量子位 insights

2025上半年AI核心成果及趋势报告

2025.7 量子位智库 QbitAI Insights 分析师 Xuanhao xuanhao@qbitai.com

序言

- 人工智能可能是人类有史以来最重要的发明，我们也正在见证AI的飞速发展，技术突破与场景落地交织推动着行业加速演进。量子位智库将在本报告中为大家介绍2025年上半年，AI领域的关键动态和趋势，旨在为决策者、从业者和创新者提供前沿洞察，帮助他们在竞争激烈的生态中把握机遇。**本报告将从应用、模型、技术、行业4个方面进行展开：**

- **应用趋势**：包括通用类Agent开始进入主流、垂类Agent开始涌现、AI编程获得市场验证高速增长、模型上下文协议（MCP）获得行业关注等应用侧核心变化

- **模型趋势**：包括推理模型能力进步、工具使用能力落地、模型多模态能力增强、小模型加速应用普及、模型评估加速演化等模型层面的核心变化

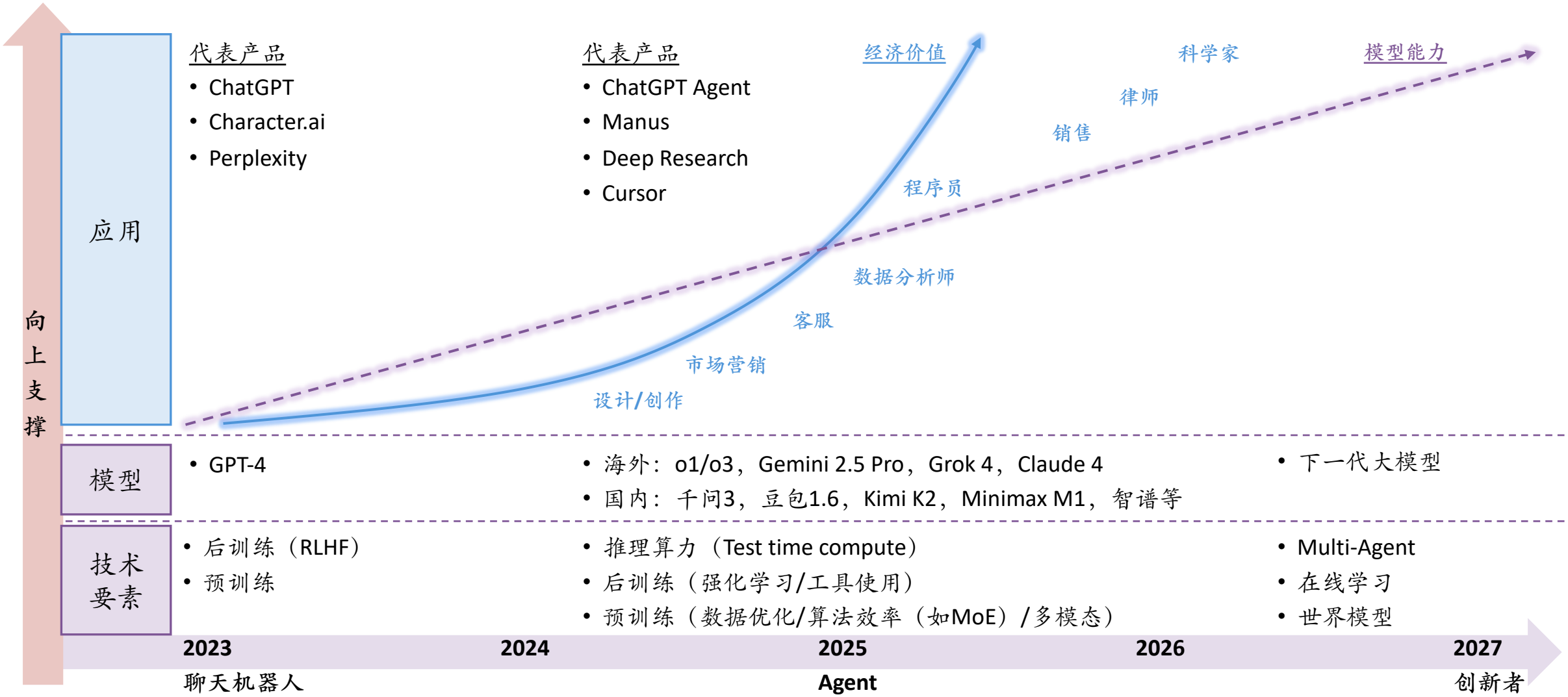
- **技术趋势**：包括模型不同训练阶段的重心变化、强化学习的重要性、多智能体（Multi-Agent）系统和在线学习的优势、新型模型架构迭代和应用情况等技术范式的核心变化

- **行业趋势**：主要介绍AI领域的行业核心动态，包括头部玩家在模型层的差距正在缩小、 OpenAI领先优势缩小，谷歌和xAI在上半年的竞争中迎头赶上、中美大模型的竞争差距缩小、AI编程成为目前必争之地等核心动态

目录

01 应用趋势

AI行业发展的底层逻辑是技术范式带来更强的模型能力，进而解锁更大的应用空间，加速价值创造



信息来源：量子位智库

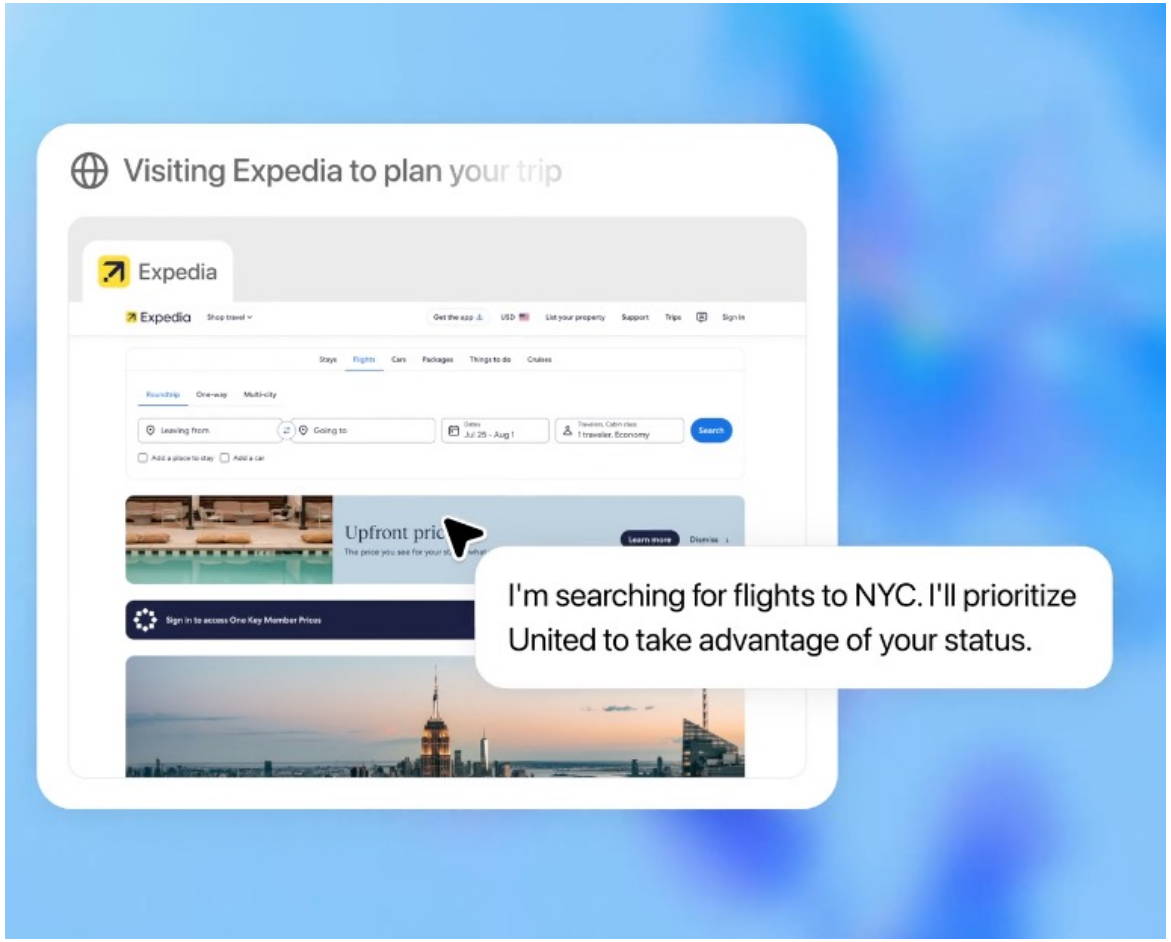
通用类Agent产品深度整合工具使用，主打完成场景多样的深度研究类任务，交付内容更加丰富，成为2025上半年应用亮点

阶段		核心技术	交付内容深度	交付形式	工作量	案例
聊天机器人		• 预训练模型 • RLHF • SFT监督对齐	• 文字，仅通过对话可以完成的简单任务，例如草拟简单的文字模版和语言翻译	文字报告	• 完成数分钟人类工作量，辅助完成知识类任务	ChatGPT OpenAI o1 deepseek
Agent	LLM	• Agent Planing框架：基于提示词和context对任务进行分解，生成执行步骤	• 检索互联网：搜索数十个甚至上百个信息源获取充足信息 • 调用工具获取数据：例如连接数据库获得准确、丰富的信息 • 深度生成：可以生成详尽完整的数千字深度报告	文字报告 图文报告 视频素材 网页文件 PPT	• 完成数小时人类工作量，自动化部分生产力	OpenAI ChatGPT Agent manus MiniMax Agent Kimi-Researcher Genspark flowith 扣子空间 秘塔 META SOTA
	工具	• 工具调用：调用或集成现有软件，如API 、搜索引擎、数据库				
	记忆	• 记忆能力：包括长期和短期的记忆能力，对话历史、文件知识库等				
	环境	• 沙盒环境：Agent具体执行任务的安全云端环境				

信息来源：量子位智库

以视觉操作为核心的Computer Use Agent（CUA）开始推向市场，代表了通用类Agent的另一条路径，正在与基于文本的深度研究类Agent融合

CUA技术示意图



分析

- CUA的基本原理是通过截取屏幕图像，利用模型的视觉能力，识别图形用户界面（GUI）中的按钮、菜单、文本字段等元素，通过虚拟光标和键盘输入与界面交互，执行点击、输入文本、滚动等操作

优势

- 多样化工具使用：让AI模拟人是AI接入互联网最快的方式，可以解决当前AI工具能力匮乏的问题，商业上也可以加快落地，应用基本无需改造即可让AI使用
- 打破数据孤岛：CUA能够访问到在不同应用上的所有信息，收集更多context，帮助用户作出更智能的决策

局限

- 运行成本高：依赖模型的视觉能力，图片处理导致成本较高；异步化难：CUA技术依赖屏幕截取，需要将计算机控制权交给AI（沙盒化虚拟机除外），C端场景下无法自动的完成身份验证；准确率不高：CUA 在简单网页任务上表现优异，但在复杂本地操作中仍有短板

