AIGC 行业全景篇——

算力、模型与应用的创新融合

胡国栋

招商银行研究院

行业研究员

☎：0755-83169269

✉：huguodong@cmbchina.com

相关研究报告

■ **人工智能的发展历程与 AIGC 的市场机遇**。人工智能的发展经历了三次浪潮，从最初的逻辑推理和专家系统，到机器学习和深度学习，人工智能技术不断演进。AIGC 利用生成式 AI 技术，创造出多样化的内容，展示出巨大的商业潜力。AIGC 产业链可划分为基础层、模型层和应用层。预计到 2030 年，AIGC 市场规模将增至 9810 亿美元，推动全球经济增长 4.9 万亿美元，累计产生的经济影响达到 19.9 万亿美元。

■ **基础层:AIGC 的快速发展推动算力需求激增，算力存储网络成为投资的主赛道。**随着 AIGC 技术的快速发展，特别是基于 Transformer 的大模型对算力需求急剧增加，全球互联网巨头纷纷加大对 AIGC 基础设施的投资，以推动创新和保持竞争优势。GPU 系统、HBM 存储和高性能网络基础设施在 AIGC 计算中发挥着关键作用，满足了对高速并行计算的需求，成为硬件基础设施投资的主赛道。

■ **模型层：算法进步、性能成本优化与商业模式多元化的融合**。AIGC 技术的进步得益于生成算法、预训练模型和多模态技术的创新。在大语言模型的竞争中，性能和成本是两个核心要素，性能的提升和成本的降低使得 AIGC 的应用更加广泛。AIGC 公司通过订阅服务、API 接入等多元化商业模式来拓展收入渠道，从而增强自身的市场竞争力。随着企业逐渐认识到 AIGC 技术的潜力，预计 B2B 服务将在整体 AIGC 市场中占据主导地位。全球大语言模型市场将显著增长，OpenAI 凭借 ChatGPT 的成功在市场上处于领先地位，而科技巨头如微软、亚马逊和谷歌也正通过技术创新和产品整合来追赶。

■ **应用层：技术创新应用推动市场发展和行业变革**。AIGC 技术正推动 ToC 和 ToB 领域的创新与多元化应用，覆盖 Chatbot、社交、游戏和内容创作等多个场景，并在企业层面提供提高效率、降低成本的解决方案。在电子设备领域，AIGC 技术引发革新，特别是在智能手机、汽车和人形机器人的智能化创新中表现突出。各类 AIGC 应用爆发式增长，其中 AI 广告市场份额最大,药物研发、网络安全和 IT 服务市场增速最快。

■ **业务建议**。随着市场需求的不断增长，AIGC 应用有望在多个行业实现突破，带来长远的发展机会，建议重点关注 AIGC 应用的长期机会。（本部分有删减，招商银行各行部请参照文末联系方式联系研究院）

■ **风险提示**。（1）伦理道德的风险。（2）技术缺陷的风险。（3）监管与法律的风险。（4）商业化不确定的风险。（5）市场竞争加剧的风险。（6）宏观经济波动的风险。

目录

图目录

表目录

AIGC 是一种基于生成式 AI 技术的新型内容创作方式。本篇报告围绕 AIGC 的发展历程与市场机遇、算力基础设施的发展趋势、大模型算法与商业模式的融合以及 AIGC 应用市场的创新发展来分析 AIGC 产业链的相关机会，最后阐明商业银行在 AIGC 赛道的业务机会与风险。

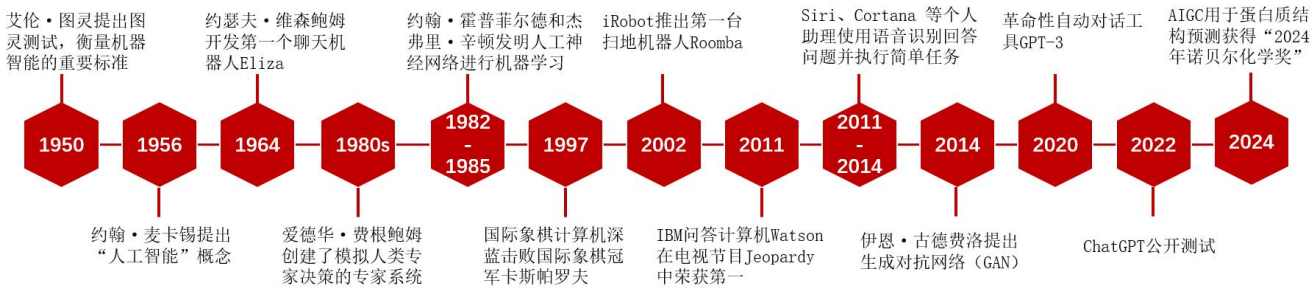
1. 人工智能的发展历程与 AIGC 的市场机遇

1.1 人工智能产业发展历程，从图灵测试到生成式 AI 的演进

人工智能（Artificial Intelligence，AI）作为计算机科学的一个重要分支，旨在深入探讨智能的本质，创造出能够模拟人类思维与反应的智能机器。经过多年的不断进化，人工智能如今已演变为一门涵盖机器人技术、语言识别、图像识别、自然语言处理及专家系统等多个研究领域的综合性学科。

人工智能的历史可以追溯到 20 世纪 50 年代。1950 年，被誉为“人工智能之父”的艾伦·图灵（Alan Turing）提出了著名的图灵测试，借助“问”与“答”的方式来评估机器是否具备智能。1956 年，约翰·麦卡锡（John McCarthy）在达特茅斯会议上首次正式提出“人工智能”这一术语，这一时刻标志着人工智能作为一门独立学科的诞生。

图 1：人工智能发展历史时间线



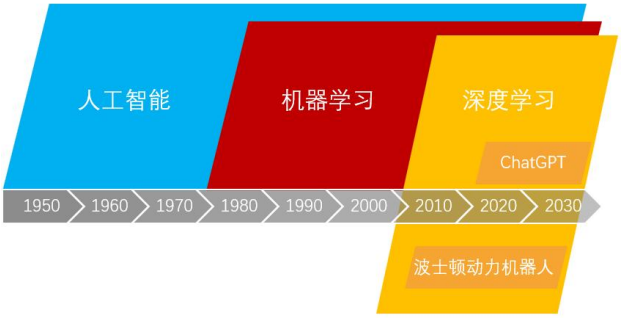
资料来源：招商银行研究院

人工智能的发展历程可以划分为三次浪潮，从最初的逻辑推理和专家系统，到机器学习和深度学习，人工智能技术不断演进。第一次浪潮（1950s-1970s）期间，研究主要集中在利用符号逻辑和推理来模拟人类智能，然而由于对技术能力的期望过高与实际进展之间的落差，到 70 年代中期，人工智能进入了“第一次 AI 之冬”。在第二次浪潮（1980s-2000s）期间，随着计算能力的提升和知识表示技术的发展，专家系统在 80 年代兴起，能够模拟特定领域的专家决策能力。90 年代，机器学习这一分支迅速崛起，使计算机能够从数据中

