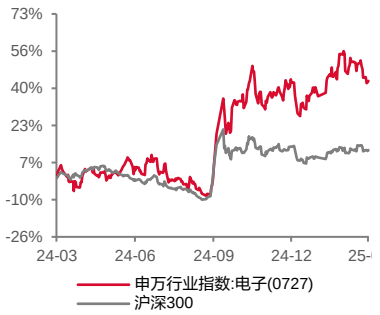
2025年03月27日

AI大模型竞赛方兴未艾，OpenAI与DeepSeek引领行业生态重构

——半导体行业深度报告（十二）



相关研究

1. 乐鑫科技(688018): AIOT次新品显著放量，产品矩阵拓展布局新市场——公司深度报告
2. 海外科技股2024Q4业绩持续回暖，DeepSeek大模型引燃AI云与端热情——半导体行业2月份月报
3. AI大模型风起云涌，半导体与光模块长期受益——半导体行业深度报告（十）

投资要点：

- **2024年全球AI市场规模有望达到6.16万亿美元，同比增长30.1%，2027年有望扩张至11.64万亿美元，CAGR为23.65%。**AI概念于1956年达特茅斯会议首次提出，是一种模拟人类智能的技术，按照智能程度划分，主要分为狭义人工智能、通用人工智能和超级人工智能，目前通用人工智能还处于理论阶段。AI具有算力、算法、数据三大要素，算法决定了AI如何处理数据和解决问题，数据决定了算法是否能得到有效的训练和优化，算力提供了执行算法和处理数据所需的计算资源。从AI产业链看，整体涵盖基础设施层、模型层、平台层、应用层及服务层多个环节，基础设施层主要包括与芯片、计算、存储、网络、软件、连接与通信等多个上游领域，模型层可分为通用大模型、行业大模型等。根据Frost & Sullivan，自2020年起，全球AI市场规模以高于20%的同比增速呈现迅猛增长的态势，从2019年的1.91万亿美元有望扩张至2024年的6.16万亿美元，同比增速逐年上升，整体市场有望在2027年扩张至11.64万亿美元，体现出全球AI行业井喷式的发展速度。

- **未来五年全球大模型行业市场规模的CAGR有望达到36.23%，AI Agent或将成为继API调动和模型推理部署后新的商业化形式，大模型行业竞争格局也将逐步收敛至头部厂商。**AI大模型作为AI产业链中的核心环节，经过大规模数据和强大的计算能力训练，通常具有高度的通用性和泛化能力，可以应用于自然语言处理、图像识别、语音识别等领域。深度学习是机器学习的重要分支，主要涵盖预训练、后训练、推理等阶段，Scaling Law是预训练阶段驱动模型进步的第一性原理，“涌现”现象进一步证明了模型参数量、数据、计算量大小对于模型性能提高的重要性。大模型的商业化落地形式主要包括通过API调用收费以及定制化的模型推理部署，前者市场价格竞争较为激烈，后者是国内的核心业务模式，尤其是云端部署，从金额来看，在政务、教科领域落地的大模型项目较多。随着AI Agent发展，未来基于结果和价值创造的商业模式有望逐步落地。从行业供给格局看，大模型竞争日趋白热，模型之间差距逐步缩小，护城河不清晰，厂商需要持续大量投入，海内外竞争格局都将逐步收敛至头部厂商，部分规模较小的模型厂商或聚焦于垂直化的细分场景。

- **GPT与OpenAI o1系列模型分别验证了算力投入在训练侧和推理侧的重要性，而DeepSeek通过创新性的训练方法和架构实现了较低的模型训练成本，在未来大模型不断创新迭代的背景下，性能提升与成本下行或成为两条重要主线。**基于GPT-3.5的ChatGPT的发布推动了AI技术的普及和AI产业的变革，是人工智能的重要里程碑之一。ChatGPT的创始人OpenAI自成立起先后发布了GPT系列模型和以OpenAI o1、o3为代表的深度推理模型，GPT系列模型注重预训练阶段的Scaling Law，整体来说更适合解决通识类知识，目前已经迭代至GPT-4系列，从最初单一的文本模态迭代成为多模态大模型，参数规模、训练数据、上下文窗口大小相比前代呈指数级增长，模型性能相应也有显著提升。OpenAI o1模型引入了思维链，证明了推理侧的算力资源投入同样重要，Scaling Law在推理阶段或同样适用，未来，GPT系列与o1为代表的深度推理系列模型或将互相补充。近期，DeepSeek大模型的发布进一步拉动了AI热潮，DeepSeek-R1发布后仅用七天用户增长一亿，海内外头部厂商纷纷入场布局。DeepSeek-V3性能对齐海外领军闭源模型，但依靠引入MLA机制和创新性的DeepSeekMoE架构实现了远低于行业平均的训练成本和定价。DeepSeek-R1在后训练阶段大规模使用了强化学习技术而不依赖监督微调，性能对齐OpenAI-o1正式版，同时证明了蒸馏技术能够将大模型的推理能力转移到更小的模型上，提升它们的表现。