OpenAI ChatGPT Agent | CLAUDE COMPUTER USE | Operator | GLM-PC

信息来源：量子位智库，OpenAI

受益于大模型在语义理解、多模态等方面的能力提升，垂直应用场景开始Agent化，自然语言操控功能正在成为垂类工作流的一部分

旅行

飞猪旅行

- 飞猪推出“问一问”功能，多个Agent协同工作，例如路线制定、交通票务查询、出行攻略以及酒店规划等Agent相互协同
- 可用自然语言在对话框提出、更改各类出行需求

- 技术基础：大模型能力提升，可以准确调用工具查询数据（例如机票、酒店信息），指令遵循能力增强可以理解用户意图

设计

LOVART

- 以自然语言交互为核心，同时整合大量专业设计功能，重塑传统视觉的工作流，一句话即可生成接近生产级的海报或视频

- 技术基础：图像生成模型能力提升，仅通过自然语言就能实现精准的图片生成和编辑（例如GPT-4o和Gemini的图像生成模型，以及其他3D资产生成模型）

创作

MINIMAX

- 以自然语言交互为核心，通过简单语言和图片输入，视频创作Agent就能自动分析、构思并生成具有专业水准、富有观看价值的完整内容，提升了创作效率

- 技术基础：新一代视频生成模型有更强的指令遵循、语义理解能力和编辑灵活度，生成视频的物理规律理解、对象一致性更佳

时尚

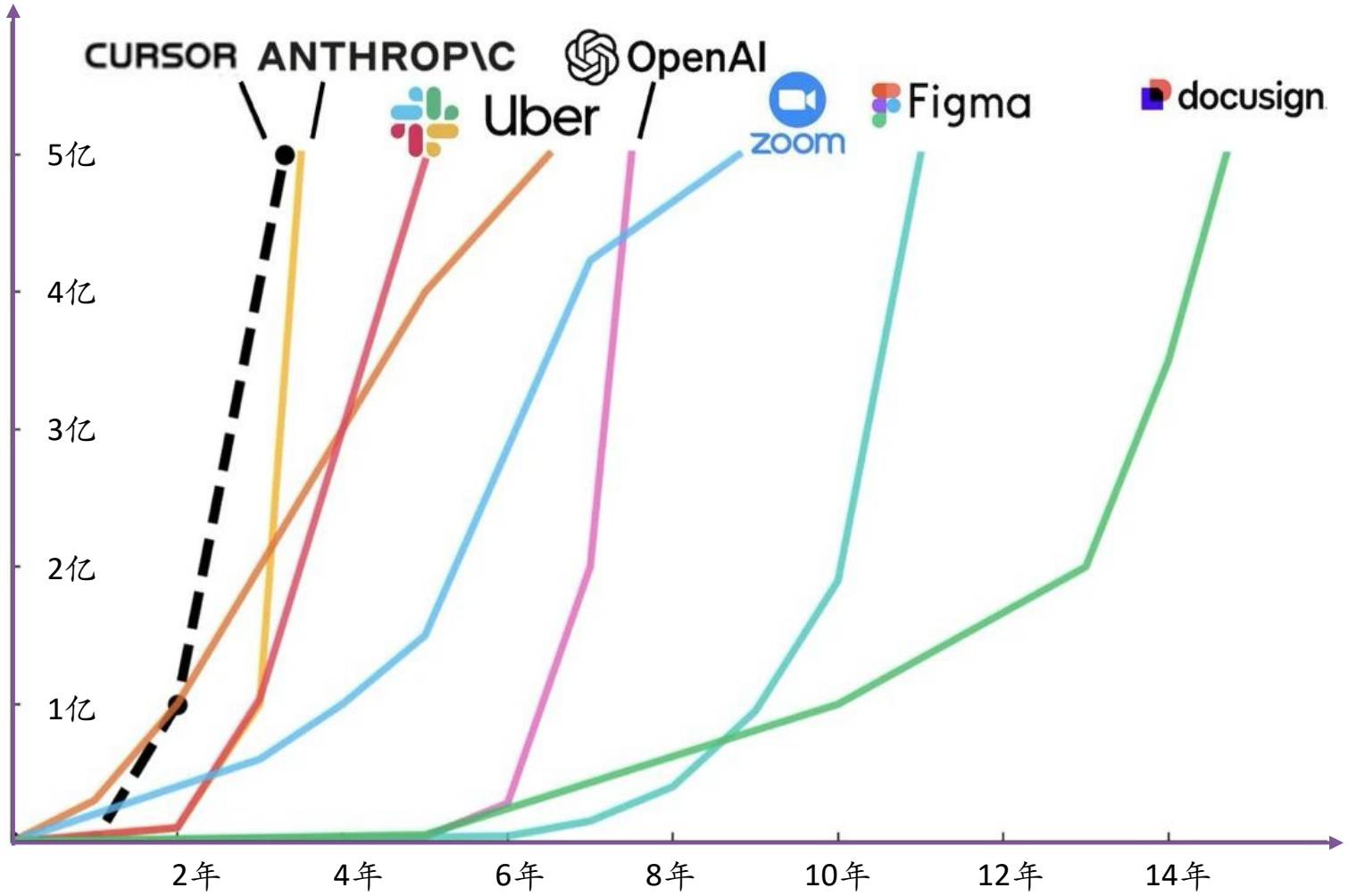
GENSMO

- 通过自然语言描述可以生成时尚穿搭，匹配相应的生活、工作、娱乐场景，让用户看到整体穿搭效果，也可以通过自然语言和用户图片一键生成成套搭配

- 技术基础：数字人技术的成熟、大模型语义理解能力和世界知识的增强，模型美学效果提升

AI编程成为当前最核心的垂类应用领域，正在从源头改变软件生产方式，头部编程应用收入增长速度创纪录，获得市场有效验证

不同应用达到5亿美元年收入所需时间



分析

- Cursor ARR[1] 突破5亿美金，证明了AI编程的价值空间，产品演化大概分为以下几个阶段：

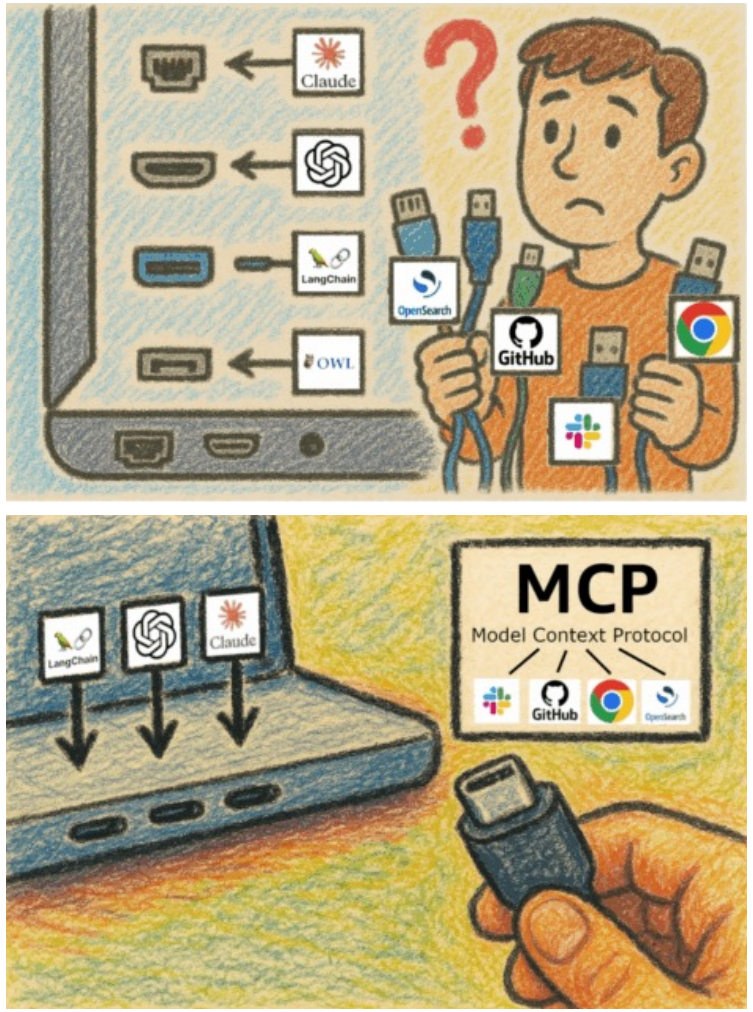
1. **代码补全**：通过理解代码上下文，预测用户的下一步编辑，主要是向后补全
2. **单文件代码编辑**：根据最近的修改和上下文，提供跨越多行的代码建议，适用于编辑单个文件或特定区域
3. **多文件同时编辑**：自动检索上下文，通过自定义的检索模型能够理解整个代码库，减少用户手动提供上下文的需要。可自动编写运行终端命令，创建、删除和修改文件，完成更复杂任务
4. **端到端交付**：后台运行任务，保留用户接管能力，适合并行处理多个任务，全流程云端容器化，用户聚焦验证和优化

自动化程度增加

信息来源：量子位智库，1）Annual Recurring Revenue，年度经常性收入

模型上下文协议MCP加速大模型应用普及，赋能模型获取大量外部信息、操控现有软件应用，打开更大应用空间，但尚未达到规模化生产级水平

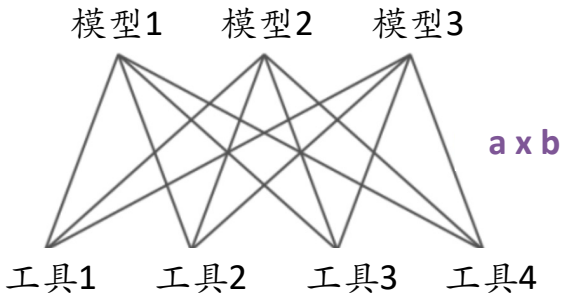
MCP示意图

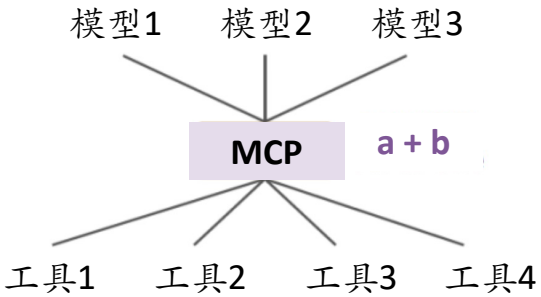


分析

优势分析

- MCP可为大模型提供标准化接口，主打高效安全调用外部数据和工具，为Agent提供技术底座和生态支持。传统的API调用面临调用方和被调用方之间急剧增加的复杂度，MCP则尝试把规范整合到一个通信协议中。MCP生态主要有3类玩家：客户端（MCP Client）、服务端（MCP Server）、MCP聚合平台





