SuperCLUE
中文大模型综合性测评基准

中文大模型基准测评2025年9月报告

—— 2025年中文大模型阶段性进展9月评估

SuperCLUE团队

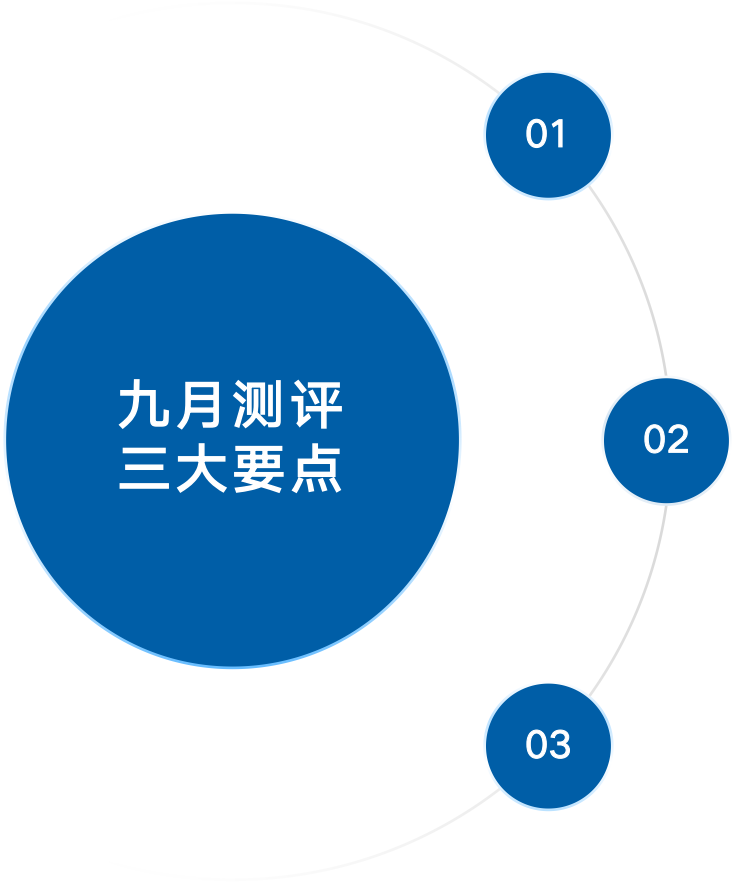
2025.10.16

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.

报告摘要



1. 海外模型保持头部优势，国内模型继续追赶。

在本次9月通用测评中，海外模型占据了榜单前6，其中GPT-5(high)以69.37分遥遥领先，o4-mini(high)（65.91分）、Claude-Sonnet-4.5-Reasoning（65.62分）、Claude-Opus-4.1-Reasoning（64.87分）、Gemini-2.5-Pro（64.68分）等紧随其后。国内的DeepSeek-V3.2-Exp-Thinking、Doubao-Seed-1.6-thinking-250715分别以62.62分和60.96分并列国内第一。

2. 国内开源模型优势显著。

国内的DeepSeek-V3.2-Exp-Thinking（62.62分）、openPangu-Ultra-MoE-718B（58.87分）和Qwen3-235B-A22B-Thinking-2507（57.73分）分别位于开源模型榜单前三，大幅度领先海外开源最好模型gpt-oss-120b（53.05分）。

3. 国内模型更具性价比，海外模型推理效率更高。

国内模型的API价格大多数处于0-10元/百万Tokens，平均API价格为3.88元/百万Tokens，而海外模型的API价格比较分散，从2-200元/百万Tokens不等，海外模型平均API价格为20.46元/百万Tokens，是国内模型API价格的5倍以上。

国内推理模型平均每题的推理耗时为101.07秒，而海外推理模型仅有41.60秒，海外推理模型的推理效率远高于国内推理模型。

SuperCLUE九月中文大模型基准测评简介

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9月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1260道简答题，测评集的介绍如下：

SuperCLUE-9月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。本次评估的人类一致性约98%。

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。本次评估的人类一致性约98%。

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），本次评估的人类一致性约99%。

4.智能体Agent

介绍：主要考察在中文场景下基于可执行的环境，LLM作为执行代理在对话中调用工具完成任务的能力。包括单轮对话和多轮对话。涉及的中文场景包括但不限于汽车控制、股票交易、智能家居、旅行规划等10余个场景。

评价方式：结合任务完成与否、系统状态比对的0/1评估，本次评估的人类一致性约99%。

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，本次评估的人类一致性约99%。

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。

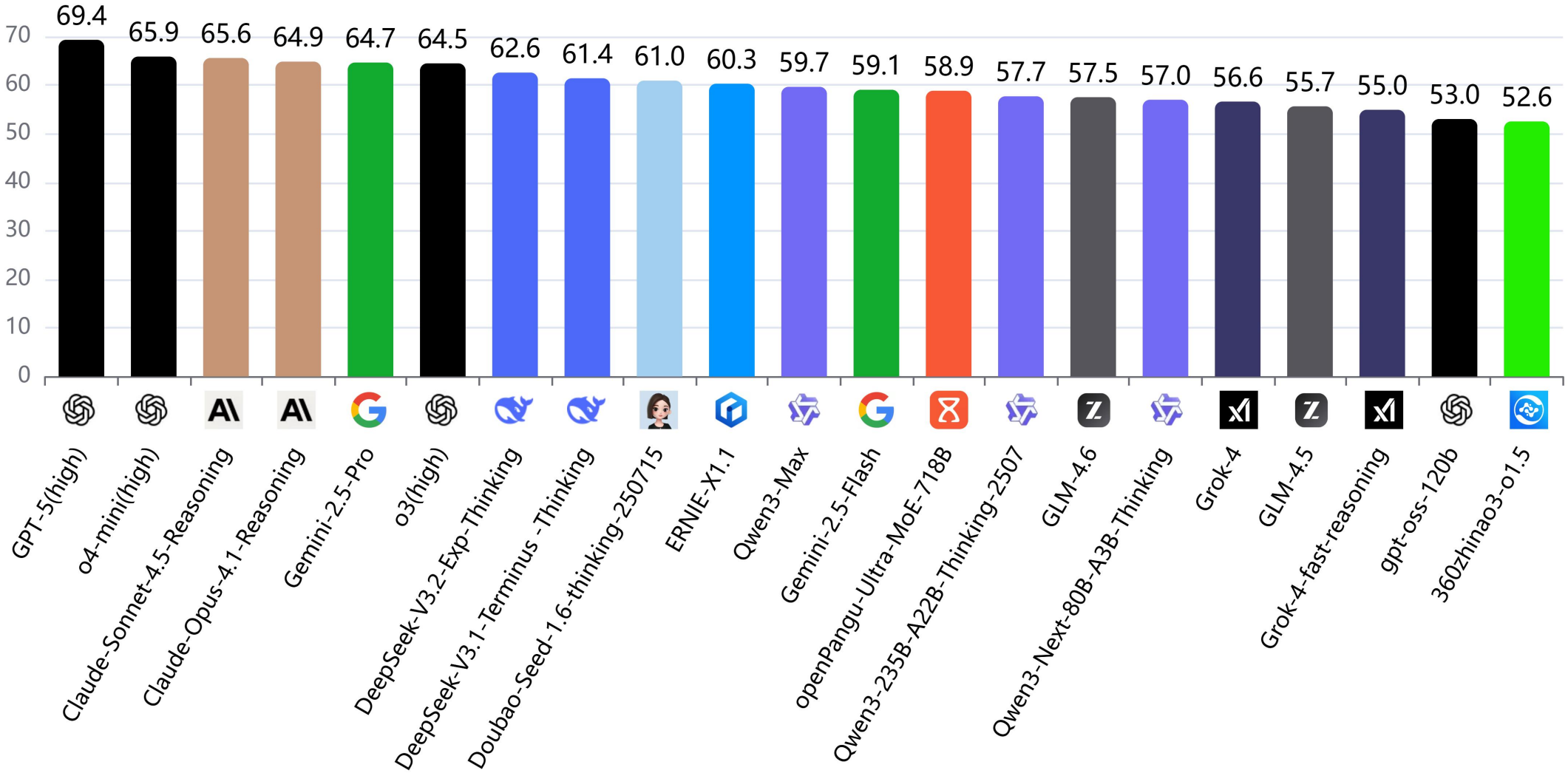
评价方式：基于人工校验参考答案的、对每个句子是否存在幻觉进行0/1评估，本次评估的人类一致性约95%。