- **投资建议：**AI大模型时代下，AI算力需求高速扩张，从而驱动AI芯片、存储、服务器、光模块、PCB等上游产业链半导体板块的需求快速增长，相关标的有望长期受益。（1）云端AI芯片板块关注寒武纪、海光信息、龙芯中科等；（2）端侧AI芯片板块关注恒玄科技、乐鑫科技、中科蓝讯、晶晨股份、瑞芯微、全志科技、炬芯科技、国科微等；（3）存储板块关注兆易创新、佰维存储、德明利、江波龙、澜起科技、东芯股份、聚辰股份、普冉股份、北京君正等；（4）光模块、光器件、光芯片板块关注中际旭创、天孚通信、新易盛、光迅科技、源杰科技等；（5）PCB板块关注鹏鼎控股、胜宏科技、深南电路、沪电股份、东山精密、景旺电子等；（6）服务器（含液冷）板块关注浪潮信息、工业富联、紫光股份、中石科技、光迅科技、川环科技、国芯科技等；（7）电源板块关注麦格米特、光宝科技、中国长城、新雷能、欧陆通等。
- **风险提示：（1）AI需求不及预期风险；（2）行业竞争过度风险；（3）国际贸易政策的变化风险。**

正文目录

1. AI 市场高速扩张，有望引领新一代工业革命 6

1.1. AI 推动生产变革，行业步入蓬勃发展期 6
1.2. AI 产业链涵盖基础设施到应用落地多个环节 8

2. AI 大模型是 AI 变革的重要环节之一 10

2.1. “Scaling Law”驱动大模型不断进步 10
2.2. 大模型商业化模式有望通过 AI Agent 实现转型 14
2.3. 大模型竞争日趋白热，未来玩家格局或将逐步收敛 18

