



SuperCLUE

中文大模型综合性测评基准

中文大模型基准测评2025年3月报告

— 2025中文大模型阶段性进展3月评估

SuperCLUE团队

2025.03.18

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年度关键进展及趋势

- 2025年大模型关键进展
- 2025年值得关注的中文大模型全景图
- 2025年国内外大模型差距

三、总体测评结果与分析

- SuperCLUE通用能力测评总榜
- SuperCLUE模型象限
- SuperCLUE推理模型总榜
- SuperCLUE推理任务总榜
- SuperCLUE基础模型总榜
- 性价比区间分布
- 综合效能区间分布
- 国内大模型成熟度-SC成熟度指数
- 评测与人类一致性验证
- 开源模型榜单
- 10B级别小模型榜单
- 端侧5B级别小模型榜单

二、年度通用测评介绍

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

四、DeepSeek系列模型深度分析

- DeepSeek-R1及其蒸馏模型对比
- DeepSeek-R1和国内外头部推理模型对比
- DeepSeek-R1第三方平台联网搜索能力测试（网页端）
- DeepSeek-R1第三方平台稳定性测试（网页端、App端、API端）

报告摘要（一）



SuperCLUE模型象限（2025）

- o3-mini(high)稳居第一，推理能力全面领先

o3-mini(high) 在 SuperCLUE-3月评测中表现卓越，以 76.01 分和 84.51分稳居总榜和推理任务总榜第一，双榜均领先国内第一近5分，展现出强大的通用推理能力。

- 推理模型较基础模型优势明显，且国内较海外头部推理模型略有领先

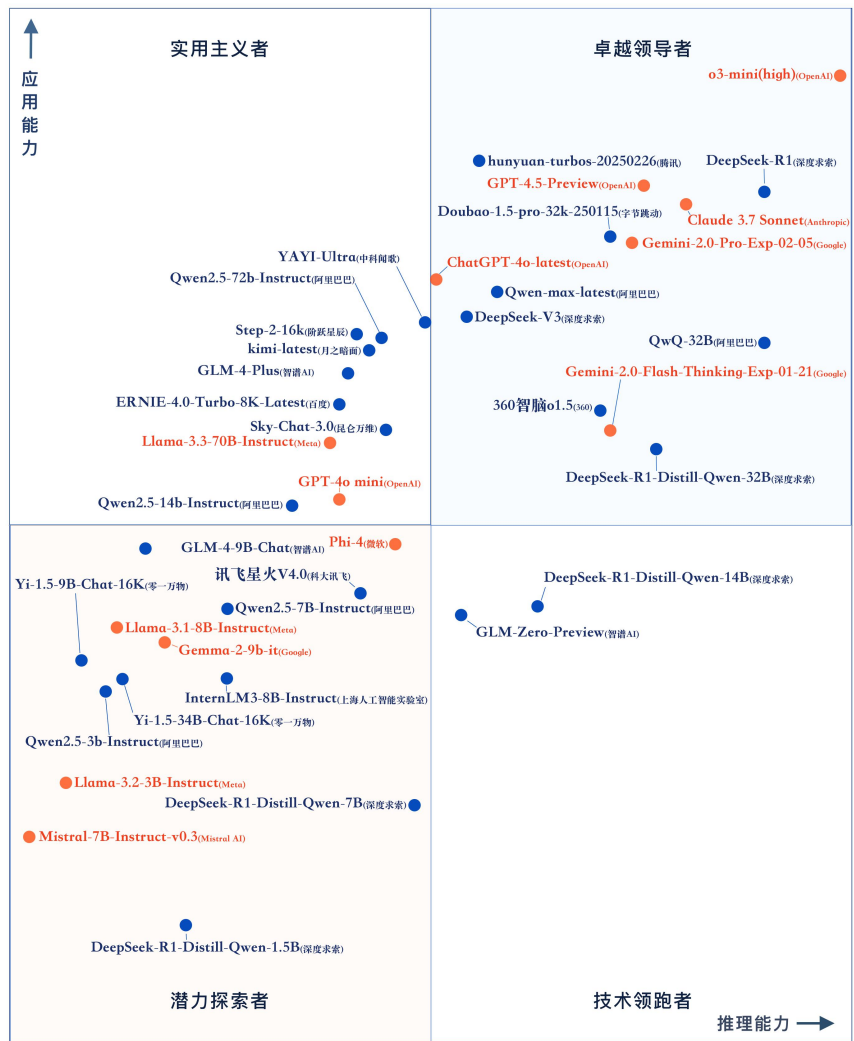
国内模型推理DeepSeek-R1在总榜上领先Claude 3.7 Sonnet近2.32分，QwQ-32B在推理任务榜单上领先Gemini-2.0-Flash-Thinking-Exp-01-21近7.65分。

- 大模型在推理速度和性价比的两级分化逐渐明显

国内推理模型DeepSeek-R1和QwQ-32B虽然推理任务上表现出色，但推理时间也随之变长；在性价比方面，DeepSeek、Qwen系列模型具有极高的性价比优势，但Claude 3.7 Sonnet由于高性能仍保持较高的API价格。

- 小参数模型表现超出预期

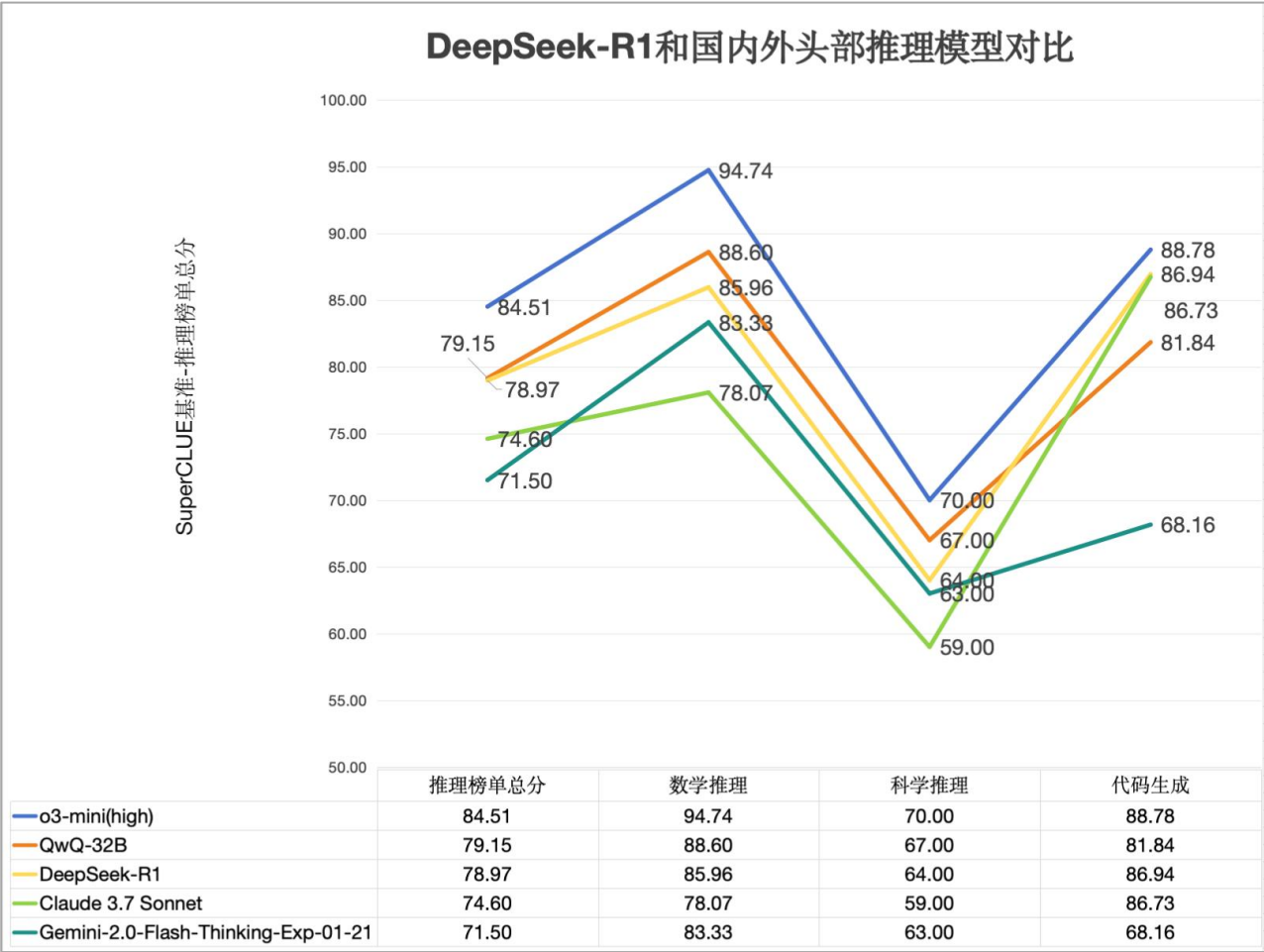
多款开源小参数量模型展现出惊人潜力。尤其是DeepSeek-R1-Distill系列，其中7B和14B版本在数学推理任务上分别取得了77.23分和79.46分的高分，超越了众多闭源大模型。



数据来源：SuperCLUE，2025年3月18日；

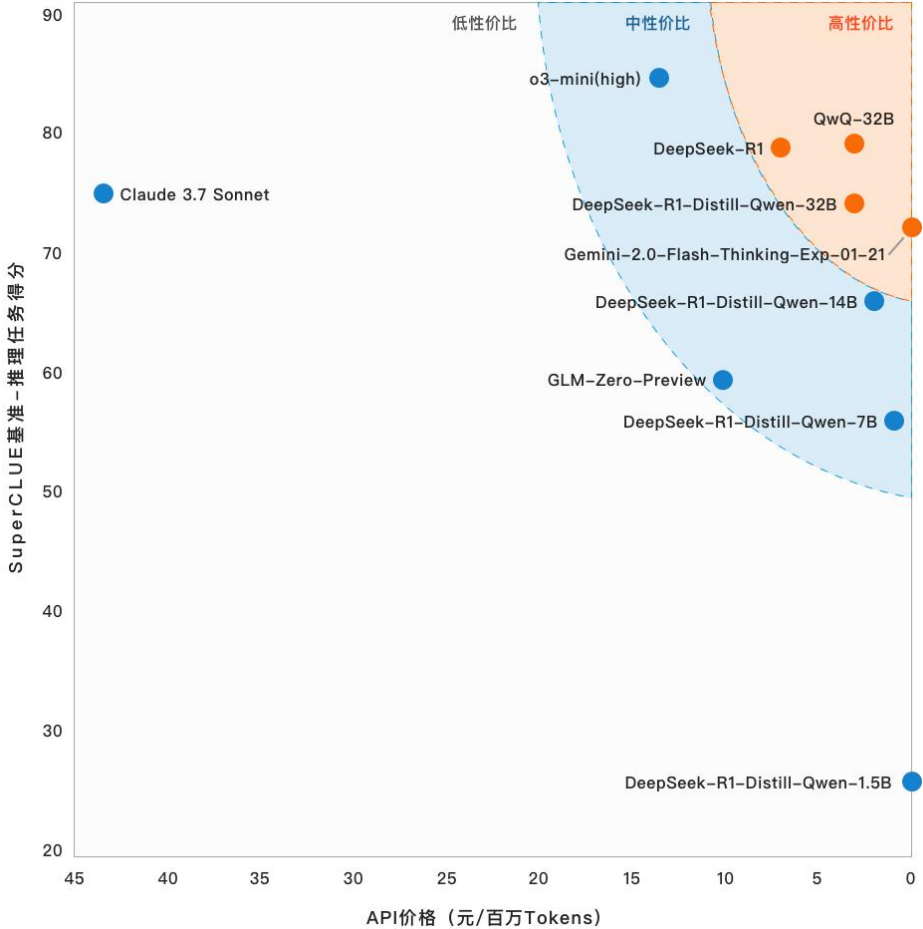
注：两个维度的组成。推理能力包含：数学推理、科学推理、代码；应用能力包括：文本理解与创作、指令遵循、Agent能力。

DeepSeek-R1和国内外推理模型能力对比



数据来源：SuperCLUE，2025年3月18日。

大模型性价比分布



数据来源：SuperCLUE，2025年3月18日；推理任务得分为推理任务总分：数学推理、科学推理和代码的平均分。开源模型如QwQ-32B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3月的标准价格（非优惠价格）。

第1部分

2025年度关键进展及趋势

1. 2025年大模型关键进展
2. 2025年值得关注的中文大模型全景图
3. 2025年国内外大模型差距

◆ 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年有了实质性的突破。具体可分为：准备期、跃进期、繁荣期和深化期。

SuperCLUE：AI大模型2025年关键进展



SuperCLUE: 2025年最值得关注的中文大模型全景图

文本

多模态

行业

通用闭源

文心一言

通义千问

腾讯混元

商汤日日新

BlueLM

360智脑

天工

MiLM

中科闻歌

紫东太初

澜舟科技

通用开源

字节豆包

Kimi.ai

MINIMAX

盘古大模型

DeepSeek

阶跃星辰

openbayes

Transn传神

推理

智谱·AI

云和声

山海

百川智能

零一万物

AndesGPT

ZTE中兴

讯飞星火

天翼AI

西湖心辰

实时交互

星火极速

智谱清言

海螺AI

豆包

文小言

通义APP

日日新

Kimi

文生视频

可灵AI

即梦AI

清影

Vidu

PixVerse

海螺AI

HiDream.ai

通义万相

视觉理解

腾讯混元

阶跃星辰

Qwen2-VL

Doubao-vision

SenseChat-Vision

海螺AI

GLM-4v

书生·万象

文生图

即梦AI

混元-DiT

快手可图

CogView

讯飞星火

meitu

通义万相

文心一格

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-01

部分领域

医疗

百度灵医

医联MedGPT

百川AI全科医生

讯飞晓医

.....

汽车

理想MindGPT

DriveGPT

极氪Kr大模型

易车大模型

.....

教育

MathGPT

作业帮

子曰

.....

金融

蚂蚁金融大模型

妙想金融大模型

轩辕大模型

HithinkGPT

.....

工业

奇智孔明AlInno-15B

华为盘古工业大模型

SMORE LrMo

羚羊工业大模型

.....

更多行业

营销: 探迹SalesGPT

文化: 阅文集团妙笔大模型

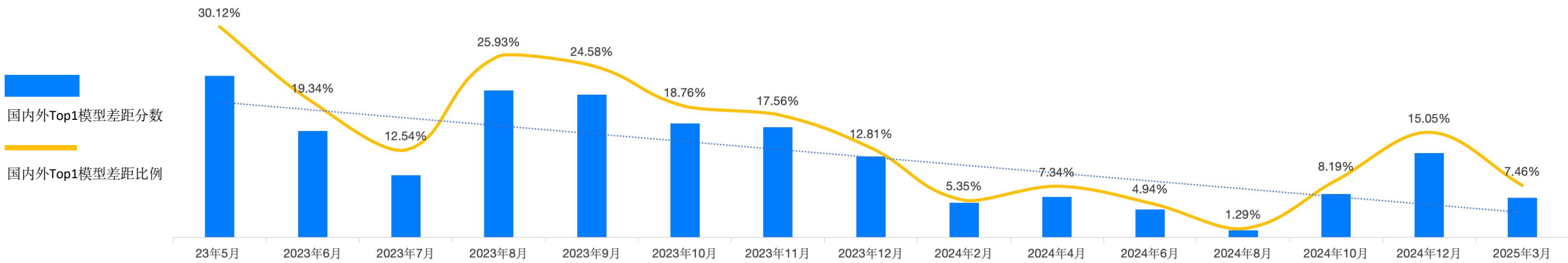
法律: Chat Law

AI4S: DP* 深势分子大模型

.....

