

海外科技行业深度研究报告

模型加速迭代，有望促进应用端快速发展

- **AI大模型是人工智能预训练大模型的简称，其融合了“预训练”和“大模型”两个核心概念，开创了一种全新的智能模式。**具体而言，AI大模型通过在海量数据集上进行预训练，能够在仅使用少量数据进行微调，甚至无需微调的情况下，直接支持多种应用场景。模型通常采用多层神经网络架构，并借助高级优化算法和强大的计算资源进行训练，从而具备出色的泛化能力、通用性和实用性，广泛应用于自然语言处理、计算机视觉、智能语音等多个领域，并在其中取得了突破性的性能提升。
- **大模型是实现通用人工智能（AGI）的关键路径之一。**缩放法则表明，随着模型规模的不断扩大，其在各项任务上的表现会持续提升。而涌现性则意味着，当模型规模达到一定阈值时，其对某类任务的效果会突然出现显著的性能增长，并涌现出新的能力。当大模型能够存储并动态连接人类的全部知识时，其展现出的智能水平将远超人们的预期。AGI的实现将推动人工智能从“数据飞轮”向“智慧飞轮”的演进，最终迈向人机共智的未来。现有的AI体系主要依赖数据飞轮，而AGI催生了新的研究范式——智慧飞轮。通过强化学习和人类反馈，智慧飞轮能够不断解锁基模型的新能力，从而更高效地解决海量的开放式任务。
- **对比海外各家大模型厂商，各厂商进化重点与路径不一：**

 1. **OpenAI:** 使命是确保通用人工智能（AGI）能够造福全人类。OpenAI的AGI使命主要聚焦于两个核心目标：一是开发出能够解决人类水平问题的系统；二是构建安全且有益的AGI。
 2. **LLaMA:** 目标是通过持续优化核心能力（如推理、编码）和扩展多模态边界（如加入语音功能和视频理解能力），打造“一模多能”的AI。未来，LLaMA有望在多模态问答系统、智能助手、教育、医疗等领域发挥更大作用。
 3. **Anthropic:** 致力于将Claude打造成一个更智能、更灵活的AI助手。Claude 4将推出“混合AI”模型，通过滑动条灵活调整处理token的数量，以适应从简单到复杂的各种任务。
 4. **xAI:** Grok系列模型在自然语言处理、复杂推理、代码生成以及内容创作等领域展现出强大实力，尤其在实时知识集成和可解释性方面具有显著优势。未来，xAI将重点拓展多模态能力，集成图像和视频处理功能，进一步提升模型在视觉任务中的表现。
 5. **Gemini:** 谷歌明确表示，Gemini 2.0是为“代理时代”设计的模型，具备自主理解、规划和执行任务的能力。
- **模型不断迭代，有望利好应用侧快速发展，建议关注以下标的：**1)办公：MSFT；2)金融：AFRM、UPST、SOFI、LMND；3)工业：CDNS、SNPS；4)医疗：TEM、RXRX、DOCS、HIMS、VEEV；5)创意：ADOBE；6)教育：DUOL；7)电商：SHOP、SEA；8)企业服务软件：SAP、CRM、NOW、TEAM、ORCL、WDAY、HUBS、FRSH、CXM；9)安全：PANW、CRWD、NET、FTNT；10)数据基建：ESTC、SNOW、MDB；11)云服务提供商：AMZN、META、MSFT、GOOG、ORCL；12)广告：APP、U、ZETA等。
- **风险提示：**竞争加剧，商业化不及预期，技术更新较快。

推荐（维持）

华创证券研究所

证券分析师：吴鸣远

邮箱：wumingyuan@hcyjs.com

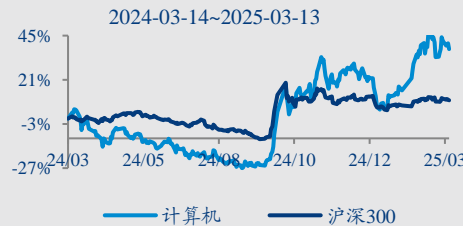
执业编号：S0360523040001

行业基本数据

		占比%
股票家数(只)	336	0.04
总市值(亿元)	49,060.09	4.91
流通市值(亿元)	41,734.84	5.24

相对指数表现

	%	1M	6M	12M
绝对表现	2.1%	84.8%	35.6%	
相对表现	1.9%	61.0%	26.1%	



相关研究报告

《海外科技行业重大事项点评：港股科技：AI+SaaS带来资产重估》

2025-03-10

《计算机行业周报（20250303-20250307）：AI+军工，推动国防信息化建设》

2025-03-09

《AI+专题系列点评（三十）：DeepSeek开源周，重塑AI基础设施效率》

2025-03-02

投资主题

报告亮点

本报告深入探讨了 AI 大模型的发展现状与未来趋势，指出 AI 大模型作为实现通用人工智能（AGI）的关键路径，正推动人工智能从“数据飞轮”向“智慧飞轮”演进。通过对 OpenAI、LLaMA、Anthropic、xAI 和 Gemini 等主要厂商的最新产品和技术路线的对比分析，报告指出各厂商在多模态能力、推理性能和行业应用方面的差异化发展。

投资逻辑

通过对 OpenAI、LLaMA、Anthropic、xAI 和 Gemini 等主要厂商的深入对比，报告揭示了各厂商在多模态能力、推理性能和应用场景上的差异化优势：OpenAI 以强大的多模态处理能力领先；LLaMA 专注于多语言对话和代码生成；Anthropic 在推理成本控制和企业级应用中表现出色；xAI 在数学、科学和编程任务中表现优异；Gemini 则凭借其自主理解和任务执行能力，为“代理时代”做好准备。报告认为，模型路线分化有望引领不同场景的创新应用，建议关注底层模型更新进展，从而把握应用与算力层的投资机会。

目 录

一、 AI大模型掀起时代浪潮，加速通用人工智能 AGI 时代的来临	5
(一) 人工智能发展进入以 AGI 为代表的新里程碑阶段	5
(二) 大模型是实现通用人工智能 (AGI) 的关键路径之一	6
二、 大模型进入百模大战阶段，发展趋势逐渐清晰	7
(一) 三轮驱动下，AI 模型迎来良好发展契机	7
(二) AI 大模型产业趋势确定，Mass 将重构商业化生态。	9
1、 大模型将趋于“通用化”与“专用化”并行	9
2、 大模型将趋于“平台化”与“简化”并行	10
3、 大模型发展路线逐渐清晰，MaaS 将重构商业化生态	10
三、 海外 AI 大模型主要厂商盘点：发展情况或将分化	11
(一) 总结：当前各家模型最新版本对比	11
(二) Open AI	12
(三) LLAMA	14
(四) Anthropic	16
(五) xAI	17
(六) Gemini	19
四、 投资建议	20
五、 风险提示	20

图表目录

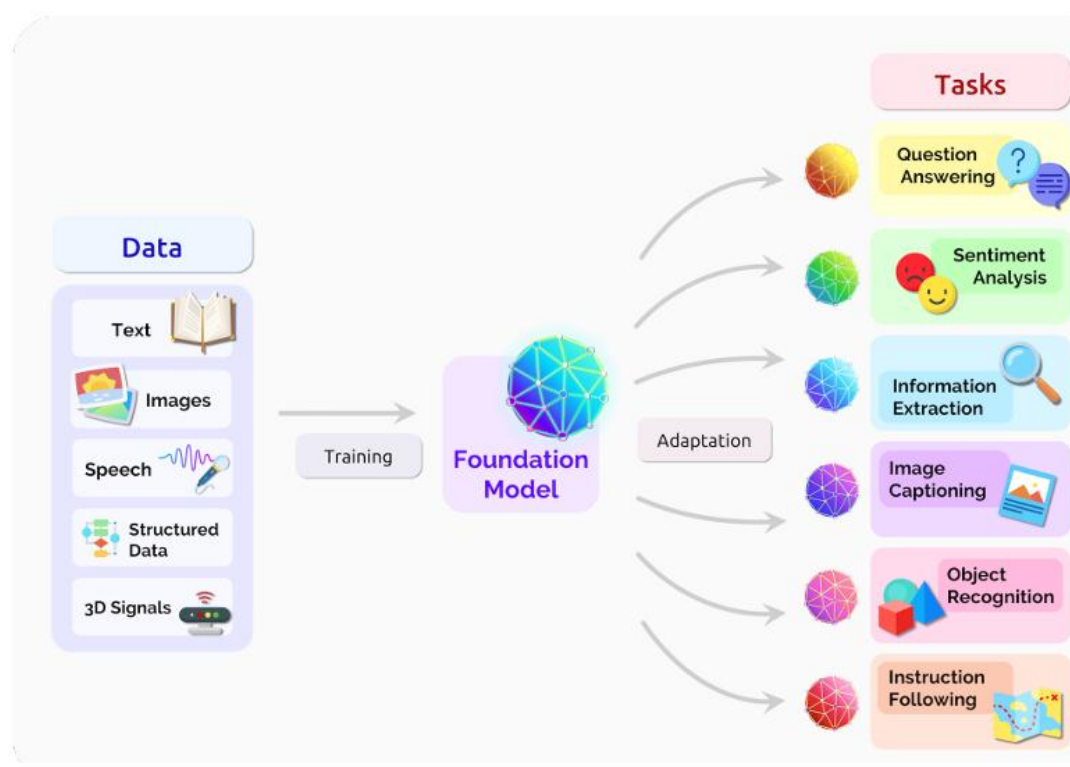
图表 1	AI 大模型的内涵与特征.....	5
图表 2	AGI 包含多个方面以及多个要素.....	6
图表 3	大模型将由数据飞轮向智慧飞轮升级演进.....	7
图表 4	优秀的大模型能够赋能各行各业开放任务.....	7
图表 5	LLM 研究的发展路径.....	8
图表 6	AI 大模型的技术架构.....	9
图表 7	AI 大模型将通用化与专用化并行.....	9
图表 8	大模型平台化与简易化并进.....	10
图表 9	MaaS 产业一体化架构.....	11
图表 10	各模型当前智力、速度、价格对比（截至 2025 年 3 月）.....	12
图表 11	各模型当前智力与价格的对比（截至 2025 年 3 月）.....	12
图表 12	Open AI 模型系列与价格.....	13
图表 13	Open AI 将 AGI 目标放在其官网上.....	13
图表 14	Open AI 最新产品 GPT-4.5 与 GPT-5 相关安排.....	14
图表 15	LLama 各个模型对比.....	15
图表 16	Llama 3.2 使用的多模态 LLM 方法的插图.....	15
图表 17	Anthropic 各个模型对比.....	16
图表 18	Claude Chat 形式.....	17
图表 19	XAI 发展时间线.....	18
图表 20	Grok3 功能实例.....	19
图表 21	Gemini 各个模型对比.....	19
图表 22	Gemini2.0 意图切入 Agent 时代.....	20