学习并不断改进。第三次浪潮（2010s-至今）以来，现代人工智能技术广泛应用，如卷积神经网络（CNN）、递归神经网络（RNN）和生成对抗网络（GAN）等，特别是在自然语言处理（NLP）领域的突破，例如 BERT 和 GPT 系列模型，使得机器在理解和生成自然语言方面取得了显著进展。

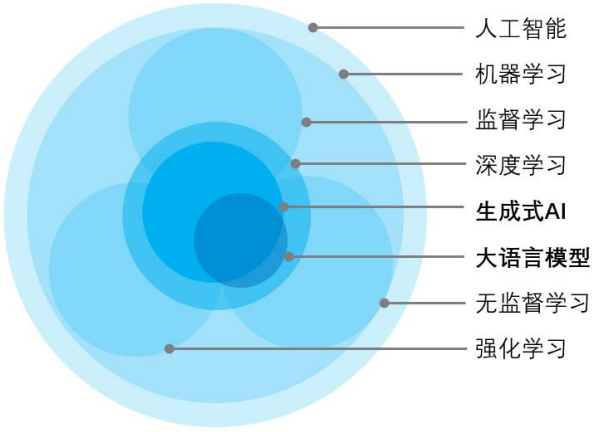
机器学习是人工智能的一个重要分支，使计算机系统能够从数据中汲取知识，进而做出预测或决策。该领域涵盖多个子领域，包括监督学习、无监督学习和强化学习等。监督学习通过利用带标签的训练数据来训练模型，使其能够对未见过的数据进行标签预测，常见算法有逻辑回归、支持向量机（SVM）、决策树和随机森林、神经网络等。无监督学习处理未标记的数据，旨在发现数据的内在结构或模式，常见的算法包括 K-均值聚类、层次聚类和主成分分析（PCA）等。强化学习通过与环境的交互学习如何采取行动，以实现长期奖励最大化，常见算法包括 Q 学习、Sarsa 和深度 Q 网络（DQN）等。机器学习的应用领域非常广泛，涵盖自然语言处理、计算机视觉、医疗健康、金融及推荐系统等多个领域。

图 2：人工智能技术发展经历三次浪潮



资料来源：infoDiagram、招商银行研究院

图 3：人工智能技术路线关系图



资料来源：CSDN、招商银行研究院

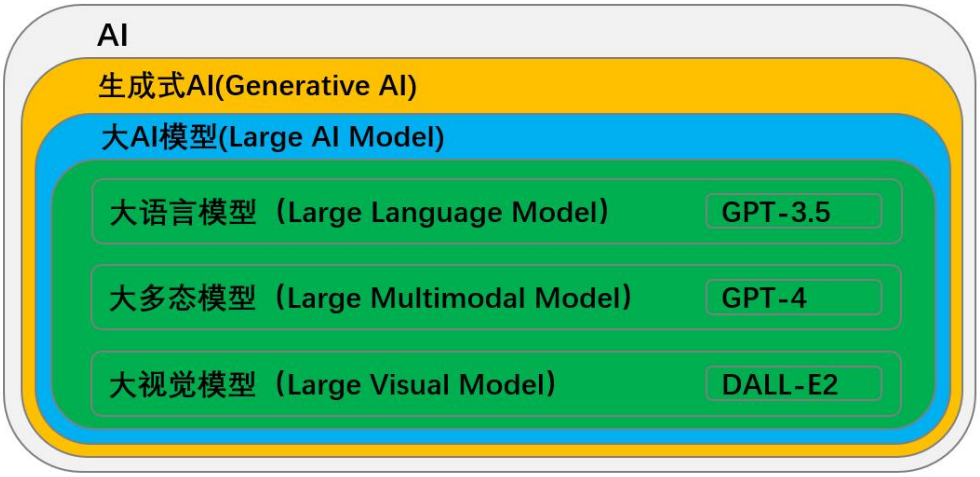
深度学习作为机器学习的一个重要分支，专注于利用深层神经网络解决复杂问题。它受人脑生物神经网络的启发，构建了由许多简单神经元组成的系统。每个神经元负责接收输入信号，进行加权求和，并通过激活函数生成输出。深度学习通过多层次的神经网络结构，能够自动提取数据的深层次特征。常见的算法包括前馈神经网络（FNNs）、卷积神经网络（CNNs）、循环神经网络（RNNs）以及 Transformer 等。深度学习的应用领域广泛，涵盖图像识别、语音识别、自然语言处理、生成模型、游戏与强化学习等，已成为当今人工智能技术的核心驱动力。

生成式 AI（Generative AI）作为深度学习的一个重要分支，是一种能够基于用户请求创造原创内容的人工智能技术。它通过深度学习模型识别现有内容的模式和结构，这些模型在未标记的原始数据上进行训练，旨在发现并编码大量数据中的模式与关系，以理解自然语言请求并生成相应的新内容。生成式 AI 的应用领域极为广泛，涵盖文本生成、图像合成、音乐创作以及语音合成等多个方面。

1.2 生成式 AI 技术推动内容创作的变革与大语言模型的发展

AIGC（Artificial Intelligence Generated Content）是一种利用生成式 AI 技术自动创作内容的新型生产方式。与传统 AI 主要关注于识别和预测现有数据模式不同，AIGC 则专注于创造全新的、有创意的数据。其核心原理在于学习和理解数据的分布，从而生成具有相似特征的新数据，能够生成文本、图像、音频、视频等多种形式的内容。