SuperCLUE全球大模型中文综合能力排行榜（2025年9月）

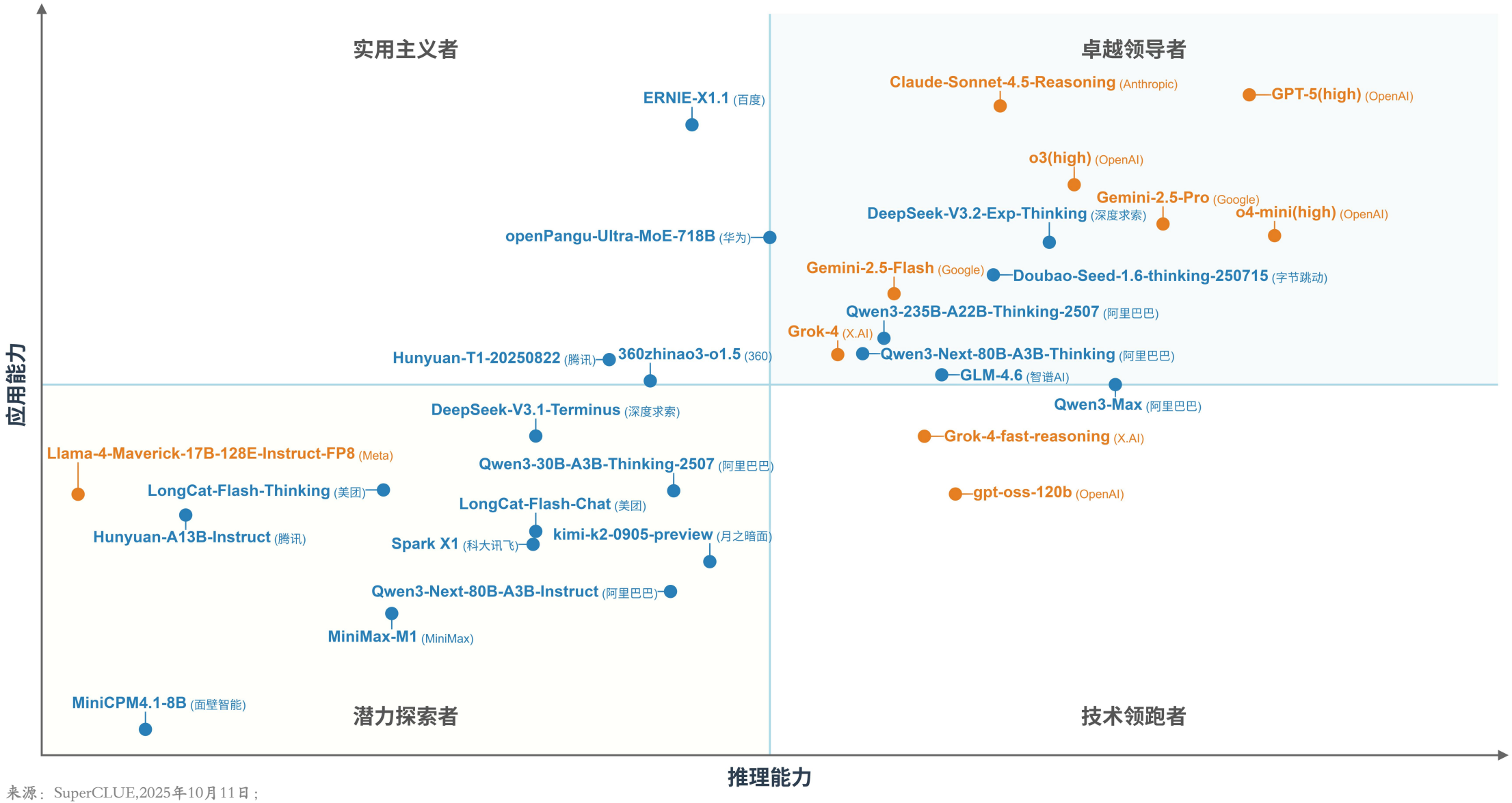
SuperCLUE
中文大模型综合性测评基准

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Agent（多轮工具调用）、幻觉控制、精确指令遵循。题目量为1260道新题，共测评33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



SuperCLUE模型象限（202509）



来源：SuperCLUE,2025年10月11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Agent；
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。

SuperCLUE九月中文大模型基准测评——总榜（附六大维度分数）

SuperCLUE测评基准2025年9月总体表现（包括补测模型）

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Agent	属地	使用方式	测评日期
-	GPT-5(high)	OpenAI	闭源	**69.37**	73.64	78.35	44.83	73.26	70.85	75.28	海外	API	2025.9.28
-	o4-mini(high)	OpenAI	闭源	**65.91**	72.73	66.41	54.31	67.65	64.38	70.00	海外	API	2025.9.28
-	Claude-Sonnet-4.5-Reasoning	Anthropic	闭源	**65.62**	52.73	78.61	42.24	69.25	73.65	77.22	海外	API	2025.10.11
-	Claude-Opus-4.1-Reasoning	Anthropic	闭源	**64.87**	49.09	85.24	45.69	63.10	70.85	75.28	海外	API	2025.9.28
-	Gemini-2.5-Pro	Google	闭源	**64.68**	71.82	73.17	50.00	49.73	60.33	83.06	海外	API	2025.9.28
-	o3(high)	OpenAI	闭源	**64.51**	77.27	77.47	35.34	67.91	62.16	66.94	海外	API	2025.9.28
🥇	DeepSeek-V3.2-Exp-Thinking	深度求索	开源	**62.62**	**60.91**	71.26	45.69	**55.61**	**66.12**	76.11	国内	API	2025.10.11
-	DeepSeek-V3.1-Terminus-Thinking	深度求索	开源	**61.44**	60.00	68.92	47.41	54.81	61.68	75.83	国内	API	2025.9.28
🥈	Doubao-Seed-1.6-thinking-250715	字节跳动	闭源	**60.96**	60.55	77.78	**52.59**	37.43	54.92	**82.50**	国内	API	2025.9.28
🥈	ERNIE-X1.1	百度	闭源	**60.33**	52.45	**78.79**	33.62	**64.91**	53.86	**78.33**	国内	API	2025.9.28
🥉	Qwen3-Max	阿里巴巴	闭源	**59.68**	**62.73**	70.93	**48.28**	26.47	**67.18**	**82.50**	国内	API	2025.9.28
-	Gemini-2.5-Flash	Google	闭源	**59.07**	62.73	72.37	41.38	41.44	55.69	80.83	海外	API	2025.9.28
🥉	openPangu-Ultra-MoE-718B	华为	开源	**58.87**	53.64	**81.29**	37.93	51.07	57.92	71.39	国内	API	2025.9.28
4	Qwen3-235B-A22B-Thinking-2507	阿里巴巴	开源	**57.73**	58.18	61.37	**46.55**	51.60	54.25	74.44	国内	API	2025.9.28
4	GLM-4.6	智谱AI	开源	**57.55**	54.55	73.08	43.97	40.37	**65.25**	68.06	国内	API	2025.10.11
4	Qwen3-Next-80B-A3B-Thinking	阿里巴巴	开源	**57.02**	**66.36**	60.75	35.34	**56.95**	55.50	67.22	国内	API	2025.9.28
-	Grok-4	X.AI	闭源	**56.65**	52.73	73.52	37.93	39.57	64.48	71.67	海外	API	2025.9.28
-	GLM-4.5	智谱AI	开源	**55.67**	52.73	66.78	43.10	36.10	66.41	68.89	国内	API	2025.9.28
-	Grok-4-fast-reasoning	X.AI	闭源	**54.95**	62.73	67.54	39.66	25.94	59.94	73.89	海外	API	2025.9.28
-	gpt-oss-120b	OpenAI	开源	**53.05**	70.91	55.03	50.86	50.00	43.15	48.33	海外	API	2025.9.28
5	360zhinao3-o1.5	360	闭源	**52.55**	50.00	68.45	43.10	39.84	41.70	72.22	国内	API	2025.9.28
5	Hunyuan-T1-20250822	腾讯	闭源	**52.29**	49.09	**77.96**	40.52	45.45	40.15	60.56	国内	API	2025.9.28
6	Qwen3-30B-A3B-Thinking-2507	阿里巴巴	开源	**48.63**	54.55	54.94	36.21	42.25	46.91	56.94	国内	API	2025.9.28
6	DeepSeek-V3.1-Terminus	深度求索	开源	**48.03**	45.45	71.44	41.38	24.06	33.88	71.94	国内	API	2025.9.28
7	kimi-k2-0905-preview	月之暗面	开源	**46.51**	46.36	62.12	35.34	18.45	60.42	56.39	国内	API	2025.9.28
8	Qwen3-Next-80B-A3B-Instruct	阿里巴巴	开源	**44.50**	53.77	55.41	30.43	17.91	53.09	56.39	国内	API	2025.9.28
8	LongCat-Flash-Chat	美团	开源	**44.17**	36.36	50.34	37.93	26.74	46.43	67.22	国内	API	2025.9.28
8	Spark X1	科大讯飞	闭源	**43.59**	42.73	61.96	30.17	20.59	47.49	58.61	国内	API	2025.9.28
9	LongCat-Flash-Thinking	美团	开源	**42.72**	37.27	60.14	29.31	41.98	35.42	52.22	国内	API	2025.9.28
10	MiniMax-M1	MiniMax	开源	**37.90**	31.82	62.22	30.17	23.26	41.03	38.89	国内	API	2025.9.28
10	Hunyuan-A13B-Instruct	腾讯	开源	**37.66**	29.09	64.76	17.24	24.60	31.37	58.89	国内	API	2025.9.28
-	Llama-4-Maverick-17B-128E-Instruct-FP8	Meta	开源	**36.30**	16.36	68.39	18.10	23.53	30.02	61.39	海外	API	2025.9.28
11	MiniCPM4.1-8B	面壁智能	开源	**28.17**	33.72	48.09	13.83	18.45	25.19	29.72	国内	模型	2025.9.28

注：数据来源SuperCLUE，2025年10月11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。海外模型及补测模型的旧版本仅作对比参考，不参与排名。标红为国内前三。

测评分析

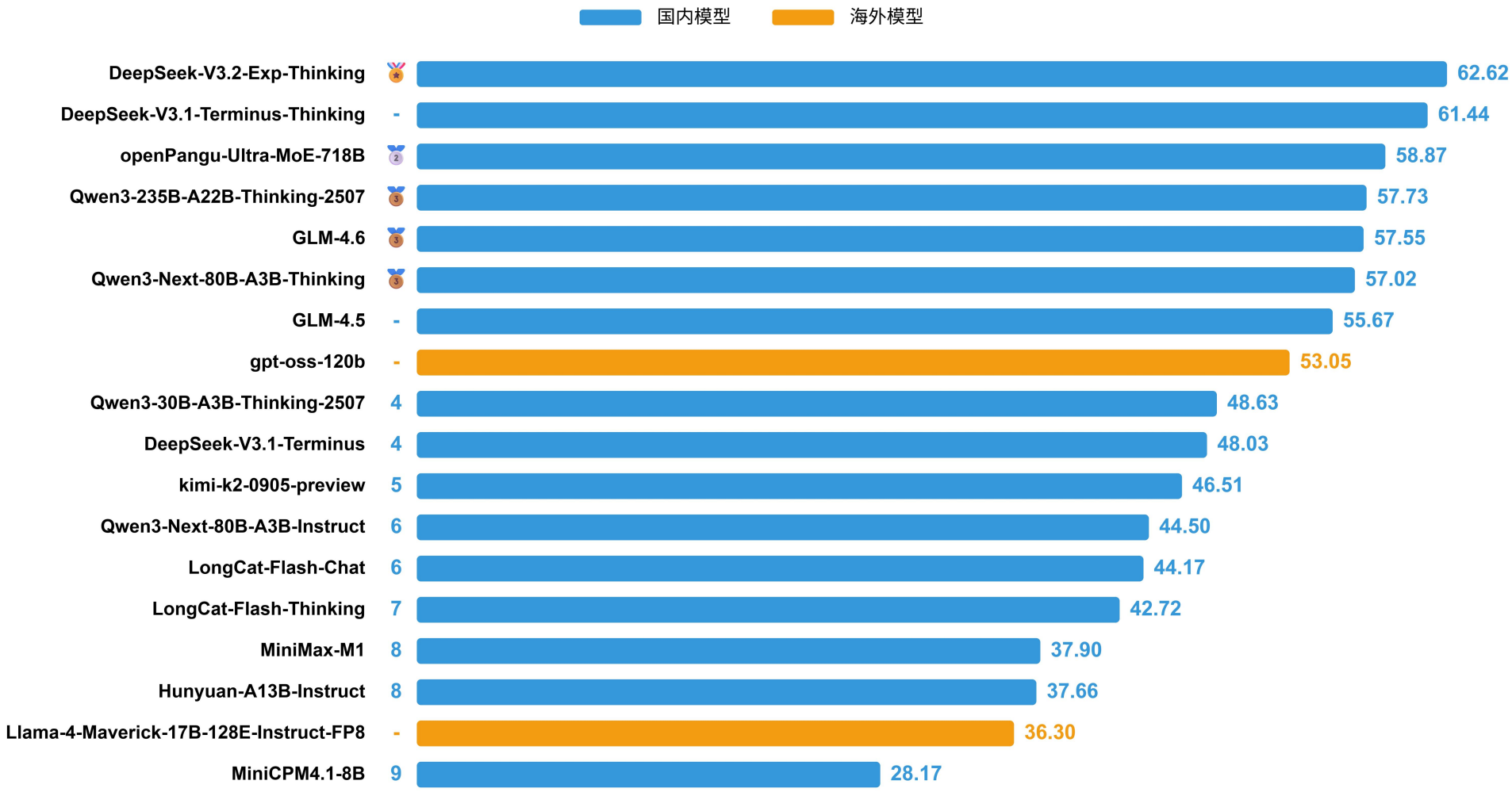
1. 国内外头部大模型的差距依旧存在。

在本次9月通用测评中，TOP6均为海外模型，TOP7-11为国内大模型。海外TOP5平均分为66.09分，国内TOP5平均分为61.01分，相差近5分，差距依然存在。