- 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、o1、o3-mini、GPT-4.5的多个版本的迭代升级。国内模型也经历了波澜壮阔的23个月的迭代周期。但随着DeepSeek-R1的发布，差距从15.05%缩小至7.46%。

SuperCLUE基准：过去23个月国内外TOP大模型对比趋势



模型	23年5月	23年6月	23年7月	23年8月	23年9月	23年10月	23年11月	23年12月	24年2月	24年4月	24年6月	24年8月	24年10月	24年12月	25年3月
GPT最新模型 (GPT3.5、4、4-Turbo、4o、o1、o3-mini、GPT-4.5)	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	80.4	76.01
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	68.3	70.34
国内TOP2	49.52	62.58	59.35	55.70	62.61	70.42	72.88	76.54	86.77	72.58	76.00	76.24	69.00	68.3	66.38
国内TOP3	46.45	59.80	58.02	53.43	62.12	69.57	71.87	75.04	85.70	72.45	76.00	74.63	68.91	67.4	64.69

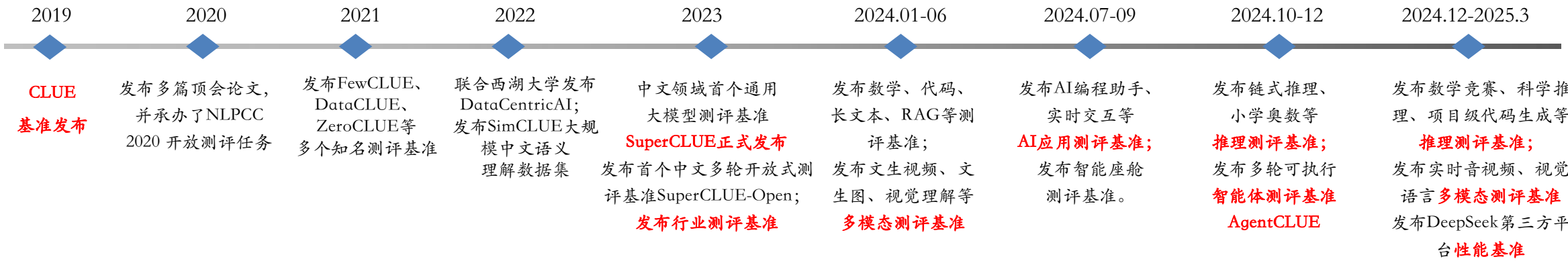
来源：SuperCLUE, 2023年5月 ~ 2025年3月，期间发布的15次大模型基准测评报告。

第2部分

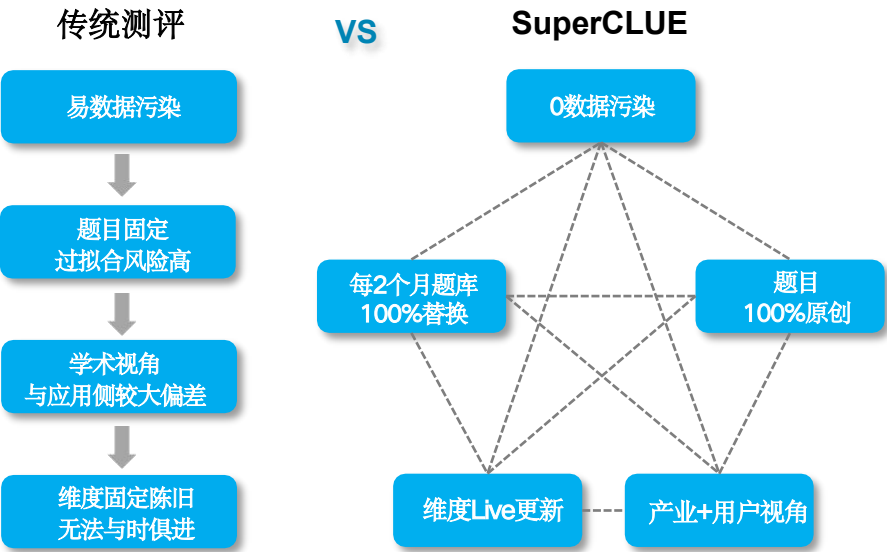
年度通用测评介绍

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

SuperCLUE是大模型时代背景下CLUE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（The Chinese Language Understanding Evaluation）**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

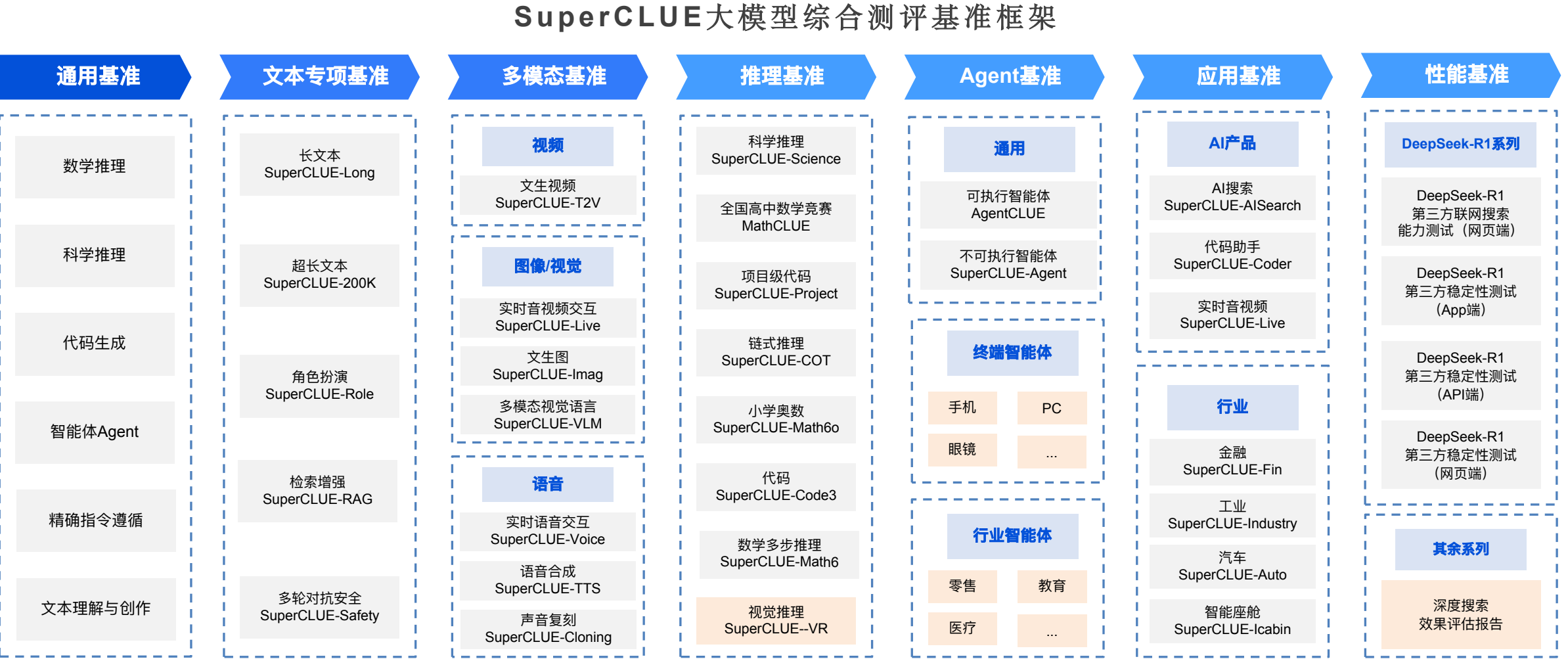
测评方式与用户交互一致

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、文本专项系列基准、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。



已发布

即将发布

注：通用基准介绍可在报告中查看，其余可点击可跳转最新的测评文章

本次2025年3月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1509道多轮简答题。

【SuperCLUE通用数据集】分为数学推理、科学推理、代码生成、智能体Agent、精确指令遵循、文本理解与创作；

【SuperCLUE评价方式】分为基于人工校验参考答案的评估（0-1得分）、基于代码单元测试的评估（0-1得分）、结合任务完成与否、系统状态比对的评估（0-1得分）、基于规则脚本的评估（0-1得分）、人工校验参考答案的、多维度评价标准的评估。

SuperCLUE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。包括但不限于平面几何、代数、初等数论、组合问题等竞赛级的数学推理数据集。

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：主要考察模型在处理编程任务时理解和生成代码的能力。HumanEval的中文升级版，涵盖数据结构、基础算法、数学问题、数据科学等多种类型的代码数据集。

评价方式：基于代码单元测试的评估（0-1得分）

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，在多轮对话中调用工具完成任务的能力。包括两大任务类型：常规单轮对话和常规多轮对话。

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。

评价方式：基于规则脚本的评估（0-1得分）

6.文本理解与创作

介绍：主要考察模型在处理文本相关任务时的理解和创作能力。包括但不限于文本摘要、阅读理解、指代消解、长文本等基础语义理解和生成创作数据集。

评价方式：人工校验参考答案的、多维度评价标准的评估

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、初等数论、概率与统计、三角学、数列与极限数据集。

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 x, y 为实数，且满足 $\sin x + \sin y = \sin(x + y)$. 求 $\cos x + \cos y$ 的所有可能取值。

模型回答： 本题可先利用三角函数两角和公式对已知条件进行化简，然后通过设未知数的方式求解 $\cos x + \cos y$ 的取值范围。

- 步骤一：对 $\sin x + \sin y = \sin(x + y)$ 进行化简
根据两角和公式 $\sin(A + B) = \sin A \cos B + \cos A \sin B$ ，将 $\sin x + \sin y = \sin(x + y)$ 展开可得：

$$\sin x + \sin y = \sin x \cos y + \cos x \sin y$$

<篇幅有限省略>

- 当 $\sin(\frac{x+y}{2}) = 0$ 时：
此时 $\frac{x+y}{2} = n\pi, n \in \mathbb{Z}$ ，即 $x + y = 2n\pi, n \in \mathbb{Z}$ ，那么 $y = 2n\pi - x$ 。
将 $y = 2n\pi - x$ 代入 $\cos x + \cos y$ 可得：

$$\cos x + \cos y = \cos x + \cos(2n\pi - x)$$

根据诱导公式 $\cos(2k\pi - \alpha) = \cos \alpha$ ，可得 $\cos x + \cos(2n\pi - x) = \cos x + \cos x = 2 \cos x$ 。

因为 $-1 \leq \cos x \leq 1$ ，所以 $-2 \leq 2 \cos x \leq 2$ ，即 $-2 \leq \cos x + \cos y \leq 2$ 。

综合以上三种情况， $\cos x + \cos y$ 的取值范围是 $[-2, 2]$ 。

所以，答案为 $[-2, 2]$ 。

评价打分：

问题及改进的反馈意见： 模型的回答逻辑清晰，步骤详尽，且最终结论与参考答案一致。唯一的改进建议是可以进一步简化部分推导过程，以提高可读性。

打分与评价开始：

1. 基于标准1的<结论准确性>：

模型的最终结论为 $\boxed{[-2, 2]}$ ，与参考答案 $[-2, 2]$ 一致。模型的推导过程正确，结论准确。

基于标准1的得分：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：一艘航天器以恒定加速度 $a = 2.5 \text{ m/s}^2$ （在其自身参考系中测量）从地面出发，沿直线运动，目标是相距 $D = 1.2 \times 10^{16} \text{ m}$ 远的一颗行星。假设航天器不断以相同加速度加速至中点，然后在到达一半距离后立即对称地切换为以相同大小的减速度（相对于出发点参考系加速度数值近似恒定；若需严格分析需用狭义相对论处理），令其在抵达目标时恰好速度降至零。试计算：在地面观测者的参考系中，航天器完成此次旅行所需的时间约为多少？（以年为单位）

模型回答： 计算步骤：

1. 计算加速阶段的时间：

在加速阶段，航天器从静止开始，以恒定加速度 a 加速到中点。我们可以使用运动学公式：

$$s = \frac{1}{2}at^2$$

其中 $s = 6 \times 10^{15} \text{ m}$, $a = 2.5 \text{ m/s}^2$ 。解这个方程可以得到时间 t ：

$$6 \times 10^{15} = \frac{1}{2} \times 2.5 \times t^2$$
$$t^2 = \frac{6 \times 10^{15} \times 2}{2.5} = 4.8 \times 10^{15}$$
$$t = \sqrt{4.8 \times 10^{15}} \approx 6.928 \times 10^7 \text{ s}$$

<篇幅有限省略>

3. 将时间转换为年：

1年大约有 3.154×10^7 秒，因此：

$$T \approx \frac{1.3856 \times 10^8}{3.154 \times 10^7} \approx 4.39 \text{ 年}$$

最终答案：

4.39 年

评价打分：

问题及改进的反馈意见： AI助手的回答清晰地分解了问题，使用了正确的公式和步骤来计算航天器的旅行时间，并且最终结果与参考答案一致。唯一的不足是，虽然答案的精度足够，但可以进一步明确地说明其与参考答案的近似关系。

打分与评价开始：

1. 基于标准1的<结论准确性>；

AI助手的最终答案为 $\boxed{4.39 \text{ 年}}$ ，与参考答案“约4.4年”在数值上本质一致，符合小数与近似值的转换规则，因此回答是正确的。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

主要考察模型在处理编程任务时理解和生成代码的能力。**HumanEval**的中文升级版，涵盖数据结构、基础算法、数学问题、数据科学等多种类型的代码数据集。

测评方法

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。

示例-代码生成

问题：

```
from typing import List
```

```
def task_func(matrix: List[List[int]]) -> List[List[int]]:
```

```
    """
```

给定一个 $m \times n$ 的矩阵，表示每个单元格的高度，该矩阵代表一个大陆的高度图。

“北冰洋”接触矩阵的上边，“印度洋”接触矩阵的下边。

水流只能通过四个方向（上、下、左、右）从一个单元流向高度更低或相等的相邻单元。

请找出所有可以同时流向北冰洋和印度洋的单元格坐标。

要求：

- 返回的坐标列表应按唯一顺序排列（按从左到右、从上到下的顺序排列）。

```
    """
```

测试功能函数：

```
assert candidate([[1, 2, 2, 3, 5], [3, 2, 3, 4, 4], [2, 4, 5, 3, 1], [6, 7, 1, 4, 5], [5, 1, 1, 2, 4]])==[[2, 2], [3, 1]]
```

计算模型得分：测试通过，得1分

模型回答：

```
from typing import List
```

```
def task_func(matrix: List[List[int]]) -> List[List[int]]:
```

```
    if not matrix or not matrix[0]:
```

```
        return []
```

```
    m, n = len(matrix), len(matrix[0])
```

```
    # 定义可以流向北冰洋和印度洋的集合
```

```
    can_flow_to_north = set()
```

```
    can_flow_to_south = set()    return False
```

```
    if ju[2] <= 0:
```

```
        return False
```

```
    for i in range(3):
```

```
        if ju[i] == 0:
```

```
            continue
```

```
        max_amount = math.floor(ju[i] / 2)
```

```
        res = math.ceil(target / max_amount)
```

```
        if res > 0:
```

```
            return True
```

```
    return False
```

<篇幅有限省略>

```
    # 找出可以同时流向北冰洋和印度洋的单元格
```

```
    result = []
```

```
    for i in range(m):
```

```
        for j in range(n):
```

```
            if (i, j) in can_flow_to_north and (i, j) in can_flow_to_south:
```

```
                result.append([i, j])
```

```
    return result
```

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 多轮对话中调用工具完成任务的能力。包括两大任务类型：常规单轮对话和常规多轮对话。

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8个API，本题用到两个API。

- 博客API：该工具属于博客API，该工具为在博客上发布博文，转发，评论和关注用户等提供了核心功能。
- 旅游出行：该工具属于旅行系统，该系统允许用户预订航班，管理信用卡并查看预算信息。

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>

【第一轮问题】：我想要发布一条博文，内容为“3月10日，前往拉萨旅游”，然后使用标签“#欢乐”和“#特种兵旅游”。

【模型回复1】：`[post_tweet(content="3月10日，前往拉萨旅游", tags=["#欢乐", "#特种兵旅游"])]`

【模型得到的反馈】：第1个函数调用：`post_tweet(content="3月10日，前往拉萨旅游", tags=["#欢乐", "#特种兵旅游"])`，执行返回结果为：`{"id": 1, "username": "boke123", "content": "3月10日，前往拉萨旅游", "tags": ["#欢乐", "#特种兵旅游"], "mentions": []}`

【模型回复2】：博文已成功发布，内容为“3月10日，前往拉萨旅游”，并使用了标签“#欢乐”和“#特种兵旅游”。

..... <篇幅有限省略其余对话>

【第二轮问题】：我想要用 card_id 为 card001，access_token 为 abc123xyz123 的信用卡购买2025年3月10日从杭州飞拉萨的经济舱机票。

【模型回复1】：`[get_nearest_airport_by_city(location="杭州"),get_nearest_airport_by_city(location="拉萨")]`

【模型得到的反馈1】：第1个函数调用：`get_nearest_airport_by_city(location="杭州")`，执行返回结果为：`{"nearest_airport": "HGH"}`第2个函数调用：`get_nearest_airport_by_city(location="拉萨")`，执行返回结果为：`{"nearest_airport": "LXA"}`

【模型回复2】：`[get_flight_cost(travel_from="HGH", travel_to="LXA", travel_date="2025-03-10", travel_class="经济舱")]`

【模型得到的反馈2】：第1个函数调用：`get_flight_cost(travel_from="HGH", travel_to="LXA", travel_date="2025-03-10", travel_class="经济舱")`，执行返回结果为：`{"travel_cost_list": [460.0]}`

【模型回复3】：`[book_flight(access_token="abc123xyz123", card_id="card001", travel_date="2025-03-10", travel_from="HGH", travel_to="LXA", travel_class="经济舱", travel_cost=460.0)]`

【模型得到的反馈3】：第1个函数调用：`book_flight(access_token="abc123xyz123", card_id="card001", travel_date="2025-03-10", travel_from="HGH", travel_to="LXA", travel_class="经济舱", travel_cost=460.0)`，执行返回结果为：`{"booking_id": "3426812", "transaction_id": "45451592", "booking_status": true, "booking_history": {}}`