一、AI 大模型掀起时代浪潮，加速通用人工智能 AGI 时代的来临

(一) 人工智能发展进入以 AGI 为代表的新里程碑阶段

如何理解大模型？ AI 大模型是人工智能预训练大模型的简称，其融合了“预训练”和“大模型”两个核心概念，开创了一种全新的智能模式。具体而言，AI 大模型通过在海量数据集上进行预训练，能够在仅使用少量数据进行微调，甚至无需微调的情况下，直接支持多种应用场景。模型通常采用多层神经网络架构，并借助高级优化算法和强大的计算资源进行训练，从而具备出色的泛化能力、通用性和实用性，广泛应用于自然语言处理、计算机视觉、智能语音等多个领域，并在其中取得了突破性的性能提升。

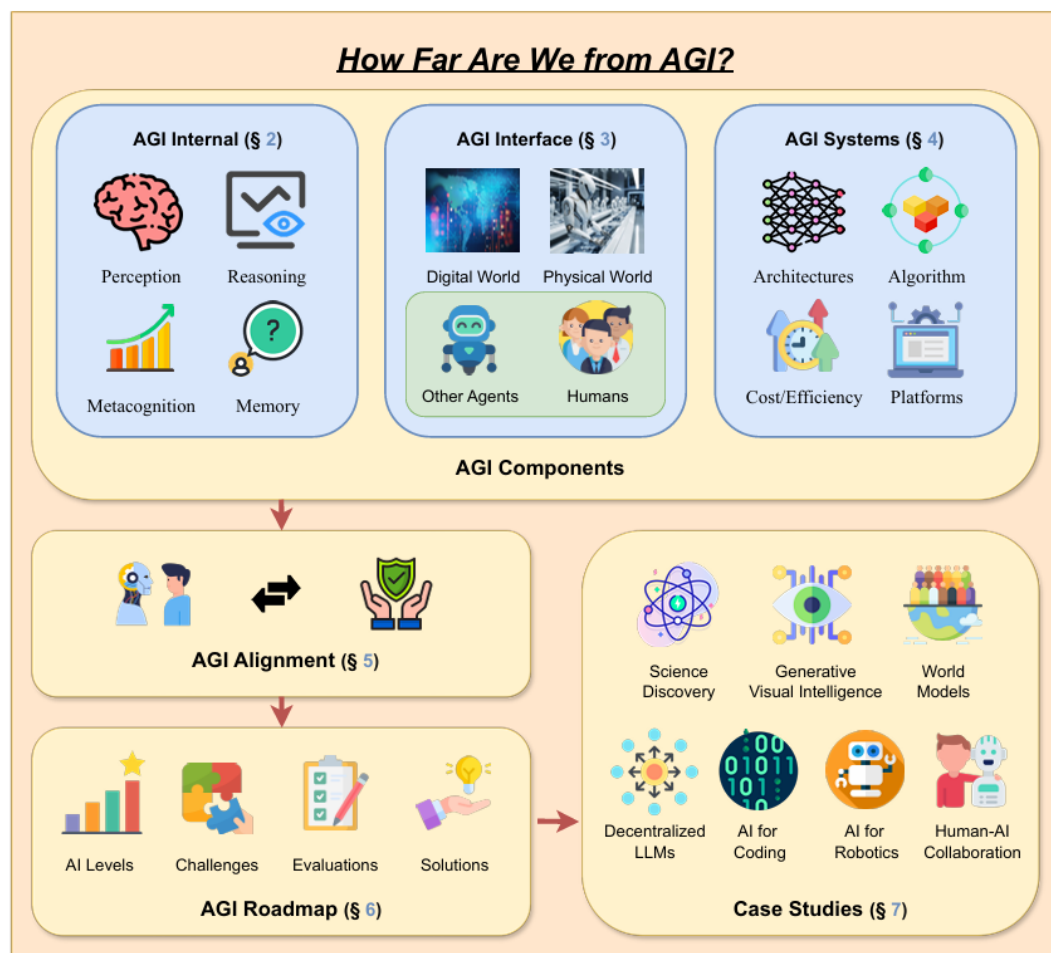
图表 1 AI 大模型的内涵与特征



资料来源：CRFM 《On the Opportunities and Risks of Foundation Models》

如何理解 AGI 技术？ AGI 技术能够精准识别人类情绪意图、理解人类语言、学习人类知识并进行类脑推理与创造，当前通往 AGI 的技术路径多元，目前大模型是最佳实现方式。借 Sam Altman 对 AGI 的定义：如果 AI 模型具有一个“普通人”学习解决问题的综合技能，能够在任何领域变得优秀，那就拥有了 AGI。

图表 2 AGI 包含多个方面以及多个要素



资料来源: Tao Feng 等《How Far Are We From AGI: Are LLMs All We Need?》

(二) 大模型是实现通用人工智能 (AGI) 的关键路径之一

大模型是实现通用人工智能 (AGI) 的关键路径之一。缩放法则表明, 随着模型规模的不断扩大, 其在各项任务上的表现会持续提升。而涌现性则意味着, 当模型规模达到一定阈值时, 其对某类任务的效果会突然出现显著的性能增长, 并涌现出新的能力。当大模型能够存储并动态连接人类的全部知识时, 其展现出的智能水平将远超人们的预期。AGI 的实现将推动人工智能从“数据飞轮”向“智慧飞轮”的演进, 最终迈向人机共智的未来。现有的 AI 体系主要依赖数据飞轮, 而 AGI 催生了新的研究范式——智慧飞轮。通过强化学习和人类反馈, 智慧飞轮能够不断解锁基模型的新能力, 从而更高效地解决海量的开放式任务。

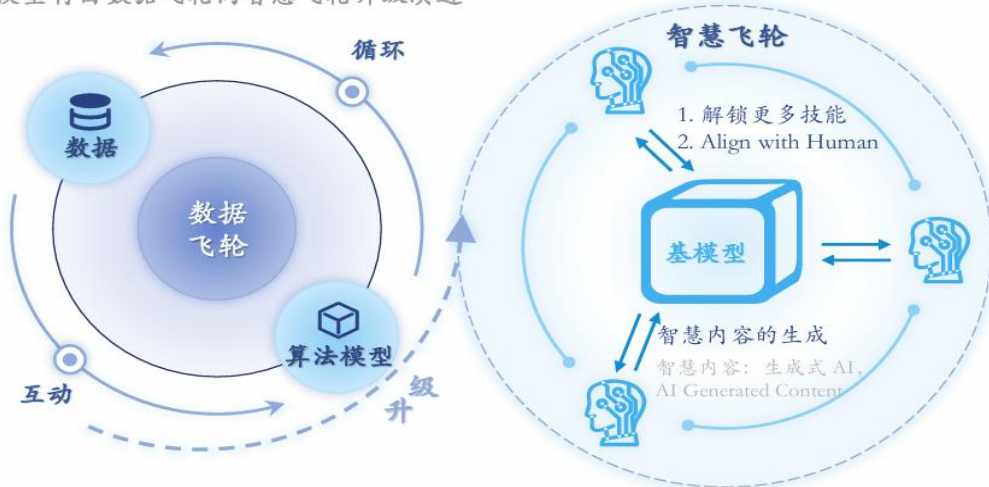
- **数据飞轮**: 现有 AI 体系主要通过前端获取大量数据并进行人工标注, 随后利用更新后的模型反馈到前端以获取更高质量的数据。然而, 这一过程存在研发时间长、成本高的问题。
- **智慧飞轮**: AGI 体系则强调人与模型的互动, 基模型能够不断理解人类意图, 解锁更多技能, 并实现自动化标注。其成本仅为现有 AI 体系的约 1%, 能够推动数据的快速迭代与优化, 从而输出更高质量的内容。