3. 大模型创新迭代，性能提升与成本下行或成为两条主线 23

3.1. GPT 与 o1 验证了训练侧和推理侧算力投入的重要性 23
3.2. DeepSeek 创新性地实现了成本更低的训练 26
3.3. AI 大模型产业链半导体相关重点厂商梳理 32

4. 投资建议与风险提示 37

4.1. 投资建议 37
4.2. 风险提示 37

图表目录

图 1 人工智能发展历程 6
图 2 按智能程度划分的三类人工智能 7
图 3 Gen AI 的工作原理 7
图 4 Gen AI 在各领域的应用效果 7
图 5 AI 的三大要素 8
图 6 AI 算力的相关常用名词及其含义 8
图 7 AI 产业链 8
图 8 全球 AI 市场规模（十亿美元）及同比增速 9
图 9 头部主要厂商大模型迭代时间轴 10
图 10 MLLM 的架构示意图 11
图 11 训练与推理示意图 11
图 12 模型性能与计算量、数据大小、参数量的关系 12
图 13 大模型的涌现现象 12
图 14 GPT 系列模型迭代参数规模的变化 13
图 15 海外云厂商 2024Q1-Q4 资本开支（亿美元） 13
图 16 2020-2029E 全球大模型市场规模（亿美元）及增速 14
图 17 大模型商业化模式 15
图 18 模型推理部署四种主要形式的优劣 17
图 19 2024 年国内各行业大模型公开披露的落地项目数量（单位：个） 17
图 20 2024 年国内各行业大模型公开披露的落地项目金额（单位：亿元） 17
图 21 字节跳动 Coze 智能体创建界面 18
图 22 大模型区别于互联网时代的竞争特点 19
图 23 海外主流 AI 大模型基准评分差距逐步缩小 19
图 24 海外模型厂商竞争格局 20
图 25 大模型区别于互联网时代的竞争特点 21
图 26 2020-2024 年阿里云业务营收占比 21
图 27 国内模型厂商竞争格局 22
图 28 OpenAI 发展历程以及重要模型发布节点 23
图 29 OpenAI GPT 系列模型迭代相关性能参数 24
图 30 以 GPT-4 为例的 GPT 系列模型路径 24
图 31 OpenAI o1 与其他头部模型评分对比 25
图 32 OpenAI o1 在训练和推理阶段算力资源的投入与模型性能的关系 25
图 33 o1 模型相比 GPT-4o 在推理密集型任务上的改进 26
图 34 未来 GPT 系列与 o1 系列模型或将收敛融合 26
图 35 DeepSeek 发展历程以及重要模型发布节点 26
图 36 DeepSeek 用户增长速度 27
图 37 海内外接入 DeepSeek 的厂商 27
图 38 DeepSeek-V3 多项评测能力与海内外头部模型对比 28
图 39 DeepSeek 模型性能与价格比处于最优范围内 28
图 40 DeepSeek-V3 训练成本 28
图 41 DeepSeek-V3 的 MLA 和 DeepSeekMoE 架构 29
图 42 DeepSeek-R1 在数学、代码、自然语言推理等任务上的性能表现 30
图 43 DeepSeek-R1-Zero 在训练过程中的 AIME 准确性不断上升 31
图 44 蒸馏后的小型模型在数学、代码、自然语言推理等任务上的性能表现 32
图 45 AI 服务器产业链 33
图 46 寒武纪 2020-2024 年总营收和归母净利润与各自同比增速 33

图 47 海光信息 2020-2024 年总营收和归母净利润与各自同比增速 33
图 48 恒玄科技 2020-2024 年总营收和归母净利润与各自同比增速 34
图 49 乐鑫科技 2020-2024 年总营收和归母净利润与各自同比增速 34
图 50 兆易创新 2020 年-2024Q1-Q3 总营收和归母净利润与各自同比增速 34
图 51 澜起科技 2020-2024 年总营收和归母净利润与各自同比增速 34
图 52 中际旭创 2020-2024 年总营收和归母净利润与各自同比增速 35
图 53 天孚通信 2020 年-2024Q1-Q3 总营收和归母净利润与各自同比增速 35
图 54 鹏鼎控股 2020-2024 年总营收和归母净利润与各自同比增速 35
图 55 胜宏科技 2020-2024 年总营收和归母净利润与各自同比增速 35
图 56 浪潮信息 2020 年-2024Q1-Q3 总营收和归母净利润与各自同比增速 36
图 57 工业富联 2020-2024 年总营收和归母净利润与各自同比增速 36
图 58 麦格米特 2020 年-2024Q1-Q3 总营收和归母净利润与各自同比增速 36
图 59 欧陆通 2020 年-2024Q1-Q3 总营收和归母净利润与各自同比增速 36

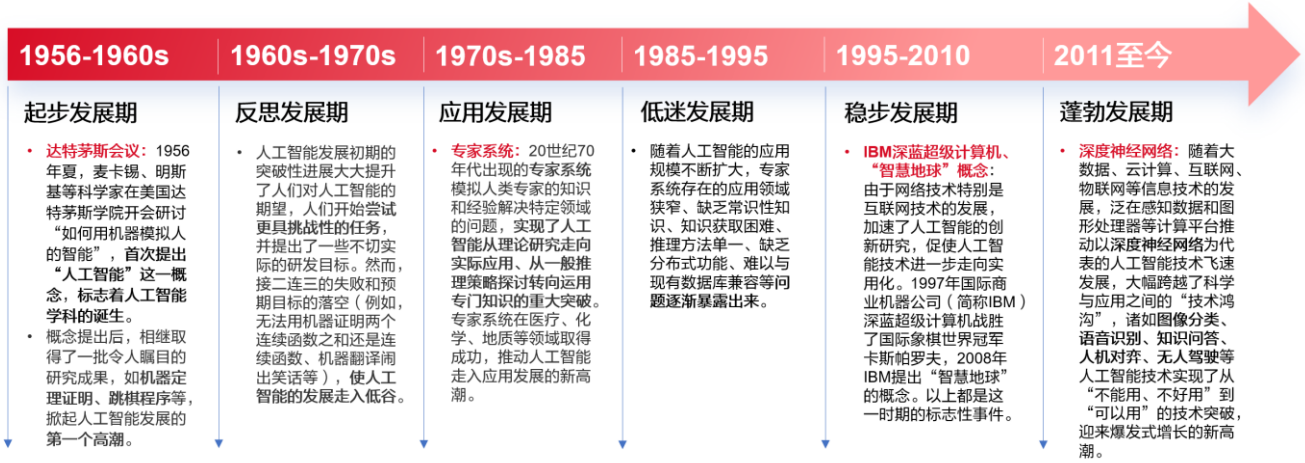
表 1 Transformer 架构和 MoE 架构的对比 13
表 2 模型 API 服务的构成、重要性和主要指标 15
表 3 海内外代表大模型 Token 定价 16

