大模型落地与前沿趋势研究报告

2024. 量子位智库 Qbit AI Insights 分析师 Xuanhao xuanhao@qbitai.com

目　录

01

大模型市场落地概况

大模型市场落地概况：

大模型业务模式概览

大模型市场宏观和落地分布

信息来源：量子位智库

大模型业务模式概览：主要包括应用开发部署平台、模型API服务及模型服务三部分，目前模型服务和模型API是核心业务

		描述	示例
应用层	应用场景	• 最终产生价值的落地场景	• 文案写作、代码生成、信息检索、海报制作、视频制作，智能客服、智能营销
	AI产品	• 基于大模型构建的AI产品	perplexity Claude 秘塔 Midjourney runway
重点展开 模型层	应用搭建	• 包括AI应用的开发、部署、管理，帮助用户构建AI应用，核心组件包括提示词工程、RAG、插件生态、应用编排等	Dify
	模型API	• 为用户提供API调用，主要包括托管模型丰富度（自研及第三方模型）、模型能力、价格、API性能、安全合规等方等因素	推理平台：硅基流动 大模型厂商：智谱·AI、Moonshot AI、阶跃星辰、MINIMAX、零一万物 互联网/云厂商：阿里云、百度智能云、腾讯云、火山引擎、华为云
	模型服务	• 数据准备：训练数据的采集、选择、标注、清洗等 • 模型训练：包括预训练（pre-train）和微调（fine-tune） • 模型推理部署：帮结合业务场景需求帮助客户在云端和本地进行模型推理部署 • 模型社区：搭建开源模型服务平台，提供多样化模型库、模型API服务、训练数据集等	
Infra+云	GPU及配套基础设施	• 算力服务 • GPU集群搭建 • 互联设施 • 数据中心建设	

信息来源：量子位智库

模型服务：数据准备和模型训练是模型服务的关键内容，对交付服务的深度和细致程度有较高要求，是国内市场目前最重要的商业模式

构成	描述	应用现状	重要性
数据准备	• 环节：包括数据采集，数据评估、数据选择、数据标注、数据清洗、数据回流等 • 类型：行业数据，即整个的垂直行业的相关数据，例如医疗、金融、制造、政务等行业；场景数据，即垂直行业内某一场景的数据，例如客服、培训、产品开发、疾病问诊；企业数据，即和企业业务和自身属性相关的数据，例如产品信息、运营信息等	• 数据准备是目前模型服务最核心的问题，需要客户厘清数据的种类需求、格式需求等，在此过程中模型服务商需要和客户紧密合作，帮助客户梳理、准备数据 • 模型微调的数据需求在数百GB级别，以及至少上万条的对话数据、交互数据，数据的质量直接决定模型的表现	
模型训练	• 微调（Fine-tuning）：在已经预训练好的大模型基础上基于特定数据集进一步调优，对算力（百卡级）和数据（数百GB）的需求小，例如LORA，Adapter layer、Prefix Tuning等技术 • 预训练（Pre-training）：从头开始进行预训练，要求有大量的垂直相关数据资源和算力，包括文字、图像、视频、交互记录及其他特殊格式数据	• 微调是目前最主流的服务方式，效果好成本低，相应技术较成熟 • 预训练模型较少见，成本高挑战大，主要用来解决特殊问题，例如进行蛋白质结构预测的AlphaFold	
模型推理部署	• 云端部署：模型在云厂部署，使用模型时调用模型API，由云厂商负责所有运维 • 本地部署：自主可控，响应快、服务稳定保响应时长和调用频率、隐私、安全性强 • 混合部署：兼顾两种部署模式，具体依照业务场景的需求决定	• 根据客户偏好和需求决定模型部署方式，云端部署是主流方式	
模型社区	• 汇聚各类模型信息、数据集、模型竞赛、技术内容分享的社区平台	• 主要目的是构建围绕大模型的开发者生态，促进生态繁荣，例如阿里的魔塔社区ModelScope	

分析

- 需求方画像：以G端客户、中大型B端客户为主，国央企、金融等行业是主要客户
- 模型服务是国内大模型市场的核心部分，贡献了目前行业的大部分营收
- 商业模式较重，需要模型厂商进行交付服务，全流程服务客户，本质上是一个[人*天]投入业务模式，但相较于软件、云业务的定制化交付服务要更加轻量化（不同客户的服务内容基本相同）

信息来源：量子位智库

模型API：国内各厂商模型能力没有明显差异化，API市场的低价竞争阶段将长期持续，整体用量正在快速增加但难以贡献营收

构成	描述	关键指标	重要性
模型库	• 包括模型厂商的自研模型和第三方开源模型（主要适用于云厂商，会提供MaaS[1]服务）	• 可选模型的种类和数量，包括语音、图像等其他模态，以及模型供应商数量	◔
模型能力	• 推理能力：衡量大模型智能的核心指标，也包括指令遵从、内容合规、用户意图理解等 • 上下文长度（Context window）：模型支持的上下文窗口大小，决定模型可以处理的任务复杂度上限 • 多模态能力：包括文字、图像、视频、音频等模态的理解、生成、交互表现	• 调用量：由真实的市场需求用脚投票产生 • 静态评测：各类评测榜单，如MMLU、MATH、GPQA、HumanEval、GSM8K等 • 动态评测：用户双盲实测榜单，如LmSys的 Chatbot Arena Leaderboard、LiveBench等	●
价格	• API价格总体上呈快速下降趋势，和具体和调用量、调用方式有关	• API价格，包括输出价格（Input Tokens）和输出（Output Tokens）价格，以及更便宜的Batch API	◕
API性能	• API服务的各项性能，包括延迟、吞吐量、输出速度等	• 延迟：FTL（First Token latency，从发出请求到收到第一个Token的时间，也称Time to first Token ） • 吞吐量：RPM（Request per minute，每分钟请求次数）、TPM（Tokens per minute，一分钟输出Token量） • 输出速度：Tokens per second（每秒输出Token量） • 稳定性：各项指标稳定性、波动水平、故障频率	◕
安全合规	• 用户的数据隐私保护、内容安全合规性	• 关于隐私保护、安全合规的关键举措和安全系统构建成熟度 • 安全、合规事故的频率	◑

关键分析

- 需求方画像：行业属性非常碎片化，包括来自各个行业的个体开发者、中小企业、初创公司等，对客户自身技术能力有一定要求
- 难以贡献营收：
  - API在海外是大模型的核心商业模式
  - 国内市场由于模型能力缺乏差异化、能力不成熟、行业低价竞争趋势等因素作用，市场规模较小，目前难以成为模型厂商的主要收入来源

信息来源：量子位智库，1）Model as a Service

应用搭建平台：旨在释放大模型的应用潜力，但目前的产品形态尚未获得市场验证，用户规模依然较小，未来1-2年将更加成熟覆盖更多用户

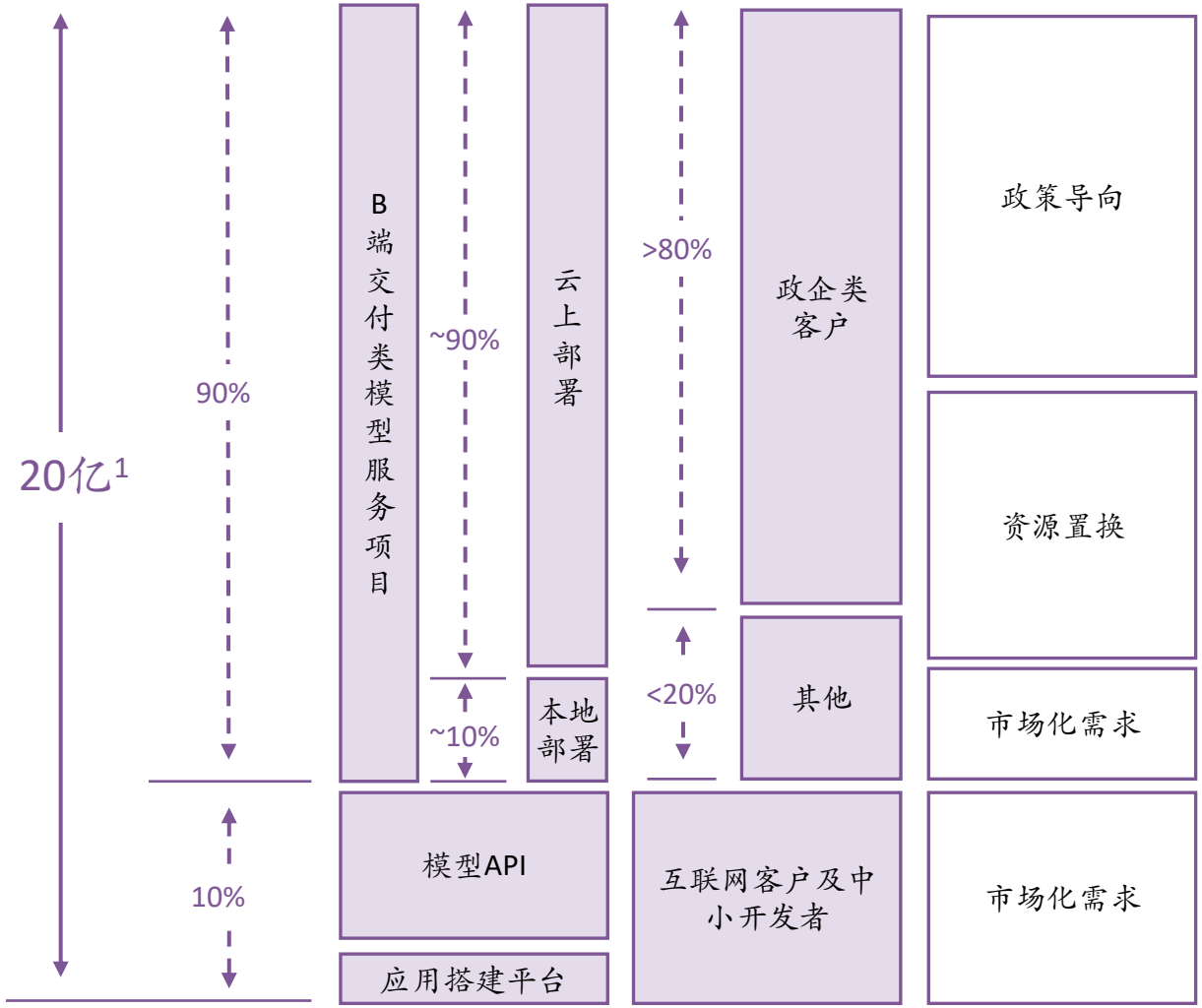
构成	描述	示例/特点	重要性
提示词工程	• 通过优化提示词来引导模型生成所需输出，可以调整提示词的措辞、结构和内容来提高模型响应的准确性和相关性 • 随着模型能力的不断增强（上下文窗口的增加、推理能力增强、幻觉程度下降），提示词工程可以解决的问题集也在不断扩大，有大量应用场景可以仅通过提示词工程来解决	• Meta prompting、系统提示词，Prompt模板库 • Chain-of-Thought、Tree-of-Thought等提示词研究	●
RAG[1]	• 原理是从大型知识库或文档集合（向量化后）中检索相关信息，然后使用大模型（如GPT-4o）对检索到的信息进行加工和扩展，以生成更丰富、更准确、幻觉更少的回答，兼顾了检索系统的准确性和生成模型的灵活性，适用于处理复杂问答和知识密集型任务	• 信源可解释，可靠性高 • 可以实现对敏感数据的隐私权限控制 • 数据知识库可即时更新	●
插件搭建	• 通过多样的插件调度，系统可以在需要时动态调用不同的功能模块，以满足复杂的应用需求 • 包括插件的注册、依赖关系的管理、执行顺序的安排等，确保各插件能够协同工作，提高系统的整体性能和灵活性	• 代码解释器、可视化类插件、图表插件、搜索引擎、及执行类插件（订票、更改日程等）	◑
应用编排	• 可以自动化地管理和协调多个应用组件和服务，以实现特定业务流程或工作流的目标 • 通过应用编排，开发者可以定义复杂应用的逻辑和操作步骤，确保各组件按预期协同工作，帮助提升系统的可管理性和扩展性	• Assistant API、LangChain、LlamaIndex	◔

分析

- 需求方画像：应用开发部署平台的用户主要是应用开发者群体、ISV技术人员，包括专业开发者及技术能力较弱的业务人员
- 底层大模型扮演技术基座的角色，但从模型到产品仍有较大的差距，需要在产品层针对业务、场景需求进行开发设计
- 产品尚未验证PMF：应用开发部署平台本质上类似低代码开发平台，真实的市场需求尚未验证，技术人员偏好使用代码开发应用，而非技术人员入门上手依然有一定挑战，目前用户依然高度集中在产品、开发类人员

信息来源：量子位智库，1）RAG（Retrieval-Augmented Generation），检索生成技术

国内大模型市场结构：以B、G端客户为绝对主力的市场结构在短期不会改变，但长期随技术进步和行业发展，市场化需求的份额将逐渐增加



描述	特点
• 国家、地方政府出台推动大模型发展落地的相关政策，给予相关经济支持，国字头、中字头机构有政策上的指标和预算 - 例如《北京市推动“人工智能+”行动计划（2024-2025年）》将对重大标杆应用工程和示范性应用，按照项目工程投资予以一定比例支持；面向商业化应用，北京对多个细分场景下的大模型应用落地项目发放模型券	• 客户自身技术能力较弱，需要外部模型厂商进行全流程的交付服务 • 客单价较高，有付费能力，且一般不考核模型落地的实际ROI
• Infra资源置换：在促进人工智能产业发展的整体背景下，模型厂商和大模型产业链的相关参与方进行合作，相互采购服务。例如和算力建设、运营方进行合作，相互采买旗下模型和算力服务 • 股权置换：和潜在大型客户进行股权置换来获得订单	• 需要在B端、G端的资源网络促成互惠合作
• 客户主要来自金融、医疗等行业	• 对模型能力要求较高，ROI驱动
• 需求方画像较为碎片，包括中腰部的各类型互联网客户，也包括一些中小型的下游AI应用开发者、独立开发者以及普通的C端用户	• API用户一般具有基本的技术开发能力，可以自行做模型选择和调试