【模型回复4】：您的航班已成功预订。您将于2025年3月10日从杭州（HGH）飞往拉萨（LXA），乘坐经济舱。预订ID为3426812，交易ID为45451592。祝您旅途愉快！

（本题共3轮对话，此处省略第三轮对话，并由于篇幅限制，只展示关键信息。本次答对3轮对话，记3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。

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：请撰写一篇关于“书籍与旅行”的短文，文章中必须使用“远方”、“故事”和“灯火”这三个词语。文章中每个句子（用中文句号或问号或感叹号分隔）的最后一个字必须以“也”字结尾。请确保文章只能有4个段落（不要出现标题），每个段落之间用“~~”符号分隔，其中第2段的第8个汉字必须是“桥”字。

回答：

远方的书籍，会带来许多故事的**魅力**，人们可以通过阅读这些**故事**，体验到**远方**的文化和历史，也可以从中获得深深的**体验**。

~~

当我们旅行到**远方**，看到的**灯火**会让我们感到一种深深的感伤，灯火的闪烁，会让我们感到一种温暖的感觉，也会让我们感受到一种被怀念的**感觉**。

~~

在这些故事中，我们会看到许多桥，连接着不同的人 and 不同的事物，灯火的闪烁，会让我们感到一种被连接的感觉，也会让我们感受到一种被融合的**感觉**。

~~

远方的书籍和灯火，会让我们感到一种被 **transported** 的感觉，人们可以通过阅读这些故事和看到灯火的闪烁，体验到远方的文化和历史，也可以从中获得深深的**体验**。

评价打分：

未遵循指令内容：文章中每个句子（用中文句号或问号或感叹号分隔）的最后一个字没有以“也”字结尾；第2段的第8个汉字不是“桥”字。

测评结果：**测试不通过，不得分**

SuperCLUE-文本理解与创作评测数据集

文本理解与创作

主要考察模型在处理文本相关任务时的理解和创作能力。包括但不限于文本摘要、阅读理解、指代消解、长文本等基础语义理解和生成创作数据集。

测评方法

评测流程：

我们采用高级AI模型（例如GPT-4o）来评估不同任务的表现，根据设定的评价标准进行打分（1-5分）。

以【生成与创作】任务为例，设定的评价标准包括：符合基本要求、语言与逻辑连贯性、推理链条完整度。

假设在一个具体的题目中，模型A的得分如下：符合基本要求3分、语言与逻辑连贯性4分、推理链条完整度5分。首先，我们计算这一题目的平均得分：

$$\frac{3+4+5}{3} = 4$$

因此，该题目的具体得分为4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，每个任务的得分都能清晰地反映模型在不同标准下的表现。

示例-文本理解与创作

问题：你是一家人工智能芯片研发公司的研发部门经理，需要撰写一份《AI芯片性能优化项目延期情况说明》。该项目原计划在6个月内完成新一代7nm制程AI芯片的性能提升和功耗优化，但目前已超期2个月。请向技术副总裁提交一份正式的项目延期情况说明，详细解释延期原因（包括技术瓶颈、供应链风险等），并提出具体的改进措施。说明文件需要采用技术管理报告的规范格式，包含数据分析和里程碑调整计划。

模型回答： AI芯片性能优化项目延期情况说明 <篇幅有限省略部分内容>

提交人：研发部门经理

提交日期：XXXX年XX月XX日

一、项目基本情况

项目名称：AI芯片性能优化项目

原计划周期：6个月（XX年XX月-XX年XX月）

二、延期原因分析

1. 技术瓶颈

- 电路优化复杂度提升：在低功耗模式下优化计算单元的性能，发现寄生参数建模误差较大，导致实际能效比达不到预期目标（优化目标15%，当前仅提升9%）。

四、调整后的里程碑规划

阶段	原计划完成时间	调整后完成时间	备注
计算核心优化	XX月	XX+2月	优化建模精度，提高能效比
存储架构优化	XX月	XX+1月	采用动态缓存策略
EDA工具验证	XX月	XX+1月	提前验证，减少返工

五、结论与总结

当前AI芯片性能优化项目由于技术瓶颈和供应链风险已超期2个月，但团队已采取一系列优化措施，包括技术调整、供应链协调和测试策略优化，以确保项目在合理期限内推进。

预计在调整后的周期内完成各阶段的优化与流片，最终目标仍保持芯片性能提升15%和存储访问延迟优化

评价打分：

1. 基于标准1的符合基本要求：模型的回答完全满足了用户问题的所有基本要求，针对延期原因、改进措施和调整后的计划都予以了详细说明，且遵循了技术管理报告的规范格式。回答直接且完全覆盖了用户需求，没有遗漏。得分：5分

2. 基于标准2的创造性：模型体现了较高的技术分析能力且提供了详实的具体解决方案，如动态缓存方法、寄生参数建模优化等，但创新性略显不足。尽管内容详细，但大部分回复属于常规应对措施，没有显著超出一般技术报告的期望范围（如缺乏对后续类似问题的预防机制分析）。得分：4.5分

3. 基于标准3的语言与逻辑连贯性：回答语言流畅，逻辑清晰，段落布局井然有序，各部分彼此间的过渡与衔接自然。数据和技术细节也支撑了逻辑推演，整体呈现了较强的可读性和说服力，并符合正式报告的规范要求。缺点是部分段落的冗长可能削弱阅读效率。得分：4.5分

4. 基于标准4的事实正确性：回答中所引用的技术性数据、对EDA工具的描述、以及流程中的常见瓶颈因素，均与当前人工智能芯片开发领域实际情况一致。从总体上看，模型展示出对技术内容的准确把握，没有明显误导性或错误性信息。得分：5分

综合以上，AI助手的回答的综合得分（平均分）为[[4.75]]

本次测评数据选取了**SuperCLUE-3月**测评结果，模型选取了国内外有代表性的**45个大模型在3月份的版本**。

模型	机构	简介	模型	机构	简介
1.o3-mini(high)	OpenAI	OpenAI在2025年1月31日发布的推理模型，使用官方API： o3-mini-2025-01-31	24.Llama-3.3-70B-Instruct	Meta	Meta发布的Llama3.3版本70B模型，调用的together.ai的接口
2.DeepSeek-R1	深度求索	深度求索在2025年1月20日推出的DeepSeek-R1,使用官网API： deepseek-reasoner	25.Phi-4	微软	官方开源的Phi-4版本。对应huggingface 仓库名称： microsoft/phi-4
3.Claude 3.7 Sonnet	Anthropic	Antropic在2025年2月25日发布的版本，使用POE： Claude-3.7-Sonnet-Reasoning	26.GPT-4o mini	OpenAI	OpenAI在2024年7月18日发布的版本，使用官方API： gpt-4o-mini-2024-07-18
4.GPT-4.5-Preview	OpenAI	OpenAI在2025年2月27日发布的版本，使用POE： GPT-4.5-Preview	27.讯飞星火V4.0	科大讯飞	科大讯飞发布的API版本： Spark4.0 Ultra
5.QwQ-32B	阿里巴巴	基于Qwen2.5-32B模型训练的QwQ推理模型，使用阿里云公开的API： qwq-32b	28.Qwen2.5-14b-Instruct	阿里巴巴	Qwen2.5开源系列14B模型，使用阿里云公开的API： qwen2.5-14b-instruct
6.Gemini-2.0-Pro-Exp-02-05	Google	官方在2025年2月5日发布的Gemini 2.0 Pro模型的实验版本。	29.DeepSeek-R1-Distill-Qwen-7B	深度求索	基于Qwen2.5-Math-7B的蒸馏模型，使用阿里云API： deepseek-r1-distill-qwen-7b
7.Doubao-1.5-pro-32k-250115	字节跳动	官方公开的Doubao-1.5-pro-32k，使用方式为API： Doubao-1.5-pro-32k-250115	30.Qwen2.5-7B-Instruct	阿里巴巴	Qwen2.5开源系列7B模型。使用阿里云公开的API接口： qwen2.5-7b-instruct
8.hunyuan-turbos-20250226	腾讯	官方在2025年2月26日发布的旗舰模型，使用官方API： hunyuan-turbos-20250226	31.InternLM3-8B-Instruct	上海AI实验室	上海人工智能实验室发布的InternLM3-8B-Instruct开源版本
9.DeepSeek-R1-Distill-Qwen-32B	深度求索	基于Qwen2.5-32B的蒸馏模型，使用阿里云API： deepseek-r1-distill-qwen-32b	32.GLM-4-9B-Chat	智谱AI	官方开源的GLM-4-9B-Chat版本
10.Qwen-max-latest	阿里巴巴	通义千问系列效果最好的模型，使用阿里云公开的API： qwen-max-latest	33.Gemma-2-9b-it	Google	官方开源的gemma-2-9b-it版本
11.Gemini-2.0-Flash-Thinking-Exp-01-21	Google	官方在2025年1月21日发布的Gemini 2.0 Flash Thinking模型的实验版本	34.Llama-3.1-8B-Instruct	Meta	Meta发布的Llama3.1版本8B开源模型，调用的together.ai的接口
12.360智脑o1.5	360	官方提供的小范围内测版本，使用方式为API	35.Yi-1.5-34B-Chat-16K	零一万物	官方开源的Yi-1.5-34B-Chat-16K版本
13.DeepSeek-V3	深度求索	深度求索在2024年12月26日发布的版本，使用官方API： deepseek-chat	36.Qwen2.5-3b-Instruct	阿里巴巴	Qwen2.5开源系列3B模型。使用阿里云公开的API接口： qwen2.5-3b-instruct
14.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	37.Yi-1.5-9B-Chat-16K	零一万物	官方开源的Yi-1.5-9B-Chat-16K版本
15.YAYI-Ultra	中科闻歌	官方提供的小范围内测版本，使用方式为API	38.DeepSeek-R1-Distill-Qwen-1.5B	深度求索	基于Qwen2.5-Math-1.5B的蒸馏模型，使用阿里云API:deepseek-r1-distill-qwen-1.5b
16.Qwen2.5-72B-Instruct	阿里巴巴	Qwen2.5开源系列72B模型，使用阿里云公开的API接口： qwen2.5-72b-instruct	39.Llama-3.2-3B-Instruct	Meta	Meta发布的Llama3.2版本3B开源模型，调用的together.ai的接口
18.kimi-latest	月之暗面	与Kimi 智能助手产品使用的大模型同版本，使用API： kimi-latest	40.Mistral-7B-Instruct-v0.3	Mistral AI	官方开源的Mistral-7B-Instruct-v0.3版本
18.Step-2-16k	阶跃星辰	官方公开发布的API版本： step-2-16k	41.Grok-3	X.AI	X.AI在2025年2月19日推出的模型版本，使用官网网页
19.DeepSeek-R1-Distill-Qwen-14B	深度求索	基于Qwen2.5-14B的蒸馏模型。使用阿里云API： deepseek-r1-distill-qwen-14b	42.Step R-mini	阶跃星辰	阶跃星辰推出的推理模型，使用官网网页，模型选择“Step R-mini”
20.Sky-Chat-3.0	昆仑万维	昆仑万维发布的千亿级别 MOE 模型，使用官方的API接口。	43.Kimi-K1.5	月之暗面	月之暗面推出的推理模型，使用官网网页，开启“K1.5长思考”模式
21.GLM-4-Plus	智谱AI	智谱自研的GLM系列大模型最新版本，使用官方API： GLM-4-Plus	44.QwQ-Max-Preview	阿里巴巴	基于Qwen2.5-Max训练的推理模型，使用网页
22.ERNIE-4.0-Turbo-8K-Latest	百度	百度文心系列中效果最强大的模型，使用官方API： ERNIE-4.0-Turbo-8K-Latest	45.Skywork o1 Preview	昆仑万维	昆仑万维推出的o1版推理模型，使用官网网页
23.GLM-Zero-Preview	智谱AI	智谱在2024年12月31日发布的推理模型，使用官方API： GLM-Zero-Preview	/	/	/

第3部分

总体测评结果与分析

1. SuperCLUE通用能力测评总榜
2. SuperCLUE模型象限
3. SuperCLUE推理模型总榜
4. SuperCLUE推理任务总榜

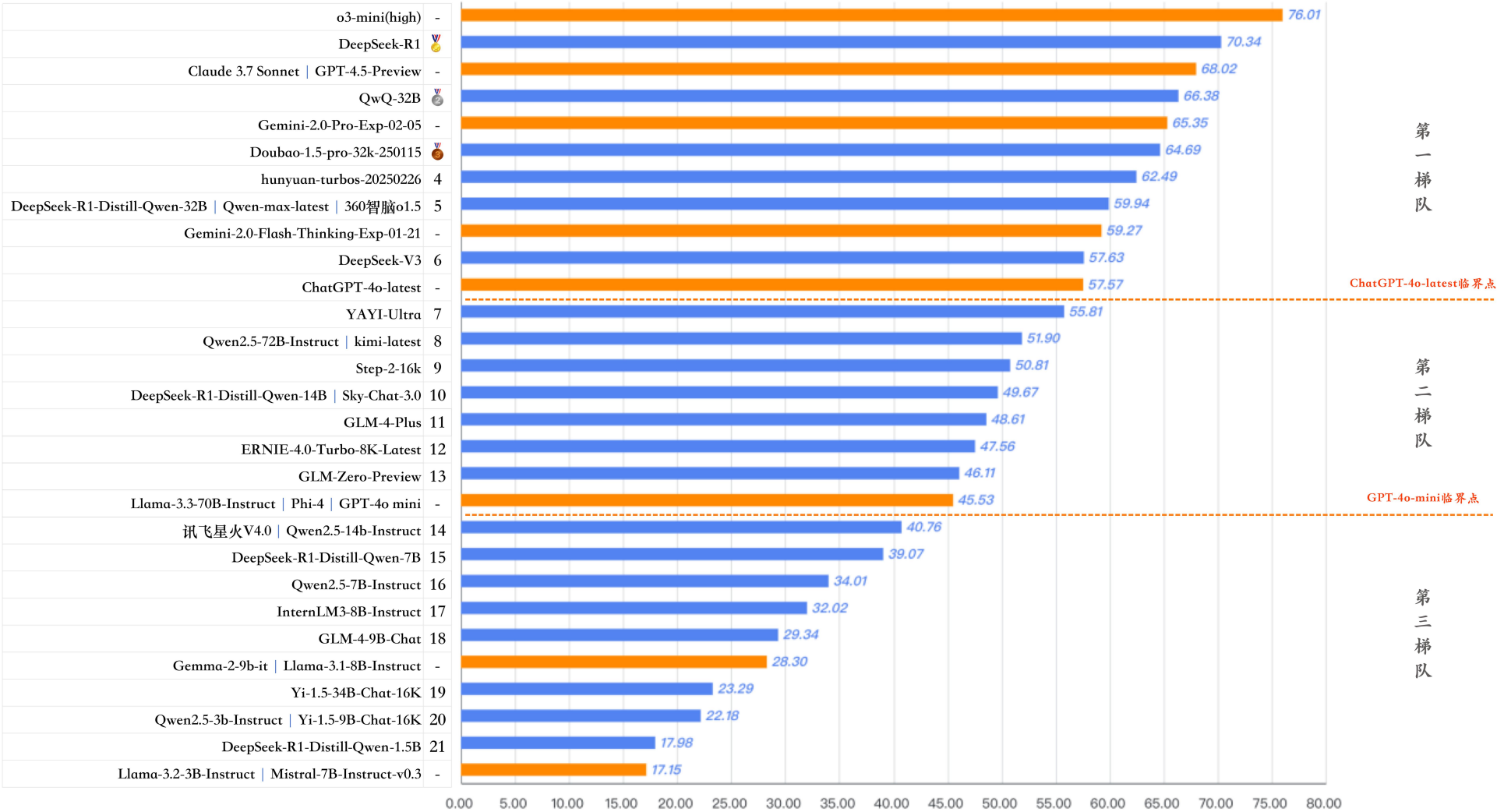
5. SuperCLUE基础模型总榜
6. 性价比区间分布
7. 综合效能区间分布
8. 国内大模型成熟度-SC成熟度指数