局限分析

- **MCP生态技术侧尚未成熟：**在大规模的生产级场景中落地较少，客户端（MCP Client）现在支持的调用数量相对有限（20-30个调用），服务端（MCP Server）虽然数量快速增加，但稳定性和可靠性参差不齐，限制应用普及
- **激励机制不完善：**部分软件供应商希望拥有自己的流量入口和用户关系，并不想成为被MCP抽象的API，没有动机积极开放、打磨自己的MCP服务端

前景展望

- 目前海内外头部互联网公司，如谷歌、亚马逊、阿里、字节等公司都在积极推动MCP生态发展，构建生态社区，随模型能力增强MCP将成为AI核心生态组件

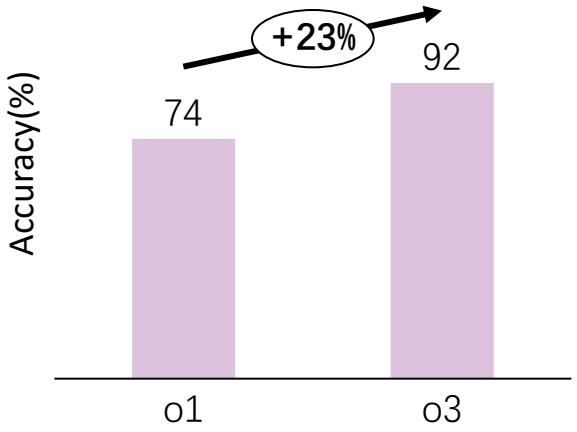
信息来源：量子位智库

02

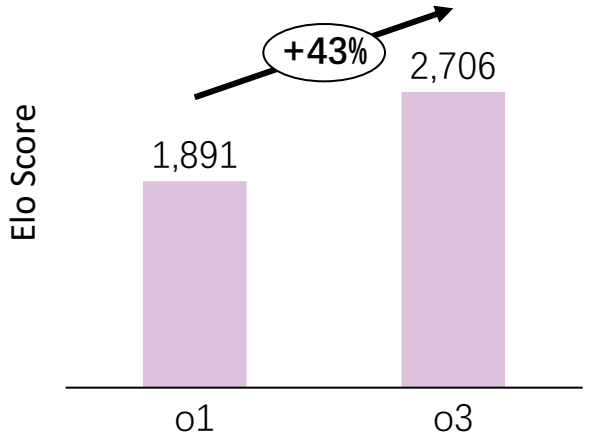
模型趋势

模型推理能力在思维链范式下，依然可以通过堆积更多算力持续提升模型能力，数理类、代码类问题提升尤其显著

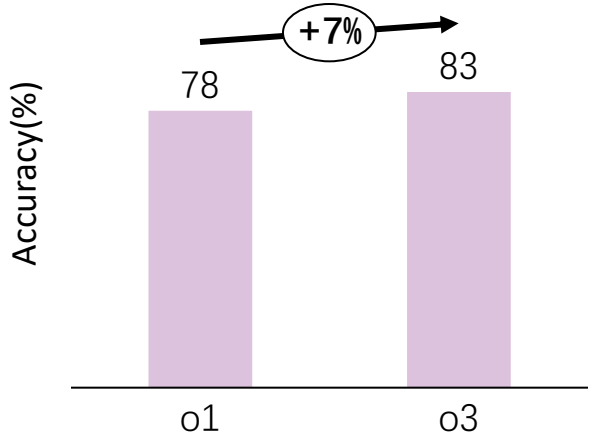
AIME 25（美国数学邀请赛）



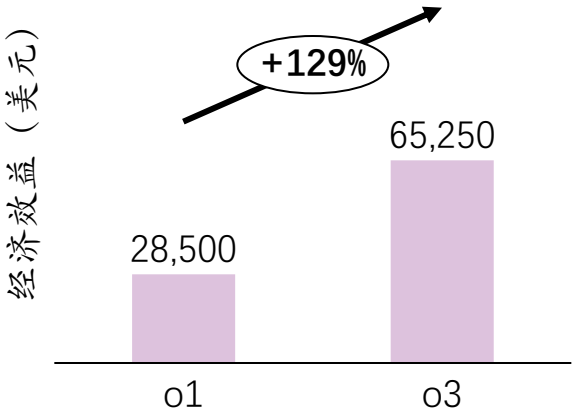
Codeforce 代码竞赛排名



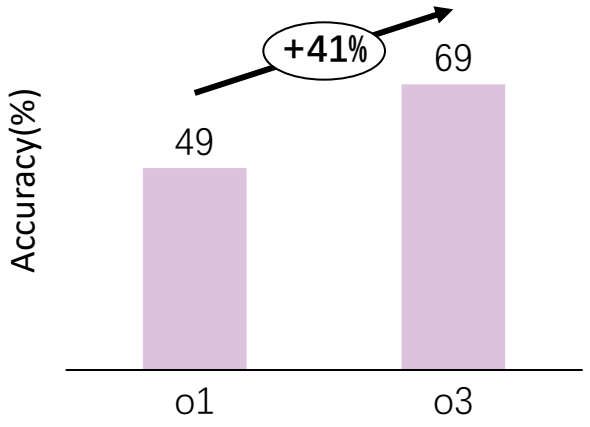
GPQA Diamond



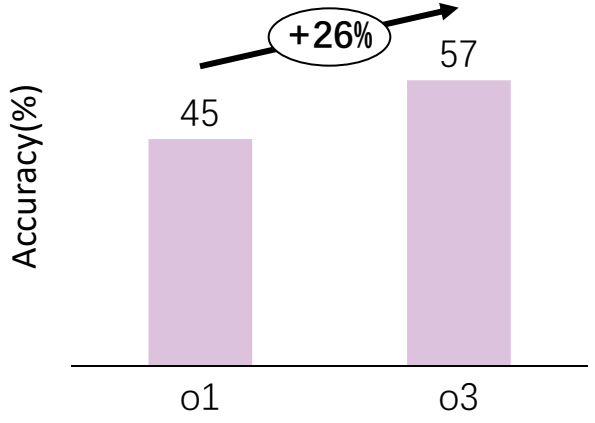
SWE-Lancer



SWE-Bench



多轮指令遵从[1]



分析

- 最近半年模型的智能上限在继续提升，自2024年底以思维链技术为核心的推理模型通用推理能力持续提升，证明了基于纯自然语言进行通用推理也能达到极高的智能水平
- 此外谷歌和OpenAI的实验模型已经可以用自然语言在IMO[2]中取得金牌水平，模型推理能力进展迅速

Google OpenAI

信息来源：OpenAI，1）榜单为：Scale MultiChallengeMulti-turn instruction following，2）International Mathematical Olympiad，国际奥林匹克数学竞赛

大模型开始走向Agentic，对工具使用进行端到端训练集成，相比仅基于文本的思维链推理有重大提升，可完成更复杂困难的任务

无工具使用能力	使用现有工具	发明新工具
• 利用模型已有知识进行推理，知识更新有限，解决真实世界问题的能力有限，主要能力在解决用户泛化的日常问询和低难度推理问题	• 模型可以在思维链中使用工具来增强其能力，例如在思考过程中裁剪或转换图像、搜索网页或用Python编译器分析数据	• 尚未解锁 • 未来模型将会像人类一样，不仅有能力使用现有的工具，也会自己开发合适的工具来解决问题
• 例如GPT-4o/OpenAI o1/DeepSeek V3/Gemini/Grok 3等模型	• OpenAI o3/o3 Pro • ChatGPT Agent	• 暂无
聊天机器人	智能体/Agent	创新型AI

Humanity's Last Exam榜单表现[1]

信息来源：量子位智库，1）以ChatGPT Agent为例

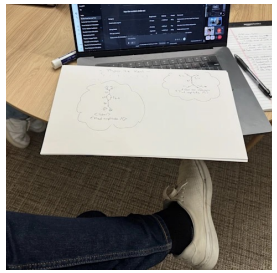
大模型开始端到端融合视觉和文本走向多模态推理，以语言为中枢逐渐解锁多模态推理的系统2[1]慢思考

工作流类视觉推理

- **VisProg**：视觉问答框架，通过大模型生成符号化程序来解决视觉任务，主要利用上下文学习能力，将复杂问题分解为可执行的子步骤，调用预定义的视觉工具 API（如目标检测、图像分割等）来完成任务
- **ViperGPT**：视觉推理框架，利用大语言模型生成可执行程序，但与 VisProg 不同的是，它直接生成 Python 代码，调用预定义的视觉 API（如目标检测、图像分割等）来回答基于图像的问题
- **Visual Sketchpad**：多模态语言模型框架，模拟人类绘制草图辅助推理的行为，允许模型通过生成代码调用绘图工具（如画线、框、标记等）或视觉模型（如目标检测、分割模型）来创建视觉草图，并根据这些草图进行动态规划和推理

无法Scale ✗

端到端视觉推理



- 多次调整放大读取手写的量子电动力学题目，精确提取文本和图表和专业公式，再运用思维链进行深度推理解决问题

- 搜索图像，找到公交车相关信息，放大精确读取文本，再通过网页搜索地理位置和车辆、站点信息，给出车辆通勤频率和运营时间表

- 根据图像信息推测真实位置，和大模型庞大的知识能力打通，搜索网页进行信息核实

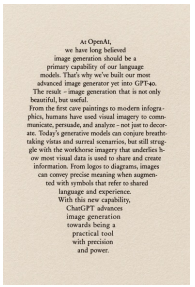
- OpenAI的o3模型尤其擅长视觉推理类任务，例如走迷宫、推箱子、做数独、图片找不同等，可以像侦探一样推理，放大照片局部细节同时调用工具进行多次推理检查
- 不足之处在于模型的性能不稳定，依然会出现较多幻觉问题，可靠性有限

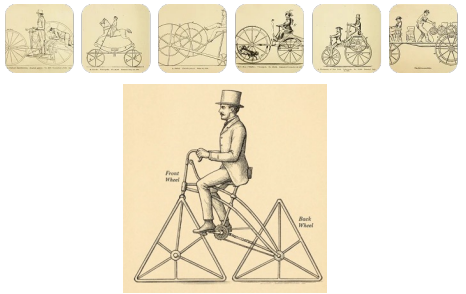
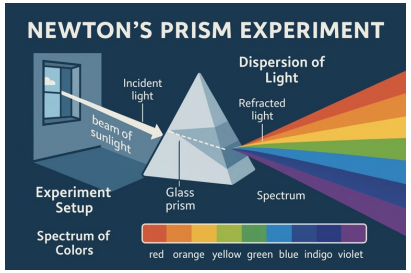
可以scale ✓

信息来源：量子位智库，1）“系统1”和“系统2”是心理学家丹尼尔·卡尼曼（Daniel Kahneman）在其著作《思考，快与慢》中提出的概念，用以描述人类思维的两种不同模式