信息来源：量子位智库，1）2024年

大模型部署方式：云上部署是目前大模型最普遍的部署方式，成本和运维复杂性相对较低，尤其受中小型客户欢迎，上云趋势未来仍将持续

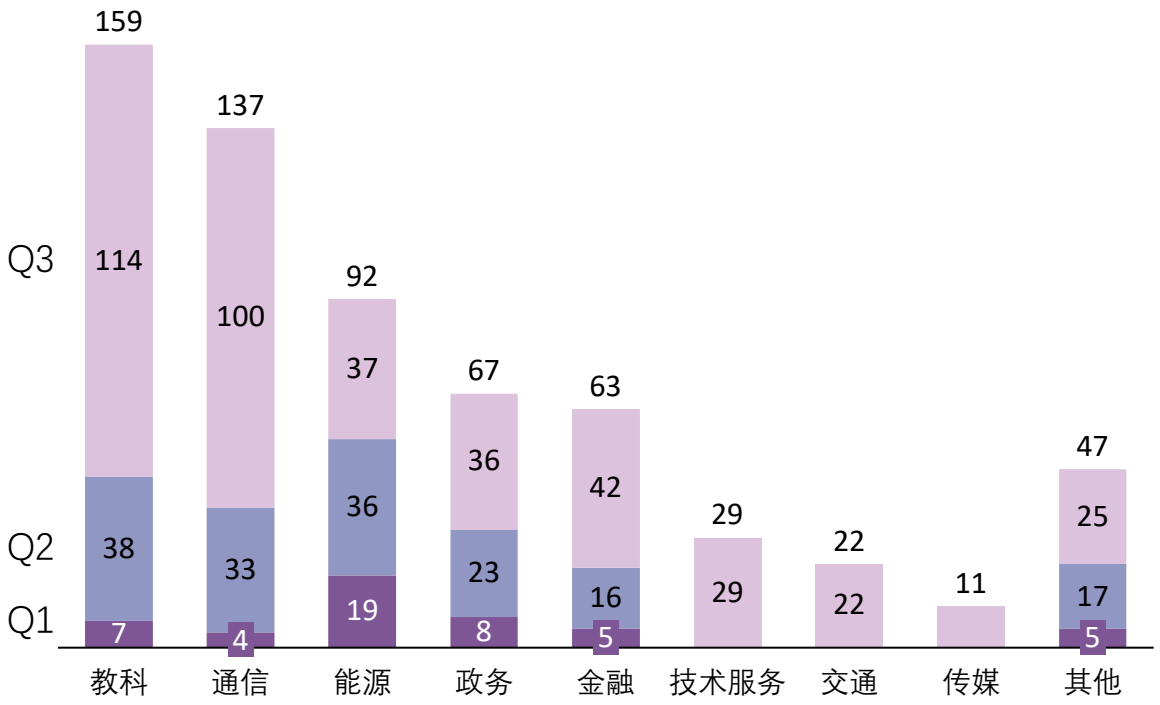
项目制交付

		描述	适用需求	核心价值	应用成本
云上部署	应用开发平台	• 普通用户注册即用，可快速搭建智能体、AI应用	• 面向中小开发者，方便探索不同类型的应用	• Token和存储空间	
	模型通用API	• 普通用户注册即用，快速接入通用大模型能力，按照用量收费，无固定成本	• 面向中小开发者。可以帮客户快速完成场景验证，跑通业务流，同时客户对模型能力没有垂直化需求	• API Token用量	
	模型服务 微调+独家算力	• 为客户提供模型微调训练服务和相关部署方案，同时保留独家算力，支持高并发、快响应等需求，不受通用API用量波动造成的影响	• 面相通用模型不能满足的场景需求，场景对垂直领域知识、模型的输出模式有特别需求，例如金融、医疗及等领域，同时客户需要对大模型Token用量、并发请求规模有规划和估计	• 交付服务以及Token用量	
	模型服务 微调+云端私有化	• 满足客户对于数据隐私的需求，在独占算力的基础上增加数据安全相关的额外协议措施，增加云厂商背书，确保客户数据安全	• 面向没有能力和意愿自建算力，但有高并发应用场景或对数据的隐私、安全性有较高要求的行业客户	• 交付服务、安全协议，以及Token用量	
私有化部署	模型服务 微调+私有化部署	• 部署在客户自己的私有云上，拥有最高级别的安全性、自主可控性	• 客户可完全控制数据的存储和处理过程，确保敏感数据不会离开企业服务器，同时客户可以自主配置软硬件资源，适配业务需求提高效率，主要面向例如政企、金融、制造业的大客户	• 深度交付服务	
	私有化应用搭建平台				

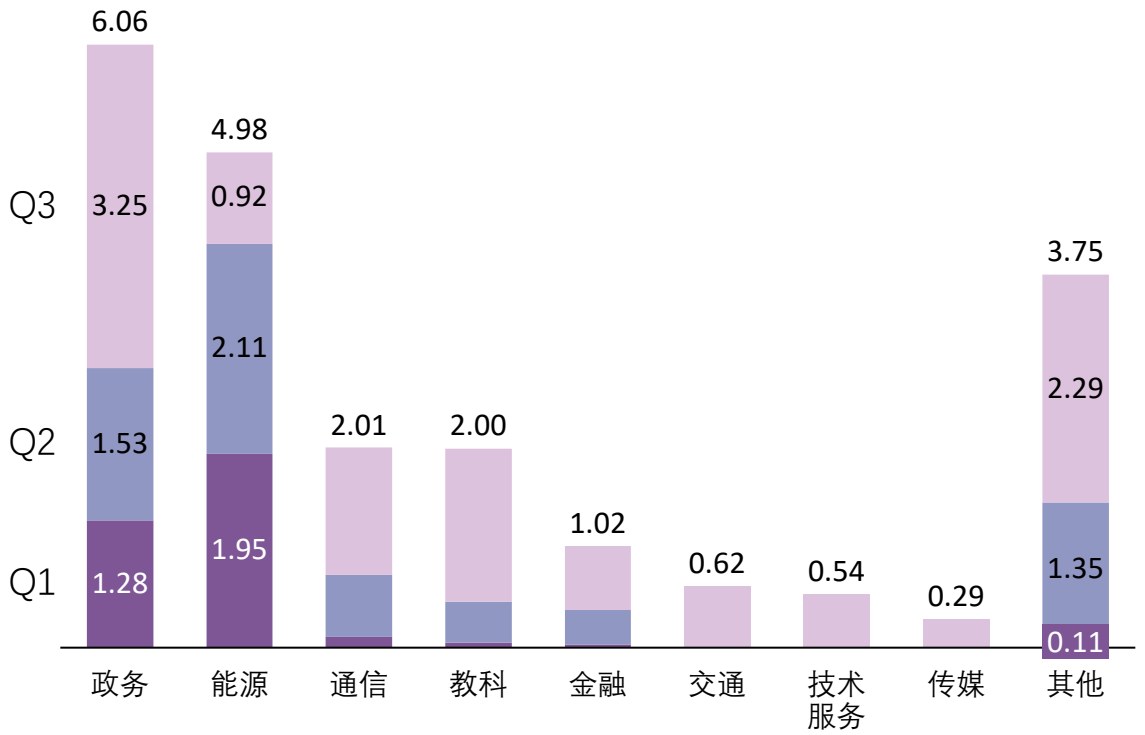
信息来源：量子位智库

大模型市场行业分布情况[1]：大模型项目在教科、通信、能源、政府、金融等行业落地较多，全行业在今年第二、三季度增长明显

2024年各行业大模型公开披露的落地项目数量（单位：个）



2024年各行业大模型公开披露的落地项目金额（单位：亿元）



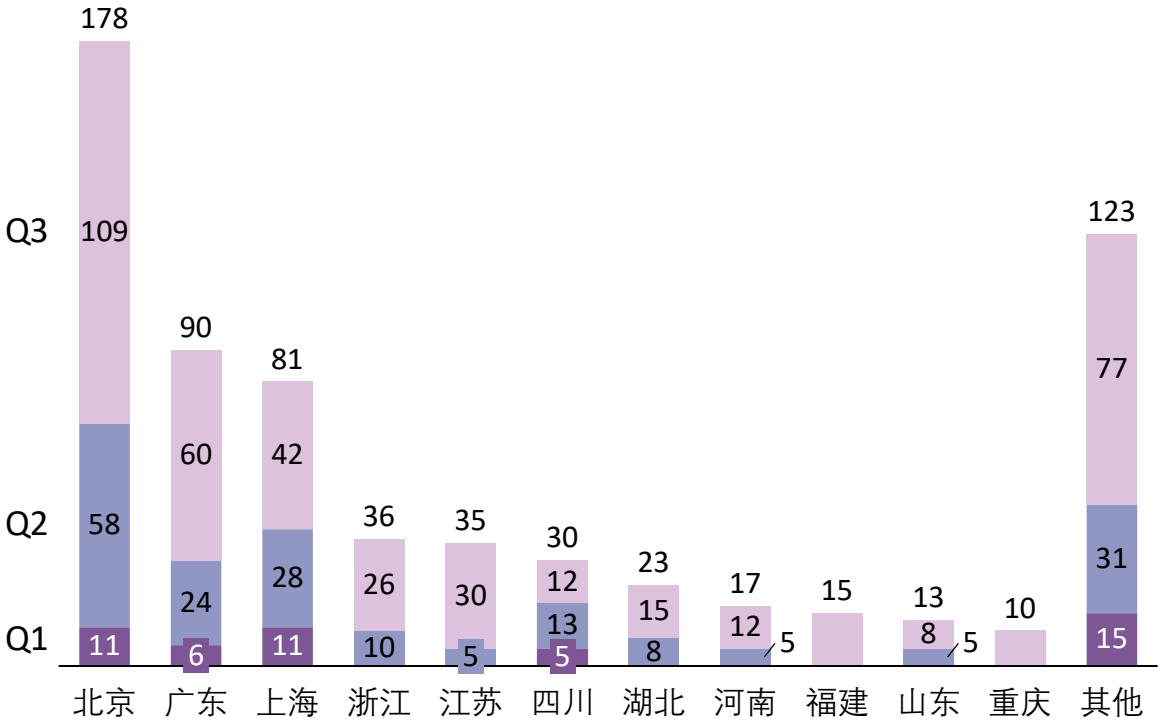
- 从数量上看，教科类、通信（运营商）、能源、政务、金融是目前公开披露落地项目最多的行业，在今年第二、第三季度落地项目数量出现了明显增长

- 从金额上来看，政务、能源、通信（运营商）、教科类，金融是项目落地总成交金额最大的行业，其中政务和能源的单项目金额较大，其他行业在第三季度出现快速增长

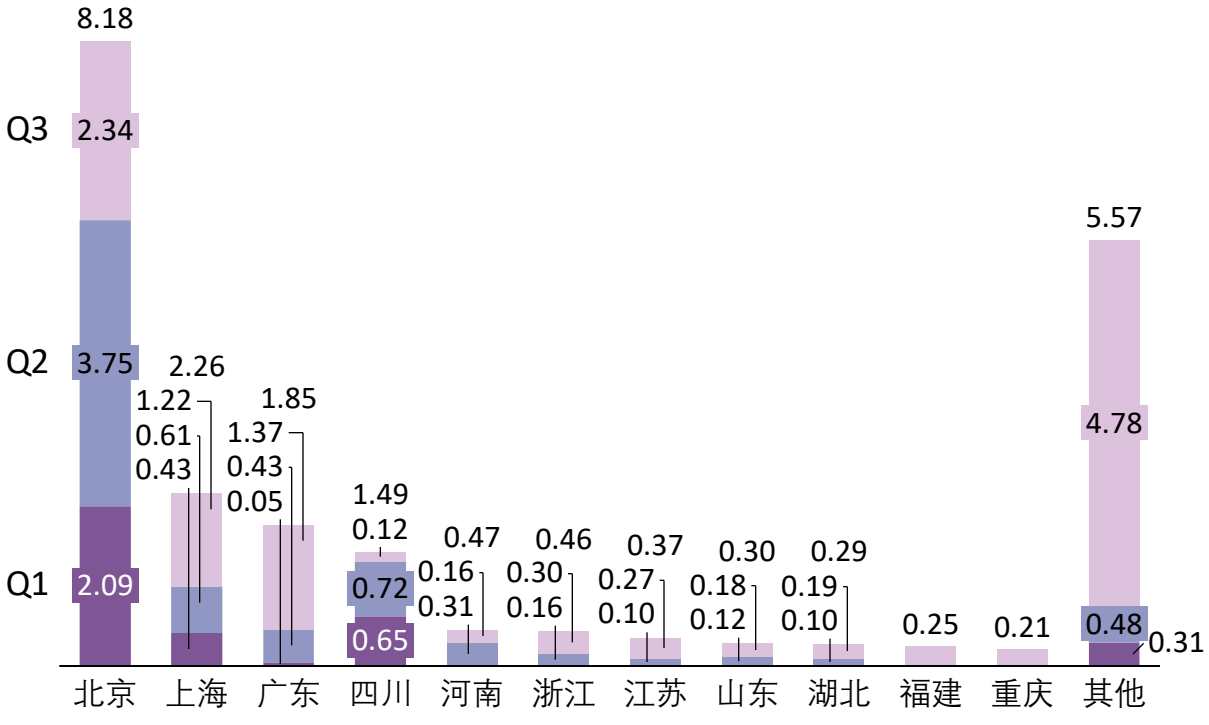
信息来源：智能超参数，仅统计公开信息供参考，1）考虑到一般落地项目多为整体解决方案，其中可能包括大模型及与之匹配的应用、数据、算力等