9. 评测与人类一致性验证
10. 开源模型榜单
11. 10B级别小模型榜单
12. 端侧5B级别小模型榜单

国内外通用大模型SuperCLUE基准榜单

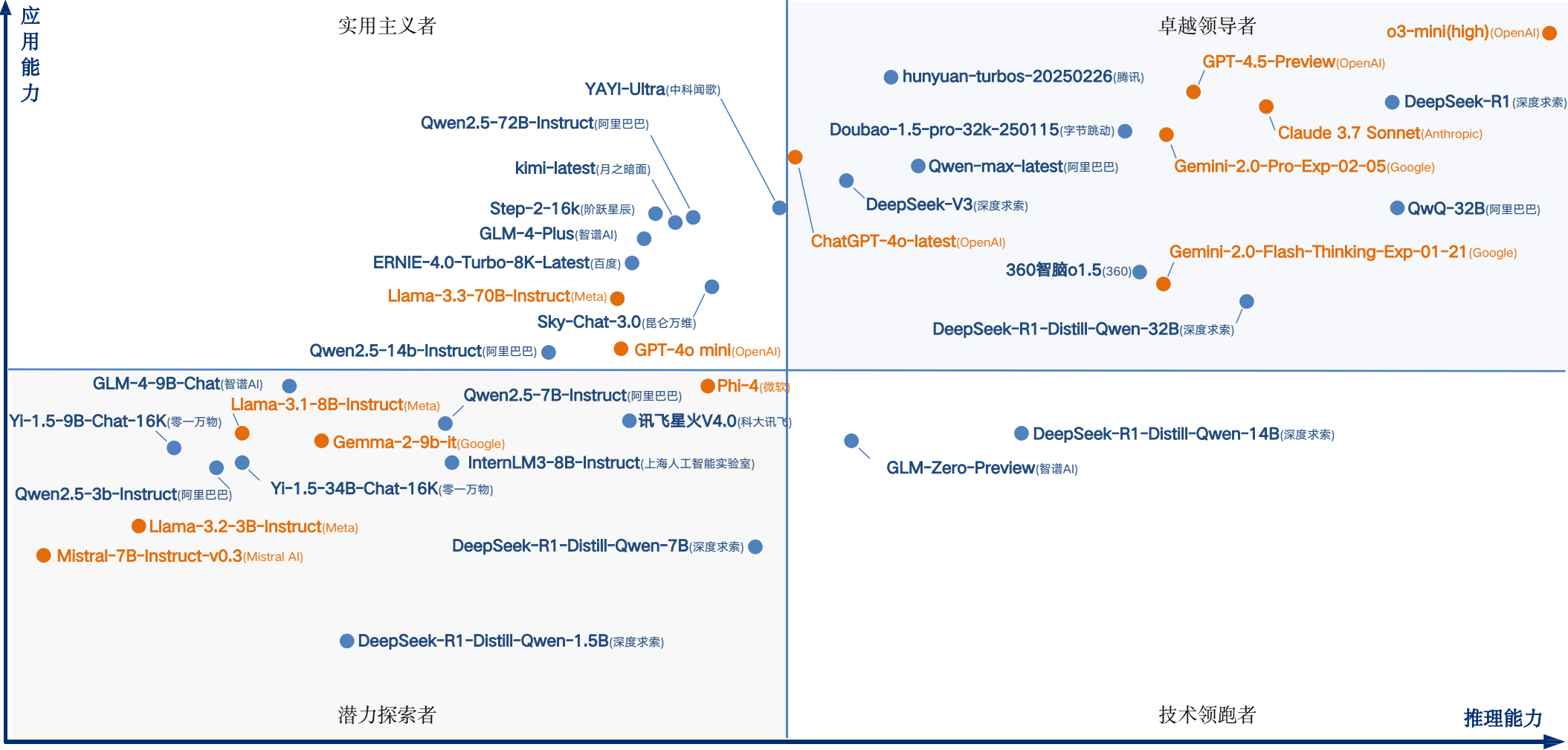
国内模型

海外及其他对比模型



来源：SuperCLUE，2025年3月18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模型象限（2025）



来源：SuperCLUE，2025年3月18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；应用能力包括：文本理解与创作、指令遵循、Agent能力；

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-总榜

测评分析

1. o3-mini(high)总分断层领先5.67分，数学推理近满分、指令遵循能力表现卓越

在2025年3月SuperCLUE测评中，OpenAI的o3-mini(high)以总分76.01分稳居榜首，与国内排名第一的模型DeepSeek-R1（70.34分）拉开5.67分差距。尤其在数学推理任务中，o3-mini(high)以94.74分的成绩大幅领先其他模型，指令遵循能力也与第二名拉开近20分的差距。

2. 国产模型崭露头角，DeepSeek-R1、QwQ-32B等加速缩小与海外模型的差距

国产模型整体表现亮眼，DeepSeek-R1等国内头部模型在关键能力维度上已能与国际领先模型媲美，甚至在某些领域实现超越。QwQ-32B在数学推理任务中取得88.60分，超越了GPT-4.5-Preview、Claude 3.7 Claude等众多海外模型。与此同时，字节跳动的Doubao-1.5-Pro在科学推理方面（70.00分）表现突出，比肩o3-mini(high)，腾讯的hunyuan-turbos则凭借70.09分刷新了Agent能力榜单成绩，展现了国产模型在多个关键指标上的强劲实力。

3. DeepSeek-R1-Distill系列模型在推理任务中的表现接近于闭源大模型。

从测评结果可见，多款开源小参数量模型展现出惊人潜力。尤其是DeepSeek-R1-Distill系列，在数学推理任务中表现十分突出，其中7B和14B版本分别取得了77.23分和79.46分的高分，不仅大幅领先同级别的小模型，还超越了众多闭源大模型，展现出接近甚至媲美旗舰大模型的卓越实力。

SuperCLUE测评基准2025年3月总体表现										
排名	模型名称	机构	总分	数学推理	科学推理	代码生成	智能体Agent	指令遵循	文本理解与创作	使用方式
-	o3-mini(high)	OpenAI	76.01	94.74	70.00	88.78	57.14	66.40	79.01	API
	DeepSeek-R1	深度求索	70.34	85.96	64.00	86.94	65.18	39.52	80.41	API
-	Claude 3.7 Sonnet(Extended)	Anthropic	68.02	78.07	59.00	86.73	56.62	48.92	78.77	POE
-	GPT-4.5-Preview	OpenAI	67.46	67.54	70.00	79.18	71.88	35.75	80.40	POE
	QwQ-32B	阿里巴巴	66.38	88.60	67.00	81.84	48.66	29.92	82.27	API
-	Gemini-2.0-Pro-Exp-02-05	Google	65.35	65.79	70.71	77.76	64.88	33.60	79.34	API
	Doubao-1.5-pro-32k-250115	字节跳动	64.69	62.28	70.00	76.94	54.46	46.77	77.66	API
4	hunyuan-turbos-20250226	腾讯	62.49	47.37	63.00	74.49	70.09	41.13	78.88	API
5	DeepSeek-R1-Distill-Qwen-32B	深度求索	59.94	85.85	62.89	73.43	36.77	23.18	77.53	API
5	Qwen-max-latest	阿里巴巴	59.34	42.98	68.00	76.33	58.48	29.38	80.88	API
-	Gemini-2.0-Flash-Thinking-Exp-01-21	Google	59.27	83.33	63.00	68.16	26.34	33.60	81.16	API
5	360智脑o1.5	360	59.08	83.33	57.00	71.43	36.61	26.34	79.78	API
6	DeepSeek-V3	深度求索	57.63	48.25	63.00	68.78	63.39	23.39	78.99	API
-	ChatGPT-4o-latest	OpenAI	57.57	35.96	66.00	73.06	56.70	32.80	80.89	API
7	YAYI-Ultra	中科闻歌	55.81	42.11	62.00	69.39	59.38	23.39	78.57	API
8	Qwen2.5-72B-Instruct	阿里巴巴	51.90	33.33	58.00	62.86	55.80	22.91	78.52	API
8	kimi-latest	月之暗面	51.47	27.19	54.00	70.61	62.05	19.89	75.10	API
9	Step-2-16k	阶跃星辰	50.81	26.32	58.00	62.45	59.38	18.55	80.17	API
10	DeepSeek-R1-Distill-Qwen-14B	深度求索	49.67	79.46	63.27	55.79	7.14	16.85	75.51	API
10	Sky-Chat-3.0	昆仑万维	49.17	38.60	63.00	55.10	38.84	21.83	77.66	API
11	GLM-4-Plus	智谱AI	48.61	26.32	53.00	61.84	49.55	21.77	79.17	API
12	ERNIE-4.0-Turbo-8K-Latest	百度	47.56	29.82	48.00	61.22	50.45	19.35	76.54	API
13	GLM-Zero-Preview	智谱AI	46.11	74.56	64.00	41.02	8.48	16.94	71.64	API
-	Llama-3.3-70B-Instruct	Meta	45.53	21.05	52.00	62.86	39.29	26.08	71.92	API
-	Phi-4	微软	45.26	35.09	61.00	60.20	23.83	15.05	76.37	模型
-	GPT-4o mini	OpenAI	43.80	21.05	53.00	63.06	29.02	20.43	76.22	API
14	讯飞星火V4.0	科大讯飞	40.76	39.82	49.00	51.22	16.52	12.63	75.36	API
14	Qwen2.5-14b-Instruct	阿里巴巴	40.70	21.05	48.00	50.61	32.59	15.09	76.87	API
15	DeepSeek-R1-Distill-Qwen-7B	深度求索	39.07	77.23	58.06	34.50	2.68	6.47	55.45	API
16	Qwen2.5-7B-Instruct	阿里巴巴	34.01	21.05	39.00	40.00	17.41	10.51	76.11	API
17	InternLM3-8B-Instruct	上海AI Lab	32.02	32.74	43.00	25.31	8.93	8.60	73.53	模型
18	GLM-4-9B-Chat	智谱AI	29.34	7.02	21.00	33.88	30.36	9.14	74.66	模型
-	Gemma-2-9b-it	Google	28.30	2.63	31.00	37.35	10.27	16.67	71.88	模型
-	Llama-3.1-8B-Instruct	Meta	25.42	1.75	19.00	31.02	23.66	10.48	66.63	API
19	Yi-1.5-34B-Chat-16K	零一万物	23.29	6.14	22.00	23.27	7.14	7.80	73.41	模型
20	Qwen2.5-3b-Instruct	阿里巴巴	22.18	13.16	20.00	12.65	7.59	6.20	73.49	API
20	Yi-1.5-9B-Chat-16K	零一万物	21.94	4.42	19.00	14.49	14.75	7.53	71.42	模型
21	DeepSeek-R1-Distill-Qwen-1.5B	深度求索	17.98	37.72	35.00	3.88	0.00	1.62	29.64	API
-	Llama-3.2-3B-Instruct	Meta	17.15	7.89	5.00	18.78	3.57	5.48	62.17	API
-	Mistral-7B-Instruct-v0.3	Mistral AI	11.78	1.75	5.00	2.86	1.34	4.30	55.43	模型

注：数据来源SuperCLUE，2025年3月18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。

SuperCLUE-推理模型总榜

SuperCLUE测评基准2025年3月推理模型表现							
排名	模型名称	机构	推理榜单总分	数学推理	科学推理	代码生成	使用方式
-	o3-mini(high)	OpenAI	84.51	94.74	70.00	88.78	API
	QwQ-32B	阿里巴巴	79.15	88.60	67.00	81.84	API
	DeepSeek-R1	深度求索	78.97	85.96	64.00	86.94	API
	QwQ-Max-preview	阿里巴巴	76.01	86.84	64.65	76.53	网页
-	Claude 3.7 Sonnet	Anthropic	74.60	78.07	59.00	86.73	POE
	DeepSeek-R1-Distill-Qwen-32B	深度求索	74.06	85.85	62.89	73.43	API
-	Gemini-2.0-Flash-Thinking-Exp-01-21	Google	71.50	83.33	63.00	68.16	API
4	360智脑o1.5	360	70.59	83.33	57.00	71.43	API
4	Kimi-K1.5	月之暗面	70.19	78.07	67.00	65.51	网页
5	DeepSeek-R1-Distill-Qwen-14B	深度求索	66.17	79.46	63.27	55.79	API
5	Step R-mini	阶跃星辰	65.20	73.45	61.00	61.16	网页
6	GLM-Zero-Preview	智谱AI	59.86	74.56	64.00	41.02	API
7	DeepSeek-R1-Distill-Qwen-7B	深度求索	56.60	77.23	58.06	34.50	API
8	Skywork o1 Preview	昆仑万维	47.67	55.17	44.21	43.63	网页
9	InternLM3-8B-Instruct	上海人工智能实验室	33.68	32.74	43.00	25.31	模型
10	DeepSeek-R1-Distill-Qwen-1.5B	深度求索	25.53	37.72	35.00	3.88	API

注：数据来源SuperCLUE，2025年3月18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。

测评分析

1.国内头部推理模型在推理任务上表现出色，但与o3-mini(high)还有一定差距。

QwQ-32B和DeepSeek-R1在本次推理模型评测中表现突出，分别以79.15分和78.97分的优异成绩摘得榜单金牌，但与海外最优模型o3-mini(high)仍存在一定差距。另外，QwQ-Max-preview、DeepSeek-R1-Distill-Qwen-32B、360智脑o1.5以及Kimi-K1.5等模型同样展现出强劲实力，均取得了70分以上的优异成绩。

2.国内头部推理模型较海外头部推理模型略有领先。

国内排名前三的推理模型平均得分为78.04分，海外排名前三的推理模型平均得分为76.87分，国内排名前三推理模型较海外高出1.17分。国内头部推理模型在性能方面相较海外同类模型展现出一定的竞争力，并略具优势。

3.国内推理模型在数学推理任务上表现较好，但在科学推理和代码两大任务上还有进步空间。

在本次测评中，国内推理模型在数学推理任务上的平均得分为72.23分，表现良好。但在科学推理和代码两大任务上的平均得分仅为57.78分和55.46分，均未超过60分的及格线；另外，在科学推理任务上，国内最佳模型的成绩尚未达到70分，而在其他两大任务上，国内最佳模型均有超过80分的表现，表明国内的推理模型在科学推理和代码任务上还有一定的进步空间。

SuperCLUE-推理任务总榜

测评分析

1. 推理模型较基础模型优势明显，前十名中仅有两席为基础模型。

根据推理任务榜单可以发现，除了GPT-4.5-Preview和Gemini-2.0-Pro-Exp-02-05 这两个综合实力强劲的通用大模型外，其余八个席位几乎被DeepSeek-R1系列、QwQ系列以及o3-mini(high)等专注于推理能力优化的模型所包揽。这些推理模型凭借其在数学、科学、代码等细分领域的精湛表现，展现出相较于通用模型更为突出的专业实力，在对模型推理深度与广度要求更高的场景下，呈现出明显的领先态势，也预示着在特定领域的精耕细作，是模型性能突破的关键路径。

2. 同类任务中模型性能差距显著，能力分化明显

例如在数学推理这一细分领域，模型的性能差距尤为显著，呈现出明显的能力分化趋势。榜单前三的模型，如o3-mini(high)、DeepSeek-R1和QwQ-Max-Preview，在数学推理任务中均取得了超过85分的高分，展现出强大的推理能力。而相比之下，排名靠后的模型则显得较为薄弱，其中一些模型的得分甚至仅为个位数，差距十分悬殊。

3. 蒸馏技术成效显著，小参数模型表现超预期。

QwQ-32B作为百亿级参数模型，以79.15的总分位居榜单前两名，国内与DeepSeek-R1并列第一，仅次于OpenAI的专用推理模型。更值得注意的是，DeepSeek-R1-Distill-Qwen系列蒸馏模型在不同参数规模下的梯度表现。从32B（74.06）到14B（66.17）、7B（56.60）以及1.5B（25.53），充分展示了蒸馏技术在知识传递方面的高效性。

