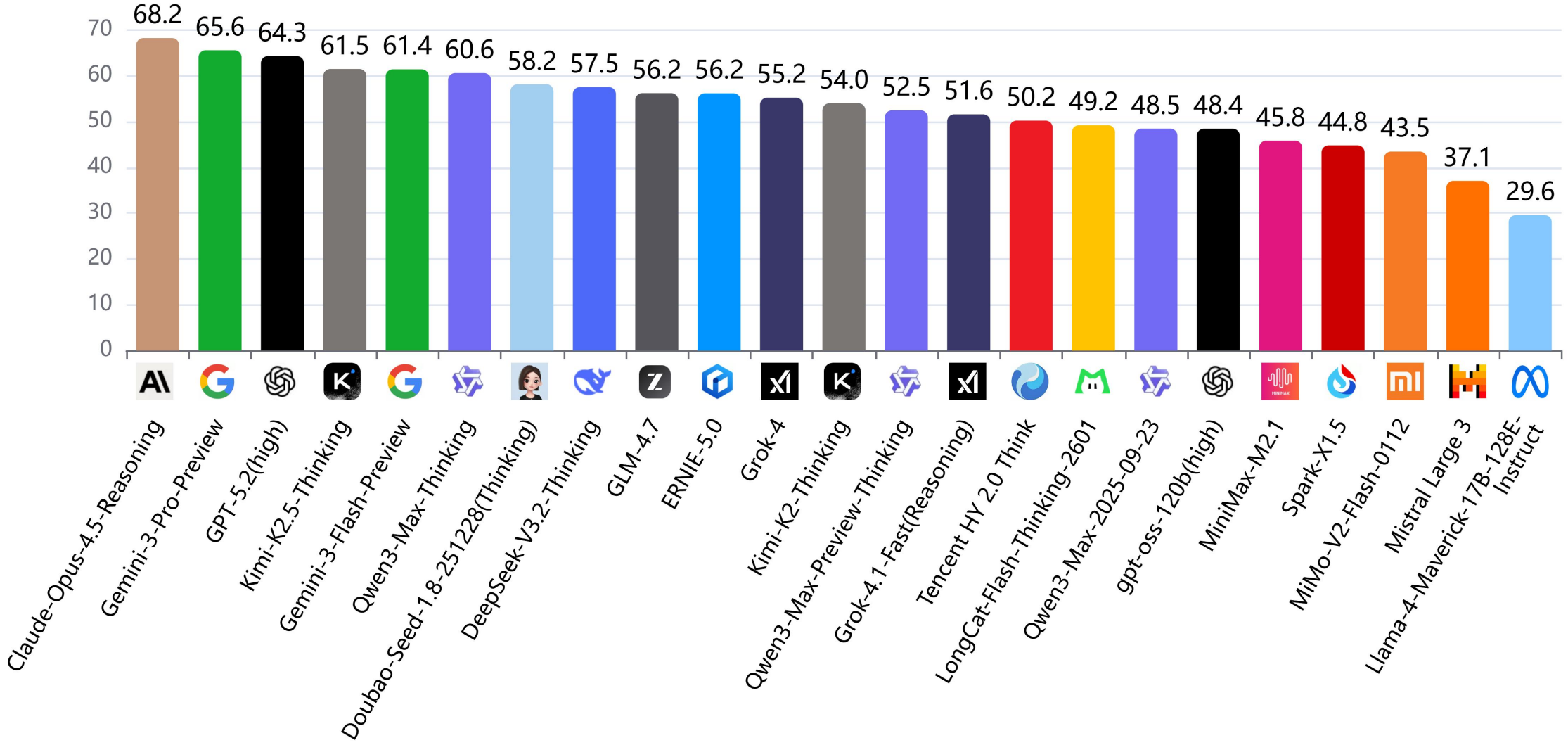
介绍：主要考察模型在执行中文生成任务时应对忠实性幻觉的能力。包括但不限于文本摘要、阅读理解、多文本问答和对话补全等基础语义理解与生成创作数据集。

评价方式：基于人工校验参考答案的、对每个句子是否存在幻觉进行0/1评估。

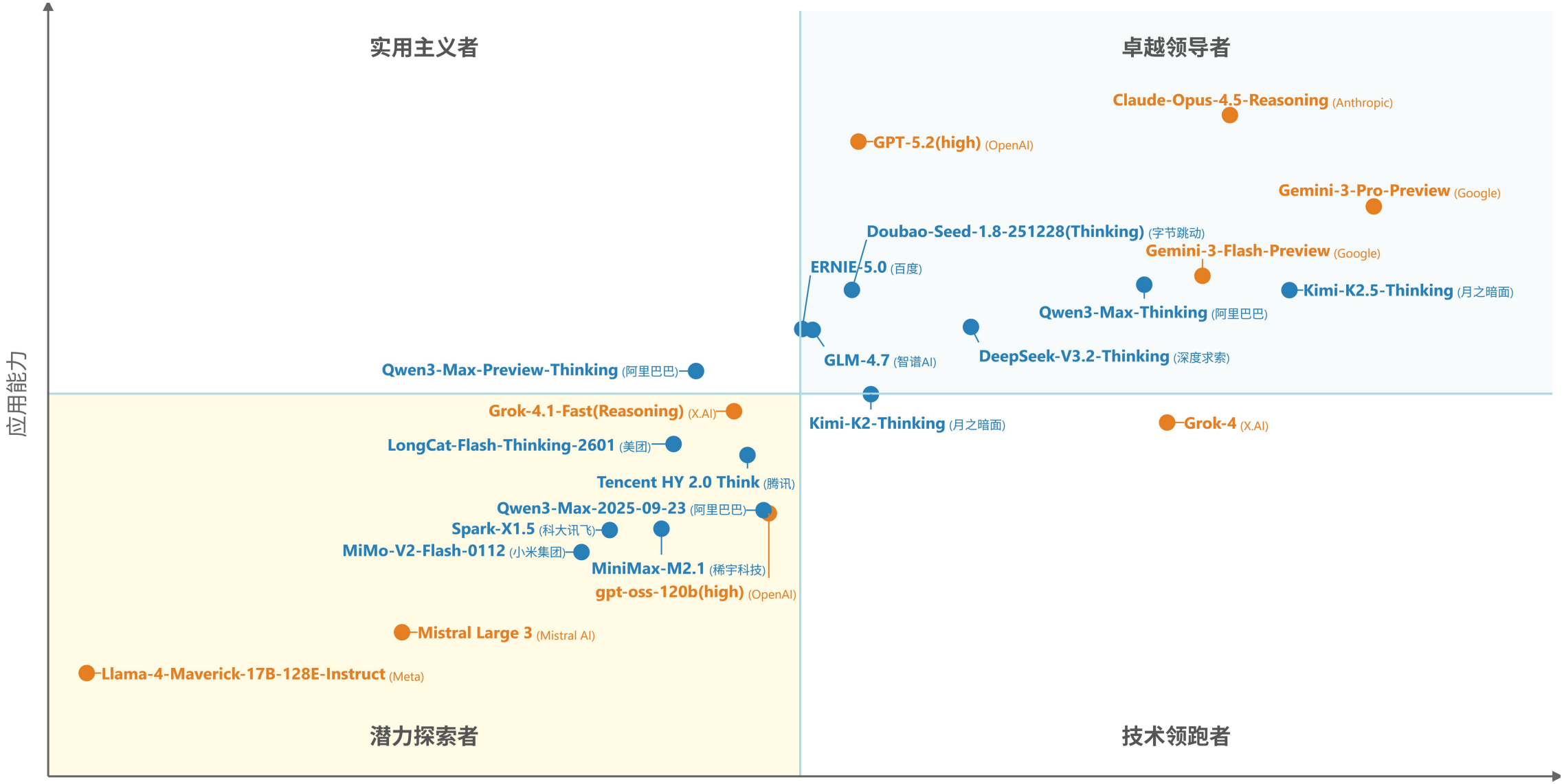
2025年全球大模型中文智能指数排行榜

本次测评包括六大任务：数学推理、科学推理、代码生成（含web开发）、智能体（任务规划）、幻觉控制、精确指令遵循。测评集共998道题，共测评23个国内外大模型，最终得分取各任务平均分。

SuperCLUE官网地址：SuperCLUE.ai



2025年SuperCLUE模型象限

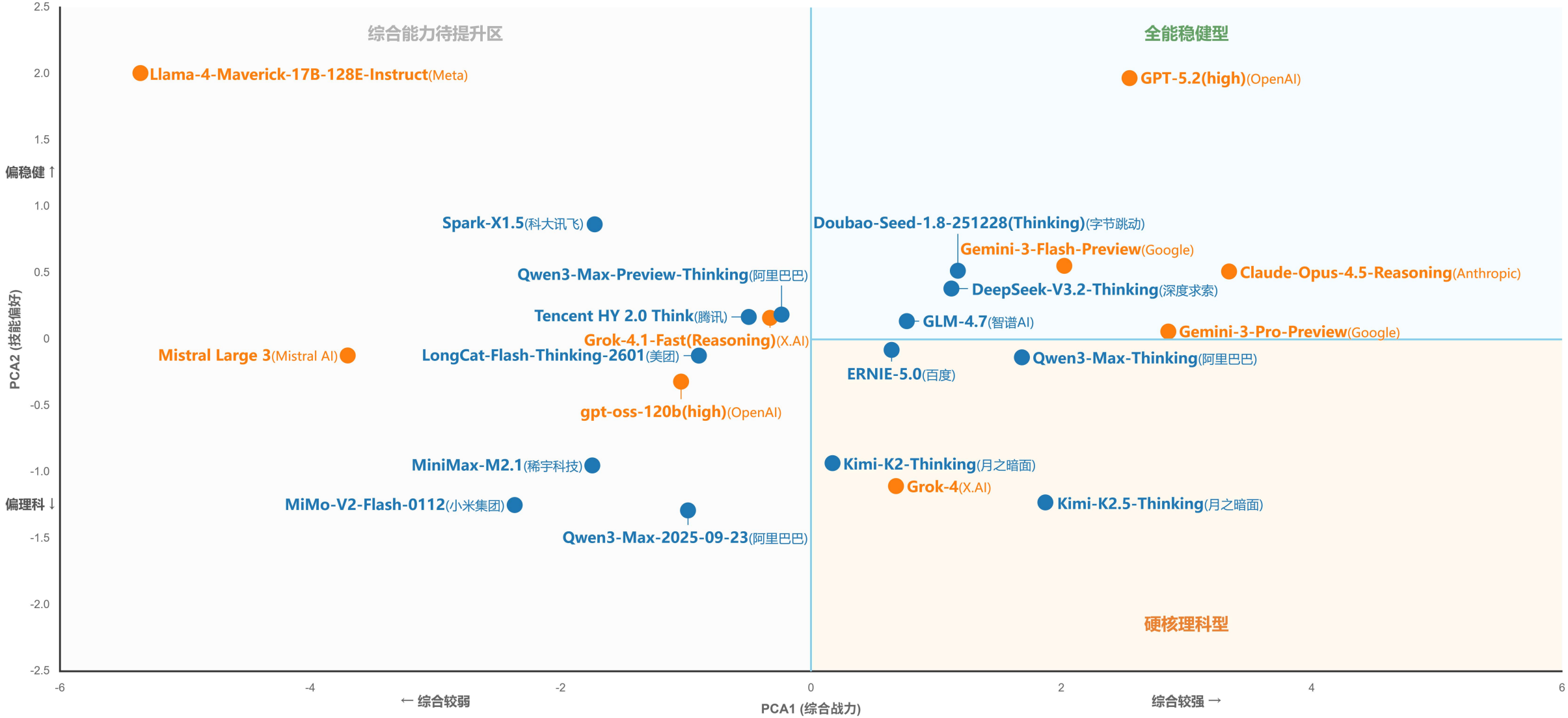


来源：SuperCLUE,2026年1月29日；

注：1. 两个维度的组成。推理能力包含：数学推理、科学推理、代码生成；应用能力包括：幻觉控制、精确指令遵循、智能体(任务规划)；
2. 四个象限的含义。它们代表大模型所处的不同阶段与定位，其中【潜力探索者】代表模型正在探索阶段未来拥有较大潜力；【技术领跑者】代表模型在基础技术方面具备领先性；【实用主义者】代表模型在场景应用深度上具备领先性；【卓越领导者】代表模型在基础和场景应用上处于领先位置，引领国内大模型发展。

2025年SuperCLUE模型能力格局

SuperCLUE
中文大模型综合性测评基准



注：1. 来源与计算：本图基于SuperCLUE2025年年度中文大模型基准测评数据，使用主成分分析(PCA)算法生成。将六大能力维度（数学推理、科学推理、代码生成、智能体(任务规划)、精确指令遵循、幻觉控制）的数据进行标准化处理（Z-Score）后，通过降维将高维信息投影至二维平面。

2. 坐标轴含义：**横轴**（PCA1）- 综合战力（绝对强弱）：代表模型的**整体实力**。越往右，综合得分越高，意味着该模型在大多数维度上都比左侧模型更强。**纵轴**（PCA2）- 技能偏好（相对侧重）：代表模型**自身技能树的结构差异**，而非绝对能力的短板。**上方**：相较于其自身水平，在"智能体-任务规划/防幻觉"等稳健型任务上表现更为突出。**下方**：相较于其自身水平，在"数学推理、科学推理"等理科任务上表现更为突出。

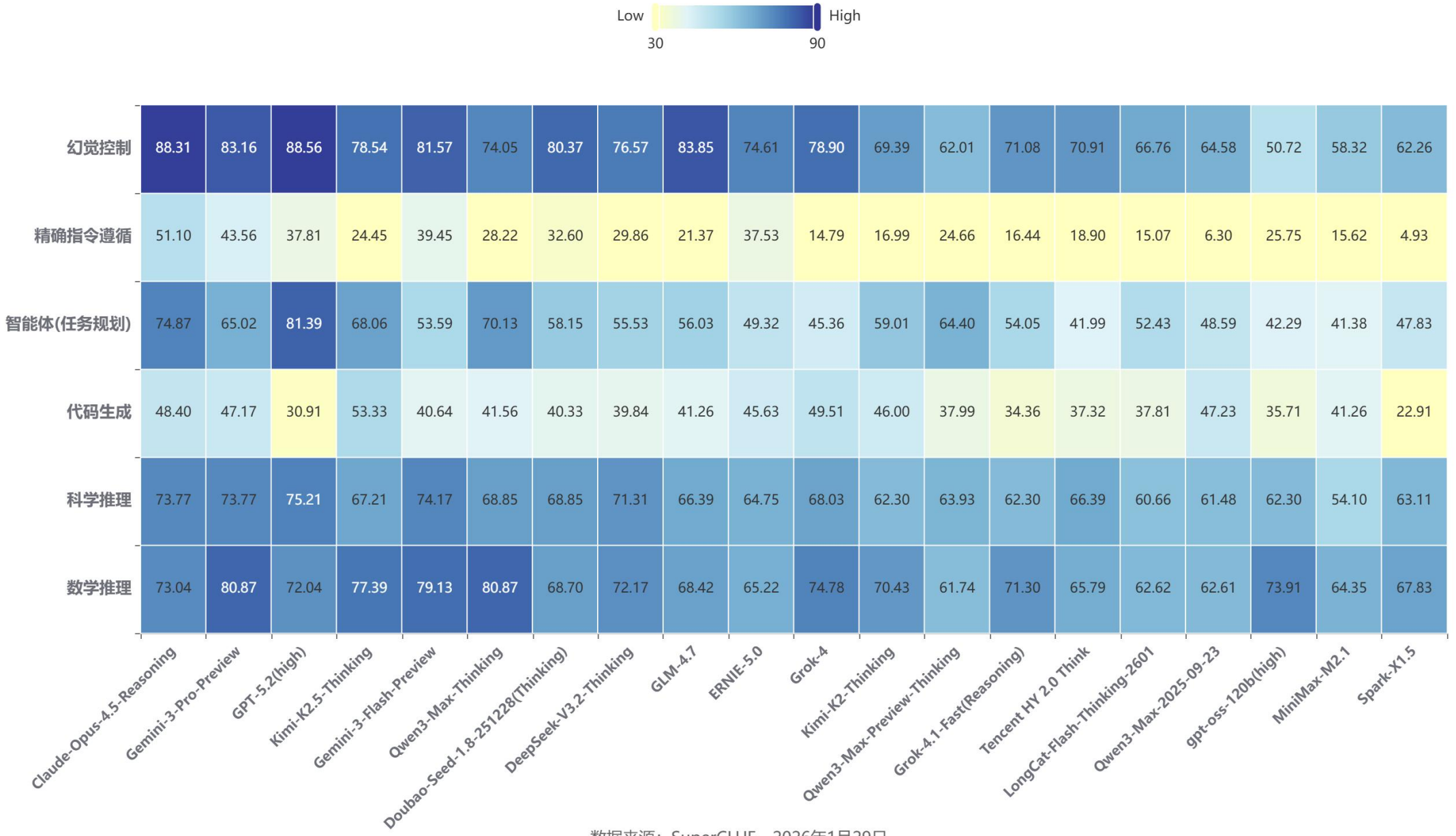
3. 区域解读：特别说明：坐标位置反映的是相对偏好。例如，右上角的模型虽然偏向稳健，但由于其处于右侧（综合强），其理科能力很可能依然强于左下方（综合弱且偏理科）的模型。

- 全能稳健型（右上）：综合顶级，且技能点更侧重于长链路规划与精准执行（精确指令遵循、智能体等）。
- 硬核理科型（右下）：综合顶级，且技能点更侧重于深度思考与逻辑计算（数学推理、代码生成等）。
- 综合能力待提升区（左侧）：整体各项能力仍有较大提升空间。

SuperCLUE2025年年度测评六大任务国内Top3

测评任务	海外第一	国内第一	国内第二	国内第三
数学推理	Gemini-3-Pro-Preview	Qwen3-Max-Thinking	Kimi-K2.5-Thinking	DeepSeek-V3.2-Thinking
科学推理	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
代码生成	Grok-4	Kimi-K2.5-Thinking	Qwen3-Max-2025-09-23	Kimi-K2-Thinking、 ERNIE-5.0
智能体(任务规划)	GPT-5.2(high)	Qwen3-Max-Thinking	Kimi-K2.5-Thinking	Qwen3-Max-Preview-Thinking
精确指令遵循	Claude-Opus-4.5-Reasoning	ERNIE-5.0	Doubao-Seed-1.8-251228(Thinking)	DeepSeek-V3.2-Thinking
幻觉控制	GPT-5.2(high)	GLM-4.7	Doubao-Seed-1.8-251228(Thinking)	Kimi-K2.5-Thinking

SuperCLUE2025年年度测评六大任务国内外Top20热力图



数据来源：SuperCLUE，2026年1月29日

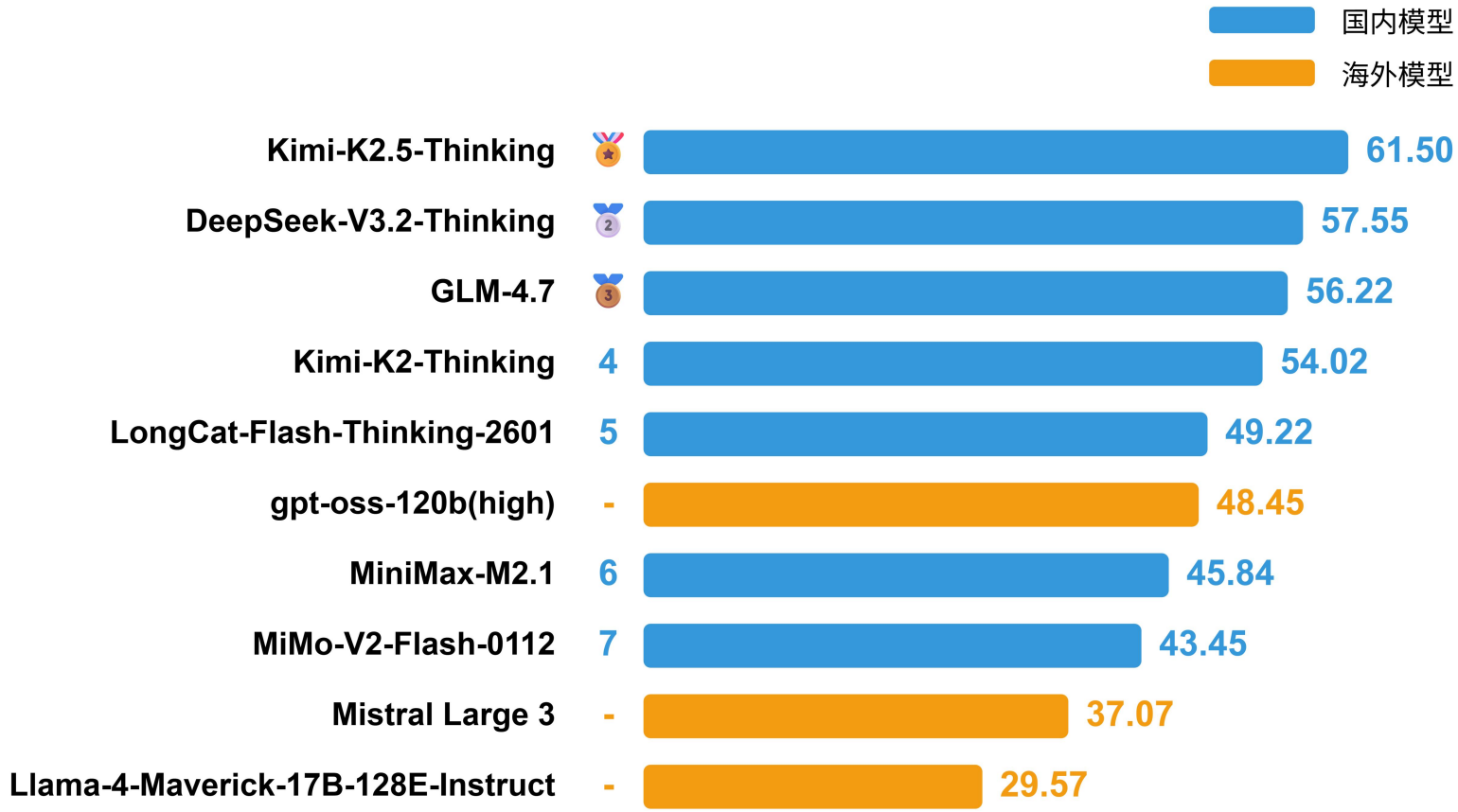
2025年年度中文大模型基准测评——总榜

SuperCLUE测评基准2025年年度总体表现（包括3个补测模型）

排名	模型名称	机构	开/闭源	总分	数学推理	幻觉控制	科学推理	精确指令遵循	代码生成	智能体（任务规划）	属地	使用方式
-	Claude-Opus-4.5-Reasoning	Anthropic	闭源	**68.25**	73.04	88.31	73.77	51.10	48.40	74.87	海外	API
-	Gemini-3-Pro-Preview	Google	闭源	**65.59**	80.87	83.16	73.77	43.56	47.17	65.02	海外	API
-	GPT-5.2(high)	OpenAI	闭源	**64.32**	72.04	88.56	75.21	37.81	30.91	81.39	海外	API
1	Kimi-K2.5-Thinking	月之暗面	开源	**61.50**	**77.39**	**78.54**	67.21	24.45	**53.33**	**68.06**	国内	API
-	Gemini-3-Flash-Preview	Google	闭源	**61.43**	79.13	81.57	74.17	39.45	40.64	53.59	海外	API
1	Qwen3-Max-Thinking	阿里巴巴	闭源	**60.61**	**80.87**	74.05	**68.85**	28.22	41.56	**70.13**	国内	API
2	Doubao-Seed-1.8-251228(Thinking)	字节跳动	闭源	**58.17**	68.70	**80.37**	**68.85**	**32.60**	40.33	58.15	国内	API
2	DeepSeek-V3.2-Thinking	深度求索	开源	**57.55**	**72.17**	76.57	**71.31**	**29.86**	39.84	55.53	国内	API
3	GLM-4.7	智谱AI	开源	**56.22**	68.42	**83.85**	66.39	21.37	41.26	56.03	国内	API
3	ERNIE-5.0	百度	闭源	**56.18**	65.22	74.61	64.75	**37.53**	45.63	49.32	国内	API
-	Grok-4	X.AI	闭源	**55.23**	74.78	78.90	68.03	14.79	49.51	45.36	海外	API
4	Kimi-K2-Thinking	月之暗面	开源	**54.02**	70.43	69.39	62.30	16.99	**46.00**	59.01	国内	API
5	Qwen3-Max-Preview-Thinking	阿里巴巴	闭源	**52.46**	61.74	62.01	63.93	24.66	37.99	**64.40**	国内	API
-	Grok-4.1-Fast(Reasoning)	X.AI	闭源	**51.59**	71.30	71.08	62.30	16.44	34.36	54.05	海外	API
6	Tencent HY 2.0 Think	腾讯	闭源	**50.22**	65.79	70.91	66.39	18.90	37.32	41.99	国内	API
7	LongCat-Flash-Thinking-2601	美团	开源	**49.22**	62.62	66.76	60.66	15.07	37.81	52.43	国内	API
7	Qwen3-Max-2025-09-23	阿里巴巴	闭源	**48.46**	62.61	64.58	61.48	6.30	**47.23**	48.59	国内	API
-	gpt-oss-120b(high)	OpenAI	开源	**48.45**	73.91	50.72	62.30	25.75	35.71	42.29	海外	API
8	MiniMax-M2.1	稀宇科技	开源	**45.84**	64.35	58.32	54.10	15.62	41.26	41.38	国内	API
9	Spark-X1.5	科大讯飞	闭源	**44.81**	67.83	62.26	63.11	4.93	22.91	47.83	国内	API
10	MiMo-V2-Flash-0112	小米集团	开源	**43.45**	61.74	60.24	48.36	3.01	40.52	46.84	国内	API
-	Mistral Large 3	Mistral AI	开源	**37.07**	45.22	51.08	51.64	8.77	33.37	32.34	海外	API
-	Llama-4-Maverick-17B-128E-Instruct	Meta	开源	**29.57**	32.17	57.30	48.36	5.48	13.85	20.29	海外	API

注：数据来源SuperCLUE，2026年1月29日；为减少波动影响，本次测评将相差1分内的模型视为并列。海外模型仅作对比参考，不参与排名。标红为国内前三。

SuperCLUE2025年年度基准测评开源模型总分对比



数据来源：SuperCLUE，2026年1月29日。

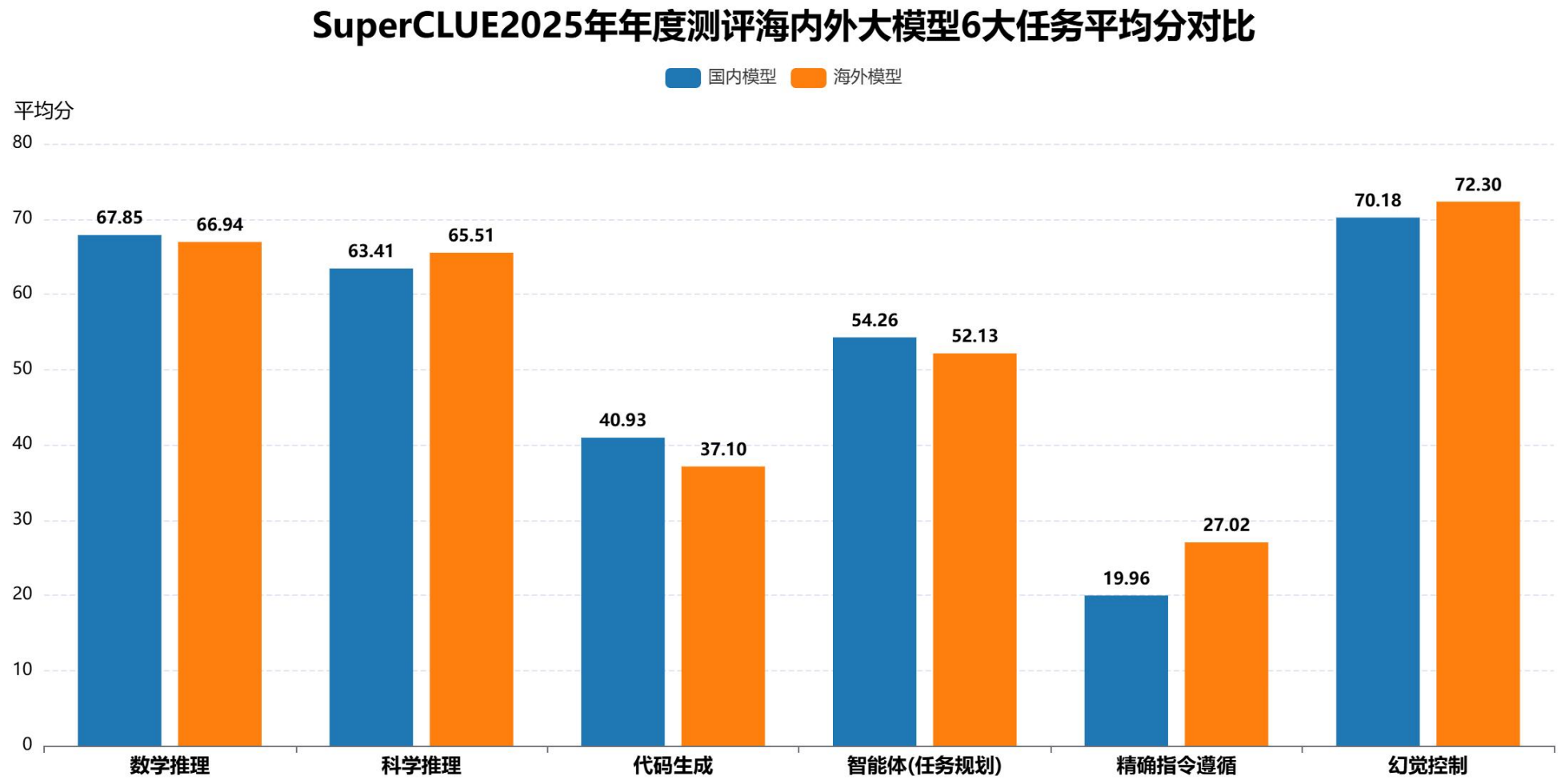
注：由于部分模型分数较为接近，为了减少问题波动对排名的影响，本次测评将相距1分区间的模型定义为并列。海外模型仅对比参考，不参与排名。

测评分析

国内开源模型全面领先海外开源模型。

开源模型榜单Top5均为国内模型，其中Kimi-K2.5-Thinking以61.50分取得开源第一，领先第二名近4分。DeepSeek-V3.2-Thinking和GLM-4.7跻身开源Top3，大幅领先海外最佳开源模型gpt-oss-120b(high)。

海内外大模型对比分析



数据来源：SuperCLUE，2026年1月29日。

测评分析

1. 推理能力整体已高度对齐。

在数学推理和科学推理两大任务上，海内外整体的平均分差距不大，整体的推理能力相当。具体而言，国内模型在数学推理有微弱领先，海外模型在科学推理领先较多，主要是头部模型的领先。

2. 国内模型在代码和智能体任务上整体表现更佳。

在代码生成和智能体(任务规划)两大任务上，国内模型平均超过海外两分以上。具体而言，在代码生成任务上，国内模型整体趋于中上游的位置，更有国内顶尖模型摘得桂冠；在智能体任务上，国内模型整体表现不俗，但国内头部模型与海外头部模型还存在一定的差距。

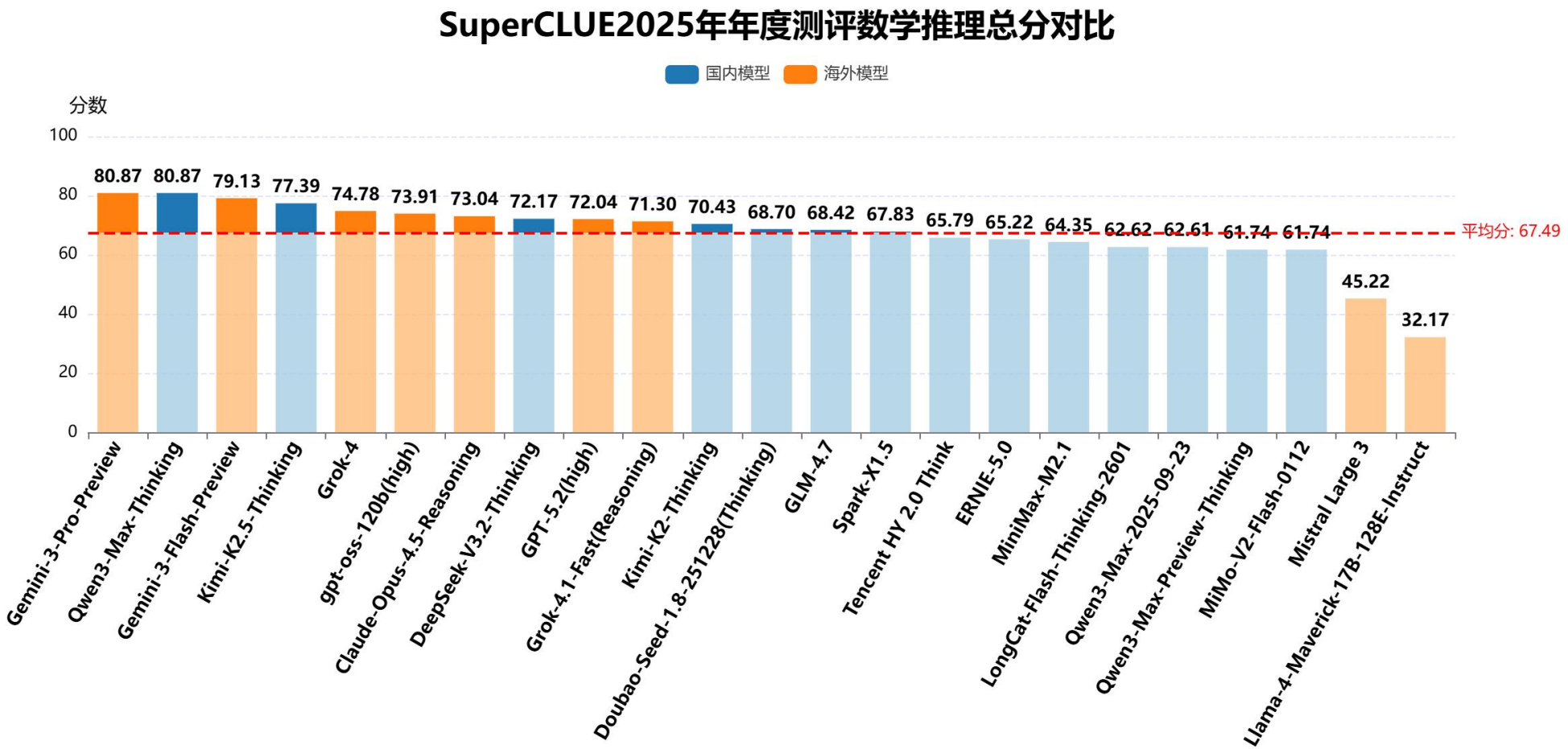
3. 精确指令遵循和幻觉控制是国内模型的短板。

国内模型在这两大任务上整体表现不如海外模型。具体而言，精确指令遵循是二者相差最大的维度，平均分差超过7分，幻觉控制平均分差近2分。

海内外大模型对比分析——数学推理任务

介绍：主要考察模型运用数学概念和逻辑进行多步推理和问题解答的能力。包括但不限于几何学、代数学、概率论与数理统计等竞赛级别数据集。

评价方式：基于参考答案的0/1评估，模型答案与参考答案一致得1分，反之得0分，不对回答过程进行评价。



数据来源：SuperCLUE，2026年1月29日。

测评分析

1. 国内头部模型追平。

国内Qwen3-Max-Thinking在数学推理任务上与Gemini-3-Pro-Preview均取得80.87分，并列全球第一。Kimi-K2.5-Thinking也以77.39分位居全球第四，体现了国内模型在数学推理领域的突破性进展。

