



SuperCLUE

中文大模型综合性测评基准

# 中文大模型基准测评2025年上半年报告

—— 2025年中文大模型阶段性进展半年度评估

SuperCLUE团队

2025.08.04

# 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

## 一、2025年上半年关键进展及趋势

1. 2025年上半年大模型关键进展
2. 2025年最值得关注的中文大模型及智能体全景图
3. 2025年国内外大模型差距
4. 近一年SuperCLUE通用基准测评开闭源模型最好成绩对比

## 二、七月通用测评介绍

1. SuperCLUE基准介绍
2. SuperCLUE大模型综合测评体系
3. SuperCLUE通用测评基准数据集及评价方式
4. 各维度测评说明及示例
5. 测评模型列表

## 三、总体测评结果与分析

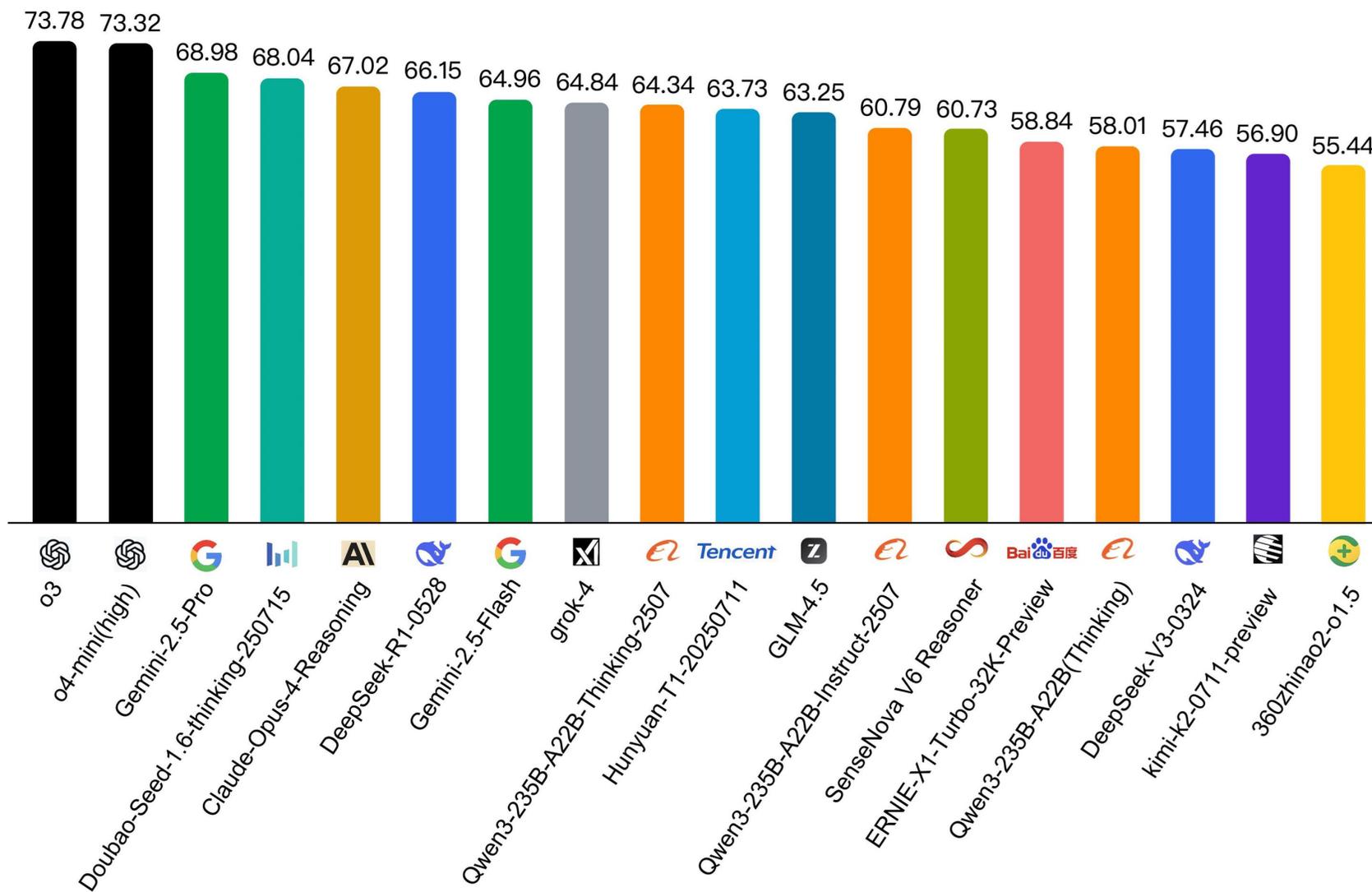
1. SuperCLUE模型象限 (202507)
2. SuperCLUE通用能力测评榜单
3. SuperCLUE性价比区间分布
4. SuperCLUE大模型综合效能区间分布
5. 代表性模型分析
6. 国内大模型成熟度-SC成熟度指数
7. 评测与人类一致性验证

## 四、专项测评基准介绍

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

## SuperCLUE全球大模型中文综合能力排行榜 (2025年7月)

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1288道新题，最终得分取各任务平均分。



- o3以73.78的总分取得总榜第一，领跑全球。

海外头部模型o3、o4-mini(high)和Gemini-2.5-Pro在本次七月通用基准测评中取得了73.78分、73.32分和68.98分的总成绩，分别位于榜单前三。Doubao-Seed-1.6-thinking-250715以68.04的总分取得国内第一、全球第四的成绩。

- 国内外头部模型之间的差异较大，海外模型在推理任务上的优势尤其明显。

海外模型在推理任务上的优势尤其显著，o3和o4-mini(high)在推理任务上分别取得了75.02和72.68的分数，领跑推理任务榜单。国内推理任务成绩最好的模型分别是DeepSeek-R1-0528和Doubao-Seed-1.6-thinking-250715，均有超过65分的表现，但与海外头部模型依旧相差近10分。

- 国内开源模型相较于海外开源模型优势显著。

DeepSeek-R1-0528、Qwen3-235B-A22B-Thinking-2507和GLM-4.5分别以66.15分、64.34分和63.25分取得了开源榜单的前三名，海外开源模型最好成绩仅有46.37分，与国内开源模型最好成绩相差近20分，国内开源模型的优势显著。

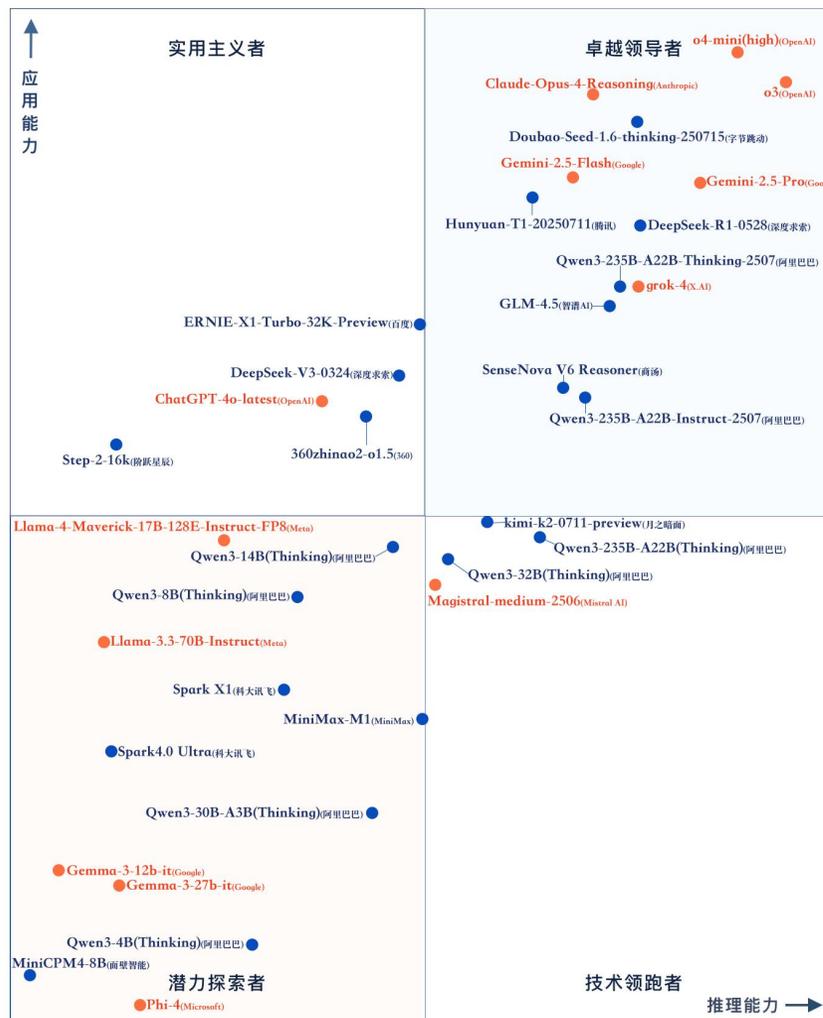
- 国内大模型在智能体Agent和幻觉控制任务上的表现良好。

在智能体Agent任务上，Doubao-Seed-1.6-thinking-250715以90.67分领跑全球，GLM-4.5和SenseNova V6 Reasoner以83.58分并列国内第二。在幻觉控制任务上，Doubao-Seed-1.6-thinking-250715、ERNIE-X1-Turbo-32K-Preview和Hunyuan-T1-20250711分别位于国内前三。

- Qwen3系列的开源小参数量模型表现亮眼。

Qwen3系列的多款开源小参数量模型展现出惊人潜力。其中8B、4B和1.7B版本分别在10B级别和端侧5B级别的榜单中遥遥领先。

## SuperCLUE模型象限 (202507)

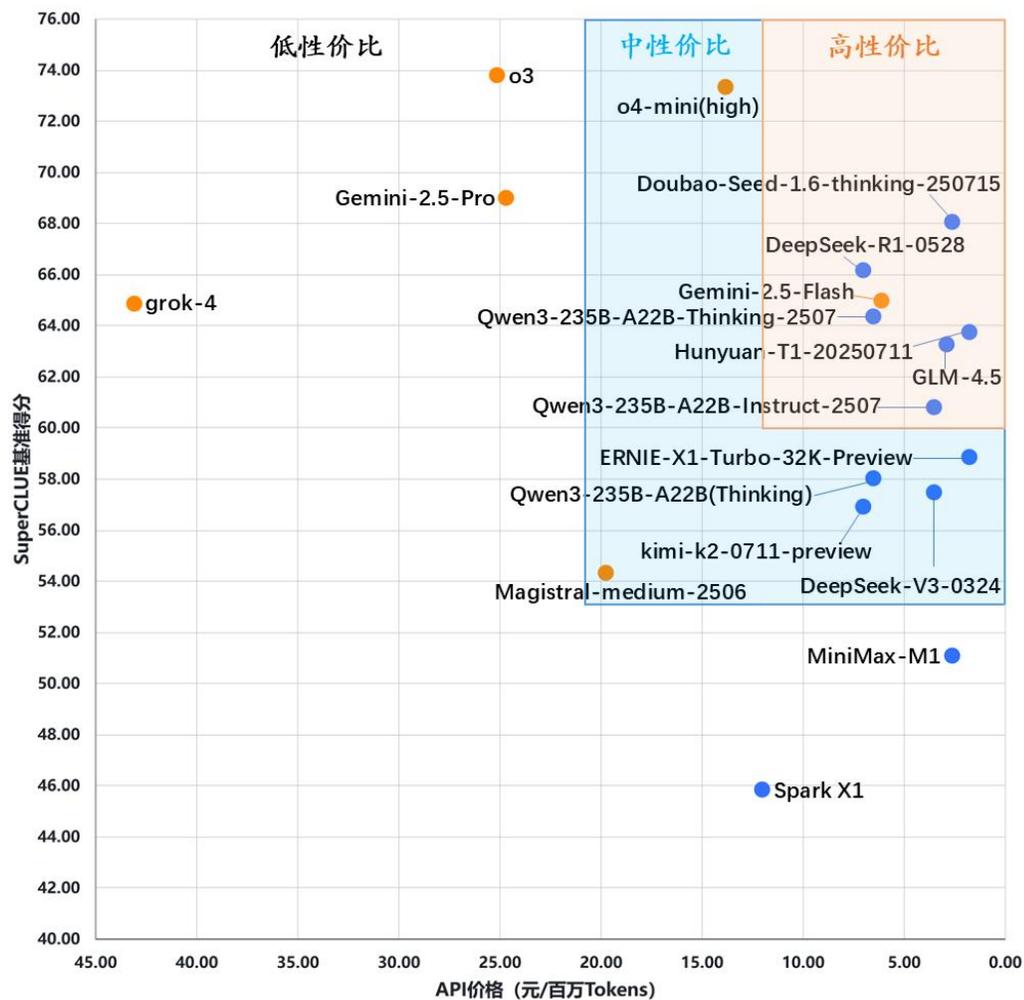


数据来源: SuperCLUE, 2025年7月31日;

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;

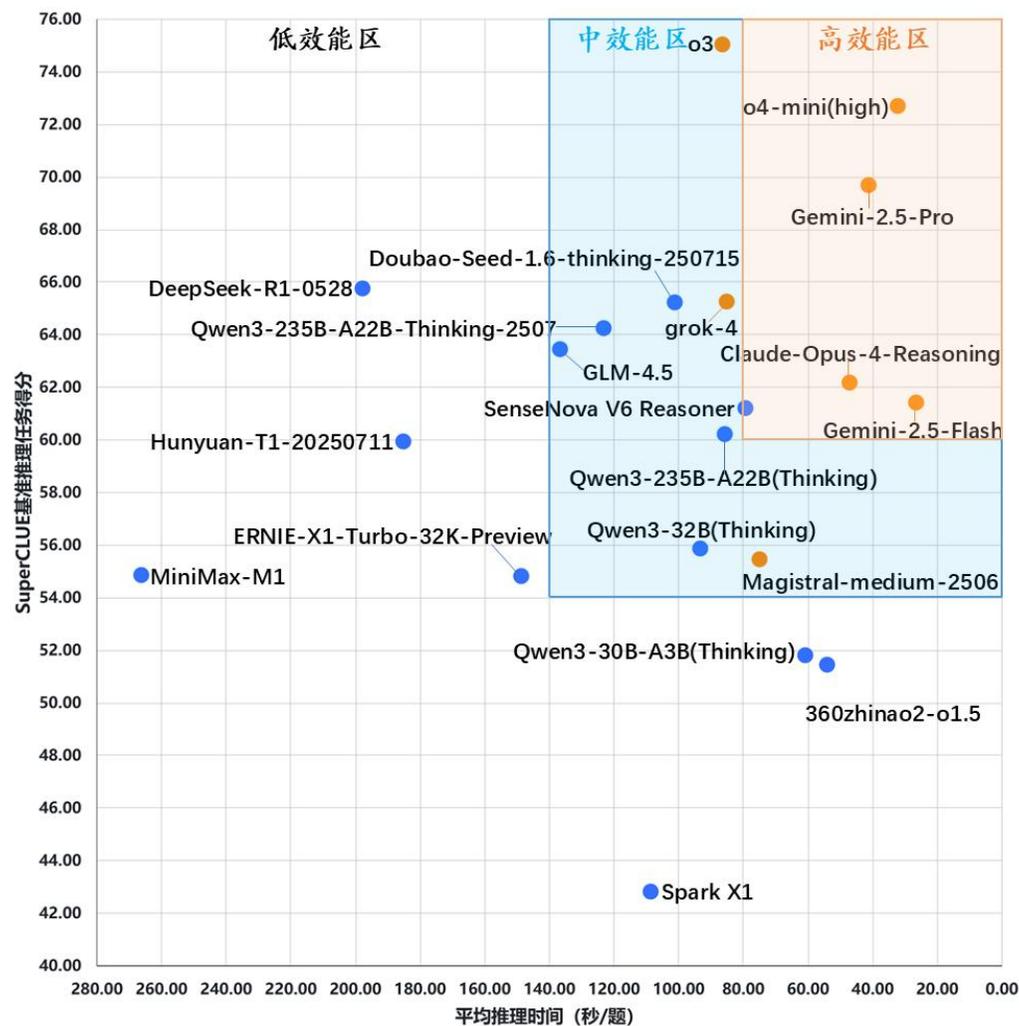
2. 四个象限的含义。它们代表大模型所处的不同阶段与定位, 其中【潜力探索者】代表模型正在探索阶段未来拥有较大潜力; 【技术领跑者】代表模型在基础技术方面具备领先性; 【实用主义者】代表模型在场景应用深度上具备领先性; 【卓越领导者】代表模型在基础和场景应用上处于领先地位, 引领国内大模型发展。

## 大模型性价比区间分布



数据来源: SuperCLUE, 2025年7月31日; 开源模型如Qwen3-235B-A22B(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7月的标准价格 (非优惠价格)。

## 推理模型推理效能区间分布



数据来源: SuperCLUE, 2025年7月31日; 模型推理速度选取7月测评中具有公开API的模型。平均推理时间为所有测评数据推理时间的平均值 (秒)。推理任务得分为推理任务总分: 数学推理、科学推理和代码生成的平均分。

# 第一部分

## 2025年上半年关键进展及趋势

1. 2025年上半年大模型关键进展
2. 2025年最值得关注的中文大模型及智能体全景图
3. 2025年国内外大模型差距
4. 近一年SuperCLUE通用基准测评开闭源模型最好成绩对比

◆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2年半有了实质性的突破。具体可分为：技术引爆与生态构建期、多模态与深度推理拓展期、智能体崛起与AGI探索期。

## SuperCLUE: AI大模型2025年上半年关键进展

关键进展

### 📦 技术引爆与生态构建期

- OpenAI发布ChatGPT及GPT-4，迅速点燃全球对大模型的关注并成为现象级应用；
- Meta开源Llama2，激活开发者生态，降低技术门槛，推动全球长尾创新；
- **中国首批大模型集中亮相**。百度文心一言1.0、阿里通义千问、讯飞星火V1.0、360智脑、ChatGLM系列等快速响应，标志着中国进入核心竞争梯队；
- **中国开源模型爆发**。百川Baichuan-7B、智谱ChatGLM2、通义千问Qwen等形成“模型矩阵”，加速技术民主化。

### 🎯 多模态与深度推理拓展期

- OpenAI发布Sora，实现高质量时序连贯视频生成，引发全球视频AIGC创业潮；
- OpenAI o1系列引入“慢思考”机制，AI大模型的发展重心进一步深化，开始攻克更复杂的推理和逻辑思考难题；
- **国内多模态领域**快速跟进与创新，并在部分领域领先海外。可灵AI、Vidu、Pixverse、海螺视频等视频生成模型陆续发布，并在海外取得较大的应用进展；
- **国内推理模型集中涌现**。k0-math、DeepSeek-R1-Lite、QwQ-32B-Preview、GLM-Zero-Preview等，力求在这些高价值场景取得突破。

### 🤖 智能体崛起与AGI探索期

- 顶级大模型能力持续突破，**性能与效率并进**。o3、o4系列、Gemini 2.5系列、Claude 4系列、Grok 4等顶级模型发布，进一步提升了推理的深度、准确性和可靠性，模型的工具调用能力也提升明显；
- Monica.im发布全球首款通用型AI Agent: **Manus**爆火出圈，成为 AI 领域的重大突破；
- 国内**AI智能体**潜力被深度挖掘，大量AI Agent产品涌现。比如AutoGLM、OpenManus、扣子空间、纳米AI超级搜索智能体、心流AI助手、天工超级智能体、MiniMax Agent、心响等；
- 国内模型加速迭代与开源，国内外差距动态变化。**DeepSeek-R1**以开源、高性价比以及媲美同期顶级闭源模型的特点引爆全球；**Kimi-k2**、**MiniMax-M1**、**GLM-4.5**等厂商继续开源新模型，缩小国内外差距。

2022.12

2023.12

2024.12

2025.07

时间

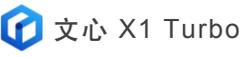
# SuperCLUE: 2025年最值得关注的中文大模型及智能体全景图

## 文本

**通用闭源**  字节豆包  通义千问  文心一言  腾讯混元  商汤日日新  盘古大模型  Kimi.ai  ZTE中兴  讯飞星火  MiLM

 智谱·AI  360智脑  MINIMAX  DeepSeek  天工  阶跃星辰  云从科技  云知声  山海

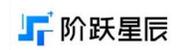
**通用开源**  Qwen  DeepSeek  GLM-4.5  面壁小钢炮  Hunyuan-Large  MiniMax-01  书生·浦语  ERNIE-4.5  K2

**推理**  Qwen3  DeepSeek-R1  K1.5长思考  Step R-mini  文心 X1 Turbo  GLM-4.5  Hunyuan T1  MiniMax-M1

## 多模态

**实时交互**  日日新  智谱清言  海螺AI  豆包  文小言  星火极速  Kimi

**文/图生视频**  ByteDance  海螺AI  可灵AI  Vidu  清影  通义万相  PixVerse

**视觉理解**  腾讯混元  Doubao-vision  QvQ-Max  SenseChat-Vision  阶跃星辰  GLM-4v

**文生图**  即梦AI  快手可图  文心一格  通义万相  CogView  讯飞星火

**语音合成/声音复刻**

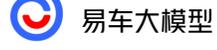
 Doubao-语音合成  百度TTS

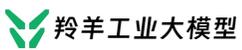
 讯飞语音合成  CosyVoice

 Fish Audio  Speech-02

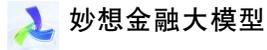
## 行业

**医疗**  百度灵医  医联MedGPT  讯飞晓医

**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

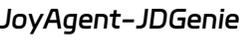
**教育**  MathGPT  子曰  作业帮

**金融**  蚂蚁金融大模型  妙想金融大模型  轩辕大模型

**法律**  Chat Law  CHINESE LAW 元典智库  得理

## 智能体

**通用闭源**  AutoGLM 沉思  扣子空间  MINIMAX  天工  心流

**通用开源**  OWL  OpenManus  Genie  JoyAgent-JDGenie

**深度研究**  智谱清言  纳米AI搜索  秘塔AI搜索  夸克

 DeepResearch  Kimi-Researcher  豆包 深入研究  Qwen 深入研究

总体趋势上，国内外第一梯队大模型在中文领域的通用能力差距正在缩小。2023年5月至今，国内外大模型能力持续发展。其中GPT系列模型为代表的海外最好模型经过了从GPT3.5、GPT4、GPT4-Turbo、GPT4o到o1系列、o3系列以及o4系列多个版本的迭代升级。国内模型也经历了波澜壮阔的25个月的迭代周期。本次测评我们可以发现国内外第一梯队的大模型从2023年5月通用基准测评10.42%的差距缩小到7.78%。

SuperCLUE17次大模型通用基准测评国内外顶尖模型成绩汇总

测评日期	2023年5月	2023年6月	2023年7月	2023年8月	2023年9月	2023年10月	2023年11月	2023年12月	2024年2月	2024年4月	2024年6月	2024年8月	2024年10月	2024年12月	2025年3月	2025年5月	2025年7月
国外TOP1	76.67	78.76	70.89	81.03	83.20	87.08	89.79	90.63	92.71	79.13	81.00	79.67	75.85	80.40	76.01	70.51	73.78
国内TOP1	53.58	63.53	62.00	60.02	62.75	70.74	74.02	79.02	87.75	73.32	77.00	78.64	69.64	68.30	70.34	63.16	68.04