2. 国内头部模型之间竞争激烈。

国内的DeepSeek-V3.2-Exp-Thinking、Doubao-Seed-1.6-thinking-250715和ERNIE-X1.1分别位于本次测评的国内前三，Qwen3-Max和openPangu-Ultra-MoE-718B紧随其后，整体差距较小。

SuperCLUE2025年9月中文大模型基准测评开源模型总分对比（加入补测模型）



数据来源：SuperCLUE，2025年10月11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。海外模型及补测模型的旧版本仅对比参考，不参与排名。

测评分析

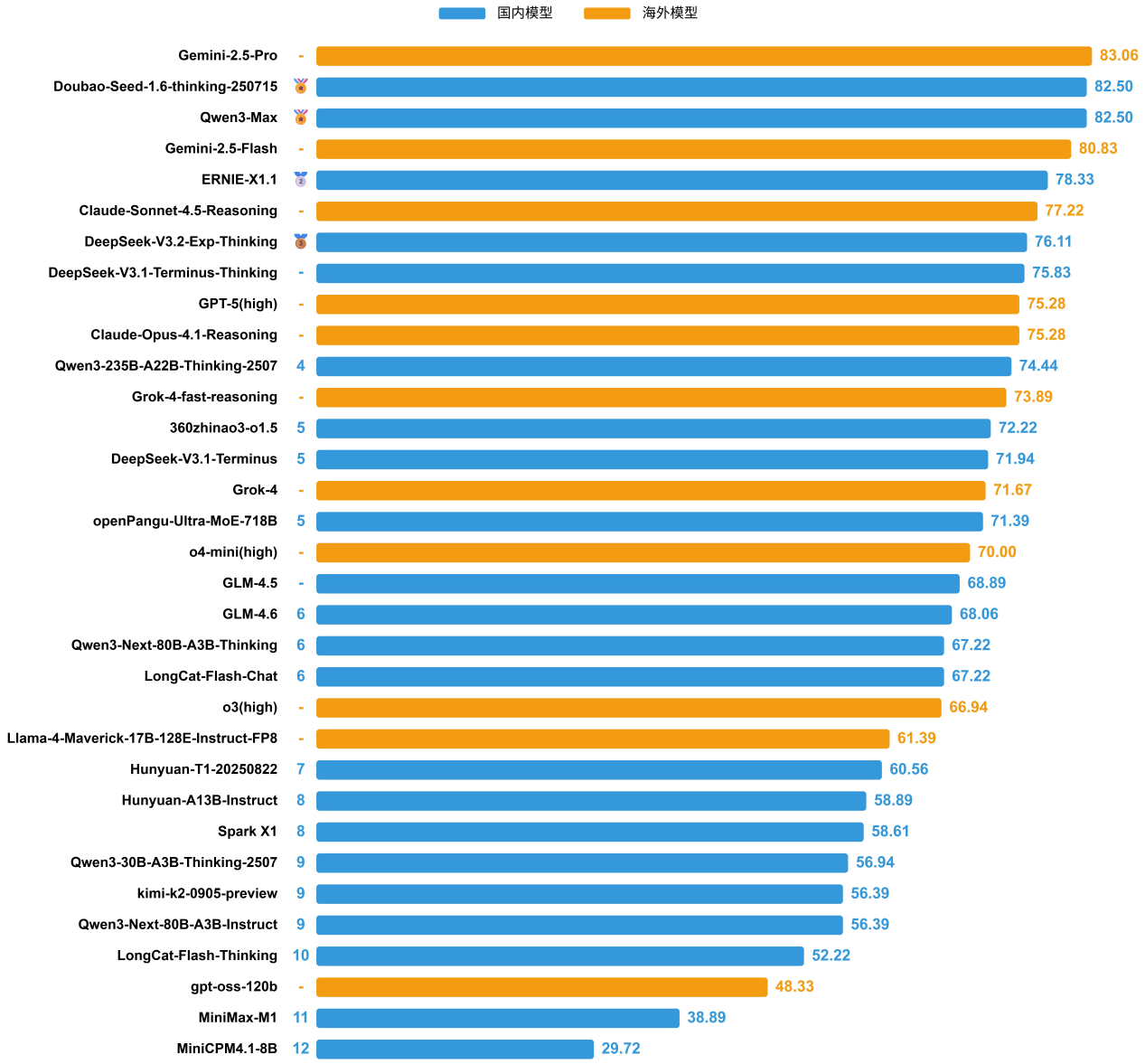
1. 国内开源模型全面领先海外开源模型。

在开源模型榜单中TOP10国内模型占据9席，仅有一个海外开源模型进入TOP10。其中DeepSeek-V3.2-Exp -Thinking以62.62分夺得开源榜首，openPangu-Ultra-MoE-718B位居开源第二，Qwen3系列以及GLM-4.6并列开源第三。

2. 国内外模型在不同任务上各有领先。

在数学和科学推理任务上，海外开源模型有一定的优势，但在代码生成、幻觉控制、智能体Agent、精确指令遵循四大任务上，国内开源模型均有领先，且领先幅度较大。

SuperCLUE2025年9月中文大模型基准测评智能体Agent任务总分对比（加入补测模型）



数据来源：SuperCLUE，2025年10月11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。海外模型及补测模型的旧版本仅对比参考，不参与排名。