大模型图像生成能力全方位增强，语言理解能力升级和审美提升是最大亮点，普通用户可以仅通过自然语言进行完整创作

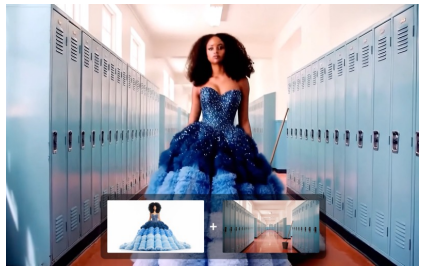
GPT-4o 图像生成

	介绍	图示
文字渲染增强	• 文字生成的控制能力显著增强，可以生成文字清晰的段落，渲染效果好 • 例如直接将大段文字食谱渲染在菜单中	
复杂指令理解	• 指令遵循能力强，可以理解复杂指令并在生成结果中体现 • 例如在一次生成中同时遵循16个细节指令	
多轮对话编辑	• 可以通过多轮对话对生成内容进行连续编辑、再生成，加速创意工作流程 • 例如将之前生成的文字内容无缝贴合到新背景	 

	介绍	图示
强上下文关联	• 可关联多种输入，例如输入多张图片进行风格学习，按用户要求进行类似风格图片的创作 • 例如生成类似风格手稿	
审美品味提升	• 生成内容的艺术性、审美显著提升，例如生成逼真的吉卜力动漫风格 • 例如生成高拟真度的照片和自拍、吉卜力风格图片	
知识理解增强	• 知识增强，能够从大模型中直接获取到世界知识，生成与现实世界知识相符的图像 • 例如一句话生成解释物理原理的信息海报	

信息来源：量子位智库

视频生成模型整合原生配音，可控性和编辑灵活度增加，生成视频的物体一致性和物理规律协调性增强，AI视频商业化和普及度进展积极

		介绍	图示
Veo3	原生音画同步生成	• 视频生成可以同步匹配背景音效和人物语言 • 例如生成右图视频，海浪的背景声和人物对话均由Veo 3生成，且唇动精准度和语言高度吻合	
	运动更加可控	• 可以选中图片中的多个物体进行精细运动效果控制 • 例如右图视频，可以选中多个立方体，用箭头标示物体运动方向	
	参考方式多样	• 可以用多张图片相互参考，用自然语言决定组合方式 • 例如右图视频，可以将两张图片组合为一个视频，方便用户利用现有素材	

		介绍	图示
海螺视频		• 增强了对于复杂物理运动的理解，模型学到了更多物理规律，一致性更强 • 例如可以生成体操运动等之前视频模型有缺陷的输出场景	
快手可灵		• 发布可灵2.0/2.1，增强了生成视频的细节精细度，商业化方面有积极进展，月收入已达到1400万美元	
字节Seed Dance		• 字节的Seeddance 1.0模型在Artificial Analysis的榜单中视频生成功能中排名第一，目前已经整合到即梦供创作者使用	

信息来源：量子位智库

模型智能密度持续提升，模型厂商积极推出小模型实现极致性价比，降低模型部署硬性门槛，加速模型应用普及

玩家	模型	介绍
阿里巴巴	• **Qwen 3 系列小模型**	• Qwen3-0.6B/1.7B/4B：上下文窗口为32K。Qwen3-8B/14B/32B：上下文窗口为128K。稠密模型，可用于边缘计算等场景，能在低算力设备上运行，性能全面
ByteDance 字节跳动	• **Seed-Coder小模型**	• Seed-Coder-8B-Base/Instruct/Reasoning三个版本，专注于代码生成，通过自身生成和筛选高质量训练数据，可大幅提升模型代码生成能力
智谱·AI	• **GLM系列小模型**	• GLM-4-9B-0414/GLM-Z1-9B-0414/GLM-4.1V-9B-Thinking：分别主打对话和推理，上下文长度32K，其中GLM-4.1V-9B-Thinking在多项测评中拿下SOTA
deepseek	• 蒸馏系列模型	• DeepSeek-R1-Distill-Qwen-1.5B/7B、DeepSeek-R1-Distill-Llama-8B等，基于千问模型较小的base model进行微调，方便AI社区部署
xiaomi	• **Xiaomi MiMo**	• 为推理（Reasoning）而生，联动预训练到后训练，全面提升推理能力；在数学推理（AIME 24-25）和代码竞赛（LiveCodeBench v5）上表现优异
Google	• **Gemma 3系列模型**	• 轻量级模型，支持超过35种语言，具备分析文本、图像及短视频的多模态能力，有1B、4B、12B、27B等多个版本，上下文为128K。Gemma 3n，专为低资源设备打造的多模态AI模型，仅需2GB RAM即可在手机、平板和轻薄笔记本上流畅运行，有音频处理能力，支持文本、图像、视频和音频的实时处理
Microsoft	• **Phi 4系列模型**	• 包括Phi-4 Reasoning/Mini-Reasoning/Reasoning-plus等多个版本，参数规模在140亿左右，可以在消费级硬件上运行，上下文窗口为32K

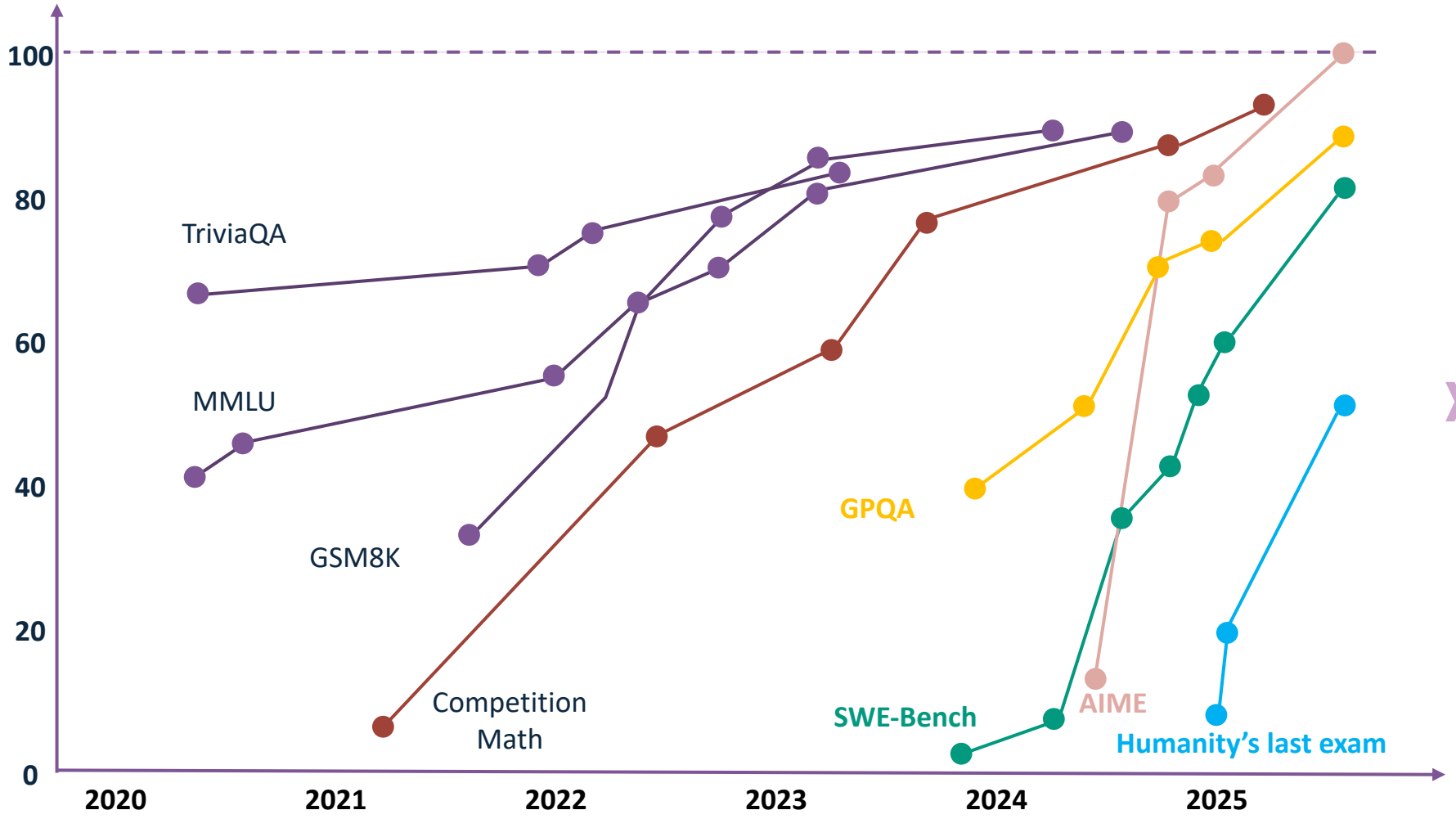
分析

- 运行小模型对于硬件的要求极低，可以在端侧设备上部署试用
- 小模型的输出成本极低，性价比远超闭源的大模型，适用于对模型能力要求低但token用量大的场景，例如AI陪伴、AI搜索等场景

信息来源：量子位智库，1）Sim2Real（Simulation to Reality）是指将人工智能模型从模拟环境（simulation）训练中获得的知识和能力转移到现实世界（real world）中应用的过程

模型评估加速演化，传统评估榜单快速饱和，可以动态更新，能在真实世界产生使用价值任务成为重要评估方向

主流模型基准测试评估变化趋势（Accuracy，%）



分析

- 随着模型能力增强，真实反映AI的客观能力正变得越来越困难，传统的静态榜单已经趋于饱和，未来将不再是最重要的评估方式
- 未来的模型评估，除对单纯智能的考验外，也需要一套对齐现实世界专家能力的实用性任务体系，重点考察实用性任务、商业价值或者经济产出

OpenAI
- 推出HealthBench：衡量AI在医疗健康领域能力的全新基准测试

HONGSHAN 红杉中国
- 推出xBench：衡量AI在HR、销售等领域落地商业价值的基准测试

信息来源：量子位智库

03

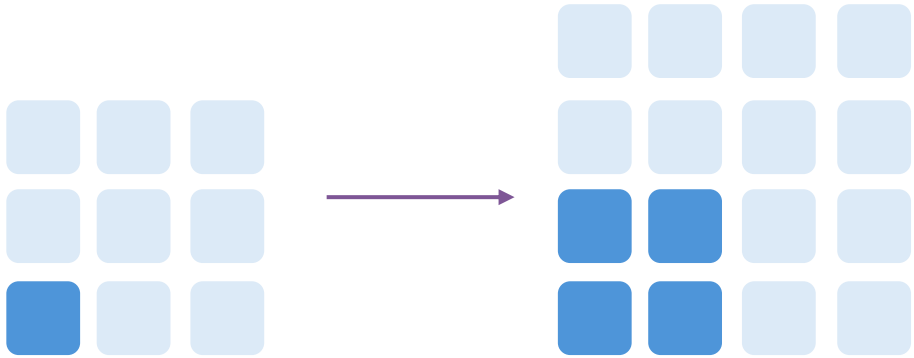
技术趋势

训练阶段上，资源投入向后训练和强化学习倾斜，但预训练仍然有充足的优化空间，二者最终共同决定模型能力

模型能力 =

预训练

隐性能力 显性能力

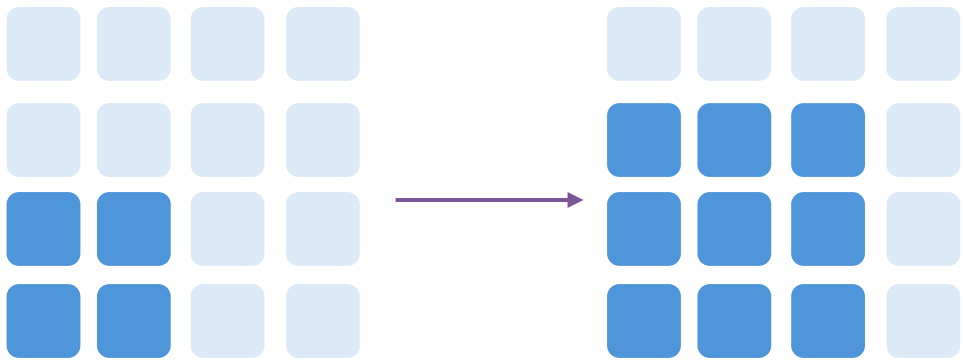


- 预训练阶段的主要作用：让模型学会更多基础能力和世界知识，但很多能力并不会在基座模型阶段直接表现出来，或者表现出来的能力比较弱（例如推理能力）
- 从资源投入的角度看，预训练阶段的投入ROI在下降，但由于直接影响后训练和推理阶段的模型能力，整体来看投入的必要性依然很高