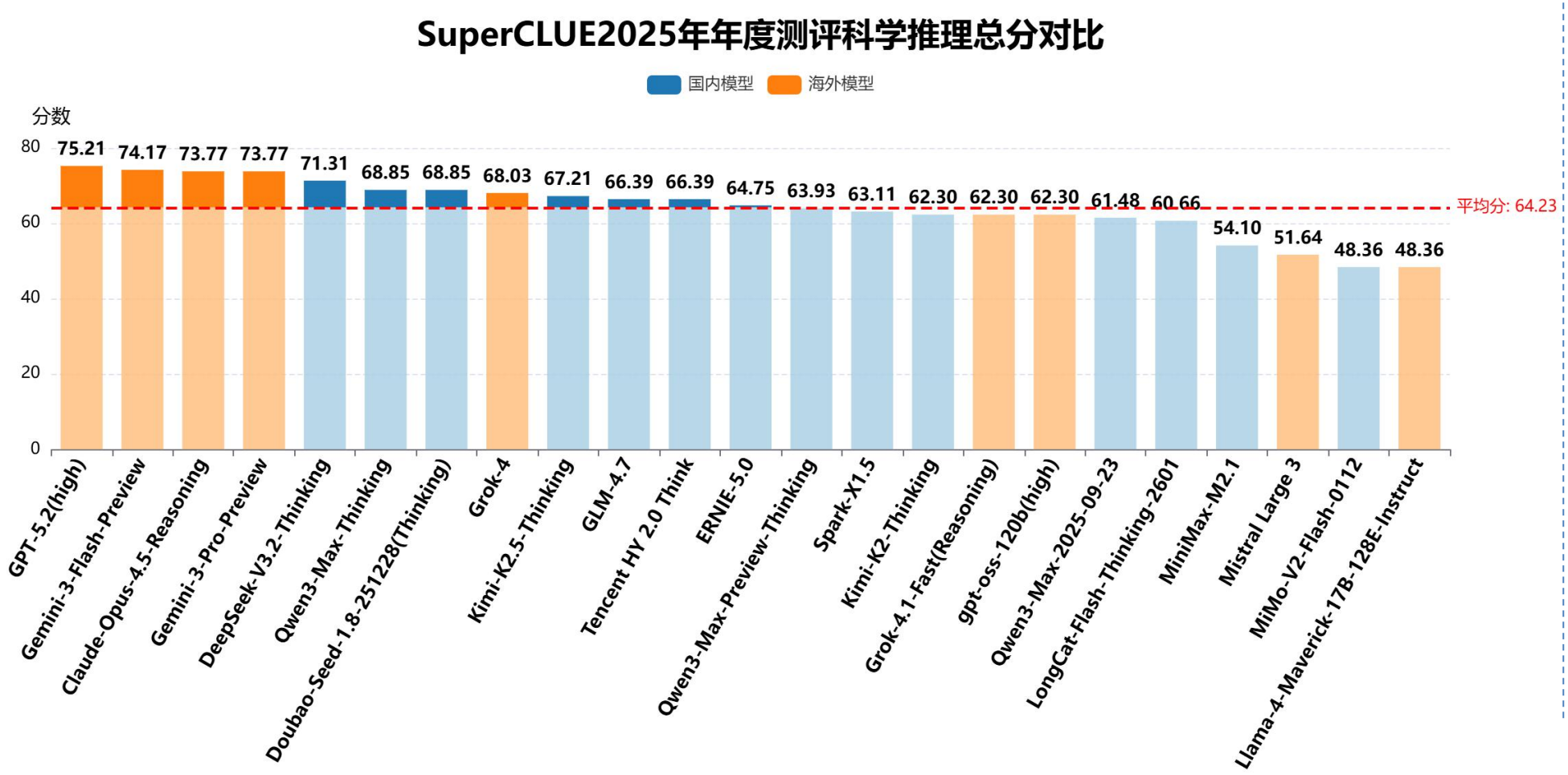
2. 国内整体梯队靠后。

数学推理Top10中，国内模型仅占4席，海外模型密度更高；国内大多数模型集中在平均分（67.49分）附近的后半段，与国际顶尖模型差距不小。

海内外大模型对比分析——科学推理任务

介绍： 主要考察模型在跨学科背景下理解和推导因果关系的能力。包括物理、化学、生物等在内的研究生级别科学数据集。

评价方式： 基于参考答案的0/1评估，模型答案与参考答案一致得1分，反之得0分，不对回答过程进行评价。



数据来源：SuperCLUE，2026年1月29日。

测评分析

1.海外头部垄断。

在科学推理任务中，海外模型包揽了前四席，分别是GPT-5.2(high)(75.21分)、Gemini-3-Flash-Preview(74.17分)、Claude-Opus-4.5-Reasoning(73.77分)和Gemini-3-Pro-Preview(73.77分)，国内仅有DeepSeek-V3.2-Thinking进入前五名，Qwen3-Max-Thinking和Doubao-Seed-1.8-251228(Thinking)紧随其后。

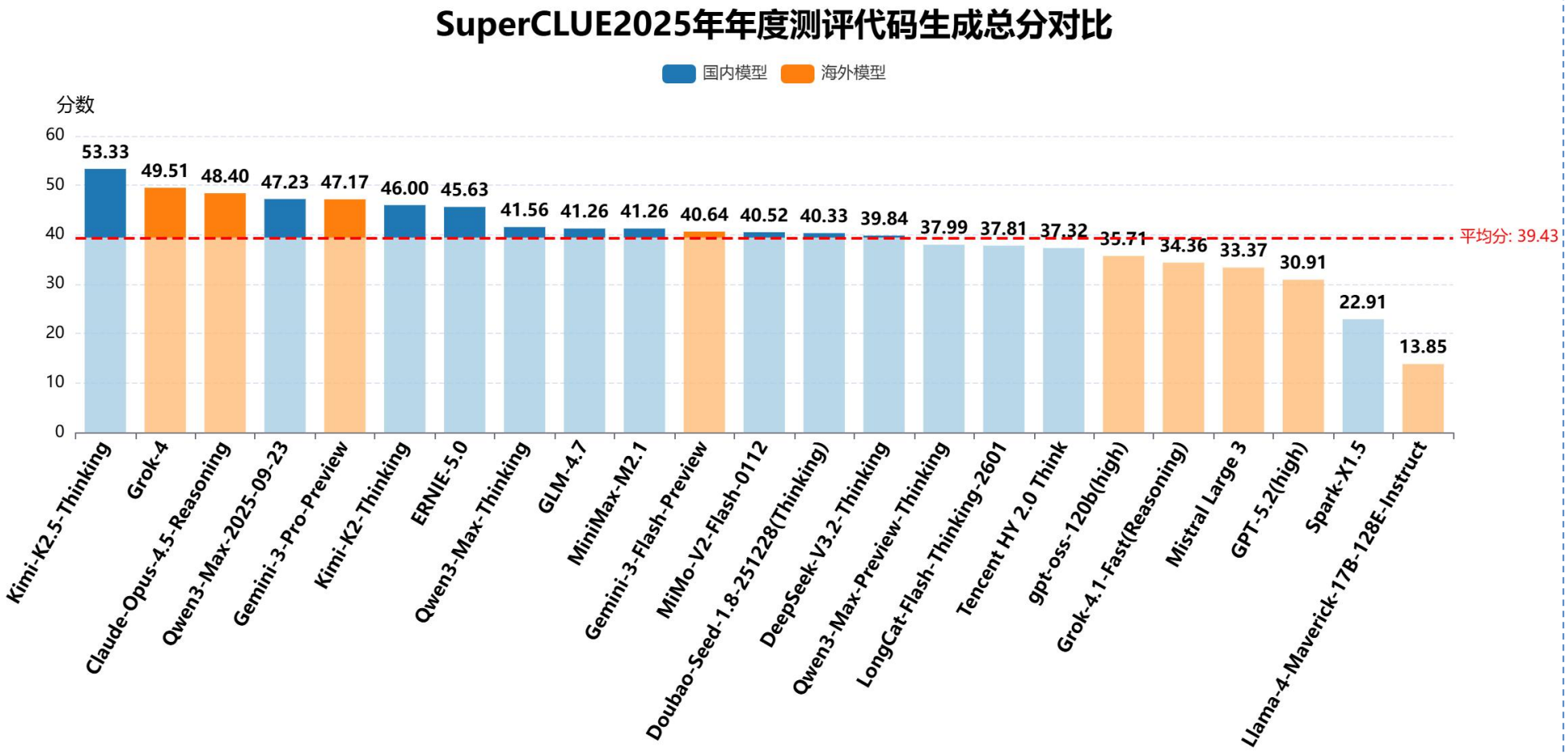
2. 国内整体分布重心偏向中部。

相较于在数学推理任务上国内模型整体偏后，在科学推理任务上，国内整体更偏向于中部的位置，国内大多数模型均分布于平均线附近。

海内外大模型对比分析——代码生成任务

介绍：该任务分为两大类型：一是独立功能函数生成，生成覆盖数据结构、算法等领域的独立函数。二是Web应用生成，要求模型构建旅游订票、电商、社交媒体等完整的交互式网站。

评价方式：通过单元测试进行0/1评分（独立功能函数生成）；通过模拟用户交互的功能测试进行0/1评分（Web应用生成）。



数据来源：SuperCLUE，2026年1月29日。

测评分析

1. 国产模型表现亮眼。

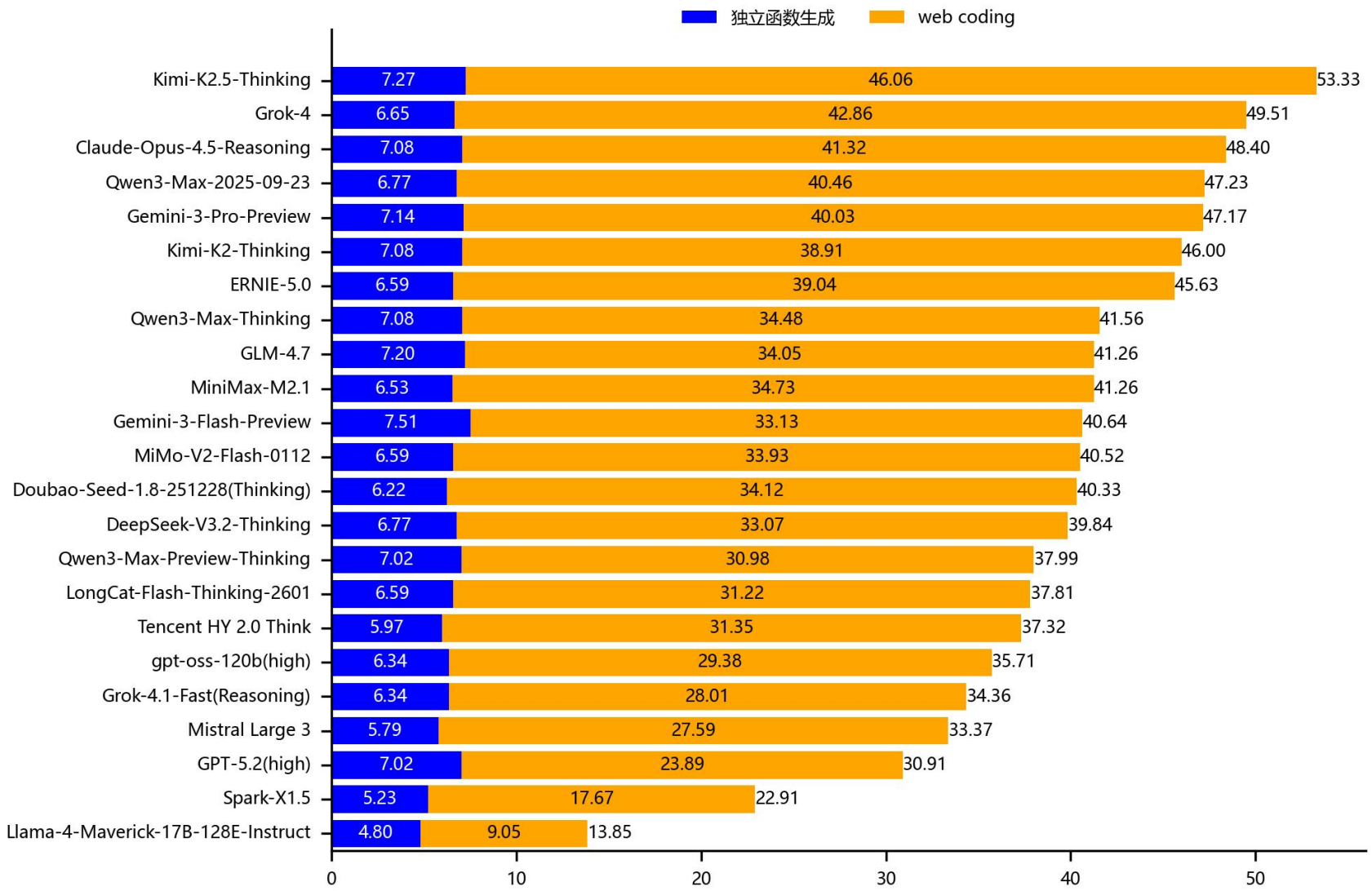
国产开源模型Kimi-K2.5-Thinking以53.33分位居全球第一，超越Grok-4、Claude-Opus-4.5-Reasoning等一众海外顶尖模型，Qwen3-Max-2025-09-23也以47.23分跻身Top5。说明国产头部模型已经在代码生成（独立功能函数生成和Web Coding）领域实现了从追赶到齐平甚至微弱领先的跨越。

2. GPT-5.2(high)的“滑铁卢”

GPT-5.2(high)在代码生成任务中仅有30.91分，排名倒数第三，这与我们的测评机制有关，我们每题设置了最大推理时长（30分钟），超过该时长还会有两次重试机会。如果三次机会全部超时未获取到答案，那么该题将被记0分。GPT-5.2(high)由于推理时间过长，存在许多超时被记0分的题目。

海内外大模型对比分析——代码生成任务

代码生成两大子任务加权后在总成绩中的占比



数据来源：SuperCLUE, 2026年1月29日。

注：1. 代码生成任务最终得分由独立函数生成任务与Web Coding任务加权后得到，各子任务的权重 = 各子任务的测试用例数量 / 总测试用例数量；

2. 蓝色条形中间的数字为独立函数生成任务加权后的分数，橙色条形中间的数字为Web Coding任务加权后的分数，条形右边的数字为两大子任务求和得到的总分。

（图中展示时子任务的分数会保留两位小数，求和时会产生累计误差，但总分按照所有通过测试用例的数量除以总测试用例数量进行计算，不会产生累计误差）

测评分析

1. Web Coding：国产模型的超车区。

Kimi-K2.5-Thinking在Web Coding子任务上取得了46.06的高分，位居第一，领先第二名3.2分，是其总分跃居榜首的关键因素，这与其原生的多模态架构设计息息相关。

其他国产头部模型如Qwen3-Max-2025-09-23、ERNIE-5.0在Web Coding子任务上与国际顶尖模型（Grok-4、Claude-Opus-4.5-Reasoning）分差均在3分左右，差距较小。

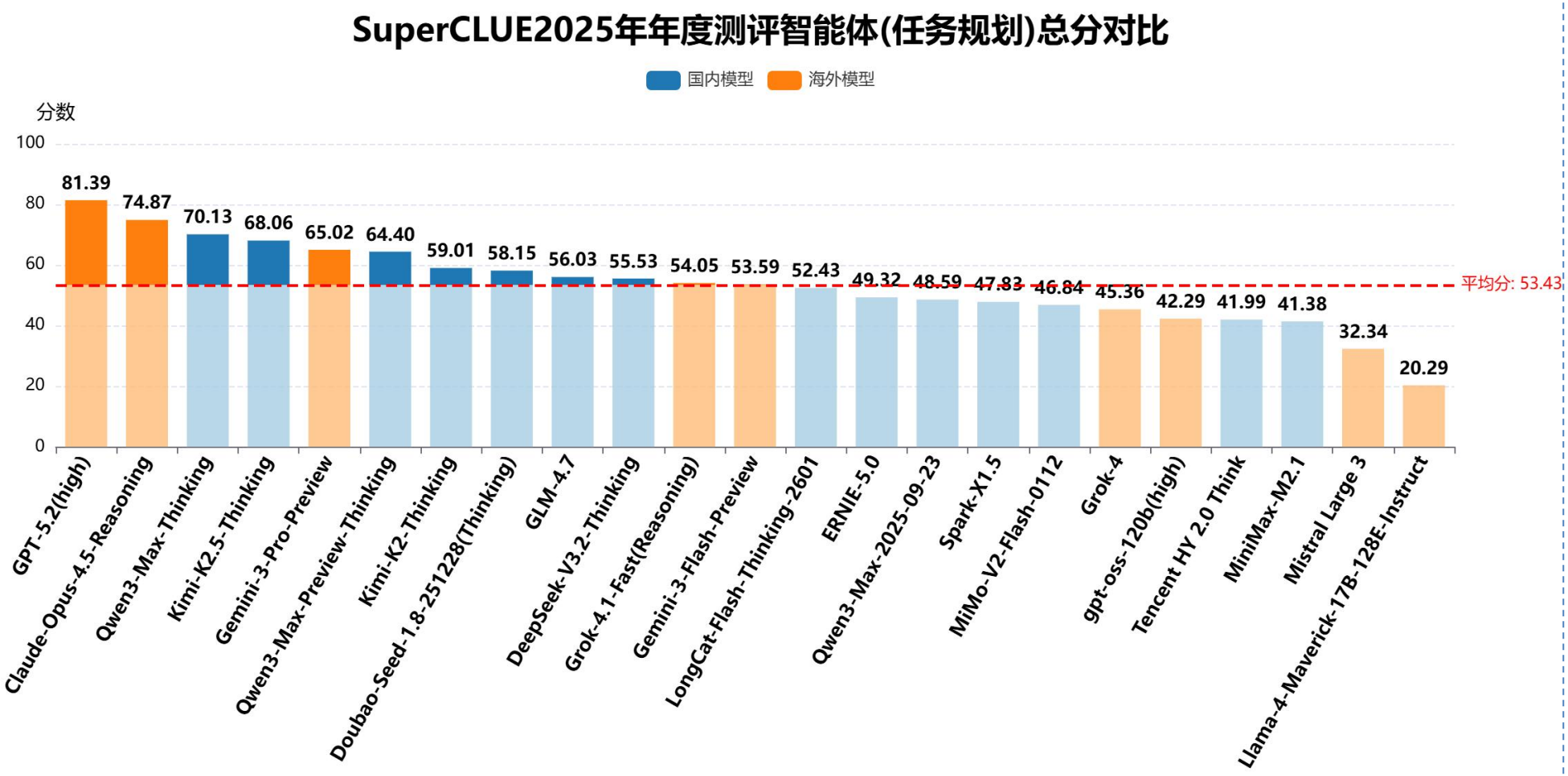
2. Web Coding子任务的区分度显著更高。

所有模型在独立函数生成子任务上的差距并不显著，标准差仅有0.66，但在Web Coding子任务上的标准差达到了8.23，是拉开模型在代码生成任务上差距的主要原因。

海内外大模型对比分析——智能体(任务规划)

介绍：主要考察模型在复杂任务场景中制定结构化行动方案的能力，包括且不限于生活服务、工作协作、学习成长、健康医疗等。要求模型基于给定目标和约束条件，生成逻辑连贯、步骤清晰、可执行的行动计划。

评价方式：利用裁判模型根据行动方案对预设检查点的完成情况进行离散判定（0/1），或对方案整体质量进行连续评分（0–100）。



数据来源：SuperCLUE，2026年1月29日。

测评分析

1. 海外头部模型优势显著。

海外头部模型GPT-5.2(high)以81.39分领跑榜单，Claude-Opus-4.5-Reasoning以74.87分紧随其后。国内Qwen3-Max-Thinking（70.13分）和Kimi-K2.5-Thinking（68.06分）分居第三和第四，海内外头部模型的差距超过10分，国产模型在任务规划领域还有一定的进步空间。

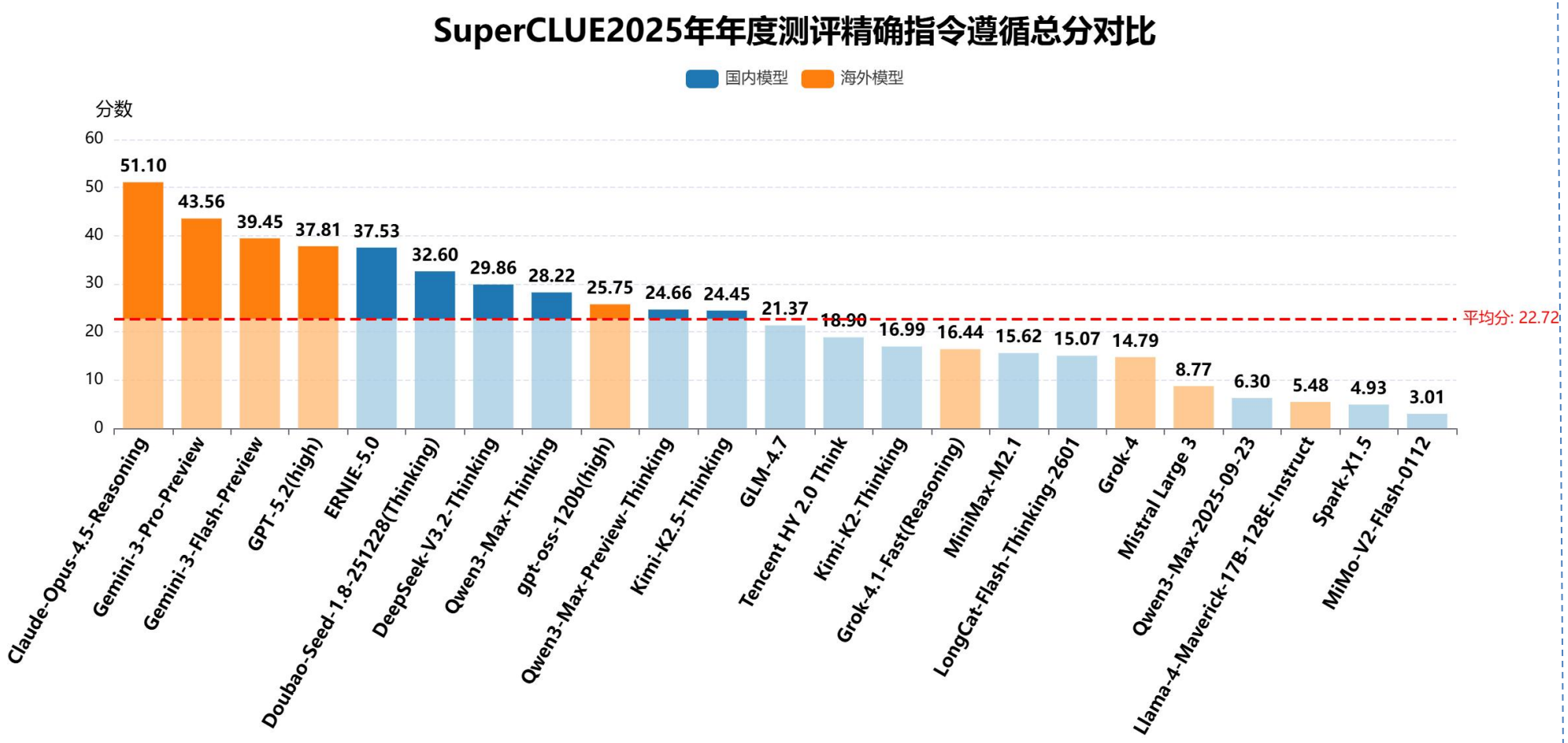
2. 行业整体水平跨度巨大，长尾效应明显。

当前大模型在智能体(任务规划)方面的发展极不平衡，整体水平跨度巨大，两极分化比较严重，最高分和最低分相差了4倍之多。此外，该任务的标准差是六大任务中最大的，达到了13.78，说明该任务对于当前大模型依然是极具挑战性的，是区分大模型能力的关键。

海内外大模型对比分析——精确指令遵循任务

介绍：主要考察模型的指令遵循能力，包括但不限于定义的输出格式或标准来生成响应，精确地呈现要求的数据和信息。涉及的中文场景包括但不限于结构约束、量化约束、语义约束、复合约束等不少于4个场景。

评价方式：基于规则脚本的0/1评估。



数据来源：SuperCLUE，2026年1月29日。

测评分析

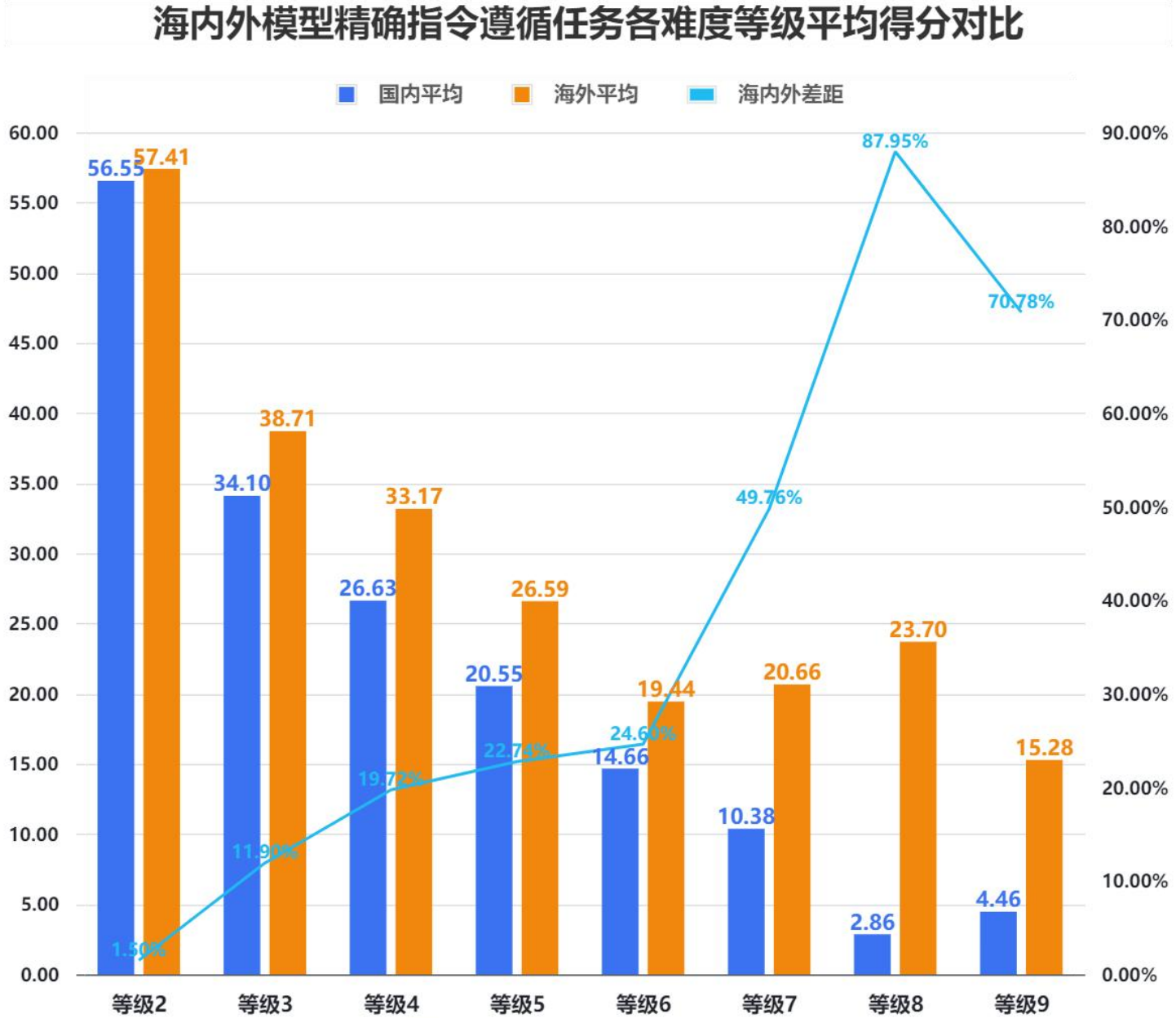
1. 梯度分化明显，海外头部领先显著。

前四名均为海外模型且分差较大，Claude-Opus-4.5-Reasoning以51.10分位居榜首，与第二名拉开了近8分的差距。国产模型ERNIE-5.0、Doubao-Seed-1.8-251228(Thinking)等紧随其后，与第一名差距超过13分。国内模型在该任务上的平均分为19.97分，海外模型的平均分为27.02分，相差近7分，还有一定的提升空间。

2. 整体表现有待提升，两极分化严重。

精确指令遵循任务涵盖8个难度等级，难度从等级2到等级9（等级2代表该题有两个指令，依此类推），整体题目难度较大，所有模型在该任务上的整体平均分仅22.72分，超过一半的模型未达到平均水平。

海内外大模型对比分析——精确指令遵循任务



数据来源：SuperCLUE，2026年1月29日。

测评分析

1. 难度与得分的“负相关”显著。

从难度等级2到等级9，海内外模型整体上的得分呈现出极其明显的指数级下降趋势，在高难度等级（L7-L9）的得分甚至来到了个位数。这说明目前的模型在面对极高复杂度、多重嵌套约束的任务时，很难完全满足用户的所有要求。

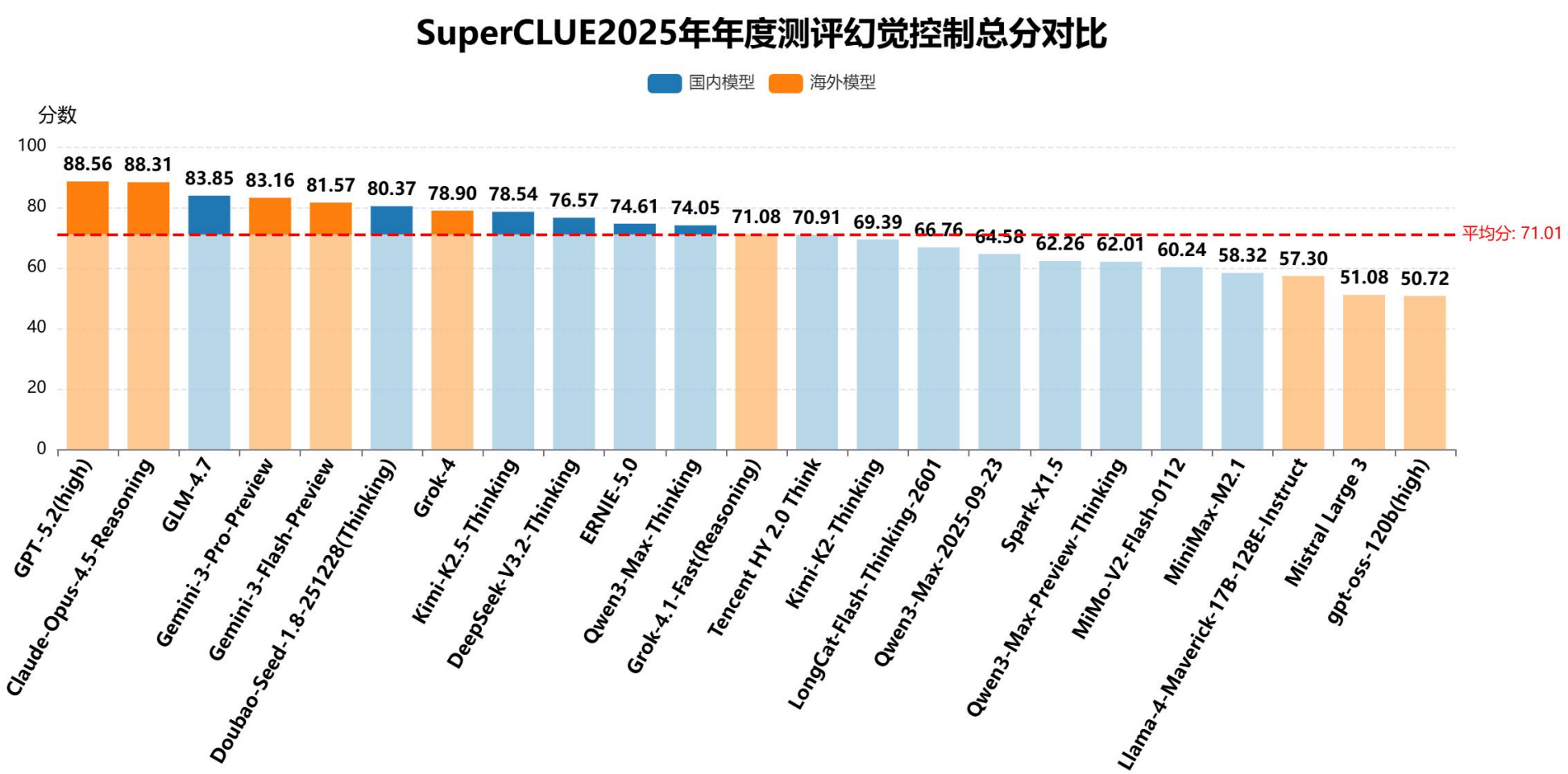
2. 就鲁棒性而言，海外模型整体优于国内模型。

整体来看，随着指令数量的递增，海内外模型的差距在逐步拉大。具体而言，在低难度等级（L2-L6）上，海内外模型的差距整体上比较稳定，差距均在25%以内。但从等级7开始，海内外模型的差距随着指令的增加，差距显著拉大。