测评分析

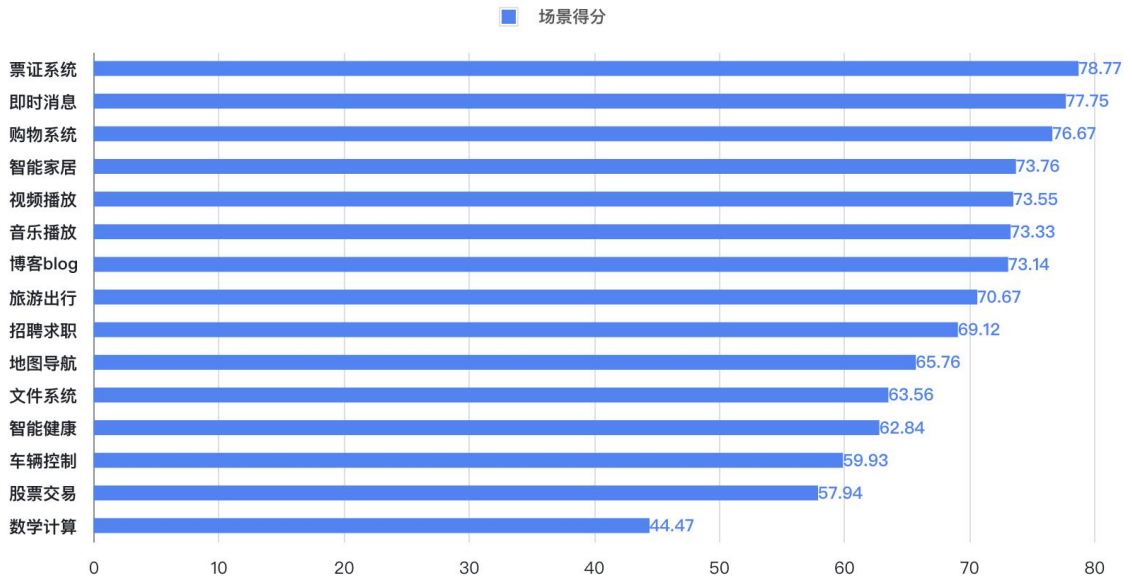
1. 在智能体Agent任务上，国内外头部大模型整体表现相当。

在本次智能体Agent任务测评中，国内TOP5的平均分为79.05分，海外TOP5的平均分为78.33分，国内外头部大模型整体表现相当。

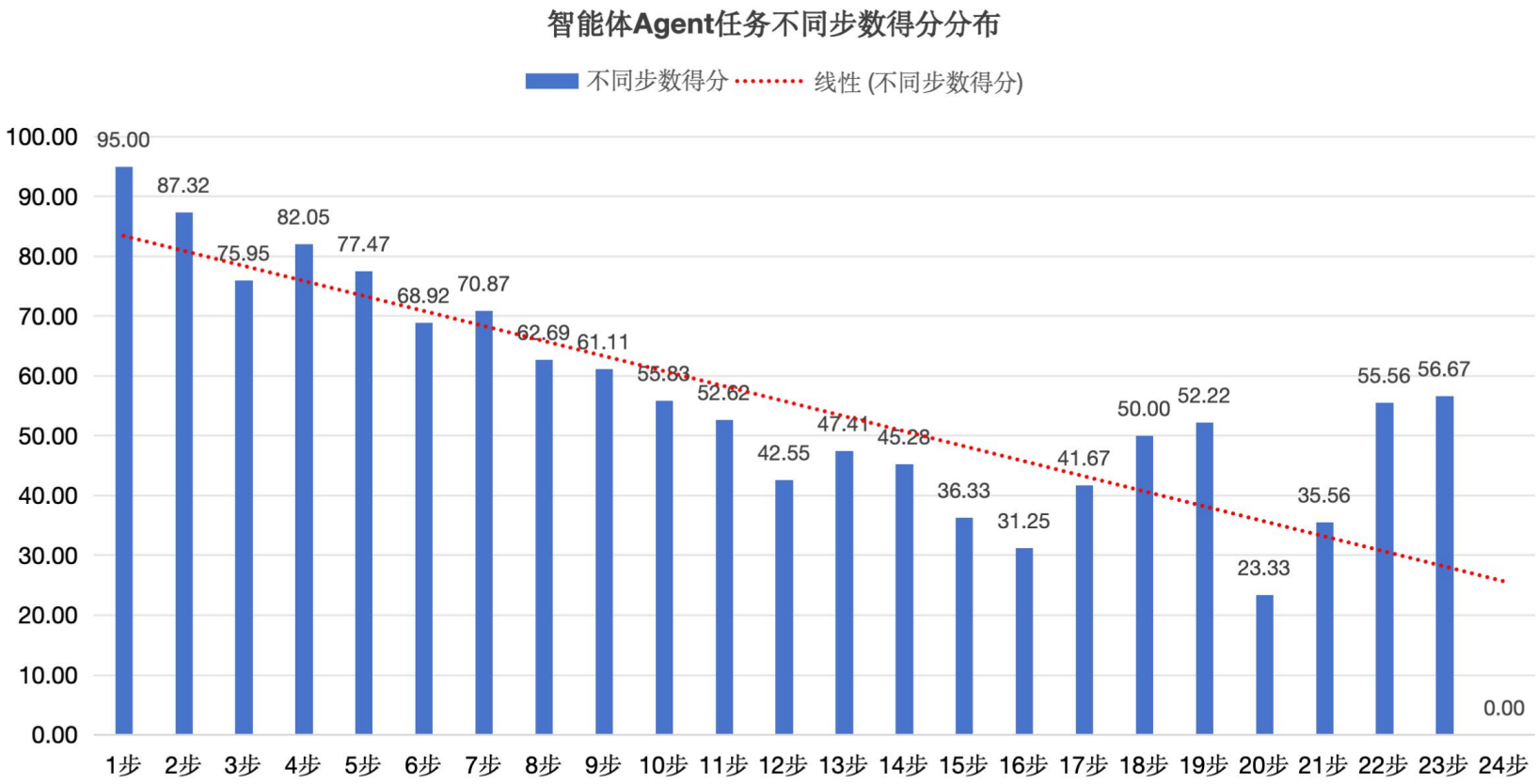
2. 模型在不同场景中的表现存在一定差异。

本次智能体Agent任务共设计了15大场景，国内外大模型在票证系统、即时消息、购物系统三大场景表现更好，总体平均分达到了75分以上。但在车辆控制、股票交易、数学计算三大场景上的表现还有待提高。总体来看，对于需要模型进行较多反思推理和数学计算的场景，得分普遍较低；而对于日常生活中较为简单的工具调用任务，如购票、发送消息等，模型的表现更佳。

智能体Agent任务15大场景得分分布



数据来源：SuperCLUE，2025年10月11日。

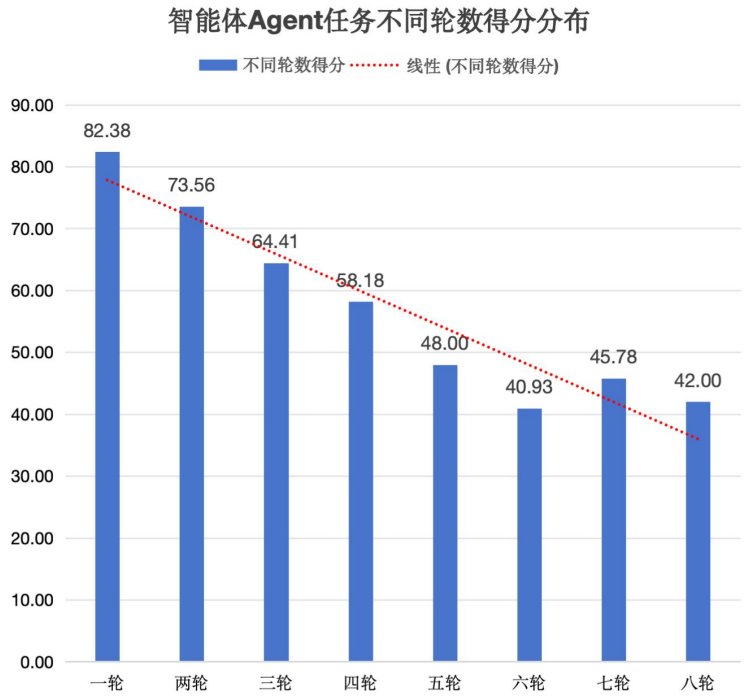


数据来源：SuperCLUE，2025年10月11日。

测评分析

3. 随着题目轮数和步数的增加，模型得分整体呈下降趋势。

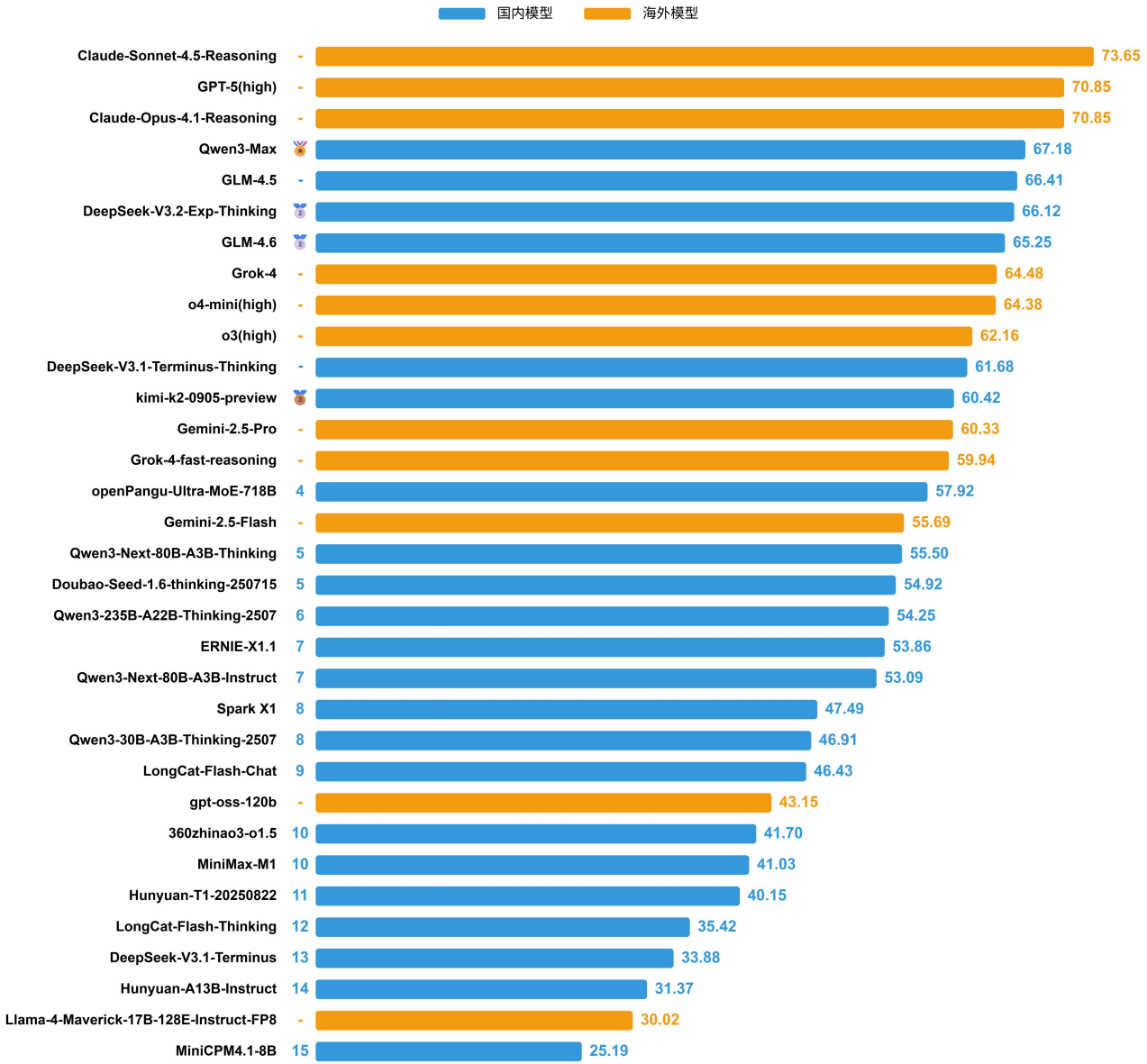
本次智能体Agent任务的测评题轮数从1到8轮不等，步数从1到24步不等。从评测结果可以看出，随着交互步数和轮数的增加，模型的平均得分整体呈现下降趋势。



数据来源：SuperCLUE，2025年10月11日。

SuperCLUE九月中文大模型基准测评——代码生成任务

SuperCLUE2025年9月中文大模型基准测评代码生成任务总分对比（加入补测模型）



数据来源：SuperCLUE，2025年10月11日。

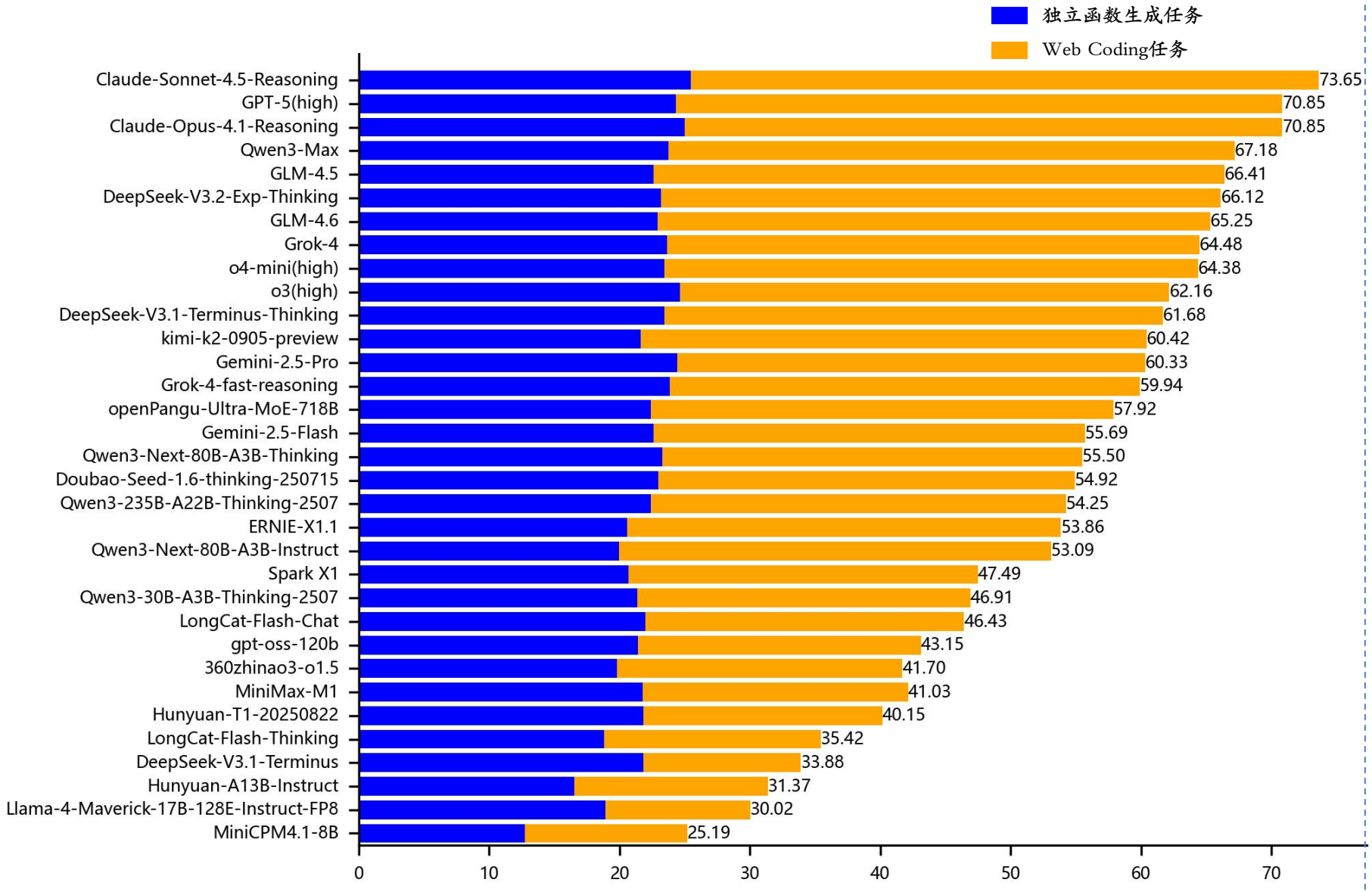
注：由于部分模型分数较为接近，为了减少问题波动对排名的影响，本次测评将相距1分区间的模型定义为并列。海外模型及补测模型的旧版本仅对比参考，不参与排名。