大模型市场地域分布情况[1]：大模型项目在一线城市和沿海省份落地较多，大规模项目主要集中在北京，其他地区今年第二、三季度增长明显

2024年各地域大模型公开披露的落地项目数量（单位：个）



2024年各地域大模型公开披露的落地项目金额（单位：亿元）



- 从数量上看，北京、广东（深圳）、上海、浙江、江苏等地大模型的落地数量最多，高度集中在经济发达一线城市
- 第二、第三季度落地数量大幅增加，有明显增长趋势

- 从金额上来看，落地项目高度集中在北京，头部大客户数量较多，尤其政策驱动类的政务客户在北京较为密集。北京落地节奏最快，其他地区主要在第二、三季度增长明显

信息来源：智能超参数，仅统计公开信息供参考，1）考虑到一般落地项目多为整体解决方案，其中可能包括大模型及与之匹配的应用、数据、算力等

02

大模型落地与前沿发展趋势

大模型落地与前沿发展趋势：

大模型落地逻辑	模型能力拆解	大模型技术趋势	前沿应用趋势	开源模型

信息来源：量子位智库

大模型落地逻辑：模型能力是关键要素，现有模型已经可以实现多行业的广泛覆盖，但进一步深度整合需要模型能力大幅提升

落地需求	落地需求展开			技术供给分析		
	描述	举例	覆盖度	限制因素	可用模型	
松耦合阶段（Co-pilot）	• AI作为生产力工作流的单点工具协助用户工作	• 以ChatGPT为代表的通用类智能助手 • B、G端针对垂直场景需求的智能助手或服务类Agent，例如Github Copilot	• 广度：已经在绝大多数领域覆盖落地，目前绝大多数行业、场景均有覆盖，已经历近2年的市场初期教育、认知阶段 • 深度：非常有限，以辅助类角色为主，需要由用户来主导	• 成熟度较高，已经获得大量用户和市场验证	• GPT-4o • Claude 3.5 • 豆包 • Yi-lighting	
半自动阶段(Agent)	• AI成为生产力工作流的一部分，基于Agent能力向端到端完成工作的方向发展	• Anthropic的通用类Agent应用Computer Use • 目前可以解决垂直场景下（如代码）的部分问题	• 广度：同上 • 深度：主要是把已经覆盖的场景做深做透，整合集成更多任务上下游的规划、推理、执行环节（例如代码开发的部分环节）	• **模型可靠性**：幻觉问题依然难以解决，在高价值、高精度、低容错的场景下无法打开市场 • **推理能力**：推理能力目前难以适应复杂的生产级场景，尤其涉及多步的长程推理 • **多模态**：目前模型在多模态方面的生成和理解也不够成熟可靠 • 无法持续学习：有知识截断问题无法实时更新	需要更先进的模型解锁	
自动化阶段(Co-worker)	• AI可以**端到端完成工作，有更多主动性**	• 无	• 广度和深度同时增加，在大多数场景可以覆盖完整工作流 • 从用户主导的Copilot、Agent进阶到由Co-worker			

市场需求逐渐展开

信息来源：量子位智库

大模型落地逻辑：模型能力决定落地应用形态、商业模式以及定价权，2025年将会实现下一个阶段性跨越，解锁更多应用空间

商业价值快速增加

	描述/应用形态	核心技术	商业模式	举例	商业价值	时间
相对成熟 L1	• **对话型聊天机器人**，帮助用户进行基础的知识查询，简易内容、代码撰写	• GPT范式 • RLHF	• 软件订阅制 • 模型API • 模型微调等服务	• GPT-4o • Claude 3.5 • 豆包 • Kimi	• 从C端智能助手的订阅价格来看，目前L1和弱L2的模型市场市场价格为20美元每月 • GPT-4o API的价格约2.5美元元/百万Token（输入），10美元/百万Token（输出）	2023
构筑基础 正在发展 L2	• **推理型AI**，有规划反思能力，可解决复杂的逻辑推理、代码、技术问题	• CoT深度推理 • 强化学习	• 软件订阅制 • 模型API	• OpenAI o1	• 模型的价格和用量同步提升驱动市场规模增长 • OpenAI o1 API[1]的价格约15美元/百万Token（输入），60美元/百万Token（输出），远高于普通模型	2024
构筑基础 正在发展 L3	• **AI Agent**，能够帮助用户采取行动、自动化操纵软件、编排工作流	• 多Agent推理	• 更高的订阅价格 • 模型API • 基于结果定价	• 无	• 和现有的工作场景深度结合，放大存量价值和场景，价格和用量进一步提升	2025
早期探索 L4	• **创新型AI**，有高阶创新能力的AI，可以发明新的科学理论和工程技术	• 未知	• 软件订阅制 • 模型API • 产值抽成	• 无	• 从上游提升生产力，打开新场景、新空间，商业价值没有上限	?
展望中 L5	• **AI组织**，可以完成人类机构、组织的工作	• 未知	• 软件订阅制 • 模型API • 未知新形势	• 无	• 显著改变现有的经济结构，较为遥远	?

信息来源：量子位智库，1）包含用于推理的中间步骤Token消耗

大模型落地逻辑：模型能力提升需要大量资源支撑技术突破，进而自底向上打开市场，国内玩家需要积极追赶海外领先实践

向上支撑

	海外 2023	海外 2024	海外 2025E	国内 2023	国内 2024	国内 2025E
市场规模[1](美元)	5亿	20亿	50亿	2亿	3亿	10亿
应用规模[2]	• 5000万周活用户 • 200万开发者	• 2.5亿周活用户 • 1100万C端付费用户 • 300万开发者	• 5亿周活用户 • 2000万C端付费用户 400万开发者	• 无	• 3000万月活+ • API收入千万级RMB	• 近亿月活 • API收入千万级RMB
增量技术(模型能力)	• **万亿参数Moe** • 多模态理解 • 自监督学习	• 数万亿参数MoE • 多模态生成 • **推理能力** • **强化学习**	• **Agent** • **新产品形态** • 下一代旗舰模型	• 千亿参数 • 文本模态	• 万亿参数MoE • 多模态理解 • 多模态生成	• 推理能力 • 强化学习 • Agent
资源投入[3]	• 头部模型研发投入接近10亿美元 • 万卡A100及H100集群	• 头部模型研发投入为30-50亿美元 • 数万卡H100集群	• 头部模型研发投入达到100亿美元 • 十万卡英伟达H/B系列集群	• 头部模型研发投入达数千万美元 • 万卡A100集群	• 头部模型研发投入达数千万美元 • 万卡英伟达GPU集群及国产集群适配	• 头部模型研发投入上亿美元 • 数万卡英伟达及国产算力适配

信息来源：量子位智库，1）仅讨论API及模型服务的市场规模，不包括应用层收入，2）以头部应用为例，3）以头部玩家资源投入为例

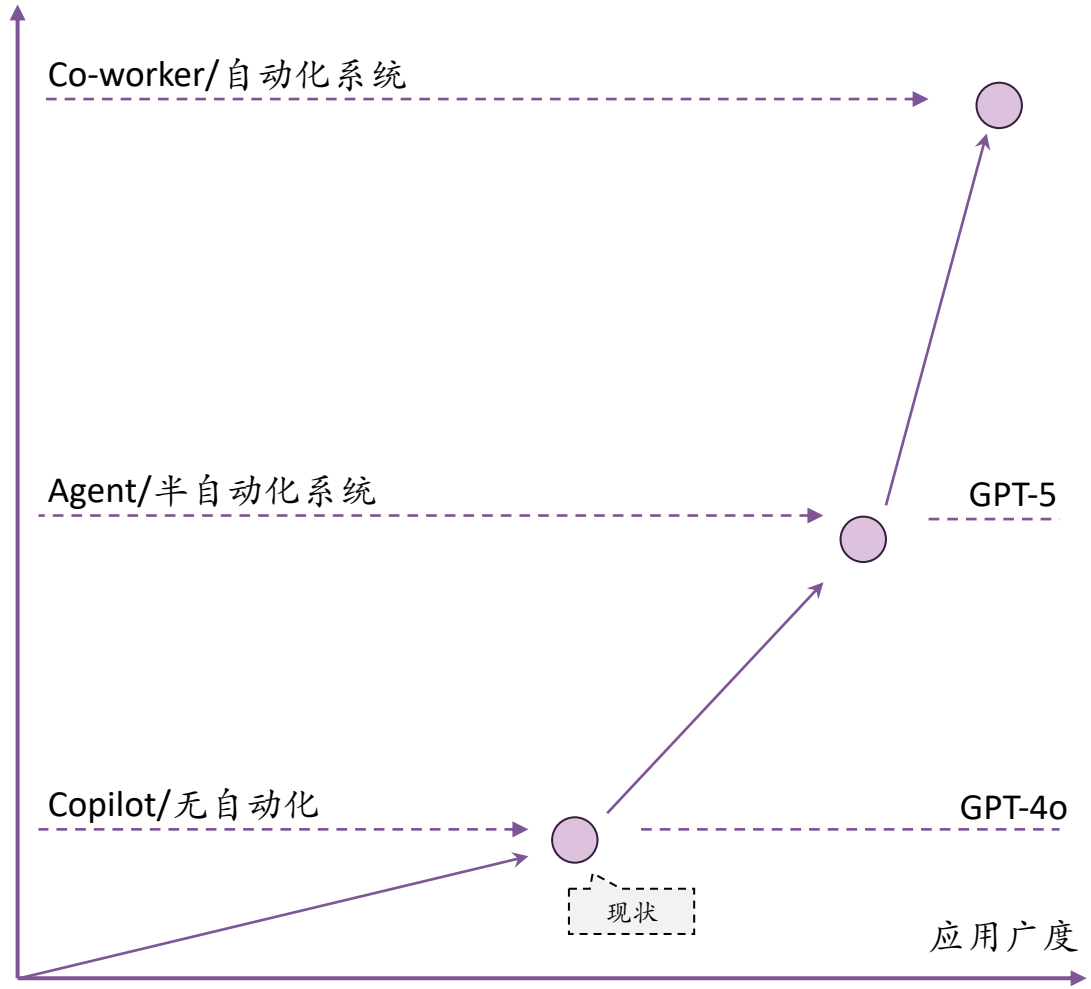
大模型落地逻辑：多个技术方向驱动模型能力提升，推动应用深度和应用广度增加，长期将向自动化系统演进

技术驱动

具体方向		成熟度
预训练	• 自监督学习 • 2018年GPT-1出现，已经过多次迭代，行业可以大规模复现	
后训练	• 强化学习为主，也有其他监督学习方式 • 技术路径尚未收敛，目前处于“GPT-1”阶段	
推理	• 深度推理正在由特定领域扩展到通用领域 • 技术路径尚未收敛	
模型优化/降本	• 目前主要是剪枝、量化、蒸馏等方式，已经有成熟方式，可以显著降低推理成本	

模型能力

大模型应用发展路径

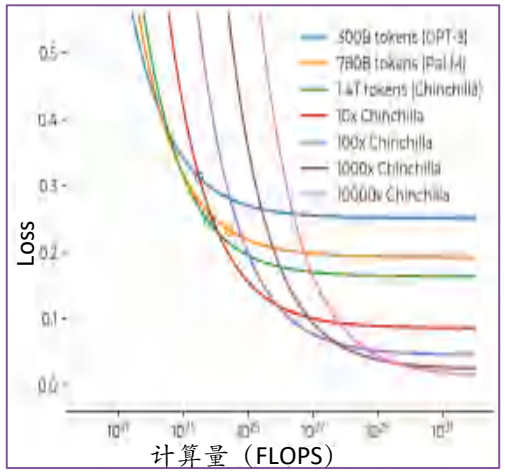


信息来源：量子位智库

模型能力-预训练：由于硬件资源和数据的限制，资源投入的边际提升已经放缓，但仍是驱动模型能力提升的核心因素之一，领军玩家积极投入

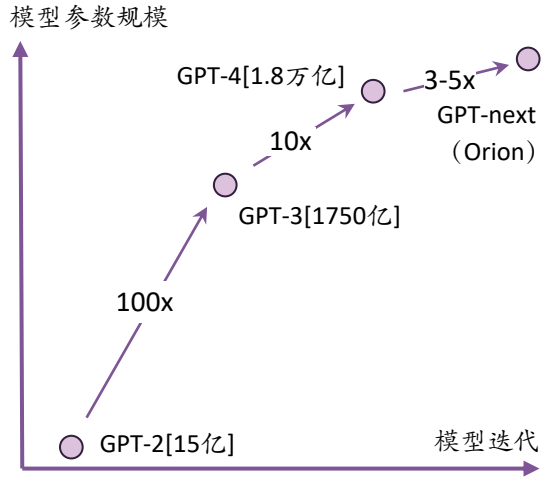
模型能力 → 预训练 → 后训练 → 推理

1 预训练阶段的Scaling Law依然是核心主线



- 预训练阶段的Scaling Law主要指模型能力随着模型参数规模、训练数据、计算量的增加不断提升，是目前GPT 范式中成本最高的训练阶段（99%[1]的计算在预训练）
- Scaling law依然是大模型发展的核心驱动因素，是模型厂商需要持续探索研发的重要领域，下一阶段的预训练投入巨大

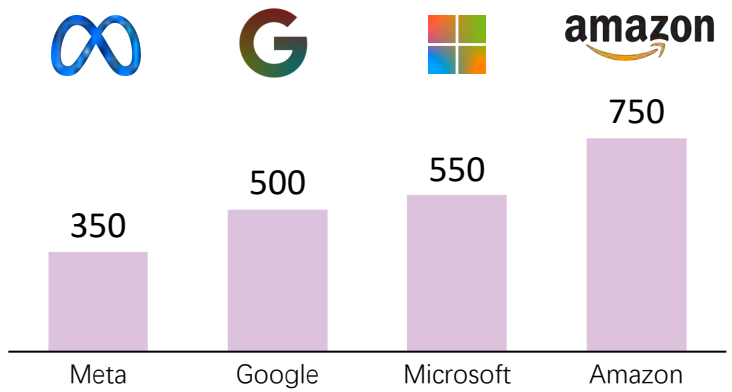
2 旗舰模型规模增速放缓



- 大模型参数规模提升速度已经放缓，GPT-4采用了MoE架构，并不是一个稠密模型（Dense Model），下一代模型在参数规模上可能增加3-5倍
- 目前模型规模增加的主要限制因素在硬件层面，模型参数过大对于GPU训练集群的内存要求和通信要求极高，物理基础设施能力的提升比软件更慢

3 2024年领军玩家扩建更大规模的数据中心，为未来大模型继续增加规模积极准备

北美云厂商2024年资本支出投入预期（亿/美元）



“我们在扩大规模的同时没有看到边际收益的递减，但每隔几年才能对Scaling law进行一次采样，因为建造超级计算机和在其上训练模型都需要一段时间。”