这种从数据驱动到智慧驱动的转变, 将为人工智能的发展带来质的飞跃, 加速 AGI 的实现

二、大模型进入百模大战阶段，发展趋势逐渐清晰

图表 3 大模型将由数据飞轮向智慧飞轮升级演进

大模型将由数据飞轮向智慧飞轮升级演进

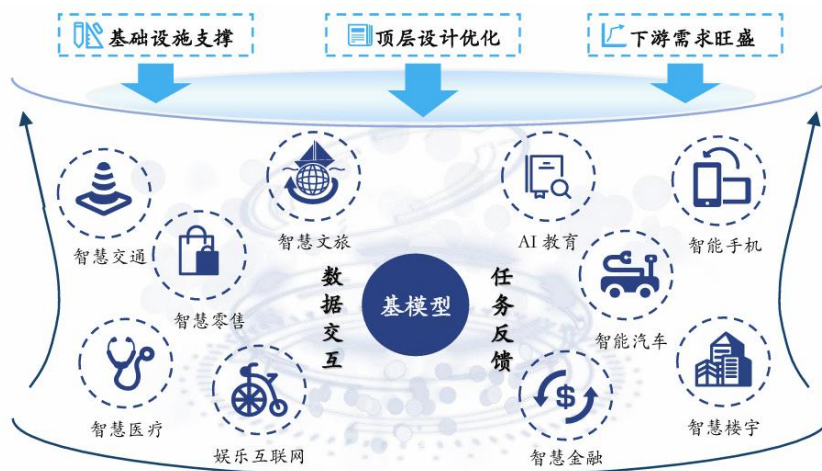


资料来源:《Beyond the imitation game: Quantifying and extrapolating the capabilities of language models》, 商汤, 转引自弗若斯特沙利文《AI 大模型市场研究报告(2023)——迈向通用人工智能, 大模型拉开新时代序幕》

(一) 三轮驱动下, AI 模型迎来良好发展契机

在“基础设施支撑+顶层设计优化+下游需求旺盛”三轮驱动下, AI 大模型迎来了良好的发展契机。通过数据交互和任务反馈, 优秀的大模型能够赋能各行各业开放任务, 满足对未来 AI 应用的期待。展望未来, 大模型“训练基础设施-底层技术-基础应用垂直应用”发展路线逐渐清晰, 随着底层技术逐步革新, 基模型和领域大模型持续完善, 大模型应用边界不断拓宽, 将加速赋能交通、医疗、金融等各个行业和领域, 引发一场以强人工智能和通用人工智能为代表的新一轮智能革命浪潮, 大幅提高生产和生活效率, 带来深刻的经济、社会和产业变革。

图表 4 优秀的大模型能够赋能各行各业开放任务

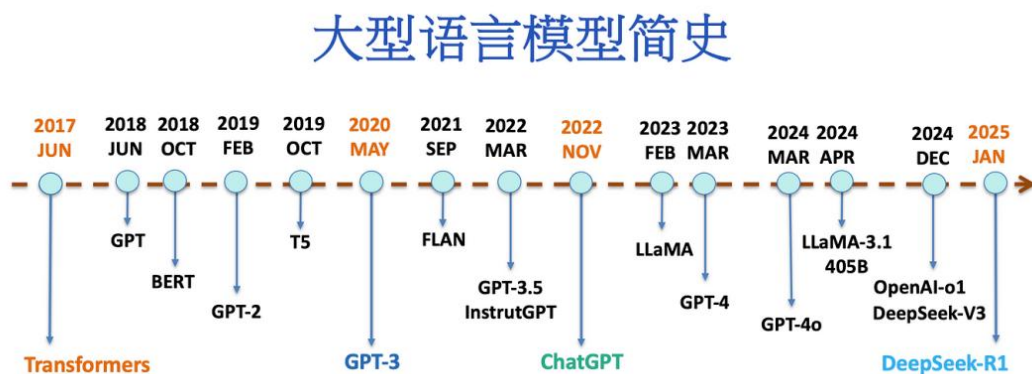


资料来源: 弗若斯特沙利文《AI 大模型市场研究报告(2023)——迈向通用人工智能, 大模型拉开新时代序幕》

AI 基础架构及基础设施快速发展, 助推大模型应用落地。从基础架构的角度来看,

Transformer 是推动 AI 大模型发展的关键基石。它由论文《Attention is All You Need》首次提出，是一种全新的、简洁的网络架构，遵循 Encoder-Decoder 的框架流程来实现任务目标。与传统的循环神经网络（RNN）架构不同，Transformer 完全基于注意力机制构建，摒弃了循环和卷积操作。目前 Transformer 已逐步取代 LSTM 等 RNN 模型，成为 NLP 问题的首选模型，并有逐步统一图像处理等领域的趋势。

图表 5 LLM 研究的发展路径



资料来源：大语言模型简史：从 Transformer（2017）到 DeepSeek-R1（2025）的进化之路，CSDN

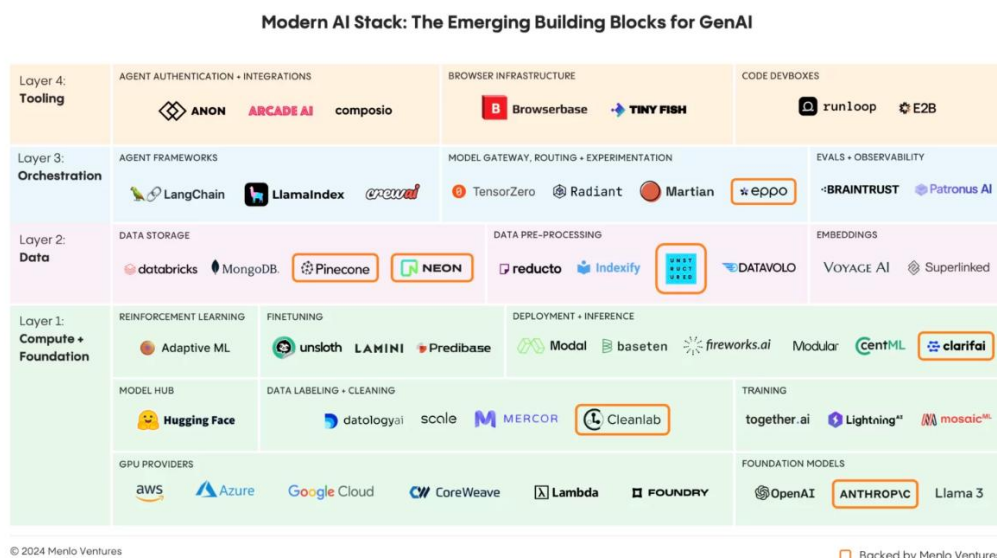
核心技术层协同发展，共同赋能 AI 大模型生态。 AI 大模型的技术架构通常由基础层、技术层、能力层、应用层和终端层五大板块构成。其中，技术层是 AI 大模型的核心，它通过将 AI 技术群与大模型进行融合创新，为各行业提供深度赋能。

基础层： AI 大模型的基础层主要包括硬件基础设施以及数据、算力和算法模型三大核心要素。随着大模型规模的持续扩大，对计算资源的需求也在不断攀升。因此，高性能硬件设备、海量场景数据、强大的算力基础以及不断升级迭代的算法模型，成为推动 AI 大模型发展的关键支撑。深度学习模型的持续升级与迭代，显著提升了 AI 算法的学习能力，而开源模式则为 AI 大模型成为海量应用、网络和服务的基础提供了可能。

技术层： AI 大模型的技术层主要聚焦于模型构建。目前，Transformer 架构在该领域占据主导地位，例如 BERT、GPT 系列等。AI 大模型涵盖自然语言处理（NLP）大模型、计算机视觉（CV）大模型以及多模态大模型等。这些模型普遍采用预训练和微调的策略，先在大量无标注数据上学习语言或图像的基本表示，再针对特定任务进行微调。

能力层、应用层及终端层： 在基础层和技术层的支持下，AI 大模型具备了文字生成、音频处理、图像识别、视频编辑、代码生成、策略制定以及多模态融合等多种能力。这些能力广泛应用于金融、电商、传媒、教育、游戏、医疗、工业、政务等多个领域，为不同类型的用户——包括企业级用户、政府机构用户和大众消费者——提供多样化的产品和服务。