测评分析

1. 国内头部大模型表现优异，但与海外顶尖大模型还有一定差距。

总体来看，国内大模型在代码生成任务上表现优异，Qwen3-Max、GLM系列、DeepSeek-V3.2-Exp-Thinking均有超过Grok-4、o4-mini(high)等海外模型的表现。但相较于在编程领域TOP3级别的海外顶尖模型，如Claude-Sonnet-4.5-Reasoning、GPT-5(high)、Claude-Opus-4.1-Reasoning而言，还有一定的差距。

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, 2025年10月11日。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。

测评分析

2.Web Coding是模型之间拉开差距的主要原因。

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2.51，但在Web Coding子任务上的标准差达到了10.84，是拉开模型在代码生成任务上差距的主要原因。

3. 模型在Web Coding子任务上的表现相较于独立函数生成子任务更差。

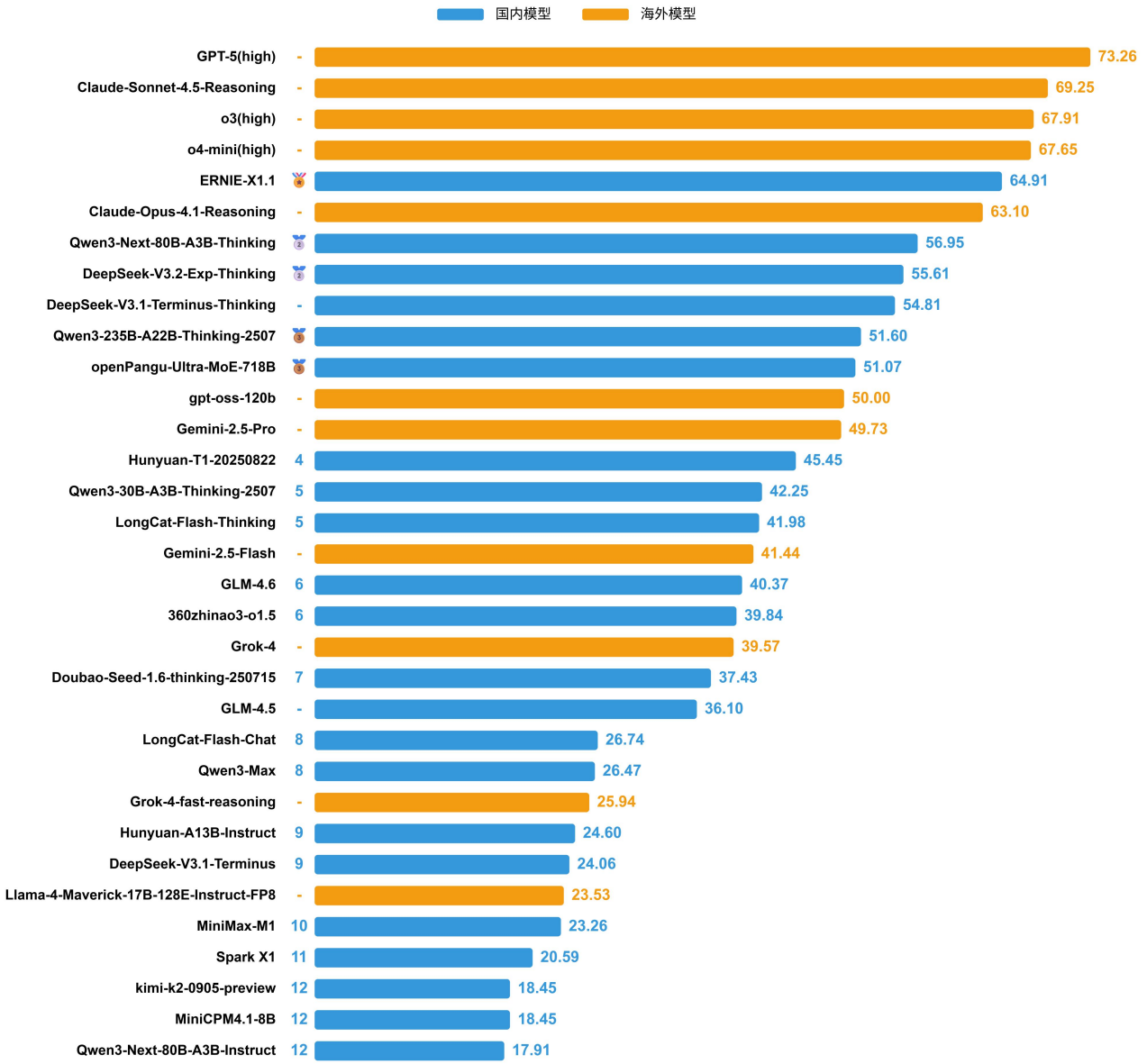
从单一子任务来看，国内外大模型在独立函数生成任务上的平均得分为83.88分，而在Web Coding任务上的平均得分仅为42.63分，相差超过41分。国内外大模型在Web Coding任务上还有较大的提升空间。（左图为两大子任务加权后的总分，非单一子任务得分。）

4. 国内模型在Web Coding和独立函数生成两个子任务上均与海外模型存在一定差距。

国内大模型在独立函数生成和Web Coding子任务上的平均分分别为81.03分和39.46分，海外大模型分别为89.57分和48.97分，差距均在8.5分以上。

SuperCLUE九月中文大模型基准测评——精确指令遵循任务

SuperCLUE2025年9月中文大模型基准测评精确指令遵循任务总分对比（加入补测模型）



数据来源：SuperCLUE，2025年10月11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。海外模型及补测模型的旧版本仅对比参考，不参与排名。

测评分析

1. 海外头部大模型的表现普遍优于国内头部大模型。

在本次精确指令遵循任务中，海外头部大模型GPT-5(high)、Claude-Sonnet-4.5-Reasoning、o3(high)等占据榜单前四，国内大模型ERNIE-X1.1以64.91分位居国内第一、全球第五，表现亮眼。总体而言，国内大模型和海外大模型相比，在精确指令遵循上的表现差距较大。