×

后训练

隐性能力 显性能力

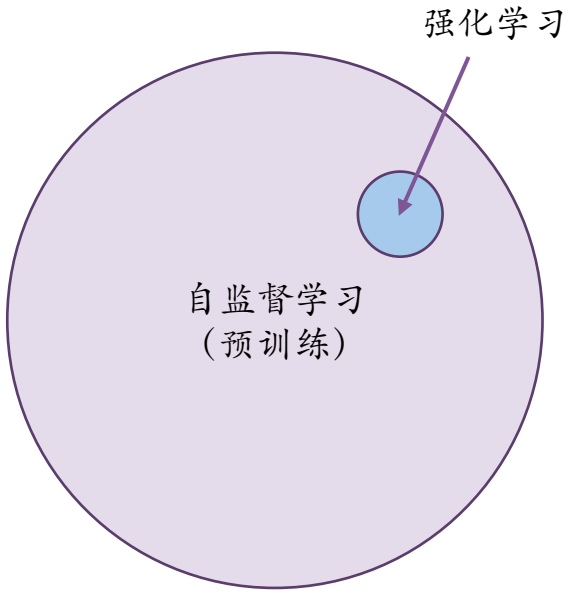


- 后训练/RL阶段的主要作用：激发基座模型的能力，把基座模型的已有能力放大，或者激发新的能力
- 但后训练和强化学习对模型智能上限的提升存在争议，主要作用在于用更高的概率选出基座模型中的正确推理路径，可能不创造新能力，例如在GPT-2/3上进行大量的后训练也无法获得好的推理能力

信息来源：量子位智库

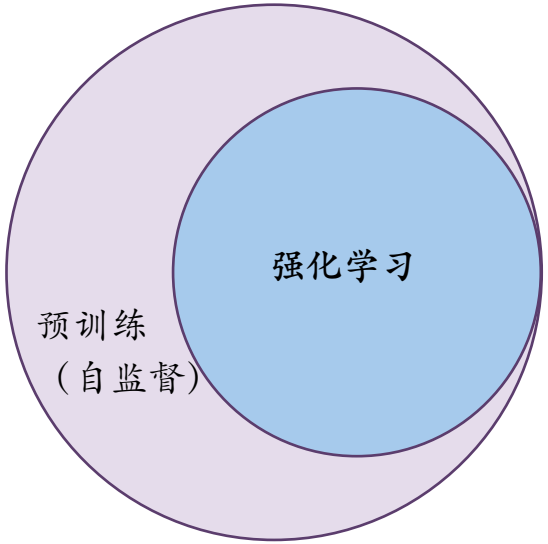
强化学习的重要性继续提升，算力消耗在未来会超过以自监督学习为核心的预训练，未来会从数学、代码等奖励清晰的领域向其他领域泛化

o1



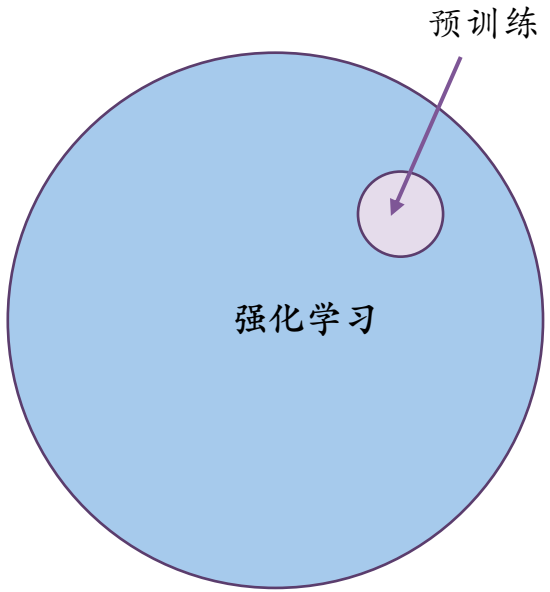
- OpenAI的o1模型是大规模强化学习在大模型领域落地的里程碑

o3



- OpenAI的o3模型、DeepResearch和Agent类产品，开始在训练过程中加入工具使用等复杂能力，对于强化学习的算力要求更高

未来



- 强化学习在未来将成为驱动模型智能的核心，从训练方式上看会成为算力消耗最大的部分

分析

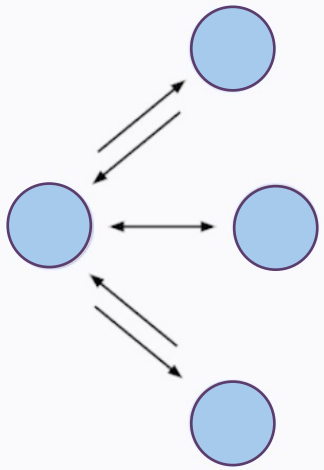
- 强化学习的关键在于如何设置奖励模型，对于代码、数学等有明确答案的领域，奖励模型的设置相对容易
- 对于没有清晰反馈的领域，目前采用的方法是通过专门的模型对表现进行评分
- 模型在集成复杂的工具调用等其他能力后，强化学习训练评估的难度也在增加

信息来源：量子位智库

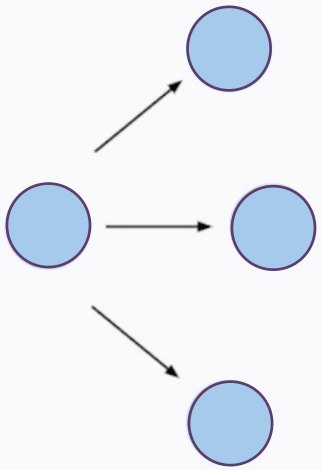
多智能体（Multi-Agent）系统可能成为继思维链推理模型之后的下一个前沿范式，继续提高智能上限

Multi-Agent基本模式介绍

领导型Agent + 执行型Agent

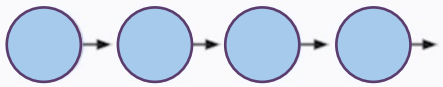


串行

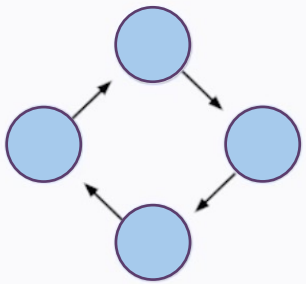


并行

固定工作流

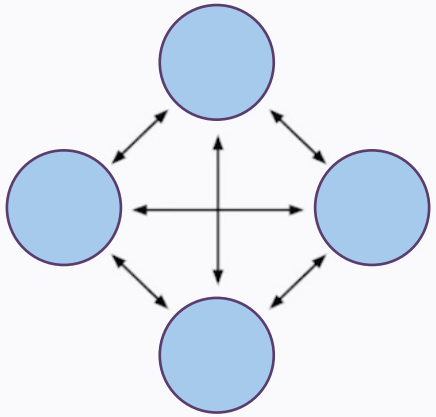


串行



环形

Agent群



多对多交互

分析

- Multi Agent有如下优势：
- 分布式处理并行工作，提高效率和计算速度，适合大规模动态环境
- 高效利用上下文：单个Agent基于局部信息决策，减少对全局信息的依赖，避免上下文污染
- 能力多样化：不同Agent有不同的知识、技能，可通过信息共享产生更优的解决方案
- 鲁棒性与容错性：单个Agent的故障不会导致整个系统失效

- 上下文限制：单Agent接收过多上下文有无法聚焦关键信息的问题
- 工具调用限制：模型可以调用的工具非常多，单个模型很难做出有效调用决策
- 领域知识限制：单个模型没有垂直领域的知识建构

单Agent劣势

Grok 4 Claude manus

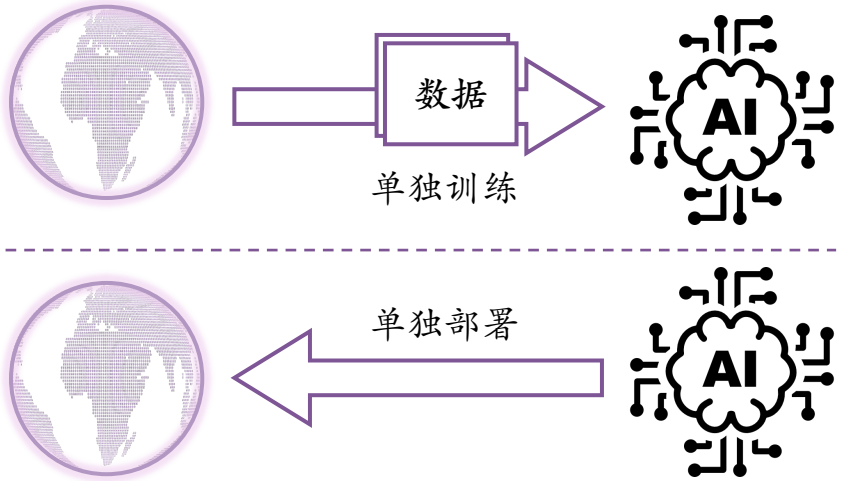
- Grok 4 Heavy已采用Multi-Agent架构
- Claude的Research功能已采用Multi-Agent架构
- Manus已采用Multi-Agent架构