图表 6 AI 大模型的技术架构



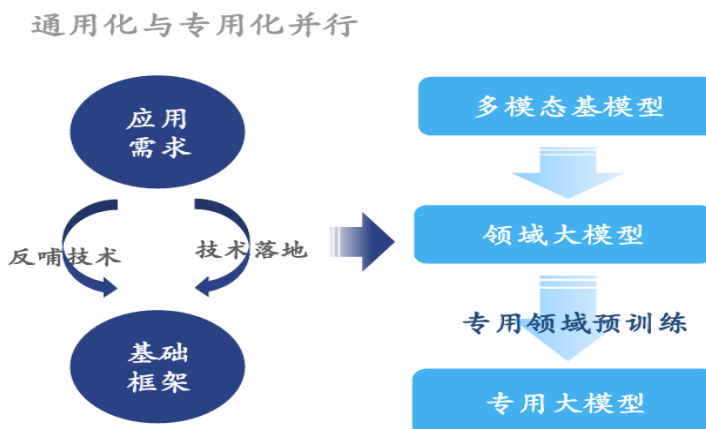
资料来源: Menlo Ventures

(二) AI 大模型产业趋势确定, Mass 将重构商业化生态。

1、大模型将趋于“通用化”与“专用化”并行

通用大模型是一种多模态基础模型,其核心在于统一架构、统一知识表达以及统一任务处理。这种模型通过统一的框架和学习模式,构建适用于多种模态的通用词表,并将所有任务转化为序列到序列的任务形式。以 GPT 系列模型为例,在自然语言处理领域表现出色,能够胜任文本生成、问答、摘要、翻译等多种任务。同样,BERT 模型也广泛应用于自然语言处理的多个任务,如文本分类、命名实体识别和问答等,展现了强大的通用性和适应性。专用大模型通过通用预训练和专用预训练相结合的方式,实现对特定业务场景的优化应用。专用大模型可以分为领域大模型和行业大模型。其中,领域大模型主要针对特定的技术领域,如自然语言处理(NLP)和计算机视觉(CV);而行业大模型则专注于特定的行业需求,如金融、能源等。

图表 7 AI 大模型将通用化与专用化并行

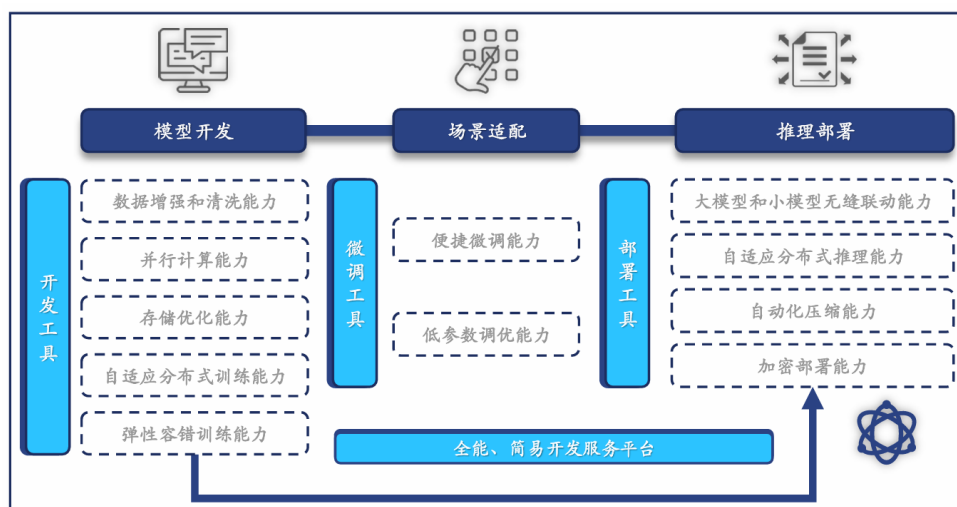


资料来源: 弗若斯特沙利文《AI 大模型市场研究报告(2023)——迈向通用人工智能,大模型拉开新时代序幕》

2、大模型将趋于“平台化”与“简易化”并行

大模型的发展呈现出平台化和简易化两大趋势。平台化主要是指为开发者和研究者提供 AI 模型开发与应用的完整解决方案。在 AI 技术的加持下，GPT 内部插件的用户体验优于直接使用外部原生应用，用户的使用场景也逐渐向 GPT 模型内部迁移。目前，GPT 插件商店内的应用主要由 OpenAI 邀请的第三方开发者构建，涵盖办公协作、电商、旅游等多个领域。随着外部成熟应用的交互圈层不断扩大，GPT 与用户日常生活的结合将更加紧密。参考移动互联网时代从 PC 端到移动端的迁移趋势，主流应用预计将加速向 AI 平台迁移，即迅速适应并布局 GPT 内的 AI 插件，探索 AI 加持下的新应用场景。这些应用可以借助 GPT 的 AI 技术提升用户体验，而 GPT 则通过这些应用吸引更多用户，网络效应进一步强化了这种相互促进的趋势，从而提升 GPT 平台的整体生态价值。

图表 8 大模型平台化与简易化并进



资料来源：弗若斯特沙利文《AI 大模型市场研究报告（2023）——迈向通用人工智能，大模型拉开新时代序幕》

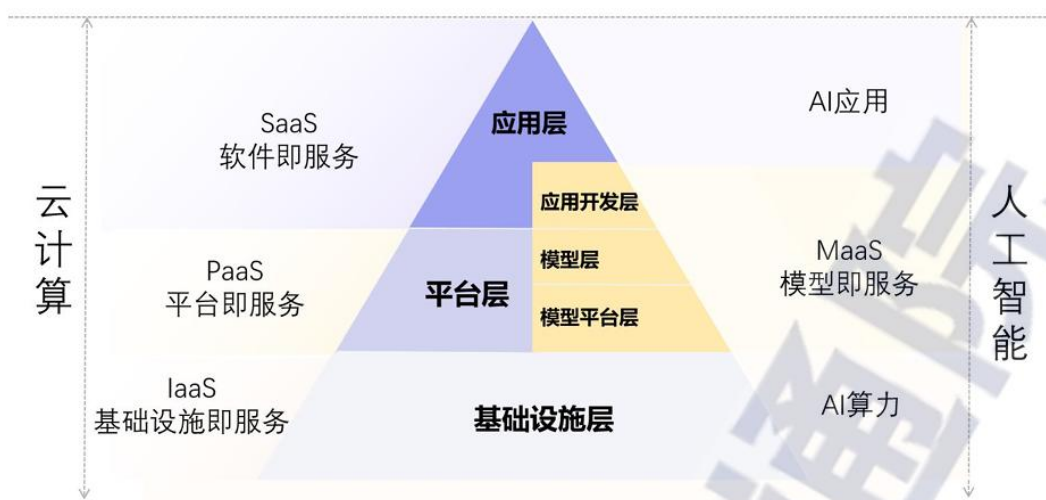
3、大模型发展路线逐渐清晰，MaaS 将重构商业化生态

国内大模型的商业应用大幕正徐徐拉开，厂商们加速战略布局，加大资金和技术投入，持续打磨大模型，努力追赶全球大模型产业化的浪潮。MaaS（Model as a Service，模型即服务）模式能够降低 AI 应用开发的门槛，重构 AI 产业的商业化结构生态，激发新的产业链分工和商业模式不断涌现，有望成为未来大模型的主流商业模式。

MaaS 模式由基础层、中间层和应用层三部分组成：

- 基础层提供多模态基模型以及 CV、NLP、多模态等领域大模型，输出泛化能力，并对外开放大模型的调用接口。
- 中间层即平台层付费使用基础层提供的接口，直接调用基础模型，基于行业特色数据与知识进行精调，开发行业大模型，精准适配 AI 具体应用需求，如汽车领域的自动驾驶、医疗领域的蛋白质解析等，以及企业私有模型。
- 应用层则进行多样化的 SaaS 应用以及新一代智能硬件的开发，未来有望出现“杀手级”应用和现象级产品。

图表 9 MaaS 产业一体化架构



资料来源：中国信息通信研究院 《MaaS 框架与应用研究报告(2024 年)》

MaaS 模式在 B/G 端和 C 端的商业化落地存在差异：

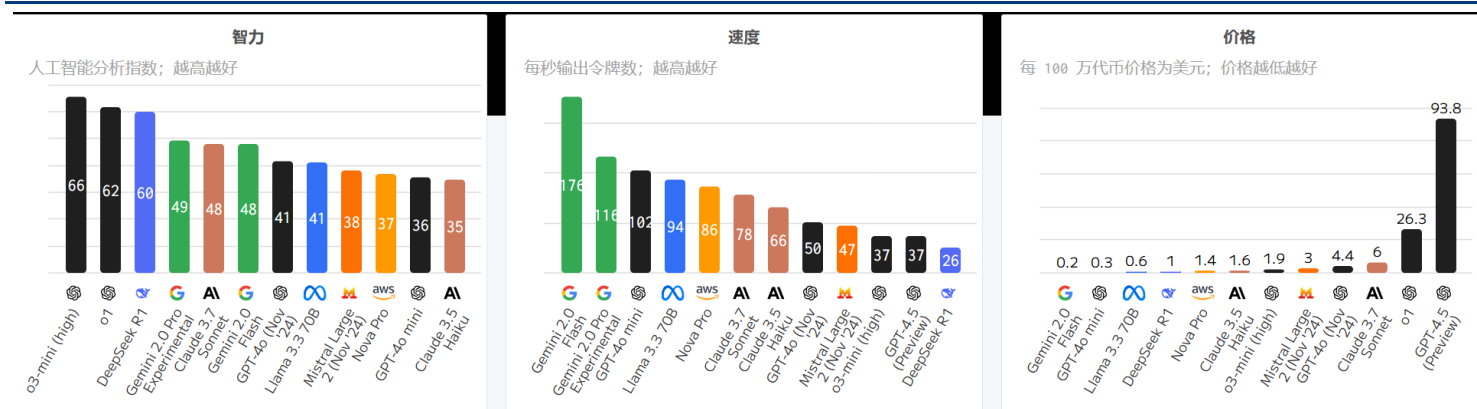
- C 端用户数量庞大，工具使用门槛较低，落地速度以及未来成长速度都较快，“应用商店”等创新商业模式不断涌现。通过“对话 + 插件”的形式，将大模型单点工具接入厂商自有或第三方应用，能够打造新的用户交互界面和入口。未来 MaaS 模式下 C 端的商业化落地形式将不断创新，颠覆传统，拥有广阔的想象空间。
- B/G 端则需要针对行业领域和业务场景进行大量的工程工作，尤其是传统行业，知识获取和积累需要较长时间，即使是同一行业，不同细分场景的痛点也各不相同，导致 AI 大模型的渗透率存在明显差异。应用场景碎片化的特点使得构建低成本、易用、泛化能力较强的能力平台需要较长周期，但 B/G 端客户付费能力较强，未来盈利空间和成长空间较为广阔。