---Kevin Scott（CTO，微软）

“我们目前没有发现基于Transformer的模型在规模扩展上遇到瓶颈，我们将会持续扩大规模（scale up），建造更多的计算集群、生成更多的合成数据来训练模型，我认为规模扩展将仍然持续一段时间，基础设施方面投入将达到上千亿美元。”

---Mark Zuckerberg（CEO，Meta）

信息来源：量子位智库，OpenAI，1）仅针对GPT-4这一代的预训练模型，不考虑后训练方面的新范式

模型能力-后训练：强化学习成为后训练阶段的关键技术，相比预训练技术路径尚未收敛，需要高质量数据带来边际提升

模型能力
预训练
后训练
推理

A GPT路径

Base Model
- 大规模预训练后得到的模型
- 万卡需求，10万亿Token
- 自监督学习

SFT model
- 数万条问答对（QA）数据
- 人工标注，量少质量高

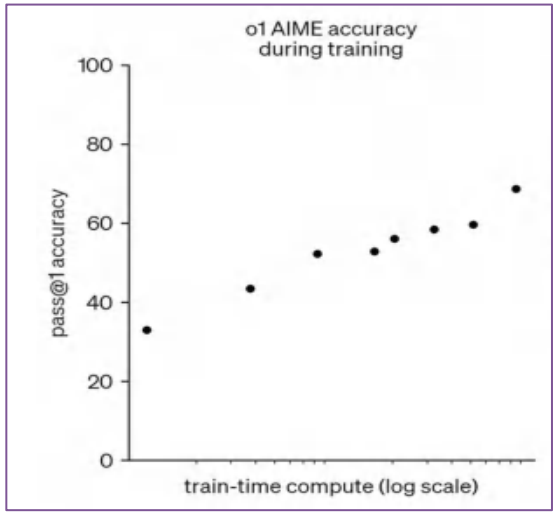
Reward Model
- 数十万条人工偏好选择数据，用人工打分选择模型产生的回答来训练奖励模型
- 人工标注，量少质量高

RL model
- 生成最大化奖励模型Token，可以直接部署
- 例如ChatGPT、Claude

后训练

- 后训练（Post-train）概念较宽泛，没有清晰的划分，一般指在初始预训练（通常是自监督学习）完成后，对模型进行的一系列优化和微调过程，以便更好地适应特定任务或提升模型的实际应用效果，主要包括SFT、RLHF环节

B o1路径



- 以强化学习为核心的后训练Scaling law已经出现，大规模强化学习算法可以教会模型如何在数据高效的训练过程有效思考
- 随着强化学习的训练算力增加，o1 类推理模型的性能会持续提高

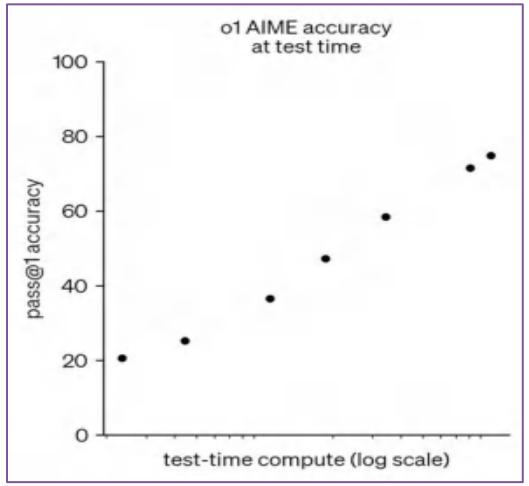
- 技术未收敛：强化学习的技术难度比较高，GPT系列的技术扩散相对比较充分，但强化学习相关技术并未收敛，没有清晰的范式，行业经验集中在OpenAI、Anthropic、Google等头部玩家，追赶难度较大
- 数据要求高：关注高质量的增量边际数据，对数据质量的要求比GPT系列更高，模型训练可能需要大量人工标注的详尽推理数据（例如领域专家、高级科研人员等），标注难度大、耗时长、成本高，此外可能需要大量的合成数据来进行训练
- 算力需求高：o1的算力需求超过万卡H100，未来强化学习部分的算力占比和需求量可能更高

信息来源：量子位智库，OpenAI

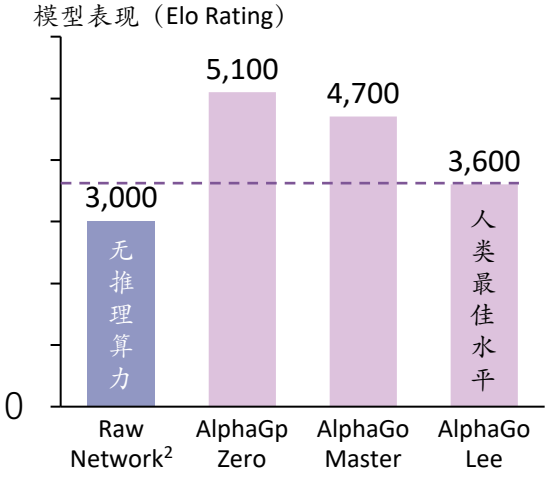
模型能力-推理：o1标志着深度推理范式走向台前，推理侧的资源需求正在快速增加，可以高效提升模型能力，未来空间广阔

模型能力 = 预训练 × 后训练 × 推理

1 推理的Scaling Law已经出现

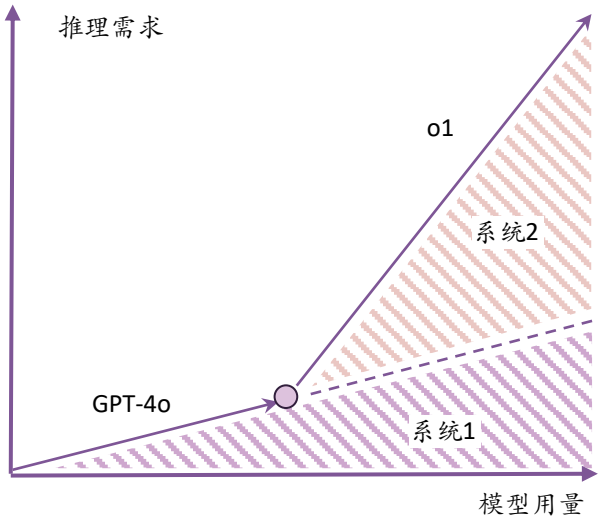


- 随着推理算力（Test-time compute）的增加，模型可以思考更长的时间，生成更多的CoT（Chain-of -Thoughts）数据，从而进行复杂问题的拆解和推理，得到更好的推理效果，思考时间越长，效果越好
- 目前o1在数学、代码类问题上的表现尤其突出

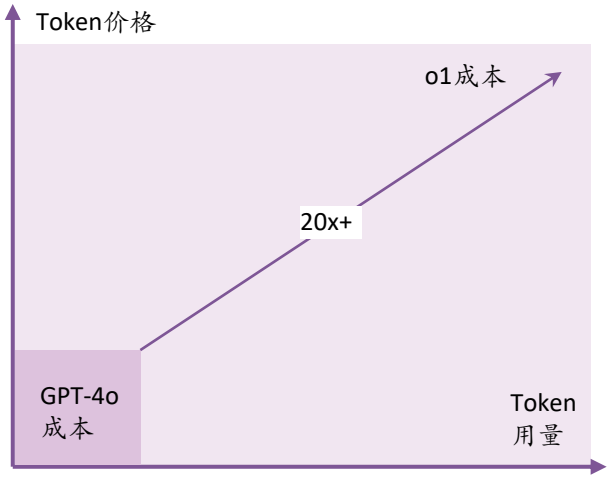


- 推理Scaling law开始泛化：增加推理时间获得性能提升已经在AlphaGo、Alpha Geometry等专注于解决特定问题的模型上获得验证（模型推理的时间越长表现越好），现在这一特性正在泛化到更通用的推理场景

2 在类o1的模型范式下，推理将变得更加重要，成为算力重心



- 随着大模型应用范围和深度增加，GPT系列模型的推理市场需求在迅速增长
- 类似o1的需推理模型将进一步刺激已经在快速增长的推理市场，系统2思考[1]、合成数据需求将带来更多推理求



- 价格贵用量大：推理模型在Token消耗量（包括用于隐藏思维链的Token）和单位Token价格上都远高于基础的GPT模型
- 综合的使用成本会相比GPT-4o会高一个数量级，主要用于高价值、同时对低延迟敏感度低的场景

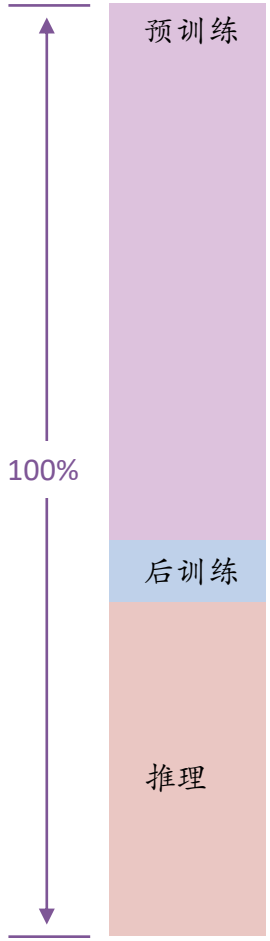
信息来源：量子位智库，OpenAI，1）指认知科学中的系统2，慢思考，2）指没有深度推理、搜索的原始模型

模型技术趋势：o1模型代表的新范式对技术创新能力、工程能力和算力资源有更高要求，并驱动算力分布向后训练和推理加码

A GPT-4路径

	描述
预训练	• 数据需求：10T以上Token，数据类型比较多样化 • 参数规模：1.8万亿参数（MoE架构） • 算力需求：2.5万张A100训练3个月，占据模型训练阶段的绝大多数算力
后训练	• SFT：由人来标注用于问答数据，数量少，质量高 • RLHF：由人类提供偏好反馈数据来训练强化学习的奖励模型，再进一步优化模型回答 • 算力需求低：数百张A100训练数十天
推理	• 类似系统1的思考模式，简单推理，单次推理的算力需求和Token的消耗量都很低，推理算力基本和调用需求量成正比

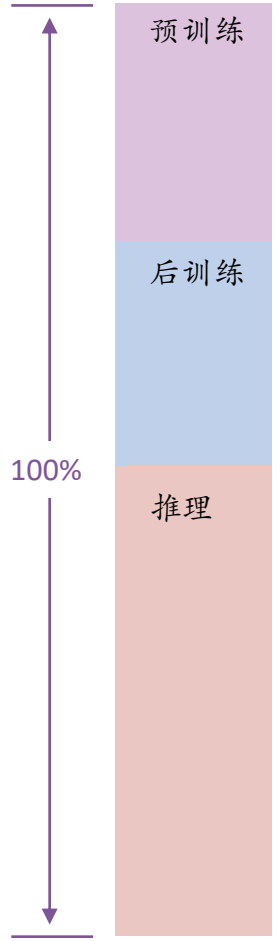
算力分布[1]



B o1路径及未来

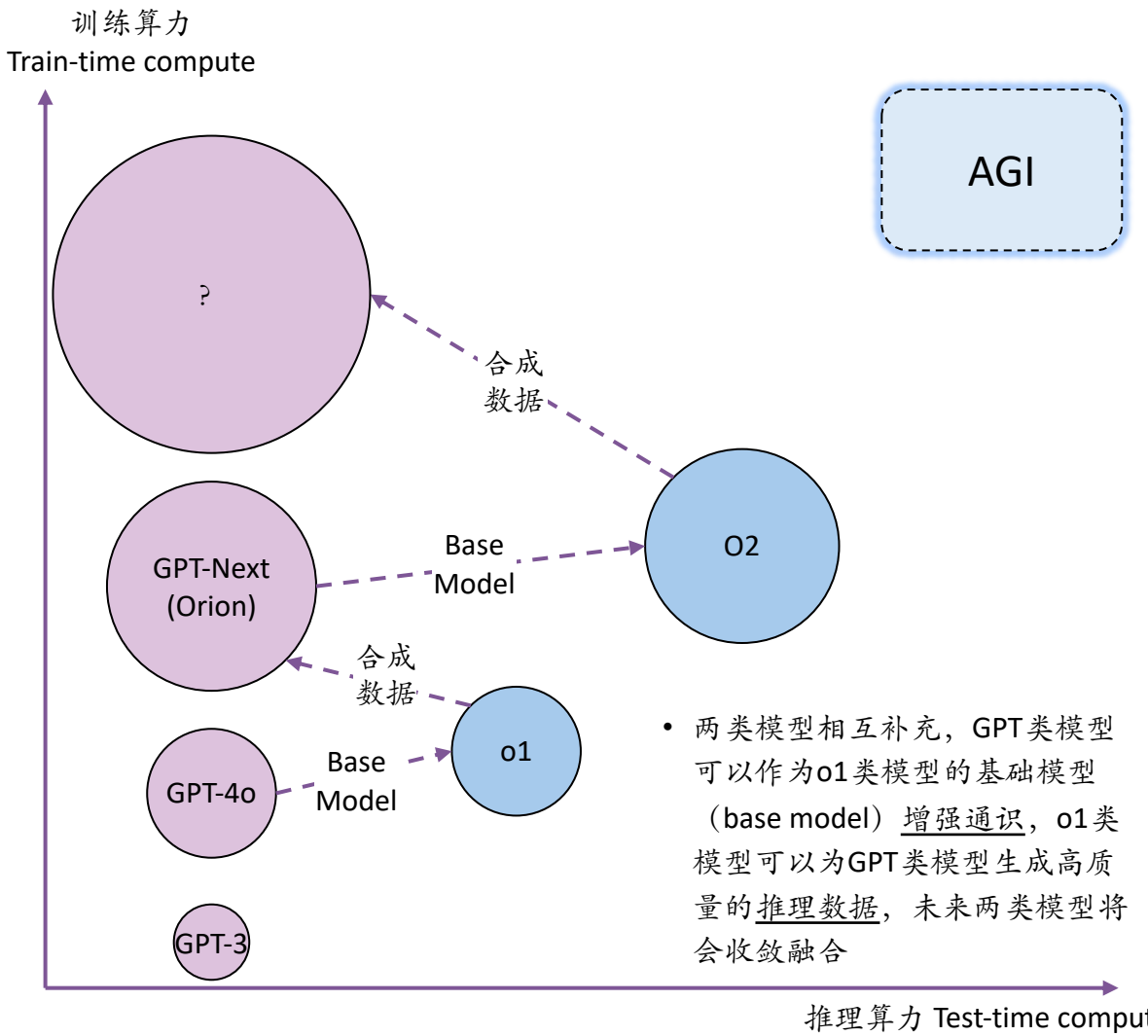
	描述
预训练	• 整体类似GPT路径，但可能需要细颗粒度的推理数据进行增强，比如通过强化学习生成的合成数据来来调整预训练阶段的数据构成、增强推理能力 • 可能复用GPT系列的基础模型（base model）
后训练	• 强化学习成为后训练阶段的关键技术范式，但具体技术实现路径尚未清晰，实现难度大 • 利用RL生成大量的高质量数据可以进行模型优化，但对数据质量的要求远高于GPT系列 • 算力需求很高，万卡集群起，算力需求不低于GPT类模型的预训练阶段
推理	• 类似系统2的思考模式，在推理阶段生成大量中间阶段的思考数据，每次调用的Token消耗量剧增 • 除用量提升外，模型思考时间越长，算力需求越大，未来的长程思考将延伸到小时级甚至数天 • 推理本身会成为强化学习和合成数据等关键技术的一部分