SuperCLUE测评基准2025年3月推理任务表现							
排名	模型名称	机构	推理榜单总分	数学推理	科学推理	代码生成	使用方式
-	o3-mini(high)	OpenAI	84.51	94.74	70.00	88.78	API
-	QwQ-32B	阿里巴巴	79.15	88.60	67.00	81.84	API
-	DeepSeek-R1	深度求索	78.97	85.96	64.00	86.94	API
-	QwQ-Max-preview	阿里巴巴	76.01	86.84	64.65	76.53	网页
-	Claude 3.7 Sonnet	Anthropic	74.60	78.07	59.00	86.73	POE
-	DeepSeek-R1-Distill-Qwen-32B	深度求索	74.06	85.85	62.89	73.43	API
-	GPT-4.5-Preview	OpenAI	72.24	67.54	70.00	79.18	POE
-	Gemini-2.0-Flash-Thinking-Exp-01-21	Google	71.50	83.33	63.00	68.16	API
-	Gemini-2.0-Pro-Exp-02-05	Google	71.42	65.79	70.71	77.76	API
4	360智脑o1.5	360	70.59	83.33	57.00	71.43	API
4	Kimi-K1.5	月之暗面	70.19	78.07	67.00	65.51	网页
5	Doubao-1.5-pro-32k-250115	字节跳动	69.74	62.28	70.00	76.94	API
-	Grok-3	X.AI	68.86	58.77	67.00	80.82	网页
6	DeepSeek-R1-Distill-Qwen-14B	深度求索	66.17	79.46	63.27	55.79	API
6	Step R-mini	阶跃星辰	65.20	73.45	61.00	61.16	网页
7	Qwen-max-latest	阿里巴巴	62.44	42.98	68.00	76.33	API
7	hunyuan-turbos-20250226	腾讯	61.62	47.37	63.00	74.49	API
8	DeepSeek-V3	深度求索	60.01	48.25	63.00	68.78	API
8	GLM-Zero-Preview	智谱AI	59.86	74.56	64.00	41.02	API
-	ChatGPT-4o-latest	OpenAI	58.34	35.96	66.00	73.06	API
9	YAYI-Ultra	中科闻歌	57.83	42.11	62.00	69.39	API
10	DeepSeek-R1-Distill-Qwen-7B	深度求索	56.60	77.23	58.06	34.50	API
10	Sky-Chat-3.0	昆仑万维	52.23	38.60	63.00	55.10	API
-	Phi-4	微软	52.10	35.09	61.00	60.20	模型
11	Qwen2.5-72B-Instruct	阿里巴巴	51.40	33.33	58.00	62.86	API
12	kimi-latest	月之暗面	50.60	27.19	54.00	70.61	API
13	Step-2-16k	阶跃星辰	48.92	26.32	58.00	62.45	API
14	Skywork o1 Preview	昆仑万维	47.67	55.17	44.21	43.63	网页
14	GLM-4-Plus	智谱AI	47.05	26.32	53.00	61.84	API
14	讯飞星火V4.0	科大讯飞	46.68	39.82	49.00	51.22	API
15	ERNIE-4.0-Turbo-8K-Latest	百度	46.35	29.82	48.00	61.22	API
-	GPT-4o mini	OpenAI	45.70	21.05	53.00	63.06	API
-	Llama-3.3-70B-Instruct	Meta	45.30	21.05	52.00	62.86	API
16	Qwen2.5-14b-Instruct	阿里巴巴	39.89	21.05	48.00	50.61	API
17	InternLM3-8B-Instruct	上海人工智能实验室	33.68	32.74	43.00	25.31	模型
17	Qwen2.5-7B-Instruct	阿里巴巴	33.35	21.05	39.00	40.00	API
18	DeepSeek-R1-Distill-Qwen-1.5B	深度求索	25.53	37.72	35.00	3.88	API
-	Gemma-2-9b-it	Google	23.66	2.63	31.00	37.35	模型
19	GLM-4-9B-Chat	智谱AI	20.63	7.02	21.00	33.88	模型
-	Llama-3.1-8B-Instruct	Meta	17.26	1.75	19.00	31.02	API
20	Yi-1.5-34B-Chat-16K	零一万物	17.14	6.14	22.00	23.27	模型
21	Qwen2.5-3b-Instruct	阿里巴巴	15.27	13.16	20.00	12.65	API
22	Yi-1.5-9B-Chat-16K	零一万物	12.64	4.42	19.00	14.49	模型
-	Llama-3.2-3B-Instruct	Meta	10.56	7.89	5.00	18.78	API
-	Mistral-7B-Instruct-v0.3	Mistral AI	3.20	1.75	5.00	2.86	模型

注：数据来源SuperCLUE，2025年3月18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。

SuperCLUE-基础模型总榜

SuperCLUE测评基准2025年3月基础模型总体表现										
排名	模型名称	机构	总分	数学推理	科学推理	代码生成	智能体Agent	指令遵循	文本理解与创作	使用方式
-	GPT-4.5-Preview	OpenAI	67.46	67.54	70.00	79.18	71.88	35.75	80.40	POE
-	Gemini-2.0-Pro-Exp-02-05	Google	65.35	65.79	70.71	77.76	64.88	33.60	79.34	API
	Doubao-1.5-pro-32k-250115	字节跳动	64.69	62.28	70.00	76.94	54.46	46.77	77.66	API
	hunyuan-turbos-20250226	腾讯	62.49	47.37	63.00	74.49	70.09	41.13	78.88	API
	Qwen-max-latest	阿里巴巴	59.34	42.98	68.00	76.33	58.48	29.38	80.88	API
4	DeepSeek-V3	深度求索	57.63	48.25	63.00	68.78	63.39	23.39	78.99	API
-	ChatGPT-4o-latest	OpenAI	57.57	35.96	66.00	73.06	56.70	32.80	80.89	API
5	YAYI-Ultra	中科闻歌	55.81	42.11	62.00	69.39	59.38	23.39	78.57	API
6	Qwen2.5-72B-Instruct	阿里巴巴	51.90	33.33	58.00	62.86	55.80	22.91	78.52	API
6	kimi-latest	月之暗面	51.47	27.19	54.00	70.61	62.05	19.89	75.10	API
7	Step-2-16k	阶跃星辰	50.81	26.32	58.00	62.45	59.38	18.55	80.17	API
8	Sky-Chat-3.0	昆仑万维	49.17	38.60	63.00	55.10	38.84	21.83	77.66	API
8	GLM-4-Plus	智谱AI	48.61	26.32	53.00	61.84	49.55	21.77	79.17	API
9	ERNIE-4.0-Turbo-8K-Latest	百度	47.56	29.82	48.00	61.22	50.45	19.35	76.54	API
-	Llama-3.3-70B-Instruct	Meta	45.53	21.05	52.00	62.86	39.29	26.08	71.92	API
-	Phi-4	微软	45.26	35.09	61.00	60.20	23.83	15.05	76.37	模型
-	GPT-4o mini	OpenAI	43.80	21.05	53.00	63.06	29.02	20.43	76.22	API
10	讯飞星火V4.0	科大讯飞	40.76	39.82	49.00	51.22	16.52	12.63	75.36	API
10	Qwen2.5-14b-Instruct	阿里巴巴	40.70	21.05	48.00	50.61	32.59	15.09	76.87	API
11	Qwen2.5-7B-Instruct	阿里巴巴	34.01	21.05	39.00	40.00	17.41	10.51	76.11	API
12	GLM-4-9B-Chat	智谱AI	29.34	7.02	21.00	33.88	30.36	9.14	74.66	模型
-	Gemma-2-9b-it	Google	28.30	2.63	31.00	37.35	10.27	16.67	71.88	模型
-	Llama-3.1-8B-Instruct	Meta	25.42	1.75	19.00	31.02	23.66	10.48	66.63	API
13	Yi-1.5-34B-Chat-16K	零一万物	23.29	6.14	22.00	23.27	7.14	7.80	73.41	模型
14	Qwen2.5-3b-Instruct	阿里巴巴	22.18	13.16	20.00	12.65	7.59	6.20	73.49	API
14	Yi-1.5-9B-Chat-16K	零一万物	21.94	4.42	19.00	14.49	14.75	7.53	71.42	模型
-	Llama-3.2-3B-Instruct	Meta	17.15	7.89	5.00	18.78	3.57	5.48	62.17	API
-	Mistral-7B-Instruct-v0.3	Mistral AI	11.78	1.75	5.00	2.86	1.34	4.30	55.43	模型

注：数据来源SuperCLUE，2025年3月18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。

测评分析

1.国内头部基础模型表现优异，得分均超过ChatGPT-4o-latest。

在本次测评的基础模型总榜中，Doubao-1.5-pro-32k-250115以64.69分的优异成绩荣登榜首，摘得金牌；hunyuan-turbos-20250226和Qwen-max-latest分别以62.49分和59.34分位列第二和第三。此外，DeepSeek-V3以57.63分位居第四，表现同样亮眼，上述模型均领先于海外知名模型ChatGPT-4o-latest。

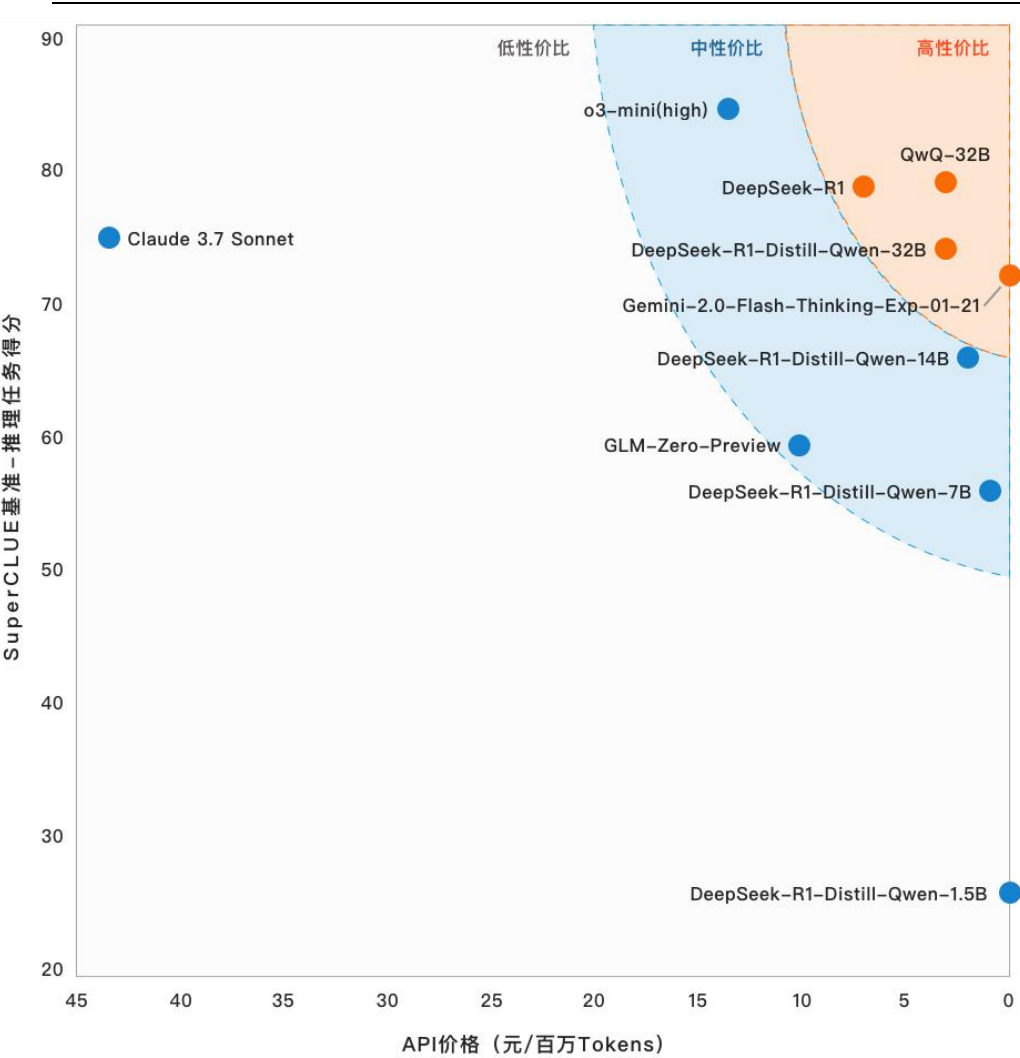
2.国内头部基础模型较海外头部基础模型略有落后，但差距很小。

国内排名前三的基础模型平均得分达62.17分，而海外排名前三的基础模型平均得分则为63.46分，国内基础模型较海外低1.29分。具体而言，国内基础模型在指令遵循任务方面展现出相对优势，但在数学推理任务上存在一定差距。其中，Doubao-1.5-pro-32k-250115与hunyuan-turbos-20250226在指令遵循任务中表现卓越，得分均超过40分。但在数学推理任务上，仅有Doubao-1.5-pro-32k-250115得分突破60分。

3.各基础模型的表现呈现出明显的任务差异性。

具体而言，在文本理解与创作任务上，模型表现相对稳定，标准差为5.76分；而在其他任务维度上，模型性能的离散程度显著增大，标准差均超过10分。尤其是在科学推理、代码处理以及Agent这三个任务，模型表现的差异性尤为突出，标准差均超过20分。

推理模型性价比分布



数据来源：SuperCLUE，2025年3月18日；推理任务得分为推理任务总分：数学推理、科学推理和代码的平均分。开源模型如QwQ-32B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3月的标准价格（非优惠价格）。

趋势分析

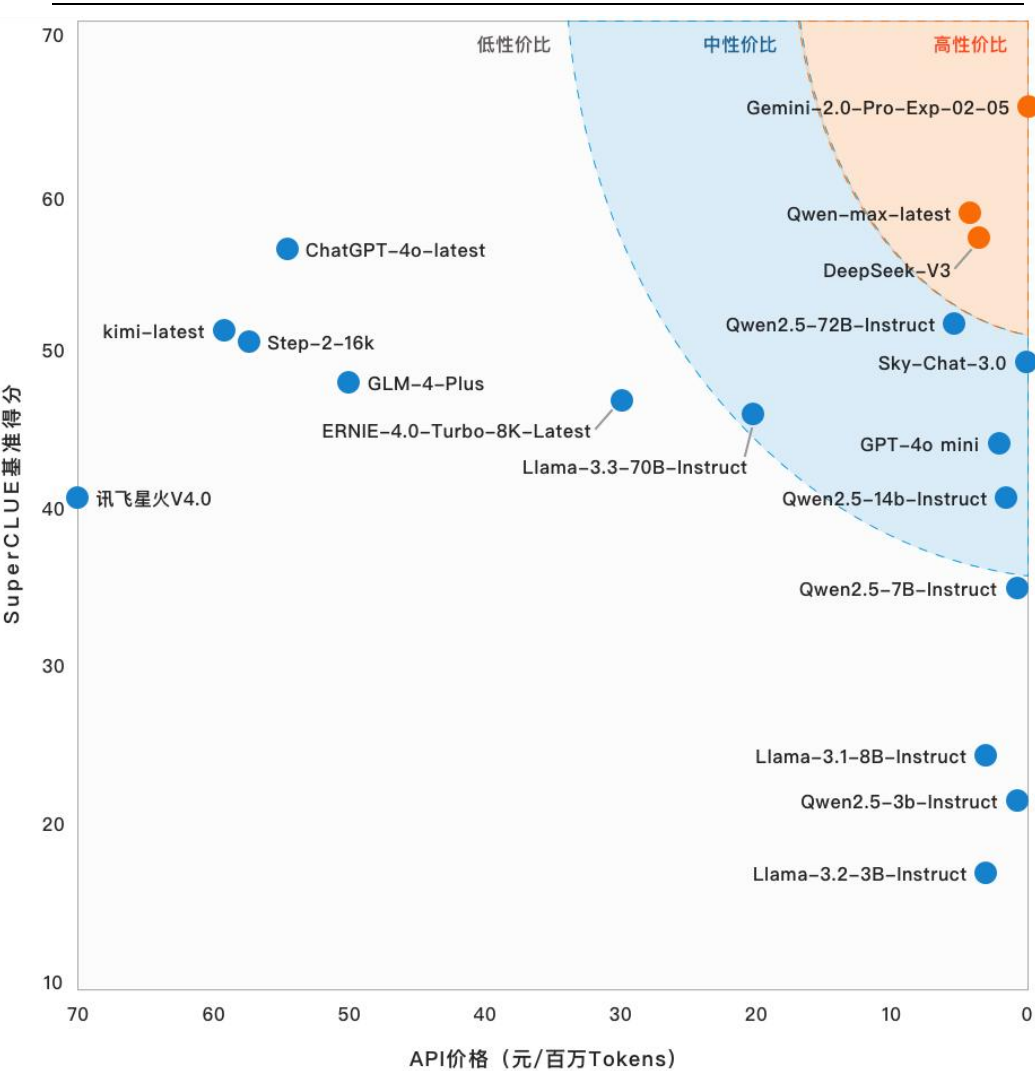
1. 国产推理模型凭借较低的价格实现高质量输出，展现出显著的性价比优势。

国产推理模型QwQ-32B、DeepSeek-R1和DeepSeek-R1-Distill-Qwen-32B在性价比方面展现出强大竞争力。它们在保持高水平性能的同时，保持了极低的应用成本，展现出较好的落地可用性。而海外模型Gemini 2.0 Flash Thinking exp（暂时免费）也具备高性价比，但在推理任务上的表现略逊色于其他国产高性价比模型。

2. 推理模型的推理能力与其API价格无关，其中o3-mini-high以中等的价位领先其他模型。

在推理任务上，得分较高的推理模型大多集中在中高性价比区域。这些模型不仅在推理能力上有所增强，同时价格上也展现出显著的优势。o3-mini-high以中等的价位表现最佳；相比之下，Claude 3.7 Sonnet的成本最高，但在输出质量方面不是最佳，属于低性价比模型。

基础模型性价比分布



数据来源：SuperCLUE，2025年3月18日；总分为六大任务的平均分。开源模型如QwQ-32B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3月的标准价格（非优惠价格）。

趋势分析

1. 国产基础模型在性价比（价格+效果）上有较大优势

国产基础模型如Qwen-max-latest和DeepSeek-V3在性价比上展现出强大的竞争力。这些基础模型在较高性能的基础上，能够保持低应用成本，在应用落地方面展现出友好的可用性。