三、海外 AI 大模型主要厂商盘点：发展情况或将分化

(一) 总结：当前各家模型最新版本对比

各模型优势领域不一，根据 artificial analysis 的测评，在智力方面，O3mini (high) 和 o1 表现最佳，得分分别为 66 和 62，在速度方面，Gemini 2.0 Flash 速度最快，每秒输出令牌数达到 176，远超第二名的 Gemini 2.0 Pro。价格方面，Gemini 2.0 最具成本效益，每 100 万代币价格为 0.2 美元。综合智力与性价比，claude3.7 Sonnet Thinking, O3mini (high) 等模型位于具有强吸引力象限中。

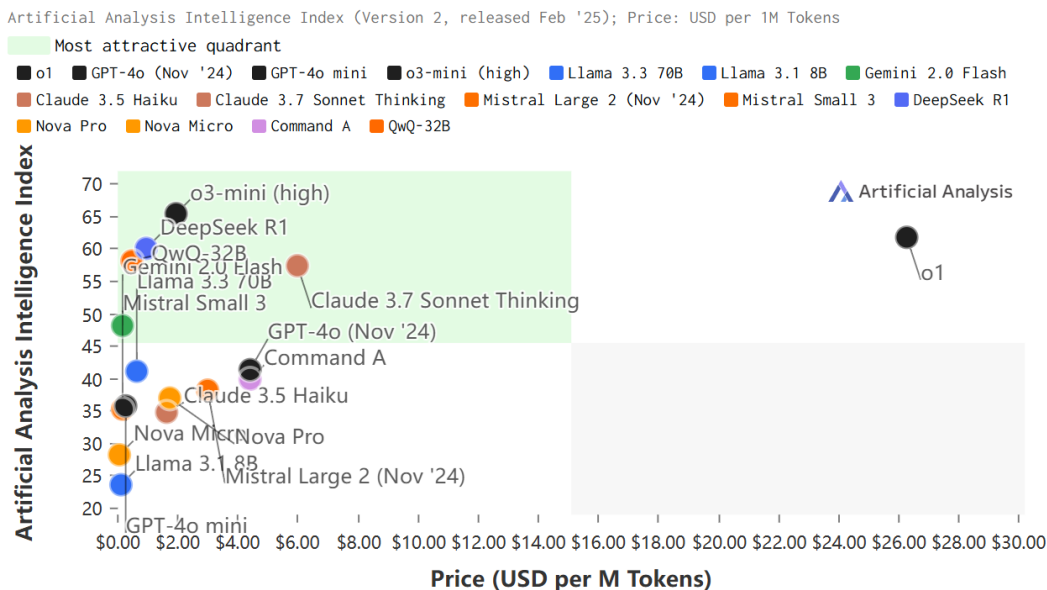
图表 10 各模型当前智力、速度、价格对比（截至 2025 年 3 月）



资料来源: artificial analysis

图表 11 各模型当前智力与价格的对比（截至 2025 年 3 月）

Intelligence vs. Price



资料来源: artificial analysis

(二) Open AI:

OpenAI 的发展始于 2015 年，由埃隆·马斯克、山姆·奥特曼等科技界人士共同创立，旨在推动人工智能技术的前沿研究并确保其安全性和普惠性。2018 年，OpenAI 发布了 GPT-1，这是其首个基于 Transformer 架构的大型语言模型，标志着其在自然语言处理领域的初步探索。2019 年，GPT-2 发布，凭借更高的性能和更强大的语言生成能力，引发了广泛关注。2020 年，GPT-3 的推出进一步巩固了 OpenAI 在 AI 领域的领先地位，其强大的多语言处理能力和广泛的应用场景为后续发展奠定了基础。2023 年，OpenAI 发布了 GPT-4，这一模型在性能、效率和多模态处理能力上实现了质的飞跃，同时开放 AIP，开启了

AI 在消费级市场的广泛应用。2024 年，OpenAI 继续创新，推出了 GPT-4o、o1 系列等新模型，进一步优化了性能和成本。2025 年，OpenAI 计划推出 GPT-4.5 和 GPT-5，预计将在性能和效率上再次突破，引领 AI 技术的未来发展方向。

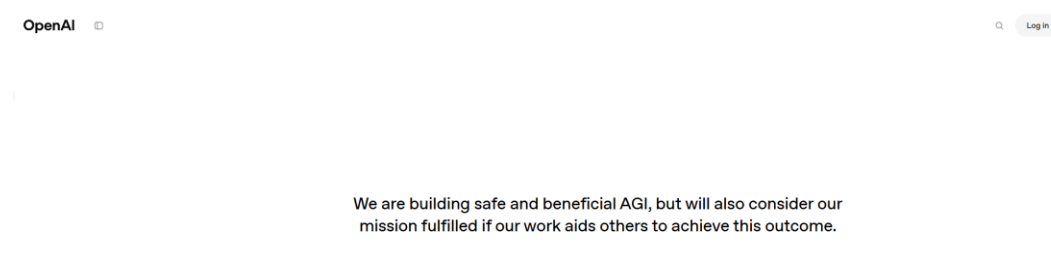
图表 12 Open AI 模型系列与价格

模型名称	发布时间	输入价格 (每 100 万 tokens)	输出价格 (每 100 万 tokens)
GPT-1	2018 年 6 月	-	-
GPT-2	2019 年 2 月	-	-
GPT-3	2020 年 6 月	-	-
GPT-3.5	2022 年	-	-
GPT-4	2023 年 3 月	\$30.00	\$60.00
GPT-4 Turbo	2023 年 11 月	-	-
GPT-4o	2024 年 5 月	\$2.50	\$10.00
GPT-4o mini	2024 年 7 月	\$0.15	\$0.60
o1 系列 (o1、o1-mini)	2024 年 9 月	\$15.00	\$60.00
o3-mini	2025 年 2 月	\$1.10	\$4.40
GPT-4.5	2025 年 2 月	\$75.00	\$150.00

资料来源: Open AI 官网, 华创证券

以实现 AGI 为最终目标, 产品线将持续丰富。OpenAI 的使命是确保通用人工智能 (AGI) 能够造福全人类。AGI 是一种具备与人类同等智慧水平的人工智能系统。它能够在人类从事的所有专业领域中, 执行任何人类能够完成的智力任务, 甚至在许多具有经济价值的工作中超越人类。OpenAI 的 AGI 使命主要聚焦于两个核心目标: 一是开发出能够解决人类水平问题的系统; 二是构建安全且有益的 AGI。

图表 13 Open AI 将 AGI 目标放在其官网上





资料来源: open AI 官网


实现 AGI 路径中, Open AI 将实现以下产品路线图: 从 2023 年 12 月底奥特曼 (OpenAI) 的发言中可以看出, 其未来产品线将涵盖 GPT-5、更先进的语音交互模型、更高效的网页浏览 (网页 Agent) 以及个性化服务 (PC Agent)。目前, 其视频生成模型 (如 Sora) 已经发布。此外, OpenAI 两次发表文生 3D 模型相关的论文, 包括 Shape-E 和 Point-E。目前, OpenAI 的产品线正沿着通用人工智能 (AGI) 的方向发展, 已经发布了涵盖文字、语音、图片和视频的多种模型。未来, 预计其将在 3D 模型、Agent 以及语音交互等领域取得更多突破。

在 2024 年 12 月的“OpenAI 12 Days”活动中，OpenAI 进一步展示了其在 AI 领域的最新进展和产品更新。例如，OpenAI 在活动期间正式发布了 Sora——文本到视频生成模型，用户可以通过文本描述生成 1080p 分辨率的 20 秒视频。此外，ChatGPT 的 AI 搜索引擎也已向所有用户开放，包括免费用户。相关更新同样表明 OpenAI 正在加速其产品线的扩展和功能的增强，进一步推动人工智能技术的普及和应用。

图表 14 Open AI 最新产品 GPT-4.5 与 GPT-5 相关安排



Sam Altman 
@sama

 ...

OPENAI ROADMAP UPDATE FOR GPT-4.5 and GPT-5:

We want to do a better job of sharing our intended roadmap, and a much better job simplifying our product offerings.

We want AI to “just work” for you; we realize how complicated our model and product offerings have gotten.

We hate the model picker as much as you do and want to return to magic unified intelligence.

We will next ship GPT-4.5, the model we called Orion internally, as our last non-chain-of-thought model.

After that, a top goal for us is to unify o-series models and GPT-series models by creating systems that can use all our tools, know when to think for a long time or not, and generally be useful for a very wide range of tasks.