算力分布



信息来源：量子位智库，1）粗略描述算力分布的变化情况

模型技术趋势：o1模型代表的深度推理模型、推理算力（Test-time compute）开辟了新的模型发展路径，将与GPT系列并行发展、相互促进



- 两类模型相互补充，GPT类模型可以作为o1类模型的基础模型（base model）增强通识，o1类模型可以为GPT类模型生成高质量的推理数据，未来两类模型将会收敛融合

GPT系列

系统1

- 系统1（System 1）：思维中的浅层、快速反应系统，负责处理日常生活中的快速、直觉性反应和基础认知任务，不需要刻意思考或花费太多精力
- 模型特点：全面掌握通识类知识，适用于非深度STEM[1]类问题，有更好的多模态交互能力
- 能力要素：主要能力来自大规模Transformer在预训练阶段学习理解的大量多模态信息和知识
- 发展方向：更大的模型参数规模，更好的训练数据质量（包括合成数据，决定模型智能上限），模型架构优化（增加从数据集中提取智能的提取效率）

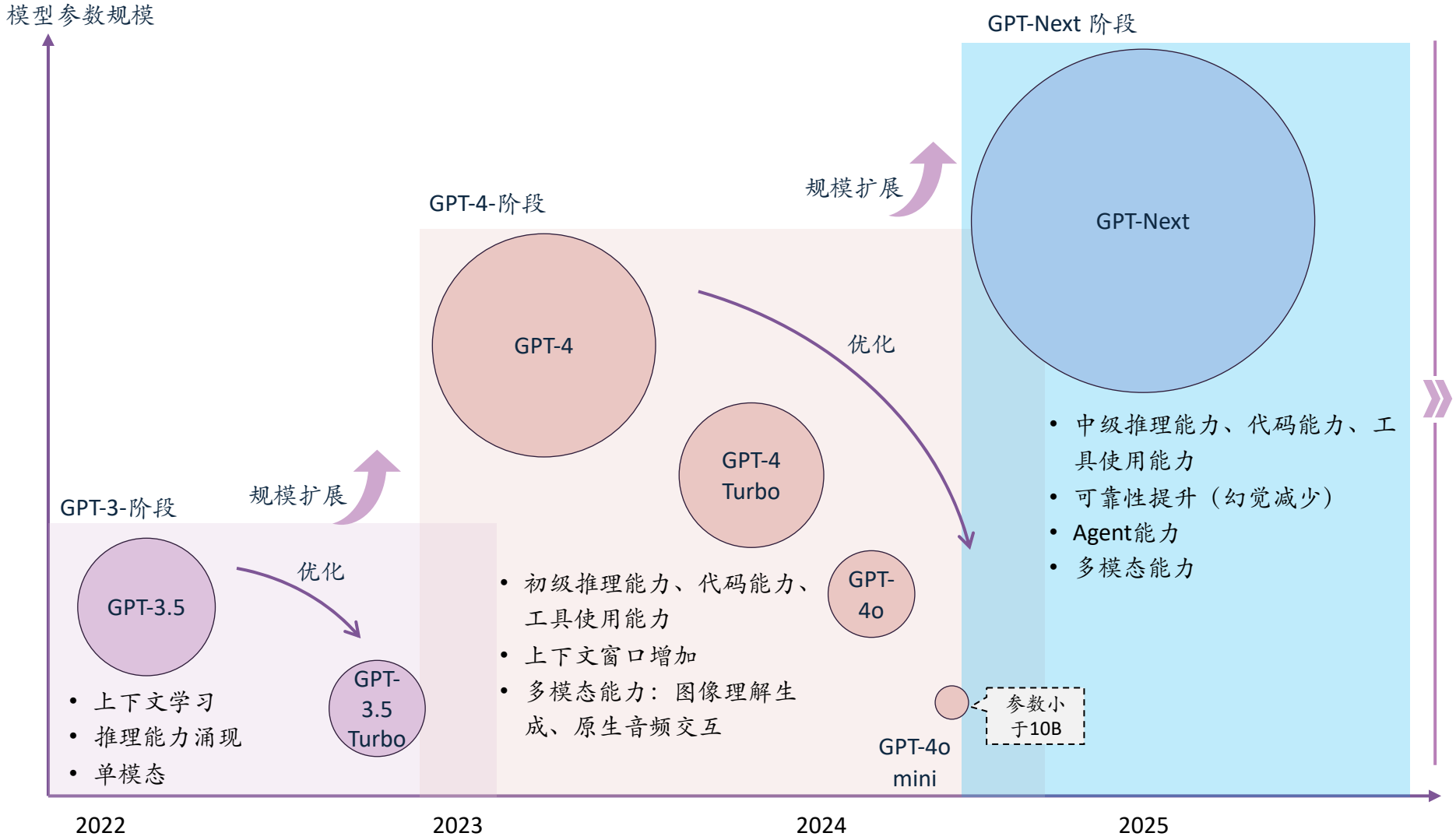
o1系列

系统2

- 系统2（System 2）：思维中的慢速、逻辑性系统，负责复杂的分析、深思熟虑的决策和解决新问题。这个系统需要消耗更多的认知资源和时间
- 模型特点：可以进行深度推理，处理复杂的逻辑分析、任务分解、规划、反思
- 能力要素：主要能力来自于在推理过程中有足够长的思考时间，在推理阶段产生大量思维链分析过程
- 发展方向：后训练阶段优化（让模型有更好的“思考方式”），更长的思考时间（可以思考数小时、数天、甚至更长）

信息来源：量子位智库，OpenAI，1）STEM指科学、科技、工程、数学

模型技术趋势：大模型的扩展规模（Scale up）和精简优化（Scale down）将同步进行，小模型将更快更便宜，但仍需训练大型模型作为基础



关键分析

- 规模扩展：主要是继续扩大模型参数规模，训练数据规模，以及有充足的算力来完成训练，模型的规模扩展（scale up）依然是发展主线
- 精简优化：目前主流方式包括：
  - 1）蒸馏：用大模型生产的高质量数据对小模型进行训练，用数据质量弥补模型规模的不足，可以取得稍逊于大模型的表现，未来蒸馏可能成为下一代模型的预训练前置过程
  - 2）量化：降低模型参数的数值精度（例如从FP8到FP4），降低对于硬件显存和通信带宽的需求，推力更快成本更低，
  - 3）剪枝：对于神经网络中权重值接近于0的参数进行删减，让模型更加紧凑高效，
  - 4）架构创新及其他技术

信息来源：量子位智库

模型应用趋势：随着模型能力提升，基于结果、价值创造的商业模式将逐渐涌现，但目前仅限于少量场景

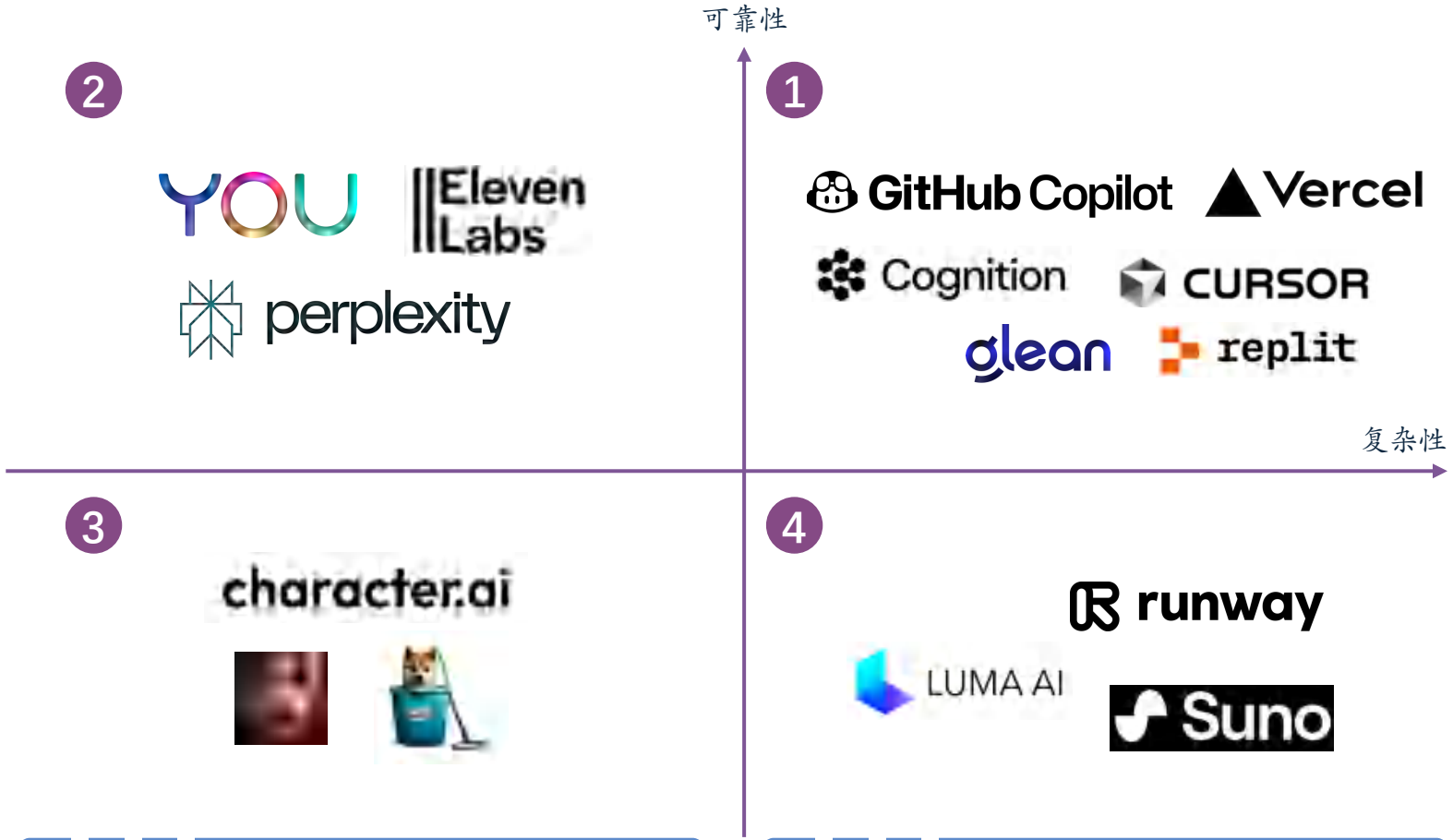
	基于用量	基于结果
核心理念	• 卖Token，通过API调用即可	• 主打“直接完成工作”，例如通过Agent自动化更多环节，最终可以端到端直接完成工作、交付价值
特点	• 简单易用，复杂性低，定制化程度低，模型能力是唯一的差异化	• 更加复杂，需要可衡量的结果和交附加值，目前只能在特定的应用场景展开，可能需要和客户进行合作，有更高的业务整合难度
应用层	• 目前大多数应用以订阅制的形式向客户收费，但本质上是基于底层模型用量的模式，再增加一部分产品层的溢价构成定价	intercom • 提供订阅制之外的AI解决方案，仅对成功解决的客服问题进行收费，每个问题0.99美元 Zendesk • 提供订阅制之外的客服自动化Agent，仅对成功解决的客服问题进行收费，每个问题0.99美元 SIERRA • 完全按解决的问题结果付费，AI执行任务、交付成果，并因成功而获得报酬，已有2000万美元ARR
模型层	• API全部基于用量	• 目前的模型厂商的产品形态还不支持直接基于模型产生的实际效益进行定价，<u>但基于结果定价可以更好地与用户对齐利益</u>，目前实现路径尚不清晰，但将成为未来的发展方向，<u>可能先从垂直场景开始探索</u>

分析

- 大模型将不断拓展在各类产品和应用上的边界，深度融合到现有产品中，<u>延伸的最终边界是直接完成业务结果</u>完成价值创造，需要两个因素成熟：
- 1）产品整合模式更加成熟，基于结果的商业模式需要更多在产品和用户侧的整合、定制化，搭建结果衡量体系，目前的通用API难以满足需求
- 2）模型能力提升，只有较高的模型能力（可靠性高，多步推力强等）和产品形态（如Agent）达到较高水平，以实现完全自动化