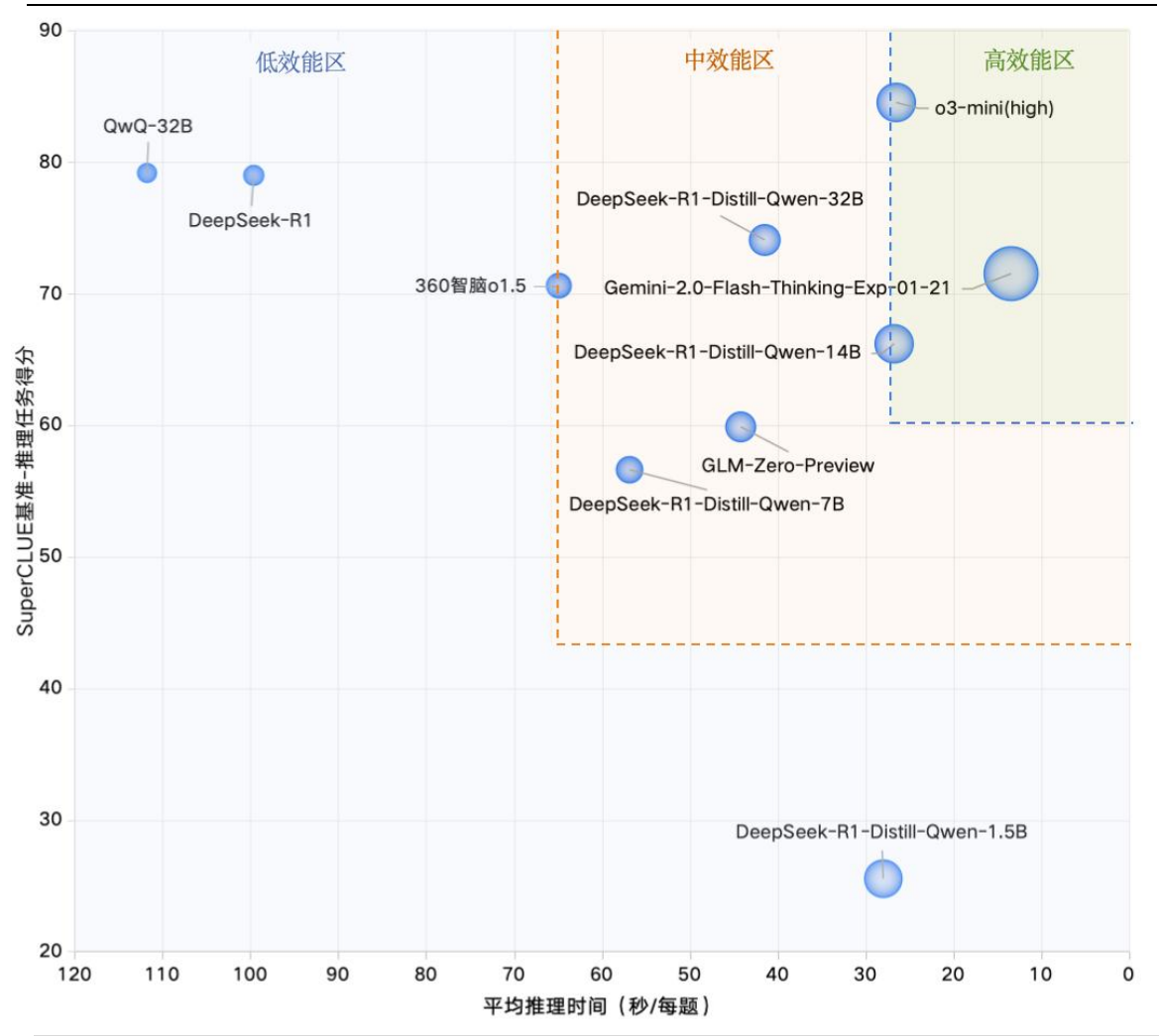
2. 位于中性价比的基础模型在价格方面取胜

中性价比模型，如Qwen2.5-72B-Instruct、Sky-Chat-3.0、GPT-4o-mini等，价格均在5元/百万tokens左右，任务得分处于中等水平，适用于广泛应用中，在价格和性能方面提供了很好的平衡。

3. 位于低性价比区间的基础模型较为集中，便可划分为高性高价和低性低价

位于低性价比区域的基础模型，在价格或性能上略逊于中高性价比模型，例如GPT-4o-latest，kimi-latest等价格较高但性能可圈可点；而Llama-3.2-3B-Instruct、Qwen2.5-3B-Instruct等的性能尚有提升空间，但价格较为实惠。表明这些模型的发展表面上显得过于片面，影响用户体验。

推理模型推理效能区间



数据来源：SuperCLUE，2025年3月18日；
模型推理速度选取3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、科学推理和代码的平均分。

趋势分析

1.海外推理模型综合效能领先

o3-mini (high) 和 Gemini-2.0-Flash-Thinking-Exp-01-21 在推理速度和基准得分的综合表现上处于领先地位，推理任务分数均在70分以上，平均推理耗时在30秒内。

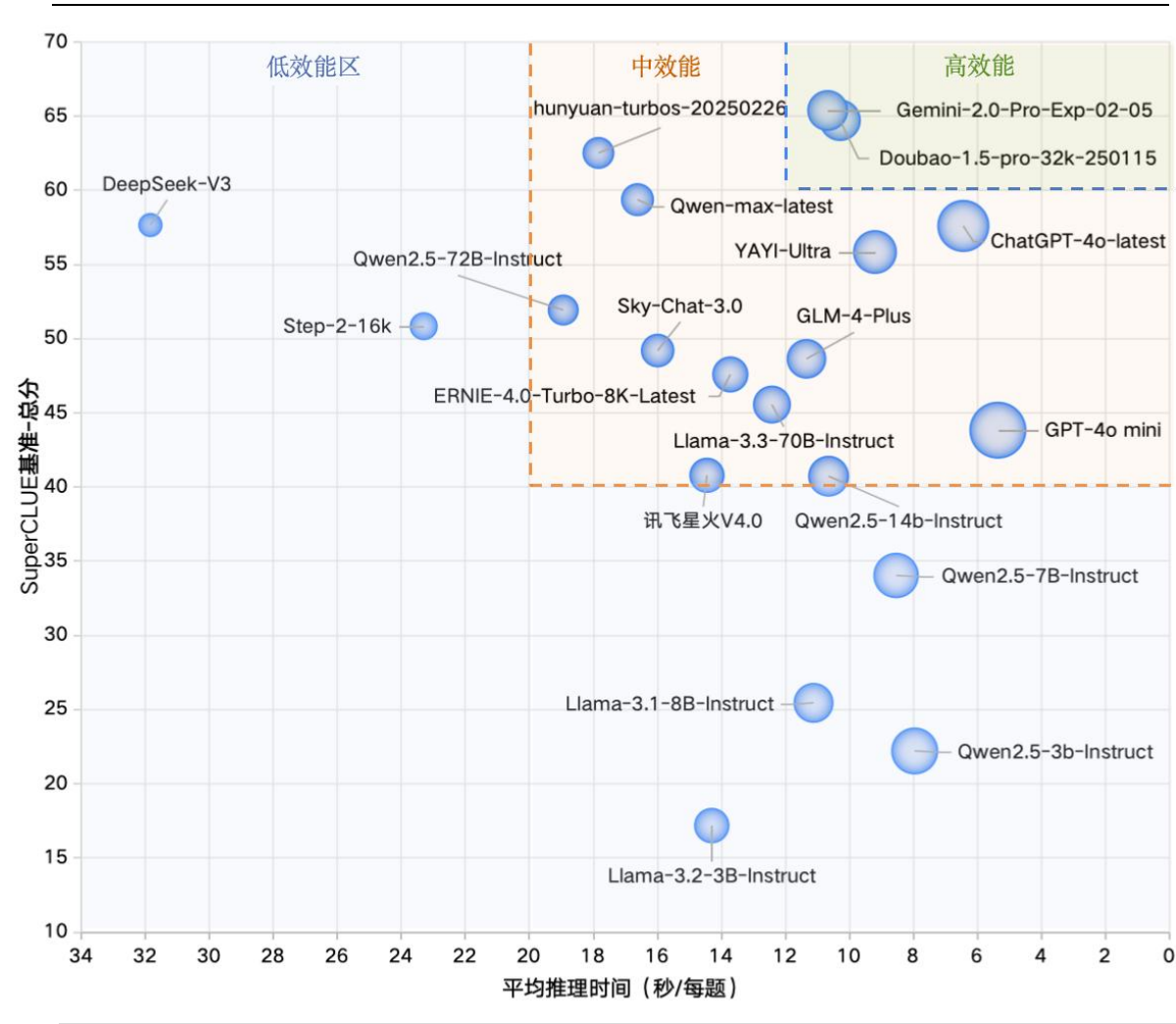
2.国内推理模型大部分处于中低效能区间

国内推理模型中，QWQ-32B推理任务得分最高（79.15分），但平均推理耗时超过110秒。推理任务得分超过70分的国内推理模型，平均推理耗时均超过40秒。

3.兼顾推理速度和准确度是重难点

即使对于综合效能较高的o3-mini (high)，其平均推理时间也达到了26秒。为了更好地将推理模型应用于实际场景，需要考虑如何在确保较高推理准确度的同时提升推理速度。

基础模型推理效能区间



数据来源: SuperCLUE, 2025年3月18日;
模型推理速度选取3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)。
总分为六大任务的平均分。

趋势分析

1.Doubao、Gemini综合效能表现领先

Doubao-1.5-pro-32k-250115和Gemini-2.0-Pro-Exp-02-05, 推理速度表现优异, 平均推理耗时均在10秒左右, 基准得分超过60分, 符合「高效能区」, 展现了较强的应用潜力。

2.国内基础模型综合效能与海外基础模型无明显优劣

在高效能区中, 国内基础模型表现与海外基础模型在伯仲之间。中效能区中, 国内基础模型表现与海外基础模型表现各有优劣, 在推理速度上, ChatGPT-4o-latest和GPT-4o-mini领先, 但在得分上只有hunyuan-turbos-20250226达到了60分。

3.基础模型在综合效能上还有提升空间

被测的基础模型推理耗时平均在20秒左右, 但得分都在70分以下, 仍然有一定的提升空间。

SuperCLUE大模型能力成熟度指数-SC指数

指数序号	能力	最高分	最低分	成熟度SC指数	成熟度区间
1	文本理解与创作	80.88	71.64	0.89	高成熟度(>0.8)
2	科学推理	70.00	48.00	0.69	中成熟度(0.5-0.8)
3	代码生成	76.94	41.02	0.53	
4	数学推理	83.33	26.32	0.32	低成熟度(0.2-0.5)
5	精确指令遵循	46.77	12.63	0.27	
6	智能体Agent	70.09	8.48	0.12	极低成熟度(0.1-0.2)

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。
- ✓ 当前国内大模型成熟度较高的能力是【文本理解与创作】，也是目前产业和用户侧大模型的重点应用场景。

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【科学推理】、【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0.2至0.5之间。
- ✓ 当前国内大模型低成熟度的能力是【数学推理】、和【精确指令遵循】。

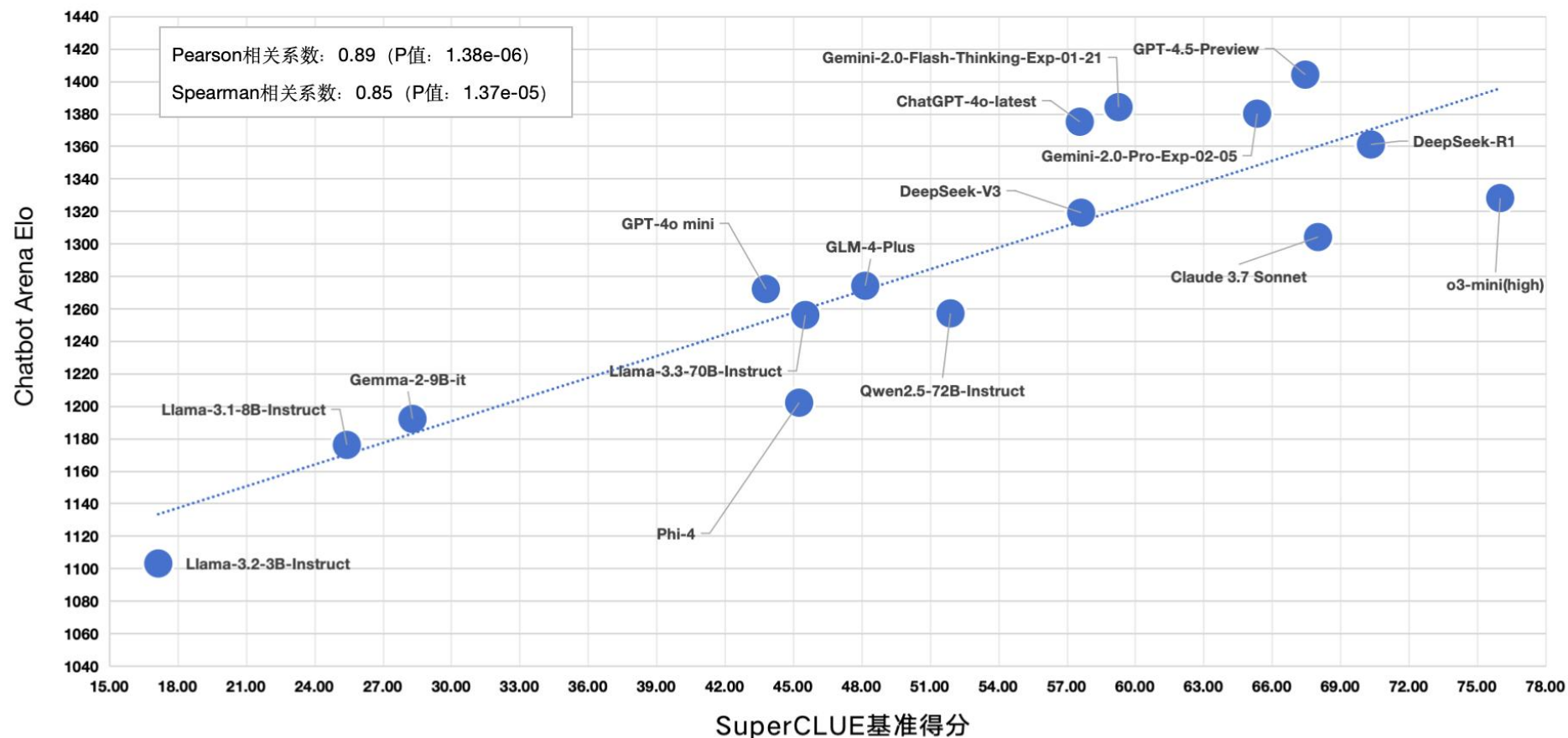
4.极低成熟度能力

- ✓ 极低成熟度指的是大多数模型无法胜任。SC成熟度指数在0.2以下。
- ✓ 当前国内大模型极低成熟度的能力是【智能体Agent】

来源：SuperCLUE, 2025年3月18日；SC成熟度指数=国内闭源模型最差成绩/国内闭源模型最好成绩

Chatbot Arena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。将SuperCLUE得分与ChatBot Arena得分进行相关性计算，得到**皮尔逊相关系数：0.89，P值：1.38e-06**；**斯皮尔曼相关系数：0.85，P值：1.37e-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Chatbot Arena为典型代表），具有**高度一致性**。

评测与人类一致性验证：SuperCLUE VS Chatbot Arena



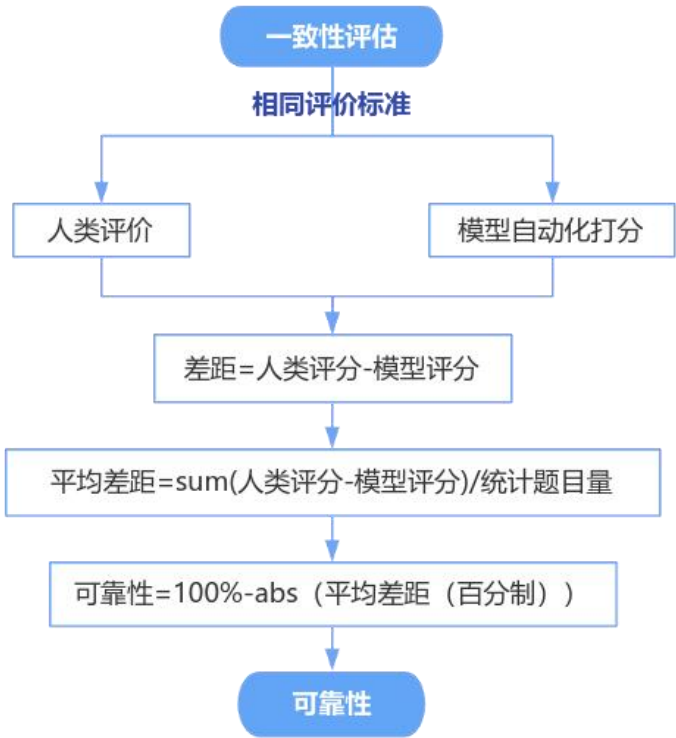
来源: SuperCLUE, 2025年3月18日;