图 4：AIGC 与大型 AI 模型的关系



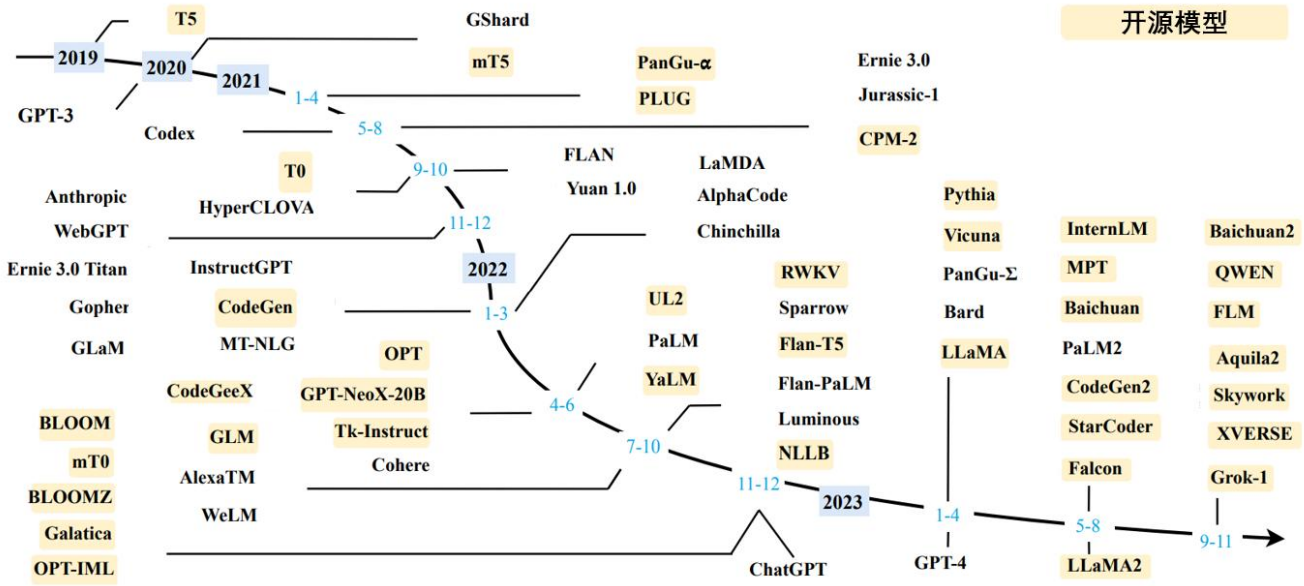
资料来源：《AIGC 的挑战和解决方案》、招商银行研究院

AIGC 涵盖了利用生成式 AI 技术生成的多种类型内容，而大型 AI 模型则是实现 AIGC 的重要技术手段。生成式 AI 通过深度学习模型在大数据集上进行训练，以创造新的文本、图像和音乐等多样化的内容。AIGC 不仅包括生成式 AI 算法，还涉及自然语言处理、计算机视觉（CV）和音频处理等核心技术。在生成式 AI 的框架中，大型 AI 模型发挥着至关重要的作用，通常采用大量参数的神经网络架构，主要包括大语言模型（LLM）、大多模态模型（LMM）和大视觉模型（LVM）。其中，大语言模型是最为核心的类型，包含数十亿以上参数的深度神经网络语言模型，运用自监督学习方法，通过大量未标注的文本进行预训练，从而掌握语言的复杂结构。需要注意的是，并非所有的大语言模型都专

注于生成任务，有些模型（如 BERT）更适合理解任务，而其他模型（如 GPT 系列）则更擅长生成任务。

大语言模型的发展经历了基础模型、能力探索和突破发展三个阶段。2017 年，Vaswani 等人提出的 Transformer 架构为大语言模型的演进奠定了坚实基础。自 2018 年以来，Google、OpenAI 和 Meta 等公司相继推出了多种大语言模型，如 OpenAI 的 GPT 系列、Google 的 Gemini 和 Meta 的 LLaMA 系列，推动了这一领域的迅猛发展。

图 5：大语言模型发展时间线



资料来源：《A Survey of Large Language Models》、招商银行研究院

在基础模型阶段（2018 年至 2021 年），研究主要集中在语言模型本身，Google 的 BERT 和 OpenAI 的 GPT-1 标志着预训练语言模型时代的开启，模型参数量实现了数量级的提升，从 GPT-1 的 1.17 亿增加到 GPT-3 的 1750 亿，这使得大模型在自然语言任务中的表现极为出色。

在能力探索阶段（2019 年至 2022 年），研究者们探索如何在不针对单一任务进行微调的情况下，充分发挥大语言模型的能力，随后引入指令微调方案（Instruction Tuning），使模型在特定任务上的表现更加精准，同时保留了在其他任务上的泛化能力。

在突破发展阶段（2022 年 11 月至今），大语言模型向多模态持续进阶，模型能力不断扩展，从文本生成延伸到图像和视频生成。自 ChatGPT 于 2022 年推出以来，它迅速在全球范围内引发了 AIGC 的广泛关注与应用热潮。以 ChatGPT 为起点，一个大语言模型即可实现回答问题、文稿撰写、代码生成和

数学解题等多项能力，这在过去需要多个小模型才能分别实现。GPT-4 作为一款开创性的多模态模型，凭借其卓越的综合实力成为行业标杆，后续推出的 GPT-4V、GPT-4-Turbo 和 GPT-4o 在性价比上不断提升。此外，Sora 文生视频模型能够根据文本提示生成视频内容，并对现有图像或视频进行编辑和扩展。

1.3 AIGC 产业链结构与未来市场增长展望

AI 产业链可分为基础层、模型层和应用层三个层面。基础层提供 AI 运行所需的底层算力资源和数据资源，其中算力资源涵盖 AI 芯片、存储、网络、安全及基础设施软件，数据资源则包括 AIGC 模型训练和优化所需的大量高质量多模态数据，以及数据分类、标记和清洗过滤的技术手段。模型层负责开发和优化模型算法，包括通用 AIGC 模型、行业应用微调模型，以及监督学习、无监督学习和强化学习等训练模型。应用层则涵盖针对企业的专用模型应用和针对个人用户的个性化服务，涉及文本、图像、音频、视频及多模态内容等多个应用服务方向。

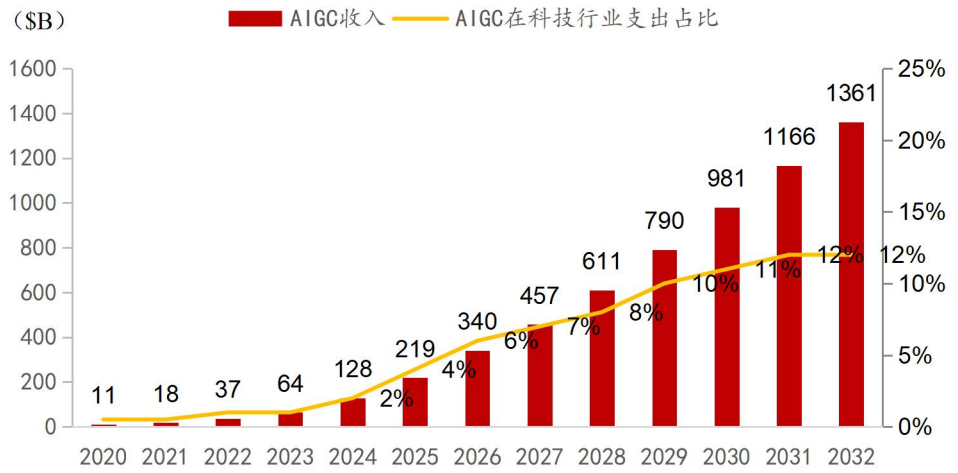
图 6：AIGC 产业链生态体系

资料来源：招商银行研究院

AIGC 技术展现出巨大的商业潜力，将成为全球经济增长的重要推动力。根据 IDC 的研究，到 2030 年，与商业相关的 AI 解决方案每投入 1 美元，预计将为全球经济带来 4.60 美元的直接和间接经济效应。预计到 2030 年，企业在采用 AI、将 AI 融入现有业务运营，以及向企业和消费者提供 AI 产品和服务的支出，将推动全球经济增长 4.9 万亿美元，累计产生的经济影响达到 19.9 万