信息来源：量子位智库

模型应用趋势：不同类型应用对于模型的需求各异，模型API和模型服务在未来将出现更多垂直优化方向以满足不同类型应用需求



- **高可靠性：**对于模型的推理能力、指令理解能力、低幻觉、生成结果精确度要求较高，例如生产力场景普遍对于模型的可靠性有较高要卡

- **高复杂：**产品的功能多样，交互逻辑复杂，模块构成丰富，用户需求多元，例如AI视频的生成编辑，软件开发流程

分析

1. **高可靠+高复杂：模型能力>生成速度>应用成本，**主要包括B端高价值场景，例如代码生成、企业内知识库搜索等，产品较为复杂，生成量相对小
   - 例如代码生成，需要模型精准实现功能，同时可以快速反馈结果，对于成本敏感度低
2. **高可靠+低复杂：应用成本>生成速度>模型能力，**主要包括功能单一、场景稳定的应用场景，例如翻译、语音转写、通用类查询等场景，生成量大
   - 例如AI搜索，需要处理大量的信息进行总结分析，对成本较为敏感，可靠性、速度需求较弱
3. **低可靠+低复杂：模型能力>生成成本>生成速度，**主要包括AI陪伴类应用，需要高质量输出维持用户满意度，同时多轮对话的生成量较大
   - 例如虚拟聊天类应用，需要定制化不同模型给用户多样化体验，同时保持多轮对话的一致性
4. **低可靠+高复杂：模型能力>生成速度>生成成本，**主要包括多模态应用，需要保证模型单次生成质量，且需要生成多次以快速反馈满用户需求
   - 例如视频生成、编辑应用，不同模型生成效果各异，生成速度较慢影响创作体验

信息来源：量子位智库

开源模型：主要由促进自有业务和模型业务B端获客驱动，将长期存在并促进生态发展、刺激下游应用构建

		描述	示例（国内|海外）	驱动强度
开源模型商业策略	促进自有业务驱动	• 补充产品商品化：不通过卖大模型本身进行商业化，而是通过用大模型更好地赋能现有应用场景，提高产品体验和商业化效率来获益 • 开源模式可构建相关技术生态提高模型使用体验	腾讯混元系列 Meta Llama系列	●
	云服务和增值服务驱动	• 赋能云业务发展：适用于云厂商，自身作为模型托管方，通过开源模型在云上的部署和调用费用盈利 • 可以同时推相关软件增值服务进行商业化，比如应用开发平台、模型数据集管理等	阿里巴巴千问系列 腾讯混元系列 Microsoft MAI/Phi系列 databricks DBRX系列	◑
	大模型业务驱动	• 模型获客工具：通过开源方式构建开发者生态，吸引开发者使用自家模型、产品 • 后期通过卖旗舰模型（闭源）的API和相应的模型训练、微调服务盈利	DeepSeek系列 零一万物Yi系列 智谱·AI GLM系列 MISTRAL AI_ Mistral系列	◔
	长期生态建设	• 塑造品牌形象：促进大模型行业发展，最终反哺业务，刺激更多的模型使用需求反哺大模型和云业务	阿里巴巴Qwen系列 Google Gemma系列	◔
开源模型应用优势		• 灵活度更高：可以训练、微调和蒸馏自己的模型，定制适配最适合业务场景的模型，使用开源工具链构建应用的迁移成本更低 • 自主可控：可以避免被少数几家闭源模型供应商束缚 • 数据安全：隐私数据无需传送到闭源厂商		

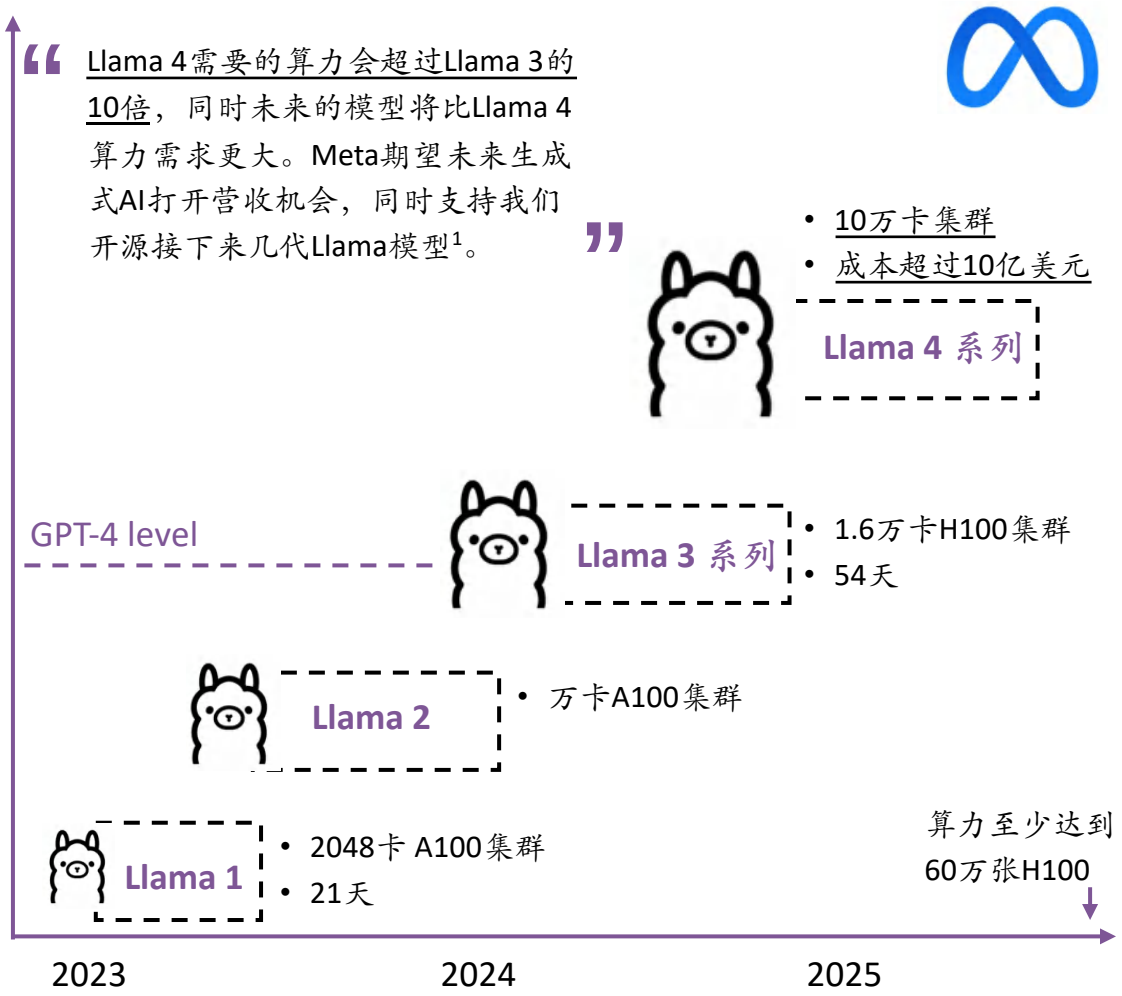
关键分析

- 目前开源模型主要是两类：
  - 1）全系开源，包括旗舰模型：目前只有Meta的Llama系列
  - 2）低配版版开源模型，一般参数规模较小，更加轻量化，主要承担模型厂商生态建设、影响力打造的功能，不贡献营收
- 随着未来模型训练成本增加，主流的开源模型格局将进一步集中，Meta的Llama系列模型将继续主导开源基础模型的格局，在模型能力上和其他开源模型拉开差距，并在相关生态构建（如推理部署）上加强优势
- 用于研究、探索类的各种小规模或者端侧开源模型将持续存在

信息来源：量子位智库

开源模型：2025年训练成本10亿美元级的开源模型（Meta Llama 4）将公开可用，促进基于开源模型的大模型市场发展

模型能力



Meta开源策略分析

基因禀赋

- 缺少B端积累：Meta基因是C端公司，没有微软、谷歌、亚马逊等云厂商在B端的深厚积累，难以打开大模型的B端市场与其他巨头竞争，同时Meta不是云厂商，没有靠大模型卖云的动力
- C端流量稳固：Meta核心业务不在大模型冲击范围内，社交网络目前没有被覆的风险，核心业务如Family of Apps（Fackbook、Instagram、Messager、WhatsApp），有非常稳定的用户网络和内容平台，有很好的AI商业化场景，Meta AI的 MAU已超5亿

竞争态势

- 缺乏技术领先度：Meta在大模型上目前尚未进入海外第一梯队，通过闭源卖API的商业化路径没有竞争优势，保持闭源策略相较OpenAI、Anthropic及谷歌没有竞争优势
- 业务自主性需求：Meta希望能够围绕自身构建开源大模型技术生态，避免被其他玩家牵制，例如和苹果的历史矛盾导致发展受限，刺激了Meta自建生态、平台的需求

过往经验

- 开源经验丰富：Meta 有长期开源项目的成功经验，硬件方面曾通过开源数据中心设计（OpenCompute）从而引领行业标准，在建设数据中心时节省数十亿美元，软件方面曾开源Pytorch 成为最成功的深度学习框架，引领了深度学习的框架标准
- Meta 同样希望 Llama 将成为开源大模型行业的标准，使自身业务生态、应用生态在未来受益

信息来源：量子位智库，1）Meta, Q2 2024 财报电话会

03

大模型竞争趋势与玩家格局

大模型竞争趋势与玩家格局：

模型竞争特点

大模型竞争趋势

大模型玩家格局

信息来源：量子位智库

大模型竞争特点：互联网时代成功应用的核心要素不再适用，大模型业务模式没有清晰的护城河，模型厂商需要持续投入竞争

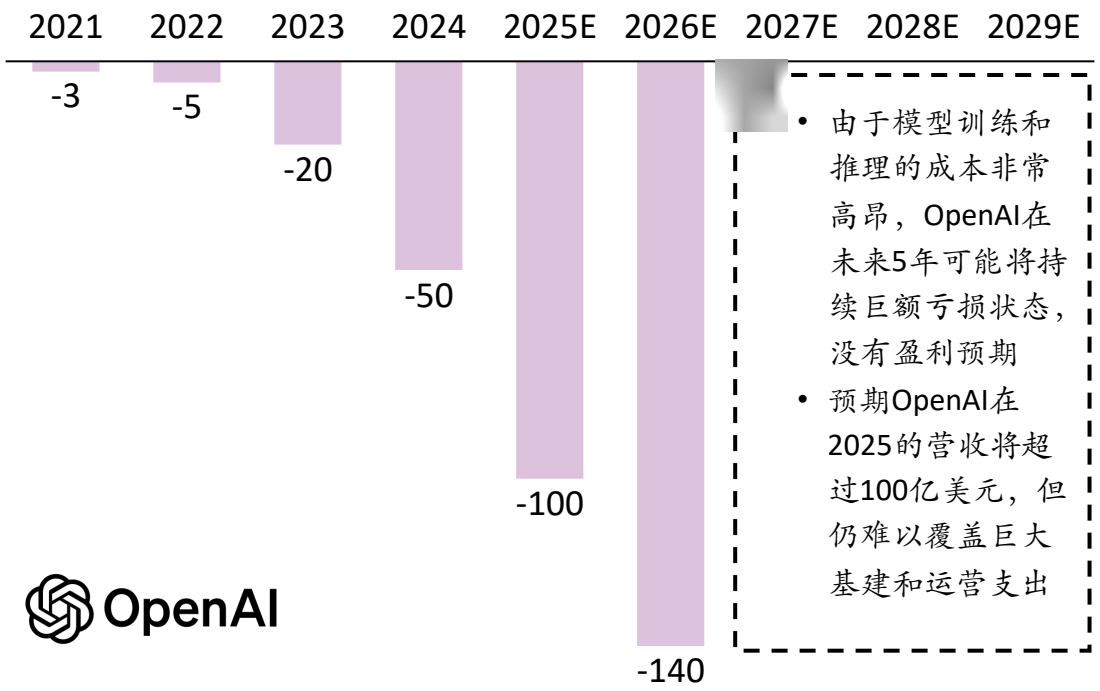
	数据飞轮	网络效应	迁移成本	规模效应	用户心智	关键分析
大模型	• 弱，目前问答数据和行为数据，验对于模型能力、用户体的提升都不明显，数据的质量和复杂度很难作为模型训练的语料 • 由模型生成合成数据或成为关键变量，但目前从外部视角观察效果并不清晰	• 无	• 低，大模型主要是API形式，和业务系统耦合很浅，用户可以轻松切换（大部分厂商都兼容OpenAI API，可以快速切换），应用功能都比较薄、同质化 • 用户最关心的还是模型的智能程度和成本	• 弱，大模型成本尚未从边际走向固定，从基础设施的角度来看，大规模的模型用量有一些负载均衡和成本方面的优势，但目前的发展的瓶颈主要是模型能力，且规模效应带来的成本降低有限	• 中，短期可以通过投流买量、补贴和先发优势打出产品认知度，中期会获得一定用户量，但长期会回归到用户体验（模型的智能程度、产品体验）	• 互联网时代应用主要依赖数据飞轮、网络效应、规模效应、用户心智等要素来建立竞争壁垒，一旦建立壁垒可以形成长期、稳定的竞争优势，并维持较高的利润水平 • 目前大量的价值创造在模型层，模型能力决定产品层的天花板 • 相比于软件，能力剧烈变化且没有上限是大模型业务的常态化特征，需要长期大量的资本投入
社交媒体	• 强，有大量内容、行为数据做推荐	• 强，用户关系网络	• 高，用户关系网络	• 强	• 强	
内容平台	• 强，同上	• 中，有用户网络但不如社交网络强	• 无	• 强	• 强	
工具软件	• 无	• 无	• 高，用户习惯、系统标准集成难迁移	• 强	• 中	
交易平台	• 中，可以积累用户数据做推荐	• 中，积累了商家和用户的双边网络	• 无	• 强	• 中	