SuperCLUE基准：过去27个月国内外TOP大模型对比趋势

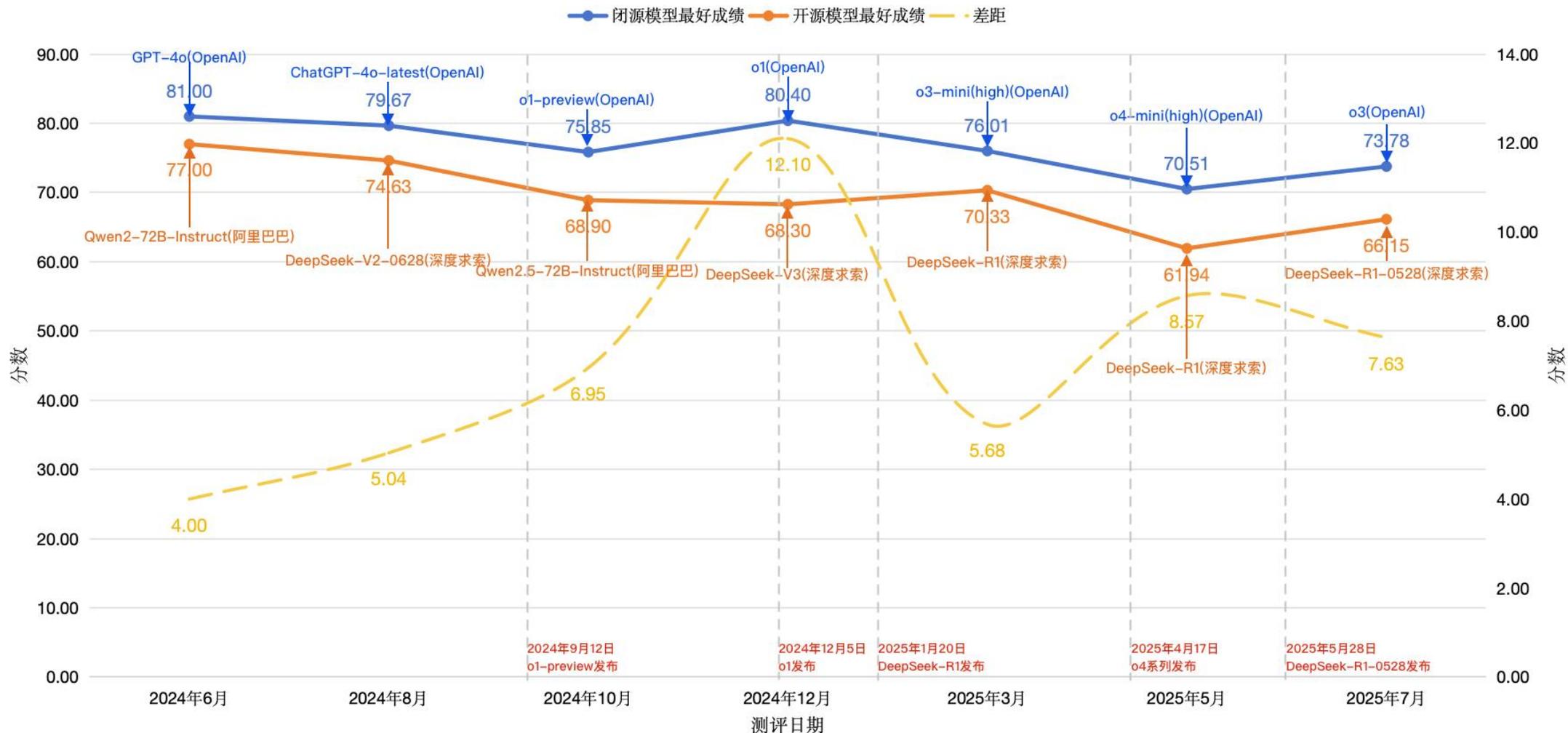
■ 差距分数 — 差距比例 - - - 线性(差距分数)



# 近一年SuperCLUE通用基准测评开闭源模型最好成绩对比

在近一年的七次SuperCLUE通用基准测评中，闭源模型与开源模型的TOP1性能差距呈现出动态变化趋势。2024年下半年，随着o1系列闭源模型的推出，两者性能差距呈现扩大态势。这一趋势在2025年初迎来转折，DeepSeek-R1开源模型的发布显著缩小了开闭源模型之间的性能差距。接着OpenAI相继推出o3、o4系列闭源模型，将开闭源模型之间的差距再次拉大。DeepSeek-R1-0528的发布又将开闭源模型之间的差距缩小。

### 近一年SuperCLUE通用基准七次测评开源和闭源模型最好成绩对比



# 第二部分

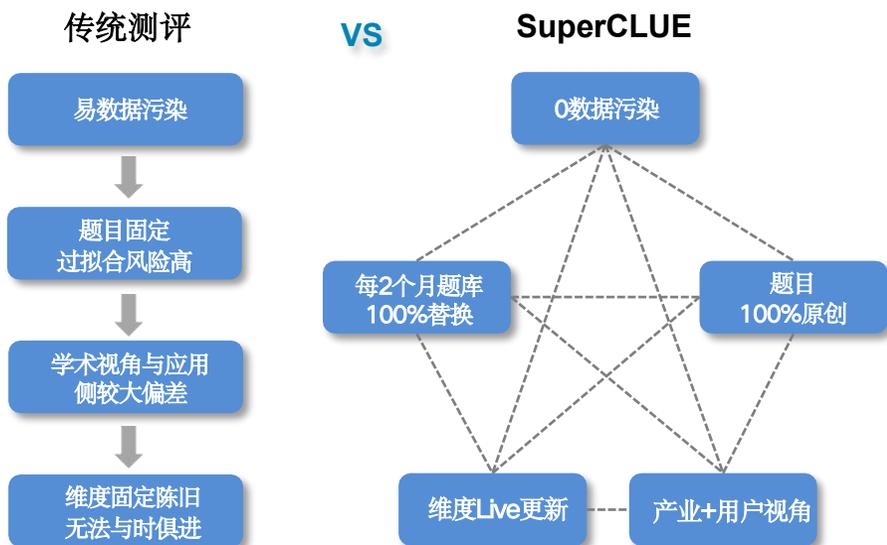
## 七月通用测评介绍

1. SuperCLUE基准介绍
2. SuperCLUE大模型综合测评体系
3. SuperCLUE通用测评基准数据集及评价方式
4. 各维度测评说明及示例
5. 测评模型列表

**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。



## SuperCLUE与传统测评的区别

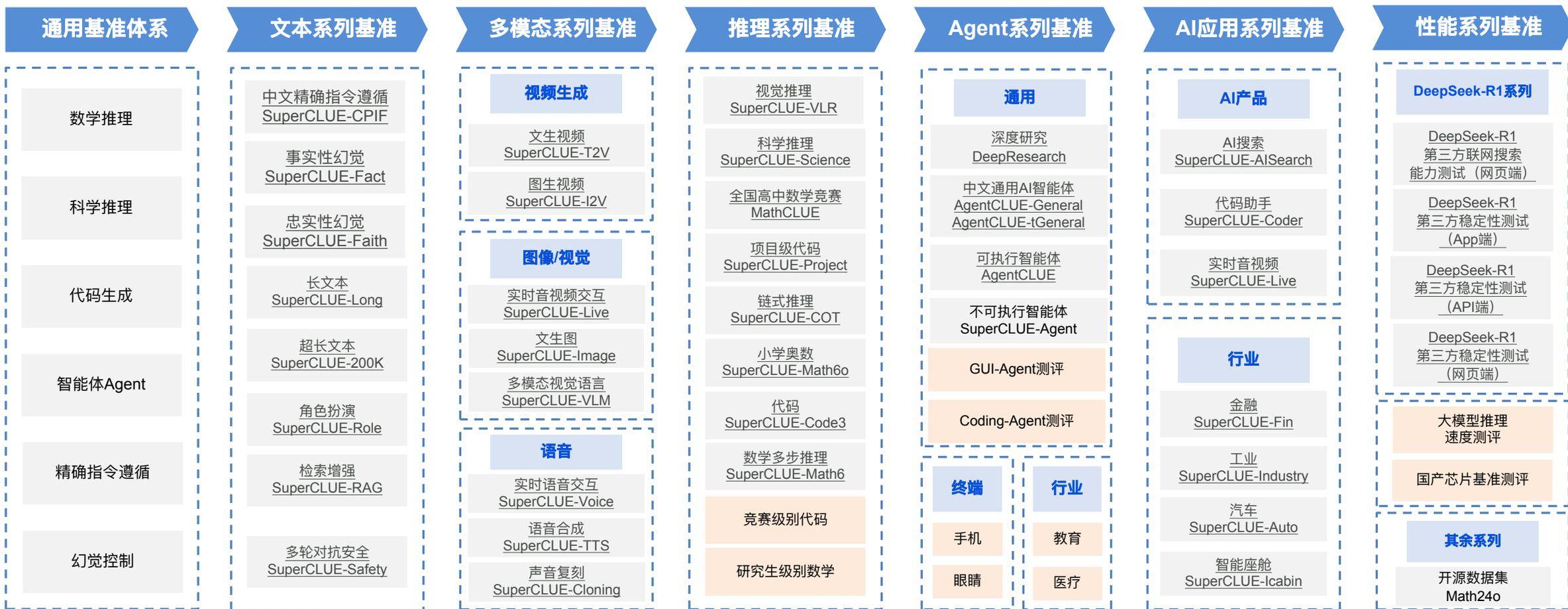


## SuperCLUE 三大特征

- 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。

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。所有基准将在本报告的第四部分进行详细介绍。

## SuperCLUE大模型综合测评基准框架



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。

本次2025年上半年度报告聚焦通用能力测评，由**六大维度**构成。题目均为**原创新题**，总量为**1288**道简答题。

## SuperCLUE-7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：**基于人工校验参考答案的评估（0-1得分）

### 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：**基于人工校验参考答案的评估（0-1得分）

### 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、系统状态比对的评估（0-1得分）

### 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：**基于规则脚本的评估（0-1得分）

### 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、对每个句子进行二元判定的评估

## SuperCLUE—数学推理评测数据集

### 数学推理

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。

## 测评方法

### 评测流程：

我们采用模型评价的方式来评估大模型在【数学推理】任务上的表现，使用高级模型（如gpt4o）判断其回答中的最终答案和参考答案是否相符，以确认大模型在一道问题上的正确率（正确或错误，即0或1）。

以【代数学】任务为例，假设在一个具体的题目中，模型A的最终答案与参考答案等价，则该题目的具体得分为1分；反之，假设在另一个具体的题目中，模型A的最终答案与参考答案不等价，则该题目的具体得分为0分。

最后，将所有题目的得分进行归一化处理，以确保满分为100分。这样最终的得分可以清晰地反映模型在【数学推理】任务中的表现。这一方法实现了完全的客观评估性。

## 示例—数学推理

### 问题：

设整数序列  $\{a_n\}_{n \geq 1}$  满足  $a_1 = 2$  且  $\prod_{i=1}^n a_i = a_{n+1} - 1$  对所有正整数  $n$  成立。计算无穷级数

$$S = \sum_{k=1}^{\infty} \left( k \sum_{n=1}^{\infty} \frac{a_{n+1}^k - 1}{(a_{n+1}(a_{n+1}-1))^k} \right) \text{ 的值。}$$

### 模型回答：1. 分析序列 $\{a_n\}$ 的性质

根据题目所给的递推关系：

$$\prod_{i=1}^n a_i = a_{n+1} - 1 \quad (1)$$

$$a_1 = 2$$

○ 当  $n = 1$  时， $a_1 = a_2 - 1$ 。因为  $a_1 = 2$ ，所以  $a_2 = 3$ 。

○ 当  $n = 2$  时， $a_1 a_2 = a_3 - 1$ 。代入  $a_1 = 2, a_2 = 3$ ，得到  $6 = a_3 - 1$ ，所以  $a_3 = 7$ 。

<篇幅有限省略>

最终答案是  $\boxed{2}$ 。

### 评价打分：

---- 逐项打分开始 ----

问题及改进的反馈意见：

模型的回答清晰地分析了问题中的递推关系，正确推导了序列  $\{a_n\}$  的性质，并且在计算无穷级数  $S$  时，合理地使用了伸缩求和和幂级数求和公式。最终答案与参考答案一致，逻辑清晰，步骤完整。然而，模型的回答中有些地方可以进一步简化，例如在推导  $S_k$  时的分解步骤稍显冗长。

打分与评价开始：

1. 基于标准 1 的 <结论一致性>：

模型的回答最终得出的答案是  $\boxed{2}$ ，与参考答案一致，且推导过程正确无误，符合题目要求。基于标准 1 的得分为 1 分。

综合以上，AI助手的回答的综合得分（平均分）为  $\boxed{[1.00]}$ 。

## SuperCLUE—科学推理评测数据集

### 科学推理

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。

## 测评方法

### 评测流程：

我们采用模型评价的方式来评估大模型在【科学推理】任务上的表现，使用高级模型（如gpt4o）判断其回答中的最终答案和参考答案是否相符，以确认大模型在一道问题上的正确率（正确或错误，即0或1）。

以【物理】任务为例，假设在一个具体的题目中，模型A的最终答案与参考答案等价，则该题目的具体得分为1分；反之，假设在另一个具体的题目中，模型A的最终答案与参考答案不等价，则该题目的具体得分为0分。

最后，将所有题目的得分进行归一化处理，以确保满分为100分。这样最终的得分可以清晰地反映模型在【科学推理】任务中的表现。这一方法实现了完全的客观评估性。

## 示例—科学推理

**问题：**一根质量为 $m$ 、长度为 $L$ 、总电阻为 $r$ 的均匀导体细杆，可绕其一端点 $O$ 在水平面内无摩擦地自由转动。整个装置处于磁感应强度大小为 $B$ 的匀强恒定磁场中，磁场方向竖直向下，与转动平面垂直。一个电容为 $C$ 的电容器，其一极板与转动轴心 $O$ 相连，另一极板与一个以 $O$ 为圆心、半径为 $L$ 的固定圆形导电轨道相连。细杆的自由端始终与该轨道保持良好电接触。初始时刻，电容器不带电，细杆被赋予一个大小为 $\omega_0$ 的初角速度。

求从运动开始直至系统达到稳定状态的整个过程中，导体杆中产生的总焦耳热。

**模型回答：** 1. 分析初末状态：

- **初始状态 (t=0):** 杆的角速度为  $\omega_0$ ，电容器不带电 (电荷量  $Q_0 = 0$ )。

系统的总能量为杆的初始动能： $E_{初} = E_{k0} = \frac{1}{2} I \omega_0^2$ ，其中  $I$  是杆绕端点转动的转动惯量。

对于质量为  $m$ 、长度为  $L$  的均匀细杆，其绕一端转动的转动惯量为  $I = \frac{1}{3} m L^2$ 。

所以， $E_{初} = \frac{1}{2} \cdot \frac{1}{3} m L^2 \cdot \omega_0^2 = \frac{1}{6} m L^2 \omega_0^2$ 。

<篇幅有限省略>

导体杆中产生的总焦耳热为  $\frac{m B^2 L^4 C \omega_0^2}{8m + 6B^2 L^2 C}$ 。

**评价打分：** ----逐项打分开始----

问题及改进的反馈意见：

AI 助手的回答逻辑清晰，步骤详尽，最终结论与参考答案一致。但回答中部分推导过程显得冗长，可以适当简化以提高可读性。

打分与评价开始：

1. 基于标准 1 的<结论一致性>；  
AI 助手的最终结论与参考答案完全一致。基于标准 1 的得分为 1 分。

综合以上，AI助手的回答的综合得分（平均分）为[[1.00]]。

## SuperCLUE—代码生成评测数据集

### 代码生成

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。

### 测评方法

#### 一、代码生成评测流程：

整体测评流程包括：**1.获取模型答案；2.提取功能函数；3.测试功能函数；4.计算模型得分。**

##### 1.获取模型答案：

使用特定的prompt要求模型按照特定格式回答，以方便后续提取。对于一个代码补全问题，每个模型获取一次答案。

##### 2.提取功能函数：

获取到模型回复之后，结合prompt要求和模型回答设定规则提取函数代码。

##### 3.测试功能函数：

对于一个代码补全问题，将提取得到的代码和该问题对应的N个测试用例组装成N个可独立运行的单元测试程序，在沙箱环境运行（沙箱环境python版本设置为3.10）。

##### 4.计算模型得分：

对于一个代码补全问题和1个测试用例，构成一个单元测试，通过测试得1分。最终计算模型在所有单元测试中得分平均分\*100作为最终的得分。

#### 二、Web应用生成评测流程：

通过模拟用户交互的功能测试进行0/1评分。

## 示例—代码生成

### 问题：

```
import collections
from typing import List, Dict, Tuple
```

```
def task_func(movies: Dict[str, Tuple[int, str]],
rooms: int, genre_conflicts: Dict[str, List[str]],
operating_hours: Tuple[int, int]) -> int:
    """
```

一家电影院需要为一个电影节排期，目标是在有限的运营时间内，放映尽可能多的不同电影。

影院有多间放映厅，排期时必须遵守以下规则：

- 每部电影有特定的时长和类型。

<篇幅有限省略>

`operating_hours (Tuple[int, int])`: 一个元组，表示影院的开始和结束运营时间(24小时制)，例如(9, 23)表示从9:00到23:00。

Returns:

int: 最多可以放映的电影数量。

"""

### 测试功能函数：

```
def check(candidate):
```

```
    assert candidate({"X": (100, "A"), "Y": (100, "B"), "Z": (100, "C"), "W": (100, "D")}, 2, {"A": ["B"], "B": ["A", "C"], "C": ["B"]}, (10, 14),) == 4
```

**计算模型得分：**测试通过，得1分。

### 模型回答：

```
import collections
from typing import List, Dict, Tuple
```

```
def task_func(movies: Dict[str, Tuple[int, str]], rooms:
int, genre_conflicts: Dict[str, List[str]], operating_hours:
Tuple[int, int]) -> int:
    # 计算总运营时间(分钟)
```

```
    total_minutes = (operating_hours[1] -
operating_hours[0]) * 60
```

<篇幅有限省略>

```
    if not conflict:
```

```
        # 可以安排这部电影
```

```
    room_schedules[room_idx].append((start_time,
end_time, genre))
```

```
        scheduled_count += 1
```

```
        scheduled = True
```

```
        break
```

```
    if scheduled:
```

```
        break
```

```
    return scheduled_count
```

## SuperCLUE—智能体Agent数据集

### 智能体Agent

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。

### 测评方法

#### 评测流程：

每一题包含N轮对话，每一轮对话为一个记分单元，答对记1分。

我们将模型回复的函数调用进行实际执行后判定是否回复正确。评价维度包括：函数调用执行结果、执行后的系统状态。

对于每一轮模型回复的评价方法：

1. 获取模型答案之后，从中提取可执行的函数调用。
2. 执行提取的函数调用得到执行结果集1，以及执行后的系统状态1；
3. 执行人类标注的ground truth(函数调用)得到执行结果集2，系统状态2；
4. 将执行结果集1与执行结果集2进行比较，系统状态1和系统状态2进行比较；

如果执行结果集2是执行结果集1的子集，系统状态1和系统状态2相同则认为模型的本轮回复正确，得1分。

最后，计算答对的记分单元的比例，转化为百分制得到模型的总分。即：

模型得分= (答对的记分单元数量/记分单元总数量) \*100

## 示例—智能体Agent

**描述：**本次Agent测评共包含12个API，本题用到2个API。

股票交易系统：该工具属于股票交易系统，该系统允许用户交易股票，管理其帐户并查看库存信息。

智能视频播放系统：该工具属于智能视频播放系统，该系统允许用户对短视频及其内容进行一系列操作。

**prompt：**你是一个组合函数完成任务的专家。你将会收到一个问题和一些可以调用的函数及其描述。基于问题，你需要调用一个或者多个函数来实现问题要求。

对于问题中的计算需求，你必须调用提供的函数进行计算，否则视为错误。

如果使用你收到的函数无法完成问题要求，请指出。如果你收到的问题缺乏调用函数所必要的参数<篇幅有限省略>

下面提供了一个列表，列表中是一些你可以调用的函数以及函数的功能和参数描述，使用json格式<篇幅有限省略>

**【问题】：**我想登录我的股票账户，用户名是“investor\_jane”，密码是“securepass987”，然后帮我查询一下“AAPL”这只股票的当前价格。另外，也帮我登录视频应用，我的用户名是“lihua”，密码是“password123”，然后搜索一下“美食制作教程”的视频id是多少。

**【模型回复1】：****【函数调用开始】**```[trading\_login(username="investor\_jane", password="securepass987"), log\_in(username="lihua", password="password123"), get\_stock\_info(symbol="AAPL"), find\_video\_by\_name(video\_name="美食制作教程")]```\b【函数调用结束】

**【模型得到的反馈1】：**下面是根据你回复的函数调用执行得到的结果,你可以据此判断是否完成了用户的需求。如果已经完成,你可以停止回复函数调用,根据执行结果回答用户的问题。如果没有完成任务,你可以决定给出新的函数调用。

下面是函数调用结果：

第1个函数调用：trading\_login(username='investor\_jane', password='securepass987'), 执行返回结果为：{"status": "Logged in successfully"}

第2个函数调用：log\_in(username='lihua', password='password123'), 执行返回结果为：{"status": "登录成功"}

第3个函数调用：get\_stock\_info(symbol='AAPL'), 执行返回结果为：{"price": 228.5, "percent\_change": 0.17, "volume": 2.552, "MA(5)": 227.11, "MA(20)": 227.09}

第4个函数调用：find\_video\_by\_name(video\_name='美食制作教程'), 执行返回结果为：{"videos": [{"id": "v002", "title": "美食制作教程", "author": "chef\_master", "likes": 8510, "description": "Howtomakedeliciousfood"}]}

注意：一旦你决定进行函数调用,你的回复中只能包括函数调用,并且你回复的所有内容都不要换行。你的函数调用(function calls)必须严格遵循以下格式给出：“**【函数调用开始】**```[func\_name1(params\_name1=params\_value1, params\_name2=params\_value2...), func\_name2(params)]```\b【函数调用结束】”，不应该包括任何其他文本

**【模型回复2】：**所有任务已完成,无需调用其他函数。“AAPL”这支股票的当前价格是228.5,“美食制作教程”的视频id是v002。

本题共1轮问题,解决问题需要调用4次工具,模型的回答中4次工具调用均正确,则本轮回答正确,记1分。

## SuperCLUE—精确指令遵循数据集

### 精确指令遵循

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。

### 测评方法

#### 评测流程：

我们完成了一个针对不同的语言模型的（文本输出）精确指令遵循表现的测试，根据设定的评估标准进行评估。评估的结果为布尔值（True 或 False）。

以【写一篇关于可再生能源对减少碳排放的作用的短文。要求文章不得使用“二氧化碳”这个词，字数不少于10个字，以JSON格式返回】任务为例：

设定的评价标准包括：遵循关键词限制、符合字数要求、输出格式正确。模型只有在命中所有指令的条件下会得到True的返回结果。

例如：

模型输出结果1：{"response": "可再生能源在减少碳排放和减轻气候变化方面起着关键作用，未来应加快推广。"}  
符合所有要求，返回True。

模型输出结果2：{"response": "使用可再生能源可以减少二氧化碳排放。"}  
违背了关键词限制，返回False。

最后，我们将所有题目的评估结果进行汇总，根据模型完全命中的题目数量计算模型的最终命中率，命中率以百分比为单位。具体来说，命中率计算公式为：

**命中率 = (模型完全命中的题目数量 ÷ 总题目数量) \* 100%**

## 示例—精确指令遵循

**问题：**随着城市化进程加快，有限的公共空间资源面临日益紧张的挑战。近年来，许多社区开始尝试在空置用地和闲置场所中建设共享花园，以提升居民参与度、改善小气候、促进邻里交流。共享花园通常由居民自发组织，通过合理的功能分区、生态种植和可持续管理，实现绿色空间与社区活动的有机结合。在设计阶段，需要考虑土壤改良、灌溉系统、道路及休憩区布局等关键要素，同时关注不同年龄层和兴趣群体的需求。