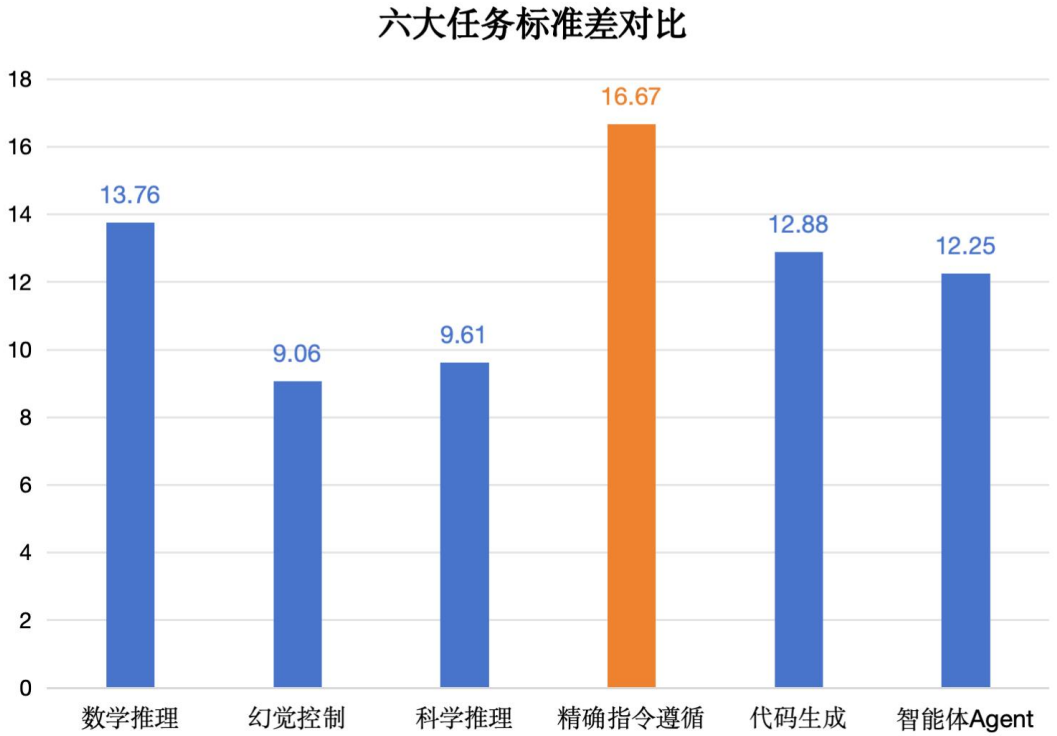
2. 大模型在该任务上的表现差异显著。

本次测评的33个大模型在该任务上的标准差达到了16.67，在六大任务中最高，表明国内外大模型在精确指令遵循任务上的表现差异显著。

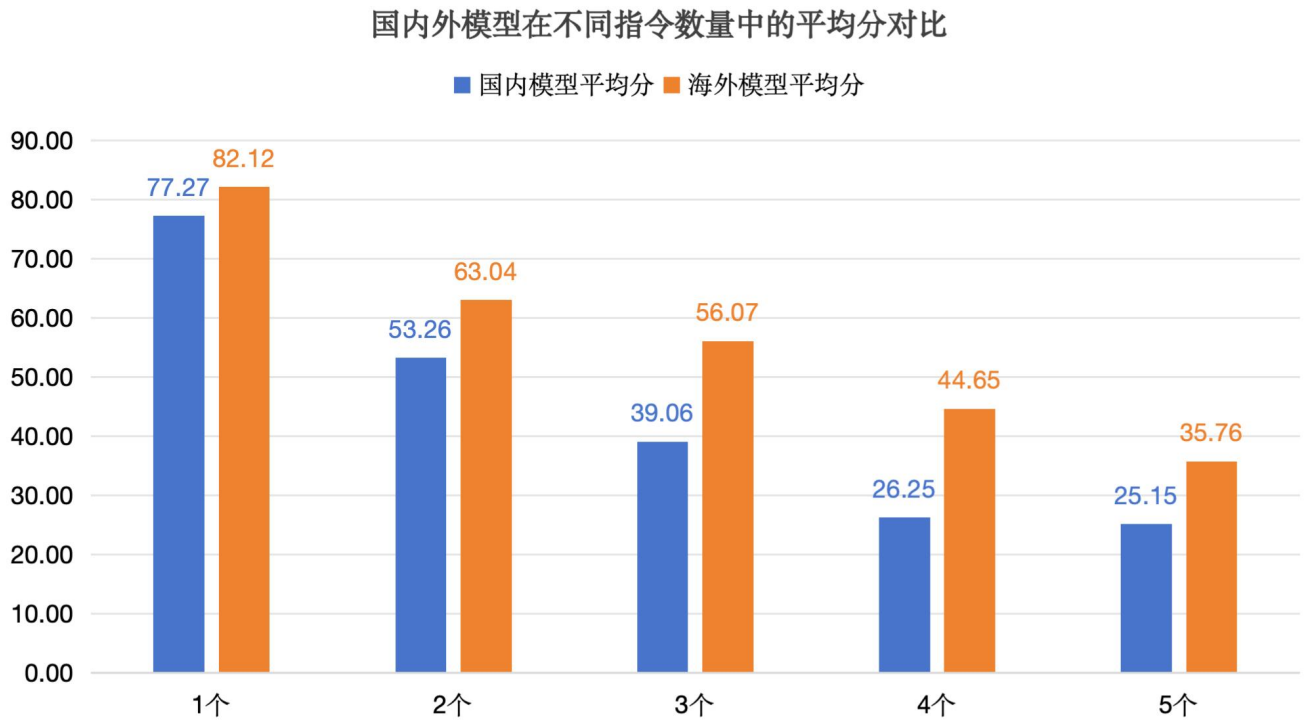
3. 随着指令数量的增加，模型得分逐渐降低，但海外模型的表现始终优于国内模型。

无论是国内还是海外模型，其平均分都随着指令数量的增加而显著下降。这揭示了当前大模型在处理简单、单一的指令时表现尚可，但随着任务复杂度的提升（指令叠加），其理解、记忆和执行能力会急剧下降，出现遗忘指令、混淆指令、执行不完整等问题。

在所有指令数量的测试中，海外模型的平均分均高于国内模型，这说明海外模型在处理复杂、多步骤任务时的鲁棒性更强。



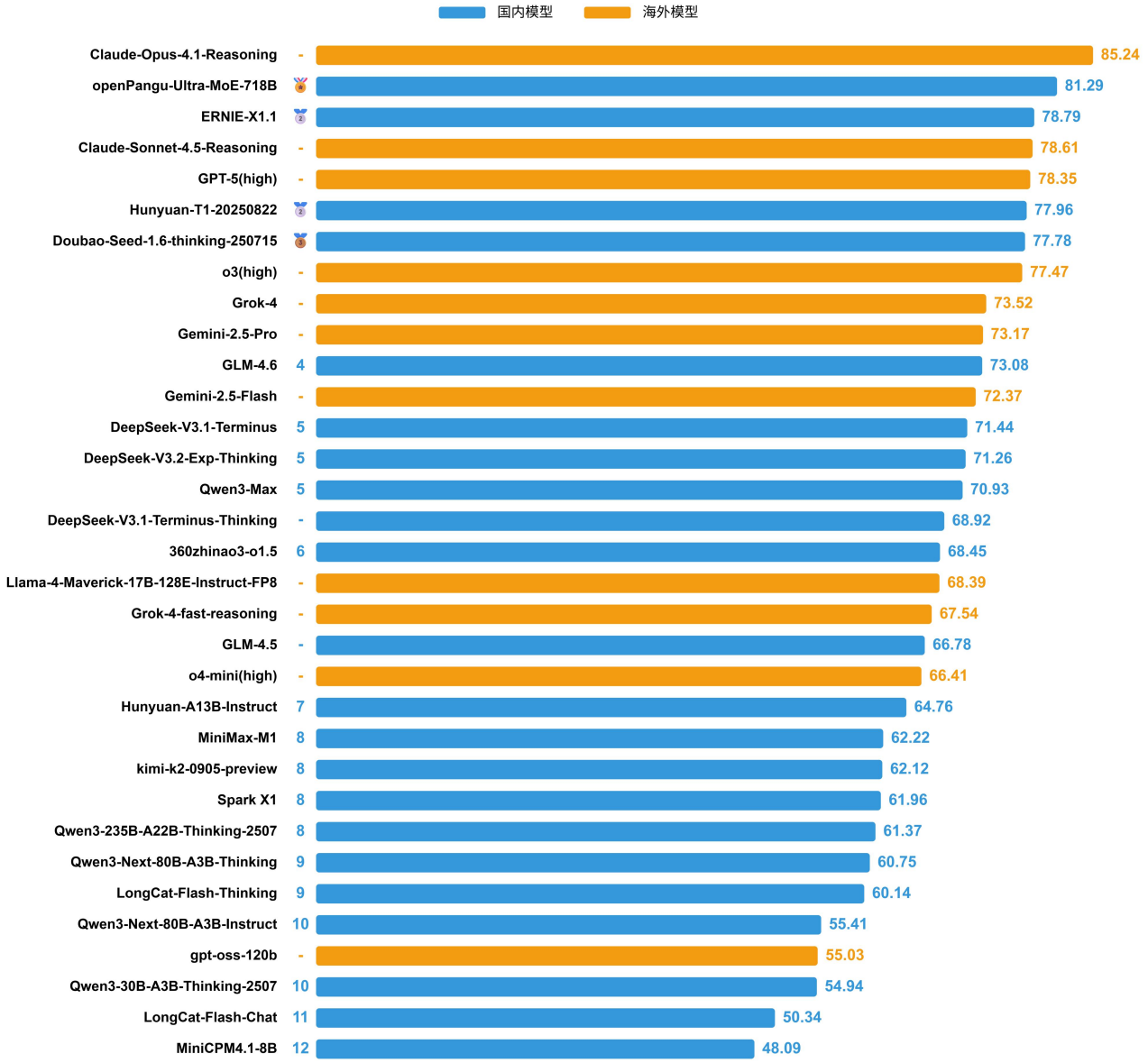
数据来源：SuperCLUE，2025年10月11日。



数据来源：SuperCLUE，2025年10月11日。

SuperCLUE九月中文大模型基准测评——幻觉控制任务

SuperCLUE2025年9月中文大模型基准测评幻觉控制任务总分对比（加入补测模型）



数据来源：SuperCLUE，2025年10月11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。海外模型及补测模型的旧版本仅对比参考，不参与排名。

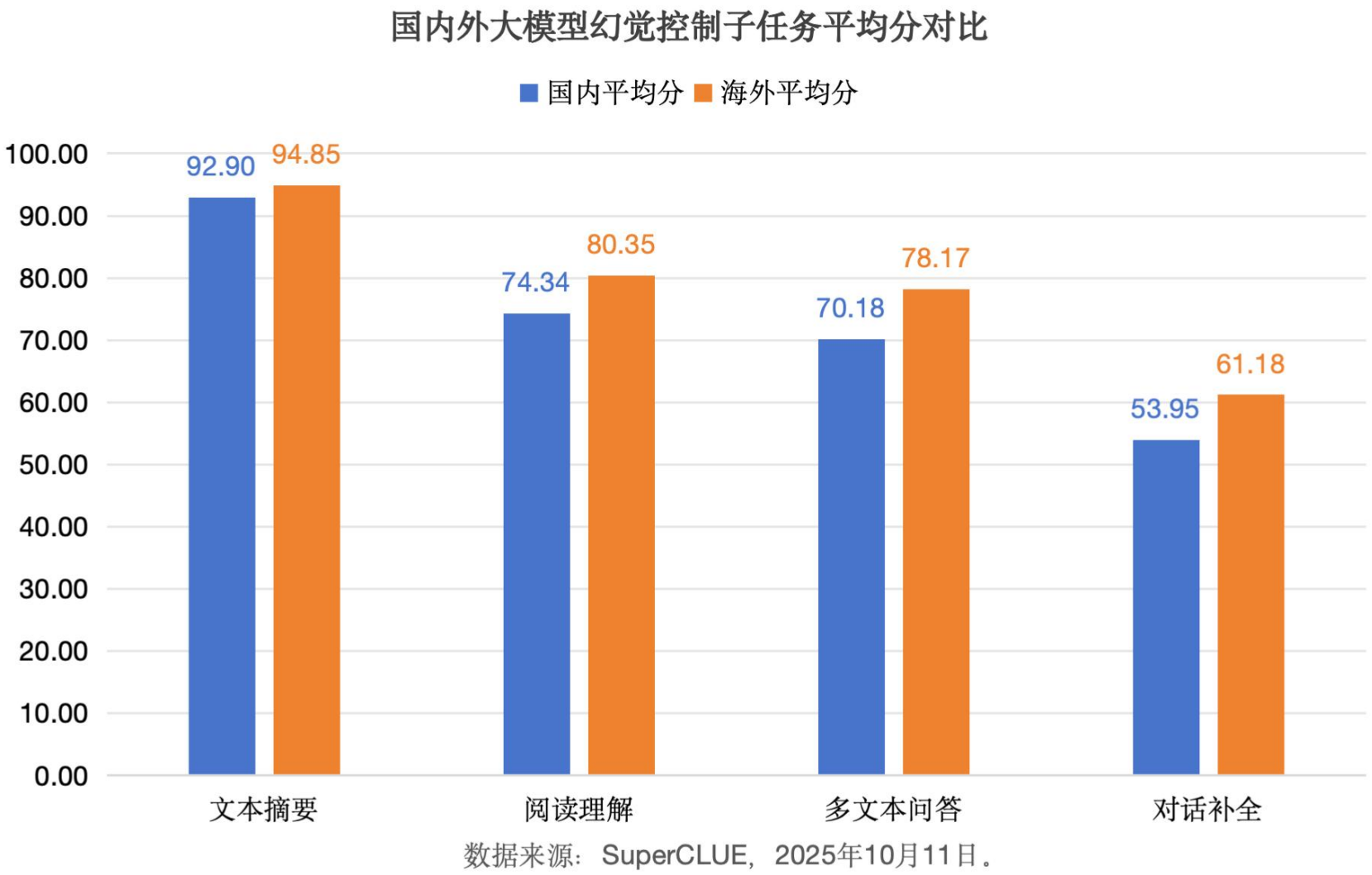
测评分析

1. 海外模型占据头部优势。

排名前十的模型中，海外模型占据了6席，且包揽了前三名中的两个席位，显示出在幻觉控制领域的整体领先地位。其中海外模型 Claude-Opus-4.1-Reasoning 以85.24 分的微弱优势夺得榜首，紧随其后的是国内模型 openPanGu-Ultra-MoE-718B（81.29分）。