指令越多，难度越高，海外模型的鲁棒性就越强，甚至在等级7和等级8，海外模型还出现了分数递增的情况，而国内模型几乎呈现完全递减的趋势。

海内外大模型对比分析——幻觉控制任务

介绍：主要考察模型在执行中文生成任务时应对忠实性幻觉的能力。包括但不限于文本摘要、阅读理解、多文本问答和对话补全等基础语义理解与生成创作数据集。
评价方式：基于人工校验参考答案的、对每个句子是否存在幻觉进行0/1评估。



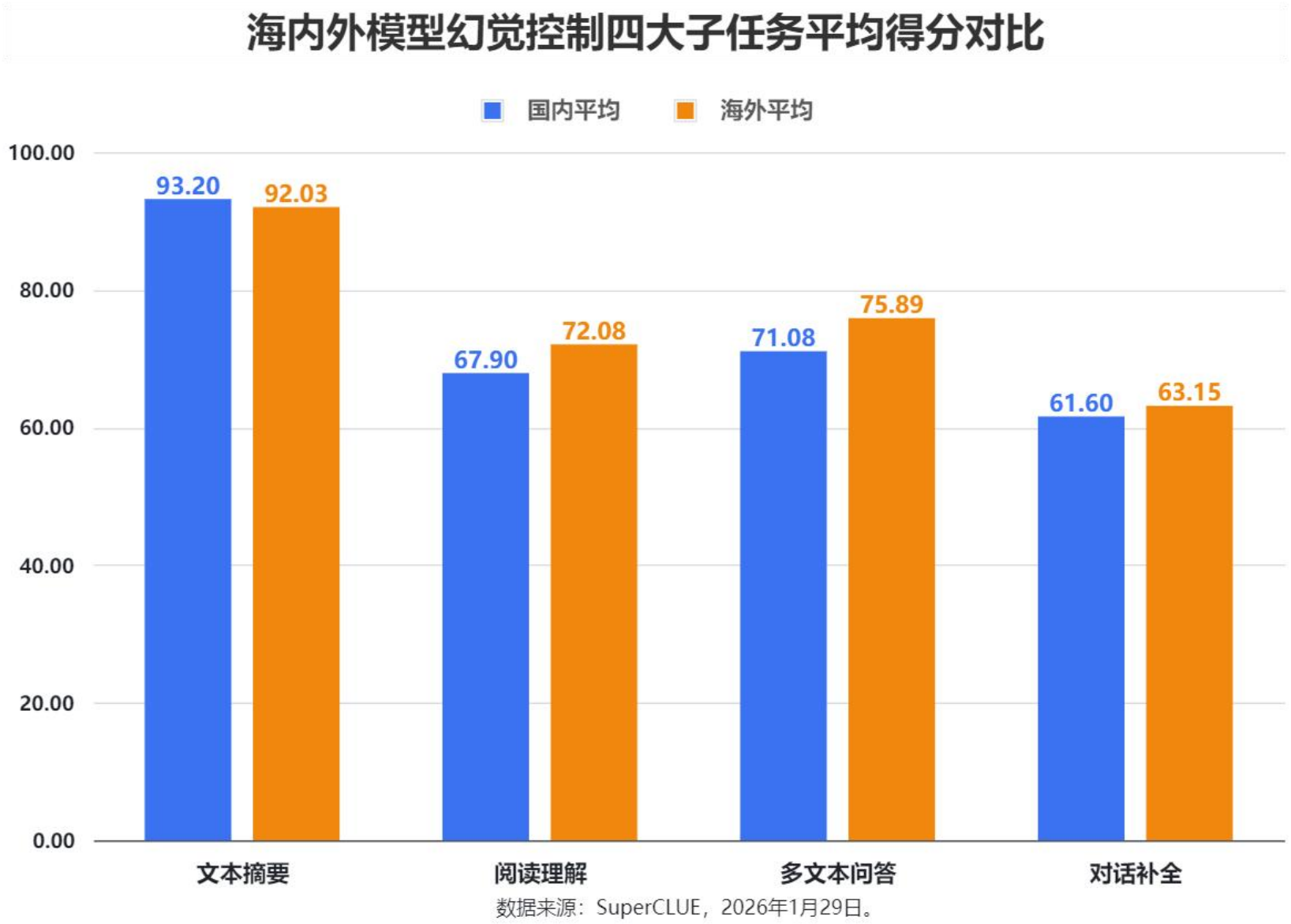
数据来源：SuperCLUE，2026年1月29日。

测评分析

海外头部模型占据显著优势，国内头部模型已有突破。

GPT-5.2(high)（88.56分）和Claude-Opus-4.5-Reasoning（88.31分）以17分以上的优势领先平均水平，展现出海外第一梯队模型在幻觉控制上的统治力。值得关注的是，GLM-4.7以83.85分跻身榜单Top3，与海外第一梯队差距缩小至5分以内；此外，Doubao-Seed-1.8-251228(Thinking)也有超过80分的不错表现，领先Grok-4，媲美国际顶尖模型Gemini-3-Flash-Preview。

海内外大模型对比分析——幻觉控制任务



数据来源：SuperCLUE，2026年1月29日。

测评分析

随着任务从“信息整合”向“开放生成”过渡，国内外大模型在幻觉控制上的得分都呈现出明显的下降趋势。

文本摘要是幻觉控制最容易的任务，得分最高，因为该任务强依赖于给定的原文，模型的任务是压缩和转述，而非创造。

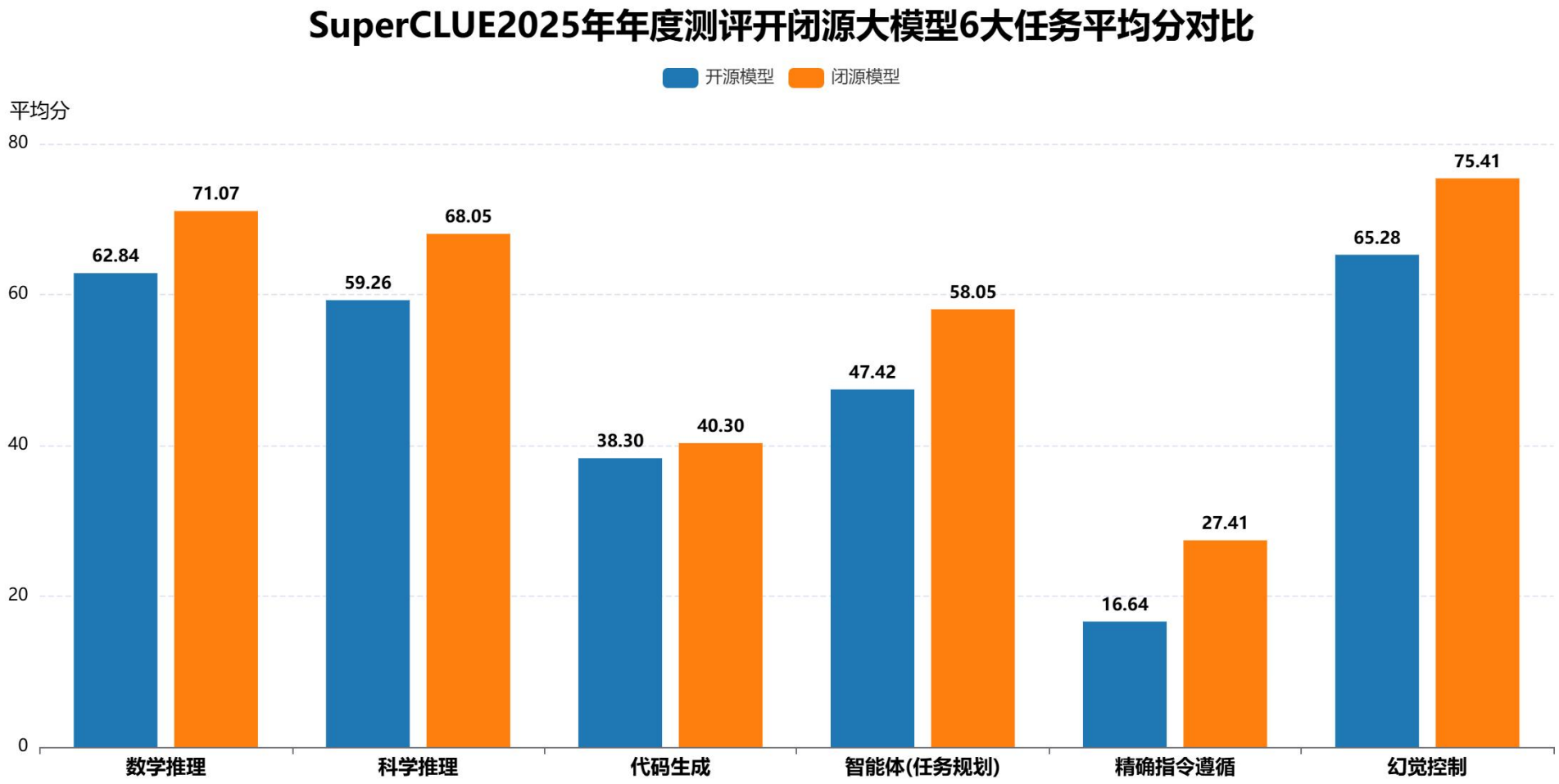
阅读理解虽然也基于给定文本，但该任务要求模型进行一定程度的推理和判断，而不仅仅是复述。这个推理过程为产生幻觉提供了空间，导致分数低于文本摘要。

多文本问答任务的挑战在于模型需要整合、比较、甚至解决多个信息源之间的冲突。信息源的增加和复杂性的提升，显著增加了模型混淆信息、错误归因的风险，从而导致幻觉。

对话补全任务是开放式和创造性的，模型往往需要根据上下文自行补充信息来使对话流畅地进行下去。这种高度的自由度也为事实性错误和无中生有的幻觉内容创造了条件。

任务越是开放，越是需要模型进行创造性生成，模型就越容易产生幻觉。

开闭源大模型对比分析



数据来源：SuperCLUE，2026年1月29日。

测评分析

1. 闭源模型全方位领先。

总体来看，在本次测评的六大任务中闭源模型的平均分均高于开源模型，尽管开源模型发展迅速，但在顶尖性能和特定复杂任务上，闭源模型依然保持着明显的领先优势。如智能体(任务规划)、精确指令遵循、幻觉控制三大维度均有超过10分的差距。

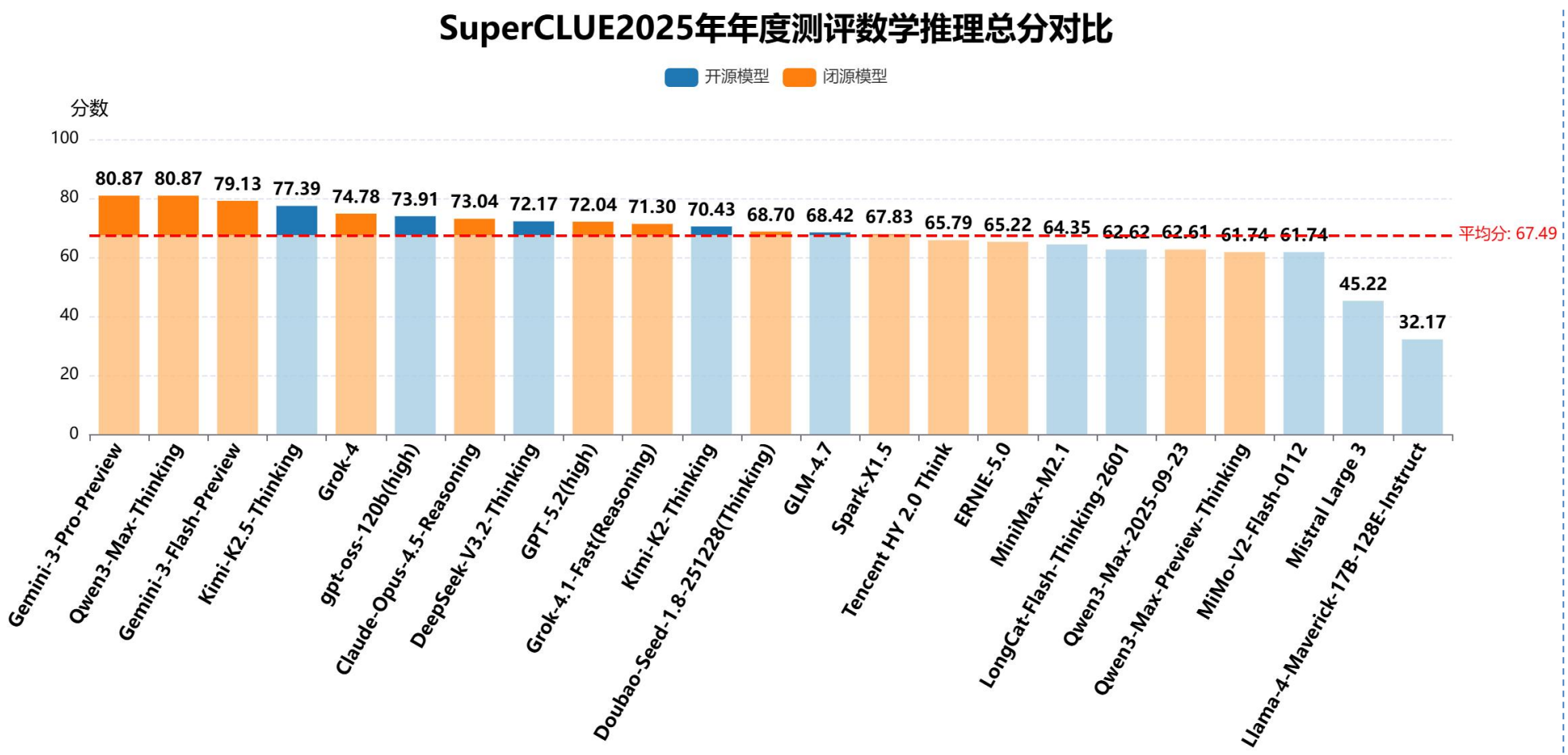
2. 开源模型在推理能力继续追赶，代码已实现单点突破。

开源模型在数学推理和科学推理两大任务上继续追赶闭源模型；在代码生成任务上的表现与闭源模型的差距比较小，仅两分左右，这与众多开源模型针对代码领域进行重点优化存在一定的关联。

开闭源大模型对比分析——数学推理任务

介绍：主要考察模型运用数学概念和逻辑进行多步推理和问题解答的能力。包括但不限于几何学、代数学、概率论与数理统计等竞赛级别数据集。

评价方式：基于参考答案的0/1评估，模型答案与参考答案一致得1分，反之得0分，不对回答过程进行评价。



数据来源：SuperCLUE，2026年1月29日。

测评分析

1. 闭源模型依然处于领先位置。

数学推理任务的Top3由闭源模型Gemini-3-Pro-Preview、Qwen3-Max-Thinking和Gemini-3-Flash-Preview占据，Top10中也仅有3个开源模型。

开源模型平均分仅为62.84分，闭源模型平均分为71.07分，差距近8分，开源模型整体上赶超闭源模型还存在一定的距离。

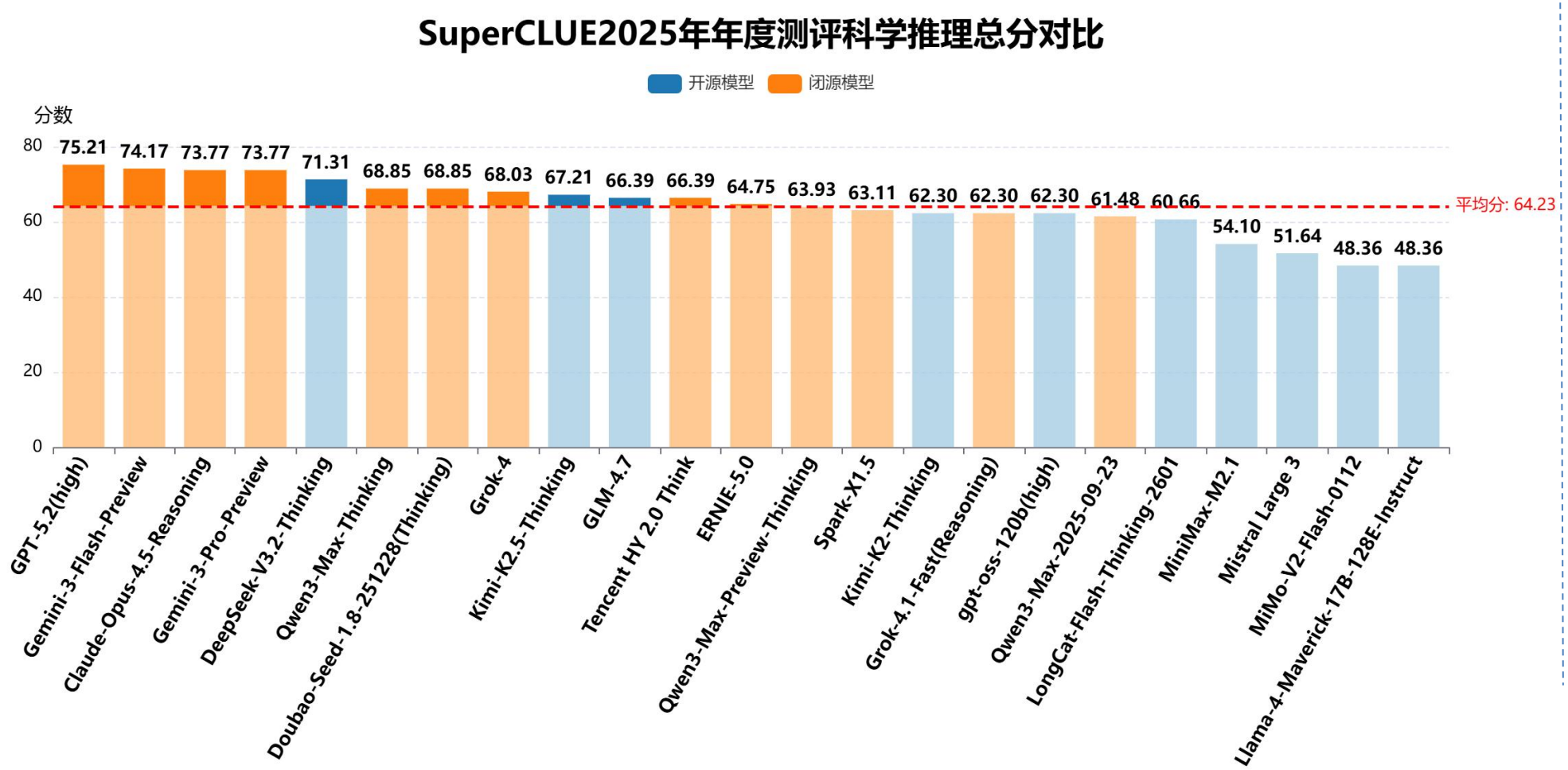
2. 头部开源模型表现强劲，有赶超顶尖闭源模型的趋势。

国产开源模型Kimi-K2.5-Thinking以77.39分位于全球第四，超过了包括Grok-4、Claude-Opus-4.5-Reasoning、GPT-5.2(high)等顶尖闭源模型。

开闭源大模型对比分析——科学推理任务

介绍：主要考察模型在跨学科背景下理解和推导因果关系的能力。包括物理、化学、生物等在内的研究生级别科学数据集。

评价方式：基于参考答案的0/1评估，模型答案与参考答案一致得1分，反之得0分，不对回答过程进行评价。



数据来源：SuperCLUE，2026年1月29日。

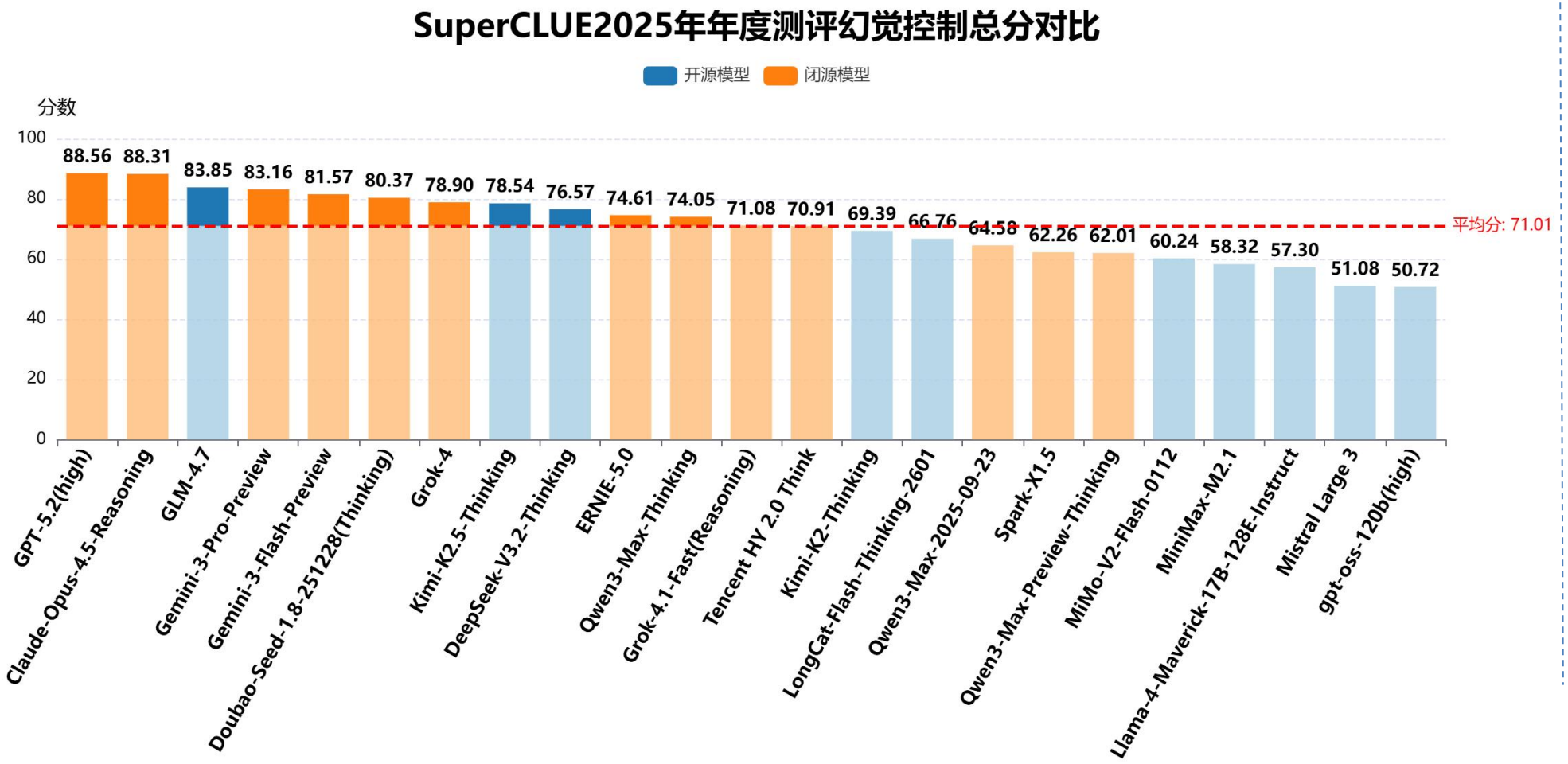
测评分析

闭源模型优势显著。

前四名GPT-5.2(high)、Gemini-3-Flash-Preview、Claude-Opus-4.5-Reasoning和Gemini-3-Pro-Preview均为闭源模型，仅有国产开源模型DeepSeek-V3.2-Thinking进入Top5。开源模型平均分为59.26分，闭源模型则达到了68.05分，分差近9分，说明在科学推理任务上，开闭源还存在着一定的差距。大多数开源模型处于榜单的中后部，大部分开源模型尚未达到平均水平。

开闭源大模型对比分析——幻觉控制任务

介绍：主要考察模型在执行中文生成任务时应对忠实性幻觉的能力。包括但不限于文本摘要、阅读理解、多文本问答和对话补全等基础语义理解与生成创作数据集。
评价方式：基于人工校验参考答案的、对每个句子是否存在幻觉进行0/1评估。



数据来源：SuperCLUE，2026年1月29日。

测评分析

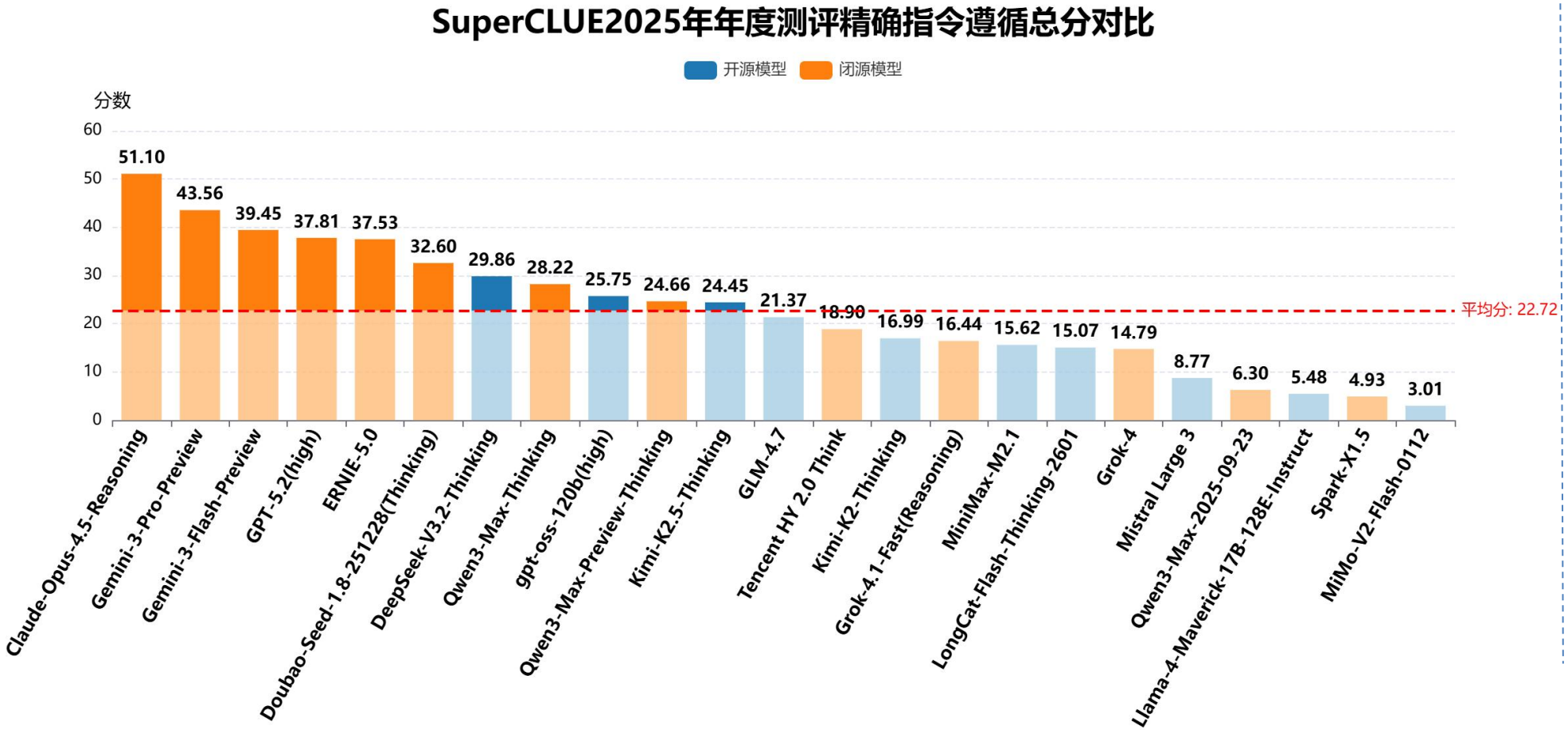
闭源模型展现出更强的可靠性。

榜单前两名均由闭源模型（GPT-5.2(high)和Claude-Opus-4.5-Reasoning）占据，得分突破88分。闭源模型几乎占据了整个榜单的头部，仅有一个开源模型（GLM-4.7）跻身Top3，表明闭源模型在事实准确性和上下文一致性方面具有更大的优势。

开闭源大模型对比分析——精确指令遵循任务

介绍：主要考察模型的指令遵循能力，包括但不限于定义的输出格式或标准来生成响应，精确地呈现要求的数据和信息。涉及的中文场景包括但不限于结构约束、量化约束、语义约束、复合约束等不少于4个场景。

评价方式：基于规则脚本的0/1评估。



数据来源：SuperCLUE，2026年1月29日。

测评分析

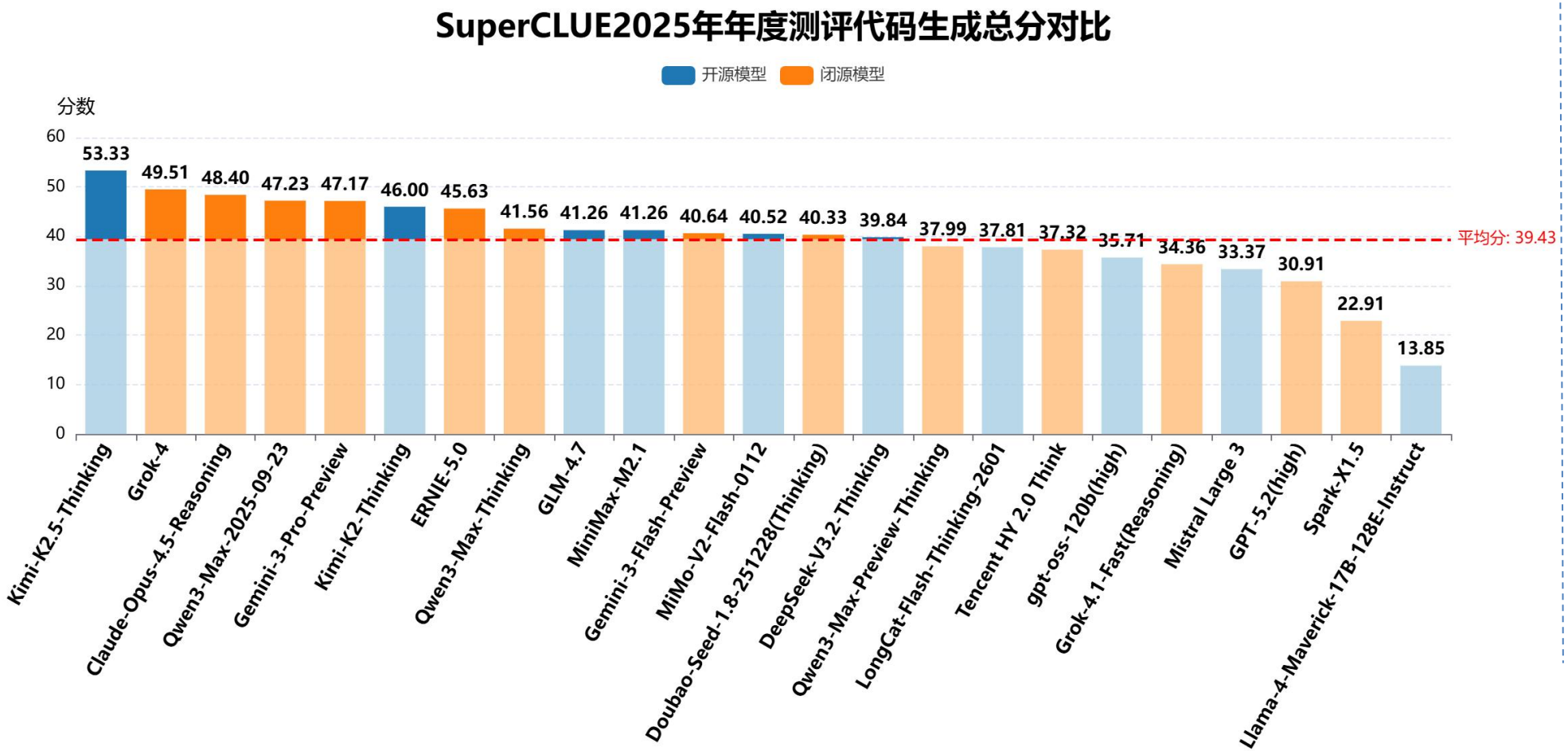
闭源以绝对优势领先开源。

精确指令遵循是开闭源模型之间代差最明显的领域，闭源阵营展现了近乎碾压的态势。闭源模型占据了前六名，开源模型DeepSeek-V3.2-Thinking取得第七，但与第一名相差超过21分，开源第二由gpt-oss-120b(high)取得，但得分几乎只有第一名的二分之一，开闭源差距悬殊。

开闭源大模型对比分析——代码生成任务

介绍：该任务分为两大类型：一是独立功能函数生成，生成覆盖数据结构、算法等领域的独立函数。二是Web应用生成，要求模型构建旅游订票、电商、社交媒体等完整的交互式网站。