请根据以上背景信息，撰写一份城市社区共享花园的设计与维护方案。要求文章以“经过深入分析，”开头。请在回复中详细解释并列其中三种主要部分，使用“&功能分区&”、“&生态益处&”和“&可行性评估&”作为主要部分的占位符。请确保文章恰好由11个句子（句子用中文句号或问号或感叹号分隔）组成。其中第7个句子的第4个汉字必须是“融”字。

**模型回答：**经过深入分析，社区共享花园的建设应结合现有空地、居民需求和管理模式进行统筹规划。

&功能分区&部分包括儿童游乐区、社区农场区、休憩漫步区和公共活动区，并需考虑无障碍通道与路面材质的选择。

&生态益处&主要体现在改善小气候、增加生物多样性、雨洪调蓄以及邻里教育功能。

&可行性评估&需要从成本投入、资源循环利用、技术支持和社区参与度等方面展开量化分析。

设计阶段应关注土壤改良方案、水源灌溉系统、植被配置和节水设施等技术细节。

同时要预留足够的通行与观赏空间，以满足不同年龄层和兴趣群体的需求。

本维护融入定期巡检、轮值制度、智能监测和社区反馈机制的有机结合。

植物种植计划可采用轮作与间作方法，减少病虫害并维持土壤肥力。

成本控制方案包括雨水收集利用、堆肥再利用以及本土苗木采购等措施。

项目实施后需定期评估社会效益与环境影响，并根据反馈及时调整管理策略。

举办季节性主题活动与培训能够增强居民归属感，确保花园长期可持续运营。

#### 评价打分：

本题存在四个精确指令考察：1）以特定文本开头；2）占位符；3）句子数量；4）汉字定位。

模型的回答符合要求，得一分。

## SuperCLUE—幻觉控制数据集

### 幻觉控制

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。

## 测评方法

### 评测流程：

我们采用高级AI模型（例如GPT-4o）三阶段自动化评估方法来评估不同任务的表现，根据设定的评价标准进行打分。

以下是具体评估流程介绍：

- 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；
- 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；
- 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。

比如：模型对于某个题目的回复存在10个句子，其中3个句子存在幻觉，这3个句子的得分都为0分，其他7个句子不存在幻觉，这7个句子的得分都为1分。那么模型在此题上的得分为： $7/10=0.70$ 分，汇总模型在所有题目上的得分，再除以总题数即为模型最终的得分。

## 示例—幻觉控制

**问题：**请总结以下文本，形成摘要：

外人眼中的茶饮生意，是窗明几净的，是香甜的，美好的，带着闲情逸致的。而真正的茶饮从业者，往往忙到没有时间坐下来休息，点单、调饮、报货、管理、运营、盘点，生意的奥秘蕴藏在无数细节与琢磨之中。<篇幅有限省略>刚开业那三天，她生意还不错，因为做了买一送一的活动，每天的流水有2000多元，当时公司派了两个人支援，加上她和丈夫两个人，四个人忙得饭都顾不上吃，等到买一送一活动结束后，销量骤降，她记得那天天快黑了，封口机上显示的数字是「54」，也就意味着她只卖出了54杯饮品，流水大约300多元。<篇幅有限省略>

**模型回答：**【茶饮行业背后的奋斗者：古茗加盟商群像观察】

（一）行业现实与理想落差

茶饮行业表面光鲜，实则充满艰辛。从业者需面对点单、调饮、运营等繁琐工作，真正的成功源于年复一年的细节打磨。本文以古茗为样本，通过多位加盟商的创业故事，揭示新式茶饮行业的生存法则。<篇幅有限省略>

（三）创业密码：长期主义与极致执行

1. **\*\*选址智慧\*\***：周勇初期误选厕所旁店铺，经总部指导后在无锡利港镇开首家店，经历**月流水300元的至暗时刻**；海瑞雪通过学生卡+明星海报精准获客。<篇幅有限省略>

**评价打分：**

打分与评价开始：

{句子：【茶饮行业背后的奋斗者：古茗加盟商群像观察】

（一）行业现实与理想落差

茶饮行业表面光鲜，实则充满艰辛。；评价：AI助手的回答片段准确地捕捉到了原文对于茶饮行业外部观感与实际存在落差描述，内容与原文高度一致。+[[1]]

{句子：从业者需面对点单、调饮、运营等繁琐工作，真正的成功源于年复一年的细节打磨。；评价：AI助手提供的这个回答片段准确地概括了文本中关于茶饮从业者工作内容艰辛和成功需要长期努力打磨细节的描述。片段内容与原文一致，信息准确。+[[1]]

<篇幅有限省略>

{句子：（三）创业密码：长期主义与极致执行

1. **\*\*选址智慧\*\***：周勇初期误选厕所旁店铺，经总部指导后在无锡利港镇开首家店，经历**月流水300元的至暗时刻**；海瑞雪通过学生卡+明星海报精准获客。；评价：AI助手的回答片段对周勇和海瑞雪的早期创业经历进行了总结，提到了周勇选址的考虑和海瑞雪的获客方式。然而，在描述周勇的流水时出现了事实性错误。+[[0]]<篇幅有限省略>

综合以上，AI助手的回答共有14个句子，其中4个句子存在幻觉，10个句子不存在幻觉，因此本题的综合得分（平均分）为[[10/14=0.71]]。

# 测评模型列表

本次测评数据选取了**SuperCLUE-7月**测评结果，模型选取了国内外有代表性的**45个大模型在7月份的版本**。

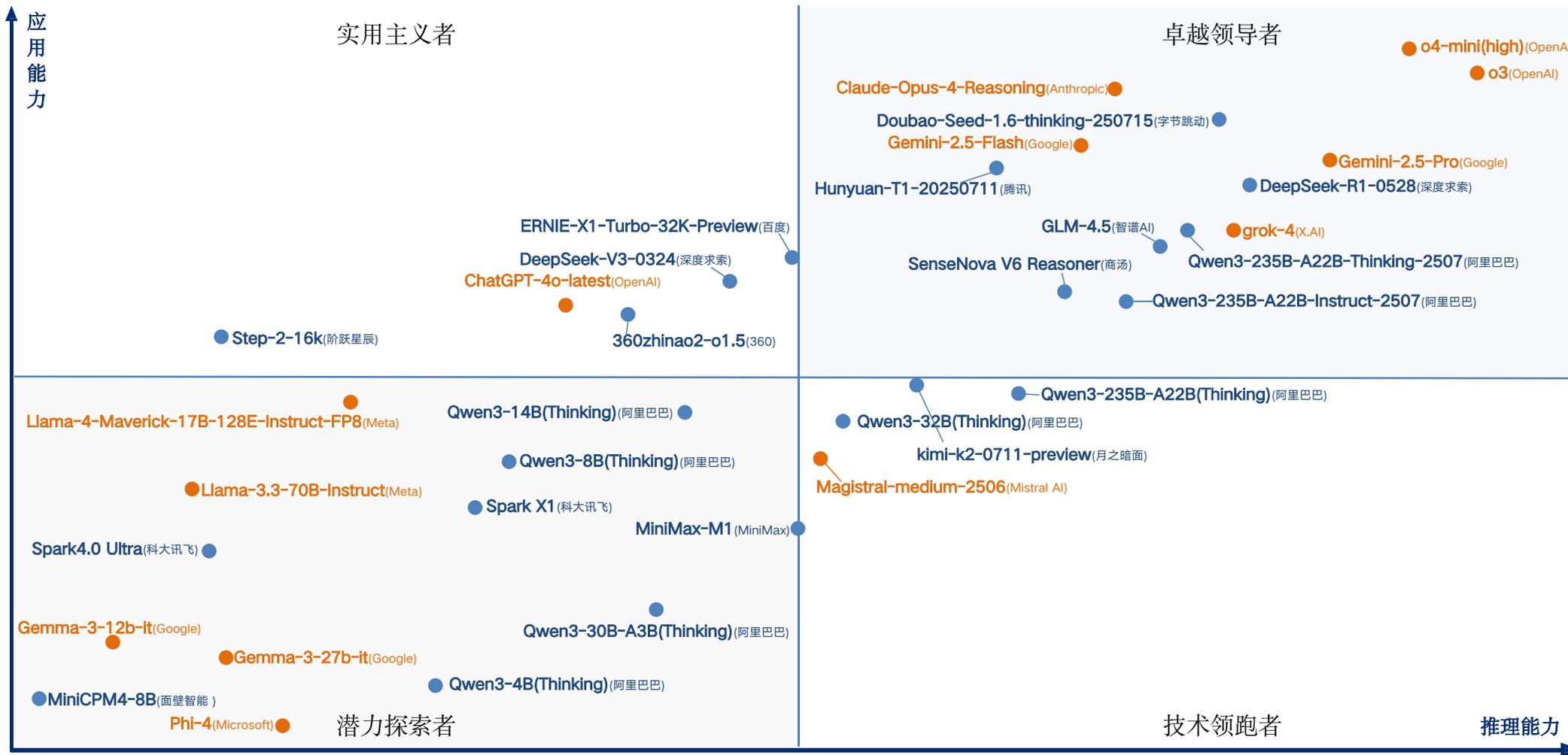
模型	机构	简介	模型	机构	简介
1.Qwen3-235B-A22B(Thinking)	阿里巴巴	官方发布的MoE推理模型，使用阿里云公开的API: qwen3-235b-a22b。	24.Gemini-2.5-Flash	Google	Gemini-2.5-Flash的正式版本，使用官方API: gemini-2.5-flash。
2.Qwen3-30B-A3B(Thinking)	阿里巴巴	官方发布的MoE推理模型，使用阿里云公开的API: qwen3-30b-a3b。	25.Gemma-3-27b-it	Google	Gemma3开源系列27B模型，使用官方API: gemma-3-27b-it。
3.Qwen3-32B(Thinking)	阿里巴巴	官方发布的Dense推理模型，使用阿里云公开的API: qwen3-32b。	26.Gemma-3-12b-it	Google	Gemma3开源系列12B模型，使用官方API: gemma-3-12b-it。
4.Qwen3-14B(Thinking)	阿里巴巴	官方发布的Dense推理模型，使用阿里云公开的API: qwen3-14b。	27.Gemma-3-4b-it	Google	Gemma3开源系列4B模型，使用官方API: gemma-3-4b-it。
5.Qwen3-8B(Thinking)	阿里巴巴	官方发布的Dense推理模型，使用阿里云公开的API: qwen3-8b。	28.Gemma-3-1b-it	Google	Gemma3开源系列1B模型，使用官方API: gemma-3-1b-it。
6.Qwen3-4B(Thinking)	阿里巴巴	官方发布的Dense推理模型，使用阿里云公开的API: qwen3-4b。	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。
7.Qwen3-1.7B(Thinking)	阿里巴巴	官方发布的Dense推理模型，使用阿里云公开的API: qwen3-1.7b。	30.Llama-3.3-70B-Instruct	Meta	使用together.ai的接口: meta-llama/Llama-3.3-70B-Instruct-Turbo。
8.Qwen3-0.6B(Thinking)	阿里巴巴	官方发布的Dense推理模型，使用阿里云公开的API: qwen3-0.6b。	31.Llama-3.1-8B-Instruct	Meta	使用together.ai的接口: meta-llama/Meta-Llama-3.1-8B-Instruct-Turbo。
9.Step-2-16k	阶跃星辰	官方发布的基础模型，使用官方API: step-2-16k。	32.ChatGPT-4o-latest	OpenAI	与ChatGPT上的GPT-4o同版本，对应OpenAI官方的API名称: chatgpt-4o-latest。
10.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。	33.o4-mini(high)	OpenAI	使用方式为AZURE OpenAI Service的API接口，reasoning_effort参数设置为: high。
11.Gemma-3n-E4B-it	Google	官方开源的模型，使用官方API: Gemma_3n_e4b_it。	34.o3	OpenAI	官方发布的推理模型o3，使用官方API: o3-2025-04-16。
12.Gemma-3n-E2B-it	Google	官方开源的模型，使用官方API: Gemma_3n_e2b_it。	35.360zhinao2-o1.5	360	360智脑最强大的推理模型，使用官方API: 360zhinao2-o1.5。
13.GLM-Z1-9B-0414	智谱AI	官方开源的模型，对应huggingface仓库名称: THUDM/GLM-Z1-9B-0414。	36.grok-4	X.AI	官方在2025年7月9日推出的最新推理模型，使用官方API: grok-4-0709。
14.DeepSeek-V3-0324	深度求索	官方发布的新版本V3模型，使用官方API: deepseek-chat。	37.ERNIE-X1-Turbo-32K-Preview	百度	官方发布的深度思考模型，使用官方API: ernie-x1-turbo-32k-preview。
15.DeepSeek-R1-0528	深度求索	官方发布的新版本开源推理模型，使用官方API: deepseek-reasoner。	38.Hunyuan-T1-20250711	腾讯	官方发布的深度思考模型，使用官方API: hunyuan-t1-20250711。
16.Qwen3-235B-A22B-Thinking-2507	阿里巴巴	官方开源的最新思考模型，使用官方API: qwen3-235b-a22b-thinking-2507。	39.Qwen3-235B-A22B-Instruct-2507	阿里巴巴	官方开源的最新非思考模型，使用官方API: qwen3-235b-a22b-instruct-2507。
17.Spark X1	科大讯飞	官方发布的推理模型，使用官方API: Spark X1。	40.Yi-1.5-9B-Chat-16K	零一万物	官方开源版本，对应huggingface仓库名称: 01-ai/Yi-1.5-9B-Chat-16K。
18.Spark4.0 Ultra	科大讯飞	官方发布的基础模型，使用官方API: Spark4.0 Ultra。	41.Magistral-medium-2506	Mistral AI	官方开源版本，对应huggingface仓库名称: mistralai/Magistral-medium-2506。
19.Phi-4	Microsoft	官方开源的模型，使用OpenRouter的API: microsoft/phi-4。	42.Ministral-8B-latest	Mistral AI	官方开源版本，对应huggingface仓库名称: mistralai/Ministral-8B-latest。
20.kimi-k2-0711-preview	月之暗面	官方发布的最新开源基础模型，使用官方API: kimi-k2-0711-preview。	43.MiniCPM4-8B	面壁智能	官方开源版本。对应huggingface仓库名称: openbmb/MiniCPM3-4B。
21.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。	44.SenseNova V6 Reasoner	商汤	官方发布的融合模态推理模型，使用官方API: SenseNova-V6-Reasoner。
22.Claude-Opus-4-Reasoning	Anthropic	官方发布的混合推理模型，使用官方API: Claude-Opus-4-Reasoning。	45.GLM-4.5	智谱AI	官方发布的最新开源推理模型，使用官方API: glm-4.5。
23.Gemini-2.5-Pro	Google	Gemini-2.5-Pro的正式版本，使用官方API: gemini-2.5-pro。	/	/	/

# 第三部分

## 总体测评结果与分析

1. SuperCLUE模型象限 (202507)
2. SuperCLUE通用能力测评榜单
3. SuperCLUE性价比区间分布
4. SuperCLUE大模型综合效能区间分布
5. 代表性模型分析
6. 国内大模型成熟度-SC成熟度指数
7. 评测与人类一致性验证

## SuperCLUE模型象限 (202507)



来源: SuperCLUE, 2025年7月31日;

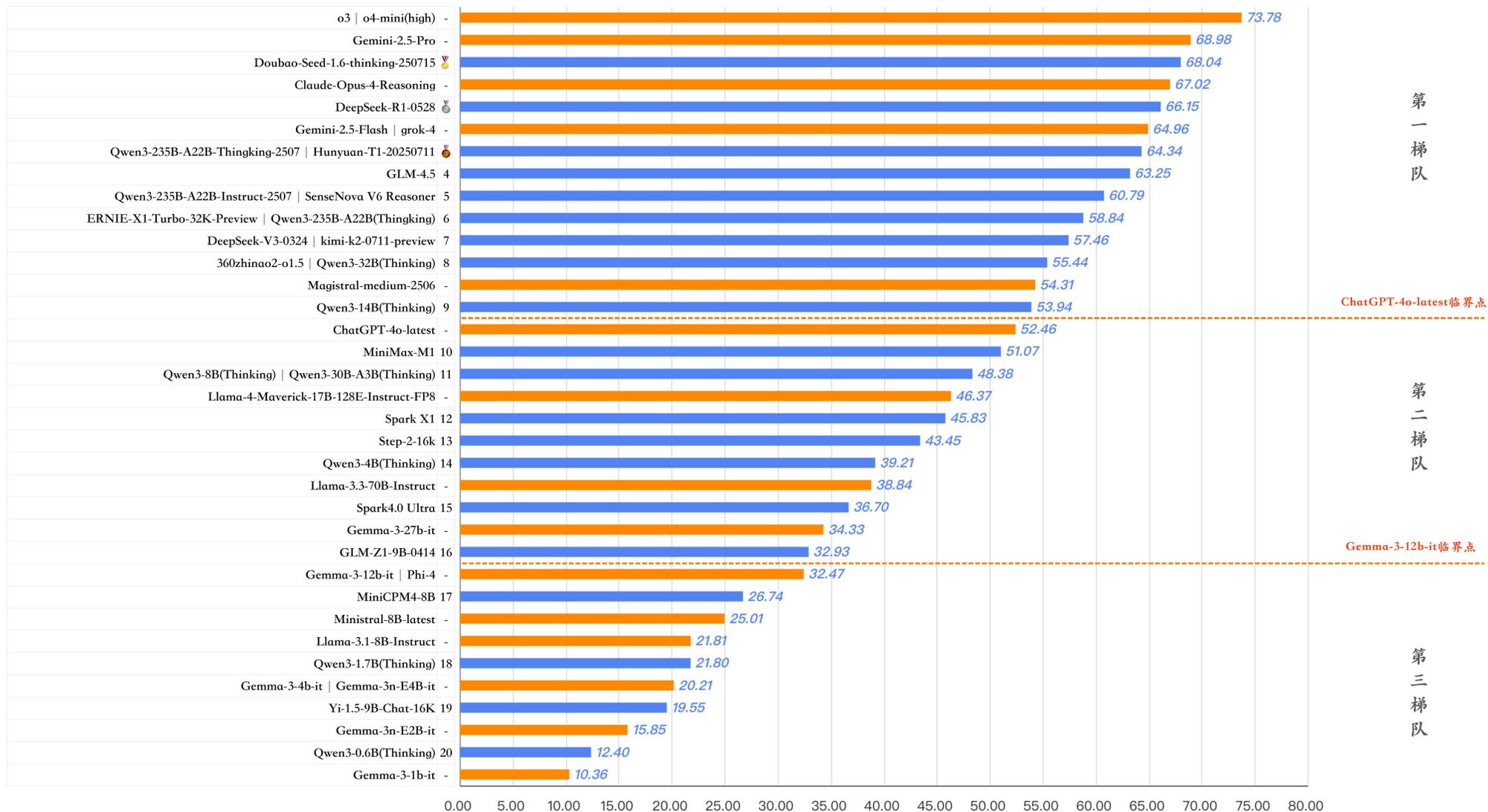
注: 1. 两个维度的组成。推理能力包含: 数学推理、科学推理、代码生成; 应用能力包括: 幻觉控制、精确指令遵循、智能体Agent;

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地位, 引领国内大模型发展。

## SuperCLUE通用大模型基准测评七月总榜

国内模型

海外及其他对比模型



来源: SuperCLUE, 2025年7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, 报告中分数展示为并列中高分。海外模型仅对比参考不参与排名。

## SuperCLUE - 总榜

## 测评分析

SuperCLUE测评基准2025年7月总体表现

排名	模型名称	机构	开源/闭源	总分	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	使用方式
-	o3	OpenAI	闭源	<b>73.78</b>	81.60	59.70	83.76	76.12	64.09	77.41	API
-	o4-mini(high)	OpenAI	闭源	<b>73.32</b>	76.34	55.56	86.14	86.36	59.64	75.90	API
-	Gemini-2.5-Pro	Google	闭源	<b>68.98</b>	74.05	52.59	82.38	82.09	39.76	83.00	API
🏆	Doubao-Seed-1.6-thinking-250715	字节跳动	闭源	<b>68.04</b>	<b>70.77</b>	40.49	<b>84.36</b>	<b>90.67</b>	32.34	<b>89.59</b>	API
-	Claude-Opus-4-Reasoning	Anthropic	闭源	<b>67.02</b>	60.16	45.93	80.40	80.60	44.81	90.24	API
🏆	DeepSeek-R1-0528	深度求索	开源	<b>66.15</b>	<b>75.00</b>	<b>48.15</b>	74.06	82.09	<b>35.31</b>	82.28	API
-	Gemini-2.5-Flash	Google	闭源	<b>64.96</b>	65.89	44.44	73.86	82.09	34.12	89.36	API
-	grok-4	X.AI	闭源	<b>64.84</b>	69.11	44.44	82.18	83.71	29.97	79.65	API
🏆	Qwen3-235B-A22B-Thinking-2507	阿里巴巴	开源	<b>64.34</b>	<b>68.70</b>	<b>42.22</b>	<b>81.78</b>	74.25	<b>43.62</b>	75.44	API
🏆	Hunyuan-T1-20250711	腾讯	闭源	<b>63.73</b>	<b>68.70</b>	37.78	73.27	76.49	<b>42.73</b>	<b>83.39</b>	API
4	GLM-4.5	智谱AI	开源	<b>63.25</b>	66.67	<b>44.62</b>	79.01	<b>83.58</b>	27.22	78.42	API
5	Qwen3-235B-A22B-Instruct-2507	阿里巴巴	开源	<b>60.79</b>	64.57	41.48	78.42	80.97	24.93	74.40	API
5	SenseNova V6 Reasoner	商汤	闭源	<b>60.73</b>	63.36	40.00	<b>80.20</b>	<b>83.58</b>	21.66	75.57	API
6	ERNIE-X1-Turbo-32K-Preview	百度	闭源	<b>58.84</b>	55.12	35.61	73.66	72.01	31.16	<b>85.51</b>	API
6	Qwen3-235B-A22B(Thinking)	阿里巴巴	开源	<b>58.01</b>	63.36	40.00	77.23	83.21	17.51	66.74	API
7	DeepSeek-V3-0324	深度求索	开源	<b>57.46</b>	52.67	36.67	72.48	81.34	19.88	81.71	API
7	kimi-k2-0711-preview	月之暗面	开源	<b>56.90</b>	53.54	37.97	80.00	76.87	19.29	73.70	API
8	360zhinao2-o1.5	360	闭源	<b>55.44</b>	46.56	31.11	76.63	75.75	28.49	74.11	API
8	Qwen3-32B(Thinking)	阿里巴巴	开源	<b>55.28</b>	57.25	34.07	76.24	78.36	18.99	66.78	API
-	Magistral-medium-2506	Mistral AI	闭源	<b>54.31</b>	55.77	35.11	75.45	61.94	23.80	73.78	API
9	Qwen3-14B(Thinking)	阿里巴巴	开源	<b>53.94</b>	57.25	29.63	72.48	77.24	18.99	68.07	API
-	ChatGPT-4o-latest	OpenAI	闭源	<b>52.46</b>	29.77	26.67	78.81	81.72	19.58	78.18	API
10	MiniMax-M1	MiniMax	开源	<b>51.07</b>	53.72	34.96	75.84	54.37	16.67	70.84	API
11	Qwen3-8B(Thinking)	阿里巴巴	开源	<b>48.38</b>	47.33	18.52	65.54	77.24	14.84	66.78	API
11	Qwen3-30B-A3B(Thinking)	阿里巴巴	开源	<b>47.51</b>	55.73	28.15	71.49	45.52	16.91	67.26	API
-	Llama-4-Maverick-17B-128E-Instruct-FP8	Meta	开源	<b>46.37</b>	18.32	20.00	72.48	70.15	21.96	75.29	API
12	Spark X1	科大讯飞	闭源	<b>45.83</b>	46.56	16.30	65.54	60.07	14.24	72.29	API
13	Step-2-16k	阶跃星辰	闭源	<b>43.45</b>	7.63	11.11	65.94	78.73	14.24	83.03	API
14	Qwen3-4B(Thinking)	阿里巴巴	开源	<b>39.21</b>	49.62	15.93	56.04	33.21	14.24	66.23	API
-	Llama-3.3-70B-Instruct	Meta	开源	<b>38.84</b>	7.63	10.37	64.16	60.82	18.99	71.07	API
15	Spark4.0 Ultra	科大讯飞	闭源	<b>36.70</b>	19.38	8.89	56.04	46.64	12.46	76.81	API
-	Gemma-3-27b-it	Google	开源	<b>34.33</b>	19.08	7.41	58.37	36.57	12.46	72.06	API
16	GLM-Z1-9B-0414	智谱AI	开源	<b>32.93</b>	54.20	20.00	28.51	46.64	10.90	37.32	模型
-	Gemma-3-12b-it	Google	开源	<b>32.47</b>	11.45	5.93	53.47	44.03	11.28	68.69	API
-	Phi-4	Microsoft	开源	<b>31.73</b>	16.79	13.33	59.01	36.94	6.82	57.51	API
17	MiniCPM4-8B	面壁智能	开源	<b>26.74</b>	7.63	4.07	42.77	39.55	5.34	61.05	模型
-	Minstral-8B-latest	Mistral AI	开源	<b>25.01</b>	6.87	2.96	39.41	30.60	2.99	67.22	API
-	Llama-3.1-8B-Instruct	Meta	开源	<b>21.81</b>	1.53	1.48	37.03	25.37	5.04	60.39	API
18	Qwen3-1.7B(Thinking)	阿里巴巴	开源	<b>21.80</b>	30.53	4.44	34.46	0.00	6.53	54.82	API
-	Gemma-3-4b-it	Google	开源	<b>20.21</b>	6.87	0.74	38.42	5.06	8.04	62.12	API
-	Gemma-3n-E4B-it	Google	开源	<b>19.70</b>	6.87	1.48	36.24	1.19	10.09	62.30	API
19	Yi-1.5-9B-Chat-16K	零一万物	开源	<b>19.55</b>	3.82	0.74	22.57	24.63	4.45	61.08	模型
-	Gemma-3n-E2B-it	Google	开源	<b>15.85</b>	6.11	0.74	22.77	0.00	8.01	57.48	API
20	Qwen3-0.6B(Thinking)	阿里巴巴	开源	<b>12.40</b>	9.16	1.48	11.49	0.75	2.67	48.88	API
-	Gemma-3-1b-it	Google	开源	<b>10.36</b>	0.00	0.00	11.49	0.00	3.28	47.37	API