1.AI 市场高速扩张，有望引领新一代工业革命

1.1.AI 推动生产变革，行业步入蓬勃发展期

（1）人工智能（Artificial Intelligence，AI）是一种模拟人类智能的技术，旨在使机器能够像人类一样思考、学习和解决问题。AI 涵盖了多种技术和方法，包括深度学习、机器学习、计算机视觉和自然语言处理等。自 1956 年达特茅斯会议首次提出 AI 概念之后，AI 经历了早期的萌芽式发展，20 世纪 70 年代出现的专家系统实现了 AI 从理论研究走向实际应用、从一般推理策略探讨转向运用专门知识的重大突破，但后续 AI 因为一系列问题陷入发展瓶颈，进入 21 世纪，随着网络技术的发展，数据的获取变得更加容易，云计算的兴起提供了强大的计算能力，为深度学习的应用提供了土壤，2010 年起，以深度神经网络为代表的 AI 技术蓬勃发展，应用落地场景多点开花，尤其在近几年，大规模预训练模型时代开启，海内外以 ChatGPT、DeepSeek 等为代表的 AI 模型竞赛如火如荼，标志着 AI 进入了一个新的纪元。

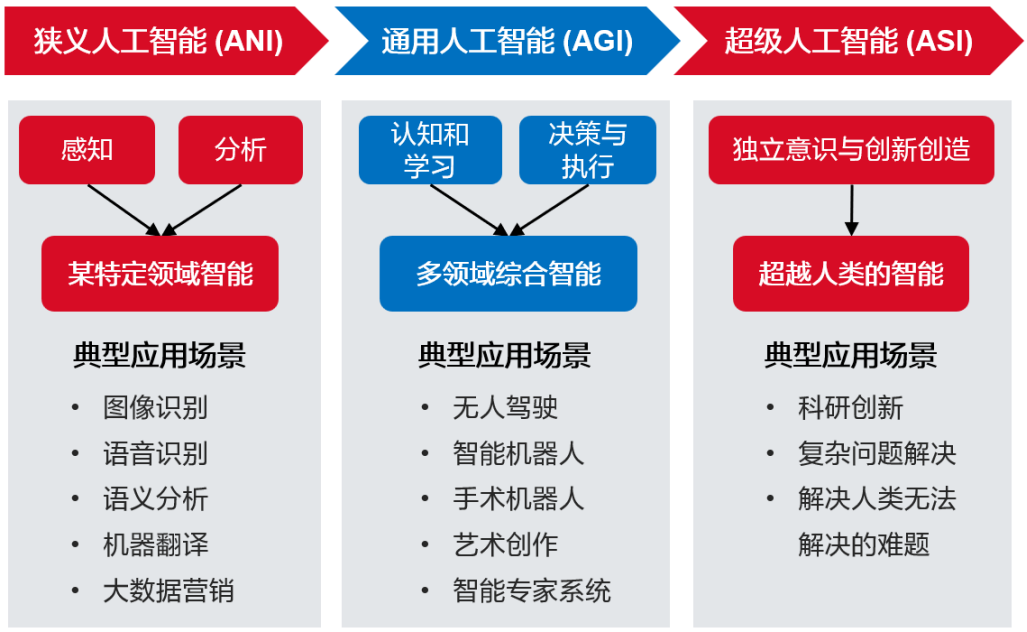
图1 人工智能发展历程



资料来源：国家互联网信息办公室

（2）按照智能程度划分，AI 主要分为狭义人工智能（ANI）、通用人工智能（AGI）和超级人工智能（ASI），目前 AGI 和 ASI 尚处于理论和探索阶段。ANI（Artificial Narrow Intelligence）又称弱人工智能指专注于特定任务的人工智能系统，能够高效执行特定功能，但其能力局限于预设任务，不具备通用智能。AGI（Artificial General Intelligence）指具备与人类相当的综合智能，能够理解、学习和执行任何智力任务，具备自主学习和推理能力。ASI（Artificial Super Intelligence）指在几乎所有领域超越人类智能的人工智能，具备自我改进能力，可能在科学、艺术等领域远超人类。目前，ANI 已广泛应用于图像和语音识别、自动驾驶等场景，AGI 尚未有实际应用，仍处于理论阶段，但 Sora 的问世无疑使我们离 AGI 更进了一步。