亿美元，占全球 GDP 的 3.5%。彭博情报预测，随着 ChatGPT 等 AIGC 应用的快速增长，AIGC 市场有望从 2022 年的 370 亿美元增长至 2032 年的 1.36 万亿美元，年均复合增长率达到 43%。此外，AIGC 在信息技术硬件、软件、服务和广告等领域的支出占比将从 2022 年的 1%提升至 2032 年的 12%。

图 7：2020-2032 年 AIGC 市场规模



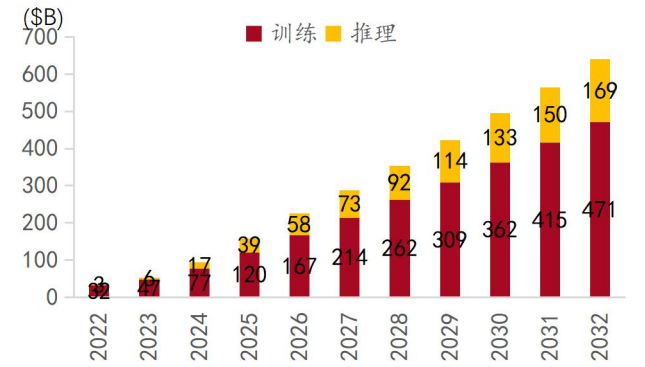
资料来源：Bloomberg Intelligence、招商银行研究院

受益于大模型算力需求，AIGC 硬件市场预计将迎来显著增长。随着 AIGC 大模型参数量的快速增加、数据规模的扩展以及对长文本处理能力的提升，算力的需求持续上升。彭博情报预测，AIGC 硬件市场将从 2022 年的 350 亿美元增长至 2032 年的 6400 亿美元，这一增长趋势反映了 AIGC 技术在训练和推理两个关键环节对算力资源的强大需求。

在训练阶段，AIGC 模型需要处理和分析庞大的数据集，这通常要求大量算力资源。预计训练硬件市场规模将从 2022 年的 320 亿美元增长到 2032 年的 4710 亿美元，年均复合增长率为 31%。而在推理阶段，通常需要较低功耗和成本的算力资源来满足用户终端的需求，预计推理硬件市场的增速将更高，从 2022 年的 30 亿美元增长至 2032 年的 1690 亿美元，年均复合增长率达到 48%。

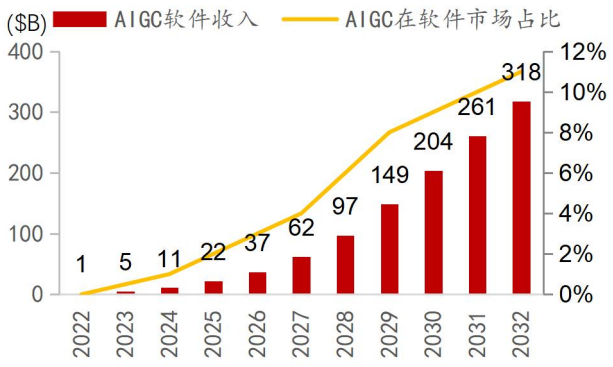
受益于 AIGC 对行业创新和效率的提升，AIGC 软件应用日益广泛。AIGC 技术正在改变影视、游戏、漫画和网络文学等领域，通过自动化和优化任务来提高生产效率并促进创意发展。例如，GitHub Copilot 基于 OpenAI 技术，向开发人员提供编码建议，从而减少编程时间，提高开发效率。集成 AIGC 助手正在成为软件行业的趋势，能够通过自动化和优化多种任务增强用户的工作效率。彭博情报预测，AIGC 软件市场将从 2022 年的 10 亿美元增长至 2032 年的 3180 亿美元，年均复合增长率高达 71%。

图 8：2020-2032 年 AIGC 硬件市场规模



资料来源：Bloomberg Intelligence、招商银行研究院

图 9：2020-2032 年 AIGC 软件市场规模



资料来源：Bloomberg Intelligence、招商银行研究院

2. 基础层：大模型的技术发展推动算力需求激增，算力存储网络成为投资的主赛道

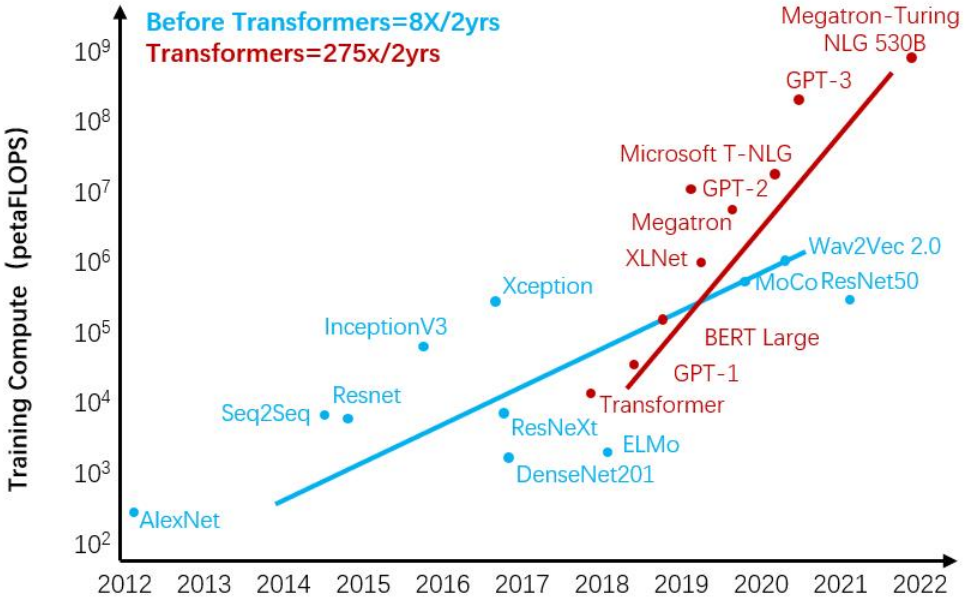
2.1 AIGC 技术迅猛发展引发算力需求激增

随着 AIGC 技术的迅猛发展，尤其是以 Transformer 为基础的大模型，对算力的需求激增。这些先进模型在训练和推理过程中，需要巨大的计算资源，包括高性能 GPU、高速存储以及高速通信网络。自 2017 年 Transformer 模型问世以来，它已成为构建大型语言模型的基石。该模型摒弃了传统的循环神经网络结构，通过自注意力机制处理序列数据，实现了对整个数据序列的并行处理，从而显著提升了训练和推理的效率。然而，这一技术进步也带来了对更高算力的迫切需求，进而增加了模型训练和部署的成本。