可靠性验证
说明

3月SuperCLUE采用自动化评价的测评方式。

为验证自动化评价的可靠性，SuperCLUE团队在进行正式测评之前，从1509道题目中针对5个模型，每个模型随机抽取了130道题目进行人工复审。

人类按照相同的评价标准，对每道题目进行评价。统计人类评分与模型评分在所有题目上的平均差距，最终计算得到可靠性。



验证结果

最终各模型可靠性指标结果如下：

模型列表	可靠性
模型1	98.39%
模型2	98.12%
模型3	91.06%
模型4	97.41%
模型5	97.20%
最终可靠性	96.44%

注：最终可靠性=sum（人类评分-模型评分）/总题目量

通过5个模型的可靠性分析验证，我们发现可靠性数据分别为98.39%、98.12%、91.06%、97.41%、97.20%，其中可靠性最低有91.06%，最高为模型的98.39%。最终有96.44%的可靠性。

所以，经过验证，SuperCLUE自动化评价有较高的可靠性。

SuperCLUE 开源榜单

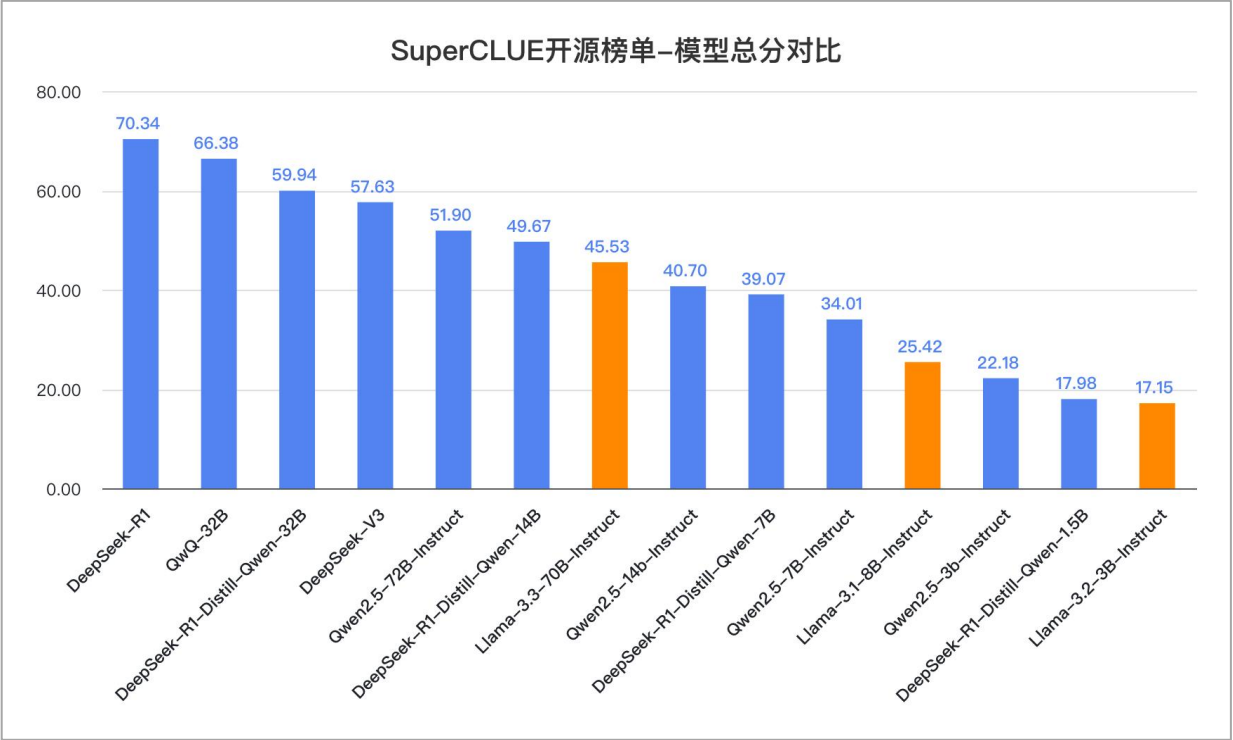
排名	模型名称	机构	参数量	总分
1	DeepSeek-R1	深度求索	671B	70.34
2	QwQ-32B	阿里巴巴	32B	66.38
3	DeepSeek-R1-Distill-Qwen-32B	深度求索	32B	59.94
4	DeepSeek-V3	深度求索	671B	57.63
5	Qwen2.5-72B-Instruct	阿里巴巴	72B	51.90
6	DeepSeek-R1-Distill-Qwen-14B	深度求索	14B	49.67
-	Llama-3.3-70B-Instruct	Meta	70B	45.53
-	Phi-4	微软	14B	45.26
7	Qwen2.5-14b-Instruct	阿里巴巴	14B	40.70
8	DeepSeek-R1-Distill-Qwen-7B	深度求索	7B	39.07
9	Qwen2.5-7B-Instruct	阿里巴巴	7B	34.01
10	InternLM3-8B-Instruct	上海人工智能实验室	8B	32.02
11	GLM-4-9B-Chat	智谱AI	9B	29.34
-	Gemma-2-9b-it	Google	9B	28.30
-	Llama-3.1-8B-Instruct	Meta	8B	25.42
12	Yi-1.5-34B-Chat-16K	零一万物	34B	23.29
13	Qwen2.5-3b-Instruct	阿里巴巴	3B	22.18
13	Yi-1.5-9B-Chat-16K	零一万物	9B	21.94
14	DeepSeek-R1-Distill-Qwen-1.5B	深度求索	1.5B	17.98
-	Llama-3.2-3B-Instruct	Meta	3B	17.15
-	Mistral-7B-Instruct-v0.3	Mistral AI	7B	11.78

来源：SuperCLUE, 2025年3月18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。
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。

开源模型分析

✓ 中文场景下，国内开源模型已具备较大优势

DeepSeek系列开源模型、Qwen系列开源模型，在3月SuperCLUE测评中表现优异，均有超过Llama-3.3-70B-Instruct的表现，引领全球开源生态。



SuperCLUE-10B级别小模型榜单

排名	模型名称	机构	参数量	总分
1	DeepSeek-R1-Distill-Qwen-7B	深度求索	7B	39.07
2	Qwen2.5-7B-Instruct	阿里巴巴	7B	34.01
3	InternLM3-8B-Instruct	上海人工智能实验室	8B	32.02
4	GLM-4-9B-Chat	智谱AI	9B	29.34
-	Gemma-2-9b-it	Google	9B	28.30
-	Llama-3.1-8B-Instruct	Meta	8B	25.42
5	Qwen2.5-3b-Instruct	阿里巴巴	3B	22.18
5	Yi-1.5-9B-Chat-16K	零一万物	9B	21.94
6	DeepSeek-R1-Distill-Qwen-1.5B	深度求索	1.5B	18.98
-	Llama-3.2-3B-Instruct	Meta	3B	18.15
-	Mistral-7B-Instruct-v0.3	Mistral AI	7B	11.78

来源：SuperCLUE, 2025年3月18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。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。

10B级别小模型分析

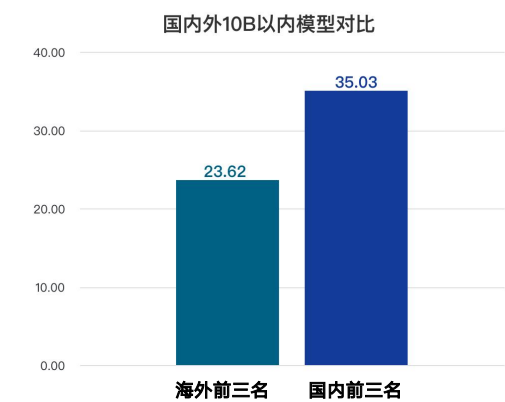
✓ 10B级别模型中，DeepSeek-R1-Distill-Qwen-7B和Gemma-2-9b-it分列国内外榜首

在本次SuperCLUE测评中，DeepSeek-R1-Distill-Qwen-7B取得39.07分，取得10B以内模型的最高分。GLM-4-9B-Chat取得29.34分排名国内第4，Qwen2.5-3b-Instruct、Yi-1.5-9B-Chat-16K并列国内5名。Gemma-2-9b-it取得28.30分，领跑海外10B以内模型。

✓ 国内10B以内模型进展迅速，展现出极致的性价比

在10B以内模型中，超过30分的模型有3个，分别为DeepSeek-R1-Distill-Qwen-7B、Qwen2.5-7B-Instruct、InternLM3-8B-Instruct均是国内大模型。展现出10B以内小参数量级模型的极致的性价比。

国内外对比



➢ 国内头部10B以内模型平均水平领先于海外模型

✓ 从国内外10B小模型能力的对比数据看，国内10B小模型有一定优势。国内10B小模型前三名模型的得分相较于国外前三名平均高出11.41分。

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
- ✓ Qwen2.5-3B-Instruct表现惊艳，取得总分22.18分的优异成绩，在SuperCLUE端侧5B小模型榜单中排名榜首。其中文本理解与创作73.49分，与同等参数量级模型Llama-3.2-3B-Instruct相比多个维度均有不同幅度的领先，展示出小参数量级模型极高的性价比。
- ✓ DeepSeek-R1-Distill-Qwen-1.5B小模型同样表现不俗，取得总分18.98分，有超过Llama-3.2-3B-Instruct的效果。

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	指令遵循	文本理解与创作	使用方式
1	Qwen2.5-3b-Instruct	阿里巴巴	3B	22.18	13.16	20.00	12.65	7.59	6.20	73.49	API
2	DeepSeek-R1-Distill-Qwen-1.5B	深度求索	1.5B	18.98	37.72	35.00	3.88	0.00	1.62	29.64	API
-	Llama-3.2-3B-Instruct	Meta	3B	18.15	7.89	5.00	18.78	3.57	5.48	62.18	API

来源： SuperCLUE, 2025年3月18日

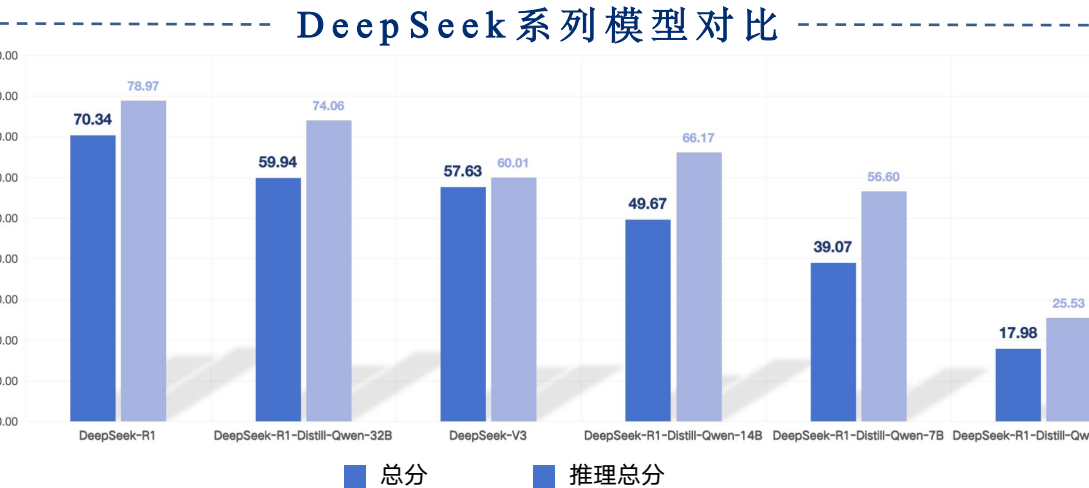
第4部分

DeepSeek系列模型深度分析

1. DeepSeek-R1及其蒸馏模型对比
2. DeepSeek-R1和国内外头部推理模型对比
3. DeepSeek-R1第三方平台联网搜索能力测试（网页端）
4. DeepSeek-R1第三方平台稳定性测试（网页端、App端、API端）

SuperCLUE-DeepSeek系列模型表现

SuperCLUE测评基准-DeepSeek系列模型表现										
模型名称	总分	推理总分	数学推理	科学推理	代码生成	智能体Agent	指令遵循	文本理解与创作	使用方式	发布日期
DeepSeek-R1	70.34	78.97	85.96	64.00	86.94	65.18	39.52	80.41	API	2025.3.18
DeepSeek-R1-Distill-Qwen-32B	59.94	74.06	85.85	62.89	73.43	36.77	23.18	77.53	API	2025.3.18
DeepSeek-V3	57.63	60.01	48.25	63.00	68.78	63.39	23.39	78.99	API	2025.3.18
DeepSeek-R1-Distill-Qwen-14B	49.67	66.17	79.46	63.27	55.79	7.14	16.85	75.51	API	2025.3.18
DeepSeek-R1-Distill-Qwen-7B	39.07	56.60	77.23	58.06	34.50	2.68	6.47	55.45	API	2025.3.18
DeepSeek-R1-Distill-Qwen-1.5B	17.98	25.53	37.72	35.00	3.88	0.00	1.62	29.64	API	2025.3.18



测评分析

1.DeepSeek-R1在各个维度均排行第一

R1在总榜和推理任务榜单上得分均超过70，在六大任务维度上明显领先其他模型。DeepSeek-R1-Distill-Qwen-32B在数学、科学推理和文本创作与理解任务上和R1得分接近，但在其余任务上相差10-30分左右。

2.推理模型在总榜和推理任务榜单上分差较大

R1和R1系列的蒸馏模型在总榜和任务榜单上的得分差距在10-20分之间，如R1在推理任务上的得分比在总榜上高出7分，DeepSeek-R1-Distill-Qwen-14B有近17分的分差，但DeepSeek-V3分差在3分之内。

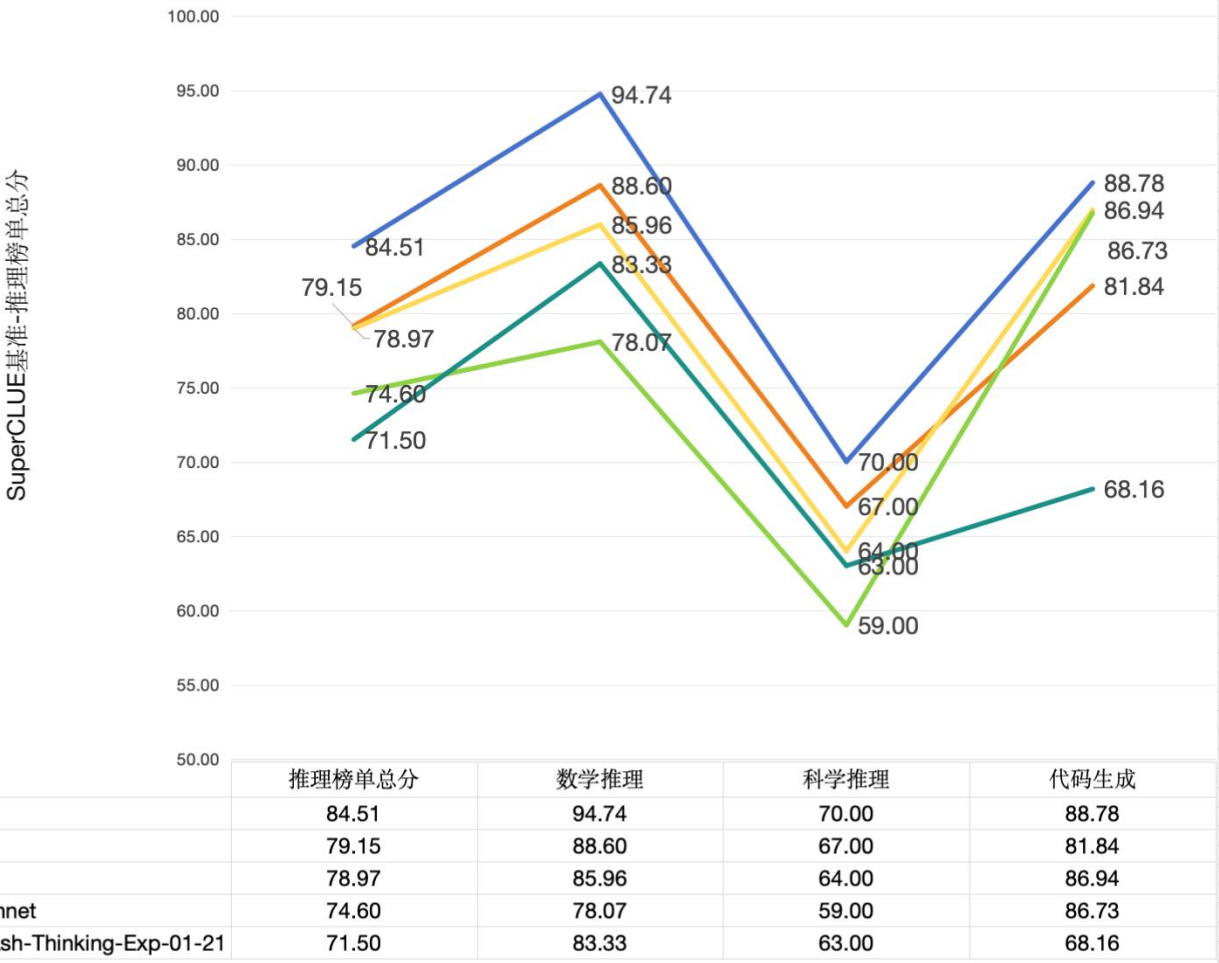
3.R1-Qwen蒸馏模型系列在推理任务中具有较高实用性

其中，DeepSeek-R1-Distill-Qwen-32B的推理任务得分接近于源模型R1的得分，而DeepSeek-R1-Distill-Qwen-14B在数学和科学推理方面表现还优于V3。这表明蒸馏模型在低能耗情况下依然能够胜任理科相关的推理任务。