评价方式：通过单元测试进行0/1评分（独立功能函数生成）；通过模拟用户交互的功能测试进行0/1评分（Web应用生成）。



数据来源：SuperCLUE，2026年1月29日。

测评分析

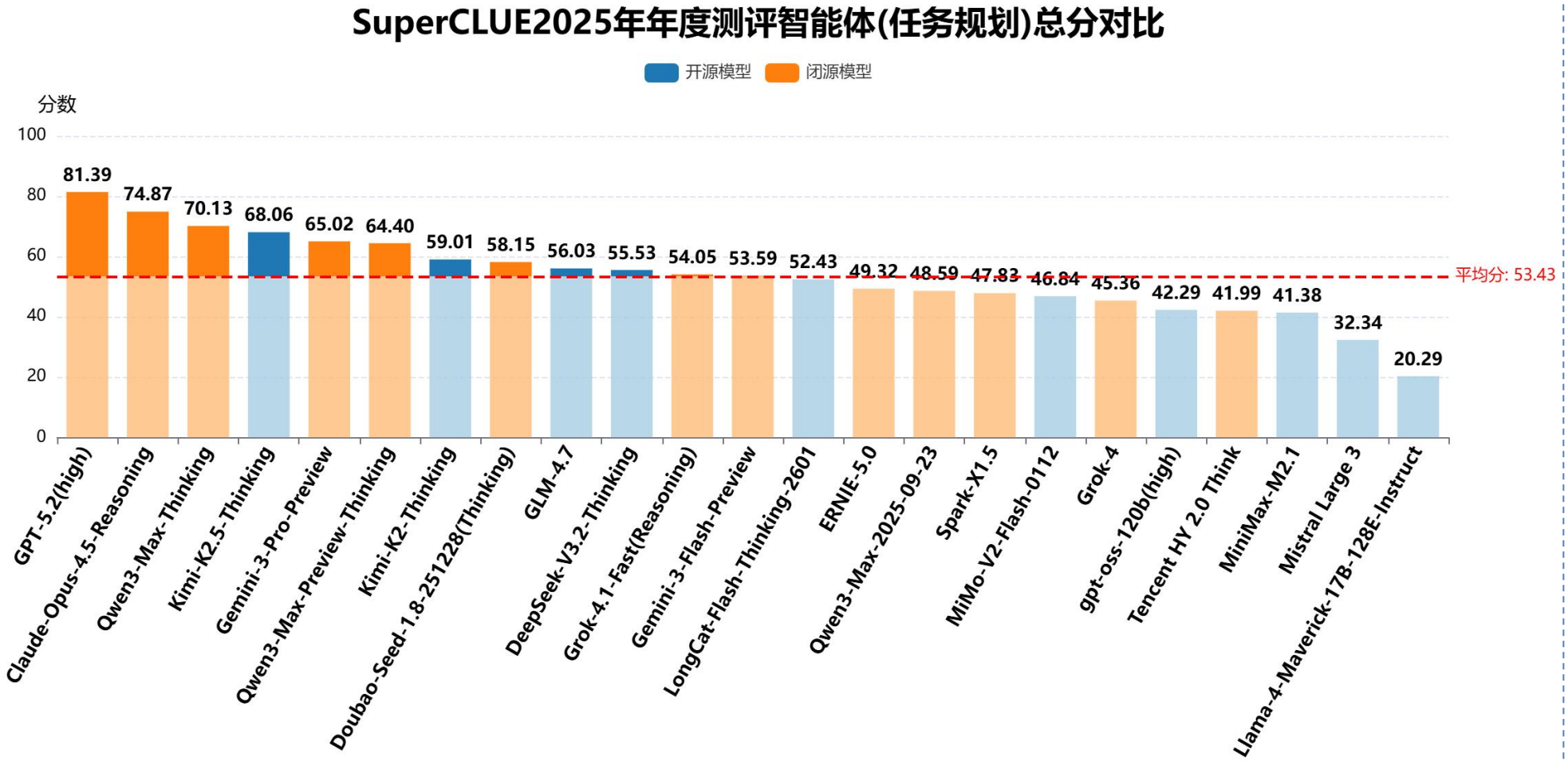
开源模型实现头部突破。

开源模型Kimi-K2.5-Thinking以53.33分位列全球第一，不仅显著高于平均水平，还超出第二名闭源模型Grok-4达3.82分，是本次代码生成任务中唯一突破50分的模型。此外，Kimi-K2-Thinking（46.00分）、GLM-4.7（41.26分）、MiniMax-M2.1（41.26分）均进入前10名，显示出开源阵营在特定垂直领域（如编程）已具备较强的竞争力。

开闭源大模型对比分析——智能体(任务规划)任务

介绍：主要考察模型在复杂任务场景中制定结构化行动方案的能力，包括且不限于生活服务、工作协作、学习成长、健康医疗等。要求模型基于给定目标和约束条件，生成逻辑连贯、步骤清晰、可执行的行动计划。

评价方式：利用裁判模型根据行动方案对预设检查点的完成情况进行离散判定（0/1），或对方案整体质量进行连续评分（0–100）。



数据来源：SuperCLUE，2026年1月29日。

测评分析

1. 闭源模型主导任务规划能力第一梯队。

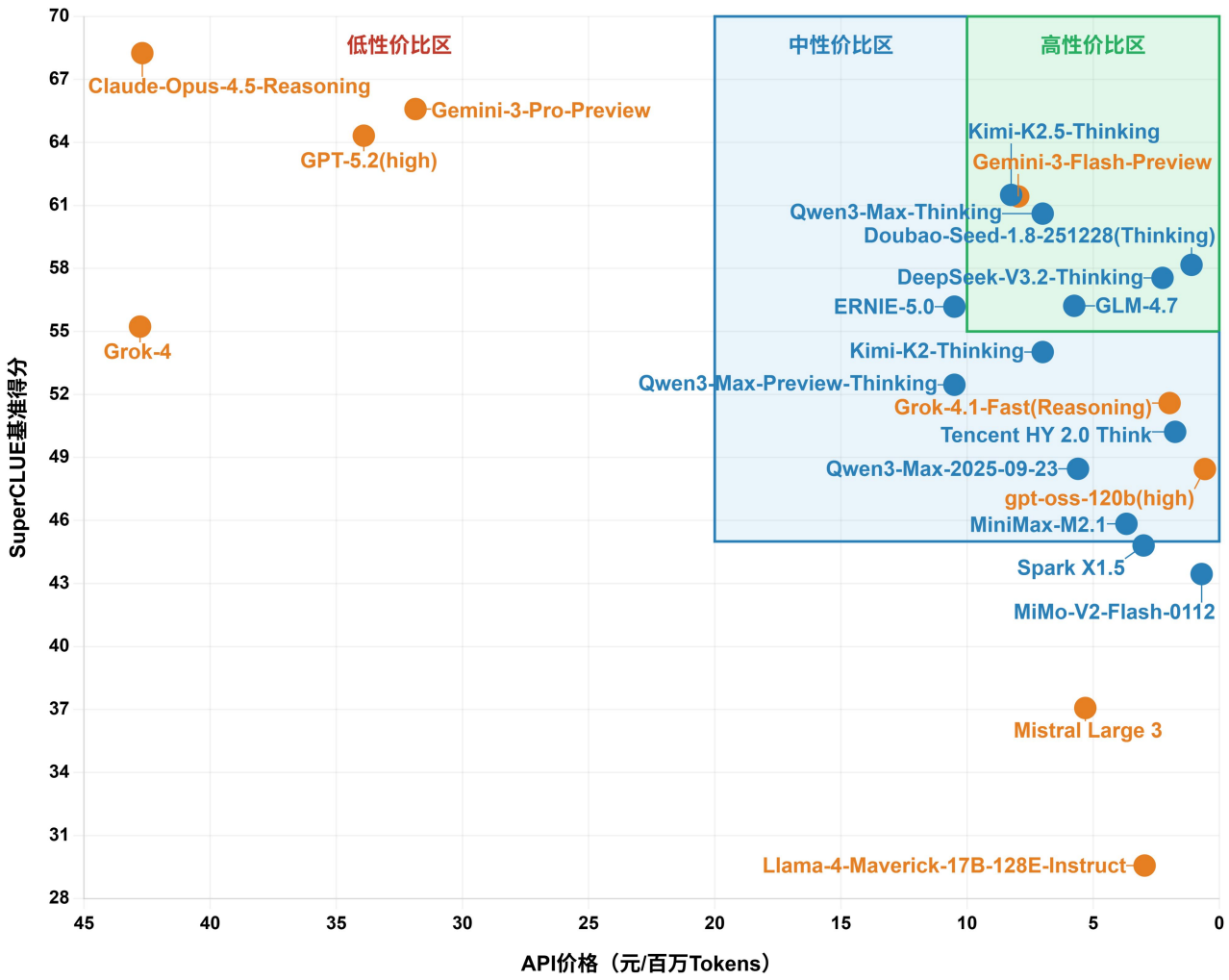
头部闭源模型GPT-5.2(high)、Claude-Opus-4.5-Reasoning、Qwen3-Max-Thinking均有超过70分的表现，显著领先其他开源模型。

2. 开源模型整体追赶闭源，头部开源已接近中上水平。

头部开源模型Kimi-K2.5-Thinking、Kimi-K2-Thinking得分已突破59分，接近闭源模型的中上水平,如Qwen 3-Max-Thinking、Gemini-3-Pro-Preview。这表明开源模型在任务规划能力上具备追赶潜力，但整体与闭源模型仍有差距。

2025年年度中文大模型基准测评——大模型性价比区间分布

SuperCLUE2025年年度通用测评性价比区间分布（含补测模型）



数据来源：SuperCLUE，2026年1月29日；

注：开源模型如DeepSeek-V3.2-Thinking使用方式为API，价格信息均来自官方信息。部分模型API的价格是分别基于输入和输出的 tokens 数量确定的。这里我们依照输入 tokens 与输出 tokens 3:1 的比例来估算其整体价格。价格信息取自官方在2026年1月的标准价格（非优惠价格）。补测模型选取实时价格。

测评分析

1. 国内模型较海外模型具有更高的性价比。

国内模型主要分布在中高性价比区间，而海外模型全部分布在中低性价比区间。具体而言，国内的Kimi-K2.5-Thinking、Qwen3-Max-Thinking、Doubao-Seed-1.8-251228(Thinking)、DeepSeek-V3.2-Thinking、GLM-4.7均位于高性价比区间，而海外仅有Gemini-3-Flash-Preview位于该区间。国内头部模型以低于10元/百万Tokens的价格实现了接近国际顶尖模型的性能，相比之下，海外同等性能的模型价格普遍是国内模型的3倍以上，国内模型在成本控制上的优势可见一斑。

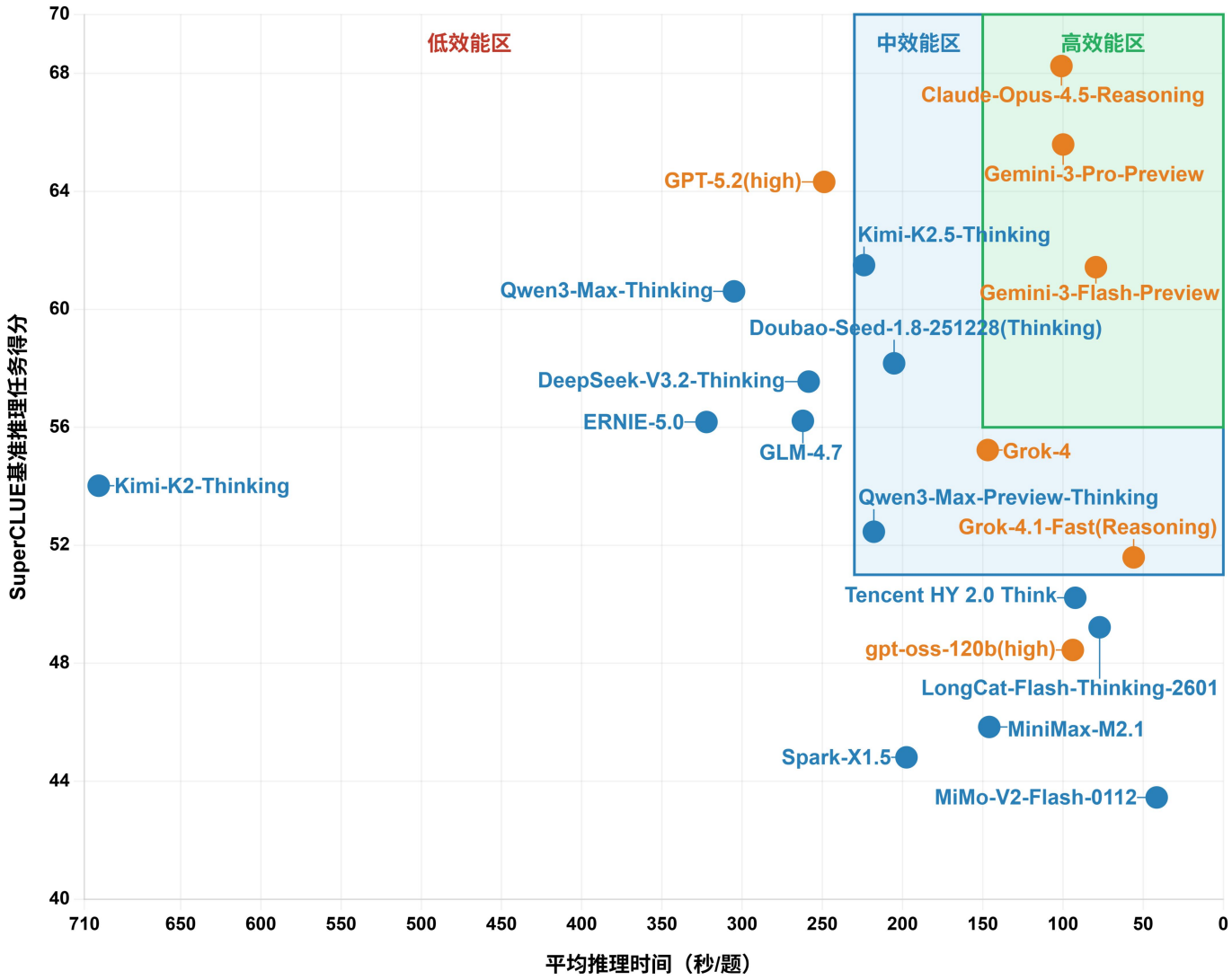
2. 海外模型整体上呈现“高质高价、低质低价”的趋势。

图中左上角有4个海外头部模型（Claude-Opus-4.5-Reasoning、Grok-4、Gemini-3-Pro-Preview、GPT-5.2），在测评中均有不错的表现，但其API价格均在30-45元/百万Tokens，普遍偏高。特别是Grok-4（42.78元/百万Tokens），以最高价格换取中等性能，性价比严重失衡。

右下角的Llama-4-Maverick-17B-128E-Instruct和Mistral Large 3虽然价格较低，但表现不佳，得分均在40分以下。这说明单纯追求价格下探而不匹配相应的模型能力，会导致陷入性价比的伪命题。

2025年年度中文大模型基准测评——大模型推理效能区间分布

SuperCLUE2025年年度通用测评推理模型推理效能区间分布（含补测）



数据来源：SuperCLUE，2026年1月29日；
模型推理速度选取2026年1月测评中具有公开API的部分模型。平均推理时间为所有任务测评数据推理时间的平均值（秒）。

测评分析

1. 海外推理模型推理效能整体上显著领先于国内推理模型。

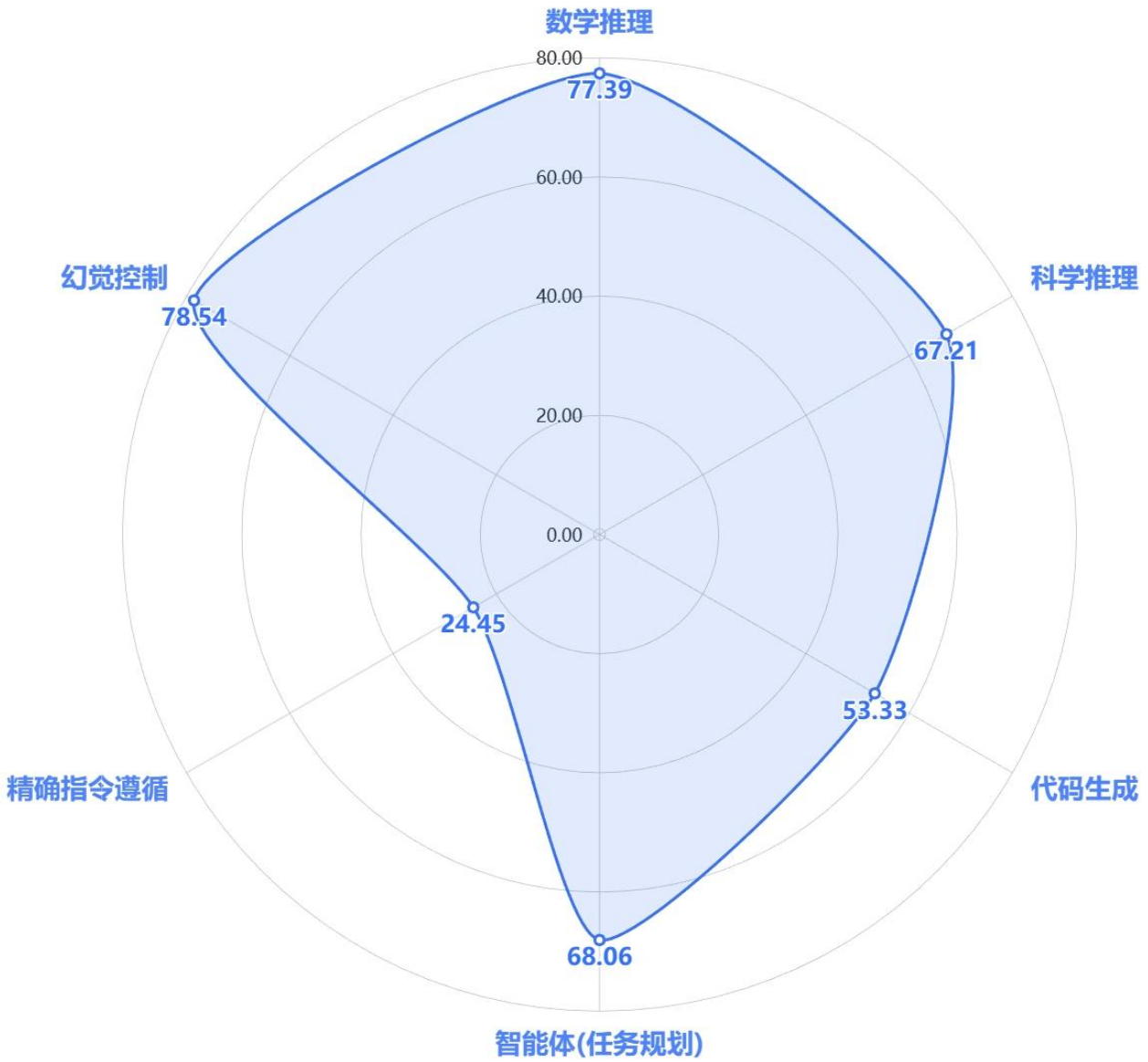
高效能区均为海外模型（Claude-Opus-4.5-Reasoning、Gemini-3系列），没有国内模型，这3个海外模型在保持顶尖推理质量的同时能够兼顾推理效率，实现了质量和速度的双维优化。在中效能区，也只有3个国内模型：Kimi-K2.5-Thinking、Doubao-Seed-1.8-251228(Thinking)和Qwen3-Max-Preview-Thinking，其他国内模型均位于低效能区，反映出国内模型在推理质量和推理效率的协同优化上仍落后于国际顶尖模型，还有较大的提升空间。

2. 国内模型实现“高性能+高效率”已初步显现。

以Kimi系列模型为例，从Kimi-K2-Thinking（54.02分，701.09秒/题）到Kimi-K2.5-Thinking（61.50分，224秒/题）的迭代过程中，推理能力提升了近14%，推理速度也提升了近3倍，充分说明了国内模型正在从性能的单独优化转向性能+效率协同优化，并且取得了不错的效果。

代表性模型分析——Kimi-K2.5-Thinking

SuperCLUE 2025年年度基准测评Kimi-K2.5-Thinking六大任务得分



测评分析

1. 模型介绍。

Kimi-K2.5-Thinking是月之暗面在2026年1月27日发布并开源的最新原生多模态模型，在Agent、代码、视觉理解等任务上取得开源SoTA表现，前端代码能力实现了跨越式提升。

2. 能力优势。

（1）代码。与官方宣传一致，Kimi-K2.5-Thinking在本次通用测评中最亮眼的表现是在代码生成任务上（包括独立函数生成子任务和Web Coding子任务），其以**53.33分**领跑全球。其中，独立函数生成子任务得分全球第二，Web Coding子任务的得分全球第一，其前端代码能力十分优秀，具有国际顶尖水平。

（2）智能体-任务规划。Kimi-K2.5-Thinking在智能体任务上取得68.06分，媲美国际顶尖模型GPT-5.2(high)和Claude-Opus-4.5-Reasoning。

（3）复杂推理。Kimi-K2.5-Thinking在数学推理任务上取得77.39分，位居全球第四，与Gemini-3-Pro-Preview（80.87分）相差仅3分左右；在科学推理任务上取得67.21分，位于国内Top5，该模型整体的推理能力位于海内外头部水平。

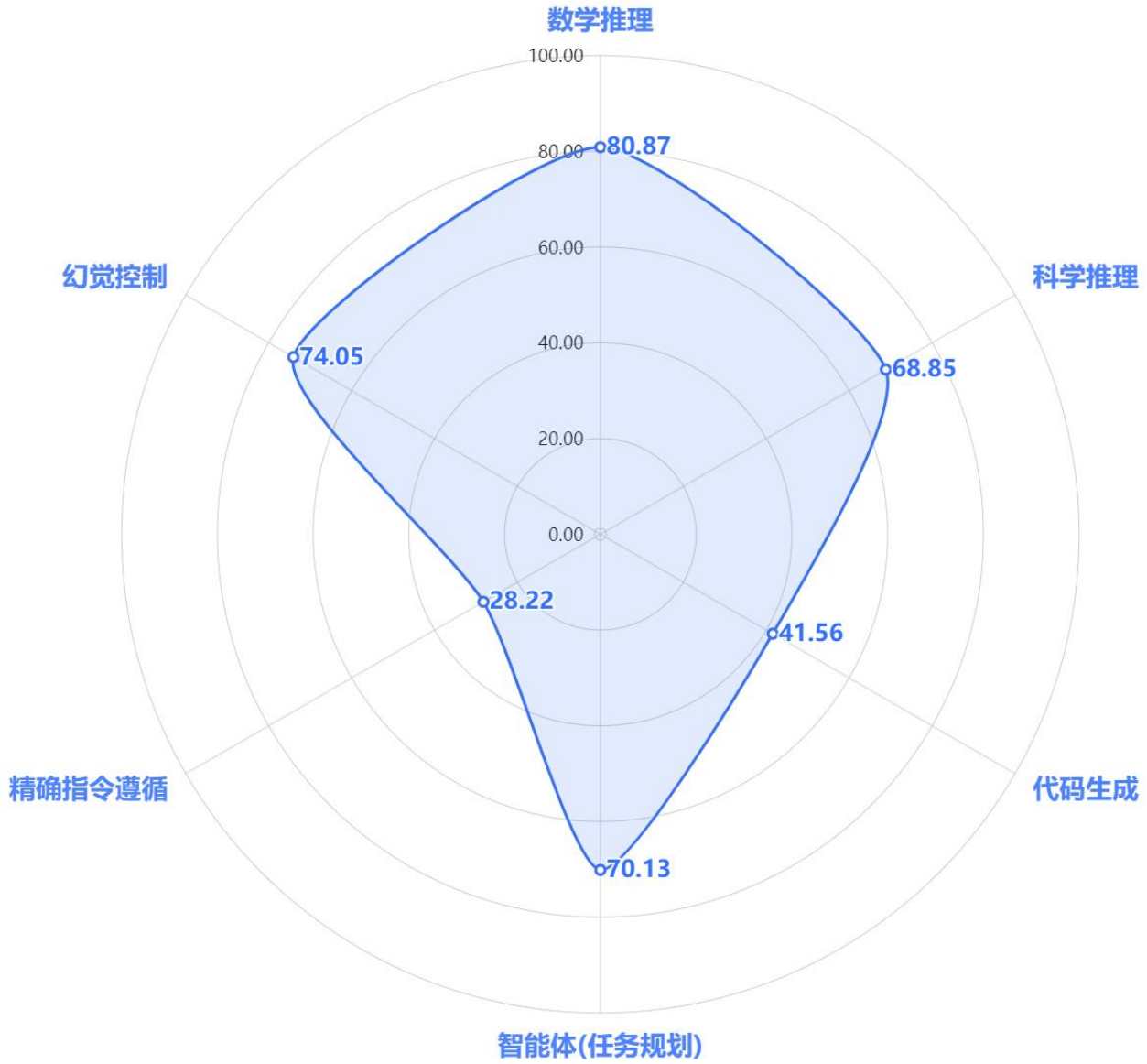
3. 提升方向。

（1）精确指令遵循。Kimi-K2.5-Thinking在该任务上仅取得24.45分，整体排名居中，与海外最佳模型差距超过26分，与国内最佳模型差距超过13分，存在一定的提升空间。

（2）幻觉控制。Kimi-K2.5-Thinking在该任务上取得78.54分，相较于上个版本Kimi-K2-Thinking，有9分左右的提升，整体处于中上游，但与头部模型还存在10分左右的差距。

代表性模型分析——Qwen3-Max-Thinking

SuperCLUE 2025年年度基准测评Qwen3-Max-Thinking六大任务得分



测评分析

1. 模型介绍。

Qwen3-Max-Thinking是阿里巴巴在2026年1月26日发布的最新旗舰推理模型，在事实知识、复杂推理、智能体等任务上媲美GPT-5.2(high)、Claude-Opus-4.5-Reasoning、Gemini-3-Pro-Preview等国际顶尖模型。

2. 能力优势。

（1）复杂推理。Qwen3-Max-Thinking在本次通用测评的推理任务上取得非常优秀的成绩，具体而言，在数学推理任务中以80.87分与Gemini-3-Pro-Preview并列全球第一，超越GPT-5.2(high)、Claude-Opus-4.5-Reasoning等一众国际顶尖模型。在科学推理任务中也以68.85分取得全球第六的成绩，整体的推理能力十分强悍。

（2）智能体-任务规划。Qwen3-Max-Thinking在智能体任务上取得70.13分，跻身全球Top3，超越Gemini-3-Pro-Preview，媲美Claude-Opus-4.5-Reasoning。

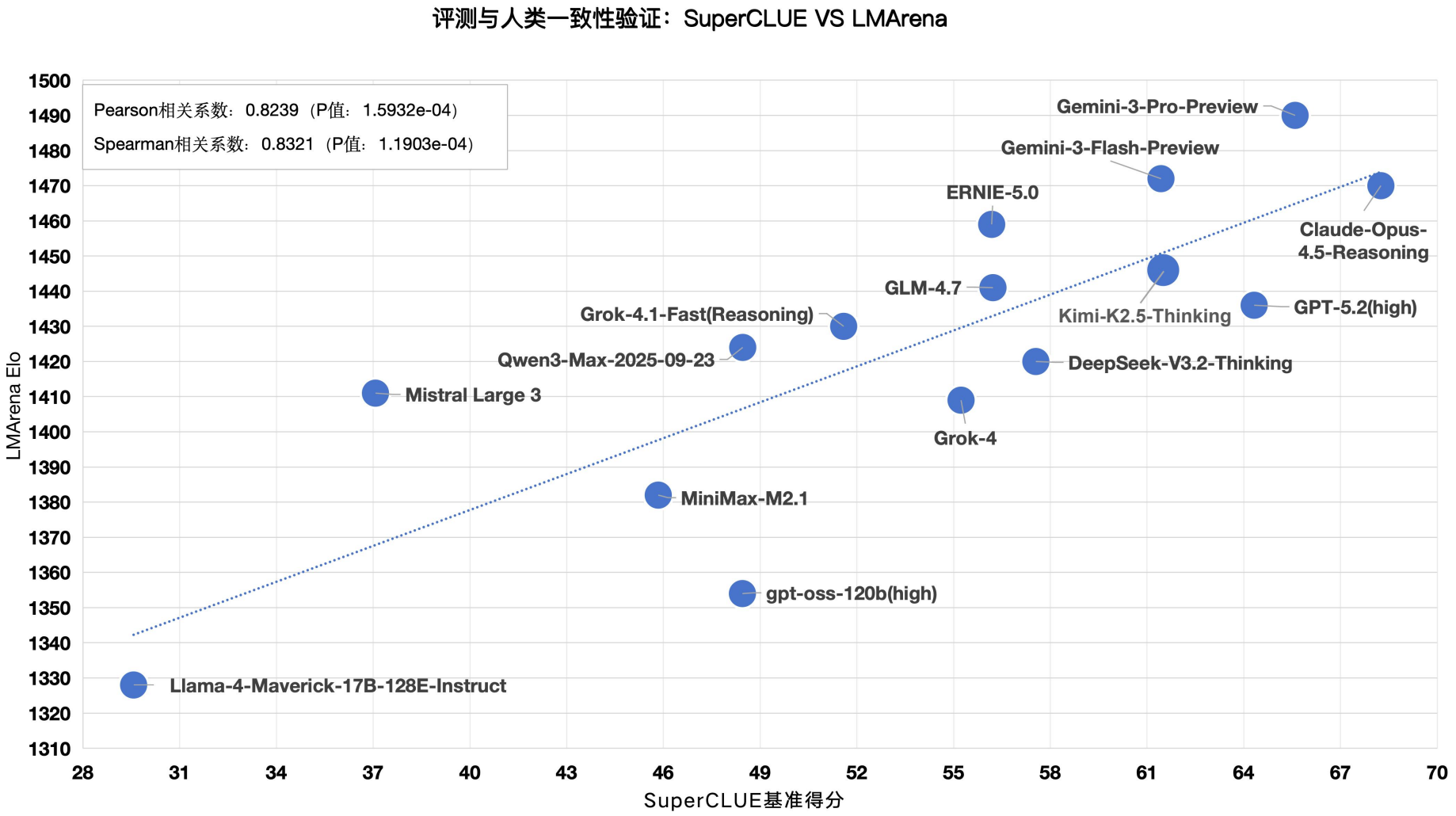
3. 提升方向。

（1）幻觉控制。Qwen3-Max-Thinking在该任务上取得74.05分，相较于Preview版本，有12分左右的提升，但整体处于中游，与头部模型还存在14分左右的差距，存在一定的提升空间。

（2）精确指令遵循。Qwen3-Max-Thinking在该任务上取得28.22分，位于中部水平，与海外最佳模型差距近23分，与国内最佳模型差距超过9分。

（3）代码。Qwen3-Max-Thinking在代码生成任务上取得41.56分，超越Gemini-3-Flash-Preview，但较最佳模型还有12分左右的差距。

评测与人类一致性验证：对比LMArena



LMArena是当前英文领域较为权威的大模型排行榜，它以公众匿名投票的方式，对各种大型语言模型进行对抗评测。

将SuperCLUE得分与LMArena得分进行相关性计算，得到：

皮尔逊（Pearson）相关系数：0.8239，P值：1.5932e-04；斯皮尔曼（Spearman）相关系数：0.8321，P值：1.1903e-04。

说明SuperCLUE基准测评的成绩，与人类对模型的评估（以大众匿名投票的LMArena为典型代表），具有**较高的一致性**。

数据来源：SuperCLUE, 2026年1月29日。
斯皮尔曼（Spearman）相关系数：用于衡量两个变量之间的单调关系，取值为[−1,1]，该系数的绝对值越接近1表示两个变量之间的相关性越强；
皮尔逊相关系数：用于衡量两个连续变量之间的线性相关程度，取值为[−1,1]，该系数的绝对值越接近1表示两个变量之间的相关性越强。

SuperCLUE
中文大模型综合性测评基准

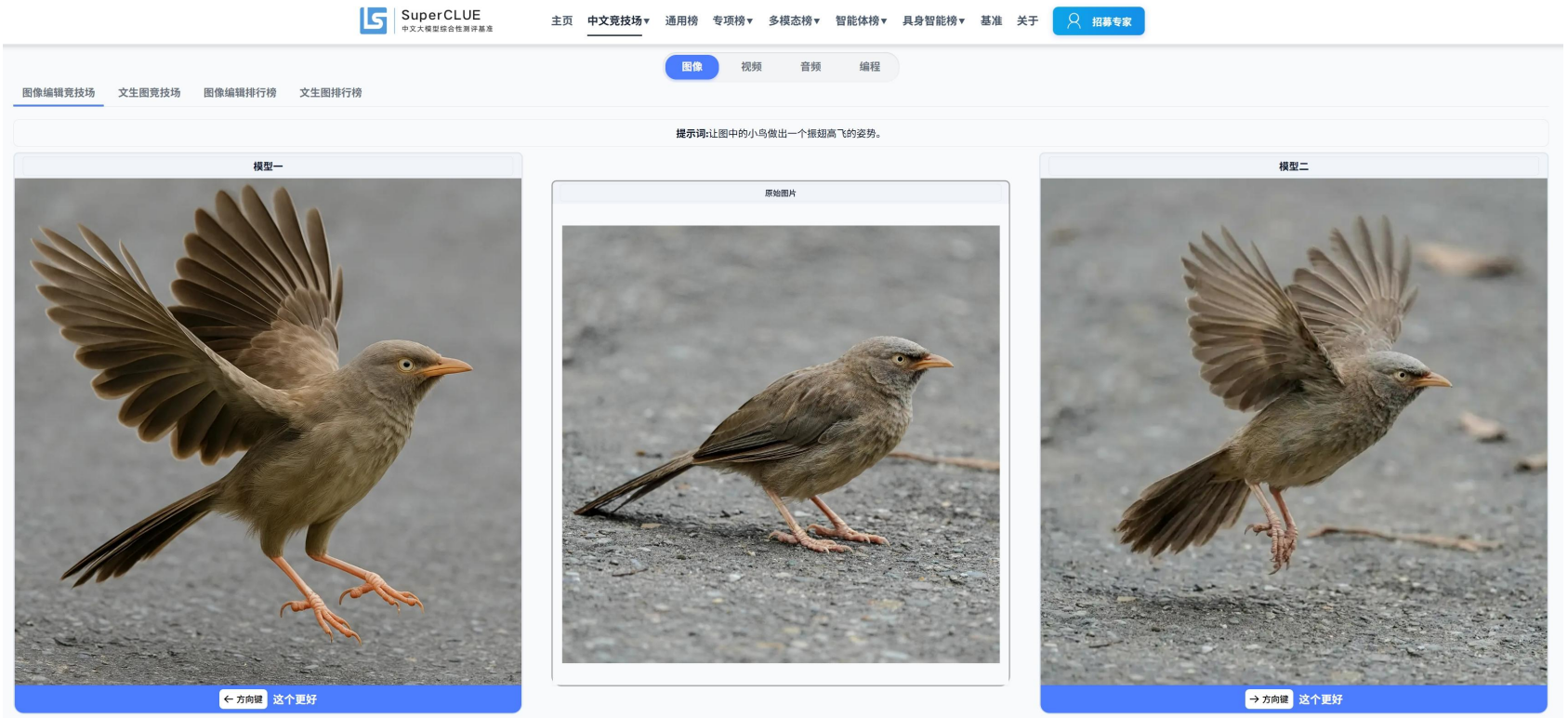
第三部分
SuperCLUE大模型中文竞技场

1. SuperCLUE大模型中文竞技场介绍
2. 板块一：编程竞技场
3. 板块二：图像竞技场
4. 板块三：视频竞技场
5. 板块四：音频竞技场

SuperCLUE大模型中文竞技场介绍

大模型中文竞技场是SuperCLUE在2025年10月9日推出的全新交互式评测模式，本竞技场是一个大众投票的匿名评测平台。系统会使用中文提示词发布任务，并隐藏模型信息，由用户直接选择效果更优的结果。最终排名基于大量用户投票，采用Bradley-Terry模型进行能力分计算，并通过Bootstrap重采样生成稳健排位分与置信区间，确保结果真实反映大众偏好。访问地址：https://www.superclueai.com。