注：数据来源SuperCLUE，2025年7月31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。海外产品仅作参考，不参与排名。标红为国内前三名。

### 1. 国内外头部模型竞争激烈，海外模型暂时领先。

OpenAI的o3以73.78分位居榜首，o4-mini(high)以73.32分和Google的Gemini-2.5-Pro以68.98分紧随其后。字节跳动的Doubao-Seed-1.6-thinking-250715和深度求索的DeepSeek-R1-0528分别以68.04分和66.15分的总分位列第四和第六，展现出国产大模型的强大实力和快速发展势头。此外，Qwen3-235B-A22B-Thinking-2507、Hunyuan-T1-20250711、GLM-4.5等国产模型也表现不俗，有接近grok-4的趋势。

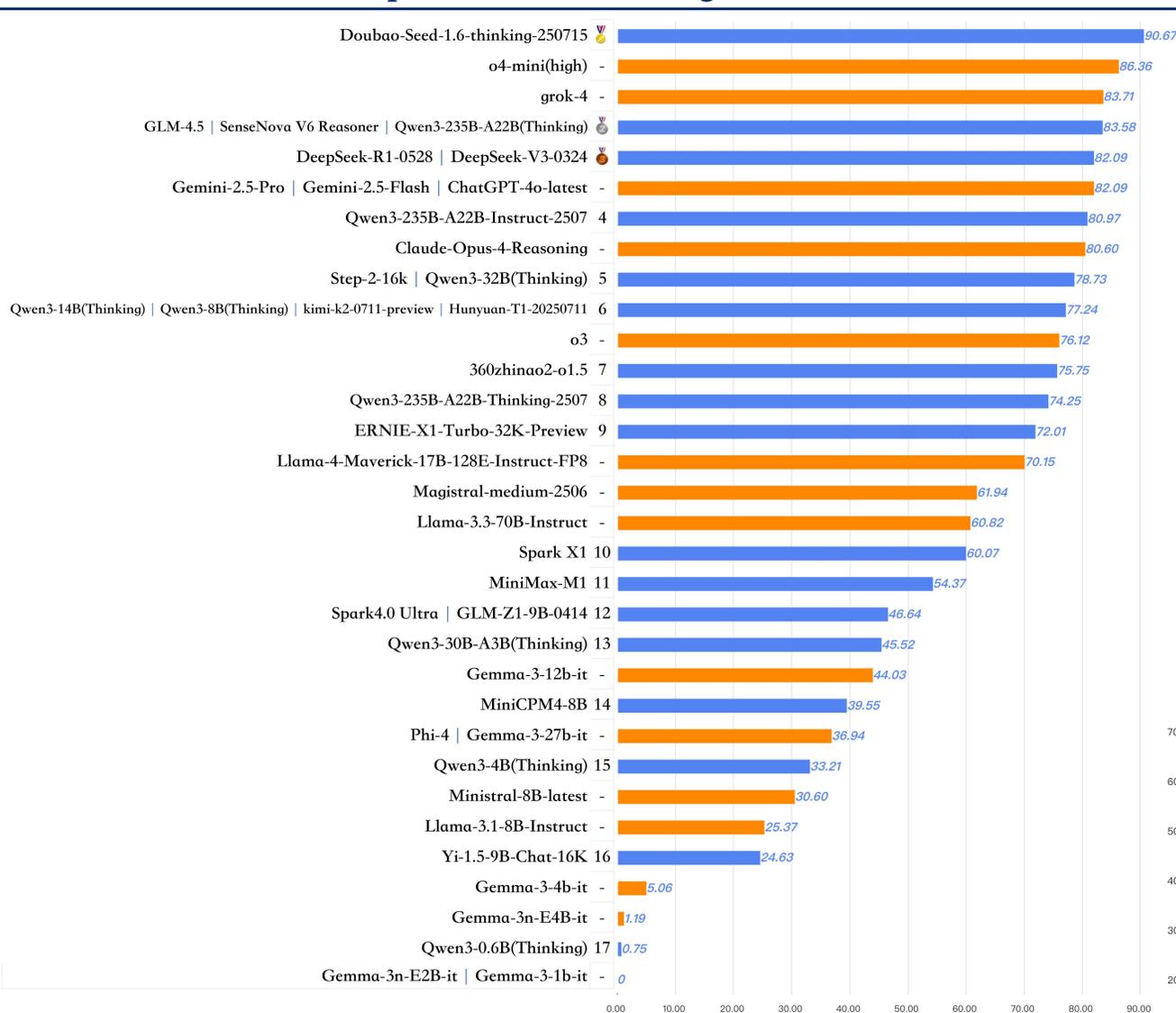
### 2. 国产开源模型崛起势头强劲。

国产开源模型在本次测评中表现出色，包括DeepSeek-R1-0528、GLM-4.5(63.25分)、多个Qwen3系列模型等。DeepSeek-R1-0528以66.15分成为开源模型第一，与闭源模型榜首仅差7.63分。在某些细分领域展现出了与顶级闭源模型相媲美的能力，如Qwen3-235B-A22B-Thinking-2507和kimi-k2-0711-preview在代码生成任务分别取得了81.78分和80分，与顶尖模型o4-mini(high)(86.14分)仅相差5-6分左右。

### 3. 国产大模型在智能体Agent和幻觉控制任务上的表现良好，但在推理任务上稍有逊色。

在智能体Agent任务上，Doubao-Seed-1.6-thinking-250715以90.67分领跑全球，GLM-4.5和SenseNova V6 Reasoner以83.58分并列国内第二。在幻觉控制任务上，Doubao-Seed-1.6-thinking-250715、ERNIE-X1-Turbo-32K-Preview和Hunyuan-T1-20250711分别位于国内前三。在推理任务上，海外最好模型为o3，推理任务的分数为75.02分，而国产最好模型DeepSeek-R1-0528仅有65.74分，相差9.28分。

## SuperCLUE-智能体Agent任务榜单



## 测评分析

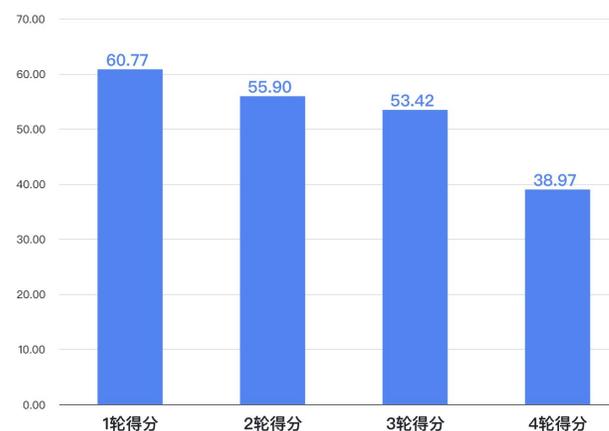
1. 国内多款大模型在智能体测评中表现亮眼，赶超国外主流模型。

Doubao-Seed-1.6-thinking-250715在智能体Agent任务中表现尤为突出，以90.67分的成绩高居榜首。国外o4-mini(high)紧随其后，但两者仍存在4分以上的差距。在智能体Agent任务中，国内多款大模型已实现对国外主流模型的赶超，如GLM-4.5、SenseNova V6 Reasoner、Qwen3-235B-A22B(Thinking)的得分均已超越Gemini-2.5-Pro，且接近grok-4。

2. 模型得分随任务复杂度上升而降低。

在智能体Agent任务中，随着交互轮次和交互步数的增加，模型的工具调用能力会有所降低。对于轮数为1-3轮的题目，模型在每轮的平均得分均在50分以上，而轮数为4轮时，模型在该轮的平均得分仅有38.97分，下降幅度超过14分。随着交互步数的增加，模型的平均得分也呈现逐渐下降的趋势。

智能体Agent任务不同轮数平均得分



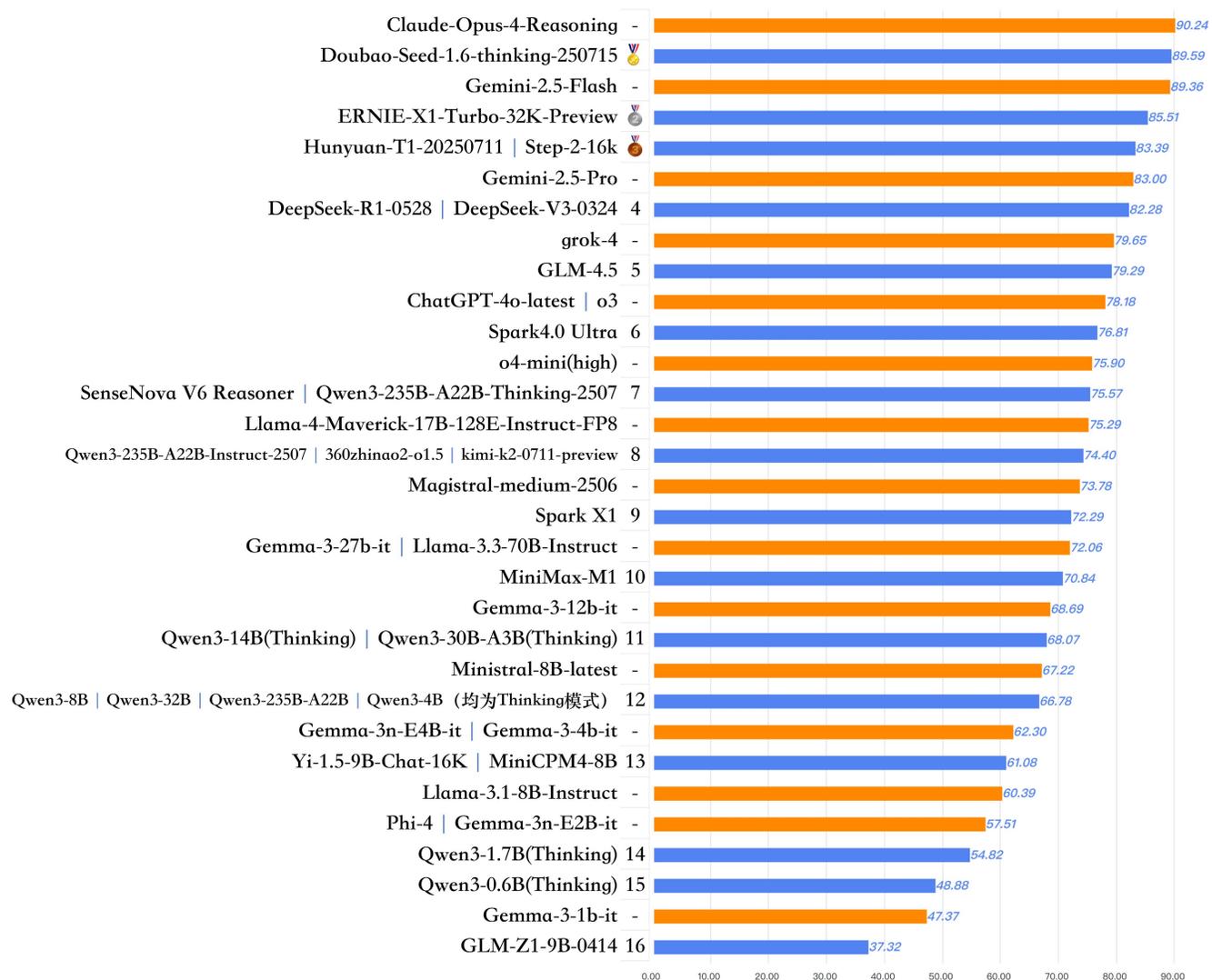
智能体Agent任务不同步数平均得分



来源：SuperCLUE, 2025年7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，报告中分数展示为并列中高分。海外模型仅对比参考不参与排名。

## SuperCLUE-幻觉控制任务榜单



来源：SuperCLUE, 2025年7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，报告中分数展示为并列中高分。海外模型仅对比参考不参与排名。

## 测评分析

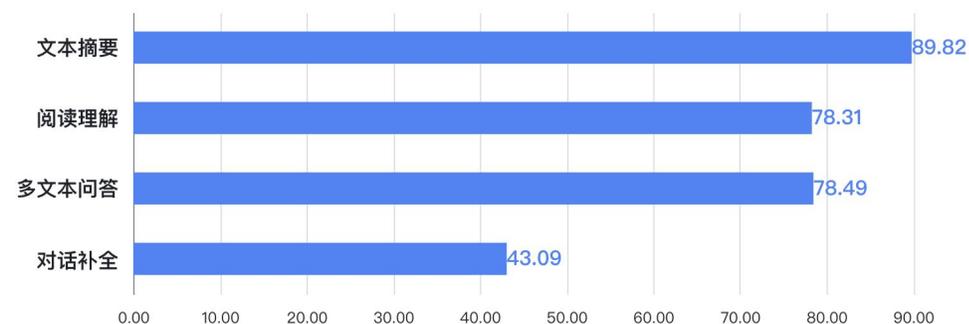
### 1. 国内外头部模型差距在幻觉控制任务上差距较小。

国外头部模型Claude-Opus-4-Reasoning以 90.24 分的高分位居榜首，国内模型Doubao-Seed-1.6-thinking-250715紧随其后，以 89.59 分的成绩与之仅相差 0.65 分。另外，国产模型ERNIE-X1-Turbo-32K-Preview、Hunyuan-T1-20250711和Step-2-16k均有超过海外顶尖模型Gemini-2.5-Pro的表现。

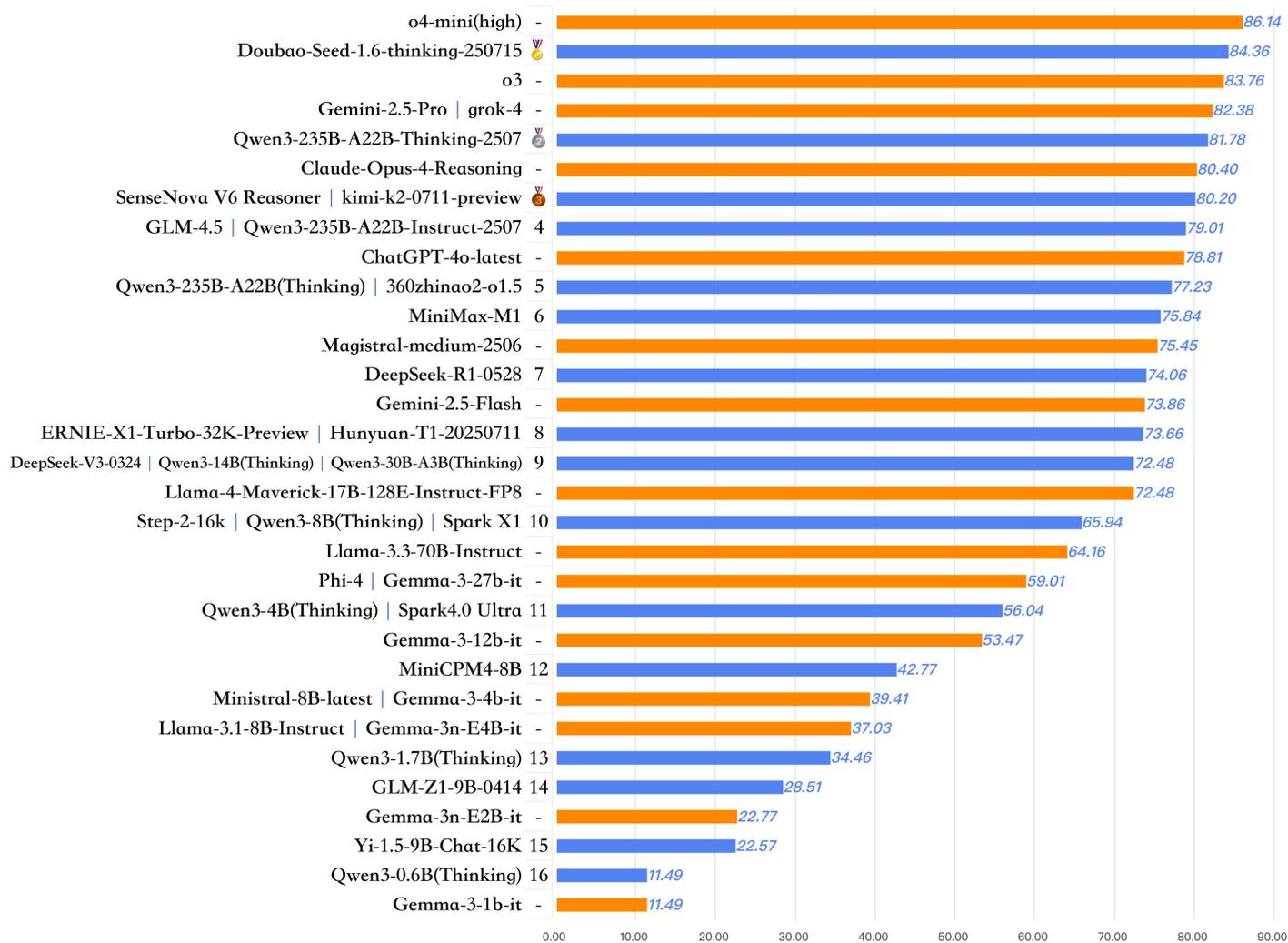
### 2. 任务开放性越高，模型的幻觉越严重。

对于幻觉控制任务中的不同类别，模型的平均得分差异显著：文本摘要（89.82分）、阅读理解（78.31分）、多文本问答（78.49分）因需紧扣原文，相对而言得分较高，幻觉率较低；而对话补全（43.09分）因需推理或自由生成，幻觉风险大幅增加，模型的平均得分相对较低。任务开放性与幻觉率呈现出一定的相关性。

幻觉控制任务各类别平均得分



## SuperCLUE-代码生成任务榜单



来源：SuperCLUE, 2025年7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，报告中分数展示为并列中高分。海外模型仅对比参考不参与排名。

## 测评分析

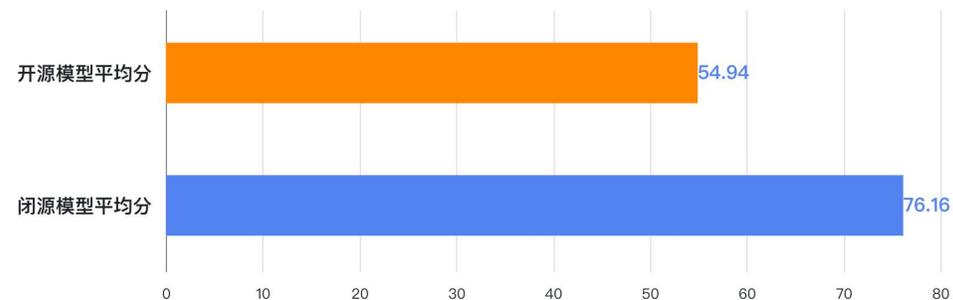
### 1. 海外头部模型相较国产模型在代码生成任务上有一定的优势。

在代码生成任务榜单中，前五名有4个海外模型，仅有一个国产模型。o4-mini(high)以86.14分位居榜首，国产模型Doubao-Seed-1.6-thinking-250715已以84.36分的成绩位于第二，o3、Gemini-2.5-Pro和grok-4三个海外模型占据代码生成任务榜单的第3-5名。国内外头部模型在该任务上有一定的差距。

### 2. 开源模型与闭源模型差距显著。

从代码生成任务的整体表现来看，所有开源模型的平均分（54.94分）与闭源模型的平均分（76.16分）相差约21分。另外，在该任务上的前5名均为闭源模型，说明当前闭源模型在代码生成能力上仍占据明显的优势，而开源模型在代码生成任务上还有一定的优化空间。