根据英伟达的数据显示，在引入 Transformer 模型之前，算力需求每两年增长约 8 倍；而采用 Transformer 模型后，这一数字飙升至每两年增长约 275 倍。为了满足不断攀升的算力需求，数据中心正朝着超大规模的发展方向迈进，以提供更强大的计算能力和更优越的可扩展性。同时，AI 服务器集群也在快速迭代与升级，以确保能够满足日益增长的算力需求。

根据 Scaling-law 法则，大语言模型的性能随着模型参数量、训练数据量和计算资源的增加而显著提升。从大模型的算力需求来看，随着参数规模、Token 数量以及训练所需算力的同步增长，模型性能不断提升。以 GPT-4 为例，其参数量从 GPT-3 的约 1750 亿提升至约 1.8 万亿，增幅超过 10 倍；而训练数据集的规模也从 GPT-3 的几千亿 Token 扩大到 13 万亿 Token。这种规模上的提升使得 GPT-4 在处理复杂问题和生成自然语言文本方面的能力得到了极大的增强。

图 10：AIGC 大模型对算力需求持续快速增长



资料来源：英伟达、招商银行研究院

随着 AIGC 大模型性能的显著提升，对计算资源的需求也呈现出指数级的增长。以 GPT-4 为例，其训练过程需要约 2.15e25 FLOPS 的运算量，这通常需要动用约 25000 块 A100 GPU，且训练周期长达 90 至 100 天。此外，数据采集、模型优化和强化学习等环节的额外开销，使得整体成本变得更加高昂。根据斯坦福大学 2024 年发布的 AI 指数报告，AIGC 模型的训练成本正在急剧上升，GPT-4 的成本从 2022 年 GPT-3 的大约 430 万美元激增至 2023 年的 7835 万美元。随着模型的不断扩展和训练过程的日益复杂，这些成本预计将继续攀升。

表 1：LLM 模型对 GPU 算力需求持续提升

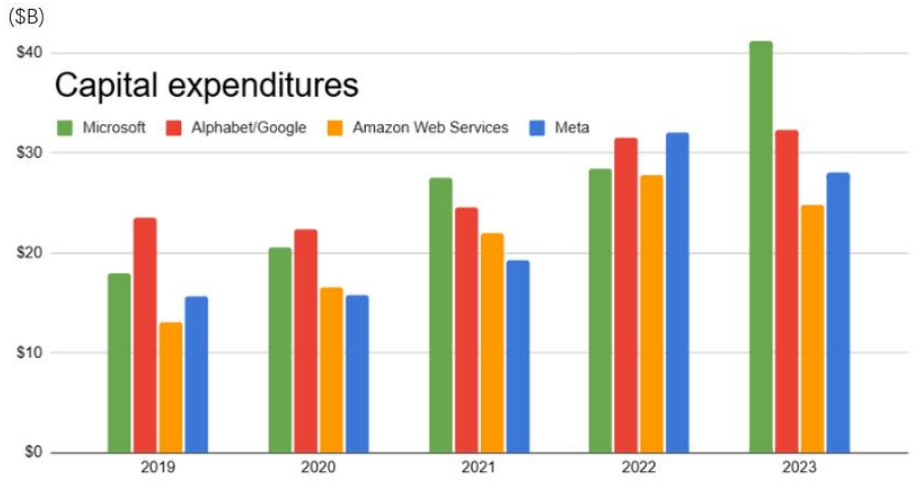
模型	参数规模（B）	Token 规模(B)	训练量 Z FLOPS	GPU 类型	GPU 算力 TFLOPS/s	GPU 利用率	训练时间	GPU 数量
GPT-3	175	300	420	H100	1600	0.3	1 周	1447
OPT	175	300	430	H100	1600	0.3	1 周	1482
LLaMA	65	1400	600	H100	1600	0.3	1 周	2067
LLaMA2	34	2000	400	H100	1600	0.3	1 周	1378
LLaMA2	70	2000	800	H100	1600	0.3	1 周	2756
GPT-4	1800	13000	21500	H100	1600	0.3	1 周	107474

资料来源：公开资料、招商银行研究院

2.2 互联网巨头加速 AIGC 投资，以推动创新和竞争

根据海外互联网巨头的资本开支计划，亚马逊、微软、谷歌和 Meta 等公司正在持续增加对 AIGC 基础设施的大规模投资。2021 至 2023 年间，这些公司的总资本支出达到 4670 亿美元，年均约 1550 亿美元。在 2024 年第二季度，资本支出达到 520 亿美元，同比增长 54%，主要用于数据中心的 AIGC 建设，以推动产品转型和技术升级。展望未来，预计在 2024 至 2027 年间，这四家公司将投入总额 8500 亿美元，年均 2125 亿美元，助力在竞争激烈的 AIGC 市场中保持领先地位并推动创新与增长。

图 11：海外互联网巨头资本支出飙升



资料来源：GeekWire、招商银行研究院

从海外互联网巨头的管理层表态来看，针对 AIGC 领域的高额资本支出将持续一段较长时期。Meta CEO 马克·扎克伯格强调维持 AI 行业领头羊地位的紧迫性，并警告投资不足可能在未来 10 至 15 年内使 Meta 处于竞争劣势。谷歌 CEO 桑达尔·皮查伊表示，公司倾向于超额投资，以确保抓住 AI 领域的收入机会，即使面临一定的资源过剩风险。亚马逊 CFO 布莱恩·奥尔萨夫斯基预计，2024 年下半年资本支出将持续增长，主要投资用于满足 AIGC 及非生成式 AI 技术的市场需求。微软 CFO 艾米·胡德则宣布，将加大 AI 基础设施建设投入，预计在 2025 财年刷新资本支出记录，以应对不断攀升的 AIGC 和云服务产品需求。

无论是在全球还是中国市场，AIGC 领域的投资都在迅速增长。IDC 数据显示，全球 AI 资本支出预计将从 2022 年的 1325 亿美元增长到 2027 年的 5124 亿美元，年均复合增长率为 31.1%。在中国市场，AI 资本支出同样展现出强劲的增长势头，预计将从 2022 年的 128 亿美元增至 2027 年的 400 亿美元，年均复合增长率为 25.6%。中国将在亚太地区人工智能市场发展中继续发挥引领作用，其 AI 资本支出占亚太地区总支出的 50%。

图 12：中国 AIGC 市场资本支出预测

资料来源：IDC、招商银行研究院

2.3 算力：GPU 引领 AIGC 技术革新，市场需求持续增长

当前，人工智能领域的 AI 芯片家族日益壮大，主要包括 GPU（图形处理器）、FPGA（现场可编程门阵列）、ASIC（专用集成电路）和 NPU（神经拟态芯片）。其中，GPU 和 FPGA 属于成熟的通用型 AI 芯片，而 ASIC 则为特定 AI 场景定制，如谷歌的 TPU、亚马逊的 Trainium 和微软的 Maia。