业界落地案例

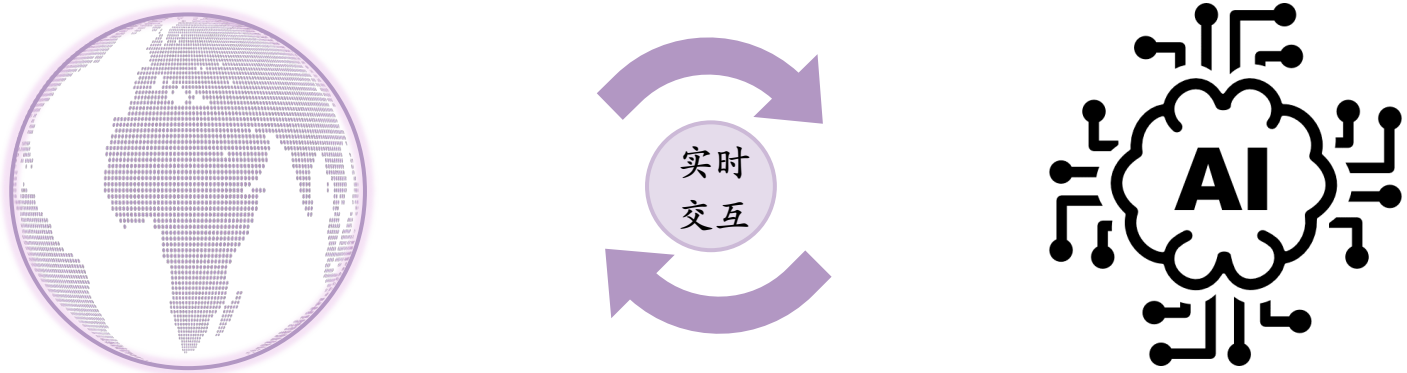
信息来源：量子位智库

从交互经验中学习有希望成为下一代模型学习方式，正在成为核心突破方向，可使模型摆脱对人类数据的依赖，提高智能上限

静态大模型



在线学习大模型



当前局限

- 高质量数据稀缺：目前模型学习范式高度依赖大量人类生成的数据，但高质量数据正变得稀缺或难以获取，继续扩展的收益正在边际递减
- 数据质量存在上限：由人类生成的数据在智能层面存在上限，难以达到超级智能水平

核心特征

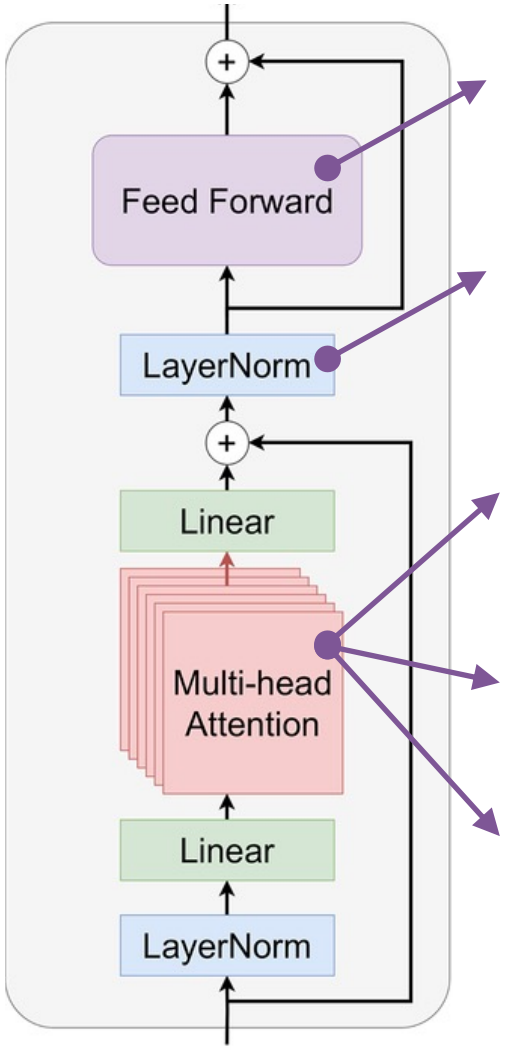
- 持续经验流：Agent在长期、连续的经验中学习，行为随过往经验自适应调整，可追求长期目标，不局限于短期交互；行动和观察能力：Agent可自主在现实世界行动，通过传感器、工具与环境交互，如操作设备、监控数据，而非依赖人类对话；吸收现实反馈：奖励信号来自环境结果（如健康指标、实验数据、用户反馈），而非人类预判，可突破人类认知局限，且能通过试错灵活调整

DeepMind

- Google Deepmind和强化学习之父Richard Sutton联合提出“经验时代（Era of Experience）”，强调从与世界实时交互中学习的重要性

信息来源：量子位智库

Transformer模型架构正在快速迭代，优化主要集中在注意力机制和前馈神经网络等层面，在工业界有多个落地案例



玩家	具体工作	核心优化点	介绍
ByteDance 字节跳动	• UltraMem	• 前馈神经网络稀疏化	• 结合大规模、超稀疏内存层来解决访存开支和推理延迟的问题，在保持模型性能的同时显著降低了推理延迟，不仅具有良好的扩展属性，而且优于传统模型
Massachusetts Institute of Technology	• Dynamic Tanh	• 替代归一化层	• 引入DyT使无需归一化的Transformer能够匹配甚至超越其归一化版本的性能，且大多无需超参数调整。在多种场景下验证了引入DyT的Transformer的有效性
deepseek	• Native Sparse Attention（NSA）	• 动态可学习注意力机制	• 采用分层设计有效减少计算量，聚焦关键信息，同时捕捉全局和局部上下文，在通用基准测试、长上下文任务和基于指令的推理中，NSA表现与全注意力模型相当甚至更优，同时针对硬件优化、
MINIMAX	• MiniMax-01 • MiniMax-M1	• 线性注意力机制 • 闪电注意力机制	• MiniMax-01首次大规模实现线性注意力机制，能够高效处理全球最长400万token的上下文，推理模型MiniMax-M1采用了闪电注意力，支持100万上下文
Moonshot AI	• Mixture of Block Attention（MOBA）	• 稀疏注意力机制	• 允许模型自主决定注意力分配，而不是依靠预定义，将专家混合（MoE）原则应用于注意力机制，能够在完整注意力和稀疏注意力之间无缝切换，在不影响性能的情况下提升效率

信息来源：量子位智库

Transformer混合架构正在涌现，以RNN变体为主，已经出现在工业界大规模应用先例

	玩家	具体工作	介绍	时间
新型RNN	Tencent 腾讯	• 腾讯混元T1模型	• 混元T1正式版沿用了混元Turbo S的创新架构，采用Hybrid-Mamba-Transformer 融合模式，这是工业界首次将混合Mamba架构无损应用于超大型推理模型，这一架构有效降低了传统Transformer结构的计算复杂度，减少了KV-Cache的内存占用，从而显著降低了训练和推理成本	• 2025.3 工业落地
	RWKV	• RWKV-7	• 一种新的序列建模架构，内存使用量和每个 token 推理时间恒定。与其他先进模型相比，训练 token数量大幅减少，但2.9B参数的小规模语言模型在多语言任务上达到了最先进水平（SOTA）	• 2025.3
新型CNN	NVIDIA	• MambaVision	• 一种新颖的混合Mamba-Transformer的模型，专门为视觉应用量身定制，核心贡献包括重新设计Mamba公式，以增强其高效建模视觉特征的能力。研究表明，在Mamba架构的最后几层配备自注意力模块显著提升了其捕捉长距离空间依赖的能力，MambaVision模型在ImageNet-1K数据集上的分类任务中SOTA性能	• 2025.3
其他	Google	• Titans	• 一个新的模型架构，结合短期记忆和长期记忆的模型架构，注意力机制由于其上下文有限但依赖建模精确，表现为短期记忆；而神经记忆由于其数据记忆能力，充当长期、更持久的记忆，Titans是基于这两个模块的新架构家族	• 2025.1

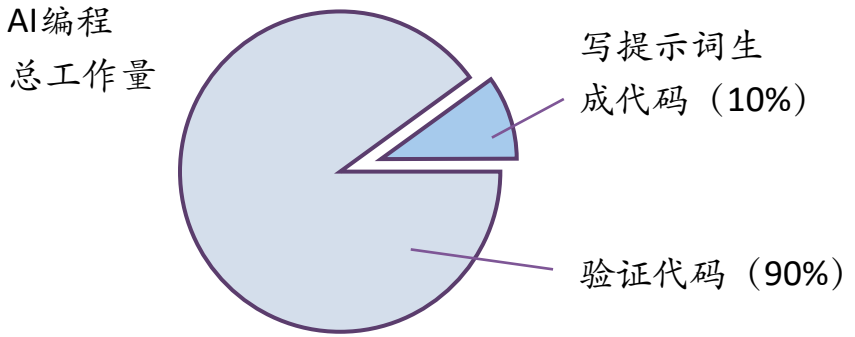
信息来源：量子位智库

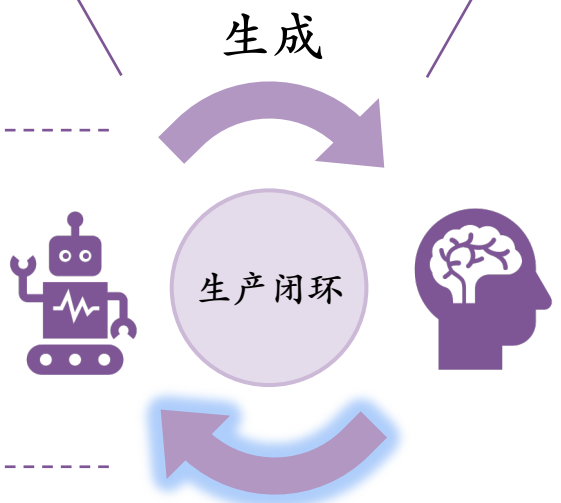
由于生成和验证难度的不对称性，代码验证成为目前AI编程自动化水平提升的前沿方向，可进一步加速软件生产自动化

生成

- 大模型极大加速了代码的生成速度，模型可以短时间生成复杂代码，但目前细节上的指令遵循、意图理解和有效性依然不足，阻碍了AI编程的大规模自动化

- 目前AI编程或者“Vibe Coding”难以直接产出生产级的代码和应用，主要是加速开发者完成细分的功能和逻辑，同时需要辅以大量人类反馈和修改，写提示词生成只占据开发者的少量注意力





来自人类的反馈验证是瓶颈

验证

- 代码和文本需要用户进行仔细阅读和推理，有效验证的成本、时间、门槛较高
- 目前解决验证问题思路是把复杂问题拆解成容易验证的多个小问题，例如绘画过程，一边作画一边进行调整，生成和验证相互交织

解决验证问题的关键原则

1. 客观事实：有公认的最优解
2. 快速验证：可以在极短时间内完成验证
3. 可扩展验证：可同时验证多个解决方案
4. 低噪音：验证与解决方案质量尽可能强相关
5. 客观事实：有公认的最优解
6. 持续奖励：针对单个问题，很容易对多个解决方案的优劣进行排序