代码生成任务开闭源模型平均得分



## SuperCLUE-基础模型榜单

## 测评分析

SuperCLUE测评基准2025年7月基础模型总体表现

排名	模型名称	机构	开源/闭源	总分	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	使用方式
1	Qwen3-235B-A22B-Instruct-2507	阿里巴巴	开源	<b>60.79</b>	<b>64.57</b>	<b>41.48</b>	<b>78.42</b>	<b>80.97</b>	<b>24.93</b>	74.40	API
2	DeepSeek-V3-0324	深度求索	开源	<b>57.46</b>	<b>52.67</b>	<b>36.67</b>	<b>72.48</b>	<b>81.34</b>	<b>19.88</b>	<b>81.71</b>	API
3	kimi-k2-0711-preview	月之暗面	开源	<b>56.90</b>	<b>53.54</b>	<b>37.97</b>	<b>80.00</b>	76.87	<b>19.29</b>	73.70	API
-	ChatGPT-4o-latest	OpenAI	闭源	<b>52.46</b>	29.77	26.67	78.81	81.72	19.58	78.18	API
-	Llama-4-Maverick-17B-128E-Instruct-FP8	Meta	开源	<b>46.37</b>	18.32	20.00	72.48	70.15	21.96	75.29	API
3	Step-2-16k	阶跃星辰	闭源	<b>43.45</b>	7.63	11.11	65.94	<b>78.73</b>	14.24	<b>83.03</b>	API
-	Llama-3.3-70B-Instruct	Meta	开源	<b>38.84</b>	7.63	10.37	64.16	60.82	18.99	71.07	API
4	Spark4.0 Ultra	科大讯飞	闭源	<b>36.70</b>	19.38	8.89	56.04	46.64	12.46	<b>76.81</b>	API
-	Gemma-3-27b-it	Google	开源	<b>34.33</b>	19.08	7.41	58.37	36.57	12.46	72.06	API
-	Gemma-3-12b-it	Google	开源	<b>32.47</b>	11.45	5.93	53.47	44.03	11.28	68.69	API
-	Phi-4	Microsoft	开源	<b>31.73</b>	16.79	13.33	59.01	36.94	6.82	57.51	API
5	MiniCPM4-8B	面壁智能	开源	<b>26.74</b>	7.63	4.07	42.77	39.55	5.34	61.05	模型
-	Minstral-8B-latest	Mistral AI	开源	<b>25.01</b>	6.87	2.96	39.41	30.60	2.99	67.22	API
-	Llama-3.1-8B-Instruct	Meta	开源	<b>21.81</b>	1.53	1.48	37.03	25.37	5.04	60.39	API
-	Gemma-3-4b-it	Google	开源	<b>20.21</b>	6.87	0.74	38.42	5.06	8.04	62.12	API
-	Gemma-3n-E4B-it	Google	开源	<b>19.70</b>	6.87	1.48	36.24	1.19	10.09	62.30	API
6	Yi-1.5-9B-Chat-16K	零一万物	开源	<b>19.55</b>	3.82	0.74	22.57	24.63	4.45	61.08	模型
-	Gemma-3n-E2B-it	Google	开源	<b>15.85</b>	6.11	0.74	22.77	0.00	8.01	57.48	API
-	Gemma-3-1b-it	Google	开源	<b>10.36</b>	0.00	0.00	11.49	0.00	3.28	47.37	API

注：数据来源SuperCLUE，2025年7月31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。海外产品仅作对比参考，不参与排名。标红为国内前三名。

### 1.国内基础模型整体表现优于国外。

国内大模型Qwen3-235B-A22B-Instruct-2507以超60分的成绩位居榜首，与ChatGPT-4o-latest（52.46分）拉开超8分的差距。国内大模型DeepSeek-V3-0324同样表现不俗，在各项任务中均稳居国内前三。

### 2.关键领域国内模型优势显著。

在数学推理方面，Qwen3-235B-A22B-Instruct-2507（64.57分）更是以压倒性优势领先ChatGPT-4o-latest（29.77分），高出34.8分。

### 3.国内模型部分任务仍需优化。

虽然国内大模型Qwen3-235B-A22B-Instruct-2507、DeepSeek-V3-0324、kimi-k2-0711-preview、Step-2-16k在总体表现上占据前三，但在代码生成、智能体Agent和幻觉控制任务中，均不同程度落后于国外表现最好的ChatGPT-4o-latest。

## SuperCLUE-推理模型榜单

## 测评分析

SuperCLUE测评基准2025年7月推理模型表现

排名	模型名称	机构	开源/闭源	总分	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	使用方式
-	o3	OpenAI	闭源	<b>73.78</b>	81.60	59.70	83.76	76.12	64.09	77.41	API
-	o4-mini(high)	OpenAI	闭源	<b>73.32</b>	76.34	55.56	86.14	86.36	59.64	75.90	API
-	Gemini-2.5-Pro	Google	闭源	<b>68.98</b>	74.05	52.59	82.38	82.09	39.76	83.00	API
🏆	Doubao-Seed-1.6-thinking-250715	字节跳动	闭源	<b>68.04</b>	<b>70.77</b>	40.49	<b>84.36</b>	<b>90.67</b>	32.34	<b>89.59</b>	API
-	Claude-Opus-4-Reasoning	Anthropic	闭源	<b>67.02</b>	60.16	45.93	80.40	80.60	44.81	90.24	API
🏆	DeepSeek-R1-0528	深度求索	开源	<b>66.15</b>	<b>75.00</b>	<b>48.15</b>	74.06	82.09	<b>35.31</b>	82.28	API
-	Gemini-2.5-Flash	Google	闭源	<b>64.96</b>	65.89	44.44	73.86	82.09	34.12	89.36	API
-	grok-4	X.AI	闭源	<b>64.84</b>	69.11	44.44	82.18	83.71	29.97	79.65	API
🏆	Qwen3-235B-A22B-Thinking-2507	阿里巴巴	开源	<b>64.34</b>	<b>68.70</b>	<b>42.22</b>	<b>81.78</b>	74.25	<b>43.62</b>	75.44	API
🏆	Hunyuan-T1-20250711	腾讯	闭源	<b>63.73</b>	<b>68.70</b>	37.78	73.27	76.49	<b>42.73</b>	<b>83.39</b>	API
4	GLM-4.5	智谱AI	开源	<b>63.25</b>	66.67	<b>44.62</b>	79.01	<b>83.58</b>	27.22	78.42	API
5	SenseNova V6 Reasoner	商汤	闭源	<b>60.73</b>	63.36	40.00	<b>80.20</b>	<b>83.58</b>	21.66	75.57	API
6	ERNIE-X1-Turbo-32K-Preview	百度	闭源	<b>58.84</b>	55.12	35.61	73.66	72.01	31.16	<b>85.51</b>	API
6	Qwen3-235B-A22B(Thinking)	阿里巴巴	开源	<b>58.01</b>	63.36	40.00	77.23	83.21	17.51	66.74	API
7	360zhinao2-o1.5	360	闭源	<b>55.44</b>	46.56	31.11	76.63	75.75	28.49	74.11	API
7	Qwen3-32B(Thinking)	阿里巴巴	开源	<b>55.28</b>	57.25	34.07	76.24	78.36	18.99	66.78	API
-	Magistral-medium-2506	Mistral AI	闭源	<b>54.31</b>	55.77	35.11	75.45	61.94	23.80	73.78	API
8	Qwen3-14B(Thinking)	阿里巴巴	开源	<b>53.94</b>	57.25	29.63	72.48	77.24	18.99	68.07	API
9	MiniMax-M1	MiniMax	开源	<b>51.07</b>	53.72	34.96	75.84	54.37	16.67	70.84	API
10	Qwen3-8B(Thinking)	阿里巴巴	开源	<b>48.38</b>	47.33	18.52	65.54	77.24	14.84	66.78	API
10	Qwen3-30B-A3B(Thinking)	阿里巴巴	开源	<b>47.51</b>	55.73	28.15	71.49	45.52	16.91	67.26	API
11	Spark X1	科大讯飞	闭源	<b>45.83</b>	46.56	16.30	65.54	60.07	14.24	72.29	API
12	Qwen3-4B(Thinking)	阿里巴巴	开源	<b>39.21</b>	49.62	15.93	56.04	33.21	14.24	66.23	API
13	GLM-Z1-9B-0414	智谱AI	开源	<b>32.93</b>	54.20	20.00	28.51	46.64	10.90	37.32	模型
14	Qwen3-1.7B(Thinking)	阿里巴巴	开源	<b>21.80</b>	30.53	4.44	34.46	0.00	6.53	54.82	API
15	Qwen3-0.6B(Thinking)	阿里巴巴	开源	<b>12.40</b>	9.16	1.48	11.49	0.75	2.67	48.88	API

注：数据来源SuperCLUE，2025年7月31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。海外产品仅作对比参考，不参与排名。标红为国内前三名。

### 1. 国际头部模型领跑，国内模型紧追不舍。

OpenAI的o3、o4-mini(high)占据榜单前两位，海外大厂模型在推理总分上优势明显，国内模型如DeepSeek-R1-0528、Doubao-Seed-1.6-thinking-250715等模型虽紧追不舍，但仍存在一定差距，清晰呈现出推理模型领域“国际头部领跑，国内奋力追赶”的梯队格局。

### 2. 国产模型亮点纷呈，部分领域接近国际顶尖水平。

国内大模型Doubao-Seed-1.6-thinking-250715在代码生成任务中斩获84.36的高分，超越了多数国内外同类模型，与国外表现最佳的o4-mini(high)（86.14分）仅相差1.78分。国内大模型Qwen3-235B-A22B-Thinking-2507则在三项任务中均有不俗表现，且每项成绩均位列国内大模型前三。

### 3. 各维度能力发展不均衡。

从各模型在不同维度的得分来看，能力发展不均衡现象明显。如代码生成维度，DeepSeek-V3-0324得分84.81，而部分模型得分较低，差距巨大。在精确指令遵循维度，模型间分数差异也较为显著。这种不均衡体现了不同模型在能力侧重上的差异，也反映出大模型在追求综合能力提升时，仍面临各维度能力协调发展的挑战。

## SuperCLUE-开源榜单

## 测评分析

SuperCLUE测评基准2025年7月开源模型总体表现

排名	模型名称	机构	参数量	总分	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	使用方式
1	DeepSeek-R1-0528	深度求索	685B	<b>66.15</b>	<b>75.00</b>	<b>48.15</b>	74.06	<b>82.09</b>	<b>35.31</b>	<b>82.28</b>	API
2	Qwen3-235B-A22B-Thinking-2507	阿里巴巴	235B	<b>64.34</b>	<b>68.70</b>	<b>42.22</b>	<b>81.78</b>	74.25	<b>43.62</b>	75.44	API
3	GLM-4.5	智谱AI	355B	<b>63.25</b>	<b>66.67</b>	<b>44.62</b>	<b>79.01</b>	<b>83.58</b>	<b>27.22</b>	78.42	API
4	Qwen3-235B-A22B-Instruct-2507	阿里巴巴	235B	<b>60.79</b>	64.57	41.48	78.42	80.97	24.93	74.40	API
5	Qwen3-235B-A22B(Thinking)	阿里巴巴	235B	<b>58.01</b>	63.36	40.00	77.23	<b>83.21</b>	17.51	66.74	API
5	DeepSeek-V3-0324	深度求索	660B	<b>57.46</b>	52.67	36.67	72.48	81.34	19.88	<b>81.71</b>	API
6	kimi-k2-0711-preview	月之暗面	1000B	<b>56.90</b>	53.54	37.97	<b>80.00</b>	76.87	19.29	73.70	API
7	Qwen3-32B(Thinking)	阿里巴巴	32B	<b>55.28</b>	57.25	34.07	76.24	78.36	18.99	66.78	API
8	Qwen3-14B(Thinking)	阿里巴巴	14B	<b>53.94</b>	57.25	29.63	72.48	77.24	18.99	68.07	API
9	MiniMax-M1	MiniMax	456B	<b>51.07</b>	53.72	34.96	75.84	54.37	16.67	70.84	API
10	Qwen3-8B(Thinking)	阿里巴巴	8B	<b>48.38</b>	47.33	18.52	65.54	77.24	14.84	66.78	API
10	Qwen3-30B-A3B(Thinking)	阿里巴巴	30B	<b>47.51</b>	55.73	28.15	71.49	45.52	16.91	67.26	API
-	Llama-4-Maverick-17B-128E-Instruct-FP8	Meta	402B	<b>46.37</b>	18.32	20.00	72.48	70.15	21.96	75.29	API
11	Qwen3-4B(Thinking)	阿里巴巴	4B	<b>39.21</b>	49.62	15.93	56.04	33.21	14.24	66.23	API
-	Llama-3.3-70B-Instruct	Meta	70B	<b>38.84</b>	7.63	10.37	64.16	60.82	18.99	71.07	API
-	Gemma-3-27b-it	Google	27B	<b>34.33</b>	19.08	7.41	58.37	36.57	12.46	72.06	API
12	GLM-Z1-9B-0414	智谱AI	9B	<b>32.93</b>	54.20	20.00	28.51	46.64	10.90	37.32	模型
-	Gemma-3-12b-it	Google	12B	<b>32.47</b>	11.45	5.93	53.47	44.03	11.28	68.69	API
-	Phi-4	Microsoft	14B	<b>31.73</b>	16.79	13.33	59.01	36.94	6.82	57.51	API
13	MiniCPM4-8B	面壁智能	8B	<b>26.74</b>	7.63	4.07	42.77	39.55	5.34	61.05	模型
-	Ministral-8B-latest	Mistral AI	8B	<b>25.01</b>	6.87	2.96	39.41	30.60	2.99	67.22	API
-	Llama-3.1-8B-Instruct	Meta	8B	<b>21.81</b>	1.53	1.48	37.03	25.37	5.04	60.39	API
14	Qwen3-1.7B(Thinking)	阿里巴巴	1.7B	<b>21.80</b>	30.53	4.44	34.46	0.00	6.53	54.82	API
-	Gemma-3-4b-it	Google	4B	<b>20.21</b>	6.87	0.74	38.42	5.06	8.04	62.12	API
-	Gemma-3n-E4B-it	Google	8B	<b>19.70</b>	6.87	1.48	36.24	1.19	10.09	62.30	API
15	Yi-1.5-9B-Chat-16K	零一万物	9B	<b>19.55</b>	3.82	0.74	22.57	24.63	4.45	61.08	模型
-	Gemma-3n-E2B-it	Google	6B	<b>15.85</b>	6.11	0.74	22.77	0.00	8.01	57.48	API
16	Qwen3-0.6B(Thinking)	阿里巴巴	0.6B	<b>12.40</b>	9.16	1.48	11.49	0.75	2.67	48.88	API
-	Gemma-3-1b-it	Google	1B	<b>10.36</b>	0.00	0.00	11.49	0.00	3.28	47.37	API

注：数据来源SuperCLUE，2025年7月31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。海外产品仅作对比参考，不参与排名。标红为国内前三名。其中模型参数量数据来源于官方披露，若模型为 MoE 架构，以总参数量为准。

### 1. 国内开源模型主导地位凸显。

在7月SuperCLUE开源模型测评中，深度求索的DeepSeek-R1-0528以66.15分位居榜首；阿里旗下多款Qwen3系列模型表现亮眼，其中Qwen3-235B-A22B-Thinking-2507、Qwen3-235B-A22B-Instruct-2507、Qwen3-235B-A22B(Thinking)分别位列第二、第四、第五，国内模型的领先优势进一步巩固。

### 2. 机构竞争格局呈现多元态势。

阿里在开源模型领域的布局成果显著，多款Qwen3系列模型跻身前列；此外，深度求索、智谱AI、月之暗面、MiniMax等机构也有模型入围，展现出行业内多主体竞逐的态势，头部机构各自拥有优势模型，竞争格局日趋丰富。

### 3. 模型专项能力各有突出侧重。

DeepSeek-R1-0528在数学推理、科学推理和幻觉控制三项任务中均以高分领跑国内外开源模型；Qwen3-235B-A22B-Thinking-2507则在代码生成和精确指令遵循任务中位居第一；智能体Agent任务中，GLM-4.5以83.58分夺冠，超出国内头部模型DeepSeek-R1-0528（82.09分）1.49分。整体来看，各模型在专项能力上的差异化竞争特征明显。

## SuperCLUE-推理任务榜单

## 测评分析

SuperCLUE测评基准2025年7月推理任务表现

排名	模型名称	机构	开源/闭源	推理任务总分	数学推理	科学推理	代码生成	使用方式
-	o3	OpenAI	闭源	<b>75.02</b>	81.60	59.70	83.76	API
-	o4-mini(high)	OpenAI	闭源	<b>72.68</b>	76.34	55.56	86.14	API
-	Gemini-2.5-Pro	Google	闭源	<b>69.67</b>	74.05	52.59	82.38	API
🏆	DeepSeek-R1-0528	深度求索	开源	<b>65.74</b>	<b>75.00</b>	<b>48.15</b>	74.06	API
-	grok-4	X.AI	闭源	<b>65.24</b>	69.11	44.44	82.18	API
🏆	Doubao-Seed-1.6-thinking-250715	字节跳动	闭源	<b>65.21</b>	<b>70.77</b>	40.49	<b>84.36</b>	API
🥈	Qwen3-235B-A22B-Thinking-2507	阿里巴巴	开源	<b>64.23</b>	<b>68.70</b>	<b>42.22</b>	<b>81.78</b>	API
🥉	GLM-4.5	智谱AI	开源	<b>63.43</b>	66.67	<b>44.62</b>	79.01	API
-	Claude-Opus-4-Reasoning	Anthropic	闭源	<b>62.16</b>	60.16	45.93	80.40	API
🥈	Qwen3-235B-A22B-Instruct-2507	阿里巴巴	开源	<b>61.49</b>	64.57	41.48	78.42	API
-	Gemini-2.5-Flash	Google	闭源	<b>61.40</b>	65.89	44.44	73.86	API
🥉	SenseNova V6 Reasoner	商汤	闭源	<b>61.19</b>	63.36	40.00	<b>80.20</b>	API
4	Qwen3-235B-A22B(Thinking)	阿里巴巴	开源	<b>60.20</b>	63.36	40.00	77.23	API
4	Hunyuan-T1-20250711	腾讯	闭源	<b>59.92</b>	<b>68.70</b>	37.78	73.27	API
5	kimi-k2-0711-preview	月之暗面	开源	<b>57.17</b>	53.54	37.97	80.00	API
6	Qwen3-32B(Thinking)	阿里巴巴	开源	<b>55.85</b>	55.25	34.07	76.24	API
-	Magistral-medium-2506	Mistral AI	闭源	<b>55.44</b>	55.77	35.11	75.45	API
7	MiniMax-M1	MiniMax	开源	<b>54.84</b>	53.72	34.96	75.84	API
7	ERNIE-X1-Turbo-32K-Preview	百度	闭源	<b>54.80</b>	55.12	35.61	73.66	API
7	DeepSeek-V3-0324	深度求索	开源	<b>53.94</b>	52.67	36.67	72.48	API
8	Qwen3-14B(Thinking)	阿里巴巴	开源	<b>53.12</b>	57.25	29.63	72.48	API
9	Qwen3-30B-A3B(Thinking)	阿里巴巴	开源	<b>51.79</b>	55.73	28.15	71.49	API
9	360zhinao2-o1.5	360	闭源	<b>51.43</b>	46.56	31.11	76.63	API
-	ChatGPT-4o-latest	OpenAI	闭源	<b>45.08</b>	29.77	26.67	78.81	API
10	Qwen3-8B(Thinking)	阿里巴巴	开源	<b>43.80</b>	47.33	18.52	65.54	API
11	Spark X1	科大讯飞	闭源	<b>42.80</b>	46.56	16.30	65.54	API
12	Qwen3-4B(Thinking)	阿里巴巴	开源	<b>40.53</b>	49.62	15.93	56.04	API
-	Llama-4-Maverick-17B-128E-Instruct-FP8	Meta	开源	<b>36.93</b>	18.32	20.00	72.48	API
13	GLM-Z1-9B-0414	智谱AI	开源	<b>34.24</b>	54.20	20.00	28.51	模型
-	Phi-4	Microsoft	开源	<b>29.71</b>	16.79	13.33	59.01	API
-	Gemma-3-27b-it	Google	开源	<b>28.29</b>	19.08	7.41	58.37	API
14	Step-2-16k	阶跃星辰	闭源	<b>28.23</b>	7.63	11.11	65.94	API
14	Spark4.0 Ultra	科大讯飞	闭源	<b>28.10</b>	19.38	8.89	56.04	API
-	Llama-3.3-70B-Instruct	Meta	开源	<b>27.39</b>	7.63	10.37	64.16	API
-	Gemma-3-12b-it	Google	开源	<b>23.62</b>	11.45	5.93	53.47	API
15	Qwen3-1.7B(Thinking)	阿里巴巴	开源	<b>23.14</b>	30.53	4.44	34.46	API
16	MiniCPM4-8B	面壁智能	开源	<b>18.16</b>	7.63	4.07	42.77	模型
-	Ministral-8B-latest	Mistral AI	开源	<b>16.41</b>	6.87	2.96	39.41	API
-	Gemma-3-4b-it	Google	开源	<b>15.34</b>	6.87	0.74	38.42	API
-	Gemma-3n-E4B-it	Google	开源	<b>14.86</b>	6.87	1.48	36.24	API
-	Llama-3.1-8B-Instruct	Meta	开源	<b>13.35</b>	5.53	1.48	37.03	API
-	Gemma-3n-E2B-it	Google	开源	<b>9.87</b>	6.11	0.74	22.77	API
17	Yi-1.5-9B-Chat-16K	零一万物	开源	<b>9.04</b>	3.82	0.74	22.57	模型
18	Qwen3-0.6B(Thinking)	阿里巴巴	开源	<b>7.38</b>	9.16	1.48	11.49	API
-	Gemma-3-1b-it	Google	开源	<b>3.83</b>	0.00	0.00	11.49	API

注：数据来源SuperCLUE，2025年7月31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。海外产品仅作参考，不参与排名。标红为国内前三名。

### 1. 榜单前三均为海外头部模型，优势显著。

o3、o4-mini(high)和Gemini-2.5-Pro分别以75.02分、72.68分和69.67分获得推理任务榜单的前三。DeepSeek-R1和Doubao-Seed-1.6-thinking-250715分别以65.74分和65.21分获得国内金牌，但与海外头部模型相比，差距还比较显著。

### 2. 国内头部模型在推理任务上竞争激烈，推理能力差距较小。

在推理任务榜单中，DeepSeek-R1和Doubao-Seed-1.6-thinking-250715获得金牌，Qwen3-235B-A22B-Thinking-2507和GLM-4.5获得银牌，Qwen3-235B-A22B-Instruct-2507和SenseNova V6 Reasoner获得铜牌，极差不到5分，国内头部模型的推理能力差距较小。

### 3. 国内开源模型的推理能力显著优于海外开源模型。

国内开源模型有5个推理任务得分在60分以上，但海外开源模型最高分不到37分，领先幅度接近23分，说明国内开源模型的推理能力要显著优于海外开源模型。

## SuperCLUE-10B级别小模型榜单

## 测评分析

SuperCLUE测评基准2025年7月10B级别小模型总体表现

排名	模型名称	机构	参数量	总分	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	使用方式
🥇	Qwen3-8B(Thinking)	阿里巴巴	8B	<b>48.38</b>	<b>47.33</b>	<b>18.52</b>	<b>65.54</b>	<b>77.24</b>	<b>14.84</b>	<b>66.78</b>	API
🥈	Qwen3-4B(Thinking)	阿里巴巴	4B	<b>39.21</b>	<b>49.62</b>	<b>15.93</b>	<b>56.04</b>	33.21	<b>14.24</b>	<b>66.23</b>	API
🥉	GLM-Z1-9B-0414	智谱AI	9B	<b>32.93</b>	<b>54.20</b>	<b>20.00</b>	28.51	<b>46.64</b>	<b>10.90</b>	37.32	模型
4	MiniCPM4-8B	面壁智能	8B	<b>26.74</b>	7.63	4.07	<b>42.77</b>	<b>39.55</b>	5.34	61.05	模型
-	Ministral-8B-latest	Mistral AI	8B	<b>25.01</b>	6.87	2.96	39.41	30.60	2.99	67.22	API
-	Llama-3.1-8B-Instruct	Meta	8B	<b>21.81</b>	1.53	1.48	37.03	25.37	5.04	60.39	API
5	Qwen3-1.7B(Thinking)	阿里巴巴	1.7B	<b>21.80</b>	30.53	4.44	34.46	0.00	6.53	54.82	API
-	Gemma-3-4b-it	Google	4B	<b>20.21</b>	6.87	0.74	38.42	5.06	8.04	62.12	API
-	Gemma-3n-E4B-it	Google	8B	<b>19.70</b>	6.87	1.48	36.24	1.19	10.09	62.30	API
6	Yi-1.5-9B-Chat-16K	零一万物	9B	<b>19.55</b>	3.82	0.74	22.57	24.63	4.45	<b>61.08</b>	模型
-	Gemma-3n-E2B-it	Google	6B	<b>15.85</b>	6.11	0.74	22.77	0.00	8.01	57.48	API
7	Qwen3-0.6B(Thinking)	阿里巴巴	0.6B	<b>12.40</b>	9.16	1.48	11.49	0.75	2.67	48.88	API
-	Gemma-3-1b-it	Google	1B	<b>10.36</b>	0.00	0.00	11.49	0.00	3.28	47.37	API