信息来源：量子位智库

大模型竞争特点：从领军玩家来看，通用基础模型的发展模式对资源投入要求高、周期长，需要极强的资源背景支撑长期发展

1 基础模型厂商投入周期长，亏损大

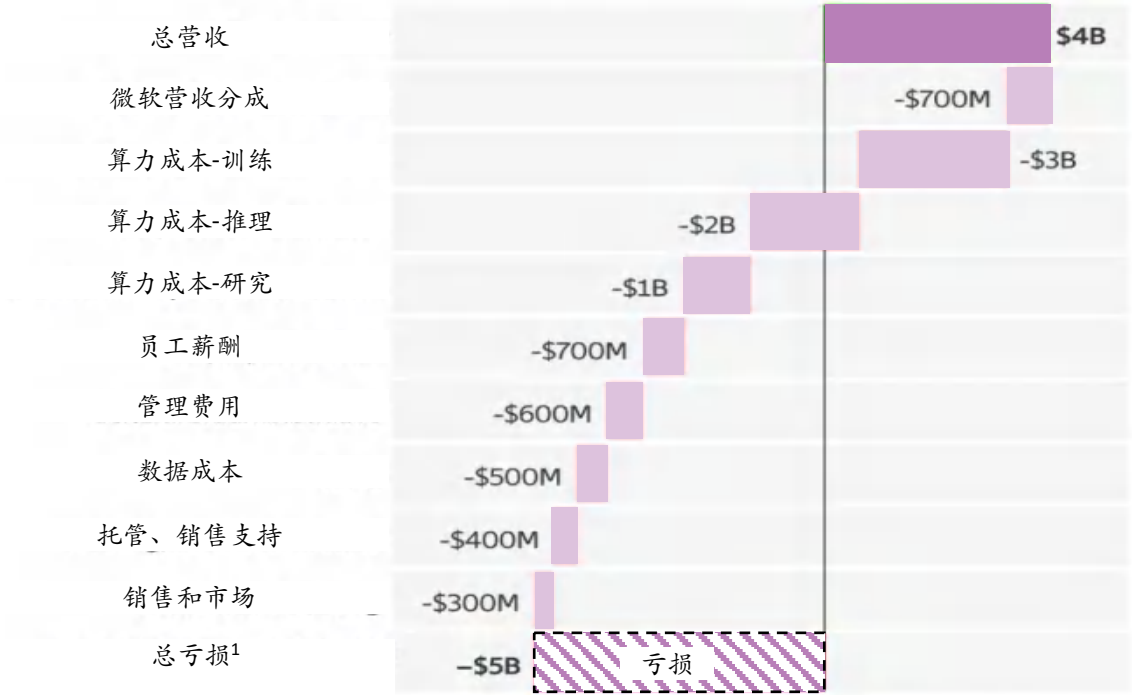
OpenAI亏损历史情况（年份，盈利[亿美元]）



- 基础模型厂商的商业模式更类似制药、光伏等行业，对于资本的投入周期有较高要求，早期的商业模式可能在数年内都无法产生利润，甚至会出现亏损逐年增加的情况

1 基础模型厂商目前普遍处于早期亏损状态

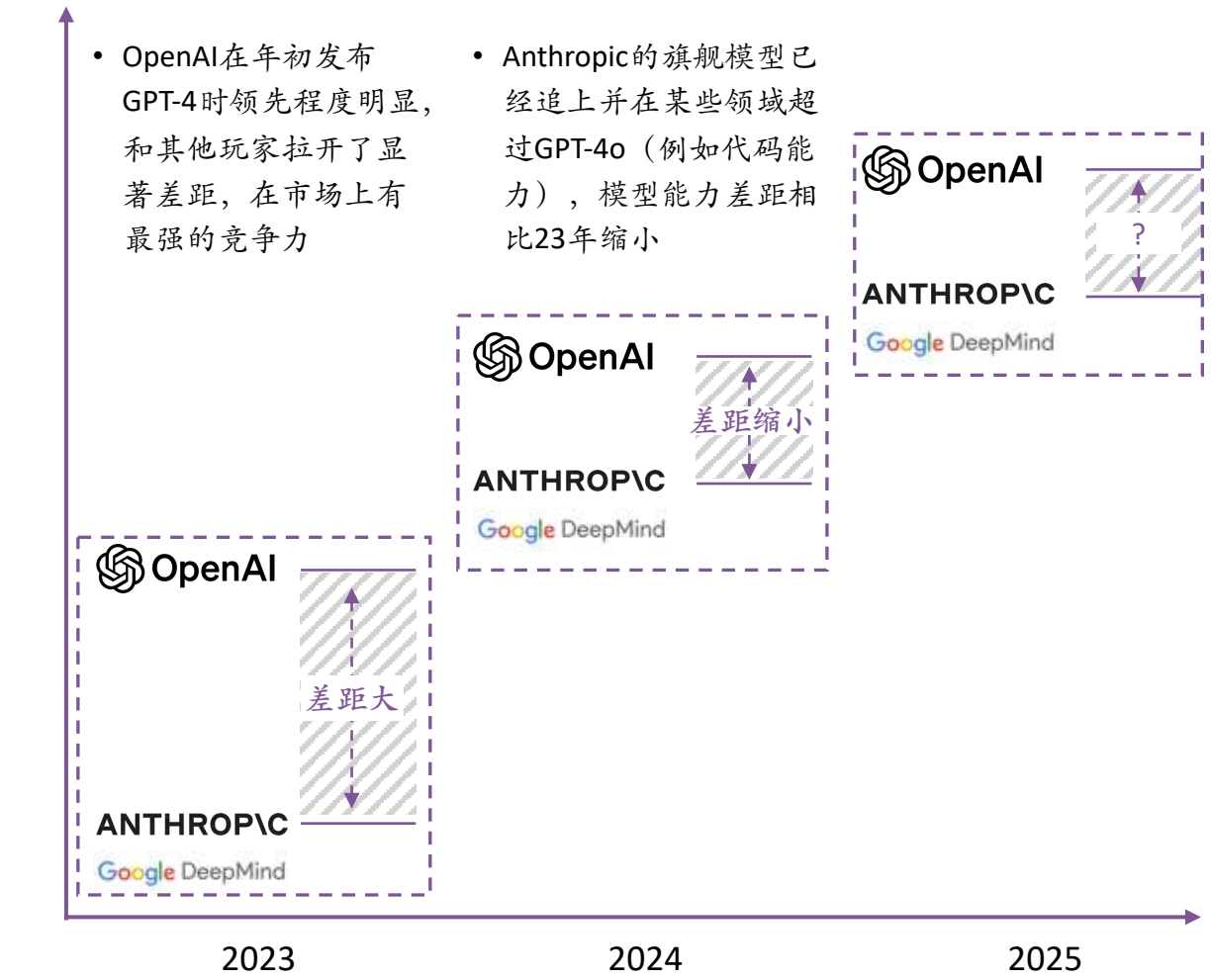
OpenAI基本财务构成情况（2024年，美元）



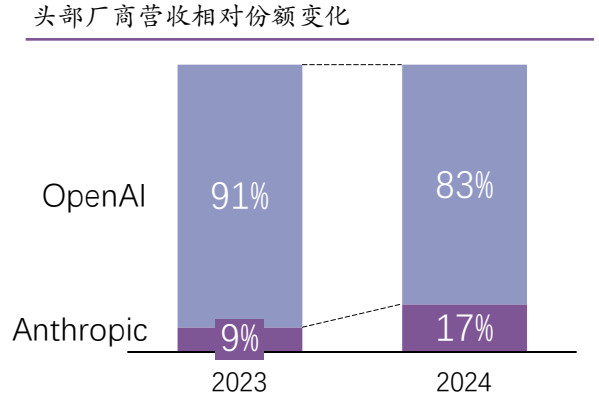
- 目前所有模型厂商都处于亏损阶段
- 模型的训练、推理、和模型预研是目前最大的成本项

信息来源：量子位智库，The Information，1）不包括员工的股权激励（PSU）

大模型竞争特点：模型能力领先程度决定市场份额，因此模型厂商需要进行持续投入，以保证产品用量和营收增长

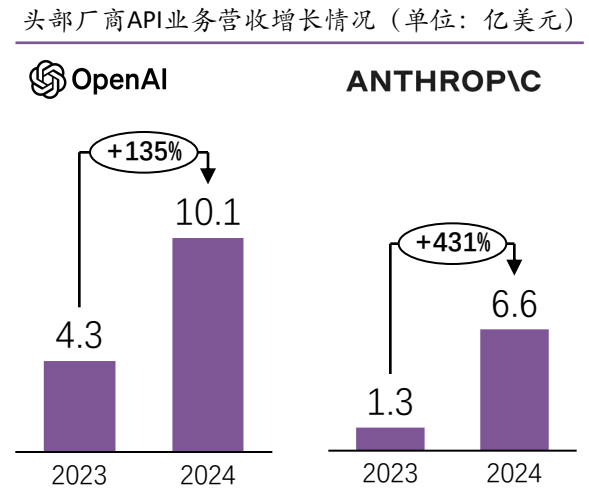


1 模型能力趋同会驱动模型厂商营收差距减小



- 由于模型能力差距的缩小，已经可以看到追赶者Anthropic在相对营收份额上的增长
- 虽然目前总营收和OpenAI差距较大（24年营收约为OpenAI的四分之一），但Anthropic在过去一年就营收增长而言要比OpenAI更快

2 API市场份额变化对于模型能力变化尤其敏感



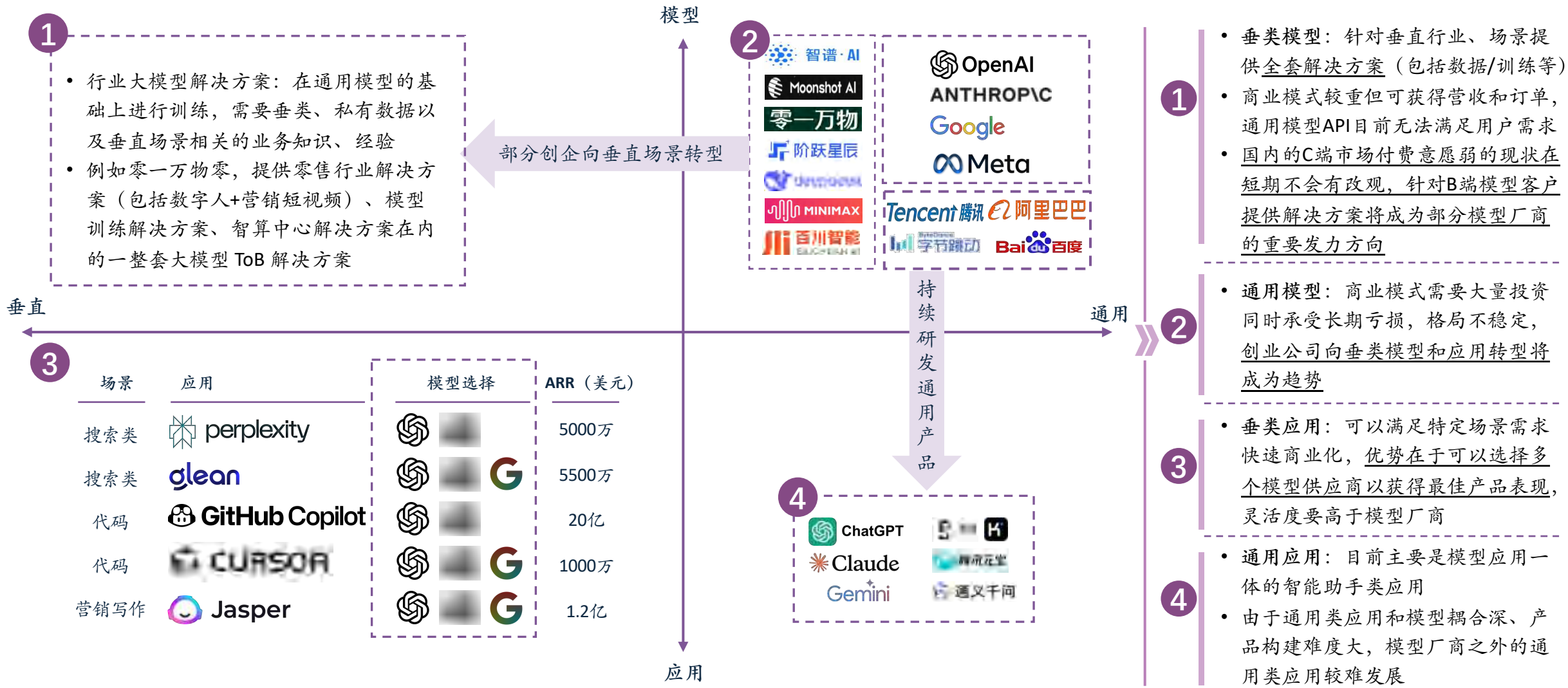
- 由于API用户主要是各类开发者和B端用户，API营收的变化可以更准确地反映市场整体上对于模型能力的评估
- 从API角度看，当追赶者模型能力接近头部水平时，营收规模差异会随着模型能力差异的缩小而收敛，Anthropic在过去一年API业务上的营收增速远超OpenAI

