

DeepSeek:
重塑全球AI格局的中国力量

国盛计算机

2025.5



国盛证券
GUOSHENG SECURITIES

目录

顶尖大模型发布进展不断

AI应用:MCP驱动Agent生态加速构建

AI应用:端侧/智驾/机器人/军工等

AI驱动中国科技资产重估

AI基建带动国产算力、云厂商需求



关注技术公众号

DeepSeek震撼科技圈：一月时间差并肩最强推理模型o1

2024.5 DeepSeek-V2 发布

提出MLA和DeepSeekMoE架构相比第一代DeepSeek 67B实现了更强的性能，节省了42.5%的训练成本，减少了93.3%的KV缓存

2024.11 推理模型 DeepSeek-R1.Lite 预览版 发布

媲美 o1-preview的推理效果并为用户展现了 o1 没有公开的思考过程

2024.12 DeepSeek-V3发布

DeepSeek-V3 671B在 2048 块 NVIDIA H800 集群上训练2个月，训练成本仅558万美元，达到GPT-4o和Claude Sonnet 3.5水准

2025.1.20 DeepSeek-R1发布

从数学（AIME/MATH）、编程(Codeforces/SWE)、学科推理(GPQA)的各个高难度benchmark结果来看，**DeepSeek-R1推理能力比肩OpenAI-o1-1217版本**。同时DeepSeek-R1蒸馏得到的Qwen和llama小模型也与OpenAI-o1-mini相当的效果。

2025.1.28 Janus-Pro、JanusFlow 发布

Janus-Pro是一款统一多模态理解与生成的创新框架，解耦视觉编码。JanusFlow 是一款通过生成流与自回归语言模型融合实现统一的框架，能生成高质量图像。

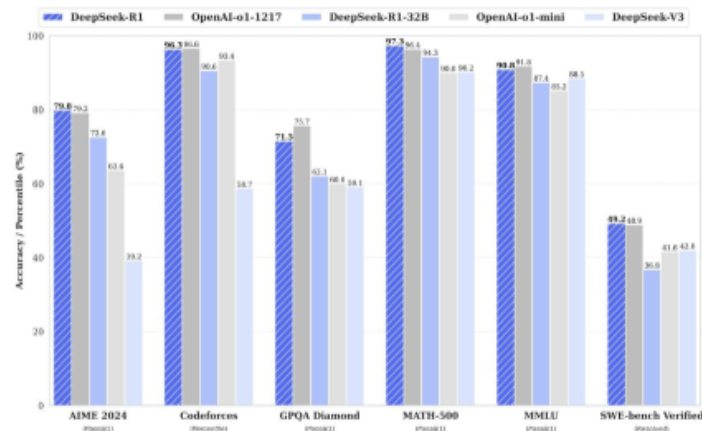


Figure 1 | Benchmark performance of DeepSeek-R1.

数学、代码能力追平o1

	AIME 2024 pass@1	AIME 2024 cons@64	MATH-500 pass@1	GPQA Diamond pass@1	LiveCodeBench pass@1	CodeForces rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759.0
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717.0
o1-mini	63.6	80.0	90.0	60.0	53.8	1820.0
QwQ-32B	44.0	60.0	90.6	54.5	41.9	1316.0
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954.0
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189.0
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481.0
DeepSeek-R1-Distill-Qwen-32B	72.6	83.3	94.3	62.1	57.2	1691.0
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205.0
DeepSeek-R1-Distill-Llama-70B	70.0	86.7	94.5	65.2	57.5	1633.0

蒸馏小模型效果超越o1-mini

DeepSeek震撼科技圈：开源、低成本倒逼OpenAI加速发布

开放的许可证和用户协议：DeepSeek-R1 统一采用标准化、宽松的 MIT License，完全开源，不限制商用，无需申请。产品协议明确可“模型蒸馏”。

OpenAI加速发布：

2月1日，OpenAI推出o3-mini, Plus和Team用户的速率限制从原来o1-mini的每天50条消息增加3倍到o3-mini的每天150条消息。
2月3日，OpenAI发布了基于OpenAI的o3模型之上开发而成的Deep Research。能够像人类分析师一样，对复杂的任务进行逐步分解，并在互联网上进行多轮的信息搜索与验证，直到找到最合适的答案
2月13日，OpenAI宣布将在未来几个月内推出GPT-5，该模型将整合OpenAI的大量技术，包括o3，在GPT-5推出之前，OpenAI计划在未来几周内先发布GPT-4.5代号“Orion”，这将是OpenAI最后一个“非思维链模型”