GPU 最初设计用于加速图形渲染和显示，广泛应用于游戏、视频制作和处理等领域。随着时间推移，因其在并行处理密集数据方面的卓越能力，GPU 逐渐成为 AI 领域的重要推动力，尤其是在深度学习训练中。其核心性能指标包括算力、显存、功耗和互联能力，成为推动 AIGC 发展的核心力量。

英伟达的 GPU 产品在 AIGC 的发展历程中扮演了至关重要的角色，其成功源于在硬件性能和软件生态方面的持续投入与创新。在硬件领域，英伟达推出了 Volta、Turing、Ampere、Hopper 和 Blackwell 等系列架构，这些架构配备了专为深度学习设计的 CUDA Core 和 Tensor Core，显著提升了 AI 训练与推理的效率。CUDA Core 负责基础运算，其数量通常与 FP32 计算单元相对应；而 Tensor Core 则在 Volta 及后续架构中引入，专门用于张量计算，与深度学习框架（如 TensorFlow 和 PyTorch）结合使用，带来了十几倍的效率提升。除了硬件创新，英伟达还构建了全面的 GPU 软件生态系统，包括 CUDA、cuDNN 和 TensorRT 等工具，大大简化了 AIGC 模型的开发和部署流程，使得 AIGC 技术的应用更加高效便捷。

表 2：英伟达主流 GPU 产品性能对比

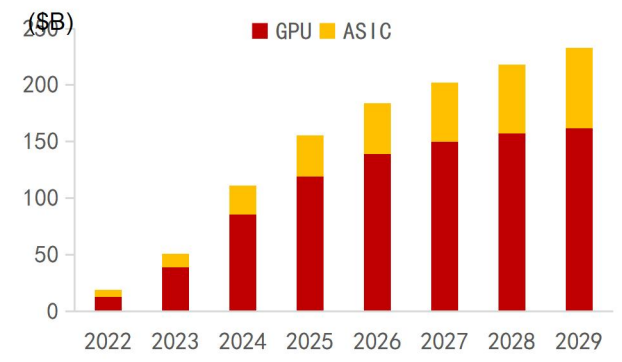
性能参数	A100	H100	H200	GB200
架构	Ampere	Hopper	Hopper	Blackwell
发布时间	2020	2022	2023	2024
FP64	9.7 TFLOPS	34 TFLOPS	34 TFLOPS	40 TFLOPS
FP64 Tensor Core	19.5 TFLOPS	67 TFLOPS	67 TFLOPS	90 TFLOPS
FP32	19.5 TFLOPS	67 TFLOPS	67 TFLOPS	180 TFLOPS
TF32 Tensor Core	312 TFLOPS	989 TFLOPS	989 TFLOPS	5 PFLOPS
BFLOAT16 Tensor	624 TFLOPS	1979 TFLOPS	1979 TFLOPS	10 PFLOPS
FP16 Tensor Core	624 TFLOPS	1979 TFLOPS	1979 TFLOPS	10 PFLOPS
FP8 Tensor Core	-	3958 TFLOPS	3958 TFLOPS	20 PFLOPS
INT8 Tensor Core	1248 TFLOPS	3958 TFLOPS	3958 TFLOPS	20 PFLOPS
GPU 内存	80 GB	80 GB	141GB	384GB
GPU 内存带宽	2.04 Tbps	3.35 Tbps	4.80 Tbps	16Tbps
互联技术	NVLink:600 GB/s PCIe Gen4:64GB/s	NVLink: 900GB/s PCIeGen5:128GB/s	NVLink: 900GB/s PCIeGen5:128GB/s	NVLink: 1.8TB/s PCIeGen6:256GB/s

资料来源：英伟达、招商银行研究院

随着 AIGC 技术在多个行业中的广泛应用，对 GPU 和 ASIC 算力的需求持续增加。全球数据中心 GPU 市场在近年来显著扩张，2023 年出货量达到 385 万颗，较 2022 年的 267 万颗增长了 44.2%。预计在经历 2023 年和 2024 年的大幅增长后，全球 AIGC GPU 和 ASIC 市场将保持稳定增长。根据 Yole 的预测，该市场规模将从 2023 年的 505 亿美元增至 2029 年的 2330 亿美元，复合年均增长率达到 29.0%。

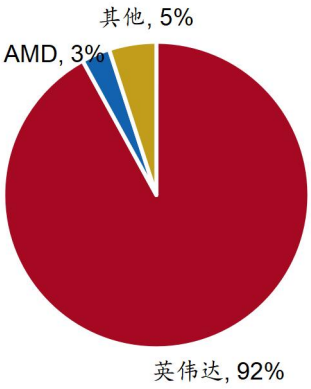
英伟达作为 GPU 市场的领导者，其产品在 AIGC 训练和推理市场占据主要份额。英伟达不断推出新的 GPU 架构和软件产品，每一代产品都在性能和能效方面持续提升，其 2023 年数据中心 GPU 销售收入达到了 362 亿美元，根据 IoT Analytics 的数据，市场份额达到 92%。与此同时，AMD 和英特尔也在数据中心 GPU 市场占有一席之地，AMD 的 MI300 系列获得了微软、Meta 等订单，市场份额达到 3%；英特尔的 Gaudi 2 则提供高性能且具成本效益的解决方案。此外，一些新兴参与者也在进入市场，推动技术创新与产品多样化。

图 13：全球 AIGC GPU 和 ASIC 市场规模预测



资料来源：Yole、招商银行研究院

图 14：AIGC GPU 市场份额（2023 年）



资料来源：IoT Analytics、招商银行研究院

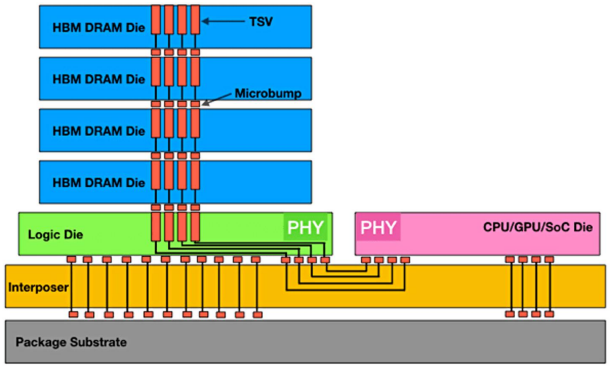
2.4 存储：HBM 凭借高带宽和低延迟推动 AIGC 计算

HBM（高带宽存储器）是一种采用 3D 堆叠技术的 DRAM，通过先进的硅通孔（TSV）封装方法，能够实现高容量、高带宽、低延时和低功耗的特性。这种设计特别适用于高性能计算和图形处理，尤其是在 AIGC 计算中，HBM 与 GPU 的结合极大提升了并行数据处理速度。

在 AIGC 计算中，GPU 需要处理大量并行数据，要求具备高算力和大带宽。通过中介层与 HBM 的互联封装，HBM 的高带宽特性为 GPU 提供了充足的内存带宽，支持其高速数据处理需求，从而加速 AIGC 模型的训练和推理过程。