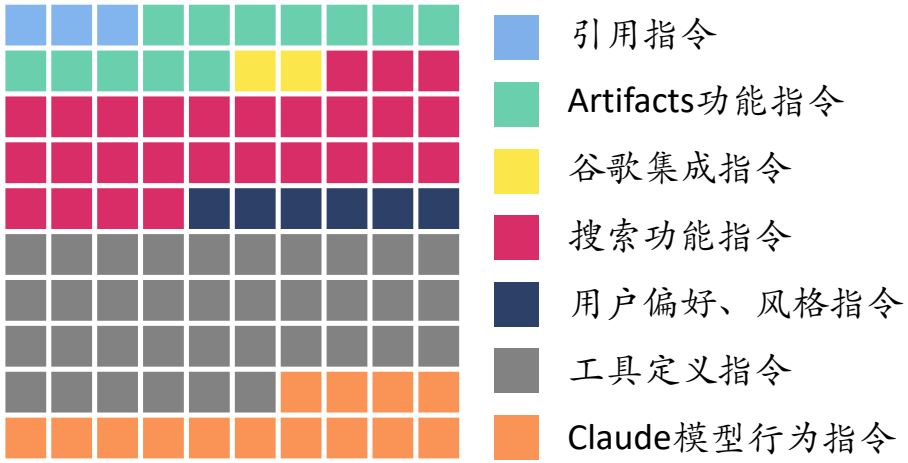
信息来源：量子位智库

系统提示词（System Prompt）正在成为决定模型用户体验的关键技术要素，相比更新大模型更加轻量化、敏捷化

用户体验 = 系统提示词 × 后训练 × 预训练

	核心改动	更新成本	主要场景
系统提示词	• 不改变模型参数	• 极低，可实时根据模型厂商和用户的反馈更新	• 提供系统级指令来指导模型的行为、角色和响应风格，同时可以实现更高程度的个性化
后训练	• 更改小部分模型参数	• 较低，需要重新训练后部署	• 通过监督微调、强化学习（如RLHF）等方法使模型适应特定任务、提升安全性、对齐人类偏好
预训练	• 更改全量模型参数	• 极高，成本高周期长	• 在海量无标签数据上从头训练模型，学习语言模式、语法、知识和世界表示等基础能力，建立模型的通用理解基础

Claude 模型系统提示词分布



- 聊天机器人不止底层的大语言模型，而是整合了工具、指令和持续优化的复杂系统
- 以如上的Claude模型系统提示词为例，一共约1.7万字，系统提示词相当于LLM的“设置菜单”，定义了语气、工具使用和上下文信息，帮助模型应对实际交互中的问题
- 未来系统提示词将走向个性化，通过各种用户数据为每位用户生成定制化提示词，高效增强用户体验

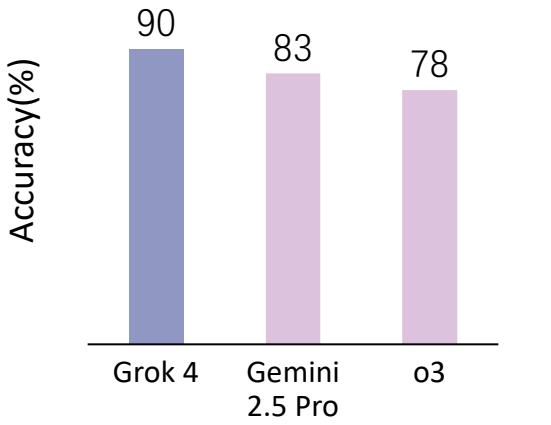
信息来源：量子位智库

04

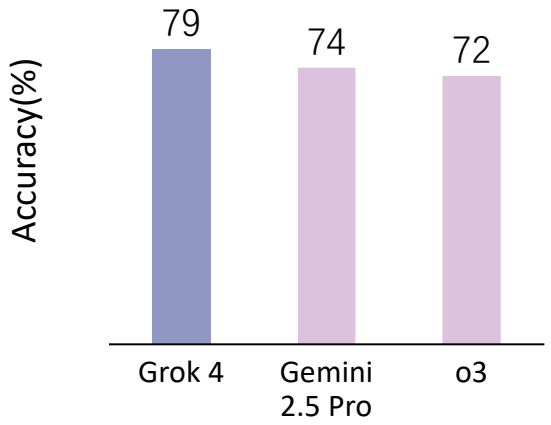
行业趋势

xAI发布Grok 4在多个领域达到SOTA水平，跻身全球大模型第一梯队，证明大模型本身没有护城河，正在改变模型层竞争格局

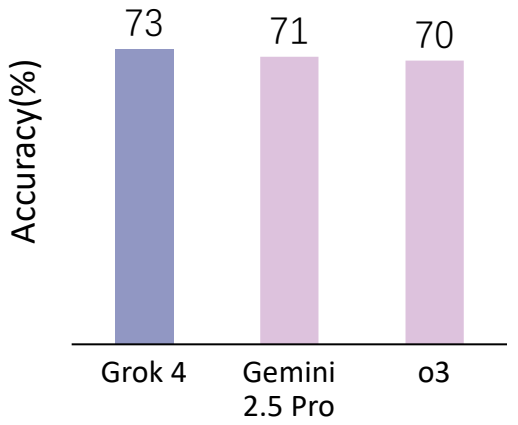
HMMT-25



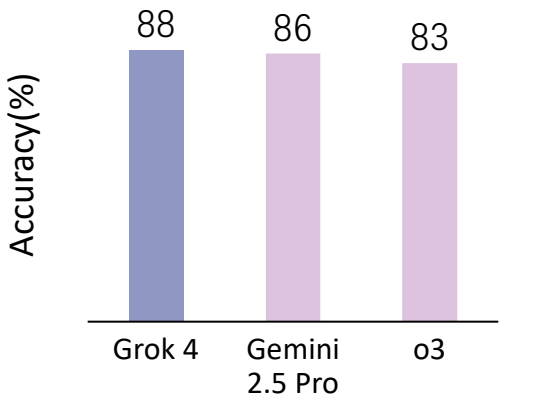
LiveCodeBench-(Jan-May)



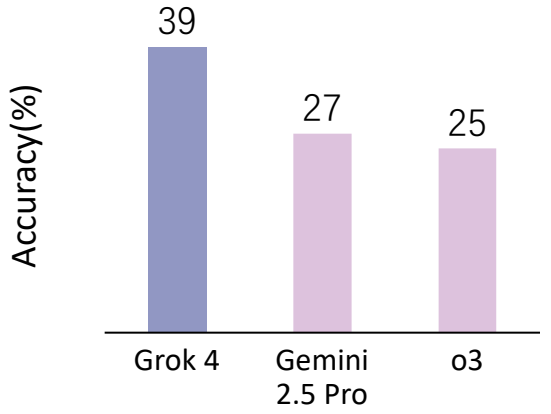
Artificial Analysis Intelligence Index



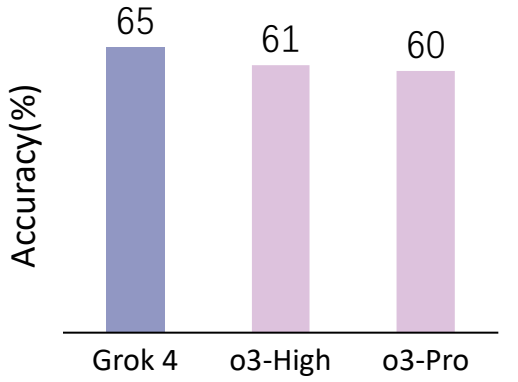
GPQA



Humanity's Last Exam



xbench-ScienceQA



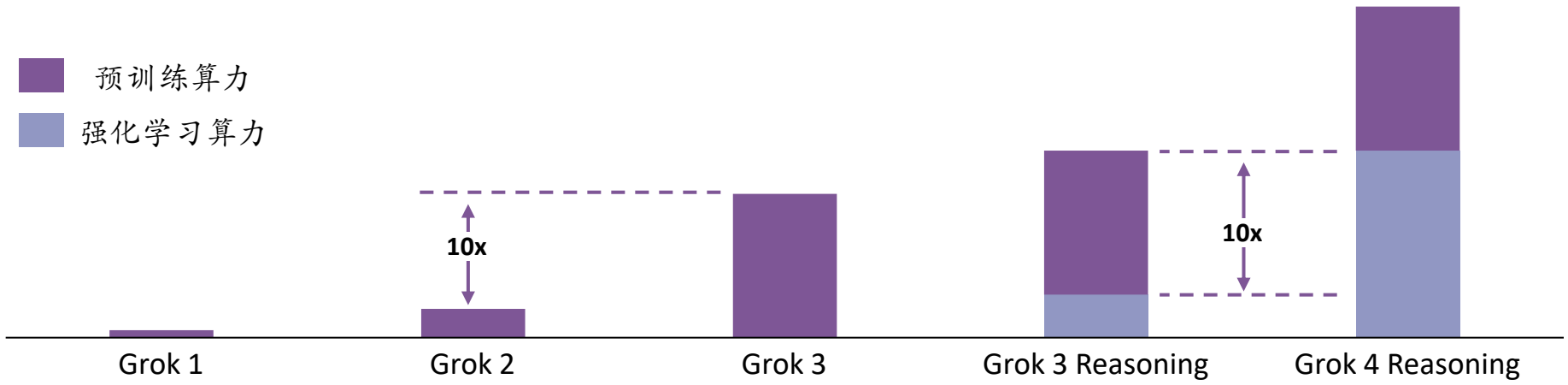
分析

- xAI成立仅2年时间，受益于充足的资本、算力、和人才支撑，在短时间内完成了对第一梯队的追赶
- 成功的核心原因主要在于公司顶尖的执行效率和来自特斯拉的工程支持，例如数据中心和电网改造
- xAI的成功证明大模型业务模式只有高壁垒，但没有护城河，模型厂商需要持续投入大量资源以应对市场竞争

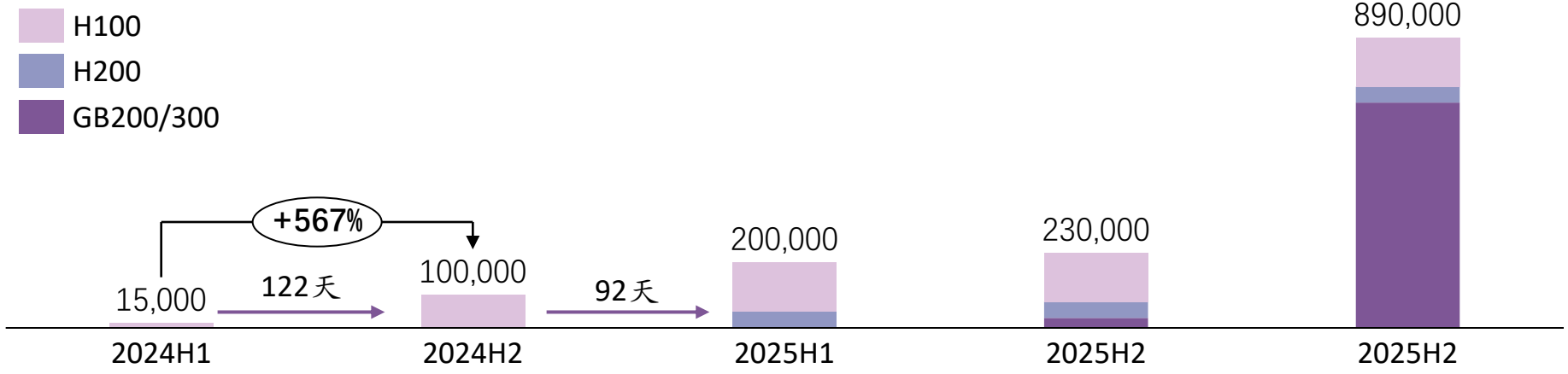
信息来源：量子位智库

算力是AI竞赛中的关键竞争要素，强化学习对算力的需求超过预训练，头部大模型玩家的计算集群已达到数十万卡规模，并在持续扩张中

Grok系列模型不同训练方式算力示意



xAI 算力集群GPU上线数量（块）



分析

- 从“Bitter Lesson[1]”的视角，AI的智能程度最终取决于算力的规模，拥有最强大的计算资源是AI竞争的核心要素
- 强化学习的重要性再次获得头部玩家落地验证
- xAI打破了大规模GPU集训的建设速度纪录，在GPU Infra领域处于全球前沿，短时间快速拉起的10万卡、20万卡集群成为帮助xAI冲击模型SOTA水平最重要的因素，未来集群的GPU数量将扩展到100万卡规模