2. 国产模型梯队分明。

国产模型内部也形成了明显的梯队。以openPanGu-Ultra-MoE-718B、ERNIE-X1.1、Hunyuan-T1-20250822等国内头部模型为代表的第一梯队已经具备与国际顶尖模型一较高下的实力，而其他模型则分布在中后段。



数据来源：SuperCLUE，2025年10月11日。

测评分析

3. 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。

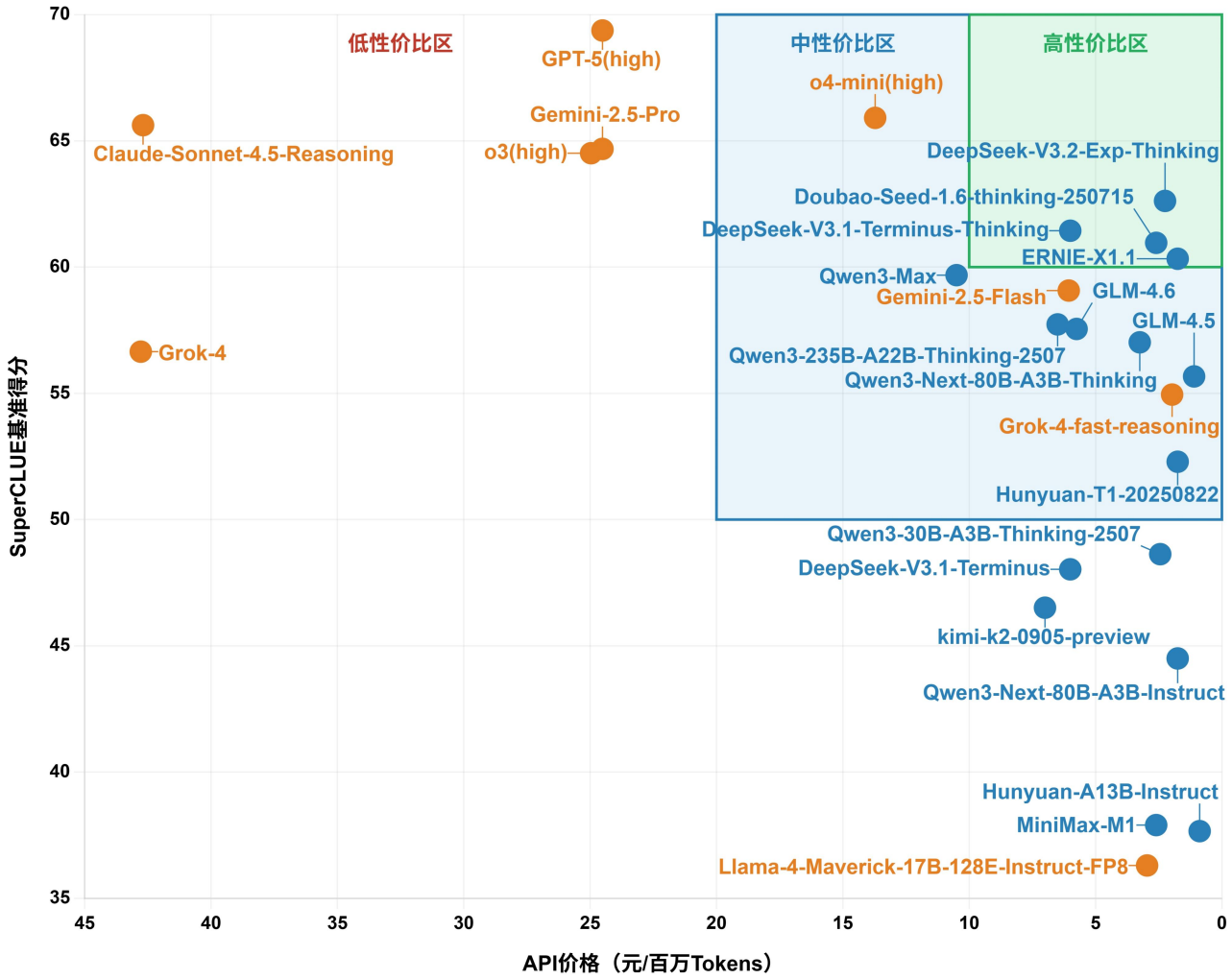
阅读理解虽然也基于给定文本，但该任务要求模型进行一定程度的推理和判断，而不仅仅是复述。这个推理过程为产生幻觉提供了空间，导致分数低于文本摘要。

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。

SuperCLUE9月中文大模型通用测评性价比区间分布



数据来源：SuperCLUE，2025年10月11日；

注：开源模型如Qwen3-235B-A22B-Thinking-2507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9月的标准价格（非优惠价格）。补测模型选取实时价格。

趋势分析

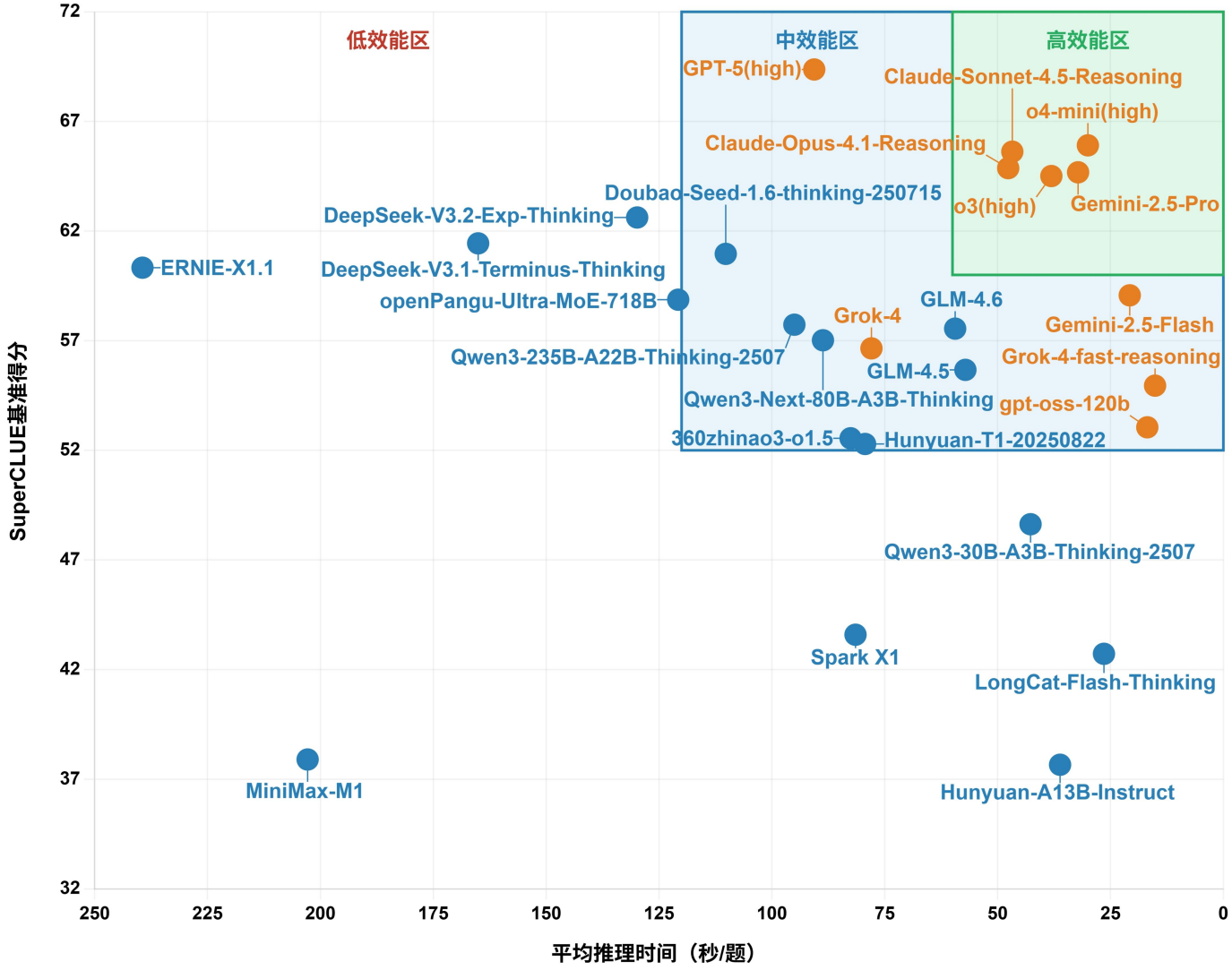
1. 国内模型具有更高的性价比。

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，没有处于高性价比区间的海外模型。国内模型平均API价格为3.88元/百万Tokens，海外模型平均API价格为20.46元/百万Tokens，海外模型的价格远高于国内模型。

2. 国内模型的API价格分布更加集中。

国内模型的API价格大多数处于0-10元/百万Tokens，而海外模型的API价格比较分散，从2-200元/百万Tokens不等（Claude-Opus-4.1-Reasoning的价格为213.9元/百万Tokens，未在图中标注）。

SuperCLUE9月中文大模型通用测评推理模型推理效能区间分布



数据来源：SuperCLUE，2025年10月11日；
模型推理速度选取9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）。