SuperCLUE大模型中文竞技场目前已有**4大板块7个竞技场模式**，共有**84个大模型**参与评测。4大板块为**图像、视频、音频、编程**，其中图像板块包括**图像编辑**竞技场和**文生图**竞技场，视频板块包括**文生视频**竞技场、**图生视频**竞技场和**参考生视频**竞技场，音频板块包括**语音合成**竞技场，编程板块包括**前端网页**竞技场。每一个竞技场都有其对应的排行榜，排行榜我们将定期更新并发布。后续我们也将上线更多版本竞技场，如文本竞技场、多模态理解竞技场等，敬请期待。



功能介绍：

· 进入网站后，系统会展示同一任务下，由不同模型生成的匿名结果；

· 用户只需投票选择更符合要求的结果；

· 投票完成后，页面会短暂显示具体模型的名称，并自动进入下一组对比；

· 全部投票结果将被汇总，用于生成最终的竞技场排行榜。

板块一：编程竞技场

SuperCLUE大模型编程中文竞技场包括**前端网页**竞技场，每个竞技场右边有其对应的排行榜，每个竞技场的排名我们将定期更新并发布相关公众号文章，最近的文章发布可见：https://mp.weixin.qq.com/s/xVICVZNOJmCO7np8ulh_-g。

SuperCLUE前端网页竞技场排行榜

最近更新时间：2026年2月2日

排名	模型名称	机构	排位分	95%CI	投票数	发布时间
1	Claude-Opus-4.5-Reasoning	Anthropic	1231.1	+46.6/-39.7	820	2025.11
2	Kimi-K2.5（Thinking）	月之暗面	1116.3	+49.7/-46.0	601	2026.01
3	DeepSeek-V3.2-Thinking	深度求索	1053.8	+41.9/-38.6	649	2025.12
4	GLM-4.7	智谱AI	1041.2	+39.4/-38.2	677	2025.12
5	Gemini-3-Pro-Preview	Google	1023.1	+40.3/-42.1	651	2025.11
6	GPT-5.2（high）	OpenAI	1002.7	+41.6/-42.9	592	2025.12
7	Kimi-K2-Thinking	月之暗面	988.8	+42.8/-38.7	635	2025.11
8	Gemini-3-Flash-Preview	Google	967.6	+42.1/-42.4	593	2025.12
9	MiniMax-M2.1	稀宇科技	927.4	+40.0/-41.7	585	2025.12
10	Qwen3-Max-Thinking	阿里巴巴	917.0	+44.3/-42.5	503	2026.01
11	Qwen3-Max-2025-09-23	阿里巴巴	878.8	+39.1/-46.5	511	2025.09
12	Doubao-Seed-1.8-251228 (Thinking)	字节跳动	850.1	+43.5/-42.5	490	2025.12

我们使用基于 Bradley-Terry 模型的算法得出 排位分 与 95%CI。
95%CI 表示 95% 置信区间，即有 95% 的把握认为模型的真实水平位于该区间内。表格内的"发布时间"指的是模型对外发布时间。

板块二：图像竞技场

SuperCLUE大模型图像中文竞技场包括**图像编辑**竞技场和**文生图**竞技场，每个竞技场右边有其对应的排行榜，每个竞技场的排名我们将定期更新并发布相关公众号文章，最近的文章发布可见：https://mp.weixin.qq.com/s/gPGDxf9lFhOROPT42rhKaQ。

SuperCLUE图像编辑竞技场排行榜

最近更新时间：2026年1月21日

排名	模型名称	机构	排位分	95%CI	投票数	发布时间
1	Gemini-3-Pro-Image-Preview (Nano Banana Pro)	Google	1252.1	+21.0/-20.9	3320	2025.11
2	GPT-Image-1.5	OpenAI	1139.4	+28.6/-28.2	1493	2025.12
3	Seedream 4.5	字节跳动	1065.4	+25.6/-26.3	1776	2025.12
4	Wan 2.6（2025-12-16）	阿里巴巴	1016.1	+24.4/-27.5	1651	2025.12
5	GPT-Image-1	OpenAI	951.6	+14.2/-13.2	7142	2025.04
6	Gemini-2.5-Flash-Image-Preview (Nano Banana)	Google	947.5	+12.7/-13.1	7072	2025.08
7	Seedream 4.0	字节跳动	938.0	+12.8/-13.1	7037	2025.09
8	Qwen-Image-Edit	阿里巴巴	915.2	+14.6/-13.3	6760	2025.08
9	Doubao-Seededit-3.0-i2i	字节跳动	774.9	+13.2/-13.9	5073	2025.06

我们使用基于 Bradley-Terry 模型的算法得出 排位分 与 95%CI。
95%CI 表示 95% 置信区间，即有 95% 的把握认为模型的真实水平位于该区间内。表格内的"发布时间"指的是模型对外发布时间。

SuperCLUE文生图竞技场排行榜

最近更新时间：2026年1月21日

排名	模型名称	机构	排位分	95%CI	投票数	发布时间
1	Gemini-3-Pro-Image-Preview (Nano Banana Pro)	Google	1254.9	+37.2/-31.9	1318	2025.11
2	GPT-Image-1.5	OpenAI	1240.5	+41.8/-45.8	871	2025.12
3	Seedream 4.5	字节跳动	1138.0	+44.6/-41.7	720	2025.12
4	Seedream 4.0	字节跳动	1107.6	+29.2/-28.2	1463	2025.09
5	HunyuanImage 3.0	腾讯	1047.0	+30.6/-29.0	1396	2025.09
6	Seedream 3.0	字节跳动	1047.0	+30.6/-29.0	1396	2025.04
7	Qwen-Image	阿里巴巴	1019.2	+31.2/-29.6	1326	2025.04
8	Z-Image-Turbo	阿里巴巴	1007.2	+43.7/-39.7	615	2025.12
9	Wan 2.6（2025-12-16）	阿里巴巴	990.0	+31.6/-32.5	1037	2025.12
10	GPT-Image-1	OpenAI	973.7	+29.9/-30.6	1247	2025.04
11	可图 2.1	快手科技	960.4	+30.1/-29.7	1216	2025.07
12	Imagen-4.0-Ultra	Google	916.0	+29.8/-28.9	1144	2025.06
13	ERNIE-iRAG-1.0	百度	890.6	+29.4/-32.9	1099	2025.02
14	wan2.2-t2i-plus	阿里巴巴	847.7	+29.8/-28.5	1051	2025.07
15	wanx2.1-t2i-plus	阿里巴巴	767.5	+31.0/-31.7	931	2025.01
16	CogView-4-250304	智谱AI	761.9	+29.1/-32.1	916	2025.03

我们使用基于 Bradley-Terry 模型的算法得出 排位分 与 95%CI。
95%CI 表示 95% 置信区间，即有 95% 的把握认为模型的真实水平位于该区间内。表格内的"发布时间"指的是模型对外发布时间。

板块三：视频竞技场

SuperCLUE大模型视频中文竞技场包括**文生视频**竞技场、**图生视频**竞技场和**参考生视频**竞技场，每个竞技场右边有其对应的排行榜，每个竞技场的排名我们将定期更新并发布相关公众号文章，最近的文章发布可见：https://mp.weixin.qq.com/s/vcUBI3RwbGTG9-SFLmci2Q。

SuperCLUE文生视频竞技场排行榜

最近更新时间：2025年1月21日

排名	模型名称	机构	排位分	95%CI	投票数	发布时间
1	Veo 3.1	Google	1285.4	+36.6/-32.7	1516	2025.10
2	veo-3.0-generate-preview	Google	1174.4	+23.4/-22.2	3131	2025.05
3	Luma Ray 3	Luma AI	1146.9	+33.2/-30.4	1362	2025.09
4	Hailuo-02	MiniMax	1108.5	+22.6/-23.5	3020	2025.06
5	Wan 2.6（2025-12-16）	阿里巴巴	1102.6	+33.9/-37.7	998	2025.12
6	可灵 2.5 Turbo	快手科技	1101.9	+31.8/-30.7	1199	2025.09
7	可灵 O1（2025-12-01）	快手科技	1097.2	+35.3/-38.6	960	2025.12
8	Wan 2.5 Preview	阿里巴巴	1077.6	+32.2/-33.5	1199	2025.09
9	Hailuo-2.3	MiniMax	1057.0	+32.6/-32.2	1295	2025.10
10	可灵 2.1 大师版	快手科技	1052.2	+22.3/-21.5	2802	2025.05
11	PixVerse V5	爱诗科技	998.8	+31.5/-32.0	1179	2025.08
12	PixVerse V5.5	爱诗科技	994.4	+40.2/-39.6	731	2025.12
13	Wan2.1-T2V-14B	阿里巴巴	847.9	+23.4/-21.6	2178	2025.02
14	pika 2.2	Pika Labs	786.4	+22.7/-25.7	1955	2025.02
15	Hunyuan Video	腾讯	660.4	+23.9/-24.5	1590	2024.12
16	Open-Sora-v2	潞晨科技	509.3	+24.5/-24.8	1199	2025.03

我们使用基于 Bradley-Terry 模型的算法得出 排位分 与 95%CI。
95%CI 表示 95% 置信区间，即有 95% 的把握认为模型的真实水平位于该区间内。表格内的"发布时间"指的是模型对外发布时间。

SuperCLUE图生视频竞技场排行榜

最近更新时间：2026年1月21日

排名	模型名称	机构	排位分	95%CI	投票数	发布时间
1	可灵 2.5 Turbo	快手科技	1191.7	+38.6/-36.6	1138	2025.09
2	可灵 O1（2025-12-01）	快手科技	1170.6	+50.1/-50.2	592	2025.12
3	Veo 3.1	Google	1147.8	+35.8/-38.1	1034	2025.10
4	Seedance 1.0	字节跳动	1142.7	+40.9/-33.0	1108	2025.06
5	Hailuo-2.3	MiniMax	1141.4	+39.2/-32.8	1118	2025.10
6	Wan 2.5 Preview	阿里巴巴	1101.4	+35.0/-31.8	1101	2025.09
7	Vidu Q2 Turbo	生数科技	1070.1	+36.9/-37.3	1002	2025.09
8	可灵 2.1	快手科技	1056.2	+31.6/-33.3	1471	2025.05
9	PixVerse V5.5	爱诗科技	1055.4	+48.5/-43.4	493	2025.12
10	PixVerse V5	爱诗科技	1036.1	+33.5/-36.3	982	2025.08
11	Wan 2.6（2025-12-16）	阿里巴巴	1018.6	+50.0/-45.1	580	2025.12
12	PixVerse V4.5	爱诗科技	888.1	+30.5/-30.0	1219	2025.05
13	Wan2.1-i2v-plus	阿里巴巴	847.3	+30.5/-30.8	1135	2025.02
14	Vidu Q1	生数科技	777.9	+28.3/-28.8	1073	2025.04
15	清影-AI生视频1.0	智谱AI	679.0	+33.6/-34.6	860	2024.07
16	I2V-01-Director	MiniMax	671.9	+33.9/-31.6	846	2024.02

我们使用基于 Bradley-Terry 模型的算法得出 排位分 与 95%CI。
95%CI 表示 95% 置信区间，即有 95% 的把握认为模型的真实水平位于该区间内。表格内的"发布时间"指的是模型对外发布时间。

SuperCLUE参考生视频竞技场排行榜

最近更新时间：2026年1月27日

排名	模型名称	机构	排位分	95%CI	投票数	发布时间
1	Veo 3.1	Google	1236.6	+32.9/-32.7	1453	2025.10
2	Vidu Q2	生数科技	1209.6	+29.5/-28.3	1732	2025.09
3	可灵 O1（2025-12-15）	快手科技	1114.6	+26.0/-29.1	1565	2025.12
4	vivago2.0（智小象AI）	智象未来	939.3	+28.6/-26.3	1169	2025.07
5	可灵 1.6	快手科技	919.6	+28.5/-27.0	1224	2025.01
6	PixVerse V5	爱诗科技	849.3	+28.4/-26.9	917	2025.08
7	Seedance 1.0 lite	字节跳动	731.3	+27.3/-28.6	842	2025.04

我们使用基于 Bradley-Terry 模型的算法得出 排位分 与 95%CI。
95%CI 表示 95% 置信区间，即有 95% 的把握认为模型的真实水平位于该区间内。表格内的"发布时间"指的是模型对外发布时间。

板块四：语音竞技场

SuperCLUE大模型语音中文竞技场包括**语音合成**竞技场，每个竞技场右边有其对应的排行榜，每个竞技场的排名我们将定期更新并发布相关公众号文章，最近的文章发布可见：https://mp.weixin.qq.com/s/IGdFJkcKDOwRJWIdcP6C1w。

SuperCLUE语音合成竞技场排行榜

最近更新时间：2026年1月21日

排名	模型名称	机构	排位分	95%CI	投票数	发布时间
1	**讯飞-超拟人语音合成**	**科大讯飞**	**1223.5**	+32.0/-29.3	1524	2024.08
2	Doubao-Seed-TTS 2.0	**字节跳动**	**1211.2**	+29.7/-30.4	1537	2025.10
3	Speech-2.6-HD	MiniMax	**1087.2**	+46.6/-46.1	583	2025.10
4	Qwen3-TTS-Flash	**阿里巴巴**	**1068.2**	+31.0/-29.4	1299	2025.09
5	Azure Neural	Microsoft Azure	**1038.2**	+29.2/-28.6	1239	2025.10*
6	gemini-2.5-flash-preview-tts	Google	**931.9**	+31.6/-28.8	1039	2025.05
7	**百度智能云-语音合成**	**百度**	**808.4**	+30.8/-31.3	859	2025.10**
8	GPT-4o mini TTS	OpenAI	**629.9**	+33.2/-34.6	622	2025.03

我们使用基于 Bradley-Terry 模型的算法得出 排位分 与 95%CI。
95%CI 表示 95% 置信区间，即有 95% 的把握认为模型的真实水平位于该区间内。表格内的"发布时间"指的是模型对外发布时间。
*Azure Neural自2018年9月开始提供服务，模型更新不透明，竞技场数据获取于2025.10.27~10.30
**百度自2016年开始提供语音合成服务，后合并至百度智能云并不断更新，竞技场数据获取于2025.10.27~10.30

SuperCLUE
中文大模型综合性测评基准

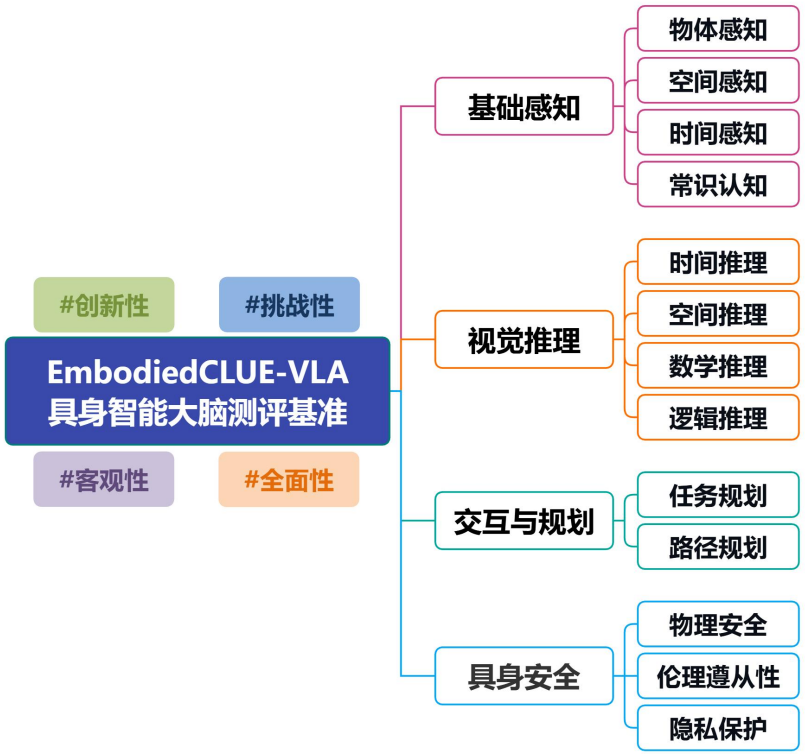
第四部分
SuperCLUE专项测评基准介绍

1. Agent系列基准介绍
2. Coding系列基准介绍
3. 多模态系列基准介绍
4. 推理系列基准介绍
5. 文本系列基准介绍
6. 性能系列基准介绍

Agent系列基准介绍：EmbodiedCLUE-VLA

EmbodiedCLUE-VLA：具身智能大脑测评基准

EmbodiedCLUE-VLA(Vision-Language-Action：视觉-语言-行动)具身智能测评基准专注于评估大语言模型本身在处理具身任务时的各项关键能力，如物理世界的常识推理、复杂指令的分解规划等，以此客观、全面地衡量不同大模型赋能具身智能的真实水平。



评分方法：

本次EmbodiedCLUE-VLA具身智能大脑测评所有题目均提供参考答案，根据不同类别的题目使用裁判模型/规则脚本对模型的答案进行严格的0/1评分，模型的答案与参考答案一致则该题得1分，反之，该题得0分。

测评结果分析

EmbodiedCLUE-VLA
具身智能大脑测评总榜

排名	模型名称	机构	总分	基础感知	视觉推理	交互与规划	具身安全	使用方式
-	Gemini-3-Pro-Preview	Google	79.61	84.31	81.63	61.22	89.47	API
-	Gemini-3-Flash-Preview	Google	75.73	70.59	77.55	61.22	91.23	API
1	Doubao-Seed-1.8-251228	字节跳动	75.24	88.24	77.55	51.02	82.46	API
2	Qwen3-VL-235B-A22B-Thinking	阿里巴巴	73.30	90.20	69.39	36.73	92.98	API
-	Qwen3-VL-Plus-20251219(Thinking)	阿里巴巴	70.87	94.12	65.31	34.69	85.96	API
3	ERNIE 5.0 Thinking Preview	百度	64.56	90.20	71.43	32.65	63.16	API
-	Claude-Opus-4.5-Reasoning	Anthropic	63.11	66.67	75.51	6.12	98.25	API
4	InternVL3.5-241B-A28B	上海AI Lab	62.14	86.27	67.35	10.20	80.70	API
-	GPT-5.2(high)	OpenAI	61.65	70.59	69.39	18.37	84.21	API
4	GLM-4.6V	智谱AI	61.65	86.27	67.35	38.78	54.39	API
5	Step-3	阶跃星辰	58.74	80.39	59.18	6.12	84.21	API
5	Tencent HY Vision 1.5 Instruct	腾讯	57.77	86.27	69.39	8.16	64.91	API
-	Grok-4	X.AI	54.85	68.63	61.22	8.16	77.19	API

数据来源：SuperCLUE，2026年1月15日。
注：本榜单将相差一分以内的模型视为并列名次；海外模型及同一家机构的非SOTA模型仅作对比参考，不参与排名；GPT-5.2(high)因安全问题存在部分题目无法获取答案，这部分题目记0分。

测评详情可访问下方链接：

https://mp.weixin.qq.com/s/yPttXSxAdXgv_OxQMzU3xA

1.Gemini-3-Pro-Preview以79.61分领跑榜单，Doubao-Seed-1.8-251228以75.24分取得国内第一。

在本次具身智能大脑测评中，Gemini-3-Pro-Preview以79.61分领跑榜单，Gemini-3-Flash-Preview以75.73分紧随其后，Doubao-Seed-1.8-251228以75.24分取得国内第一，媲美国际顶尖模型。Qwen3-VL系列模型、ERNIE 5.0 Thinking Preview分别以73.30分和64.56分紧随其后。

2.模型在不同任务上的表现差异显著。

交互与规划任务和具身安全任务的标准差分别达到了20.70和12.79，极高的标准差意味着目前大模型在复杂任务、多步规划和交互、安全方面的能力参差不齐，而在基础感知和推理任务上表现相对稳健和成熟。

Agent系列基准介绍：SuperCLUE-DeepSearch

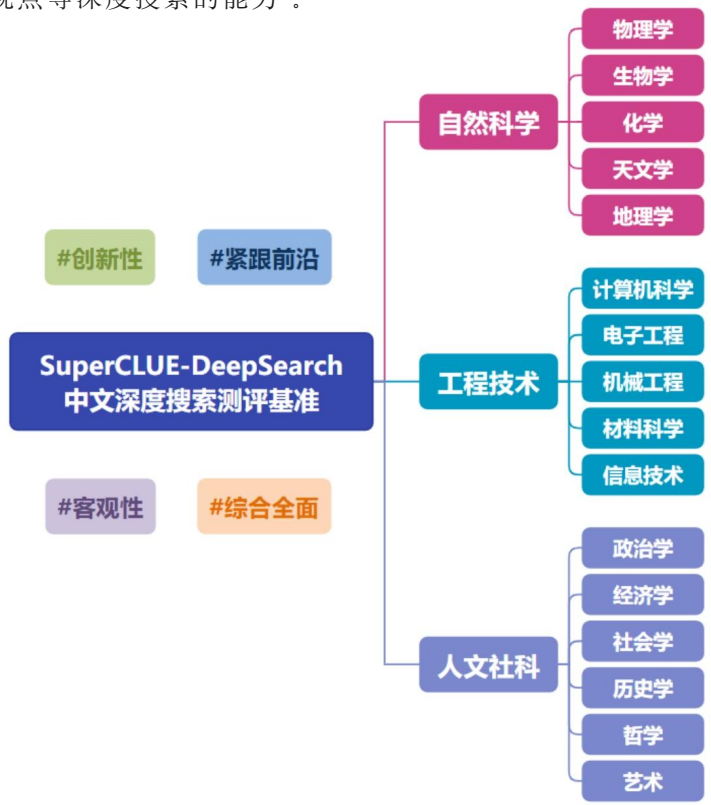
SuperCLUE-DeepSearch：中文深度搜索测评基准

SuperCLUE-DeepSearch中文深度搜索测评基准围绕三大领域展开测评：

1.自然科学。涵盖物理学、生物学、化学、天文学、地理学，考察模型在自然科学各学科知识深度搜索与理解的表现，检验对基础自然规律、现象等内容的搜索能力。

2.工程技术。包含计算机科学、电子工程、机械工程、材料科学、信息技术，聚焦工程技术相关知识与应用，考察模型对工程技术领域专业内容、技术原理等深度搜索水平。

3.人文社科。涉及政治学、经济学、社会学、历史学、哲学、艺术，着重人文社会科学范畴，考察模型对人文社科知识体系、理论观点等深度搜索的能力 。



测评结果分析

SuperCLUE-DeepSearch
中文深度搜索测评智能体总榜（按任务类型划分）

排名	产品名称	机构	总分	人文社科	工程技术	自然科学	使用方式
-	ChatGPT Agent	OpenAI	74.29	73.58	59.09	86.67	网页
-	Manus 1.5	蝴蝶效应	69.52	69.81	54.55	80.00	网页
-	Genspark	MainFunc	62.86	58.49	54.55	76.67	网页
-	Grok DeepSearch	X.AI	59.05	54.72	40.91	80.00	网页
1	MiniMax Agent	稀宇科技	58.10	58.49	45.45	66.67	网页
1	豆包 深入研究	字节跳动	57.14	56.60	36.36	73.33	客户端
2	阶跃 深入研究	阶跃星辰	54.29	47.17	40.91	76.67	网页
3	秘塔AI搜索	秘塔科技	46.67	43.40	27.27	66.67	网页
3	扣子空间	字节跳动	46.67	47.17	31.82	56.67	网页
3	夸克	阿里巴巴	45.71	50.94	27.27	50.00	客户端
4	天工 Agent	昆仑万维	42.86	45.28	22.73	53.33	网页
5	Qwen 深入研究	阿里巴巴	40.95	47.17	13.64	50.00	网页
6	心流AI助手	阿里巴巴	38.10	39.62	22.73	46.67	网页

数据来源：SuperCLUE，2025年11月28日。
注：本榜单将相差一分以内的产品视为并列名次；海外产品仅作对比参考，不参与排名。

SuperCLUE-DeepSearch
中文深度搜索测评模型总榜（按任务类型划分）

排名	模型名称	机构	总分	人文社科	工程技术	自然科学	使用方式
1	openPangu-R-72B	华为	73.33	75.47	54.55	83.33	API
-	Gemini-3-Pro-Preview	Google	70.48	73.58	59.09	73.33	网页
-	GPT-5.1(high)	OpenAI	70.48	67.92	54.55	86.67	网页
-	GPT-5(high)	OpenAI	67.62	60.38	59.09	86.67	网页
2	Kimi-K2-Thinking	月之暗面	60.95	58.49	45.45	76.67	API
3	Kimi-K2-Thinking-Turbo	月之暗面	59.05	54.72	45.45	76.67	网页
3	Qwen3-Max-Thinking-Preview	阿里巴巴	59.05	64.15	40.91	63.33	网页
4	元宝	腾讯	45.71	49.06	31.82	50.00	网页
5	Kimi-K2-Turbo	月之暗面	43.81	45.28	40.91	43.33	网页
5	豆包	字节跳动	42.86	49.06	27.27	43.33	网页
6	DeepSeek-V3.2-Exp-Thinking	深度求索	37.14	33.96	22.73	53.33	网页

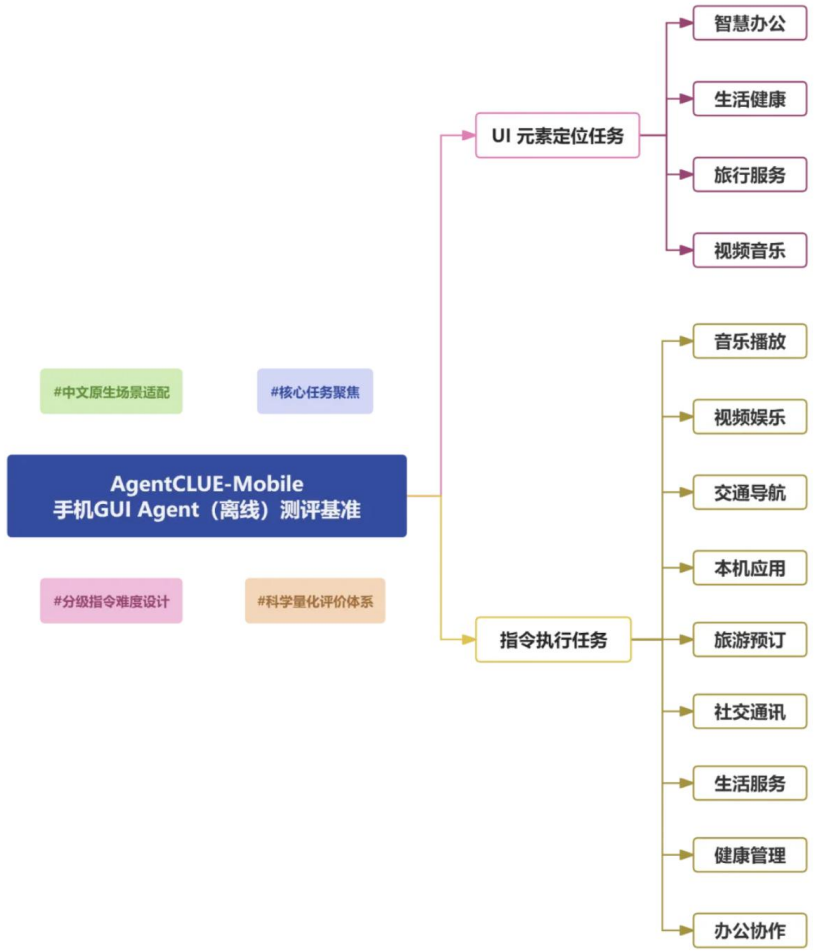
数据来源：SuperCLUE，2025年11月28日。
注：本榜单将相差一分以内的模型视为并列名次；海外模型仅作对比参考，不参与排名。

测评详情可访问下方链接：
https://mp.weixin.qq.com/s/BMp5FYEbHa6bNp9VY6MJ0w

Agent系列基准介绍：AgentCLUE-Mobile

AgentCLUE-Mobile：手机GUI Agent测评

AgentCLUE-Mobile 二期测评聚焦中文原生场景，任务基于国内用户日常及办公、健康类典型使用场景开发，贴合中文用户操作习惯，聚焦手机 GUI Agent 的 UI 元素定位与指令执行两大核心能力展开全面考察智能体在九大核心场景及新增智慧办公、生活健康等拓展场景的表现，通过跨应用难题与优化评分体系，精准衡量其手机端离线智能交互的实际水平与综合应用潜力。



测评结果分析

AgentCLUE-Mobile
手机 GUI Agent（离线）基准测评总榜概览

排名	模型	机构	总分	UI元素定位得分	指令执行得分	使用方式
1	Nebula-GUI-V2	中兴通讯	**92.27**	**98.40**	**88.18**	API
2	Doubao-Seed-1.6-thinking-250715	字节跳动	**89.86**	**96.40**	**85.50**	API
3	GLM-4.5v	智谱AI	**88.37**	94.80	**84.08**	API
4	MiMo-VL-7B-RL-2508	小米集团	**84.39**	87.95	82.01	模型
5	qwen2.5-vl-7b-instruct	阿里巴巴	**72.35**	**94.87**	57.34	模型
6	ui-tars-1.5-7b	字节跳动	**66.16**	79.92	56.99	API
7	GUI-Owl-7B	阿里巴巴	**64.21**	94.19	44.23	模型
-	Gemini-2.5-pro	Google	**60.87**	58.00	62.79	API
8	qwen2.5-vl-3b-instruct	阿里巴巴	**57.93**	89.92	36.61	模型
9	AgentCPM-GUI	面壁智能	**54.39**	60.08	50.59	模型
10	MiniCPM-V4.5-8B	面壁智能	**39.14**	74.00	15.90	模型
-	Gemma3-4B-it	Google	**8.94**	8.00	9.56	API

排名计算方式说明：为减少波动影响，榜单将分差1分值内的模型视为并列排名。国外模型及补测模型的旧版本不参与排名，只做参考。
数据来源：SuperCLUE，2025年12月28日。

测评详情可访问下方链接：https://mp.weixin.qq.com/s/8dR7ioETwjLAjll3mOSIqA

1.国产大模型能力梯度清晰，头部阵营以技术突破领跑行业，表现亮眼。

头部产品中，中兴通讯的 Nebula-GUI-V2（92.27 分）、字节跳动的 Doubao-Seed-1.6-thiking-250715（89.86分）居前，UI 元素定位与指令执行能力均衡。

2.两大核心能力的关联性兼具规律性特征与个体差异化表现。

UI 元素定位是手机 GUI Agent 的基础能力，通常 UI 元素定位得分越高，指令执行得分也越高。

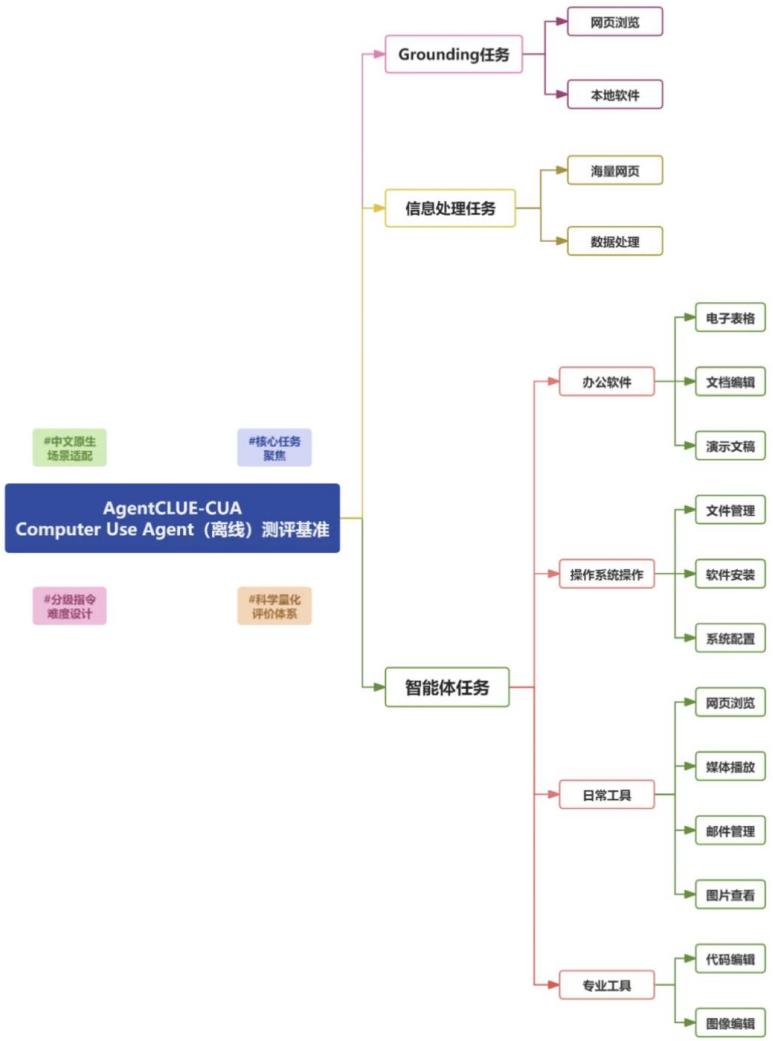
3.各产品综合指令执行质量分化显著，头部与尾部差距悬殊。

头部产品如 Nebula-GUI-V2、MiMo-VL-7B-RL-2508 等，任务完成度、平均动作类型准确率和平均动作细节准确匹配率均较高；而尾部产品像 Gemma3-4B-it、MiniCPM-V4.5-8B 等，任务完成度极低，部分甚至为0，整体综合指令执行质量亟待提升。