信息来源：量子位智库，xAI，1）苦涩的教训：由强化学习之父Richard Sutton提出的观点，强调通用方法（general methods）结合大规模计算（scale computation）总是优于依赖人类知识的特定方法

OpenAI技术领先优势明显弱化，海外头部玩家水平趋同，谷歌和xAI在2025年上半年迎头赶上，模型在多个领域达到SOTA水准

模型厂商	通用场景	视频生成	多模态助手	图像生成	代码能力
OpenAI	• o3在多模态深度推理、深度搜索方面目前依然顶尖	• Sora	• GPT-4o	• GPT-4o 图像生成是目前SOTA水平	• 头部玩家模型的代码能力也在快速趋同
Google	• Gemini 2.5 Pro的综合推理能力优秀，且有视频和音频处理多模态能力	• Veo 3	• Project Astra	• Gemini有原生图像生成能力	
xAI	• Grok 4的科学、工程类问题优秀	• 暂无，但计划2025年下半年推出	• Grok 4 语音、视频功能	• Grok图像生成	
ANTHROPIC AI	• Claude 4在通用类推理、科学推理以及日常C端任务上稍弱于另外3家	• 暂无	• 暂无	• 暂无	• 代码能力和相关产品（如Claude Code）能力顶尖
Meta	• Meta的Llama 4在各个方面效果不佳 • 目前正在进行组织调整，投资数据标注赛道龙头公司Scale AI，同时加码外部顶尖人才引进				

分析

- 上半年谷歌发布Gemini 2.5 Pro，xAI发布Grok 4，已经和OpenAI的旗舰模型处于同一水平，智能上限几乎相同，在部分能力（如工程推理、多模态等方面）上超过OpenAI
- 头部模型公司差距进一步缩小，竞争激烈

代表该领域的SOTA水平

信息来源：量子位智库

中美通用大模型技术差距缩小，中国模型公司在通用大模型之外的其他领域可以达到SOTA水平，尤其多模态领域表现出色

视频生成（Image-to-video）[1]

1	Seedance 1.0	ByteDance 字节跳动
2	Hailu-02-616	MINIMAX
3	Avenger 0.5	
4	Veo 3 Preview	Google
5	Kling 2.0	快手

图像生成&编辑[2]

1	GPT-4o Image	OpenAI
2	Seedream 3.0	ByteDance 字节跳动
3	Imagen 4 preview	Google
4	Vivago 2.0	智象未来 HiDream.ai
5	Imagen 4 Ultra Exp	Google

音频生成（Text-to-Speech）[3]

1	Speech-02-HD	MINIMAX
2	TTS-1 HD	OpenAI
3	Speech-02-HD Turbo	MINIMAX
4	TTS-1 HD	OpenAI
5	Multilingual v2	IIElevenLabs

前端代码生成[4]

1	Gemini 2.5 Pro	Google
2	DeepSeek-R1-0528	deepseek
3	Claude 4 OPus(0514)	ANTHROPIC
4	Qwen3-Coder	阿里巴巴
5	Claude 3.7 Sonnet(0219)	ANTHROPIC

长文本能力（OpenAI-MRCR）[5]

1	Gemini 2.5 Pro	Google
2	MiniMax-M1	MINIMAX
3	o3	OpenAI
4	Seed-Thinking-v1.5	ByteDance 字节跳动
5	DeepSeek-R1-0528	deepseek

深度搜索[6]

1	Kimi Researcher	Moonshot AI
2	o3	OpenAI
3	O4-mini-high	OpenAI
4	Gemini DeepResearch	Google
5	Grok 3 DeeperSearch	xAI

分析

- 上半年中国模型在多模态和代码生成领域都达到了世界一流水平，尤其多模态模型
- 除模型能力的快速追赶外，中国模型在低成本和响应速率上也有很大优势，语言模型、视觉模型、音频模型的训练和推理成本都要比海外的头部模型更低

中国模型

海外模型

信息来源：量子位智库，1）、2）、3）排名来自Artificial Analysis榜单，4）排名来自Lmsys的WebDev，5）数据来自MiniMax，6）数据来自月之暗面披露的xbench榜单

AI编程领域成为模型厂商必争之地，海外和国内头部玩家在AI编程的模型和产品领域密集布局（1/2）

模型厂商	模型	产品	并购以及其他动态	分析
OpenAI	• o3：推理模型，同时代码和工程能力显著增强 • GPT-4.1系列模型，针对编程场景进行优化，性价比更高	• OpenAI Codex：并行执行代码工程任务的工具，集成于ChatGPT • Codex CLI：可本地部署的开源命令行界面Coding Agent	• 曾计划收购AI编程初创公司Windsurf但谈判破裂	• 编程有极高概率是AI最先超越人类水平的高价值应用场景 • 对于每家模型公司来说都处于路线图的核心位置，也是目前行业共识最强的应用领域，每家公司都在做针对性强化，Anthropic的Claude模型在编程方面的领先优势也在缩小
ANTHROPIC AI	• Claude 4系列模型针对代码类工程问题进行优化，真实场景中表现优异，AI编程是Anthropic的核心战略方向	• Claude Code：在终端命令行进行“代理式编程”的工具，由内部工程研究团队孵化而来	• 积极构建AI编程生态，举办首届“Code with Claude”开发者大会，与Cursor、微软等紧密合作	
Google	• Gemini 2.5 Pro：代码能力增强 • 发布AlphaEvolve：由Gemini驱动编码代理，用于设计复杂算法和代码库	• Gemini Code：对标Anthropic的Claude Code；Jules：对标OpenAI的Codex；Stitch：AI设计工具，可直接导出前端代码	• 收购AI编程明星初创公司Windsurf的创始人和核心团队	
xAI	• Grok 4代码能力增强 • 专门用于AI编程的模型正在训练中，预计8月发布	• 暂无	• 暂无	

信息来源：量子位智库

AI编程领域成为模型厂商必争之地，海外和国内头部玩家在AI编程的模型和产品领域密集布局（2/2）

模型厂商	模型	产品
阿里巴巴	• 开源Qwen 3-Coder，主打编程部分榜单取得SOTA • Qwen 3 代码能力增强	• 通义灵码：编程Agent • Qwen Code：对标Claude Code的命令行编程工具
ByteDance 字节跳动	• 开源Seed-Coder模型，通过自身生成和筛选高质量训练数据，大幅提升模型代码生成能力	• 发布AI编程IDE Trae，全面对标明星应用Cursor，目前用户月活已破百万
Tencent 腾讯	• 暂无相关垂直模型或特别优化	• 发布AI编程IDE CodeBuddy：覆盖产品-设计-研发-部署全流程
Baidu 百度	• 暂无相关垂直模型或特别优化	• 文心快码：基于文心大模型研发的编程辅助工具，可提供代码生成、单元测试、注释生成等功能
deepseek	• DeepSeek-0528：在Web开发场景的评估分数仅次于Gemini 2.5 Pro，针对代码能力优化	• 暂无
Moonshot AI	• Kimi K2模型：具备强代码和Agent 能力的MoE架构基础模型，在部分评估榜单取得开源模型SOTA	• 暂无
智谱·AI	• 旗舰模型GLM-4.5代码能力显著提升 • 开源代码模型CodeGeeX4-ALL-9B	• 暂无

分析

- 目前国内头部的模型玩家主要采取跟随、对标海外模型厂商和明星应用的策略，例如字节推出Trae对标Cursor，阿里推出Qwen Code对标Claude Code

信息来源：量子位智库

国内大模型创业公司路线开始分化，部分厂商积极发布前沿模型产品追求智能上限，其他厂商专注垂类领域和商业化落地，放缓通用模型投入

模型厂商	核心进展
deepseek	• 2025年年初发布DeepSeek R1获得全球关注，之后开源了DeepSeek-V3-0324（推理能力、代码能力增强）和DeepSeek-R1-0528（前端代码能力增强）
Moonshot AI	• 开源K2模型，总参数量达到1亿，主打Agent类场景（Agentic编程、工具调用、数学推理），发布Kimi Researcher功能，发力深度搜索、研究功能
MINIMAX	• 开源MiniMax-01模型（主打长上下文窗口和线性注意力）和MiniMax-M1模型（主打推理能力和长上下文窗口），发布海螺视频生成模型Hailuo-02、语音模型Speech-02、MiniMax Agent等
智谱·AI	• 开源新一代旗舰模型GLM-4.5，开源GLM-4.1V-Thinking模型（可以完成多模态推理的9B参数小模型）、推理模型GLM-Z1、Agent模型 AutoGLM、图像生成模型CogView4
阶跃星辰	• 开源3D大模型Step 1X-3D、音乐大模型ACE-Step、图像大模型Step1X-Edit、Step-R1-V-mini模型（主打多模态深度推理）、Step-Video-TI2V图生视频模型、Step-1o-Audio语音交互模型等
零一万物	• 不再单独训练通用基础大模型，与阿里云联合宣布启动“产业大模型实验室”，聚焦产业大模型和B端落地，专注于训练产业大模型，提供深度的私有化部署服务 • 发布万智企业大模型一站式平台2.0版本，并推出零一万物企业级Agent智能体“万仔”，商业模式主打价值共创，致力于提高公司整体的价值产出 • 业务策略上主推“一把手打法”，首先由高层制定AI驱动的顶层战略，然后是战略、技术、业务三方团队密切配合，打造真正贴合业务需求的大模型ToB解决方案
百川智能 BAICHUAN AI	• 业务收敛，聚焦医疗大模型和医疗行业落地

战略转型

分析

- DeepSeek的现象级出圈改变了国内大模型创业公司的竞争格局，同时驱动了中国模型厂商的开源战略
- 模型创业公司分化为两条路径：
- 1）继续投入通用模型研发，发力C端或P端（Prosumer）产品，追求智能上限，保持技术驱动
- 2）追求行业落地和商业化，或者转向聚焦垂直领域

信息来源：量子位智库