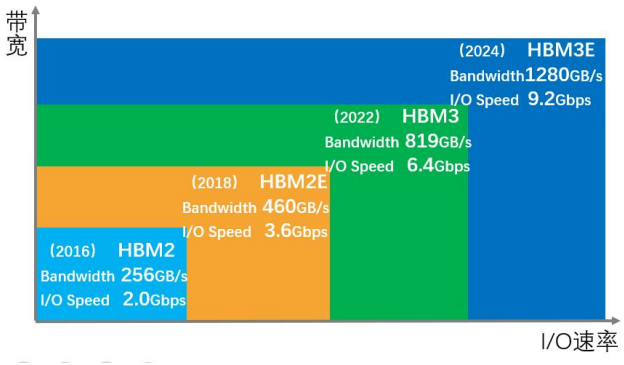
HBM 存储技术自 2013 年 SK 海力士首次推出 HBM1 以来，经历了多次重要的产品迭代，包括 HBM2、HBM2E、HBM3 和最新的 HBM3E。每一代产品在容量、带宽和功耗效率上都有显著提升，其中 HBM3E 提供高达 9.2Gbps 的 I/O 传输速度和超过 1280GB/s 的带宽。展望未来，预计 HBM4 将在 2026 年上市，将支持更广泛的内存层配置，以更好地满足不同类型的应用需求。

图 15：HBM 3D 堆叠与 GPU 封装架构



资料来源：Tech Investments、招商银行研究院

图 16：HBM 技术路线图

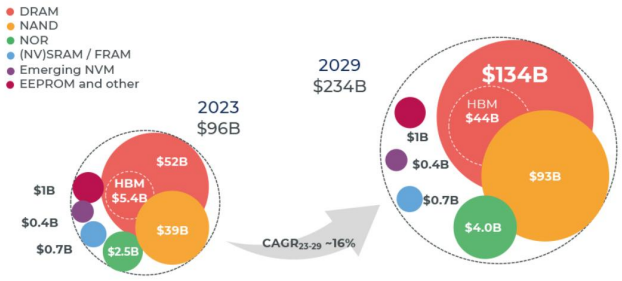


资料来源：SK 海力士、招商银行研究院

AIGC 的迅猛发展正在刺激数据中心对高速 DRAM 内存的需求持续增长。随着超大规模企业不断扩展服务器容量以支持大语言模型的训练和推理任务，高速 DRAM 和 HBM 的需求也显著上升。数据中心已成为 DRAM 需求最大的细分市场，占据整体市场份额的 50%。预计随着 HBM 和 CXL（Compute Express Link）等新技术的普及，数据中心 DRAM 市场的增速将超过整体 DRAM 市场。根据 Yole Group 的预测，DRAM 市场将从 2023 年的 520 亿美元增长到 2029 年的 1340 亿美元，复合年均增长率为 17%，而数据中心 DRAM 在 2023-2029 年间的复合年增长率将达到 25%。

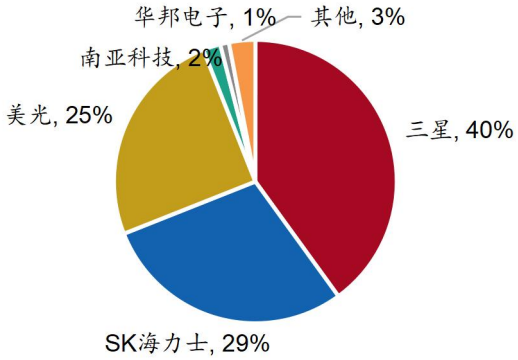
在全球 DRAM 市场中，三星、SK 海力士和美光占据了主导地位，三家公司合计市场份额高达 94%。其中，三星以 40%的市场份额继续稳居全球最大 DRAM 供应商，SK 海力士和美光分别占有 29%和 25%的市场份额。

图 17：存储行业全球市场规模预测（2023-2029）



资料来源：Yole、招商银行研究院

图 18：DRAM 全球市场份额（2023）

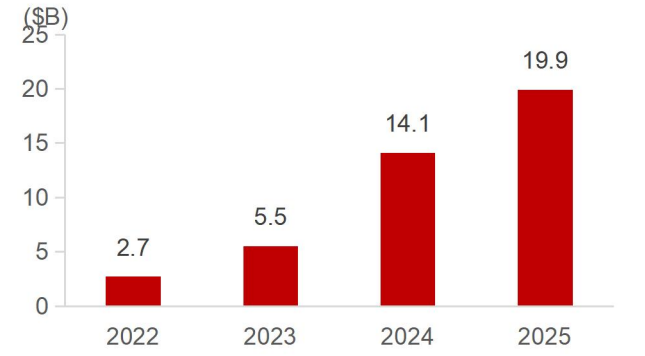


资料来源：Yole、招商银行研究院

随着 AIGC 算力需求的持续上升，全球 HBM 市场正经历快速增长。根据 Yole Group 的分析，全球 HBM 市场预计将从 2023 年的 55 亿美元增长至 2029 年的 377 亿美元，复合年均增长率达 37.8%。HBM 市场增速将显著超过整体 DRAM 市场，HBM 在整体 DRAM 出货量中的占比预计将从 2023 年的 2%增长到 2029 年的 6%，营收占比将从 2023 年的 10.4%增长到 2029 年的 32.8%。

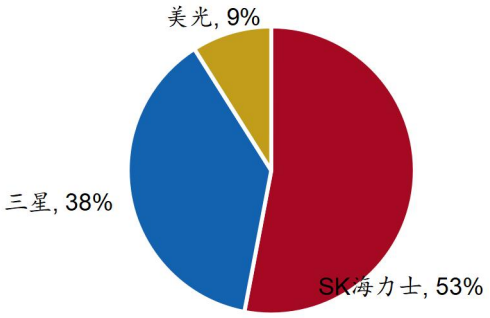
SK 海力士在 HBM 市场中占据领先地位。SK 海力士在 HBM 的开发和商业化方面处于领先地位，是 HBM3 的主要供应商，也是 Nvidia H100 和 H200 的唯一供应商。三星主要生产 HBM2E，并计划开始生产 HBM3。美光则跳过 HBM3，直接推出 HBM3E。根据 TrendForce 的数据，2023 年 SK 海力士、三星、美光在 HBM 市场的份额分别为 55%、41%、9%。