DeepSeek-R1 API

输入：
缓存命中
1元 / 百万 tokens

缓存未命中
4元 / 百万 tokens

输出：
16元 / 百万 tokens



Model	Pricing	Pricing with Batch API***
o1	\$15.00 / 1M input tokens	\$7.50 / 1M input tokens
	\$7.50 / 1M cached* input tokens	
	\$60.00 / 1M output** tokens	
Model	Pricing	Pricing with Batch API***
o3-mini	\$1.10 / 1M input tokens	\$0.55 / 1M input tokens
	\$0.55 / 1M cached* input tokens	
	\$4.40 / 1M output** tokens	

美国总统特朗普在佛罗里达表示：“中国公司发布 DeepSeek AI 应该给我们的行业敲响警钟，我们需要集中精力进行竞争。”

DeepSeek提出大量算法创新，中国从AI跟随者走到前沿探索贡献者

DeepSeek-R1:

DeepSeek-R1-Zero 提出不用监督微调直接进行强化学习，也能取得不错的效果。

DeepSeek-R1 加入少量CoT数据进行监督微调作为冷启动，然后再进行多阶段强化学习，可以取得更优的性能，同时回答更符合人类偏好。强化学习不需要进行过程监督和MCTS搜索，直接进行基于规则的奖励

DeepSeek-V3:

注意力层状态压缩(Multi-Head Latent Attention)：对attention层隐向量降维,减少推理时显存，提高推理效率

细粒度稀疏MoE架构：671B总参数，平均激活参数37B

多词元前瞻性预测(Multi-Token Prediction)：丰富训练监督信号，加速推理

混合浮点精度运算(FP8 Training)：对训练算子进行细粒度拆分，降低精度损失，首次在超大模型训练中验证FP8的有效性

PTX层优化：在比Cuda底层的编程语言上优化硬件效率

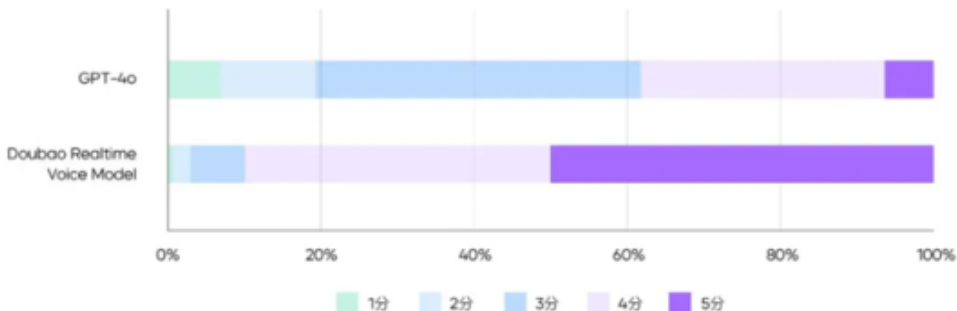
大厂模型持续进步：字节

字节豆包实时语音大模型情绪理解与表达能力突出。

1月20日豆包实时语音大模型在豆包 APP 全量开放，在情绪理解和情感表达方面与GPT-4o相比优势明显。豆包团队围绕拟人度、有用性、情商、通话稳定性、对话流畅度等多个维度进行考评。整体满意度（以5分为满分）方面，豆包实时语音大模型评分为4.36，GPT-4o为3.18。其中，50%的测试者对豆包实时语音大模型表现打出满分。在模型优点评测中豆包实时语音大模型在情绪理解和情感表达方面与GPT-4o相比优势明显。尤其是“一听就是 AI 与否”评测中，超过30%的反馈表示 GPT-4o “过于 AI”，而豆包实时语音大模型相应比例仅为2%以内。

满意度分值分布

总体满意度: Doubao Realtime Voice Model (4.36/5) > GPT-4o (3.18/5)



你好，我是豆包
我可以如何帮到你？

你可以尝试下面的示例...

AI 搜索



五一出游适合赏花的五个城市



阿联酋75年来最强降雨，天空秒变绿

PDF 问答

全球经济金融展望报告.pdf

2023年全球经济增长动力持续回落，各国复苏分化，发达经济体增速明显放缓，新兴经济体整体表现稳定。

展望2024年，预计全球经济复苏将依旧

图像生成



豆包团队评测语音大模型满意度超过GPT-4o

2025年1月，豆包MAU达到7816万，月增速10.47%