信息来源：Sacra，量子位智库

大模型竞争趋势：从成功要素上看，模型能力>生态能力>渠道能力，总体上云厂商/互联网公司优势全面，通用模型未来格局将持续向其集中

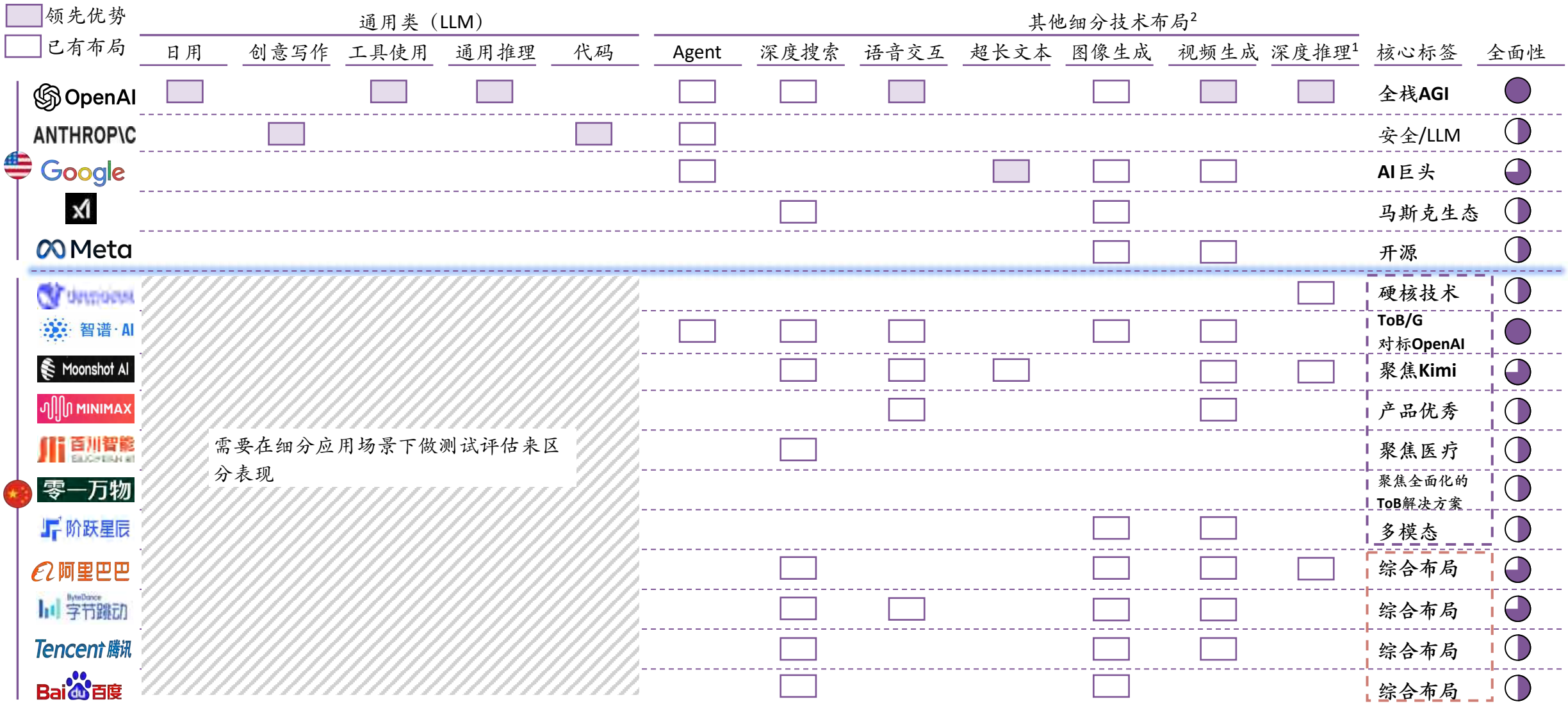
	模型能力	生态能力	渠道能力	示例
互联网公司/云厂商	• 强，目前国内云厂商的大模型和一线大模型厂商在基础模型层面没有显著差距，整体相较在同一梯队 • 现阶段在模型能力上的目标是全面达到GPT-4的水平，以及在特定场景实现超越	• 强：云厂商有最全面的业务生态，字节、阿里、腾讯、百度都有丰富的消费端场景和现有产品，可以接入自家大模型进行业务尝试，提升用量	• 强：云厂商可以依托云业务长期积累的销售渠道，触达能力更好。此外大模型可以和云产品搭配进行交叉销售（cross-sell）	阿里云 火山引擎 百度智能云 腾讯云
大模型厂商	• 同上	• 弱：需要单独做产品吸引用户，缺乏现有场景	• 弱：主要靠挖其他行业的B端销售人员构建渠道，但基础薄弱，很难与云厂商竞争，对创始团队的资源依赖也比较高	智谱·AI Moonshot AI 阶跃星辰 MINIMAX 零一万物 百川智能
推理服务平台	• 中，推理平台本身不研发模型，主要承接各类开源模型的托管推理需求，能力取决于SOTA开源模型水平 • 优势主要在于模型API的价格低廉、性能稳定，目前基本没有微调等模型服务，主打开源API	• 弱：以模型API为主，不涉及产品，主要整合开源模型生态，搭建推理平台开发者社区	• 无：主要面向开发者、中小型客户，直接卖开源模型API	硅基流动 INFINIGENCE 无问芯穹
传统技术类厂商	• 中，模型能力弱于云厂商和大模型厂商，	• 中：在自身存量业务上可以接入大模型，但规模较小	• 强：之前的存量业务主要面向B端、G端，和大模型客户画像类似，积累了较多相关的资源网络	商汤 科大讯飞 iFLYTEK

大模型竞争趋势：通用基础模型领域竞争激烈，将驱动部分模型厂商向针对垂直场景的模型服务和产品进行转型，大厂将持续押注通用类产品



信息来源：量子位智库

大模型竞争趋势：由于模型、产品能力维度多样，模型厂商难以在所有维度达到最佳水准，细分差异化是国内模型厂商的未来方向



信息来源：量子位智库，1）对标OpenAI o1的推理模型，2）主要指C端的功能，不涵盖B端布局

模型厂商概览-国内：大厂模型业务将结合自身场景推动闭环，并持续投入，创业公司开始聚焦解决商业化问题，格局未来1-2年将迎来洗牌分化

国内模型厂商竞争格局

1 互联网公司

2 头部创业公司

3 其他厂商

- 国内模型厂商的竞争格局目前尚未收敛，预期未来格局进一步集中，长期来看国内的通用模型厂商不会有第二梯队，将进一步收敛到少数玩家
- 大厂短期都会持续投入模型研发，同时加速模型应用在自由场景下实现闭环；创业公司方面，各家的融资能力和资源禀赋不同，未来策略重心会出现分化，大模型的热潮已过，之后会探索商业化前景更好的业务

1
- 云计算业务：互联网公司（一般也有云业务），大模型可以驱动云业务增长，所以云厂商是大模型的受益方，大模型业务的驱动因素包括：
  - 1）云计算是未来重要的增长曲线和战略方向，大模型相关的推理训练可以直接拉动云业务的营收增长，大模型是公有云最重要的增量市场
  - 2）由于大模型的业务模式相对传统云业务定制化成分少（例如模型微调训练，不同项目服务流程类似），模型API相当于直接卖公有云，可以获得更优的利润率
  - 3）带动其他云业务增长，如云原生AI应用开发的相关工具、向量数据库、以及传统的计算、存储、网络、安全业务
- 内部场景驱动：互联网公司旗下业务场景可接入自家模型进行迭代应用。以腾讯为例，内部有700+场景，例如腾讯会议、微信搜索、微信读书等可以直接被模型赋能，在内部实现应用

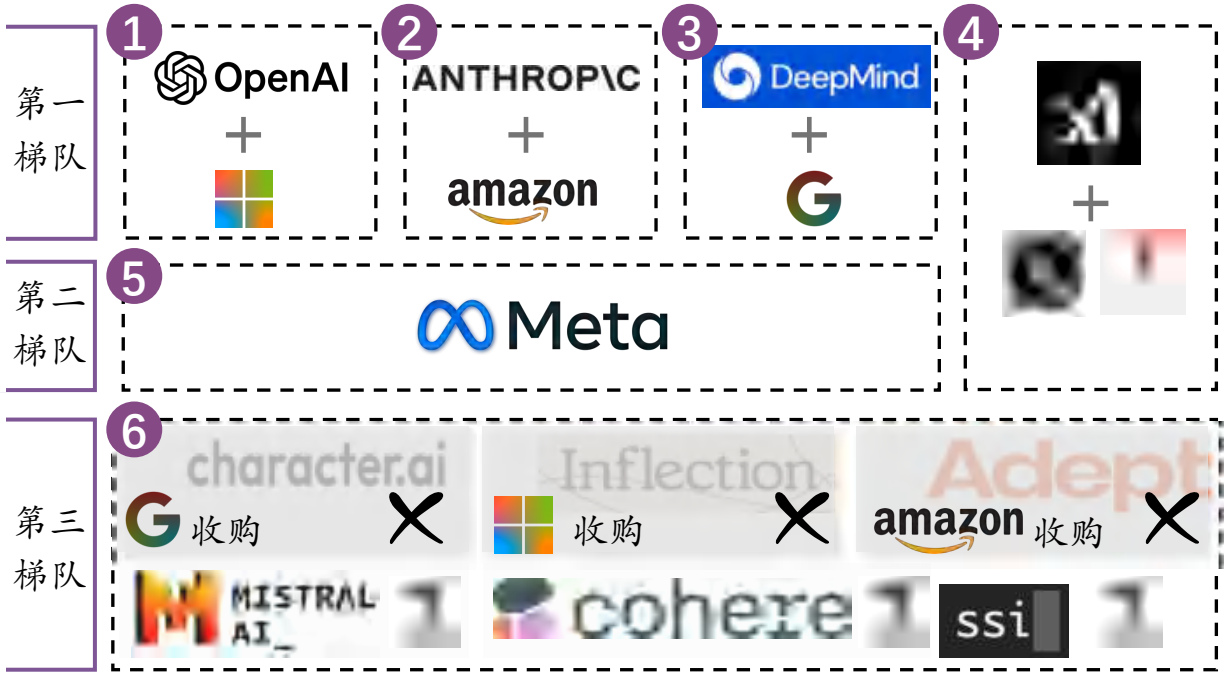
2
- 持续进行基础模型训练以保持模型能力领先成本极高，国内大模型的商业化环境很差，这将驱动创业公司更加聚焦，发力应用或者深耕垂直行业、场景会成为短期的方向，整体盈利预期遥远

3
- 传统的技术类厂商以及运营商也会做大模型，但主要思路还在已有业务和渠道上做延伸，给政府大客户、行业大客户做私有化、定制化模型服务，同时搭配其他产品服务，API的竞争力在市场上较弱，整体属于资源导向型业务，但目前在尝试从传统AI业务到大模型的转型

信息来源：量子位智库

模型厂商概览-海外：格局已经收敛到头部5家超级公司，预期未来几年持续巨量投入实现AGI，残存腰部玩家生存空间将被进一步挤压

海外模型厂商竞争格局



- 海外基础模型玩家格局已经收敛，只有巨头及巨头深度合作的玩家可以留在牌桌，资本投入需求以数百亿美元计量，竞争窗口已经关闭
- 24年第三梯队玩家已经开始淘汰出清，被巨头收购是主要的推出路径，未来基础模型不会有第3梯队，将会长期收敛到5家头部公司

1. 2024年营收约40-50亿美元，超70%来自ChatGPT订阅服务，和微软进行了深度绑定，包括产品整合、营收分成、算力供应、数据中心建设等方面，最近一年市场份额下降，Anthropic和谷歌获得了更多份额
2. 2024年营收约10亿美元，超80%来自API服务，其余来自Claude订阅服务（营收构成上和OpenAI差异较大），和Amazon进行了深度绑定，包括算力供应和产品整合，最近一年市场份额快速增加
3. 营收情况不详，算力上依靠Google Cloud，已与谷歌工作流进行整合并其他相应增值服务服务打包销售，同时发力针对中小企业的API市场，最近一年市场份额快速增加
4. 营收情况不详，目前主要产品Grok作为X平台的增值服务面向用户，API服务近期投放市场，模型能力目前尚未完全达到第一梯队水平，但算力和人才等资源增长迅速，是目前最有潜力的模型厂商
5. 选择开源路线，由其他云厂商、推理平台以及用户自己负责部署，目前尚未通过模型进行商业化，旨在构建围绕Llama模型和Meta的开源生态，最终助益自身核心业务发展
6. 创业公司商业化困难，且无法持续投入模型开发，产品竞争力弱于头部玩家，缺乏独立生存能力，多被巨头收购，预期未来1-2年剩余第三梯队玩家也将面临收购结局

信息来源：量子位智库

相关玩家及服务概览-1：当前国内大模型市场主要包括互联网公司（云厂商），大模型创业公司，模型推理平台及技术类厂商4类

	玩家	应用开发平台	模型API	模型服务（仅讨论平台产品）
国内互联网公司/云厂商	阿里巴巴	• 阿里云百炼大模型服务平台 • 钉钉AI助手	• 阿里云灵积 • 魔塔社区	• 阿里云百炼大模型服务平台 • 魔塔社区
	字节跳动	• 扣子	• 火山大模型服务平台	• 火山大模型服务平台
	腾讯	• 腾讯元器 • 大模型知识引擎	• 腾讯云Ti平台	• 腾讯云Ti平台
	百度	• 文心智能体平台Agent Builder • 百度千帆智能体平台App Builder	• 百度智能云千帆大模型平台	• 百度智能云千帆大模型平台
大模型创业公司	Deepseek	• 无	• Deepseek开放平台	• 无
	月之暗面	• 无	• Kimi 开放平台	• 无
	智谱AI	• 无	• 智谱AI开放平台	• 无
	Minimax	• 无	• Minimax开放平台	• 无
	百川智能	• 无	• 百川智能开放平台	• 无
	零一万物	• 无	• 零一万物 开放平台	• 无
	阶跃星辰	• 无	• 阶跃星辰 开放平台	• 无

信息来源：量子位智库，1）有模型推理业务就计入在内

相关玩家及服务概览-2：当前国内大模型市场主要包括互联网公司（云厂商），大模型创业公司，模型推理平台及技术类厂商4类

	玩家	应用开发平台	模型API	模型服务（仅讨论平台产品）
模型推理服务平台[1]	硅基流动	• 无	• SiliconCloud	• 无
	无问芯穹	• 无	• 无问芯穹大模型服务平台	• 无问芯穹大模型服务平台
	潞晨科技	• 无	• 潞晨云	• 无
	清程极智	• 无	• 清程云MaaS	• 无
技术类厂商	商汤科技	• 无	• 商汤日日新开放平台	• 无
	科大讯飞	• 科大讯飞星火智能体平台	• 讯飞开放平台	• 讯飞开放平台
	昆仑万维	• SkyAgents	• 天工开放平台	• 无
海外云厂商	微软	• Copolit Studio	• Azure AI Studio	• Azure AI Studio
	亚马逊	• Amazon App Studio • Amazon Q	• Amazon Bedrock	• Amazon Bedrock
	谷歌	• Vertex AI	• Vertex AI	• Vertex AI

信息来源：量子位智库，1）有模型推理业务就计入在内

量子位 智库
insights

关于量子位智库：

量子位旗下科技创新产业链接平台。致力于提供前沿科技和技术创新领域产学研体系化研究。

面向前沿AI&计算机，生物计算，量子技术及健康医疗等领域最新技术创新进展，提供系统化报告和认知。

通过媒体、社群和线下活动，基于专题技术报道及报告、专项交流会等形式，帮助决策者更早掌握创新风向。

量子位智库公众号

微信号：Qbitbot020
量子位智库小助手

关于量子位：

量子位（QbitAI），专注人工智能领域及前沿科技领域的产业服务平台。

全网订阅超过500万用户，在今日头条、知乎、百家号及各大科技信息平台量子位排名均为科技领域TOP10，
内容每天可覆盖数百万人工智能、科技领域从业者。

量子位公众号

分析师：Xuanhao（微信：feeltheagi）　智库负责人：李根（微信：ligen603）　商务合作：赵萌（微信：13343397239）