Agent系列基准介绍：AgentCLUE-CUA

AgentCLUE-CUA：Computer Use Agent测评

AgentCLUE-CUA是 Computer Use Agent （离线）测评的专项方案，旨在构建科学、全面的测评体系，精准评估 CUA 的核心能力，明确技术发展方向，为用户选择产品提供可靠依据，同时推动该领域技术的规范化、高质量发展。



测评结果分析

AgentCLUE-CUA
Computer Use Agent（离线）基准测评总榜概览

排名	模型	机构	总分	Grounding	信息处理	智能体	使用方式
	qwen3-vl-235b-a22b-thinking	阿里巴巴	**87.37**	**93.26**	**87.50**	**85.36**	API
	GLM-4.5v	智谱AI	**84.49**	**93.26**	**84.38**	**81.61**	API
	doubao-seed-1-6-vision-250815	字节跳动	**75.72**	**91.95**	**78.13**	**69.51**	API
-	Gemini-2.5-pro	Google	**59.91**	57.95	90.32	50.42	API
4	MiniCPM-V4.5-8B	面壁智能	**47.36**	67.42	56.25	37.71	模型
5	GUI-Owl-7B	阿里巴巴	**42.19**	91.01	68.75	17.06	模型
-	claude-sonnet-4.5	Anthropic	**31.57**	10.11	50.00	32.57	API
6	ui-tars-1.5-7b	字节跳动	**27.18**	60.67	46.88	9.45	API

排名计算方式说明：为减少波动影响，榜单将分差1分值内的模型视为并列排名。国外模型不参与排名，只做参考。
数据来源：SuperCLUE，2025年10月30日。

测评详情可访问下方链接：
https://mp.weixin.qq.com/s/kBnAj8FzOW138WIlmoTNgg

1.Grounding 是 CUA 基础能力，通常与智能体任务得分正相关。

得分最高的模型 qwen3-vl-235b-a22b-thinking 总分达 87.37 分，而尾部如 ui-tars-1.5-7b 等模型，总分仅 27.18 分。这清晰表明部分头部模型已具备较强的智能交互能力，但仍有大量尾部模型需在相关核心能力上进一步优化。

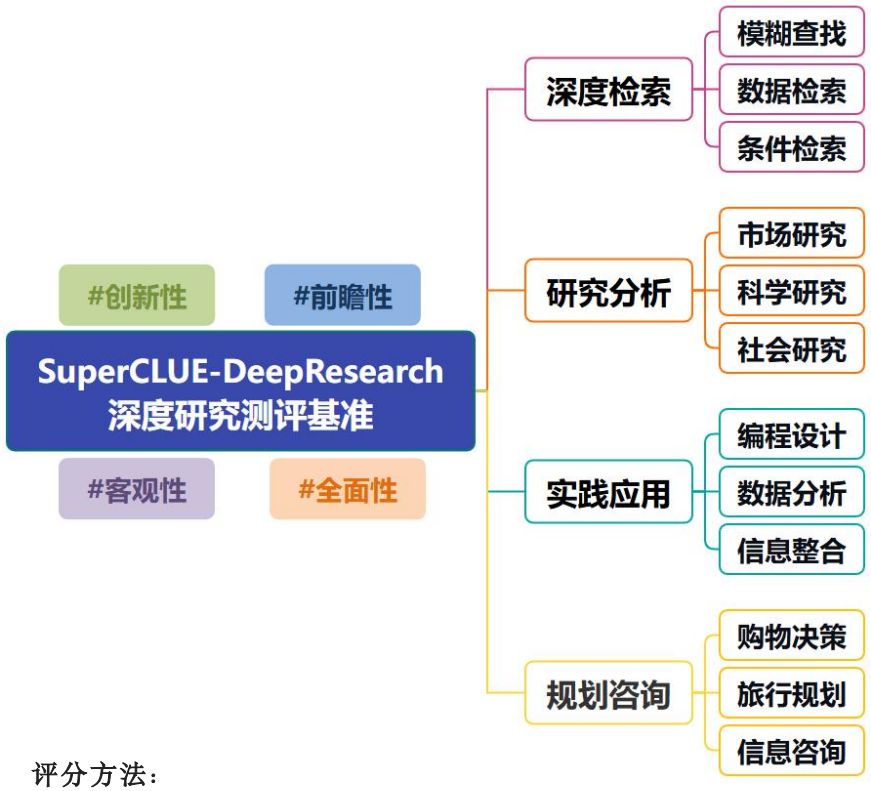
2.不同模型在综合指令执行质量上分化显著，头部与尾部差距巨大。

头部模型如 qwen3-vl-235b-a22b-thinking、GLM-4.5v 等，任务完成度、平均动作类型准确率和平均动作细节准确匹配率均较高；而尾部模型像 GUI-Owl-7B、ui-tars-1.5-7b 等，任务完成度均为 0，整体综合指令执行质量亟待提升。

Agent系列基准介绍：SuperCLUE-DeepResearch

SuperCLUE-DeepResearch：中文深度研究测评基准

SuperCLUE-DeepResearch 是一个专为深度研究产品设计的评测基准，旨在为深度研究领域提供全面且多维的能力评估参考。DeepResearch代表了AI从简单的信息检索向更高级的自主研究代理迈进的趋势，越来越多的DeepResearch产品出现在我们的视野中。为了全面客观地衡量各个深度研究产品的能力，我们推出了SuperCLUE-DeepResearch测评基准。



评分方法：

本次SuperCLUE-DeepResearch深度研究测评基准针对不同任务设置了不同的评价标准，以便更加客观公正地对产品的能力进行评价。由于评价标准的篇幅过长，不便展示，如需进一步了解可点击右方链接跳转至详细的测评文章。

测评结果分析

SuperCLUE-DeepResearch 中文深度研究测评总榜

排名	产品名称	机构	总分	深度检索	研究分析	实践应用	规划咨询
-	Deep Research(Plus)	OpenAI	**76.37**	63.64	97.33	64.09	82.58
-	Deep Research(Pro)	Google	**64.80**	27.27	98.22	46.21	92.74
-	Deep Research(Pro)	Perplexity	**62.26**	36.36	78.67	74.64	64.82
🥇	Kimi Researcher(内测)	月之暗面	**58.65**	**36.36**	**92.89**	**30.73**	**77.46**
🥈	豆包 深入研究	字节跳动	**54.27**	**27.27**	**80.22**	**37.22**	**75.94**
-	DeeperSearch(SuperGrok)	X.AI	**50.19**	27.27	65.33	36.67	73.94
🥉	秘塔AI搜索(研究模式)	秘塔科技	**49.35**	9.09	**81.56**	**37.40**	**75.40**
4	夸克 深度研究	阿里巴巴	**46.23**	9.09	**85.78**	22.78	**72.60**
5	智谱清言 沉思	智谱AI	**41.78**	0.00	80.89	**29.44**	63.66

数据来源：SuperCLUE，2025年6月30日。
注：1. 本榜单将相差一分以内的产品视为并列名次；海外产品仅作参考，不参与排名；2.OpenAI Deep Research使用底层模型为o3微调的版本进行测评，Google Deep Research使用底层模型为Gemini 2.5 Pro的版本进行测评，秘塔AI搜索使用长思考·R1+研究的先想后搜模式进行测评。

测评详情可访问下方链接：

https://mp.weixin.qq.com/s/sdOPMHf0gH2p8s1QQnn2ZQ

1.各深度研究产品表现存在显著差异。

OpenAI的深度研究产品以76.37分的总分位居榜首，与排名末位的产品分差达34分之多。Kimi Researcher以58.65分位于国内第一，研究分析任务表现十分亮眼，与排名末位的产品差距也接近17分。

2.国内外产品性能差距明显。

4款国外深度研究产品的平均得分为63.41分，显著高于5款国内产品的平均分50.06分，两者分差达13.35分，反映出明显的性能差距。

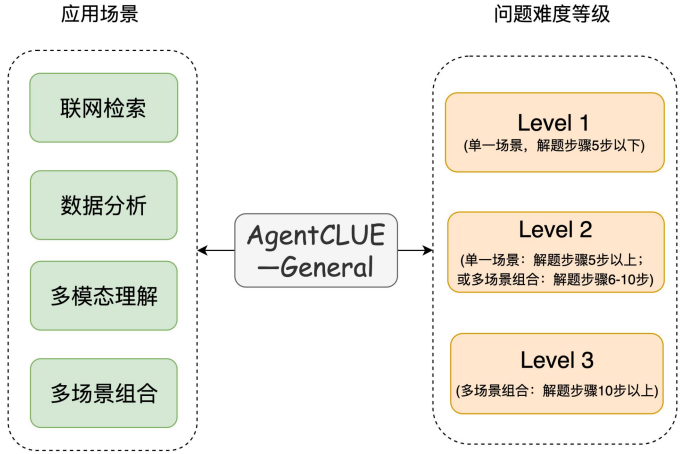
3.产品在不同任务类型表现分化显著。

研究分析类任务平均得分高达84.54分，而深度检索类任务平均分仅为26.26分。具体而言，当前深度研究产品在内容创作、报告输出等生成类主观任务上表现优异；但在需要深度搜索、大数据整合等复杂客观任务方面，仍存在较大提升空间。

Agent系列基准介绍：AgentCLUE-General系列

AgentCLUE-General：中文通用AI智能体基准

AgentCLUE-General是专注于中文通用AI智能体的测评基准。该基准立足中文应用场景，聚焦现实世界中可借助工具解决的实际问题，全面评估智能体在联网检索、数据分析、多模态理解和多场景组合四个核心应用场景的表现，并对任务根据难度进行了三个等级划分。



评分方法：

AgentCLUE-General为每个测试题目给出一个唯一的正确答案，通过人工对比Agent的答案和正确答案是否一致，来判断回答是否正确，回答正确得1分，错误得0分。对于因为智能体产品本身不支持上传文件而导致无法获取答案，也计0分。

总分计算：

我们对不同难度的题目赋予不同的重要性程度。Level 1的题目权重为1；Level 2的权重为2；Level 3的权重为3，模型的总分和每个应用场景下的总分都通过加权计算得到。具体计分规则如下：

Level 1难度的题目数量：A；Level 1产品答对的题目数量：M；
Level 2难度的题目数量：B；Level 2产品答对的题目数量：N；
Level 3难度的题目数量：C；Level 3产品答对的题目数量：Q

总分 = (M+2*N+3*Q) / (A+2*B+3*C) * 100

测评结果分析

AgentCLUE-General
中文通用AI智能体基准测评总榜

排名	模型	机构	总分	Level 1	Level 2	Level 3
🥇	Manus(Starter)	Monica	**38.46**	**71.43**	**35.00**	**33.33**
🥈	Coze(探索版)	字节跳动	**32.31**	**42.86**	**30.00**	**33.33**
🥉	Genspark(Plus)	MainFunc	**30.77**	**42.86**	**35.00**	**16.67**
4	OWL	Camel-AI	12.31	14.29	10.00	**16.67**
5	ChatGPT官网 (O4-mini-high)	OpenAI	9.23	0.00	**15.00**	0.00
5	Fellou	Fellou AI	9.23	**28.57**	10.00	0.00
6	Operator	OpenAI	4.62	14.29	5.00	0.00

数据来源：SuperCLUE，2025年4月30日。

测评详情可访问下方链接：
https://mp.weixin.qq.com/s/9wePg3wK5zNwAfdUTMOh-g

1.整体能力仍处基础阶段，头部产品表现相对领先。

当前参评的通用 AI 智能体在现实世界任务上的整体能力普遍偏弱，最高得分产品 Manus(Starter) 总分也未超过 40 分（具体为38.46 分），表明通用 Agent 技术仍处于比较基础的发展阶段，与理想状态差距较大。

2.不同难度任务能力差异显著，复杂多步骤任务是主要瓶颈。

智能体在相对简单的 Level 1 任务上表现尚可（如最高分Manus得分71.43），但随着任务难度提升至 Level 2（最高分35分） 和 Level 3（涉及更多步骤和复杂推理，最高分33.33分），智能体的得分率普遍大幅下降，处理复杂现实世界任务的能力是当前面临的主要挑战。

3.能力分布不均，结构化数据分析和通用联网检索相对突出，多模态、非结构化数据及多场景组合是显著短板。

智能体擅长处理 Excel 等结构化数据和进行日常联网检索（Manus 和 Coze 在联网检索上得分 60.00），但在非结构化文本数据处理能力不足，且在涉及图片、音频、视频等多模态任务及能力组合的多场景任务上表现尤为薄弱（多模态理解场景最高得分仅 21.43，多场景组合最高分 36.36）。

Agent系列基准介绍：AgentCLUE-General系列

AgentCLUE-tGeneral：中文通用AI智能体基准

在AgentCLUE-General的测评中，我们注意到部分智能体产品不支持上传文件或者上传文件有格式、文件大小的限制，同时考虑到智能体产品的不断涌现，以及能力不断进化，我们计划启动新的中文通用智能体测评**AgentCLUE-tGeneral。**

AgentCLUE-tGeneral测评基准定位为**纯文本输入**(名称中的t，代表文本输入)，无文件上传，输入方式更加纯粹，预期可以测评更广泛的智能体。

测评方法：

1. 评估流程：

获得问题、模型回复（文本答案、pdf报告或代码脚本等）和标准参考答案 --> 依据评分标准评价每一题的分数 --> 计算模型最终得分

2. 评分方法：为了确保评估的科学性和公正性，我们采用超级大模型进行评价。结合评估流程、评估标准、评分规则，进行细粒度评估。针对pdf报告，代码脚本文件等，均直接使用原始文件直接发送给超级模型做评价。应用这种方式，减少人为因素的干预，确保评分结果的客观性和一致性。

3.总分计算：

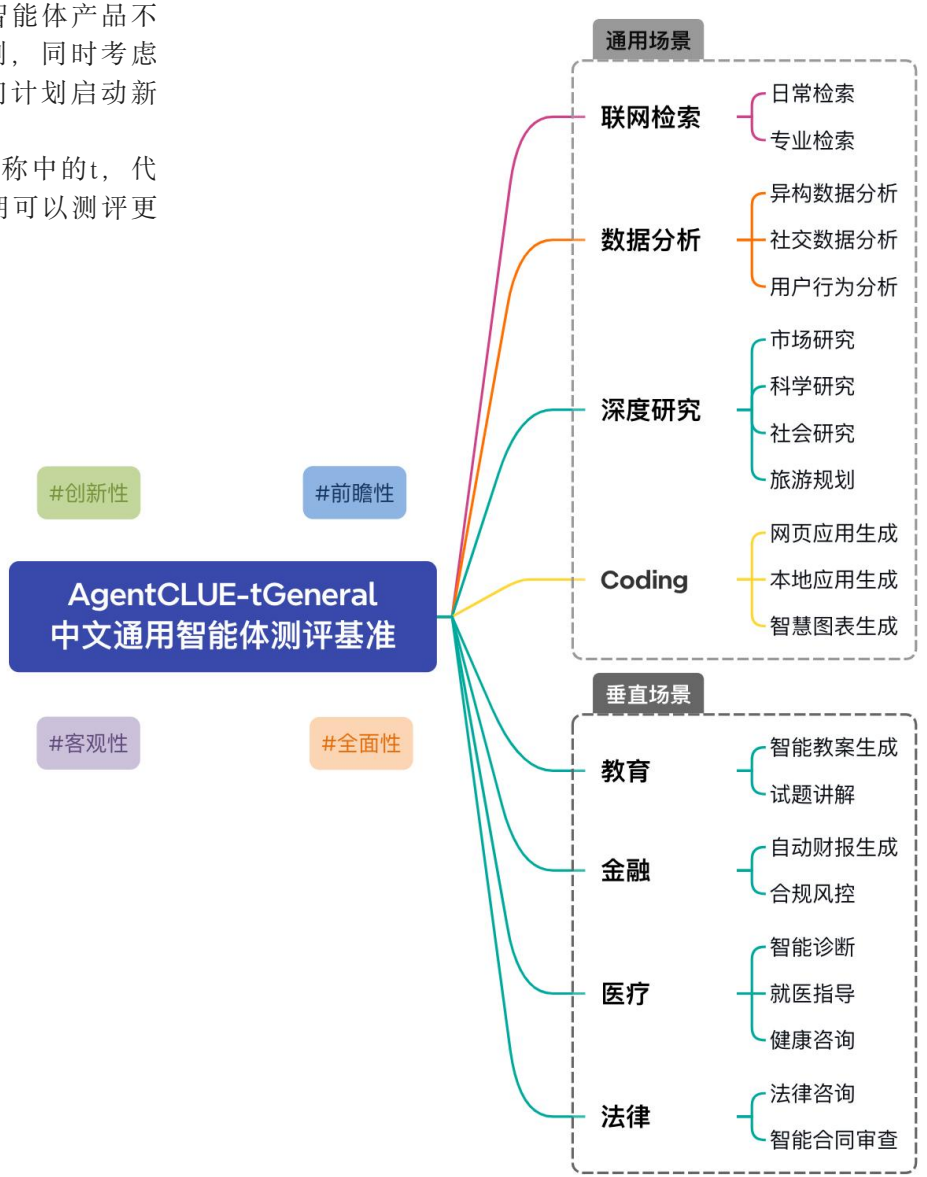
最终的总分计算：计算所有题目的平均分作为最终总分。

总分 = 八大场景总分的平均分

对每个场景的总分计算：

场景总分 = 该场景下的所有题目的平均分转化为百分制

测评方案要点



1.中文原生场景构建。

本测评采用全中文数据集，所有生成创作任务均基于典型中文使用场景设计，充分贴合国内用户的实际需求和使用习惯。

2.兼顾通用与垂直场景的多维任务体系设计。

测评不仅涵盖联网检索、数据分析、深度研究、Coding等较为通用的任务，同时包含教育、金融、医疗、法律四个垂直场景，通过多维度评估全面考察智能体产品的能力。

3.纯文本输入，测评产品更全面广泛。

所有输入问题仅包含纯文本，不涉及文件上传，只要支持文本输入即可参与测评，避免部分智能体产品因不支持上传文件、或文件上传的格式、大小限制而导致无法测评。

测评详情可访问下方链接：

https://mp.weixin.qq.com/s/qByF2RsiL7ZeEnInPsdoXQ

Coding系列基准介绍：SuperCLUE-SWE

SuperCLUE-SWE：中文软件工程测评基准

SuperCLUE-SWE项目构建了一个面向中文开发环境的软件工程评测基准，借鉴 SWE-bench 的构建理念，收集了来自中文开源项目的真实 GitHub 问题及其对应的修复方案，确保任务实例既真实可靠，又贴近中文开发场景。

测评方案要点

1.中文开发场景适配

本评测基准专为中文开发环境设计，**涵盖开源项目中各种的真实 GitHub中文问题（issue）**及其对应的拉取请求（PR）。任务实例涉及中文开发者常用的编程框架，确保评测结果贴近中文开发者的实际需求。

2.核心任务聚焦

聚焦于模型在真实开发环境中解决实际问题的能力，评估模型在理解中文问题描述、定位代码缺陷、生成修复方案以及验证修复效果等方面的表现。

3.任务难度分级

根据任务涉及代码的行数，将问题划分为**简单（1-20行）、中等（20-50行）和困难（50行以上）**三类，涵盖从单一文件修改到多文件、多步骤、多语义依赖的开发挑战。各个难度题目比例为：30%简单，40%中等，30%困难。精准区分模型在不同复杂度任务中的表现。

4.科学量化评价体系

建立明确的量化评分标准，采用 1 分制对任务表现进行评估。确保评价结果客观、可对比，为模型性能提升提供明确的方向。

测评结果分析

SuperCLUE-SWE
中文「软件工程」测评基准——bug修复能力榜单

排名	模型名称	机构	总分	给出可应用答案	未生成可应用答案	使用方式	测评时间
-	Gemini-3-Pro-Preview	Google	50.00	90.00	10.00	API	2025.12.02
-	Claude-opus-4-5	Anthropic	47.00	89.00	11.00	API	2025.12.09
-	Gemini-3-Flash-Preview	Google	45.00	83.00	17.00	API	2025.12.17
-	Claude-sonnet-4-5	Anthropic	40.00	85.00	15.00	API	2025.12.02
-	GPT-5(high)	OpenAI	35.00	63.00	37.00	API	2025.12.02
-	GPT-5.2(high)	OpenAI	28.00	61.00	39.00	API	2025.12.15
-	GPT-5.1(high)	OpenAI	27.00	60.00	40.00	API	2025.12.09
🥇	DeepSeek-V3.2	DeepSeek	25.00	62.00	38.00	API	2025.12.02
🥇	Kimi K2 Thinking	月之暗面	24.00	72.00	28.00	API	2025.12.02
🥈	Qwen3-Coder	阿里巴巴	20.00	86.00	14.00	API	2025.12.02
🥈	MiMo V2 Flash	Xiaomi MIMO	19.00	81.00	19.00	API	2025.12.17
🥉	GLM-4.6	智谱AI	18.00	73.00	27.00	API	2025.12.02
🥉	DeepSeek-V3.2-Exp-Thinking	DeepSeek	18.00	78.00	22.00	API	2025.12.02
4	Doubao-Seed-Code-Preview-251028	字节跳动	16.00	39.00	61.00	API	2025.12.02

数据来源：SuperCLUE，2025年12月17日。
注：考虑到波动影响，本排行榜将相差1分以内的模型视为并列名次。

测评详情可访问下方链接：
https://mp.weixin.qq.com/s/EBNAndUQ5066wYP57XjBYA

1.国际头部模型仍占优，但领先优势在收窄。

在本次 SuperCLUE-SWE 测评中，Gemini-3-pro 继续保持总体第一，Claude、GPT-5.1 等构成国际第一梯队，在复杂、多文件的真实 Issue 修复任务上表现最稳定、成功率最高。不过，相比早期英文基准上的“断层式”优势，这次在中文软件工程场景下，头部国际模型与国内强势模型之间的分差已经明显缩小。

2.国内代表性模型整体跟上车队，个别模型开始逼近国际第一梯队，值得重点跟踪。

从成绩看，国内主流模型在中低难度任务上，多款模型的得分已经基本追平国际模型，只是在高难度任务上拉开差距。尤其是 DeepSeek-V3.2 正式版相对 V3.2 实验版提升约 8 分，仅用几个月就把成绩追到接近 GPT-5.1，说明国内模型在代码理解与修复方面的迭代速度非常快，已具备向第一梯队发起冲击的实力。

Coding系列基准介绍：Agentic Coding

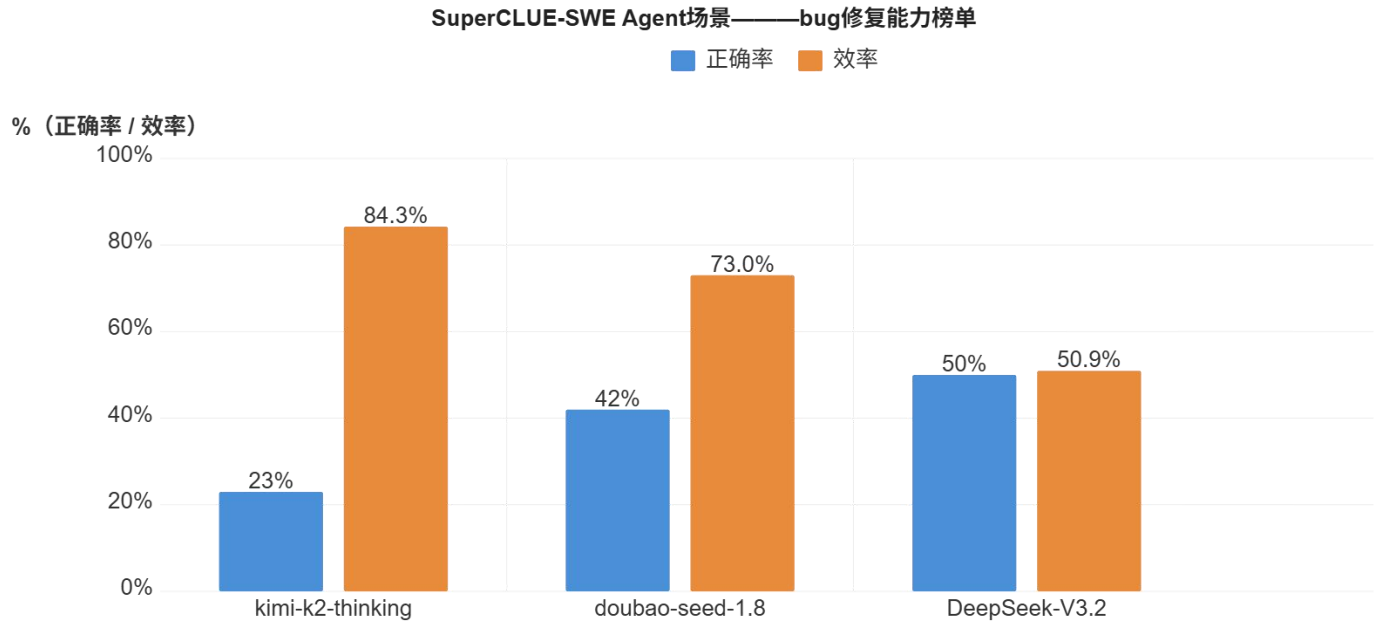
Agentic Coding：Agent 版 SC-SWE（SC-Agent）

Agent 版 SC-SWE（SC-Agent）测评是一种面向 Agentic Coding 场景的代码能力评估方案，用于评估大语言模型在真实软件工程环境中的交互式问题解决能力。

评测以 SC-SWE 软件工程问题集为基础，该数据集由真实代码仓库中的中文“修 Bug/补功能”任务构成；在测评过程中，SC-Agent 将模型视为具备感知、决策与执行能力的行动体，使其能够在完整代码仓库内开展多轮迭代，包括代码检索、文件编辑与测试执行，并根据环境反馈持续改进。评测框架提供 bash、str_replace_editor 与 submit 等工具以支持命令执行、代码修改和补丁提交；最终以各题目对应工程的单元测试是否全部通过作为成功判据，从而衡量模型在真实工程任务中的解决能力与工程行为质量（兼顾正确性与效率）。

测评详情可访问下方链接：
https://mp.weixin.qq.com/s/DGuOYq0TUQQGESnMumnf4w

测评结果分析



本文章通过答题的正确率和效率两个维度对三个模型进行了全面评估。在SC-SWE数据集上的测试结果如下：

DeepSeek：解答了最多的题目，共50道，展现了强大的广度覆盖能力。这一表现表明，DeepSeek能较好地应对多种类型的题目，尤其是在题目多样化的情境下，能够保证较高的题目解决量。尽管其平均步数较长，但在任务量上具有较明显的优势。

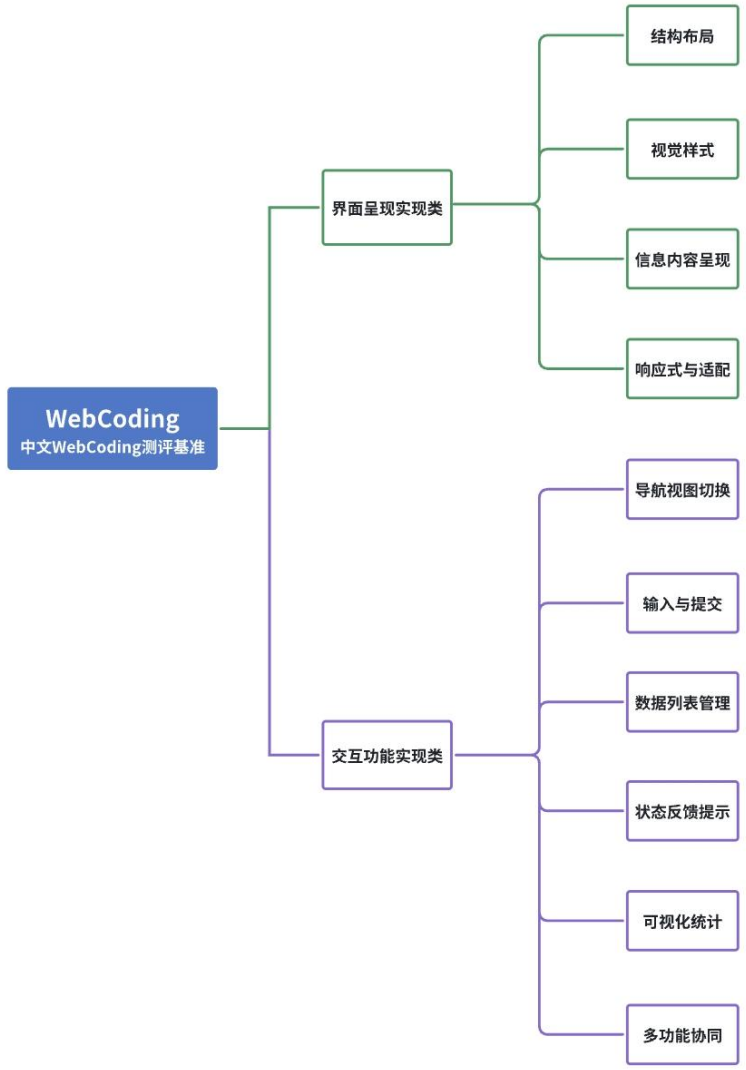
Doubao：紧随其后，共解答了42道题目，表现同样出色，尤其在简单题目上展现了较高的正确率。Doubao的解题策略可能更多依赖于对简单问题的快速识别和处理，在保证准确率的同时，也能在短时间内完成解答任务。这使得Doubao在面对大规模数据时，能够在快速答题和准确性之间找到良好的平衡。

Kimi：在效率上表现最为突出，平均步数仅为21，显著低于其他模型。这表明Kimi能够在较少的步骤内快速完成任务，具有极高的处理效率。其优势在于能够迅速收敛到正确答案，适合在对时间和计算资源要求较高的场景下使用，能有效节省时间和计算成本。

Coding系列基准介绍：SuperCLUE-WebCoding

SuperCLUE-WebCoding：中文前端开发测评基准

中文WebCoding测评基准（SuperCLUE-WebCoding），旨在分析不同模型在真实Web开发任务中的需求贴合性、功能正确性等性能，为模型编程能力开发提供精准指引。



测评结果分析

SuperCLUE-WebCoding
中文「前端开发」能力榜单

排名	模型名称	机构	总分
1	MiniMax M2.1	MinMax	75.66
-	Gemini-3-Flash-Preview	Google	72.57
-	Gemini-3-Pro-Preview	Google	72.49
-	Grok 4	X.AI	72.22
2	Qwen3-Max	阿里巴巴	71.88
-	Claude-Opus-4.5-Reasoning	Anthropic	69.80
3	Kimi K2 Thinking	月之暗面	68.90
3	Qwen3-Coder	阿里巴巴	67.93
3	GLM-4.6	智谱AI	67.92
4	GLM-4.7	智谱AI	66.87
5	DeepSeek-V3.2-Thinking	深度求索	64.37
-	GPT-5.2(high)	OpenAI	65.25
6	Qwen3-235B-A22B-Thinking-2507	阿里巴巴	60.81
7	Doubao-Seed-1.6-251015(Thinking)	字节跳动	39.85

数据来源：SuperCLUE-WebCoding，2025年12月24日
注：考虑到波动影响，本排行榜将相差1分以内的模型视为并列名次。

1.国产大模型【MiniMax M2.1】在本次中文「前端开发」基准测评中以高分差超越原第一梯队诸模型斩获头名。
2.【Gemini 3 系列】发挥稳定，依旧位于一流模型行列。

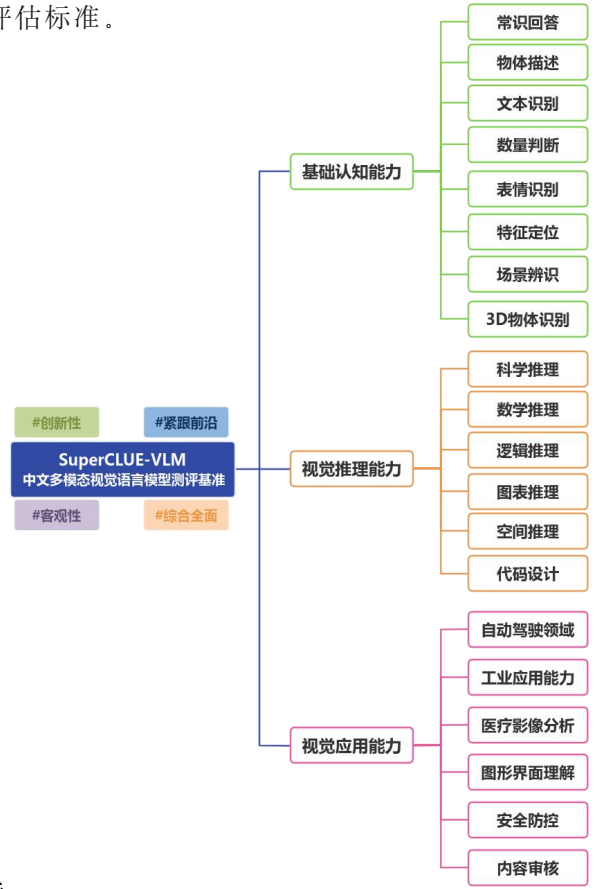
评分方法：
WebCoding测评采用端到端（E2E）功能测试，借助playwright（自动化测试库）、浏览器进行自动化功能测试，对每个测试用例进行0/1评分，对于一个测试用例，通过则记为1分，不通过则记为0分。

测评详情可访问下方链接：
https://mp.weixin.qq.com/s/JY7i-Iy93QCuJrtniG3C-Q

多模态系列基准介绍：SuperCLUE-VLM

SuperCLUE-VLM：中文多模态视觉语言测评基准

中文多模态视觉语言模型测评基准（SuperCLUE-VLM）基于中文场景特点，围绕基础视觉认知、视觉推理和视觉应用三大核心维度构建评测体系，力求为多模态视觉语言模型的发展提供客观、公正的评估标准。



评分方法：

本次测评以回答准确性作为唯一评判标准。每道题目都配有标准参考答案。为了确保评估的科学性和公正性，我们采用评价模型（Gemini-2.5-Flash），将模型的回答与参考答案进行对比，从而判断其正确性。应用这种方式，尽量减少人为因素的干预，确保评分结果的客观性和一致性。