In both ChatGPT and our API, we will release GPT-5 as a system that integrates a lot of our technology, including o3. We will no longer ship o3 as a standalone model.

The free tier of ChatGPT will get unlimited chat access to GPT-5 at the standard intelligence setting (!!), subject to abuse thresholds.

Plus subscribers will be able to run GPT-5 at a higher level of intelligence, and Pro subscribers will be able to run GPT-5 at an even higher level of intelligence. These models will incorporate voice, canvas, search, deep research, and more.

资料来源: Sam Altman 推特

(三) LLaMA:

LLaMA 模型系列自 2023 年首次发布以来，经历了从 LLaMA 1 到 LLaMA 3 的快速迭代与发展。2023 年 2 月，Meta 推出了 LLaMA 1，包括 7B、13B、30B 和 65B 四个参数版本，凭借其开源性和优异性能迅速成为开源社区的热门模型。随后，2023 年 7 月发布的 LLaMA 2 进一步扩充了训练数据量至 2 万亿 tokens，并引入了分组查询注意力机制（GQA），提升了模型性能和推理效率。2024 年 4 月，LLaMA 3 发布，支持 8K 长文本，采用更高效的 tokenizer，并使用超过 15 万亿 tokens 的语料进行训练，性能接近 SOTA。2024 年 7 月，LLaMA 3.1 进一步增强了多语言支持和安全性，巩固了其在开源大模型领域的领先地位。这一系列迭代不仅推动了开源大模型技术的发展，也为全球 AI 研究与应用提供了强大的支持。

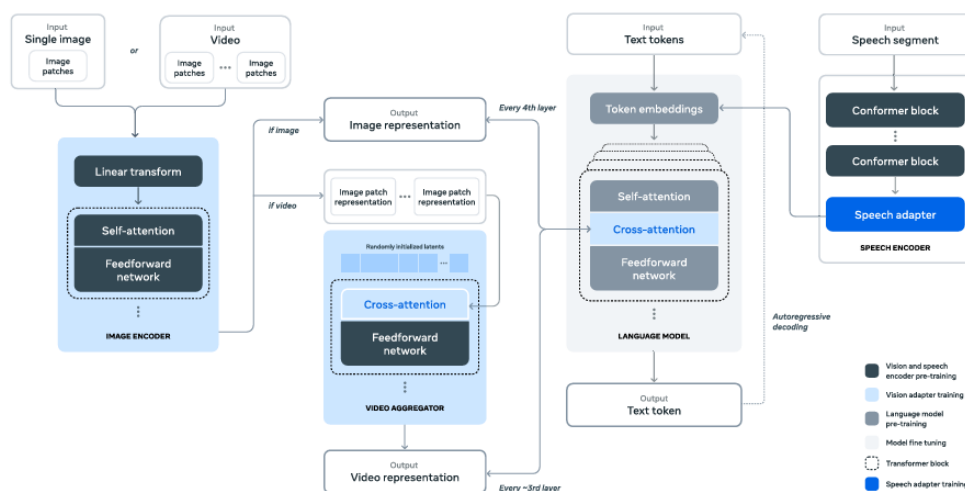
图表 15 LLaMA 各个模型对比

指标	LLaMA 1	LLaMA 2	LLaMA 3	LLaMA 3.2
规模	7B、13B、33B、65B	7B、13B、70B	8B、70B、405B	1B、3B、11B、90B
训练数据量	约 1.0T tokens	约 2.0T tokens	约 15T tokens	约 15T tokens
主要特点	引入 RMSNorm、RoPE、SwiGLU	引入分组查询注意力 (GQA)，优化对话微调	上下文长度扩展至 8192，采用 TikToken 分词器	上下文长度扩展至 128K，支持多模态
性能表现	基础性能较好	在推理、对话能力等方面优于 LLaMA 1	性能显著提升，推理速度更快，生成答案更长	多模态能力增强，适合复杂任务
适用场景	通用文本生成	对话、编程辅助	复杂推理、多语言支持	多模态任务、边缘设备
上下文窗口	2048 tokens	4096 tokens	8192 tokens	128K tokens
多模态支持	不支持	不支持	不支持	支持 (图像与文本任务)
开源许可	开源	开源	开源	开源
部署成本	部署成本低	部署成本适中	部署成本适中	多模态版本需要更多硬件资源

资料来源: CSDN, 华创证券整理

LLaMA 模型箭指多模态，未来有望成为多模态领域模型先锋。 LLaMA 的功能包括多语言对话支持 (涵盖英语、德语、法语、印地语等 8 种语言)，强大的文本生成和代码生成能力，以及长上下文窗口 (如 Llama 3.3 支持 128K tokens)。此外，LLaMA 系列还具备高效的 Transformer 架构，结合监督微调 (SFT) 和基于人类反馈的强化学习 (RLHF) 来提高模型的安全性和实用性。

图表 16 Llama 3.2 使用的多模态 LLM 方法的插图



资料来源: LlamaTeam, AI@Meta 《The Llama 3 Herd of Models》

后续发展路线: LLaMA 的目标是通过持续优化核心能力（如推理、编码）和扩展多模态边界（如加入语音功能和视频理解能力），打造“一模多能”的 AI。未来，LLaMA 有望在多模态问答系统、智能助手、教育、医疗等领域发挥更大作用。LLaMA 模型的未来发展将着重于以下几个方向：首先，模型将继续优化其多模态能力，特别是在视觉理解和生成方面，例如通过 Diffusion 技术进一步提升视觉任务的表现。其次，LLaMA 可能会进一步拓展其在特定领域的应用，如医疗领域，通过开源多语言医疗语料库来支持全球医疗 AI 的发展。此外，LLaMA 模型在性能提升方面也将持续发力，例如 LLaMA 3.3 已经展现出用较少参数达到更高性能的潜力，未来可能会继续探索更高效的架构和训练方法以降低成本并提高效率。最后，LLaMA 模型可能会在语音合成和视频理解等新领域进行探索，以满足更多元化的市场需求。

（四）Anthropic:

Anthropic 自 2021 年成立以来，迅速成为全球 AI 领域的重要参与者，专注于开发安全、可靠且可解释的 AI 系统。2023 年 3 月，Anthropic 正式发布其初代模型 Claude 1，采用独特的“宪法 AI”（Constitutional AI）技术，通过 AI 自主评估和反馈训练结果，避免了人类反馈系统的不透明性。2023 年，Claude 2 发布，性能大幅提升，支持更长文本响应，并在编程、数学和推理能力方面表现出色。同年，Claude 2.1 进一步扩展了上下文窗口至 200K，使其在长文本处理和企业级应用中更具优势。2024 年，Anthropic 推出了 Claude 3 系列，包括 Claude 3 Haiku、Sonnet 和 Opus，这些模型在多模态支持、语言能力和安全性方面取得了显著进步，成为全球首个在特定基准测试中超越了 GPT-4 模型。紧接着，Claude 3.5 系列发布，进一步优化了实时用户交流和数据提取任务的性能，并在多模态视觉处理方面取得突破。进入 2025 年，Anthropic 继于 2 月发布了 Claude 3.7 Sonnet、Claude 3.5 系列和 Claude 3.7 Sonnet 在多模态视觉处理方面取得了显著进步等，特别是在科学示意图处理方面展现出强大的应用潜力。通过持续的技术创新，Anthropic 的模型不仅在性能上不断突破，还在安全性、可靠性和可解释性方面树立了行业标杆，逐渐成为 OpenAI 的有力竞争对手。

图表 17 Anthropic 各个模型对比

	Claude 1	Claude 2	Claude 2.1	Claude 3	Claude 3.5
发布时间	2023 年 3 月	2023 年 7 月	2023 年 11 月	2024 年 3 月	2024 年 10 月
上下文窗口	9K token	100K token	200K token	200K token (特定用途可达 1M token)	200K token
多模态能力	不支持	不支持	不支持	支持（图片、图表等输入）	支持（新增计算机操作功能）
性能与能力	基础语言生成和对话	编码、数学、推理能力提升	对话优化，内容创建功能增强	多模态任务处理，复杂推理	高级编码、复杂推理，支持计算机操作

