SuperCLUE
中文大模型综合性测评基准

中文大模型基准测评2024年度报告

— 2024中文大模型阶段性进展年度评估

SuperCLUE团队

2025.01.08

SuperCLUE
中文大模型综合性测评基准
独立第三方AGI测评机构

精准量化通用人工智能（AGI）进展，定义人类迈向AGI的路线图

Accurately quantifying the progress of AGI,
defining the roadmap for humanity's journey towards AGI.

报告摘要（一）

- **OpenAI发布o1正式版，大幅领跑全球**

o1正式版的推出进一步拉大了与其他模型的差距。经12月测评，o1以80.4分大幅领跑全球，较ChatGPT-4o-latest高10.2分，较国内最好模型高12.1分。

- **国内顶尖大模型进展迅速，较为接近ChatGPT-4o-latest**

国内顶尖大模型进展迅速，其中DeepSeek-V3和SenseChat 5.5-latest取得68.3分表现出色，超过Claude 3.5 Sonnet和Gemini-2.0-Flash-Exp，较为接近ChatGPT-4o-latest（仅相差1.9分）。

- **国内模型在推理速度和性价比方面很有竞争力**

国内模型DeepSeek-V3和Qwen2.5-32B-Instruct在推理效能方面表现出色，在高水平能力的基础上，保持极快的推理速度。在性价比方面，DeepSeek-V3、Qwen2.5-72B-Instruct（阿里云）在高水平能力的基础上，保持低成本的API价格。

- **端侧小模型表现惊艳**

国内端侧小模型进展迅速，部分小尺寸模型表现要好于上一代的稍大尺寸模型，如Qwen2.5-3B-Instruct、MiniCPM3-4B，均展现出很高的性价比和落地可行性。

SuperCLUE模型象限（2024）

来源：SuperCLUE，2025年1月8日

报告摘要（二）

SuperCLUE
中文大模型综合性测评基准

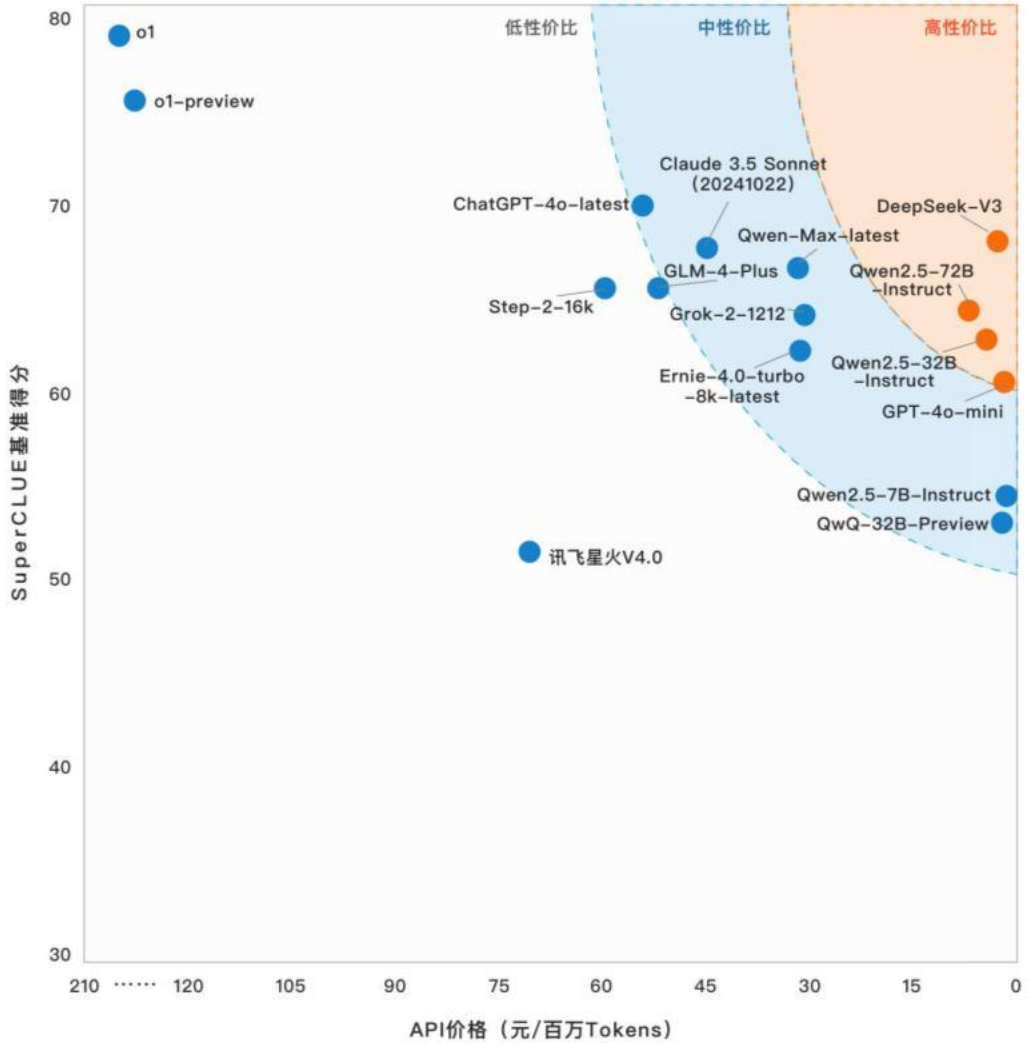
各维度国内Top3排行

一级维度	专项任务	国内TOP1	国内TOP2	国内TOP3
Hard	Agent	Step-2-16k（75.0分）	DeepSeek-V3 Qwen2.5-72B-Instruct（74.0分）	/
	指令遵循	Qwen-max-latest（35.7分）	TeleChat2-Large（34.3分）	DeepSeek-V3 SenseChat 5.5-latest（31.5分）
	深度推理	Baichuan4（60.2分）	360zhinao2-o1（59.4分）	DeepSeek-V3（58.8分）
理科	代码	Doubao-pro-32k-241215（75.2分）	DeepSeek-R1-Lite-Preview（71.2分）	DeepSeek-V2.5（70.9分）
	计算	SenseChat 5.5-latest（78.2分）	DeepSeek-V3 360zhinao2-o1（76.3分）	/
	逻辑推理	360zhinao2-o1（71.0分）	DeepSeek-V3（69.1分）	Doubao-pro-32k-241215（67.8分）
文科	语言理解	DeepSeek-V3（86.5分）	DeepSeek-R1-Lite-Preview（86.1分）	Qwen2.5-72B-Instruct TeleChat2-Large（84.7分）
	生成创作	Hunyuan-Turbo（76.2分）	NebulaCoder-V5（75.7分）	MiniMax-abab7-preview（75.6分）
	传统安全	SenseChat 5.5-latest（86.4分）	NebulaCoder-V5（82.9分）	Hunyuan-Turbo（82.5分）

来源：*SuperCLUE, 2025年1月8日*

注：专项任务排名中，当出现并列排名的情况（如并列第二），则后续排名依次顺延（第三名自动空缺）。

大模型性价比分布



数据来源：SuperCLUE，2025年1月8日；

报告目录

第1部分
2024年度关键进展及趋势

1. 2024年大模型关键进展
2. 2024年值得关注的中文大模型全景图
3. 2024年国内外大模型差距
4. 2024年国内外大模型能力趋势