图2 按智能程度划分的三类人工智能



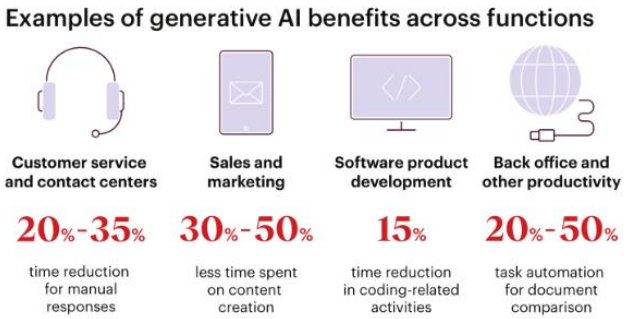
资料来源：行行查

（3）生成式人工智能（Generative Artificial Intelligence，Gen AI）是 AI 领域的重要分支，不同于传统的 AI 仅对输入数据进行处理和分析，Gen AI 能够学习并生成具有逻辑的新内容。Gen AI 可以学习并模拟事物的内在规律，是一种基于算法和模型生成具有逻辑性和连贯性的文本、图片、声音、视频、代码等内容的技术。早期 Gen AI 主要针对单一模态，如 GPT 系列生成文本、StyleGAN 生成图像。随着技术进步，Gen AI 开始结合多模态模型，依赖于复杂的机器学习模型，实现异构数据的生成式输出，创建跨模态原创内容（例如文本、图像、视频、音频或软件代码）以响应用户的提示或请求。在应用层面，Gen AI 可显著提升生产效率，根据贝恩，Gen AI 可在营销方面缩减 30%-50%内容创造所需的时间消耗，在软件开发方面缩短 15%的代码编写时间。

图3 Gen AI 的工作原理

资料来源：IBM

图4 Gen AI 在各领域的应用效果

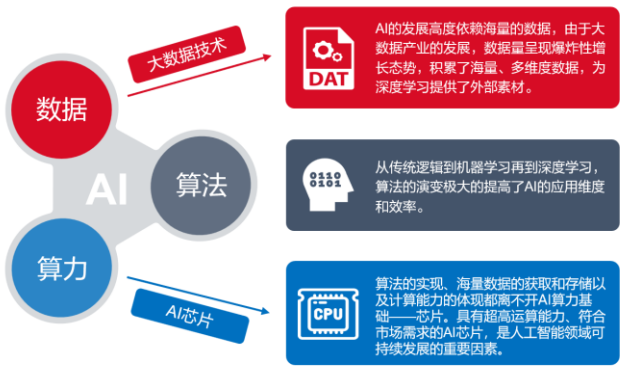


资料来源：贝恩

（4）AI 具有算力、算法、数据三大要素，其中基础层提供算力支持，通用技术平台解决算法问题，场景化应用挖掘数据价值。数据是 AI 学习和成长的基石，决定了算法是否能得到有效的训练和优化，数据的质量和数量也直接影响到 AI 模型的准确性和效率；算法是 AI 的灵魂，决定了 AI 如何处理数据和解决问题，其设计和选择直接关系到 AI 的性能和应用效果；算力是 AI 运行的动力，算力提供了执行算法和处理数据所需的计算资源，强大的算力可以支持复杂和大规模的 AI 应用。其中算力指计算设备在单位时间内处理数据的能力，