注：数据来源SuperCLUE，2025年7月31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。海外产品仅作对比参考，不参与排名。标红为国内前三名。其中模型参数量数据来源于官方披露，若模型为 MoE 架构，以总参数量为准。

### 1. Qwen3-8B (Thinking) 领跑国内10B级小模型。

在10B级别的小模型榜单中，前四均为国内小模型，其中Qwen3-8B(Thinking)以48.38分的总分位居榜首。该模型在各项任务中均跻身国内大模型前三甲，有力打破了“参数规模小=性能表现弱”的固有认知。

### 2. 任务表现分化显著，小模型“长短板”特征直观显现。

同一模型在不同任务中的得分呈现显著差异。如Qwen3-4B(Thinking)在幻觉控制任务中能取得66.23分的优异成绩，但在精确指令遵循任务上仅得14.24分，这一鲜明对比，直观展现了小模型在任务适配能力上的“长短板”特征。

### 3. 国内头部小模型大幅领先国外竞品。

国内头部小模型表现抢眼，包揽榜单前四，Qwen3-8B(Thinking)远超国外表现最佳的Ministral-8B-latest (25.01分)，领先优势达18.37分，充分彰显了国内小参数模型的强劲实力。

2025年端侧小模型快速发展，已在设备端侧（非云）上实现本地运行，其中PC、手机、智能眼镜、机器人等大量场景已展现出极高的落地可行性。

➢ 国内端侧小模型进展迅速，相比国外小模型，国内小模型在中文场景下展现出更好的性能表现

✓ 国产小模型占据主流地位。国内大模型Qwen3 - 4B(Thinking)在此领域表现抢眼，以 39.21 分的高分位居榜首，且在各项细分任务中均展现出卓越实力，全部位列第一。尤其在幻觉控制方面，该模型取得了超过60分的优异成绩，充分彰显了小模型的巨大潜力。

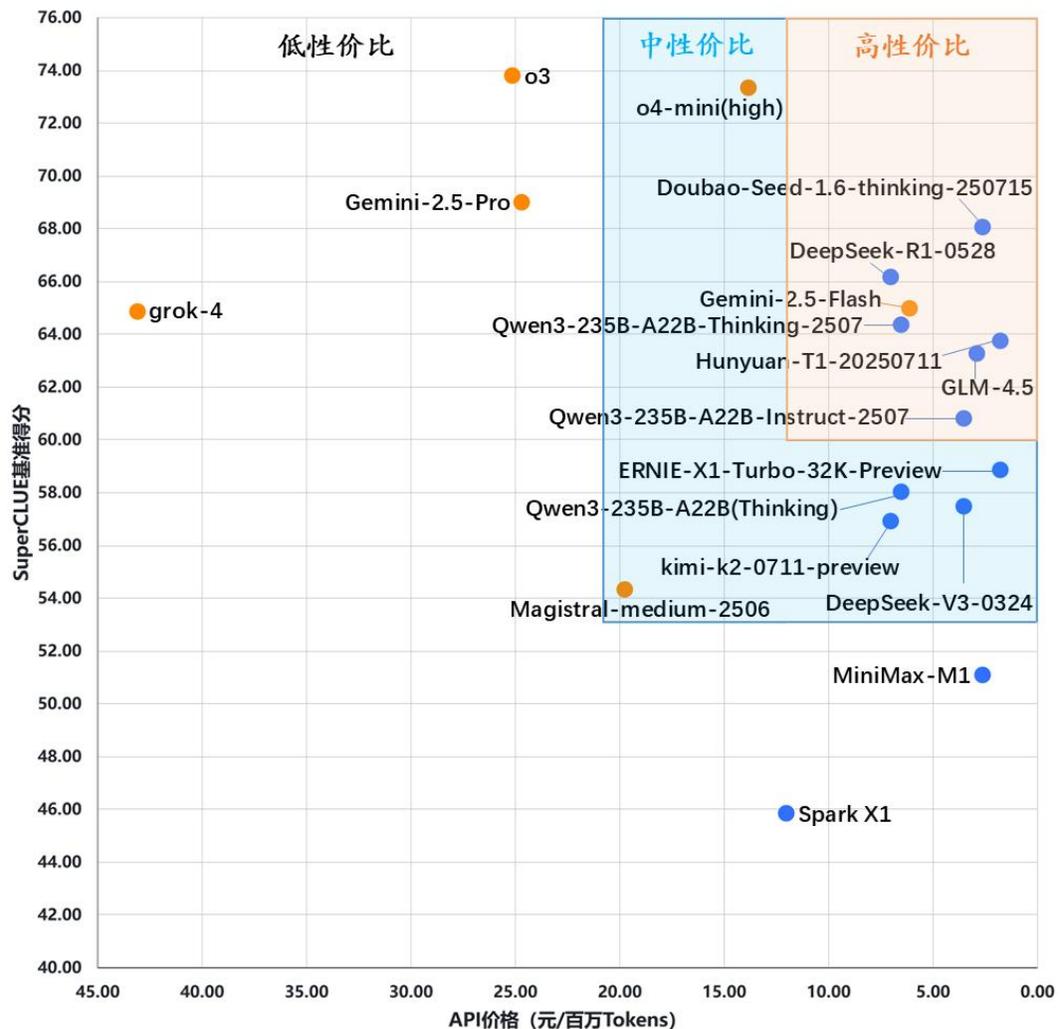
✓ 围绕数学推理、代码生成等关键任务，可清晰看到小模型在“有限参数”约束下的能力边界。如Qwen3-1.7B(Thinking)在代码生成任务中获得34.46分的不错成绩，但在智能体 Agent 任务中得分却为 0，这一差异直观暴露了端侧小模型在复杂场景适配中存在的难点。

## SuperCLUE——端侧5B级别小模型榜单

排名	模型名称	机构	参数量	总分	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	使用方式
1	Qwen3-4B(Thinking)	阿里巴巴	4B	<b>39.21</b>	<b>49.62</b>	<b>15.93</b>	<b>56.04</b>	<b>33.21</b>	<b>14.24</b>	<b>66.23</b>	API
2	Qwen3-1.7B(Thinking)	阿里巴巴	1.7B	<b>21.80</b>	<b>30.53</b>	<b>4.44</b>	<b>34.46</b>	0.00	<b>6.53</b>	<b>54.82</b>	API
-	Gemma-3-4b-it	Google	4B	20.21	6.87	0.74	38.42	5.06	8.04	62.12	API
3	Qwen3-0.6B(Thinking)	阿里巴巴	0.6B	<b>12.40</b>	<b>9.16</b>	<b>1.48</b>	<b>11.49</b>	<b>0.75</b>	2.67	<b>48.88</b>	API
-	Gemma-3-1b-it	Google	1B	10.36	0.00	0.00	11.49	0.00	3.28	47.37	API

注：数据来源SuperCLUE，2025年7月31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。海外产品仅作对比参考，不参与排名。标红为国内前三名。其中模型参数量数据来源于官方披露，若模型为 MoE 架构，以总参数量为准。

## SuperCLUE-大模型性价比区间分布



## 趋势分析

### 1. 国内头部模型较海外模型呈现出更高的性价比。

国内大模型展现出强劲的性价比优势，如Hunyuan-T1-20250711、GLM-4.5、Doubao-Seed-1.6-thinking-250715、Qwen3-235B-A22B-Thinking-2507。海外模型Gemini-2.5-Flash与国内模型Qwen3-235B-A22B-Thinking-2507、DeepSeek-R1-0528在得分和价格上不相上下，共同处于高性价比区间。

### 2. 国内中性性价比模型能力待提升。

国内处于中性性价比区间的模型，价格普遍控制在9元/百万Tokens以内，部分超低价模型甚至低于3元/百万Tokens，但这类模型的基准得分仅在56-60分之间，整体能力还有较大的提升空间。

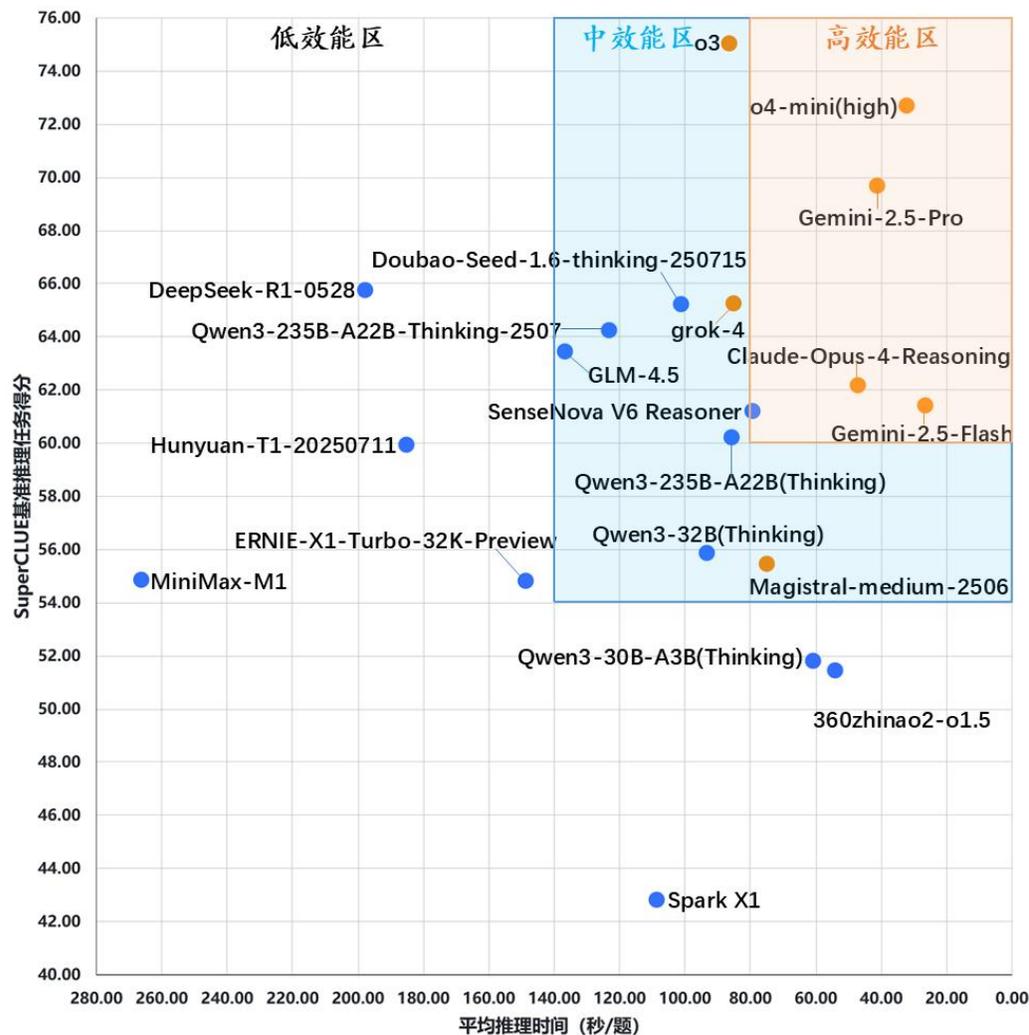
### 3. 海外头部模型性价比较低。

尽管o3、o4-mini(high)以超70分的成绩领跑，但二者价格显著高于其他模型。其中，o3的价格比国内表现最优的Doubao-Seed-1.6-thinking-250715高出了超20元/百万Tokens以上，性价比相对较低。

数据来源：SuperCLUE，2025年7月31日；开源模型如Qwen3-235B-A22B(Thinking)使用方式为API，价格信息均来自官方信息。

注：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7月的标准价格（非优惠价格）。

## SuperCLUE-大模型综合效能区间分布



## 趋势分析

### 1. 国外头部模型稳居高效能区，展现强劲应用实力。

从推理效能分布图可见，国外头部模型o4-mini(high)、Gemini-2.5-Pro、Gemini-2.5-Flash、Claude-Opus-4-Reasoning的平均推理时间均控制在60秒以内，且推理任务得分均突破60分，完全符合「高效能区」标准，充分展现出极强的实际应用效能。

### 2. 国内头部模型仅有SenseNova V6 Reasoner趋近高效能区。

国内头部模型中，SenseNova V6 Reasoner在推理时间与基准得分的综合效能方面表现最为突出，已处于「高效能区」的边界位置，彰显了国内模型的不俗实力。国内大模型Qwen3-235B-A22B(Thinking)同样表现出色，紧随其后。

### 3. 部分国内模型得分反超国外，但耗时差距明显。

国内大模型如Doubao-Seed-1.6-thinking-250715、Qwen3-235B-A22B-Thinking-2507、GLM-4.5、DeepSeek-R1-0528在推理任务得分上已实现对部分国外头部模型的反超，但在平均推理时间的耗时上却存在着较大差距。

数据来源：SuperCLUE，2025年7月31日；

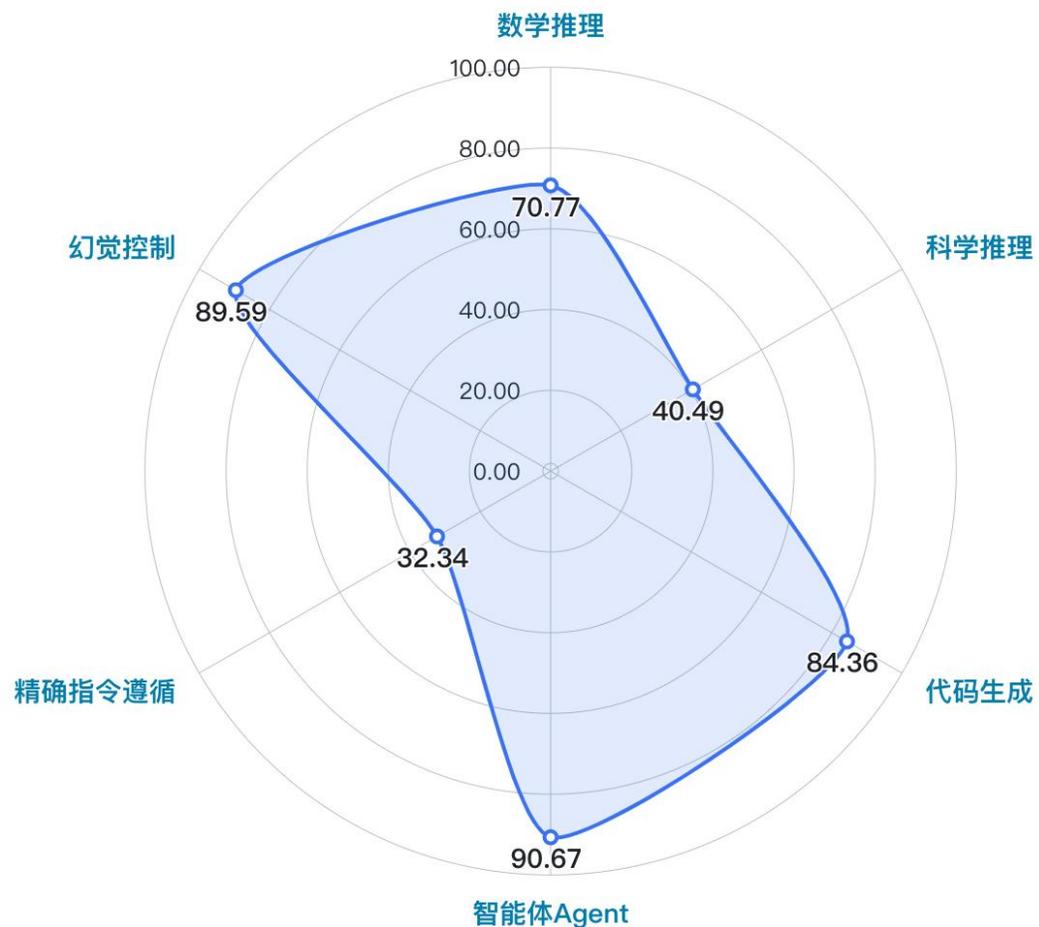
模型推理速度选取7月测评中具有公开API的部分模型。平均推理时间为所有推理任务测评数据推理时间的平均值（秒）。推理任务得分为推理任务总分：数学推理、科学推理和代码生成的平均分。

## SuperCLUE通用基准测评六大任务成绩

## 模型介绍与测评分析

Doubao-Seed-1.6-thinking-250715

Doubao-Seed-1.6-thinking-250715是由字节跳动在2025年7月推出的最新版本的多模态深度思考模型。



### 简介

#### 技术亮点

该模型在思考能力上进行了大幅强化，对比上一代深度理解模型，在编程、数学、逻辑推理等基础能力上进一步提升，支持视觉理解。

#### 能力优势

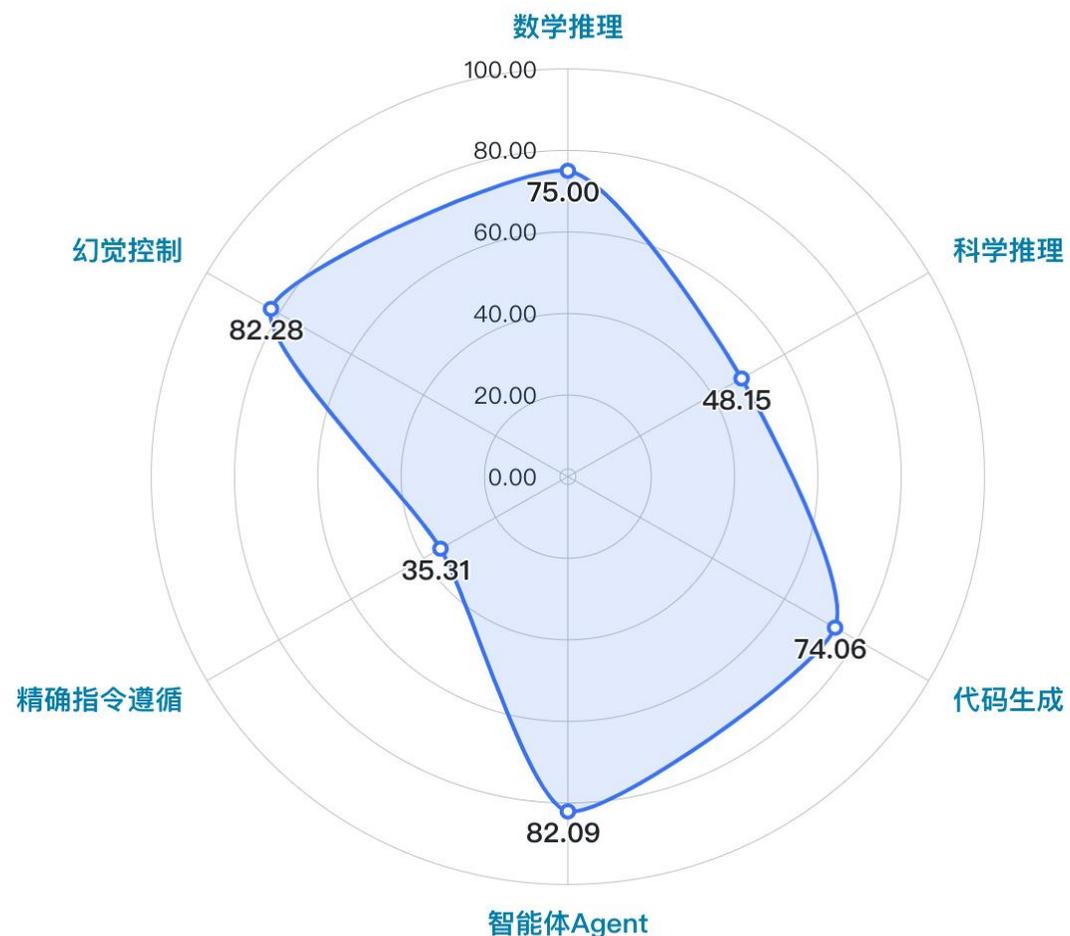
- 1.擅长通用 Agent 任务：在SuperCLUE智能体Agent任务排名全球第一；
- 2.擅长幻觉控制和代码任务：在SuperCLUE幻觉控制、代码生成任务均排名全球第二，国内第一；
- 3.擅长数学推理：在SuperCLUE数学推理任务排名国内第二。

#### 应用场景

智能体、复杂推理任务、代码编程、科学研究、中文创作

## SuperCLUE通用基准测评六大任务成绩

### DeepSeek-R1-0528



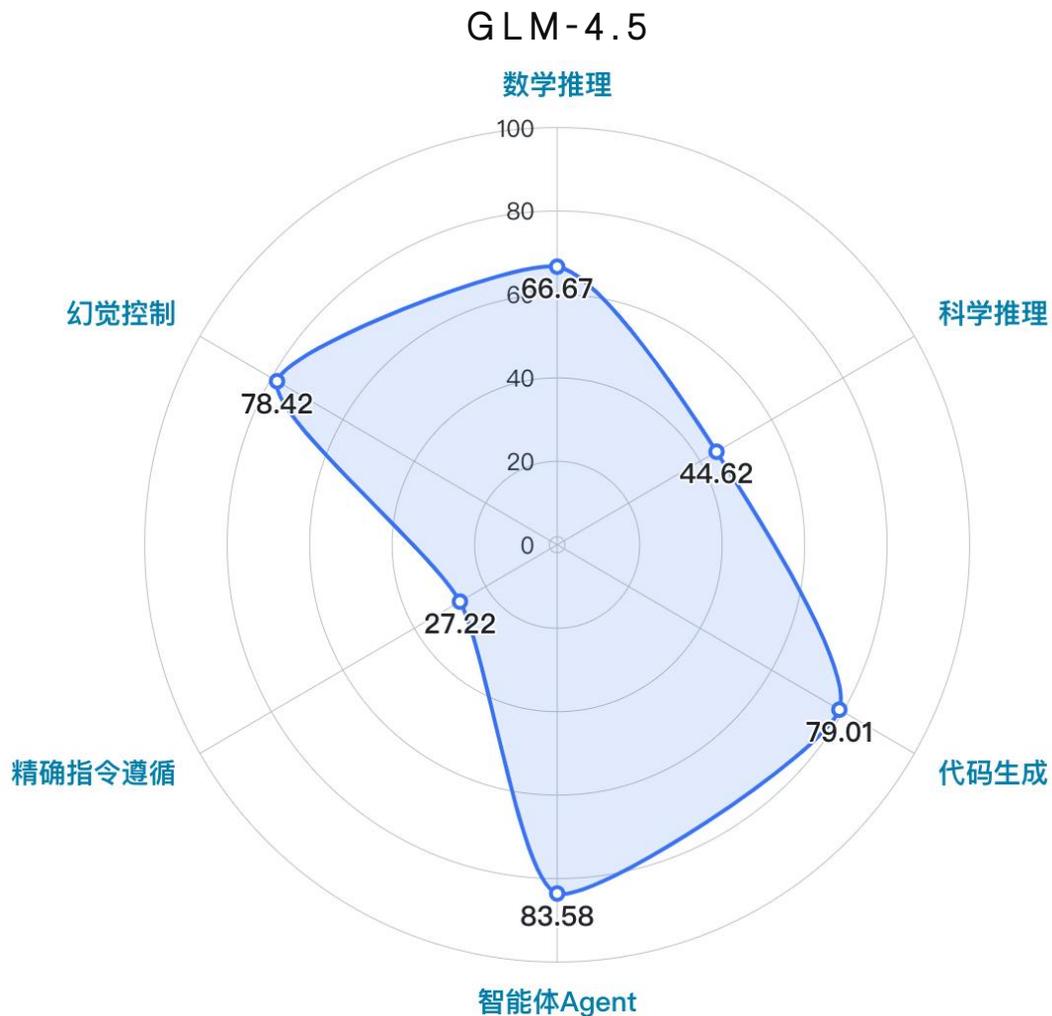
数据来源：SuperCLUE，2025年7月31日。

## 模型介绍与测评分析

DeepSeek-R1-0528是由深度求索在2025年5月28日发布的DeepSeek-R1小版本升级模型，该模型使用2024年12月所发布的DeepSeek V3 Base模型作为基座，但在后训练过程中投入了更多算力，显著提升了模型的思维深度与推理能力。