趋势分析

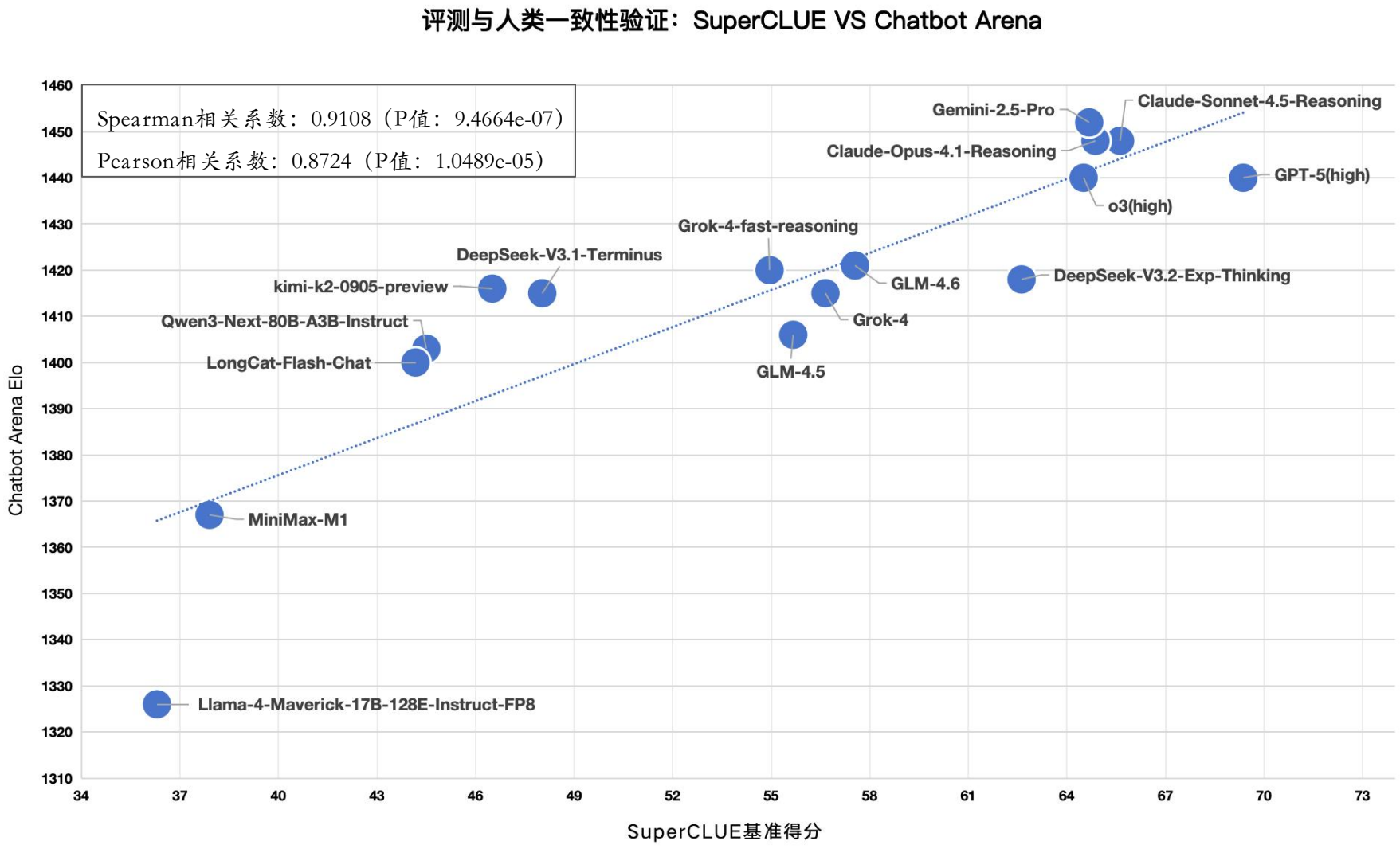
1. 海外推理模型推理效率远高于国内推理模型。

海外推理模型大多数分布在中高效能区间，而国内推理模型全部位于中低效能区间，没有进入高效能区的模型。国内推理模型平均每题的推理耗时为101.07秒，而海外推理模型仅有41.60秒，海外推理模型的推理效率远高于国内推理模型。

2. 海外推理模型的推理时间更加集中。

海外推理模型的平均推理时间集中在10-90秒的区间，而国内推理模型的平均推理时间较为分散，从10-240秒不等。

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：

斯皮尔曼（Spearman）相关系数：0.9108，P值：9.4664e-07；皮尔逊（Pearson）相关系数：0.8724，P值：1.0489e-05。

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**高度一致性**。

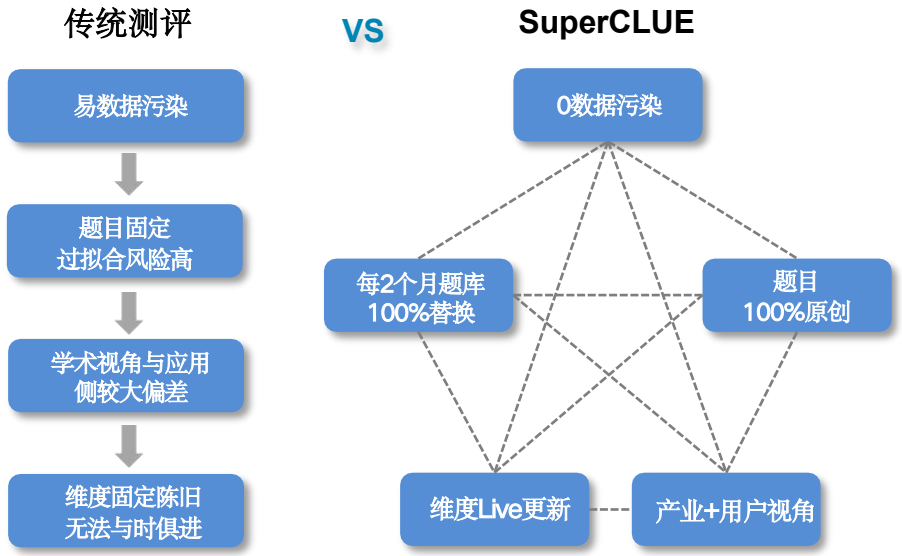
数据来源：SuperCLUE, 2025年10月11日。
斯皮尔曼（Spearman）相关系数：用于衡量两个变量之间的单调关系，取值为[−1,1]，该系数的绝对值越接近1表示两个变量之间的相关性越强；
皮尔逊相关系数：用于衡量两个连续变量之间的线性相关程度，取值为[−1,1]，该系数的绝对值越接近1表示两个变量之间的相关性越强。

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**
- **2024.01-06** — 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**
- **2024.07-12** — 发布AI编程助手等**AI应用测评基准；**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
- **2024.12-2025.09** — 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、世界模型等**多模态**系列测评基准；发布**幻觉、深度研究、精确指令遵循、深度搜索、软件工程**SWE等测评基准；发布手机和电脑GUI Agent测评基准

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%替换且全部原创，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Agent
- 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
- 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
- 具身智能 AgentCLUE-VLA
- 深度搜索 DeepSearch
- 深度研究 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搜索 SuperCLUE-AISearch
- 代码助手 SuperCLUE-Coder
- 实时音视频 SuperCLUE-Live

行业
- 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。

附录三：9月测评模型列表

SuperCLUE-9月测评选取了国内外有代表性的**33个大模型**（包括3个补测模型）。

模型	机构	简介	模型	机构	简介
1. Qwen3-Max	阿里巴巴	官方发布的最新旗舰模型，使用阿里云公开的API：qwen3-max。	18.grok-4-fast-reasoning	X.AI	官方发布的最新推理模型，使用官方API：grok-4-fast-reasoning。
2. Qwen3-Next-80B-A3B-Instruct	阿里巴巴	官方发布的基础模型，使用阿里云公开的API：qwen3-next-80b-a3b-instruct。	19.MiniCPM4.1-8B	面壁智能	官方开源版本。对应huggingface仓库名称：openbmb/MiniCPM4.1-8B。
3. Qwen3-Next-80B-A3B-Thinking	阿里巴巴	官方发布的推理模型，使用阿里云公开的API：qwen3-next-80b-a3b-thinking。	20.MiniMax-M1	MiniMax	官方发布的最新开源混合推理模型，使用官方API：MiniMax-M1。
4. Qwen3-30B-A3B-Thinking-2507	阿里巴巴	官方发布的推理模型，使用阿里云公开的API：qwen3-30b-a3b-thinking-2507。	21.LongCat-Flash-Chat	美团	官方开源的最新基础模型，使用官方API：LongCat-Flash-Chat。
5. DeepSeek-V3.1-Terminus -Thinking	深度求索	官方发布的推理模型，使用官方API。	22.LongCat-Flash-Thinking	美团	官方开源的最新推理模型，使用官方API：LongCat-Flash-Thinking。
6. DeepSeek-V3.1-Terminus	深度求索	官方发布的基础模型，使用官方API。	23.Doubao-Seed-1.6-thinking-250715	字节跳动	官方发布的深度思考模型，使用官方API：doubao-seed-1-6-thinking-250715。
7. GLM-4.5	智谱AI	官方发布的开源推理模型，使用官方API：glm-4.5。	24.ERNIE-X1.1	百度	官方发布的深度推理模型预览版，使用官方API：ERNIE-X1.1-Preview。
8. Spark X1	科大讯飞	官方发布的推理模型，使用官方API：Spark X1。	25.360zhinao3-o1.5	360	官方发布的最新推理模型，使用官方API。
9. kimi-k2-0905-preview	月之暗面	官方发布的最新开源基础模型，使用官方API：kimi-k2-0905-preview。	26.Hunyuan-T1-20250822	腾讯	官方发布的最新推理模型，使用官方API：hunyuan-t1-20250822。
10.Claude-Opus-4.1-Reasoning	Anthropic	官方发布的混合推理模型，使用官方API：Claude-Opus-4.1-Reasoning。	27.Hunyuan-A13B-Instruct	腾讯	官方开源的推理模型，使用官方API：hunyuan-a13b。
11.Gemini-2.5-Pro	Google	Gemini-2.5-Pro的正式版本，使用官方API：gemini-2.5-pro。	28.openPangu-Ultra-MoE-718B	华为	官方发布的最新推理模型，使用官方API。
12.Gemini-2.5-Flash	Google	Gemini-2.5-Flash的正式版本，使用官方API：gemini-2.5-flash。	29.Llama-4-Maverick-17B-128E-Instruct-FP8	Meta	使用together.ai的接口：meta-llama/Llama-4-Maverick-17B-128E-Instruct-FP8。
13.GPT-5(high)	OpenAI	使用方式为AZURE OpenAI Service的API接口，reasoning_effort参数设置为：high。	30.Qwen3-235B-A22B-Thinking-2507	阿里巴巴	官方开源的推理模型，使用官方API：qwen3-235b-a22b-thinking-2507。
14.o4-mini(high)	OpenAI	使用方式为AZURE OpenAI Service的API接口，reasoning_effort参数设置为：high。	31.Claude-Sonnet-4.5-Reasoning	Anthropic	官方发布的最新混合推理模型，使用官方API：Claude-Sonnet-4.5-Reasoning。
15.o3(high)	OpenAI	官方发布的新版本开源推理模型，使用官方API：deepseek-reasoner。	32.DeepSeek-V3.2-Exp-Thinking	深度求索	官方发布的最新推理模型，使用官方API。
16.gpt-oss-120b	OpenAI	官方开源的模型，使用OpenRouter的API。	33.GLM-4.6	智谱AI	官方发布的最新开源推理模型，使用官方API：glm-4.6。
17.grok-4	X.AI	官方发布的推理模型，使用官方API：grok-4-0709。	/	/	/

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9月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。