AI 算力是专门针对 AI 任务（如矩阵运算、神经网络训练）优化的计算能力，需支持高并行性和大规模数据处理，通常用浮点运算次数（FLOPS）衡量，衍生的还有 TFLOPS（万亿次/秒）、PFLOPS（千万亿次/秒）等常见单位，算力的核心硬件包括 GPU、ASIC、FPGA 等。

图5 AI 的三大要素



资料来源：行行查

图6 AI 算力的相关常用名词及其含义

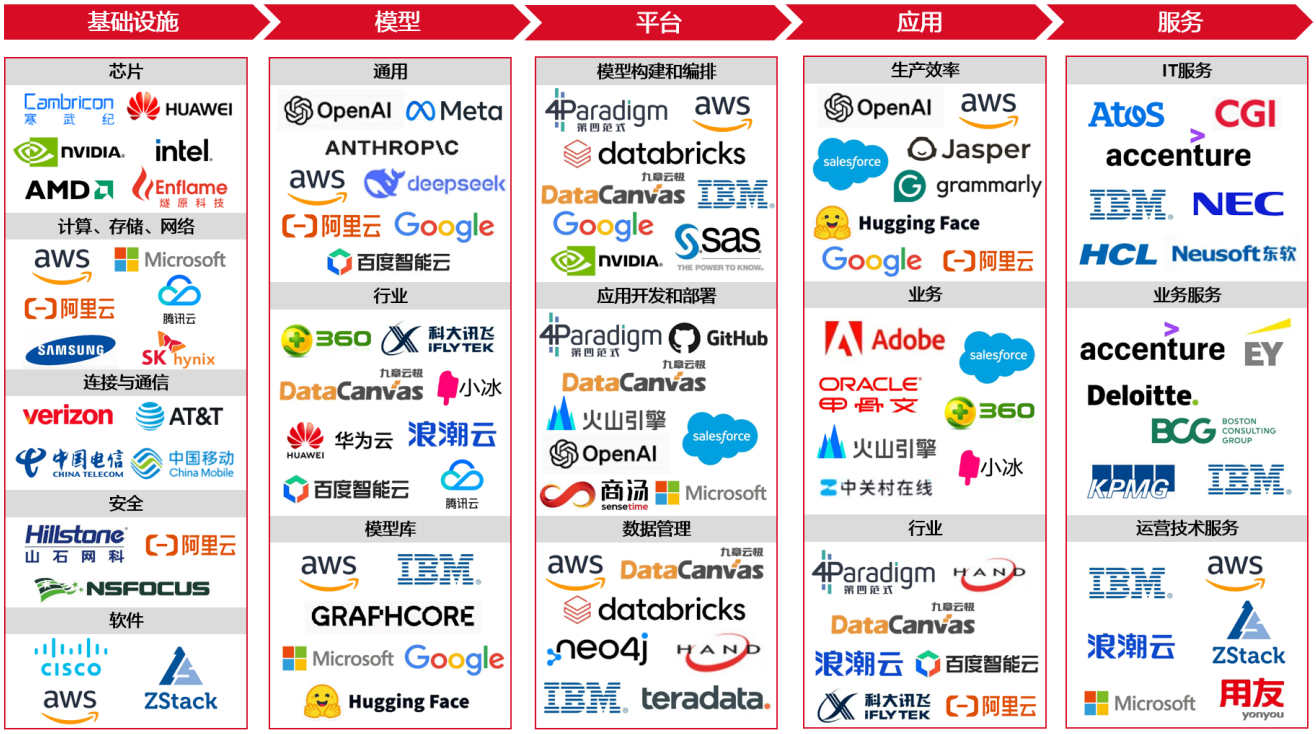
性能指标	FLOPS	Floating point number operations per second，每秒浮点运算（实数运算）次数，衡量硬件算力的核心指标
	TOPS	Tera operations per second，每秒可以处理的整型运算的万亿次数，常用于衡量整数运算（OPS，如INT8）性能，适用于推理场景
	QPS	Queries per second，每秒查询量，衡量AI系统的实际吞吐量（如每秒处理多少张图片）
	Latency	延迟，从输入到输出所需的时间
	Memory Bandwidth	显存带宽，GPU/TPU等加速器的内存数据传输速度，影响算力效率
硬件相关	GPU	图形处理器，最初用于图形渲染，后因并行计算能力强大，成为AI训练和推理的核心硬件
	TPU	张量处理器，谷歌专为AI计算设计的ASIC芯片，擅长矩阵运算，针对TensorFlow框架优化
	ASIC	专用集成电路，为特定任务（如AI推理）设计的芯片，能效比高，但灵活性低
	FPGA	现场可编程门阵列，硬件可重构，适合需要灵活性的场景（如算法迭代期）
	NPU	神经网络处理器，为深度学习设计的处理器，集成于手机/边缘设备（如手机芯片中的AI模块）