测评结果分析

SuperCLUE-VLM
多模态视觉语言基准测评总榜

排名	模型名称	机构	总分	基础认知	视觉推理	视觉应用	开/闭源
-	Gemini-3-pro	Google	83.64	89.01	82.82	79.09	闭源
1	SenseNova V6.5 Pro-20251215	商汤科技	75.35	81.66	74.31	70.08	闭源
2	Doubao-seed-1-6-vision-250815	字节跳动	73.15	82.70	64.27	72.48	闭源
3	ERNIE-5.0-Preview	百度	72.21	82.05	70.86	63.71	闭源
4	Qwen3-vl-235b-a22b-thinking	阿里巴巴	71.95	79.66	71.26	64.92	开源
-	Claude-opus-4-5-20251101	Anthropic	71.44	82.07	65.81	66.43	闭源
-	GPT-5.2(high)	OpenAI	69.16	75.18	67.35	64.96	闭源
4	Doubao-seed-1-6-251015	字节跳动	68.02	77.86	62.01	64.19	闭源
4	GLM-4.6v	智谱AI	67.68	81.74	60.83	60.48	开源
5	step-3	阶跃星辰	62.94	77.16	49.81	61.87	开源
6	Qwen3-vl-8b-instruct	阿里巴巴	61.64	74.66	52.77	57.50	开源
7	MiniCPM-V4.5	面壁智能	49.38	68.76	24.83	54.56	开源
8	InternVL3.5-8B	上海AI Lab	47.89	64.21	24.49	54.98	开源
-	Grok-4.1-fast-non-reasoning	X.AI	45.57	63.96	23.62	49.13	闭源

数据来源：SuperCLUE，2025年12月29日。
注：考虑到波动影响，本排行榜将相差1分以内的模型视为并列名次。

测评详情可访问下方链接：
https://mp.weixin.qq.com/s/2ZoQmq8isDJEgX0SMuHF-w

1.Gemini-3-pro领跑全球，国产模型紧随其后！

Google 的 Gemini-3-pro 以83.64分占据榜首，而国产模型表现亮眼 —— 商汤科技 SenseNova V6.5 Pro以75.35分领跑国内阵营，字节跳动 Doubao-seed-1-6-vision、百度 ERNIE-5.0-Preview 等也紧随其后。

2.二级任务能力：基础感知成熟，复杂领域待突破。

多模态模型在环境辨识等基础感知任务中表现优异，基础认知已趋成熟；但空间推理、医疗影像分析等复杂/专业任务得分极低，深度推理与垂直场景能力仍是短板，呈现“基础强、复杂弱”的不均衡态势。

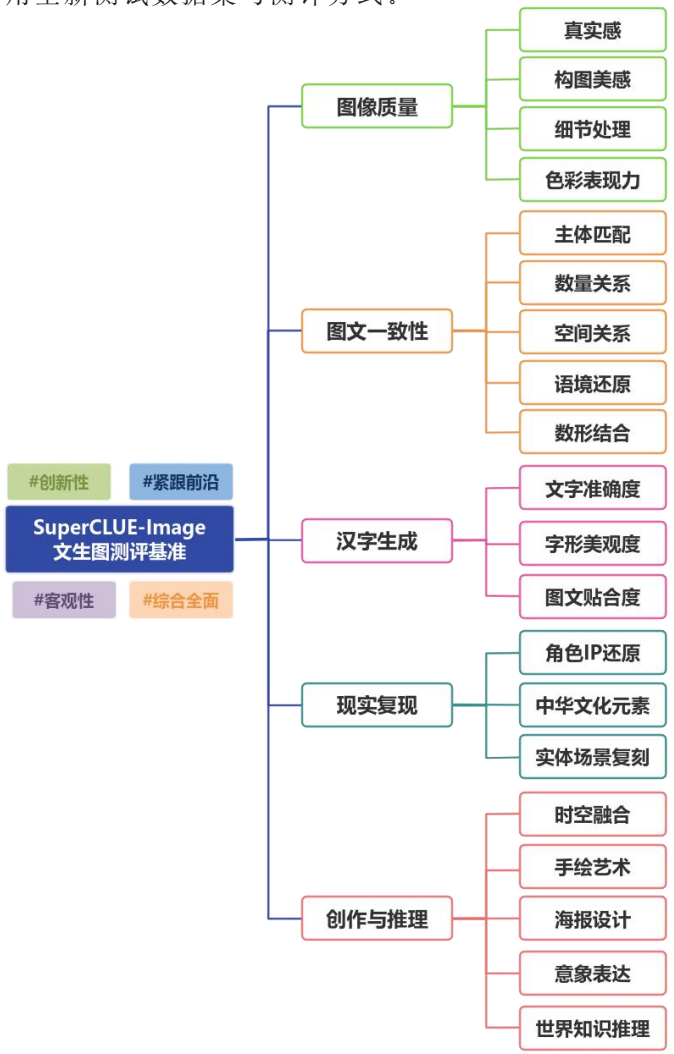
3.闭源模型整体上显著领先开源模型。

闭源模型（如 Google Gemini-3-pro、商汤 SenseNova V6.5 Pro 等）总分普遍超 70，头部闭源模型得分更是突破 80，整体实力领跑；开源模型中，阿里 Qwen3-vl-235b-a22b-thinking（71.95 分）、智谱 GLM-4.6v（67.68 分）等头部开源模型表现不错，但多数开源模型（如 MiniCPM-V4.5、InternVL3.5-8B）总分不足 50，与闭源模型仍存在明显差距。

多模态系列基准介绍：SuperCLUE-Image

SuperCLUE-Image：中文文生图测评基准

中文原生文生图模型测评基准（SuperCLUE-Image）锚定中文场景适配需求与模型综合实力，沿用“基础能力 + 应用能力”多维框架，覆盖图像质量、现实复现、创作与推理等关键维度，同时启用全新测试数据集与测评方式。



测评结果分析

SuperCLUE-Image
文生图基准测评总榜

排名	模型名称	机构	总分	基础能力			应用能力		使用方式
				图像质量	图文一致性	汉字生成	现实复现	创作与推理	
-	Gemini-3-pro-image（Nano Banana Pro）	Google	76.20	80.35	52.22	79.55	86.79	82.07	API
🥇	Seedream 4.0	字节跳动	66.87	73.87	31.67	76.90	75.06	76.85	API
-	GPT-image-1	OpenAI	66.84	72.16	36.11	68.87	78.33	78.74	API
🥈	Wan2.5-t2i	阿里巴巴	66.63	74.60	35.56	72.76	74.97	75.25	API
🥉	Pangu-T2I	华为	62.05	71.77	27.22	73.60	70.11	67.54	API
-	FLUX-2-pro	Black Forest Labs	60.61	74.45	31.67	64.42	66.78	65.71	API
🥉	Hunyuan-image-3	腾讯	58.35	73.28	22.78	64.76	75.08	55.83	API
🥉	Qwen-Image	阿里巴巴	57.78	62.59	21.11	71.68	67.08	66.43	API
4	Kling-v2-1	快手	56.29	64.60	14.44	71.28	67.06	64.08	API
5	ERNIE-iRAG-1.0	百度	45.51	50.61	3.89	65.31	60.86	46.87	API
6	CogView-4-250304	智谱	40.88	52.49	8.89	45.29	56.50	41.22	API
7	Vivago-2.1	智象未来	36.11	58.78	10.56	7.08	59.61	44.55	模型
-	Stable Diffusion 3.5 Large	Stability AI	30.74	49.93	1.67	11.98	52.42	37.70	API
-	Recraft v3	Recraft	27.69	45.96	1.67	8.67	47.36	34.81	API

数据来源：SuperCLUE，2025年12月5日。
注：考虑到波动影响，本排行榜将相差1分以内的模型视为并列名次。

测评详情可访问下方链接：
https://mp.weixin.qq.com/s/_jqt-Mv9N0JZ29NBmOS1Tg

1.差距与突破并存！Nano Banana Pro 一骑绝尘，国内头部模型紧追不舍。

本次文生图月榜中，Nano Banana Pro 以76.20分大幅领跑，拉开与其他模型的差距；国内头部厂商表现亮眼，Seedream 4.0、Wan2.5-t2i、Pangu-T2I均跻身前五。整体来看，国内头部模型在文生图领域的竞争力不俗。

2.汉字生成维度，国产模型展现明显优势。

在汉字生成能力测评中，国际模型表现普遍疲软，而国产模型集体领跑，字节 Seedream 4.0、阿里 Wan2.5-t2i、华为 Pangu-T2I均拿下 70 + 高分，生成的汉字清晰度高、辨识度强，中文适配性成为国产模型的突出长板。

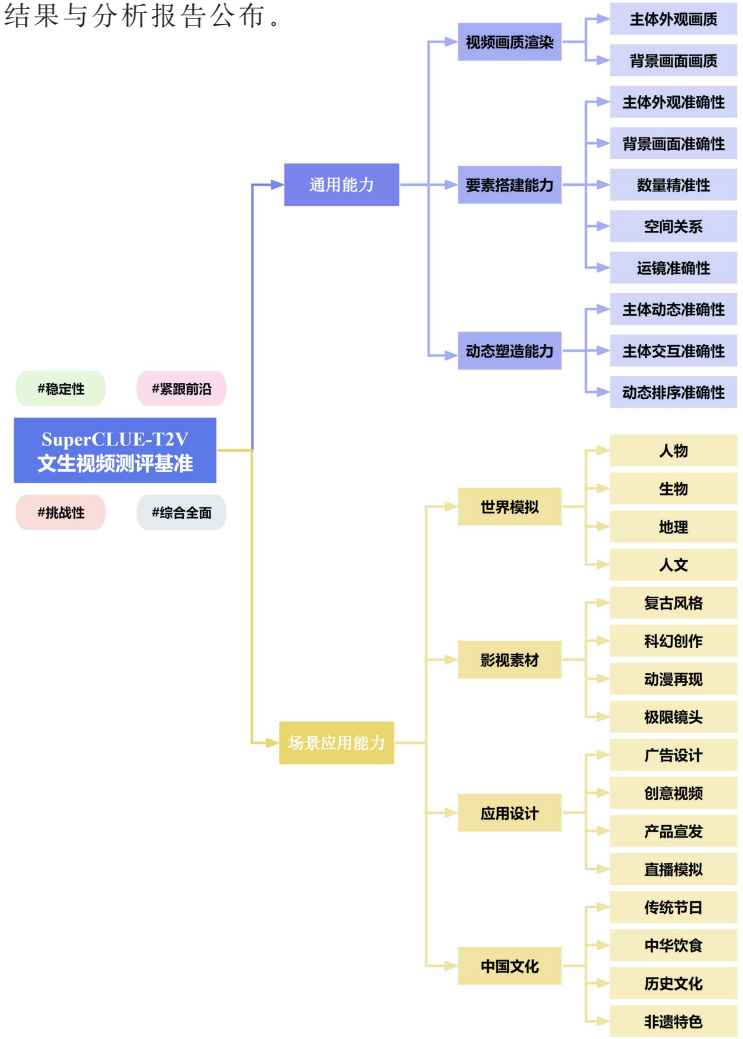
3.文生图模型基础与推理能力上国产图像质量突围，国际模型逻辑领跑。

基础能力维度中，国产模型在“图像质量”上已实现突围：阿里 Wan2.5-t2i的画面精细度超过 OpenAI 的 GPT-image-1，但“图文一致性”是明显短板。而“创作与推理”维度则是国际模型的优势场：Nano Banana Pro、GPT-image-1的逻辑连贯性更强，国产模型虽有 Seedream 4.0等中游表现，但多数得分低于70，在复杂场景的创作逻辑上仍需提升。

多模态系列基准介绍：SuperCLUE-T2V

SuperCLUE-T2V：中文文生视频测评基准

SuperCLUE-T2V新版文生视频测评方案通过提升题目的难度和精确度增加了本次测评的挑战性，从通用能力和场景应用能力两个方面对17个国内外文生视频产品进行了严格的测试与评估，并通过三次测评求平均值的方式提升测评的稳定性和可靠性，现将评估结果与分析报告公布。



测评结果分析

SuperCLUE-T2V
文生视频基准测评总榜

排名	模型名称	所属机构	总分	通用能力	场景应用	使用方式
-	veo-3.0-generate-preview	谷歌	**55.40**	**50.79**	**60.01**	API
	Hailuo-02	MiniMax	**51.67**	47.09	**56.25**	API
	Doubao-Seedance-1.0-pro	字节跳动	**49.07**	**53.93**	**44.21**	API
	kling 2.1 大师版	快手科技	44.84	**48.59**	41.09	API
-	pika 2.2	Pika	40.02	43.40	36.65	网页获取
4	Pangu-T2V	华为	39.11	40.29	37.92	API
-	Sora	OpenAI	37.99	38.33	37.65	网页获取
5	Vidu Q1	生数科技	36.66	35.65	37.67	API
6	Wan2.1-T2V-14B	阿里巴巴	33.03	37.21	28.84	本地部署
6	pixverse-4.5-video	爱诗科技	32.73	35.13	30.33	API
6	wanx2.1-t2v-plus	阿里巴巴	32.52	32.53	32.51	API
-	Mochi	Genmo	29.02	32.31	25.73	本地部署
7	CogVideoX-5b	智谱AI	25.67	25.36	25.98	本地部署
-	Open-Sora-v2	潞晨科技	25.50	27.21	23.78	本地部署
7	Step-Video-T2V	阶跃星辰	25.31	28.62	22.00	本地部署
-	Cosmos-1.0-Diffusion-14B-Text2World	英伟达	24.43	31.56	17.29	本地部署
8	Hunyuan Video	腾讯	22.32	25.54	19.11	本地部署

数据来源：SuperCLUE，2025年7月29日
注：为减少波动影响，本次测评将相差1分内的模型视为并列。海外模型仅作对比参考，不参与排名。

测评详情可访问下方链接：
https://mp.weixin.qq.com/s/YSHnIvknUxL9a9W1Q6y_Aw

1.在综合实力方面，国内头部模型均不输国外模型，呈追平或赶超趋势。

谷歌的veo-3.0-generate-preview以总分55.40位于总榜榜首，国内头部模型与之差距较小，追赶之势明显。总榜的第二名到第四名均为国内模型；后起之秀Pangu-T2V超越Sora，与排名第四的分差较小。

2.国内外模型的视频画质与要素搭建水平有所提升，动态塑造依旧是有待突破的课题。

在通用能力方面，Doubao-Seedance-1.0-pro以53.93分占据该项榜单的第一名；kling 2.1大师版以48.59分夺得第三名，同时在视频画质渲染单项任务中取得了77.22的高分，占得该项第一名。国内外模型在视频画质渲染和要素搭建能力方面的水准有所提升，动态塑造能力依旧是表现较差的环节。

3.国内外模型普遍存在基础能力优于应用能力的发展现状，产品实用性相对不足。

在场景应用能力方面，除少数的国内外头部模型可以做到基础与应用的并行发展外，绝大多数的模型普遍出现基础能力优于应用能力的现象，产品的实用性相对不足。同时，场景应用能力中，头部模型的表现跨越式领先其他模型。

多模态系列基准介绍：SuperCLUE-I2V

SuperCLUE-I2V：中文图生视频测评基准

中文原生图生视频模型测评基准（SuperCLUE-I2V）立足于中文语境，围绕运动流畅性、内容一致性、物理真实性、动漫风格、写实风格和奇幻风格六大任务构建评测体系，旨在为图生视频模型的发展提供客观、公正且具有针对性的评估标准。

评价方法：

严格按照评分细则，综合判断模型对指令的遵循情况，并结合多项预设评价维度，对生成视频的整体效果进行全面评估。每道题目采用5分制，其中1分为极差，2分为较差，3分为一般，4分为良好，5分为优秀。为了更公平地反映模型的实际表现，我们引入**回答率加权机制**，将各任务的原始得分乘以模型在该任务的回答率，以得到该任务的最终得分。具体计分方式如下：

最终得分 = 原始得分 × 回答率

其中：原始得分是模型对已答题目的平均得分；

回答率 = N_answered / N_total

N_total 表示该任务下的总题目数；

N_answered 表示模型成功生成视频的题目数。

测评结果分析

SuperCLUE-I2V
「图生视频」基准测评总榜

排名	模型名称	机构	总分	基础能力			应用能力			使用方式
				运动流畅性	内容一致性	物理真实性	动漫风格	写实风格	奇幻风格	
1	视频 3.0 Pro	字节跳动	**72.22**	**81.85**	**76.67**	**76.30**	**69.26**	**75.19**	54.07	官网
2	可灵 2.1	快手科技	**69.69**	**78.52**	**68.89**	**67.78**	**57.41**	**87.41**	**58.15**	官网
-	Pika 2.2	Pika	**63.46**	65.19	58.15	65.56	55.93	74.81	61.11	官网
3	Pixverse V4.5	爱诗科技	**57.04**	**65.56**	53.70	**61.85**	37.04	70.74	53.33	API
4	清影-AI生视频	智谱AI	**54.26**	53.33	56.30	54.44	33.70	**72.59**	**55.19**	官网
4	Vidu Q1	生数科技	**53.64**	58.52	47.41	54.81	**46.67**	66.30	48.15	API
4	阶跃视频	阶跃星辰	**53.51**	54.76	**60.00**	51.48	41.48	62.22	51.11	官网
5	wanx2.1-i2v-plus	阿里巴巴	**51.67**	51.85	42.59	52.22	36.30	70.37	**56.67**	API
6	I2V-01-Director	MiniMax	**49.20**	53.70	54.81	43.70	39.63	61.48	41.85	官网
7	WHEE	美图	**36.30**	46.30	40.74	47.78	20.37	41.48	21.11	官网
-	Sora	OpenAI	**30.31**	35.56	38.15	32.59	14.44	40.74	20.37	官网

数据来源：SuperCLUE，2025年6月11日。
注：考虑到波动影响，本排行榜将相差1分以内的模型视为并列名次。

测评详情可访问下方链接：

https://mp.weixin.qq.com/s/ulWzEsYuCvallz6aVrDd9Q

1.国内头部模型持续展现领先优势。

在本次测评中，即梦-视频3.0Pro与可灵 2.1分别以72.22、69.69的分数夺得第一名和第二名的位置，且分别与第三名的Pika 2.2产生了8.76、6.23的分差，在综合能力方面展现出领先优势。

2.各模型在物理真实性任务中表现优异，但在场景应用能力中表现不佳。

各模型在运动流畅性及内容一致性方面表现良好，多个模型突破了75分；在物理真实性方面，相比于上一次测评各模型的进步显著。但是，各模型在场景应用能力方面表现平平，尤其是在动漫风格与奇幻风格方面的表现尤为明显。

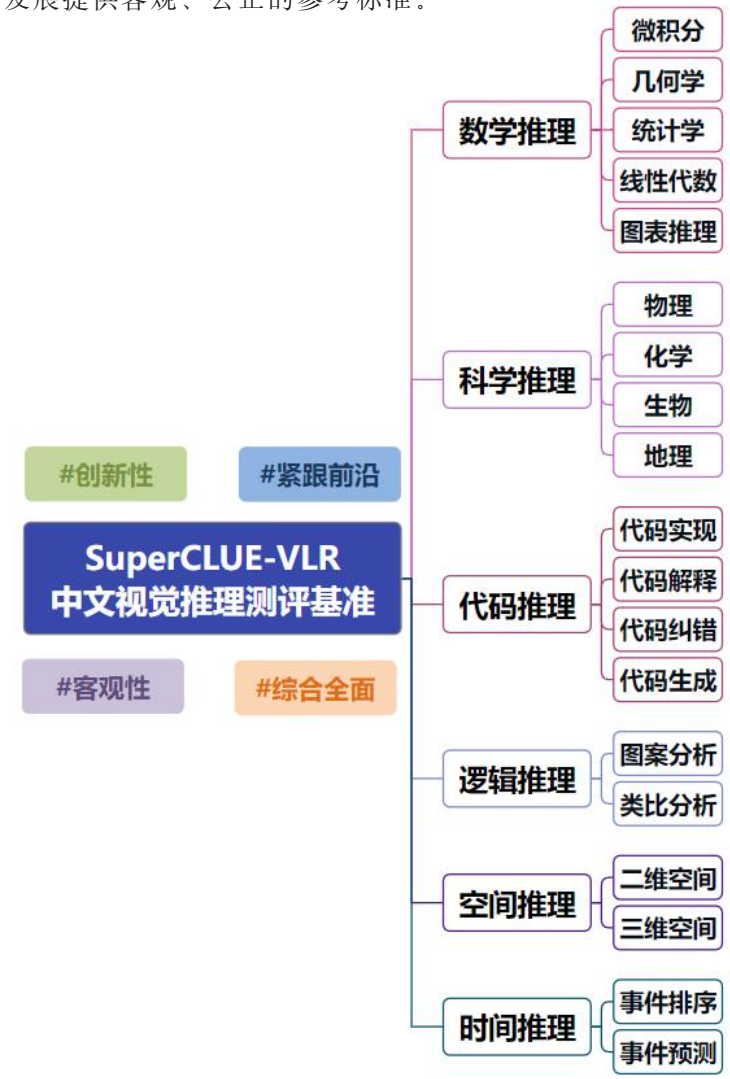
3.视频主体运动迟缓，视频生成失败以及图像分辨率低等情况时有发生且影响得分。

Sora在基础能力和场景应用中均表现不佳，说明模型对于图片的理解和执行能力的不足依旧是模型能力的严重缺失。另外，Vidu Q1所生成视频中的大多数主体部分运动缓慢，与现实场景严重不符，因此总体分数较低；WHEE在约40%的测评任务上未能成功通过图片加载出视频结果，尤其是在动画风格中，仅44.44%的视频成功生成；第三，部分模型生成的视频存在分辨率较低的情况，也会影响最终得分。

推理系列基准介绍：SuperCLUE-VLR

SuperCLUE-VLR：中文视觉推理测评基准

中文视觉推理模型测评基准（SuperCLUE-VLR）聚焦于评估视觉语言模型的推理能力，围绕数学、科学、代码、逻辑、空间、时间六大核心维度构建测评体系，旨在为视觉语言模型推理能力的发展提供客观、公正的参考标准。



测评结果分析

SuperCLUE-VLR
「视觉推理」基准测评总榜

排名	模型名称	机构	总分	数学推理	科学推理	代码推理	逻辑推理	空间推理	时间推理	使用方式
-	Gemini-2.5-Pro-Preview-05-06	Google	**72.12**	**87.20**	**65.63**	**100.00**	**47.22**	**39.81**	**92.86**	API
🥇	Doubao-1.5-Thinking-Vision-Pro	字节跳动	**65.34**	**87.90**	**59.38**	**100.00**	**36.11**	**39.35**	**69.32**	API
🥈	QVQ-Max	阿里巴巴	**56.62**	76.88	43.75	**90.00**	**42.78**	25.28	61.04	API
🥉	Hunyuan-T1-Vision	腾讯	**56.03**	72.02	53.13	**90.00**	30.56	28.15	62.34	API
4	K1.5 长思考	月之暗面	**53.71**	**83.13**	50.00	78.75	30.56	**35.65**	44.16	官网
-	Claude-3.7-Sonnet-Reasoning	Anthropic	**52.19**	78.87	46.88	**90.00**	31.67	21.57	44.16	API
4	Step-R1-V-Mini	阶跃星辰	**50.66**	76.19	40.63	85.00	16.11	21.11	**64.94**	API
-	o3	OpenAI	**47.17**	53.77	**56.25**	88.75	11.11	14.07	59.09	官网
5	InternVL3-78B	上海AI Lab	**39.65**	39.78	28.13	68.75	16.11	28.61	56.49	模型
5	GLM-4V-Plus-0111	智谱AI	**39.43**	46.33	25.00	73.75	5.56	29.44	56.49	API

数据来源：SuperCLUE，2025年5月29日。
注：考虑到波动影响，本排行榜将相差1分以内的模型视为并列名次。

评分方法：

本次测评以回答准确性作为唯一评判标准进行0-1评价。每道题目都配有标准参考答案。

为了确保评估的科学性和公正性，我们采用评价模型Gemini 2.5 Flash Preview 04-17（Thinking模式），将模型的回答与参考答案进行对比，从而判断其正确性。应用这种方式，尽量减少人为因素的干预，确保评分结果的客观性和一致性。

1.模型间的视觉推理能力分化显著，头部与末位模型分差超32分。

模型间视觉推理表现差异巨大，头部模型（如Gemini-2.5-Pro 72.12分）与末位（如InternVL3-78B等不足40分）分差超32分，凸显推理模型在该领域的显著优势。

2.模型在数学和代码任务上表现优异，但在空间和逻辑任务上普遍较弱。

参评模型在数学（如Gemini等超87分）和代码（满分）推理上表现出色，但在空间推理（最高约40分）和逻辑推理（最高约47.2分）方面普遍表现不佳。

3.o3表现未达预期，部分模型推理耗时过长影响效率。

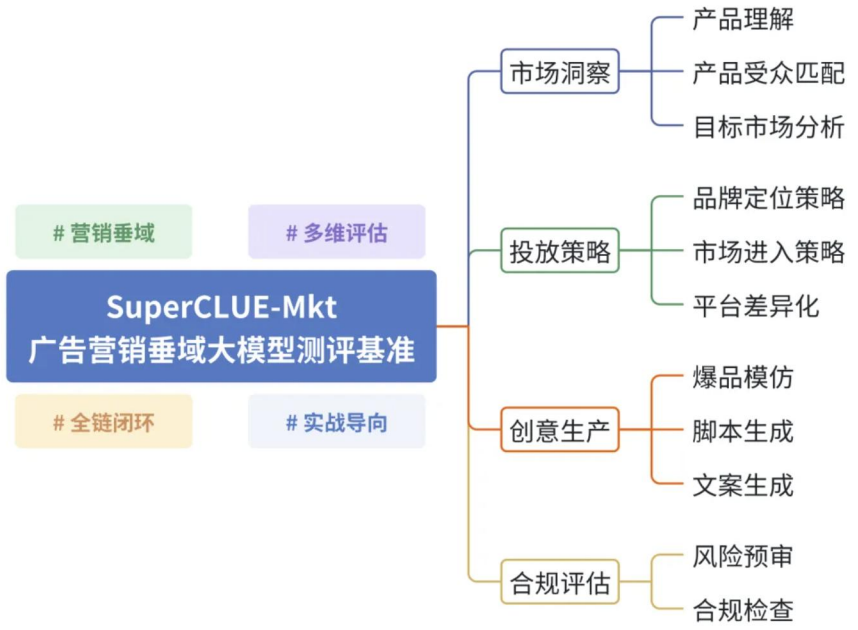
o3在复杂推理任务中表现低于预期，综合评分47.17分，处理复杂问题耗时常超10分钟，未联网模式下仍尝试检索网页，答案完整但准确性不足。

测评详情可访问下方链接：
https://mp.weixin.qq.com/s/21gGvflaZEMxT_QHv8u6ng

文本系列基准介绍：SuperCLUE-Mkt

SuperCLUE-Mkt：广告营销测评基准

SuperCLUE-Mkt是一个聚焦广告营销专业领域的文本生成能力基准，旨在评估大模型在真实营销场景下的专业表现与应用价值。该基准覆盖市场洞察、投放策略、创意生产、合规评估四大核心能力，涵盖了11个具体的细分任务。



测评方案要点

SuperCLUE-Mkt
广告营销专业大模型测评基准——总榜

排名	模型名称	机构	总分	市场洞察	投放策略	创意生产	合规评估	使用方式
1	Tec-Chi-Think-1.0-32B	钛动科技	85.82	89.47	86.97	76.54	90.28	API
1	DeepSeek-v3.2-Thinking	深度求索	85.74	88.12	89.66	72.83	92.34	API
2	Spark-X1.5	科大讯飞	83.42	84.39	85.73	72.72	90.83	API
3	Qwen3-32B(Thinking)	阿里巴巴	82.41	80.92	85.62	75.90	87.22	API
4	GLM-4.6	智谱AI	79.96	80.14	79.35	73.52	86.84	API
-	GPT-5-Nano	OpenAI	78.57	80.41	78.47	66.53	88.88	API
-	Gemma3-27B-it	Google	78.15	78.05	76.17	70.81	87.57	API
-	GPT-OSS-120B	OpenAI	77.17	77.32	75.36	71.71	84.31	API
5	GLM-4-32B-0414	智谱AI	74.35	77.74	75.39	59.40	84.88	API
6	DeepSeek-R1-Distill-Qwen-32B	深度求索	69.83	74.21	66.19	61.66	77.27	API
7	InternLM-3-8B-Instruct	上海AI Lab	64.48	68.79	59.14	55.23	74.76	API
-	Llama-3.1-70B-Instruct	Meta	52.79	54.88	40.59	58.79	56.89	API

数据来源：SuperCLUE，2026年1月23日。
注：考虑到波动影响，本排行榜将相差1分以内的模型视为并列名次。海外模型仅作对比参考，不参与排名。

测评详情可访问下方链接：
https://mp.weixin.qq.com/s/PJM4IZ55g5YIQWPgaNrCbA

摘要1：国内专业模型展现优势，领跑广告营销综合榜单。

本期测评中，钛动科技推出的Tec-Chi-Think-1.0-32B以85.82分的总成绩位居榜首，与深度求索的DeepSeek-v3.2-Thinking（85.74分）共同组成第一梯队。

摘要2：洞察与生产能力分化显著，专业模型更懂业务落地。

在“创意生产”与“市场洞察”分榜中，Tec-Chi-Think-1.0-32B均斩获第一。在创意生成任务上，它以76.54分断层领先Qwen3-32B(Thinking)和GLM-4.6，在具体业务场景中，具备更强的实战落地能力。

摘要3：策略与风控能力分数胶着，专业模型逻辑底座比肩通用顶流。

在“投放策略”与“合规评估”分榜中，DeepSeek-v3.2-Thinking拔得头筹，Tec-Chi-Think-1.0-32B与Spark-X1.5紧随其后。通用模型极具竞争力，专业模型也已具备高水准。

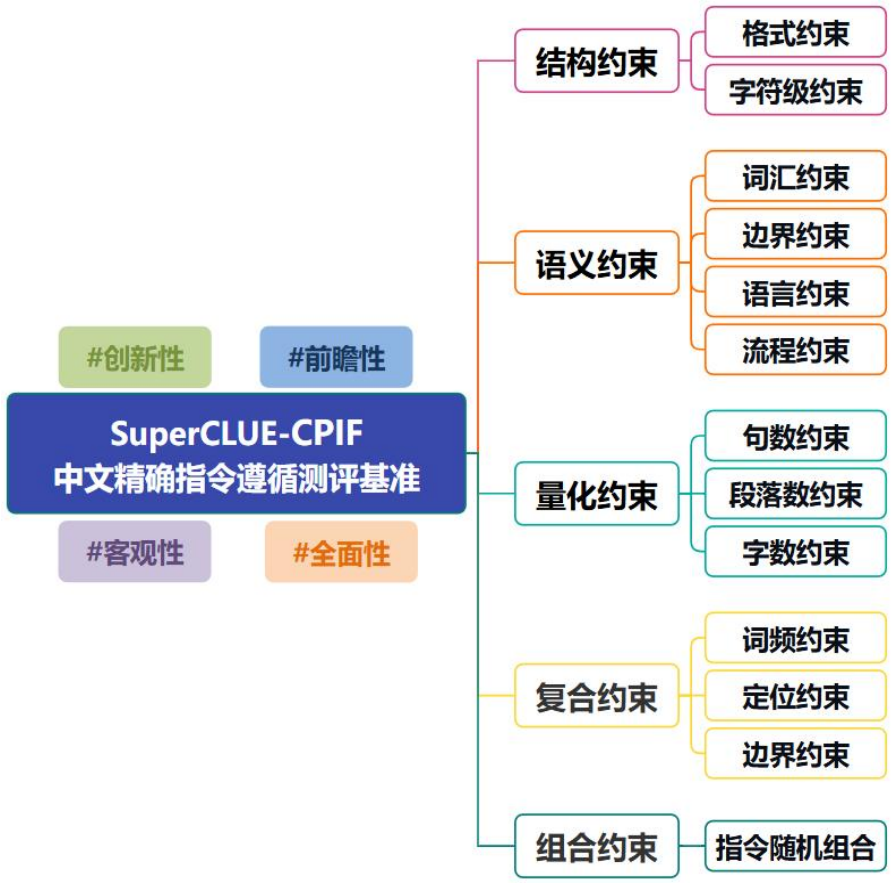
摘要4：国产模型包揽头部梯队，本土营销语境优势显著。

国内模型包揽了总榜前四名，整体表现优于海外模型。表明在处理中文互联网语境下的营销梗及社交平台规则时，国内模型具备理解与生成优势，适应国内企业的营销需求。

文本系列基准介绍：SuperCLUE-CPIF

SuperCLUE-CPIF：中文精确指令遵循测评基准

SuperCLUE-CPIF（Chinese Precise Instruction Following）是一个专为评估大型语言模型在中文环境下精确遵循复杂、多约束指令能力的评测基准。该基准通过构建一个包含多样化场景和多维度约束的高质量数据集，旨在精确度量模型将自然语言指令转化为符合所有要求的具体输出的能力。



测评方案要点

SuperCLUE-CPIF
中文精确指令遵循测评总榜（按任务类型划分）

排名	模型名称	机构	总分	语义约束	结构约束	量化约束	复合约束	组合约束	测评日期
-	GPT-5.1(high)	OpenAI	83.59	64.44	97.83	100.00	93.33	63.64	2025.11.19
-	Gemini-3-Pro-Preview	Google	83.08	60.00	100.00	92.31	93.33	69.70	2025.11.19
-	GPT-5(high)	OpenAI	82.56	60.00	89.13	100.00	95.56	72.73	2025.10.21
1	360zhinao3-o1.5	360	78.97	86.67	58.70	96.15	84.44	75.76	2025.11.19
2	ERNIE-X1.1	百度	75.90	97.78	100.00	80.77	37.78	60.61	2025.10.21
3	DeepSeek-V3.2-Exp-Thinking	深度求索	74.36	73.33	93.48	84.62	57.78	63.64	2025.10.21
-	Claude-Sonnet-4.5-Reasoning	Anthropic	72.82	60.00	80.43	92.31	68.89	69.70	2025.10.21
4	Kimi-K2-Thinking	月之暗面	67.18	57.78	67.39	80.77	82.22	48.48	2025.11.19
5	Hunyuan-T1-20250822	腾讯	66.15	53.33	93.48	88.46	42.22	60.61	2025.10.21
-	Gemini-2.5-Pro	Google	63.59	60.00	89.13	65.38	48.89	51.52	2025.10.21
6	GLM-4.6	智谱AI	60.82	55.56	93.48	80.77	37.78	37.50	2025.10.21
7	MiniMax-M2	稀宇科技	54.36	55.56	73.91	61.54	44.44	33.33	2025.11.19
-	Grok-4	X.AI	53.85	60.00	95.65	53.85	22.22	30.30	2025.10.21
8	Qwen3-Max	阿里巴巴	51.79	60.00	84.78	57.69	20.00	33.33	2025.10.21
9	Doubao-Seed-1.6-thinking-250715	字节跳动	46.67	53.33	78.26	42.31	31.11	18.18	2025.10.21