简介	
技术亮点	模型参数为685B（其中14B为MTP层），采用MLA、MoE等架构提升推理效率与长上下文处理能力，通过无监督强化学习等训练方式增强推理性能。
能力优势	1. 擅长复杂推理：在SuperCLUE数学推理、科学推理任务均排名国内第一； 2. 擅长精确指令遵循任务：在SuperCLUE精确指令遵循任务均排名国内前三； 3. 擅长通用Agent任务：在SuperCLUE智能体Agent任务排名国内前三。
应用场景	复杂推理任务、代码编程、智能体

## SuperCLUE通用基准测评六大任务成绩



数据来源：SuperCLUE，2025年7月31日。

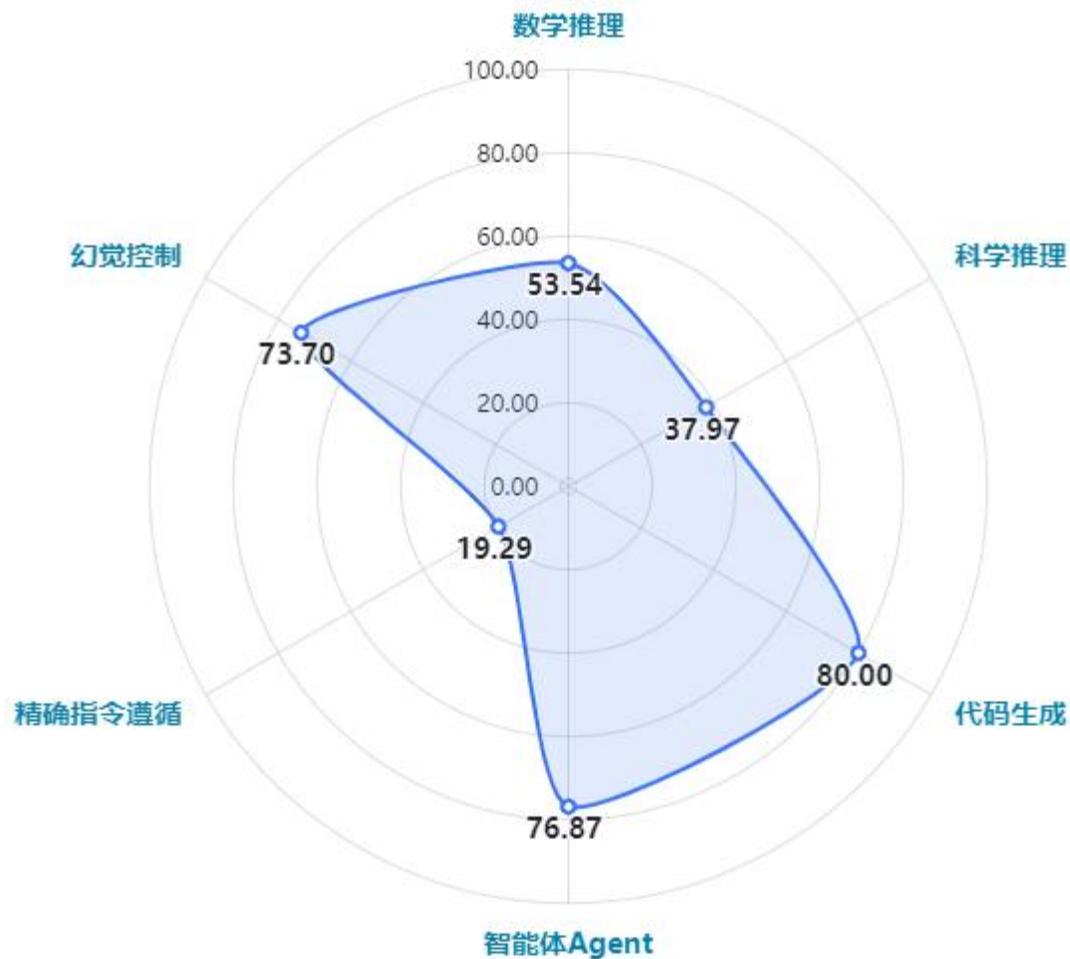
## 模型介绍与测评分析

GLM-4.5是由智谱华章在2025年7月28日正式发布并开源的最新旗舰模型，该模型采用MoE架构，该模型拥有3550亿个总参数和320亿个激活参数，将推理、编码和代理功能统一到一个模型中，以满足快速发展的代理应用日益复杂的需求。

	简介
技术亮点	在 MoE 层中使用了无损平衡路由和 S 型门控；使用Muon优化器、QK-Norm技术、分组查询注意力（Grouped-Query Attention）与部分旋转位置编码（partial RoPE），并加入多令牌预测层（MTP），提升训练效率和推理性能。
能力优势	1.擅长通用 Agent 任务：在SuperCLUE智能体Agent任务排名国内第二； 2.擅长复杂推理任务：在SuperCLUE科学推理任务排名国内第二。
应用场景	智能体、复杂推理任务、代码编程

## SuperCLUE通用基准测评六大任务成绩

kimi-k2-0711-preview



## 模型介绍与测评分析

kimi-k2-0711-preview是由月之暗面在2025年7月11日正式发布并开源的最新旗舰模型，该模型采用MoE架构，总参数 1T，激活参数 32B。

	简介
技术亮点	<ol style="list-style-type: none"><li>预训练阶段使用 MuonClip 优化器实现万亿参数模型的稳定高效训练，有效提高 Token 利用效率，找到新的 Scaling 空间；</li><li>大规模 Agentic Tool Use 数据合成；</li><li>引入自我评价机制的通用强化学习。</li></ol>
能力优势	<ol style="list-style-type: none"><li>擅长代码任务：在SuperCLUE代码生成任务排名国内第三；</li><li>擅长推理任务：在SuperCLUE数学推理、科学推理任务排名基础模型前三。</li></ol>
应用场景	智能体、复杂推理任务、代码编程

## SuperCLUE国内大模型能力成熟度指数

大模型能力成熟度指数——SC指数		
能力	成熟度SC指数	成熟度区间
数学推理	<b>0.63</b>	中成熟度 (0.5-0.8)
智能体Agent	<b>0.60</b>	
科学推理	<b>0.59</b>	
代码生成	<b>0.56</b>	
幻觉控制	<b>0.45</b>	低成熟度 (0.1-0.5)
精确指令遵循	<b>0.40</b>	

数据来源：SuperCLUE, 2025年7月31日。

SC成熟度指数计算方法说明：

- 计算相对差距：**对于每个任务，首先找出所有国产闭源模型中的最高分。然后，计算每个模型的分数与这个最高分之间的相对差距。这个差距是通过  $(\text{最高分} - \text{模型分数}) / \text{最高分}$  来计算的。这个值反映了每个模型与顶尖水平的距离。
- 计算累积差距：**将一个任务中所有国产闭源模型的相对差距加起来，然后除以模型的数量，得到一个平均的相对差距，称之为“相对差距累积值”。这个值代表了在这个任务上，所有国产闭源模型表现的平均离散程度。
- 归一化处理：**为了消除不同任务之间难度和评分标准差异带来的影响，需要对“相对差距累积值”进行归一化处理。这里采用的是最小-最大归一化方法。首先，在所有国产闭源模型的相对差距中，找出最小值和最大值。然后，通过  $(\text{相对差距累积值} - \text{最小相对差距}) / (\text{最大相对差距} - \text{最小相对差距})$  这个公式，将累积差距缩放到  $[0, 1]$  的区间内。
- 计算最终成熟度：**最后，用1减去归一化后的差距，即  $1 - \text{归一化差距}$ ，就得到了最终的任务成熟度。这个值越高，代表该任务的整体发展水平越高，模型间的平均表现越接近顶尖水平。

## 国内大模型成熟度分析

### 1. 高成熟度能力

- ✓ 高成熟度指大部分闭源大模型普遍擅长的能力，SC成熟度指数在0.8至1.0之间。
- ✓ 在本次测评的六大任务中，当前国内闭源大模型成熟度较高的任务暂时没有。

### 2. 中成熟度能力

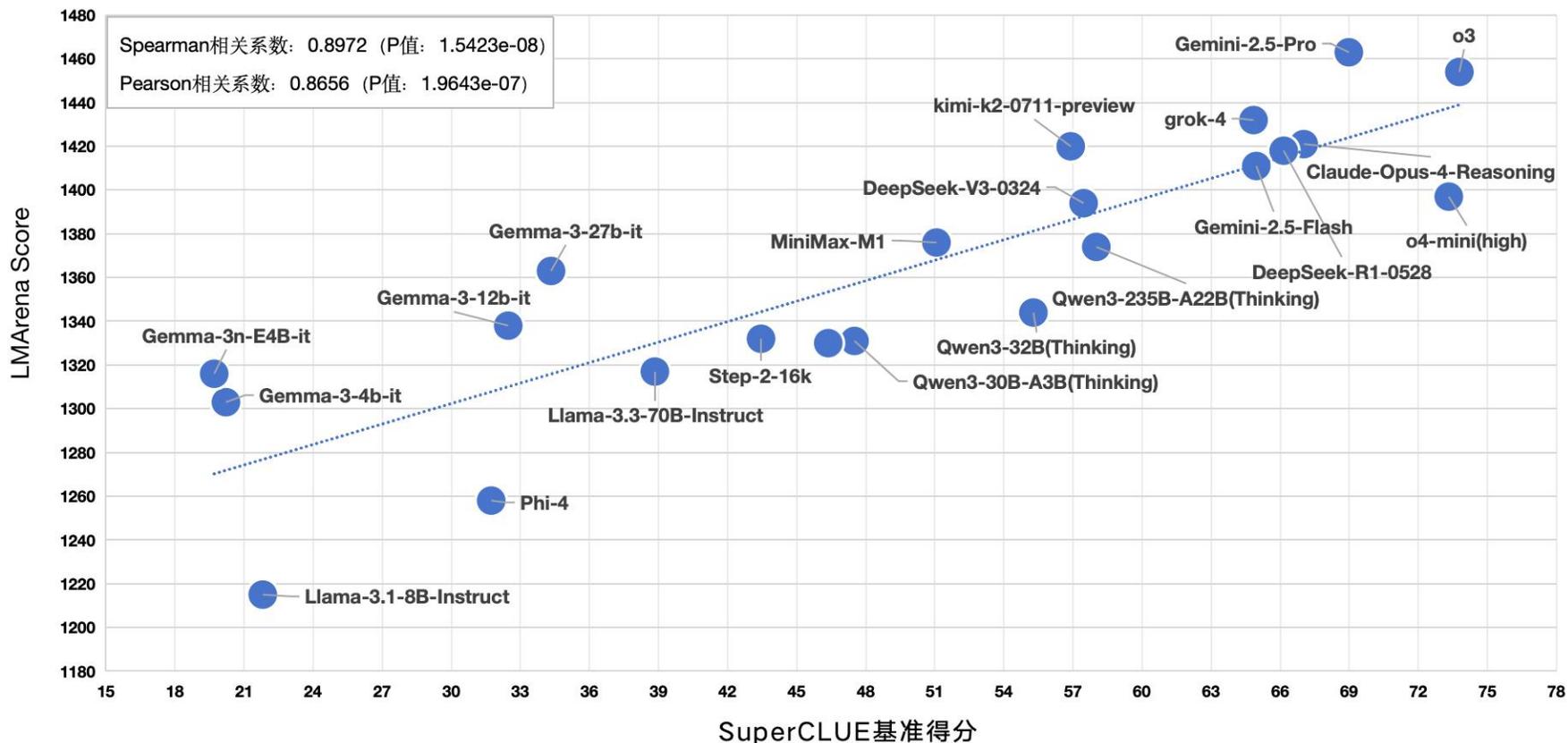
- ✓ 中成熟度指的是不同大模型能力上有一定区分度，但不会特别大。SC成熟度指数在0.5至0.8之间。
- ✓ 当前国内大模型表现出中成熟度的能力是【数学推理】、【智能体Agent】、【科学推理】和【代码生成】，还有一定优化空间。

### 3. 低成熟度能力

- ✓ 低成熟度指的是少量大模型较为擅长，很多模型无法胜任。SC成熟度指数在0.1至0.5之间。
- ✓ 当前国内大模型低成熟度的能力是【幻觉控制】和【精确指令遵循】。

LMArena是当前英文领域较为权威的大模型排行榜，由LMSYS Org开放组织构建，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。

## 评测与人类一致性验证：SuperCLUE VS LMArena



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：0.8972, P值：1.5423e-08;  
皮尔逊 (Pearson) 相关系数：0.8656, P值：1.9643e-07.

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7月31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。

# 第四部分

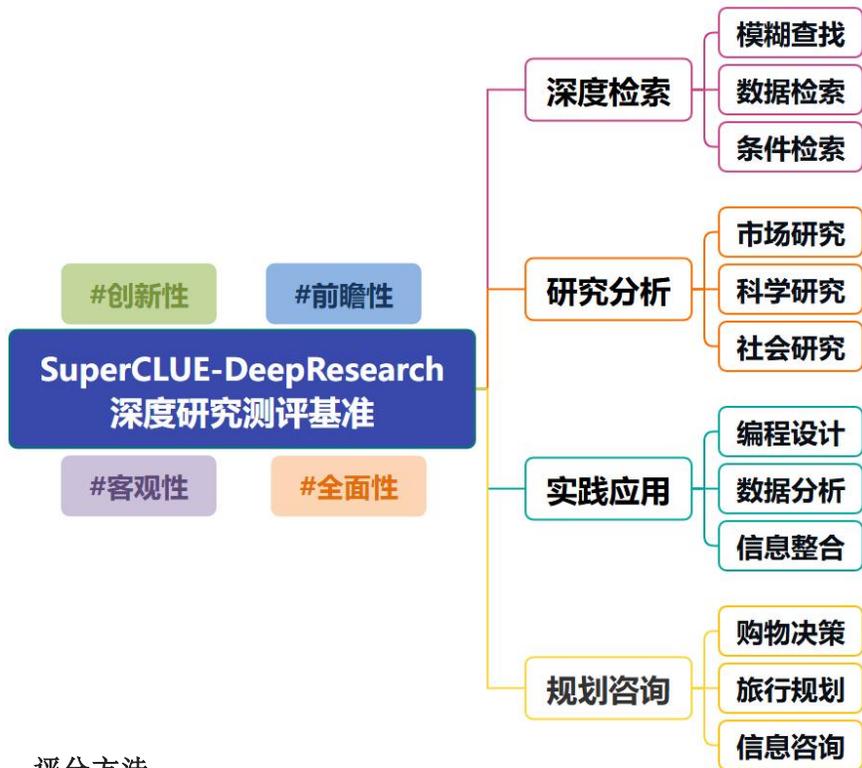
## 专项测评基准介绍

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

## SuperCLUE-DeepResearch：中文深度研究测评基准

## 测评结果分析

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。

### 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。

### 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。

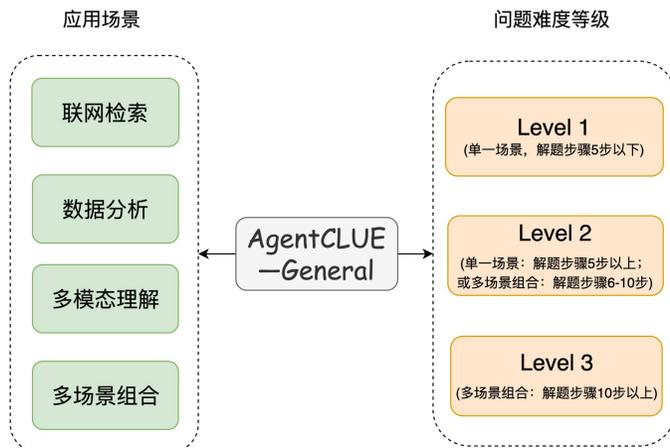
研究分析类任务平均得分高达84.54分，而深度检索类任务平均得分仅为26.26分。具体而言，当前深度研究产品在内容创作、报告输出等生成类主观任务上表现优异；但在需要深度搜索、大数据整合等复杂客观任务方面，仍存在较大提升空间。

测评详情可访问下方链接：

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

## 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难度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

### 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）。

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
1	Manus(Starter)	Monica	38.46	71.43	35.00	33.33
2	Coze(探索版)	字节跳动	32.31	42.86	30.00	33.33
3	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>

## 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。

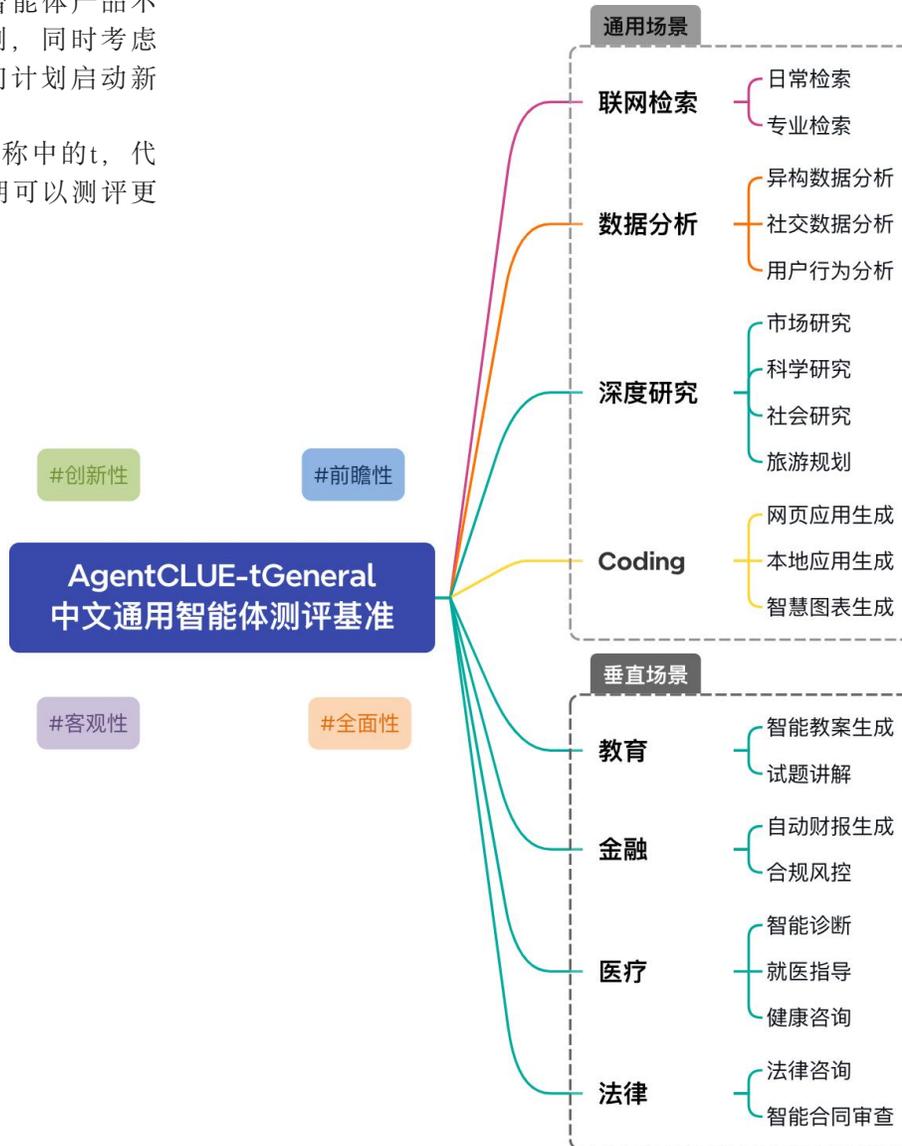
#### 3. 总分计算:

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。

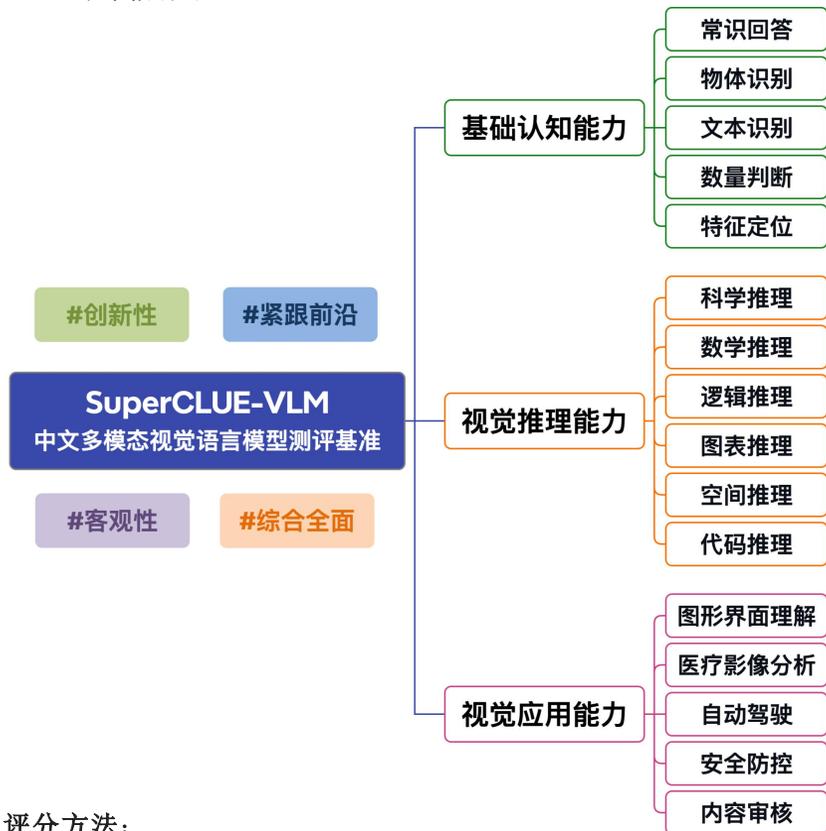
所有输入问题仅包含纯文本, 不涉及文件上传, 只要支持文本输入即可参与测评, 避免部分智能体产品因不支持上传文件、或文件上传的格式、大小限制而导致无法测评。

测评详情可访问下方链接:

<https://mp.weixin.qq.com/s/qByF2RsiL7ZeEnInPsdOXQ>

## 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	使用方式
-	Gemini-2.5-Pro	Google	74.95	74.29	76.99	73.57	API
-	o3-2025-04-16	OpenAI	69.92	68.57	70.47	70.71	API
🥇	Doubao-Seed-1.6-thinking	字节跳动	65.67	60.00	73.43	63.57	API
🥈	hunyuan-t1-vision-20250619	腾讯	63.78	60.71	68.48	62.14	API
🥉	hunyuan-turbos-vision-20250619	腾讯	60.70	61.43	62.80	57.86	API
4	SenseNova V6 Reasoner	商汤科技	58.78	57.14	63.50	55.71	API
5	QVQ-Max	阿里巴巴	55.45	54.29	53.50	58.57	API
5	step-1o-turbo-vision	阶跃星辰	54.76	58.57	44.29	61.43	API
6	Qwen2.5-VL-72B-Instruct	阿里巴巴	53.60	55.00	47.95	57.86	API
-	Claude-Sonnet-4-20250514	Anthropic	51.80	50.00	57.55	47.86	API
7	K1.5 长思考	月之暗面	51.17	48.57	61.35	43.57	官网
8	InternVL3-78B	上海AI Lab	49.80	50.71	45.11	53.57	模型
9	GLM-4V-Plus-0111	智谱AI	45.47	48.57	40.70	47.14	API
10	MiniCPM-o-2.6	面壁智能	32.66	39.29	18.69	40.00	模型

数据来源：SuperCLUE，2025年7月3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. 国外头部模型综合表现领先，国产模型紧随其后，竞争力强。**

Gemini-2.5-Pro以明显优势领跑榜单，展现了顶尖实力。同时，以豆包、混元为代表的国产模型同样表现优异，整体实力强劲，在全球范围内具备有力竞争力。

**2. 模型在基础感知上表现较好，但在高级推理和专业应用上仍是短板。**

各模型普遍擅长物体、文本识别等基础任务。但在逻辑、空间等复杂推理任务以及医疗影像分析等专业领域，表现普遍不佳，得分偏低，说明其高级推理和泛化能力亟待提升。

**3. Gemini-2.5-Pro表现全能，但部分国产模型在单项上实现超越。**

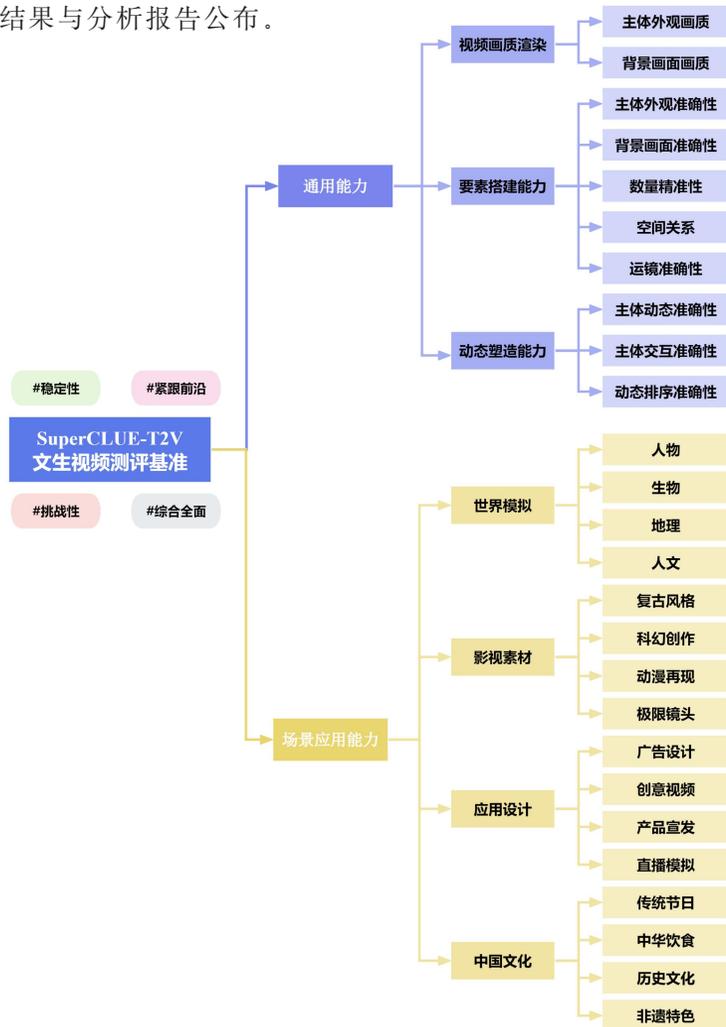
Gemini-2.5-Pro在绝大多数任务中位居前列。但部分国产模型展现出单项尖子生的特质，如Doubao-Seed-1.6-thinking在空间推理上以4分领先优势夺冠，K1.5长思考在逻辑推理上排名第一。

测评详情可访问下方链接：

<https://mp.weixin.qq.com/s/U6ARVsAnbVIYjjqy2uIX8Q>

## 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
1	Hailuo-02	MiniMax	51.67	47.09	56.25	API
2	Doubao-Seedance-1.0-pro	字节跳动	49.07	53.93	44.21	API
3	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	漪晨科技	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。海外模型仅作参考，不参与排名。

测评详情可访问下方链接：  
[https://mp.weixin.qq.com/s/YSHnlvknUxL9a9W1Q6y\\_Aw](https://mp.weixin.qq.com/s/YSHnlvknUxL9a9W1Q6y_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。

## 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 =  $N_{\text{answered}} / N_{\text{total}}$

$N_{\text{total}}$  表示该任务下的总题目数；

$N_{\text{answered}}$  表示模型成功生成视频的数目数。

## 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/uIWzEsYuCvallz6aVrDd9Q>

### 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。

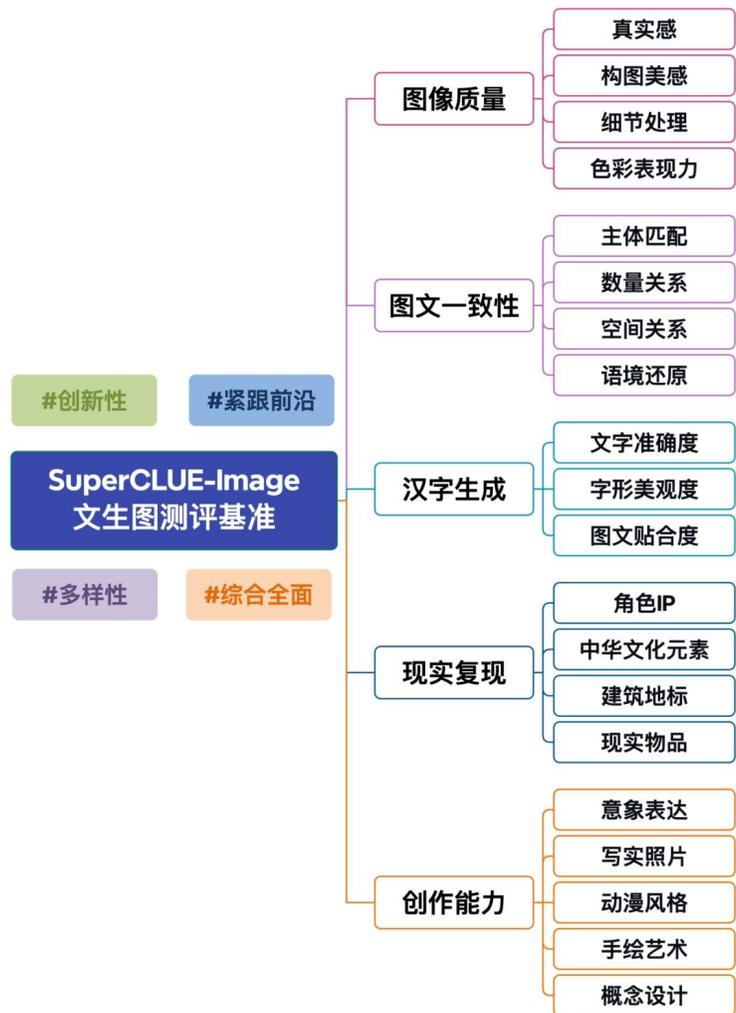
各模型在运动流畅性及内容一致性方面表现良好，多个模型突破了75分；在物理真实性方面，相比于上一次测评各模型的进步显著。但是，各模型在场景应用能力方面表现平平，尤其是在动漫风格与奇幻风格方面的表现尤为明显。

### 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。

## 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) 立足于中文场景特点，沿用以图像质量、图文一致性、汉字生成、现实复现和创作能力为核心维度的评测体系，使用全新测试集与模型版本，对多个主流图像生成大模型进行了系统性测评与榜单更新。



## 测评结果分析

总榜							
排名	模型名称	总分	图像质量得分	图文一致性得分	汉字生成得分	现实复现得分	创作能力得分
-	GPT-Image-1	68.79	76.66	60.81	66.50	69.00	70.98
-	Imagen-4.0-Ultra	64.64	80.11	64.97	38.28	69.83	70.00
🏆	Seedream 3.0	63.45	71.00	57.36	70.28	57.67	60.92
🏆	Pangu-T2I	57.90	68.39	46.58	68.95	51.95	53.65
🏆	可图 2.0	55.33	66.17	42.56	58.00	60.22	49.69
4	wanx2.1-t2i-plus	49.40	58.11	37.81	57.44	39.72	53.90
4	ERNIE-iRAG-1.0	49.25	55.16	39.42	59.95	45.94	45.80
5	CogView-4-250304	45.59	57.83	38.64	43.22	40.83	47.44
-	Recraft V3	39.81	53.22	32.67	21.83	41.95	49.38
-	Stable Image Ultra	36.91	50.55	35.61	18.61	34.16	45.64
-	FLUX.1-dev	36.81	49.28	37.22	18.94	36.88	41.74
6	HiDream-I1-Full	35.54	51.72	31.81	19.22	35.61	39.33
7	Lumina-Image-2.0	31.02	43.44	30.61	12.50	26.11	42.45
8	HunyuanDiT (2024年5月版)	28.89	42.89	17.72	8.28	32.23	43.33
9	Janus-Pro-7B	26.44	35.45	24.17	12.83	23.17	36.56

数据来源：SuperCLUE，2025年7月30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。海外模型仅作参考，不参与排名。

测评详情可访问下方链接：

<https://mp.weixin.qq.com/s/iXqQIKMPKcfmH59Prv4jfg>

### 1. 综合各项能力评分，国内模型整体占优，但国外模型占领榜首。

GPT-Image-1以68.79分的总分位列第一，展现出强劲综合实力。Seedream 3.0（字节跳动，63.45分）、Pangu-T2I（华为，57.90分）与可图2.0（快手，55.33分）等国产模型紧跟其后，表现不俗。

### 2. 国产模型在汉字生成任务中遥遥领先。

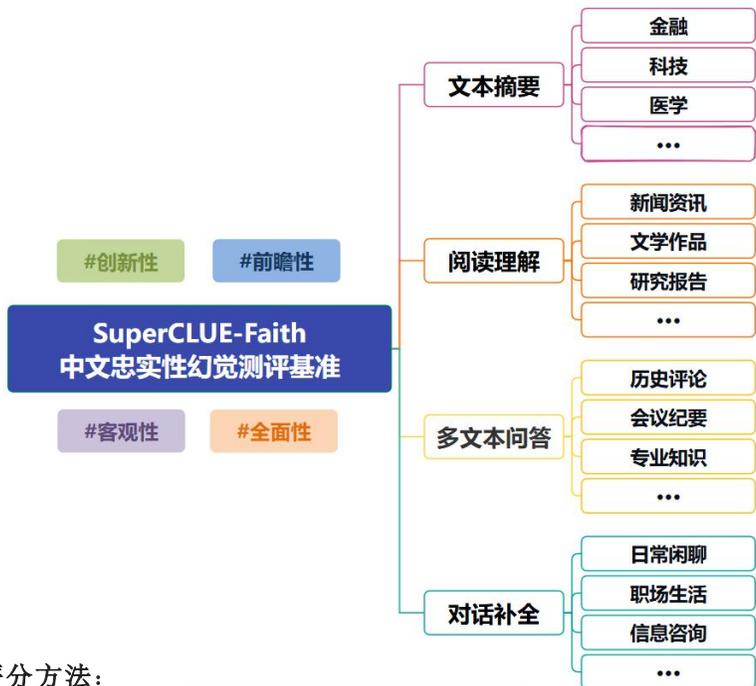
在汉字生成能力评估中，多数国际模型表现不佳，得分普遍低于40分；而Seedream 3.0（70.28分）、Pangu-T2I（68.95分）和ERNIE-iRAG-1.0（59.95分）等国产模型则表现出色，生成效果准确清晰，中文可读性强。汉字生成正成为国产文生图模型的核心竞争力之一。

### 3. 现实复现与创作能力成为拉开差距的核心维度，国外模型优势仍在。

在“现实复现”与“创作能力”两项任务中，GPT-Image-1和 Imagen-4.0-Ultra 等国际模型表现出色，得分显著高于多数国产模型，图像细节丰富、构图准确、风格自然。相比之下，国产模型虽在中文理解与汉字处理方面具备优势，但在复杂场景的真实还原、艺术风格融合等方面仍存在提升空间。

## 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

SuperCLUE-Faith  
中文忠实性幻觉测评幻觉率总榜

排名	模型名称	机构	总幻觉率 (%)	文本摘要 (%)	阅读理解 (%)	多文本问答 (%)	对话补全 (%)
1	doubao-1.5-pro-32k	字节跳动	4.11	1.65	12.68	2.05	1.45
-	gemini-2.5-pro-pre-05-06	Google	6.09	2.61	13.03	4.28	5.80
-	gemini-2.5-flash-pre-0417	Google	7.35	2.05	9.02	2.97	15.94
2	ernie-4.5-turbo-32k	百度	9.25	3.04	14.31	4.95	15.94
3	qwen-max-latest	阿里巴巴	12.35	5.33	25.76	6.16	14.49
4	DeepSeek-V3-0324	深度求索	13.83	6.16	30.39	10.39	11.59
-	ChatGPT-4o-latest	OpenAI	16.09	9.19	20.03	8.39	27.54
5	DeepSeek-R1	深度求索	21.02	16.82	39.03	9.92	20.29
6	QwQ-32B	阿里巴巴	22.70	11.16	37.54	17.98	27.54
-	o3	OpenAI	23.82	8.05	26.05	18.38	44.93
7	hunyuan-turbos-latest	腾讯	25.49	4.61	25.05	6.74	66.67
-	o3-mini(high)	OpenAI	26.14	6.21	22.03	5.48	71.01
8	Qwen3-235B-A22B	阿里巴巴	28.92	19.16	38.90	17.23	42.03
9	Qwen3-32B	阿里巴巴	30.15	16.65	41.15	15.57	49.28
-	o4-mini(high)	OpenAI	30.52	15.18	47.81	22.52	40.58
10	Qwen3-30B-A3B	阿里巴巴	32.80	11.59	27.66	9.23	82.61

数据来源：SuperCLUE，2025年5月15日。  
注：除o3外，其他测评模型均为非联网版本；本榜单将相差一分以内的模型视为并列；海外模型仅作对比，不参与排名。

### 1. 模型间的幻觉表现差异明显，极差高达28.69%。

doubao-1.5-pro-32k以95.89%的准确率和仅4.11%的幻觉率位居榜首，在四个任务上均排名第一。紧随其后的是gemini-2.5-pro-pre-05-06、gemini-2.5-flash-pre-0417和ernie-4.5-turbo-32k，其幻觉率都在10%以内，排名最后的模型幻觉率高达32.80%。

### 2. 各推理模型的幻觉比非推理模型更显著。

本次测评中推理模型的平均幻觉率为22.95%，非推理模型的平均幻觉率为13.52%，推理模型的幻觉率比非推理模型高出近9.43%，推理模型相对而言更易产生幻觉输出。

### 3. 任务开放性越高，模型的幻觉越严重。

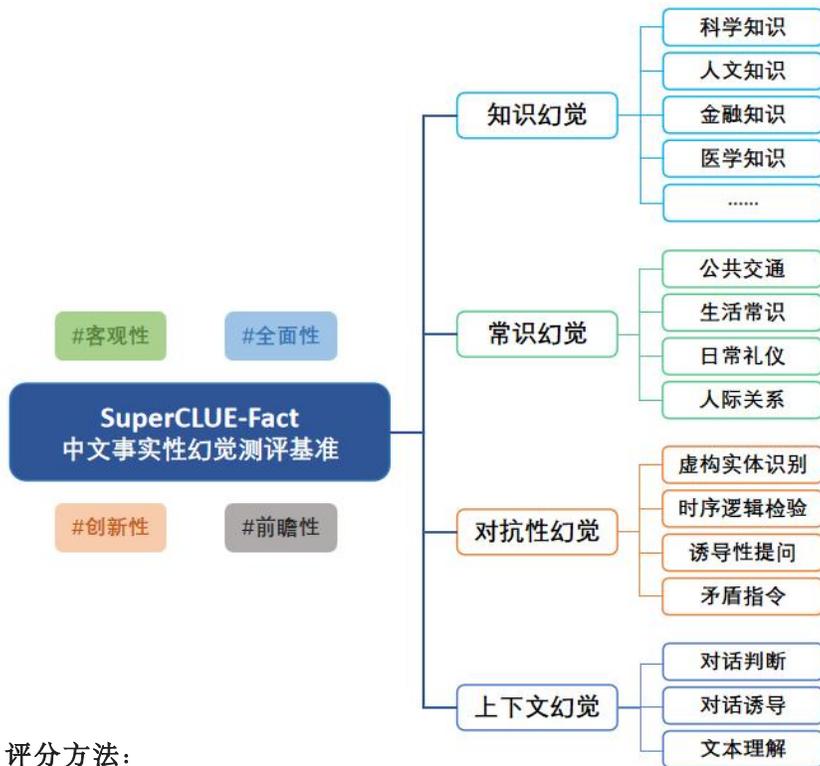
不同任务中模型幻觉率差异显著：文本摘要（8.72%）和多文本问答（10.14%）因需紧扣原文，相对而言幻觉率较低；而阅读理解（26.90%）和对话补全（33.61%）因需推理或自由生成，幻觉风险大幅增加。任务开放性与幻觉率呈现出一定的相关性。

测评详情可访问下方链接：

<https://mp.weixin.qq.com/s/0zq00-XBoUSwvmuQnNOUuw>

## 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
🥇	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
🥈	DeepSeek-V3-0324	深度求索	82.80	93.10	67.57	85.45	71.83
🥉	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。

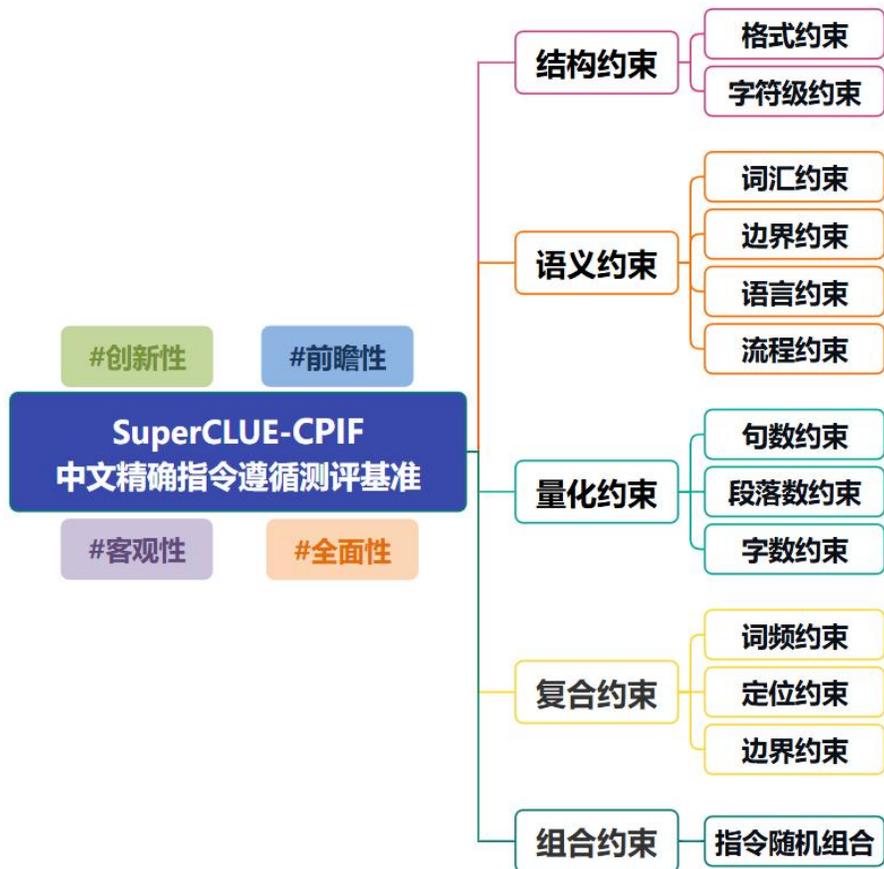
本次测评的12个模型在处理知识幻觉和对抗性幻觉方面表现相对稳健，平均得分有85分左右。然而，在常识幻觉和上下文幻觉这两类任务上，模型表现普遍较弱，平均分不足75分，差距明显。

### 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。

## 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



### 测评任务介绍：

SuperCLUE-CPIF中文精确指令遵循测评基准共有五大任务，具体介绍如下：

- 结构约束：**考察模型遵循文本外部形式、框架和字符规范的能力。包括但不限于格式约束、字符级约束。
- 语义约束：**考察模型根据指令要求，对输出的语义信息、语言类型等进行定制和控制的能力。包括但不限于词汇约束、边界约束、语言约束、流程约束。
- 量化约束：**考察模型对输出文本长度和规模进行量化控制的能力。包括但不限于句数约束、段落数约束、字数约束。
- 复合约束：**考察模型处理涵盖格式约束、语义约束以及量化约束三大任务复合指令的能力。包括但不限于词频约束、定位约束、边界约束。该任务不是对上述三大任务中的指令的拼接，而是具有独立规则的指令复合。
- 组合约束：**考察模型处理多个不同任务中复杂指令组合的能力。该任务的指令是对上述四大任务中各个指令的随机组合。

### 特点：

#### 1. 多维度的任务体系。

为确保评估的全面性，本基准构建了一个包含五大任务维度的体系，即结构约束、语义约束、量化约束、复合约束与组合约束。该体系系统性地考察了模型在不同指令类型下的表现，避免了单一任务类型带来的评估偏差。

#### 2. 阶梯式的难度梯度。

测评集被精心划分为五个难度等级 (L1-L5)。难度等级与指令约束的数量和复杂度严格对应：L1为单一基础指令，L5则包含五个层层递进的复杂指令。这种阶梯式设计能够精确地描绘出模型能力的天花板，并为模型迭代提供清晰的优化路径。

#### 3. 严格的自动化评估。

我们采用确定性的规则脚本对模型输出进行严格的二元 (0/1) 判定。任何对指令的偏离，无论多么微小，都将导致评估失败 (计0分)。这种完全自动化且无模糊地带的评估方法，排除了所有主观因素和人为干预，为所有参评模型提供了一个公平、一致且可复现的比较基线。

测评详情可访问下方链接：

[https://mp.weixin.qq.com/s/eaiWPKktrhNLMzj\\_N9va0g](https://mp.weixin.qq.com/s/eaiWPKktrhNLMzj_N9va0g)

## 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
🏆	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](https://mp.weixin.qq.com/s/21gGvflaZEMxT_QHv8u6ng)

## 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
1	腾讯元宝	腾讯	80.61	100.00	55.81	39.69
2	阶跃AI	阶跃星辰	74.49	100.00	41.86	41.10
3	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。

**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。

测评详情可访问下方链接：

[https://mp.weixin.qq.com/s/s\\_ZJP3tjxkyTVEPK\\_GucZg](https://mp.weixin.qq.com/s/s_ZJP3tjxkyTVEPK_GucZg)

## 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。

### 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%。

测评详情可访问下方链接：

<https://mp.weixin.qq.com/s/mQna2pcTeE1fnDGnLkrz6A>



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，为企业事业单位提供及时、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](mailto: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](http://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年半年度（7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。