资料来源：CSDN

1.2.AI 产业链涵盖基础设施到应用落地多个环节

（1）AI 产业链可大致分为基础设施层、模型层、平台层、应用层及服务层，其中基础设施层包含芯片、存储、网络等，模型层包含通用模型、行业模型等。上游基础设施层是 AI 产业链的基础，主要涉及数据、算力等基础软硬件，包括 AI 芯片，代表厂商寒武纪、英伟达等；计算、存储、网络方面，代表厂商亚马逊、微软、阿里、三星电子等。模型层是 AI 产业链的核心部分，包括通用大模型和行业大模型等。平台层和模型层深度绑定，使大模型更便于使用和普及。随着 AI 大模型的发展，平台中多种模型选择、如何将大模型高效且可靠地部署于生产环境是当前的核心问题。应用层是 AI 产业链的终端环节，主要涉及 AI 在各个领域的应用和落地，而大模型的不断更新升级有助于加速应用场景的创新及商业化落地。

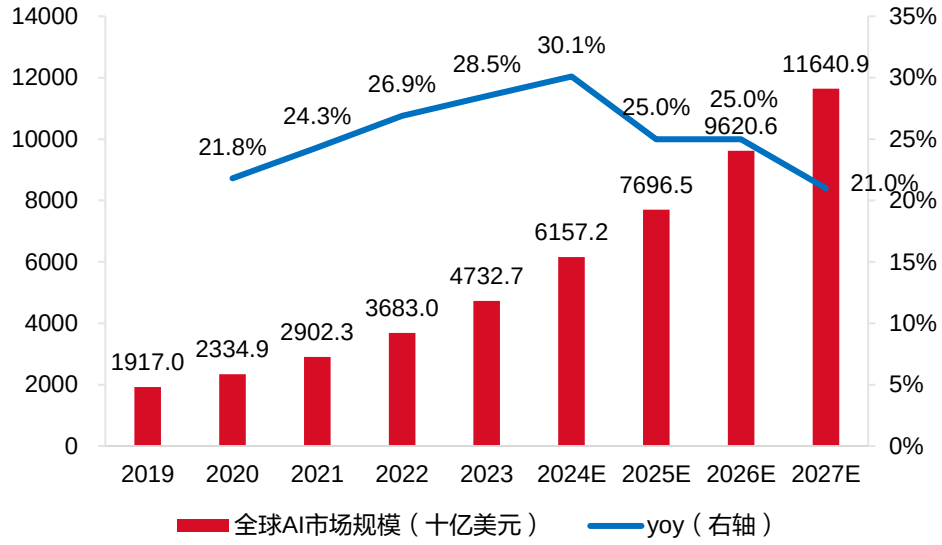
图7 AI 产业链



资料来源：IDC

（2）2024 年全球 AI 市场规模有望达到 6.16 万亿美元，同比增长 30.1%。根据 Frost & Sullivan，自 2020 年起，全球 AI 市场规模以高于 20%的同比增速呈现迅猛增长的态势，从 2019 年的 1.91 万亿美元有望扩张至 2024 年的 6.16 万亿美元，同比增速逐年上升，2025 年开始虽然预计增速同比放缓，但整体市场有望在 2027 年扩张至 11.64 万亿美元，体现出全球 AI 行业井喷式的发展速度。

图8 全球 AI 市场规模（十亿美元）及同比增速



资料来源：Frost & Sullivan