数据来源：SuperCLUE，2025年11月19日。
注：本榜单将相差一分以内的模型视为并列名次；海外模型仅作对比参考，不参与排名。

1.GPT-5.1(high)、Gemini-3-Pro-Preview、GPT-5(high)强势领跑。

GPT-5.1(high)以83.59分领跑全球，Gemini-3-Pro-Preview以83.08分位居第二，GPT-5(high)以82.56分位于第三。

2.国内模型表现亮眼，领先海外顶尖模型Claude-Sonnet-4.5-Reasoning。

360zhinao3-o1.5、ERNIE-X1.1和DeepSeek-V3.2-Exp-Thinking分别以78.97分、75.90分和74.36分位居国内前三，领先海外顶尖模型Claude-Sonnet-4.5-Reasoning。

3.海外模型在中文精确指令遵循任务上要领先于国内模型。

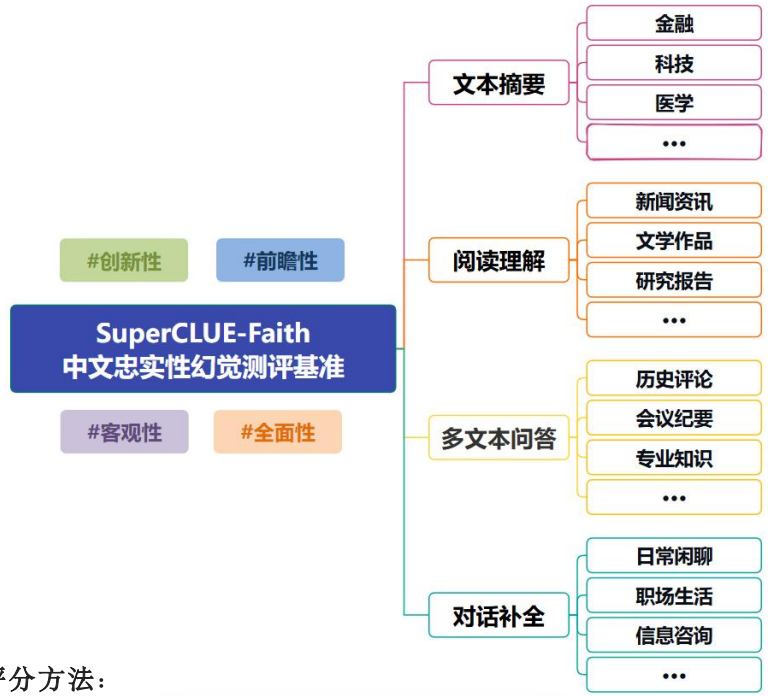
海外模型的平均分为73.25分，国内模型的平均分为63.99分，相差9.26分，海外模型在该任务上的整体表现要优于国内模型。

测评详情可访问下方链接：
https://mp.weixin.qq.com/s/pcBrykyK99dfWA5WSqj1jw

文本系列基准介绍：SuperCLUE-Faith

SuperCLUE-Faith：中文忠实性幻觉测评基准

SuperCLUE-Faith 是一个专注于评估大语言模型在中文领域忠实性幻觉表现的基准测试，该基准涵盖四大核心任务：文本摘要、阅读理解、多文本问答以及对话补全，通过多维度评测，为大语言模型的忠实性幻觉研究提供全面、客观的能力评估依据。



评分方法：

本次SuperCLUE-Faith中文忠实性幻觉测评采用大模型三阶段自动化评估方法，以下是具体评估流程介绍：

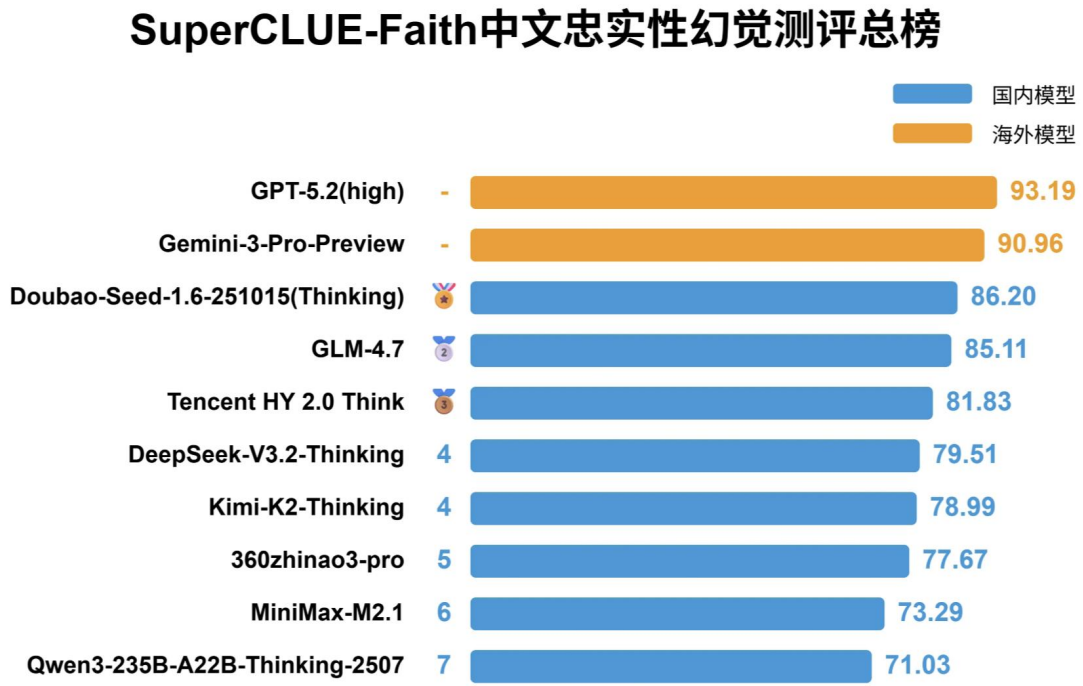
a）语句分割阶段：以中文标点符号为边界，对模型输出答案进行分句处理；

b）幻觉判定阶段：基于任务特异性评价标准，对每个分句进行二元判定：无幻觉得1分；存在幻觉得0分；

c）分数聚合阶段：单题得分（范围0-1分） = 无幻觉句子数量 / 总句子数量；总分 = 单题得分之和 / 总题数。

该评估机制通过标准化计分流程，确保结果的公平性和客观性。

测评结果分析



数据来源：SuperCLUE，2025年12月31日。

注：由于部分模型分数较为接近，为了减少问题波动对排名的影响，本次测评将相距1分区间的模型定义为并列。海外模型仅对比参考，不参与排名。

1. 忠实性幻觉是指模型的生成内容与用户指令或其他输入的分歧，以及模型生成内容内部的不一致性。

2. GPT-5.2(high)以 93.19 分领跑榜单，Gemini-3-Pro-Preview以 90.96 分紧随其后，两大顶尖海外模型在幻觉控制上表现突出。

3. Doubao-Seed-1.6-251015(Thinking)以 86.20 分位居国内第一，GLM-4.7和Tencent HY 2.0 Think分别以 85.11 分和 81.83 分位于国内第二和第三，表现不俗。

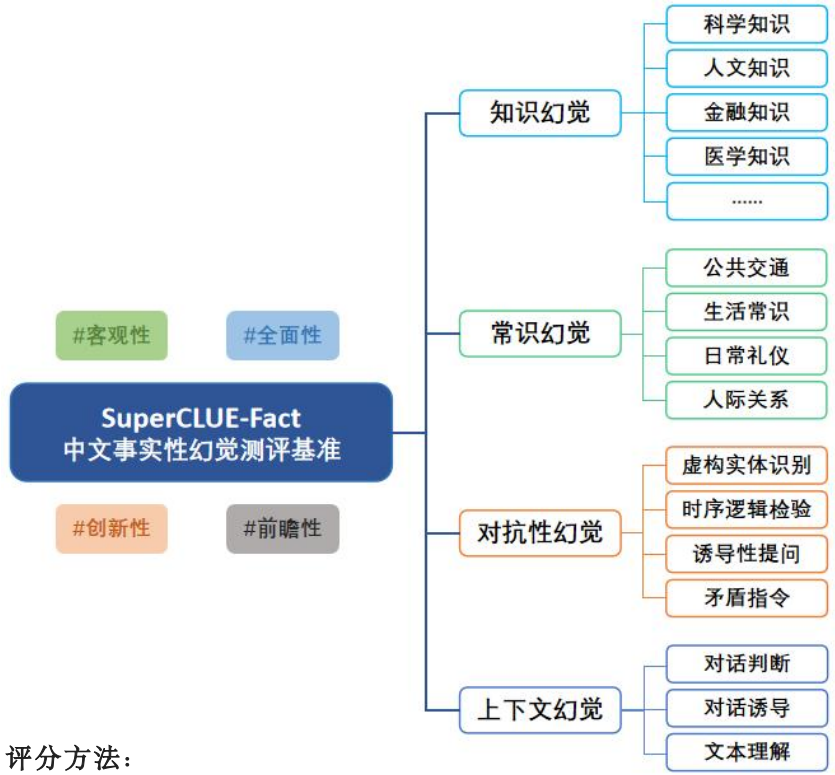
测评详情可访问下方链接：

https://mp.weixin.qq.com/s/w6DYfxNwMnWaInHC3mKv7A

文本系列基准介绍：SuperCLUE-Fact

SuperCLUE-Fact：中文事实性幻觉测评基准

SuperCLUE-Fact 是一个专注于评估大语言模型在中文领域事实性幻觉（Factual Hallucination）表现的基准测试。该基准涵盖四大核心任务：知识幻觉、常识幻觉、对抗性幻觉和上下文幻觉，重点考察模型在中文简短事实问答中的准确性，以及识别与判断事实性幻觉的能力。通过多维度评测，SuperCLUE-Fact 旨在为大语言模型的事实性幻觉研究提供全面、客观的能力评估依据。



评分方法：

为了确保评估的科学性和公正性，我们采用超级大模型进行评价。结合评估流程、评估标准、评分规则，进行细粒度评估，采用0/1评分标准，对于存在事实幻觉（答案错误）的题目评分为0，对于不存在事实幻觉（答案正确）的题目评分为1。应用这种方式，尽量减少人为因素的干预，确保评分结果的客观性和一致性。

测评结果分析

SuperCLUE-Fact
中文事实性幻觉测评总榜

排名	模型名称	机构	准确率（%）	知识幻觉	常识幻觉	对抗性幻觉	上下文幻觉
1	DeepSeek-R1	深度求索	86.02	89.66	72.97	87.27	85.92
-	GPT-4.5-Preview	OpenAI	85.30	89.66	70.27	85.45	85.92
-	gemini-2.5-pro-exp-03-25	Google	84.95	90.52	81.08	89.09	74.65
-	Claude 3.7 Sonnet(Extended)	Anthropic	84.23	87.07	78.38	90.91	77.46
-	ChatGPT-4o-latest	OpenAI	83.15	87.93	64.86	78.18	88.73
3	DeepSeek-V3-0324	深度求索	82.80	93.10	67.57	85.45	71.83
3	doubao-1.5-pro-32k	字节跳动	81.72	86.21	83.78	89.09	67.61
4	qwen-max-latest	阿里巴巴	79.93	88.79	75.68	80.00	67.61
5	QwQ-32B	阿里巴巴	78.85	82.76	70.27	81.82	74.65
-	o3-mini(high)	OpenAI	78.78	86.21	62.16	79.63	74.65
-	gemini-2.0-flash	Google	78.49	88.79	78.38	78.18	61.97
6	ernie-4.5-8k-preview	百度	77.78	86.21	86.49	81.82	56.34

数据来源：SuperCLUE，2025年4月14日。
注：测试模型均为非联网版；本榜单将相差一分以内的模型视为并列名次；海外模型仅作对比，不参与排名

测评详情可访问下方链接：
https://mp.weixin.qq.com/s/0zq00-XBoUSwvmuQnNOUUw

1.DeepSeek-R1当前领先，但头部模型差距微小。

DeepSeek-R1以86.02的总分领跑事实性幻觉榜单，GPT-4.5-Preview、gemini-2.5-pro-exp-03-25、Claude 3.7 Sonnet (Extended) 和 ChatGPT-4o-latest 也表现优异，位列前五。整个榜单的分数相对集中，尤其是在顶部梯队，显示出领先模型在事实性幻觉能力上的激烈竞争。

2.模型在不同类型的任务上表现差异显著。

本次测评的12个模型在处理知识幻觉和对抗性幻觉方面表现相对稳健，平均得分有85分左右。然而，在常识幻觉和上下文幻觉这两类任务上，模型表现普遍较弱，平均分不足75分，差距明显。

3.海外与国内模型各有优劣，海外模型总体稍领先。

整体评分上，海外模型（平均82.48分）比国内模型（平均81.18分）高出1.3分。但具体任务表现呈现差异：国内模型善于处理常识幻觉（领先3.61分）和对抗性幻觉；而海外模型则在知识幻觉和上下文幻觉方面更具优势，特别是在上下文幻觉任务上，领先国内模型6.57分，差距最为明显。

性能系列基准介绍：DeepSeek-R1系列

第三方平台DeepSeek-R1联网搜索能力测评

为了解各第三方平台接入DeepSeek-R1的联网搜索能力，本次我们针对10家接入DeepSeek-R1的第三方平台进行了联网搜索的测评，测评内容包括基础检索能力如文化生活、经济生活、实时新闻等，以及分析推理能力如推理计算、分析排序、数据检索与分析等。

测评集构建：

1. 首先从各个权威官方网站搜集不同的新闻和数据作为原始题目来源；
2. 多方核查每条原始信息的正确性，剔除错误的、存在争议的信息，然后筛选出正确的、符合客观事实的信息；
3. 使用这些原始信息按照预先规定的维度构建题目；
4. 最后对所有题目进行复查，确定最终的测评集。

答案获取：

1. 所有第三方平台皆由人工获取网页端的答案，获取答案的时间均为工作日；
2. 题目耗时皆由人工计时并记录；
3. 每道题目在获取答案前均会清除上下文信息，避免对测评产生影响。

评分方法：

1. 本次测评集的题目均为客观题，仅有唯一解，因此本次测评采取0/1的评分模式，即模型的答案与参考答案一致则记1分，模型的答案与参考答案不一致则记0分；
2. 我们将模型答案出现截断或无回复情况的题目视为未满足用户需求，该题记0分；
3. 最后的总分计算公式为：记1分的题目总数除以总题数。

测评结果分析

第三方平台DeepSeek-R1联网搜索测评总榜（网页版）

排名	名称	机构	总分	基础检索能力	分析推理能力	平均耗时
🥇	腾讯元宝	腾讯	80.61	100.00	55.81	39.69
🥈	阶跃AI	阶跃星辰	74.49	100.00	41.86	41.10
🥉	支付宝百宝箱	蚂蚁集团	73.47	96.36	44.19	45.27
4	百度AI搜索	百度	70.41	100.00	32.56	41.57
4	天工AI（高级模式）	昆仑万维	70.41	87.27	48.84	54.69
5	飞书知识问答	字节跳动	65.31	92.73	30.23	35.80
5	秘塔AI搜索（深入模式）	秘塔科技	65.31	96.36	25.58	58.58
5	纳米AI搜索	360	65.31	96.36	25.58	36.04
6	字节火山引擎	字节跳动	64.29	94.55	25.58	17.12
7	MiniMax	MiniMax	61.22	90.91	23.26	73.51

数据来源：SuperCLUE，2025年3月11日。
注：1.考虑到波动影响，本榜单将相差一分以内的第三方平台视为并列名次；
2.本榜单将截断和无回复的题目视为错误，并计入总分；
3.平均耗时表示从发送题目到回答结束所用时间，此处表示总平均耗时，单位为秒/题。

测评详情可访问下方链接：
https://mp.weixin.qq.com/s/s_ZjP3tjxkyTVEPK_GucZg

1.各平台整体表现差异较大，腾讯元宝综合实力领先。

总分相差最大的两个平台分差接近20分，联网搜索表现存在一定的差距。腾讯元宝是本次测评中唯一一个超过80分的第三方平台，以80.61分领跑联网搜索测评榜单，展现出不俗的实力。

2.基础检索能力普遍优秀，分析推理能力是不同平台之间的关键差异点。

各平台的基础检索能力平均分达到了95.45分，而分析推理能力仅有35.35分，相差近60分。在基础检索能力维度上，腾讯元宝、阶跃AI和百度AI搜索达到了100%的准确率，表现优异；但在分析推理能力维度上，仅有腾讯元宝、天工AI、支付宝百宝箱和阶跃AI超过了40分。

3.各平台的回复率普遍较高，稳定性较强。

飞书知识问答、阶跃AI、腾讯元宝和支付宝百宝箱在联网搜索回复率方面十分优秀，完整回复率均为100%，位居第一梯队。秘塔AI搜索、纳米AI搜索和天工AI紧随其后，构成第二梯队；其他平台也均有超过85%的完整回复率表现。

性能系列基准介绍：DeepSeek-R1系列

第三方平台DeepSeek-R1 API调用稳定性测评

为了给用户提供更全面、客观的参考，并帮助他们选择合适的服务平台，我们在7个服务平台上进行了DeepSeek-R1的API稳定性测评，从回复率、准确率和推理耗时等方面评估其表现。

本次测评在同一机器上对第三方平台发送请求，使用20道小学奥数推理题测试，temperature为0.6，max_token设为平台最大值或16000，采用流式输出方式记录耗时及输出token数量。每题尝试三次避免网络影响，三次失败视为获取失败。本次测评的报告仅代表测评时点的稳定性。

测评方法：

本次测评在同一机器上对第三方平台发送请求，使用20道小学奥数题测试，temperature为0.6，max_token设为平台最大值或16000，采用流式输出方式记录耗时及输出token数量。每题尝试三次避免网络影响，三次失败视为获取失败。

具体实现说明：

1.对于每个第三方平台，使用20道小学奥数题进行统一测试。为了避免网络波动造成的影响，每个模型对每个问题会尝试三次，如果三次尝试都未获取到答案，才视为获取失败。并且将测试时间设为下午开始，主要模型完成时间在工作日下午15:30-20:30之间。

2.由于测评集为推理题，输出较长，对于max_token的设置遵循以下原则：如果平台文档说明了支持的最大输出的token，我们将max_token按照平台的最大输出token进行设置；如果平台未说明，max_token统一设置为16000。对于影响生成质量的参数配置，调用时对于允许配置temperature的第三方api，我们统一采取DeepSeek的推荐参数值：0.6，其他参数保持各第三方平台默认不做配置。

3.关于推理耗时的统计方法，API的调用统一采用流式输出（调用时，将stream参数设置为True）。开始发送请求时间记录为start_time，请求开始返回数据时，记录时间chunk_time1；返回数据结束后，记录时间chunk_time2。每道题目的输出token数量记录为：completion_tokens。

测评结果分析

DeepSeek-R1第三方平台稳定性测评总榜（API版）

排名	第三方平台	机构	完整回复率	截断率	无回复率	准确率	每秒输出token数量
1	字节火山引擎	字节跳动	**100%**	0%	0%	**95.00%**	**27.94**
1	商汤大装置	商汤科技	**100%**	0%	0%	90.00%	**20.63**
1	阿里云百炼	阿里巴巴	**100%**	0%	0%	70.00%	6.90
2	硅基流动	硅基流动	95%	5%	0%	**94.74%**	11.76
2	together.ai	together.ai	95%	5%	0%	89.47%	**55.86**
2	腾讯云知识引擎	腾讯科技	95%	5%	0%	84.21%	10.97
3	微软云	Microsoft	75%	0%	25%	**93.33%**	6.90

数据来源：SuperCLUE，2025年2月20日；
注：排名代表在本次奥数推理题上的完整回复率的高低。

测评详情可访问下方链接：
https://mp.weixin.qq.com/s/mQna2pcTeE1fnDGNLkrz6A

1.各个第三方平台使用DeepSeek-R1的完整回复率表现差异不大。

除微软云的DeepSeek-R1 API外，其他的完整回复率都在95%以上。火山引擎、商汤大装置、阿里云百炼都实现了100%的完整回复率。

2.各第三方API接口输出效率差距明显，平均每秒输出token数量最低6.9个，最高55.86个。

测评显示，第三方API每秒输出token数量差异大。Together.ai以每秒55.86个token遥遥领先，文本生成效率极高；字节火山引擎次之，每秒27.94个token；阿里云百炼和微软云API则仅为每秒6.90个token。高并发或快速响应应用，宜选高生成效率平台。

3.各个第三方平台准确率上略有差异。

准确率上，字节火山引擎、硅基流动，准确率为95%左右；商汤大装置准确率在90%；阿里云百炼准确率为70%。

附录一：SuperCLUE介绍

中文通用大模型评测基准——SuperCLUE是大模型时代背景下CLUE(The Chinese Language Understanding Evaluation)基准的发展和延续，是独立、领先的通用大模型的综合性测评基准。中文语言理解测评基准CLUE**发起于2019年，**陆续推出过CLUE、FewCLUE、ZeroCLUE等广为引用的测评基准。

2019 — CLUE基准发布

2020 — 发布多篇顶会论文，并承办了NLPCC 2020开放测评任务

2021 — FewCLUE、DataCLUE、ZeroCLUE等测评基准发布

2022 — 联合西湖大学发布DataCentricAI；发布SimCLUE大规模中文语义理解数据集

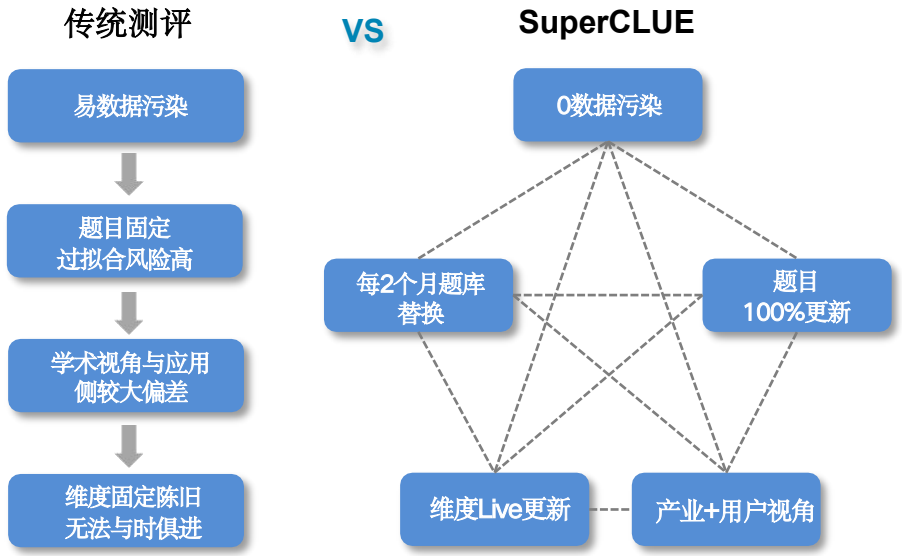
2023 — SuperCLUE正式发布；发布首个中文多轮开放式测评基准SuperCLUE-Open；发布行业测评基准

2024 — 发布数学、代码、长文本、RAG等测评基准；发布文生视频、文生图、视觉理解等多模态测评基准；发布链式推理等推理测评基准；发布多轮可执行智能体测评基准AgentCLUE

2025.01-2025.06 — 发布中文通用AI智能体测评基准AgentCLUE-General系列；发布文/图生视频等多模态系列测评基准；发布幻觉控制、深度研究等测评基准；

2025.07-2025.12 — 发布精确指令遵循、深度搜索、软件工程(SWE)、具身智能、前端开发等测评基准；发布手机GUI Agent、CUA等测评基准；推出全新交互式评测模式：SuperCLUE中文竞技场

SuperCLUE与传统测评的区别



SuperCLUE 三大特征

01 “Live”更新，0数据污染

测评题库每2个月100%替换，杜绝过拟合风险。体系维度根据大模型进展Live更新。

02 测评方式与用户交互一致

测评方法与用户交互方式保持一致，测评任务贴近真实落地场景，高度还原用户视角。

03 独立第三方，无自家模型

完全独立的第三方评测机构，不研发自家模型。承诺提供无偏倚的客观、中立评测结果。

附录二：榜单全景图

基于大模型技术和应用发展趋势以及基准测评专业经验，SuperCLUE构建出多领域、多层次的大模型综合性测评基准框架。从基础到应用覆盖：通用基准体系、文本系列基准、多模态系列基准、推理系列基准、Agent系列基准、AI应用系列基准、性能系列基准。为产业、学术和研究机构的大模型研发提供重要参考。

SuperCLUE大模型综合测评基准榜单全景图

通用基准体系
- 数学推理
- 科学推理
- 代码生成
- 智能体(任务规划)
- 精确指令遵循
- 幻觉控制

文本系列基准
- 中文精确指令遵循 SuperCLUE-CPIF
- 事实性幻觉 SuperCLUE-Fact
- 忠实性幻觉 SuperCLUE-Faith
- 长文本 SuperCLUE-Long
- 超长文本 SuperCLUE-200K
- 角色扮演 SuperCLUE-Role
- 检索增强 SuperCLUE-RAG
- 多轮对抗安全 SuperCLUE-Safety

多模态系列基准
视频生成
- 世界模型 SuperCLUE-World
- 文生视频 SuperCLUE-T2V
- 图生视频 SuperCLUE-I2V

图像/视觉
- 图像编辑 SuperCLUE-Edit
- 实时音视频交互 SuperCLUE-Live
- 文生图 SuperCLUE-Image
- 多模态视觉语言 SuperCLUE-VLM

语音
- 实时语音交互 SuperCLUE-Voice
- 语音合成 SuperCLUE-TTS
- 声音复刻 SuperCLUE-Cloning

推理系列基准
- 前端开发 SuperCLUE-WebCoding
- 软件工程 SuperCLUE-SWE
- 视觉推理 SuperCLUE-VLR
- 科学推理 SuperCLUE-Science
- 全国高中数学竞赛 MathCLUE
- 项目级代码 SuperCLUE-Project
- 链式推理 SuperCLUE-COT
- 小学奥数 SuperCLUE-Math6o
- 代码 SuperCLUE-Code3
- 数学多步推理 SuperCLUE-Math6
- 研究生级别数学

Agent系列基准
通用
- Agent版软件工程 AgentCLUE-SWE
- 具身智能 EmbodiedCLUE-VLA
- 深度搜索 SuperCLUE-DeepSearch
- 深度研究 SuperCLUE-DeepResearch
- 中文通用AI智能体 AgentCLUE-General AgentCLUE-tGeneral
- 可执行智能体 AgentCLUE
- 不可执行智能体 SuperCLUE-Agent

终端
- Computer Use Agent(离线) AgentCLUE-Mobile
- 手机 GUI-Agent（离线） AgentCLUE-Mobile

行业
- 汽车座舱智能体 AgentCLUE-ICabin

AI应用系列基准
AI产品
- AI编程工具 AgentCLUE-AICodeKit
- AI搜索 SuperCLUE-AISearch
- 代码助手 SuperCLUE-Coder

行业
- 广告营销 SuperCLUE—Mkt
- 汽车知识 SuperCLUE—AutoQA
- 金融 SuperCLUE-Fin
- 工业 SuperCLUE-Industry
- 汽车 SuperCLUE-Auto
- 智能座舱 SuperCLUE-Icabin

性能系列基准
DeepSeek-R1系列
- DeepSeek-R1 第三方联网搜索能力测试（网页端）
- DeepSeek-R1 第三方稳定性测试（App端）
- DeepSeek-R1 第三方稳定性测试（API端）
- DeepSeek-R1 第三方稳定性测试（网页端）

- 大模型推理速度测评
- 国产芯片基准测评

其余系列
- 开源数据集 Math24o

已发布　即将发布

注：通用基准介绍可在报告中查看，其余基准可点击对应链接跳转至最新的发布文章。

附录三：2025年年度测评模型列表

SuperCLUE-2025年年度测评选取了国内外有代表性的**23个大模型**。

序号	模型	机构	简介	序号	模型	机构	简介
1	Qwen3-Max-2025-09-23	阿里巴巴	官方发布的基础模型，使用阿里云公开的API：qwen3-max-2025-09-23。	13	Mistral Large 3	Mistral AI	官方发布的最新开源模型，使用官方API：mistral-large-2512。
2	Qwen3-Max-Preview-Thinking	阿里巴巴	官方发布的推理模型，使用阿里云公开的API：qwen3-max-preview-thinking。	14	Grok-4	X.AI	官方发布的推理模型，使用官方API：grok-4-0709。
3	DeepSeek-V3.2-Thinking	深度求索	官方发布的最新开源推理模型，使用官方API：deepseek-reasoner。	15	Grok-4.1-Fast(Reasoning)	X.AI	官方发布的推理模型，使用官方API：grok-4-1-fast-reasoning。
4	GLM-4.7	智谱AI	官方发布的最新开源推理模型，使用官方API：glm-4.7。	16	Llama-4-Maverick-17B-128E-Instruct	Meta	使用together.ai的接口：meta-llama/Llama-4-Maverick-17B-128E-Instruct-FP8。
5	Spark X1.5	科大讯飞	官方发布的最新推理模型，使用官方API：Spark X1.5。	17	LongCat-Flash-Thinking-2601	美团	官方发布的开源推理模型，使用官方API：LongCat-Flash-Thinking-2601。
6	Kimi-K2-Thinking	月之暗面	官方发布的开源推理模型，使用官方API：kimi-k2-thinking。	18	Doubao-Seed-1.8-251228(Thinking)	字节跳动	官方发布的最新模型，使用官方API：doubao-seed-1-8-251228。
7	MiniMax-M2.1	稀宇科技	官方发布的最新开源推理模型，使用官方API：MiniMax-M2.1。	19	MiMo-V2-Flash-0112	小米集团	官方开源的最新推理模型，使用官方API：mimo-v2-flash。
8	Claude-Opus-4.5-Reasoning	Anthropic	官方发布的混合推理模型，使用官方API：Claude-Opus-4.5-Reasoning。	20	ERNIE-5.0	百度	官方发布的最新全模态模型，使用官方API：ERNIE-5.0-Thinking-Preview。
9	Gemini-3-Flash-Preview	Google	Google发布的最新模型，使用官方API：gemini-3-flash-preview。	21	Qwen3-Max-Thinking	阿里巴巴	官方发布的最新推理模型，使用官方API：qwen3-max-2026-01-23 enable_thinking=True。
10	Gemini-3-Pro-Preview	Google	Google发布的最新模型，使用官方API：gemini-3-pro-preview。	22	Kimi-K2.5-Thinking	月之暗面	官方发布的最新开源多模态模型，使用官方API：kimi-k2.5。
11	GPT-5.2(high)	OpenAI	官方发布的最新推理模型，使用官方API：gpt-5.2-2025-12-11。	23	Tencent HY 2.0 Think	腾讯	官方发布的最新推理模型，使用官方API：hunyuan-2.0-thinking-20251109。
12	gpt-oss-120b(high)	OpenAI	官方发布的开源推理模型，使用OpenRouter的API：gpt-oss-120b。		/	/	/

SuperCLUE
中文大模型综合性测评基准
独立第三方AGI测评机构

“

为AI应用及研发团队提供专业测评服务和独立分析，助力技术选型和性能优化

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

”

联系我们

——为AI应用及研发团队提供专业测评服务和独立分析，助力技术选型和性能优化

——立足业内领先的第三方大模型测评机构，致力于为业界提供专业测评服务：

通用大模型测评

提供大模型综合性评测服务，输出全方位的评测报告，包括但不限于多维度测评结果、横向对比、典型示例、模型优化建议。

行业与专项大模型测评

聚焦测评大模型在行业落地应用效果，包括但不限于汽车、手机、金融、工业、教育、医疗等行业大模型应用能力，中文Agent能力测评、大模型安全评估、多模态能力测评、个性化角色扮演能力测评。

多模态大模型测评

多维度全方位测评多模态大模型的基础能力与应用能力，包括但不限于实时多模态交互、视频生成基准测评、文生图测评、多模态理解测评、图像编辑、语音合成、世界模型等。

Agent智能体测评

提供AI大模型落地应用及工具测评，包括但不限于AgentCLUE、AgentCLUE-General等通用Agent，代码助手、AI搜索等应用；AI PC、AI手机、XR设备及具身智能等设备端应用。

大模型深度研究报告

提供国内外大模型深度研究报告，全面调研与分析国内外大模型技术进展及应用落地情况，为企事业单位提供及时、深度的第三方专业报告。

业务合作：请简要描述需求至合作邮箱 contact@superclue.ai

SuperCLUE

交流
合作

扫码
关注

- 排行榜官方地址：https://www.superclueai.com
- 官网：www.CLUEbenchmarks.com
- Github地址：https://github.com/CLUEbenchmark
- 联系人：徐老师 18806712650（微信同号）　朱老师 18621237819（微信同号）

法律声明

- **版权声明**

本报告为SuperCLUE团队制作，其版权归属SuperCLUE，任何机构和个人引用或转载本报告时需注明来源为SuperCLUE，且不得对本报告进行任何有悖原意的引用、删节和修改。任何未注明出处的引用、转载和其他相关商业行为都将违反《中华人民共和国著作权法》和其他法律法规以及有关国际公约的规定。对任何有悖原意的曲解、恶意解读、删节和修改等行为所造成的一切后果，SuperCLUE不承担任何法律责任，并保留追究相关责任的权力。

- **免责条款**

本报告基于中文大模型基准测评（SuperCLUE）2025年年度的自动化测评结果以及已公开的信息编制，力求结果的真实性和客观性。然而，所有数据和分析均基于报告出具当日的情况，对未来信息的持续适用性或变更不承担保证。本报告所载的意见、评估及预测仅为出具日的观点和判断，且在未来无需通知即可随时更改。可能根据不同假设、研究方法、即时动态信息和市场表现，发布与本报告不同的意见、观点及预测，无义务向所有接受者进行更新。

本团队力求报告内容客观、公正，但本报告所载观点、结论和建议仅供参考使用，不作为投资建议。对依据或者使用本报告及本公司其他相关研究报告所造成的一切后果，本公司及作者不承担任何法律责任。