DeepSeek-R1和国内外推理模型能力对比

测评分析

DeepSeek-R1和国内外头部推理模型对比



数据来源：SuperCLUE，2025年3月18日。

1.综合能力接近海外头部模型

DeepSeek-R1 推理总分 78.97，分别领先 Claude 3.7 Sonnet、Gemini-2.0-Flash-Thinking-Exp-01-21近4.37、7.47分，与o3-mini (high) 相差5.54分，展现出较强的推理能力

2.数学推理和代码生成任务表现优异

DeepSeek-R1 在数学推理和代码生成任务上得分均超过80分。在数学推理上得分 85.96，超过Claude 3.7 Sonnet近7.89分，和QwQ-32B得分接近；在代码生成任务上与o3-mini(high)仅差1.84分。

3.科学推理相对薄弱，仍需优化

DeepSeek-R1在科学推理上的取得 64.00，与o3-mini (high) 相差6分，与QwQ-32B相差3分，在科学推理任务上还有一定的提升空间。

基准说明

为进一步了解各平台接入**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。

- （一）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。

- （二）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都会清除上下文信息，避免对测评产生影响。

- （三）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。

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

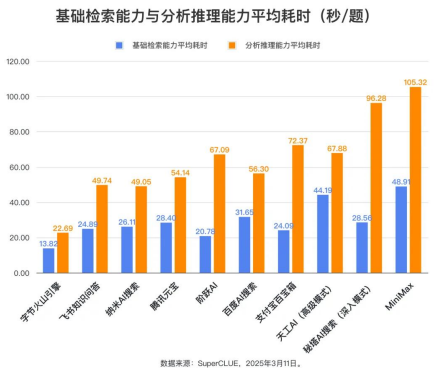
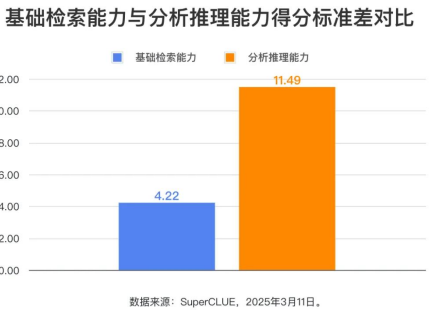
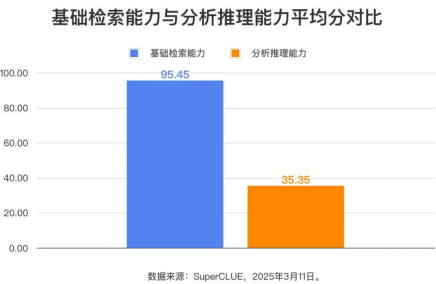
数据来源：SuperCLUE，2025年3月11日。

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。

测评结果



DeepSeek-R1稳定性测评摘要：

测评要点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。腾讯元宝以80.61分领跑联网搜索测评榜单。

测评要点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。

测评要点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。其他平台也均有超过85%的完整回复率表现。

测评要点4：各平台平均耗时差异较大，不同任务上的平均耗时差异也同样明显。

总平均耗时从最少的字节火山引擎(17.12秒/题)到最多的MiniMax(73.51秒/题)，差距非常显著。不同平台之间的搜索效率差异较大。另外，所有平台在分析推理能力上平均耗时都明显高于基础检索能力。

基准说明

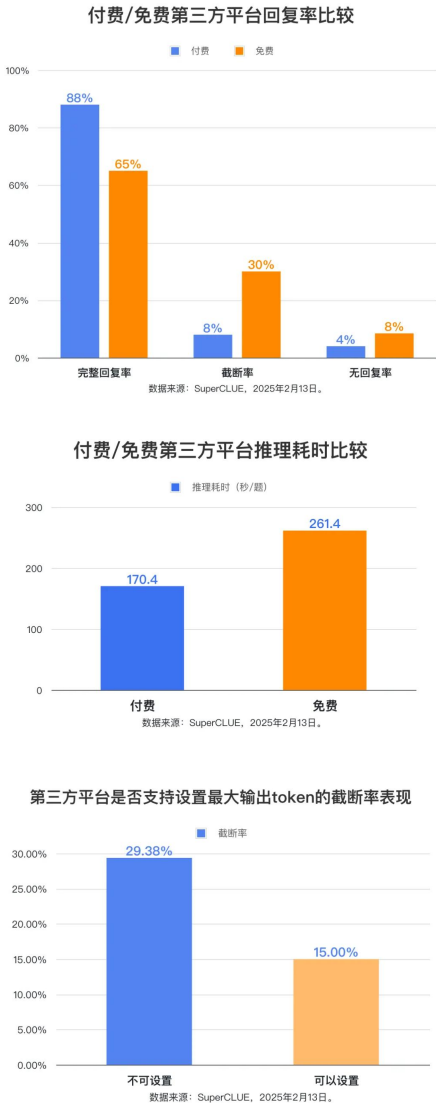
为了给用户一份全面、客观的参考，帮助他们选择最适合自身需求的服务平台，我们对支持DeepSeek-R1的多个第三方平台进行了稳定性测评。本次测评我们选择了19个第三方平台，使用20道原创小学奥数推理题来考察在各个第三方平台上使用DeepSeek-R1的实际效果，从回复率、推理耗时和准确率等方面进行评估。

完整回复率 + 截断率 + 无回复率 = 100%

- 完整回复率：模型给出完整回复，不存在截断、无响应等问题，但不考虑答案正确与否；再除以总题目数得出比例。
- 截断率：模型在回复过程中出现断开的情况，未给出完整的答案；前者再除以总题目数得出比例。
- 无回复率：模型由于特殊原因，如无响应/请求出错，未给出答案；前者再除以总题目数得出比例。
- 准确率：对于模型给出完整回复的题目，模型的答案与正确答案一致的比例；正确答案，只看最终答案，不检查解题过程。
- 推理耗时（秒/题）：对于模型给出完整回复的题目，模型推理出每题答案所用的平均时间。

第三方平台DeepSeek-R1 稳定性测评总榜（网页版）							
第三方平台名称	版本	完整 回复率	截断率	无回复率	准确率	推理耗时 (秒/题)	测评 日期
Perplexity	付费	100%	0%	0%	65%	86	2025.2.13
together.ai	付费	100%	0%	0%	85%	111	2025.2.13
字节火山引擎	免费token	100%	0%	0%	85%	392	2025.2.12
腾讯云知识引擎	免费版	95%	0%	5%	79%	216	2025.2.25
天工AI	免费版	95%	5%	0%	89%	273	2025.2.12
支付宝百宝箱	免费非联网	90%	10%	0%	89%	143	2025.2.13
商汤大装置	免费token	90%	5%	5%	78%	155	2025.2.12
硅基流动	pro版	90%	10%	0%	83%	259	2025.2.13
秘塔AI搜索	免费版	90%	10%	0%	89%	260	2025.2.12
无间芯穹	免费版	90%	10%	0%	89%	356	2025.2.12
POE	付费	75%	5%	20%	80%	130	2025.2.12
讯飞开放平台	免费token	75%	25%	0%	80%	263	2025.2.12
游晨云	VIP版	75%	25%	0%	87%	266	2025.2.13
腾讯元宝	免费版	70%	30%	0%	86%	311	2025.2.13
纳米AI搜索	免费token	55%	45%	0%	82%	163	2025.2.12
PPIO派欧云	免费版	55%	45%	0%	100%	298	2025.2.12
百度智能云	免费token	30%	40%	30%	-	-	2025.2.12
腾讯云 TI 平台	免费版	5%	95%	0%	-	-	2025.2.12
硅基流动	免费版	0%	25%	75%	-	-	2025.2.12
数据来源：SuperCLUE。 注：1.对于完整回复率不足50%的第三方平台，准确率和推理耗时不参与统计； 2.“推理耗时”指标仅反映用户感知的文本生成时间，不代表系统延时（TTFT与TOPT）等性能指标，相关数据将后续测评公布。 端到端耗时公式为：tft + tpot × (输出长度 + 推理长度)。							

测评结果



DeepSeek-R1稳定性测评摘要：

测评要点1：各个第三方平台使用DeepSeek-R1的完整回复率表现差异较大。Perplexity、together.ai和字节火山引擎完整回复率都达到了100%，表现出色，稳定性优异；腾讯云知识引擎、天工AI、支付宝百宝箱等都达到了90%及以上的完整回复率；而百度智能云、腾讯云TI平台和硅基流动普通版的完整回复率均低于50%，显示出当前稳定性具有一定的提升空间。

测评要点2：初步分析显示，国外付费第三方平台在回复率和推理耗时方面表现优于国内平台，而国内付费平台在准确率上则占据优势。具体而言，国外平台如Perplexity、together.ai和POE的完整回复率高达92%，高于国内硅基流动pro版和游晨云的83%。然而，在平均的准确率方面，国内平台以85%的成绩领先于国外平台的77%。

测评要点3：付费版第三方平台的稳定性显著优于免费版。付费版在完整回复率与推理耗时上明显优于免费版：付费版平均完整回复率88%，免费版65%；付费版推理耗时180.4秒，免费版261.4秒。准确率方面，两者均不低于80%，免费版87%略高于付费版80%。此外，同一平台的付费版相较免费版提升显著，回复率从0%飙升至90%。

基准说明

为帮助用户选择合适的第三方手机App平台，我们对支持DeepSeek-R1的10个平台进行了稳定性测评。测评使用原创小学奥数推理题，并从回复率、推理耗时和准确率等方面评估。

1. 测试时间：本次测试时间为2025年2月19日上午10：00—12：00以及下午14：00—18：00；
2. 测试过程：每人使用一部安卓手机测试一个App，对于第一个问题，所有人员同时输入问题并发送给测试的App，然后记录下模型完成每道题所花费的时间；待所有App完成该题，再同时输入第二个问题并发送给测试的App。以此类推，直至最后一题完成测试；
3. 测试环境：所有手机连接同一个WIFI，关闭其他后台程序。对于所有可调节联网设置的App，全部调整为非联网模式；对于没有该选项的App，保持默认。所有第三方App均无max_tokens参数选项，都采用其默认设置。
4. 时间记录：由于大部分App的回复不包含时间戳，所有推理耗时均按统一标准手动记录，可能存在轻微误差。

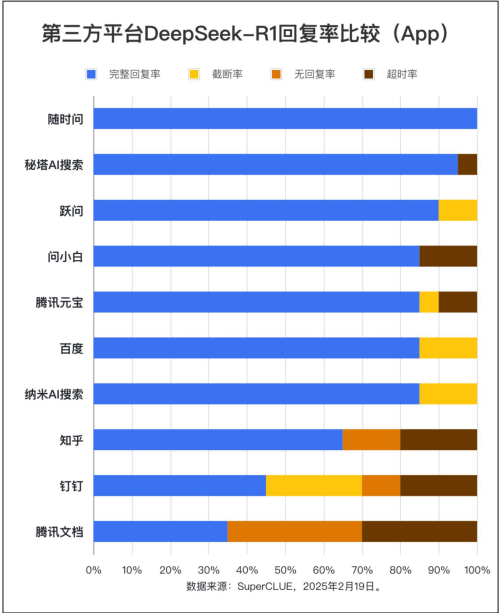
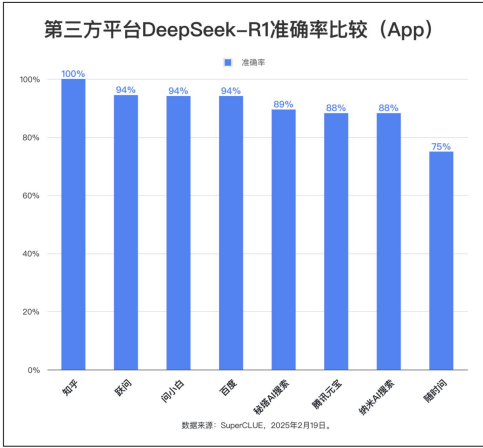
第三方平台DeepSeek-R1稳定性测评总榜（App）							
第三方平台	机构	完整回复率	截断率	无回复率	超时率	准确率	推理耗时（秒/题）
随时间	学而思	100%	0%	0%	0%	75%	129
秘塔AI搜索	秘塔科技	95%	0%	0%	5%	89%	151
跃问	阶跃星辰	90%	10%	0%	0%	94%	152
纳米AI搜索	360	85%	15%	0%	0%	88%	135
百度	百度	85%	15%	0%	0%	94%	146
问小白	元石科技	85%	0%	0%	15%	94%	213
腾讯元宝	腾讯	85%	5%	0%	10%	88%	219
知乎	知乎	65%	0%	15%	20%	100%	177
钉钉	阿里	45%	25%	10%	20%	-	-
腾讯文档	腾讯	35%	0%	35%	30%	-	-

数据来源：SuperCLUE，2025年2月19日；

注：1.对于完整回复率不足50%的第三方平台，准确率和推理耗时不参与统计；

2.推理耗时是指从问题发送给模型到模型完成回复所用时间，不包含系统延迟（TTFT和TPOT）等性能指标。

测评结果



- DeepSeek-R1稳定性测评分析：
- 各个App之间的稳定性存在较大的差异，主要体现在回复的完整性方面。
1. 在完整回复率方面，学而思的随时间、秘塔AI搜索和阶跃星辰的跃问三个App均不低于90%，其中随时间App达到了100%的回复率，均有极佳的稳定性，而钉钉和腾讯文档的完整回复率均低于50%，分别为45%和35%。
2. 在推理耗时方面，随时间、纳米AI搜索和百度App的平均每题推理耗时均在两分半以内，回复比较快。
- 出现这些差异的原因可能有以下几点：
- (1) 各个App最大输出长度的限制，导致模型的回复容易发生截断的情况；再加上题目较难和模型反思的特点会导致输出比较长；
- (2) 用户的数量也可能是稳定性存在差异的因素之一。

基准说明

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。

每一题的首token输出延迟：

ttft_per_query= chunk_time1 – start_time

每一题获取到答案的完整耗时：

latency_per_query = chunk_time2 – start_time

每一题生成输出的总时间（不包含处理输入的时间）：

latency_per_query_without_ttft = latency_per_query–
ttft_per_query

关键指标计算：

首token输出的平均耗时：

TTFT= 所有题目的ttft_per_query的平均值

获取答案的平均耗时：

Latency = 所有题目的latency的平均值

平均每秒输出的token数量：

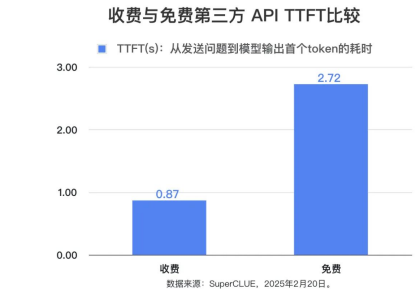
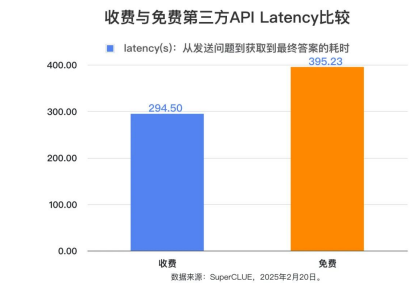
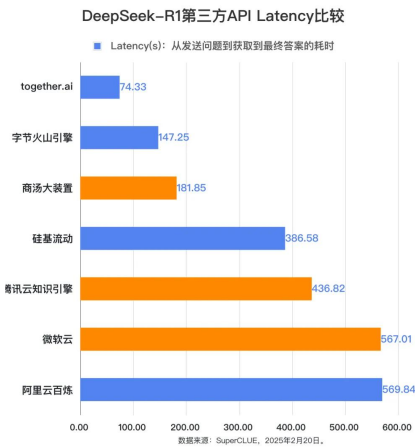
TPS = sum(completion_tokens-1) /
sum(latency_per_query_without_ttft)

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
🥇	字节火山引擎	字节跳动	100%	0%	0%	95.00%	27.94
🥈	商汤大装置	商汤科技	100%	0%	0%	90.00%	20.63
🥉	阿里云百炼	阿里巴巴	100%	0%	0%	70.00%	6.90
2	硅基流动	硅基流动	95%	5%	0%	94.74%	11.76
2	together.ai	together.ai	95%	5%	0%	89.47%	55.86
2	腾讯云知识引擎	腾讯科技	95%	5%	0%	84.21%	10.97
3	微软云	Microsoft	75%	0%	25%	93.33%	6.90

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

测评结果



DeepSeek-R1稳定性测评摘要：

测评要点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。

测评要点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。

测评要点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%。

联系我们

——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。

AI应用测评

提供AI大模型落地应用及工具测评，包括但不限于生产力工具、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

[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）3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。