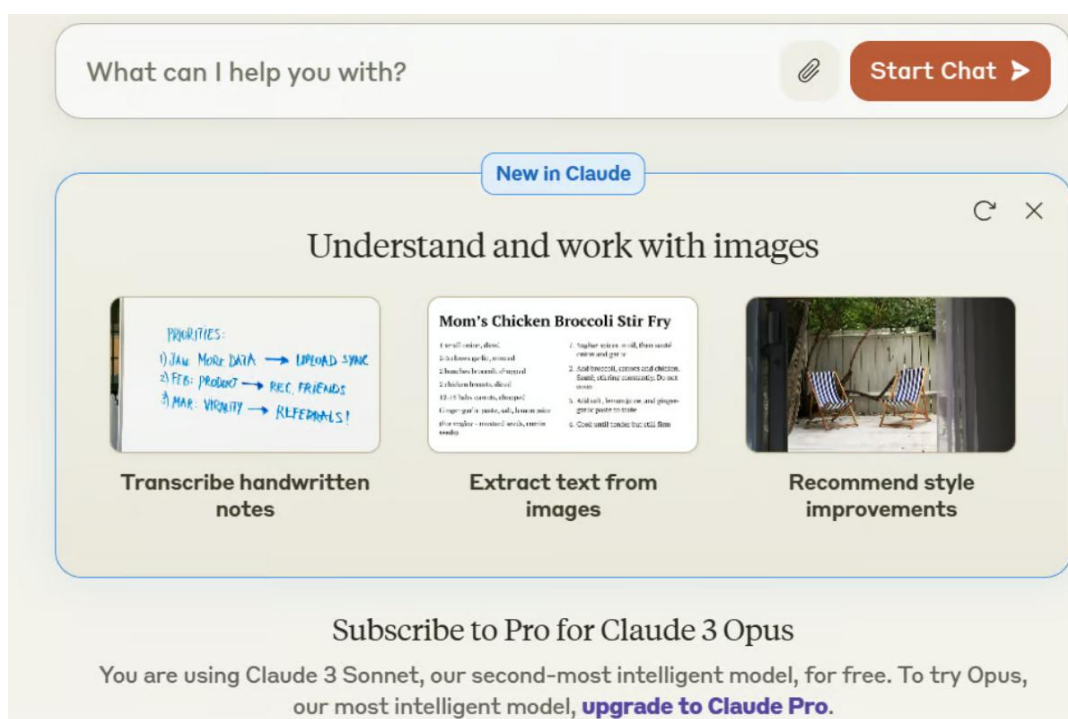
应用场景	语言生成、对话	编程辅助、文档处理	对话优化、内容创作	多模态分析、企业级应用	高级编码任务、计算机操作
安全与可靠性	安全性较低	安全性增强	安全性高，拒绝回答率较高	安全性高，拒绝回答率降低	安全性高，保持与 Claude 3 水平

资料来源：机器之心，Anthropic 官网等，华创证券

Claude 作为 Anthropic 的核心产品，已经在企业级应用、编程与开发、数据分析、多语言支持以及教育等领域展现出强大的实力。它不仅为企业提供高效、个性化的解决方案，还在编程任务中表现出色，特别是在代码生成和调试方面，显著提升了开发效率。同时，Claude 凭借其长上下文窗口和多语言能力，成为数据分析和跨文化交流中的得力助手。在教育领域，Claude 能够提供个性化的学习体验，支持内容生成，助力教育辅导。

展望未来，Anthropic 正致力于将 Claude 打造成一个更智能、更灵活的 AI 助手。Claude 4 将推出“混合 AI”模型，通过滑动条灵活调整处理 token 的数量，以适应从简单到复杂的各种任务。此外，Claude 还将具备网络访问功能，能够实时获取和处理网络信息，从而在动态环境中提供更实用的建议。记忆功能也将得到增强，使 Claude 能够跨项目甚至长期记忆用户交互信息，这对于实现虚拟协作伙伴的愿景至关重要。虽然语音交互和多模态能力的拓展目前优先级较低，但 Anthropic 也在积极探索这些方向，以进一步提升用户体验。

图表 18 Claude Chat 形式



资料来源：Claude AI 官网

(五) xAI:

xAI 自 2023 年 7 月由埃隆·马斯克宣布成立以来，迅速在人工智能领域崭露头角。2023 年 11 月，xAI 发布了其首款 AI 大模型产品——Grok-1，该模型基于混合专家系统(MoE)

架构，参数量达 3140 亿，支持长文本推理和实时知识集成。2024 年 3 月，xAI 宣布将 Grok-1 开源，成为当时参数量最大的开源大语言模型。随后，xAI 在 2024 年 3 月推出了 Grok-1.5，显著提升了数学推理和编码能力，并支持 128K 上下文处理。2024 年 8 月，xAI 发布了 Grok-2 及其轻量级版本 Grok-2 mini，进一步优化了性能并扩展了应用场景。2025 年 2 月，xAI 推出了 Grok-3，其推理能力和自然语言处理能力显著提升，并引入了“Deep Search”工具，能够整合网络搜索和 X 平台上的实时信息。

图表 19 XAI 发展时间线

时间	模型版本	主要特点
2023 年 7 月	xAI 成立，目标是打造高质量的大语言模型。	初始团队 12 人，背景强大；与 X 平台和特斯拉紧密合作。
2023 年 11 月	发布首款大语言模型 Grok-1，参数量 3140 亿。	混合专家系统（MoE）架构；支持长文本推理和实时知识集成。
2024 年 3 月	Grok-1 开源。	参数量 3140 亿，成为当时参数量最大的开源大语言模型。
2024 年 3 月	发布 Grok-1.5，显著提升数学推理和编码能力。	数学推理得分 50.6%，编码能力得分 74.1%；支持 128K 上下文处理。
2024 年 5 月	完成 60 亿美元 B 轮融资，投后估值达 240 亿美元。	资金用于算力基础设施建设和模型研发。
2024 年 6 月	推出超级计算机 Colossus，搭载 10 万块 NVIDIA H100 GPU。	计划扩展至 15 万块 H100/H200，成为全球最强 AI 训练系统。
2024 年 8 月	发布 Grok-2 及其轻量级版本 Grok-2 mini。	性能跻身全球前五；引入少量真实世界数据，提升处理真实场景的能力。
2025 年 2 月	发布 Grok-3，引入“思维链”推理能力。	性能显著提升，推理能力增强；目标是成为“地球上最聪明的人工智能”。

资料来源：xAI 官网、36 氪、机器之心、智东西等，华创证券

xAI 旗下的 Grok 系列模型在自然语言处理、复杂推理、代码生成以及内容创作等领域展现出强大实力，尤其在实时知识集成和可解释性方面具有显著优势。未来，xAI 将重点拓展多模态能力，集成图像和视频处理功能，进一步提升模型在视觉任务中的表现。同时，Grok 系列将深化与特斯拉和 SpaceX 的合作，探索在自动驾驶和人形机器人领域的应用，借助车辆传感器数据和航天技术资料，推动 AI 在更多场景的落地。此外，xAI 还将继续优化模型的推理能力和实时功能，进一步提升其在动态环境中的实用性，致力于打造“地球上最聪明的人工智能”。

图表 20 Grok3 功能实例

用户 你能把它翻译成 Python 代码吗?

格罗克 当然! 您提供的流程图描述了一个简单的猜谜游戏, 用户必须猜测它。下面是表示流程图中逻辑的 Python 代码:

```

1 import random
2
3 def guess_number(): # Generate a random number
4
5     target = random.randint(1, 10)
6
7     while True:
8         guess = int(input("Guess a number between 1 and 10: "))
9
10        if guess == target:
11            print("You won!")
12            break
13        else:
14            print("Wrong guess, try again!")
15
16 if __name__ == "__main__":
17     guess_number()

```

此代码将运行流程图中所述的猜谜游戏。系统将提示用户输入数字, 直到他们猜出正确的数字。

资料来源: XAI 官网

(六) Gemini:

Gemini 模型是谷歌推出的多模态 AI 模型, 旨在通过强大的多模态能力 (包括文本、图像、音频、视频和代码) 以及高效的推理能力, 为用户提供更智能的交互体验。其发展历程体现了谷歌在 AI 领域的持续创新和优化。2023 年 12 月, 谷歌首次发布 Gemini 1.0, 标志着其在多模态 AI 领域的重大突破。随后, 2024 年 2 月, 谷歌推出了 Gemini 1.5 系列, 显著提升了模型在长文本处理、多模态任务以及推理能力方面的表现。2024 年 5 月, 谷歌进一步优化了 Gemini Pro 1.5, 增加了对超长上下文的支持, 使其能够处理更复杂的任务。2024 年 8 月, Gemini Pro 1.5 在多个基准测试中超越了 ChatGPT-4, 展现了其在多模态处理和性能优化方面的优势。

图表 21 Gemini 各个模型对比

时间	事件	关键要素
2023 年 12 月	发布 Gemini 1.0, 谷歌称其为史上最强大、最通用的模型。	多模态能力 (文本、图像、音频、视频、代码); 优化不同尺寸版本 (Ultra、Pro、Nano)。
2024 年 2 月	发布 Gemini 1.5 系列。	性能显著提升, 支持 100 万 token 超长上下文; 推理能力增强。
2024 年 5 月	发布 Gemini Pro 1.5。	支持超长上下文 (100 万 token), 优化推理效率。
2024 年 8 月	Gemini Pro 1.5 在多个基准测试中超越 ChatGPT-4。	多模态处理能力优化; 性能提升。
2024 年 9 月	Gemini 模型在多语言任务上表现突出。	跨语言能力增强, 支持多种语言。

资料来源: Gemini 官网、CSDN、新浪财经等, 华创证券

Gemini 模型作为谷歌 DeepMind 推出的多模态 AI 系统，已经在多模态融合、长上下文处理和智能体应用等方面取得了显著进展。谷歌明确表示，Gemini 2.0 是为“代理时代”设计的模型，具备自主理解、规划和执行任务的能力。Gemini 将继续深化其多模态能力，支持更复杂的模态间交互，例如实现图像、文本、音频和视频之间的无缝转换。此外，Gemini 将通过引入强化学习等技术，具备自我学习和改进的能力，进一步提升其逻辑推理和因果推断能力。谷歌计划将 Gemini 集成到更多产品和服务中，如 Gmail、Docs 和 Sheets 等，以增强用户体验。同时，Gemini 还将优化其在边缘设备上的部署能力，使其能够在资源受限的环境中高效运行。随着技术的不断进步，Gemini 有望在智能助手、创意内容生成、教育、医疗诊断、多语言交流和科学研究等领域发挥更大的作用。未来，Gemini 还将更加注重安全性和可解释性，确保其在广泛应用中的可靠性和公平性。

图表 22 Gemini2.0 意图切入 Agent 时代



资料来源：Gemini 官网

四、投资建议

海外 AI 巨头如 OpenAI、Meta 等在模型进化中普遍依赖“算力堆叠”策略，即通过扩大模型参数量（如 GPT-4 的万亿级参数）、增加 GPU 集群规模（如 xAI 的 Grok-3 使用 20 万张 GPU）以及高昂的训练成本来追求性能突破。在模型不断更新与模型平权的催化下，应用侧有望迎来更广泛的创新与普及。随着技术的不断优化和成本的降低，更多企业和开发者能够接入高性能的 AI 模型，推动各行业智能化转型。例如，中小企业可以利用 AI 提升运营效率，开发者能基于模型快速搭建个性化应用，为用户带来更丰富、更智能的体验。

模型不断迭代，有望利好应用侧快速发展，建议关注以下标的：1)办公：MSFT；2)金融：AFRM、UPST、SOFI、LMND；3)工业：CDNS、SNPS；4)医疗：TEM、RXRX、DOCS、HIMS、VEEV；5)创意：ADOBE；6)教育：DUOL；7)电商：SHOP、SEA；8)企业服务软件：SAP、CRM、NOW、TEAM、ORCL、WDAY、HUBS、FRSH、CXM；9)安全：PANW、CRWD、NET、FTNT；10)数据基建：ESTC、SNOW、MDB；11)云服务提供商：AMZN、META、MSFT、GOOG、ORCL；12)广告：APP、U、ZETA 等。

五、风险提示

竞争加剧，商业化不及预期，技术更新较快。

团队介绍

首席研究员、组长：吴鸣远

上海交通大学硕士，曾任职于东方证券、兴业证券研究所，所在团队于 2020—2022 年连续三年获得新财富最佳分析师第三名，2023 年加入华创证券研究所。

分析师：祝小茜

中央财经大学经济学硕士。曾任职于信达证券。2024 年加入华创证券研究所。

研究员：胡昕安

工学硕士，曾任职于海康威视，2023 年加入华创证券研究所。

助理研究员：周志浩

西安交通大学金融工程学士，克拉克大学金融学硕士，曾任职于众安保险权益投资部，2024 年加入华创证券研究所。

助理研究员：张宇凡

香港大学会计学硕士。2023 年加入华创证券研究所。

助理研究员：周楚薇

香港中文大学经济学硕士。2024 年加入华创证券研究所。

助理研究员：杨玖祎

香港中文大学硕士。2024 年加入华创证券研究所。

华创证券机构销售通讯录

地区	姓名	职务	办公电话	企业邮箱
北京机构销售部	张昱洁	副总经理、北京机构销售总监	010-63214682	zhangyujie@hcyjs.com
	张菲菲	北京机构副总监	010-63214682	zhangfeifei@hcyjs.com
	张婷	华北机构销售副总监		zhangting3@hcyjs.com
	刘懿	副总监	010-63214682	liuyi@hcyjs.com
	侯春钰	资深销售经理	010-63214682	houchunyu@hcyjs.com
	顾翎蓝	资深销售经理	010-63214682	gulinglan@hcyjs.com
	蔡依林	资深销售经理	010-66500808	caiyilin@hcyjs.com
	刘颖	资深销售经理	010-66500821	liuying5@hcyjs.com
	阎星宇	销售经理		yanxingyu@hcyjs.com
	张效源	销售经理		zhangxiaoyuan@hcyjs.com
	车一哲	销售经理		cheyizhe@hcyjs.com
	郑珺丹	销售经理		zhengjundan@hcyjs.com
深圳机构销售部	张娟	副总经理、深圳机构销售总监	0755-82828570	zhangjuan@hcyjs.com
	汪丽燕	高级销售经理	0755-83715428	wangliyan@hcyjs.com
	张嘉慧	高级销售经理	0755-82756804	zhangjiahui1@hcyjs.com
	王春丽	高级销售经理	0755-82871425	wangchunli@hcyjs.com
	王越	高级销售经理		wangyue5@hcyjs.com
	温雅迪	销售经理		wenyadi@hcyjs.com
上海机构销售部	许彩霞	总经理助理、上海机构销售总监	021-20572536	xucaixia@hcyjs.com
	官逸超	上海机构销售副总监	021-20572555	guanyichao@hcyjs.com
	黄畅	上海机构销售副总监	021-20572257-2552	huangchang@hcyjs.com
	吴俊	资深销售经理	021-20572506	wujun1@hcyjs.com
	张佳妮	资深销售经理	021-20572585	zhangjian1@hcyjs.com
	郭静怡	高级销售经理		guojingyi@hcyjs.com
	蒋瑜	高级销售经理	021-20572509	jiangyu@hcyjs.com
	吴菲阳	高级销售经理		wufeiyang@hcyjs.com
	朱涨雨	高级销售经理	021-20572573	zhuzhangyu@hcyjs.com
	李凯月	高级销售经理		likaiyue@hcyjs.com
	张豫蜀	销售经理	15301633144	zhangyushu@hcyjs.com
	张玉恒	销售经理		zhangyuheng@hcyjs.com
张晨奂	销售经理		zhangchenhuan@hcyjs.com	
广州机构销售部	段佳音	广州机构销售总监	0755-82756805	duanjiayin@hcyjs.com
	周玮	销售经理		zhouwei@hcyjs.com
	王世韬	销售经理		wangshitao1@hcyjs.com
私募销售组	潘亚琪	总监	021-20572559	panyaqi@hcyjs.com
	汪子阳	副总监	021-20572559	wangziyang@hcyjs.com
	江赛专	副总监	0755-82756805	jiangsaizhuan@hcyjs.com
	汪戈	高级销售经理	021-20572559	wangge@hcyjs.com
	宋丹筠	销售经理	021-25072549	songdanyu@hcyjs.com
	赵毅	销售经理		zhaoyi@hcyjs.com

华创行业公司投资评级体系

基准指数说明:

A 股市场基准为沪深 300 指数, 香港市场基准为恒生指数, 美国市场基准为标普 500/纳斯达克指数。

公司投资评级说明:

强推: 预期未来 6 个月内超越基准指数 20% 以上;
推荐: 预期未来 6 个月内超越基准指数 10% - 20%;
中性: 预期未来 6 个月内相对基准指数变动幅度在 -10% - 10% 之间;
回避: 预期未来 6 个月内相对基准指数跌幅在 10% - 20% 之间。

行业投资评级说明:

推荐: 预期未来 3-6 个月内该行业指数涨幅超过基准指数 5% 以上;
中性: 预期未来 3-6 个月内该行业指数变动幅度相对基准指数 -5% - 5%;
回避: 预期未来 3-6 个月内该行业指数跌幅超过基准指数 5% 以上。

分析师声明

每位负责撰写本研究报告全部或部分内容的分析师在此作以下声明:

分析师在本报告中对所提及的证券或发行人发表的任何建议和观点均准确地反映了其个人对该证券或发行人的看法和判断; 分析师对任何其他券商发布的所有可能存在雷同的研究报告不负有任何直接或者间接的可能责任。

免责声明

本报告仅供华创证券有限责任公司(以下简称“本公司”)的客户使用。本公司不会因接收人收到本报告而视其为客户。

本报告所载资料的来源被认为是可靠的, 但本公司不保证其准确性或完整性。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断。在不同时期, 本公司可发出与本报告所载资料、意见及推测不一致的报告。本公司在知晓范围内履行披露义务。

报告中的内容和意见仅供参考, 并不构成本公司对具体证券买卖的出价或询价。本报告所载信息不构成对所涉及证券的个人投资建议, 也未考虑到个别客户特殊的投资目标、财务状况或需求。客户应考虑本报告中的任何意见或建议是否符合其特定状况, 自主作出投资决策并自行承担投资风险, 任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。本报告中提及的投资价格和价值以及这些投资带来的预期收入可能会波动。

本报告版权仅为本公司所有, 本公司对本报告保留一切权利。未经本公司事先书面许可, 任何机构和个人不得以任何形式翻版、复制、发表、转发或引用本报告的任何部分。如征得本公司许可进行引用、刊发的, 需在允许的范围内使用, 并注明出处为“华创证券研究”, 且不得对本报告进行任何有悖原意的引用、删节和修改。

证券市场是一个风险无时不在的市场, 请您务必对盈亏风险有清醒的认识, 认真考虑是否进行证券交易。市场有风险, 投资需谨慎。

华创证券研究所

北京总部	广深分部	上海分部
地址: 北京市西城区锦什坊街 26 号 恒奥中心 C 座 3A	地址: 深圳市福田区香梅路 1061 号 中投国 际商务中心 A 座 19 楼	地址: 上海市浦东新区花园石桥路 33 号 花旗大厦 12 层
邮编: 100033	邮编: 518034	邮编: 200120
传真: 010-66500801	传真: 0755-82027731	传真: 021-20572500
会议室: 010-66500900	会议室: 0755-82828562	会议室: 021-20572522