图 19：HBM 全球市场规模



资料来源：Yole、招商银行研究院

图 20：HBM 全球市场份额（2023）

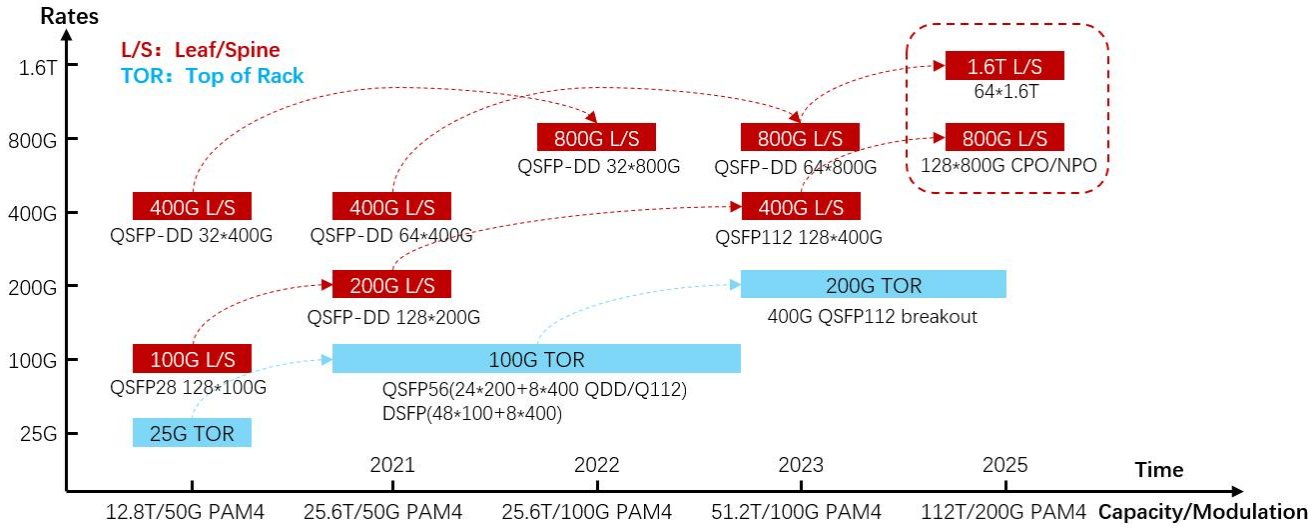


资料来源：TrendForce、招商银行研究院

2.5 网络：高性能网络基础设施推动 AIGC 发展，高速率光模块需求激增

在 AIGC 算力架构中，高性能网络基础设施扮演着至关重要的角色。网络瓶颈可能导致 GPU 集群的利用率降低、大模型训练时间延长及训练成本增加。因此，为了满足 AIGC 计算的需求，尤其是在大规模 GPU 集群中，亟需大量高效的网络交换设备，以支持高速率、低时延、高吞吐量和高能效的数据传输。800G/1600G 网络技术的发展，能够提供更高的数据传输速率和更低的传输时延，从而加速 AIGC 模型的训练与推理过程。

图 21：AIGC 发展推动数据中心向 800G 以上速率发展



资料来源：FS、招商银行研究院

AI 集群的快速发展正推动光模块需求的迅速增长。随着 GPU 性能提升和 AIGC 应用扩展，对光模块的数量和速率需求持续增加。以英伟达 DGX H100 的二层网络架构为例，每个 H100 SuperPOD 包含 4 个 SU 扩展集群，每个 SU 扩展集群由 32 个 H100 服务器和 8 个 Leaf 交换机组成，整个 SuperPOD 共有 1024 个 GPU、32 个 Leaf 交换机和 16 个 Spine 交换机。在 Leaf 层，服务器侧采用 400G 光模块，交换机侧采用 800G 光模块，总共需 1024 个 400G 光模块和 512 个 800G 光模块。在 Spine-Leaf 层采用 800 光模块互联，共需 1024 个 800G 光模块。在该架构下，GPU 与 400G 光模块的比例为 1:1，与 800G 光模块的比例为 1:1.5。当系统升级到 DGX H100 三层网络架构时，GPU 升级至 800G 网卡，Leaf 和 Spine-Leaf 层均采用 800G 光模块互联，导致光模块需求显著增加，GPU 与 800G 光模块的比例增至 1:6。 随着英伟达新一代 Blackwell 架构平台的发布，无论是小型 GB200 集群还是大型 GB200 集群，对 800G 光模块的需求比例均有所上升：小型集群的 GPU 与 800G 光模块比例达到 1:2，而大型集群则为 1:4.5。

表 3：英伟达 GPU 与光模块需求测算

GPU	NIC	Switch	网络类型	200G 光模块	400G 光模块	800G 光模块
A100	200G	200G	Layer 3	1:6	-	-
A100	200G	200G	Layer 2	1:1	-	1:0.75
H100	400G	400G	Layer 2	-	1:1	1:1.5

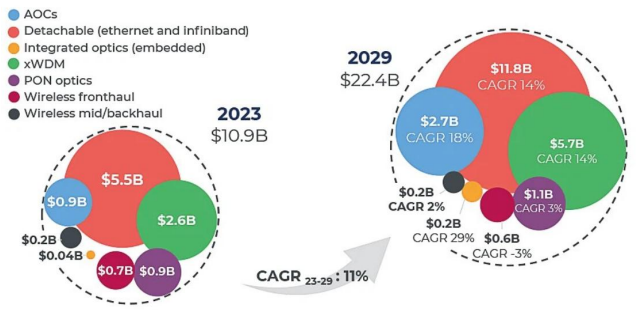
H100	800G	800G	Layer 3	-	-	1:6
GB200	800G	1600G	Layer 2	-	-	1:2
GB200	800G	1600G	Layer 3	-	-	1:4.5

资料来源：Fibermall、英伟达、招商银行研究院

AIGC 技术的迅速发展正在引领光模块市场的扩张。2023 年 5 月，谷歌和英伟达成为首批显著增加光模块采购以支持 AIGC 数据中心的大客户，随后其他领先的云计算公司也纷纷加入光模块竞争。根据 LightCounting 的预测，用于 AI 集群的光模块市场预计将从 2023 年的 20 亿美元增长至 2029 年的 120 亿美元，年均复合增长率达到 34.8%。在 2025 至 2029 年间，该市场规模将超过 520 亿美元。

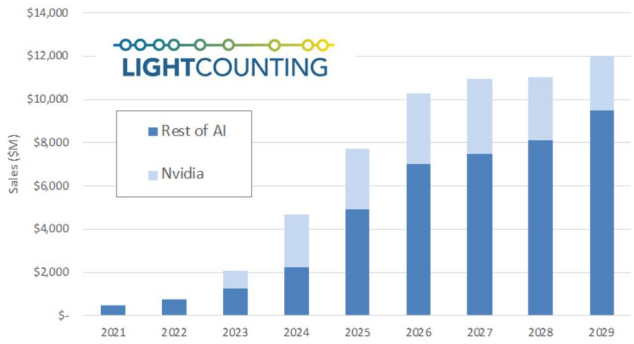
随着 AIGC 大模型训练对更快数据传输的需求不断上升，高速光模块的需求随之激增。预计到 2024 年，800G 光模块将成为市场主流，1.6T 光模块也将进入试产阶段。预计到 2029 年，1.6T 和 3.2T 光模块的市场规模将达到 100 亿美元，成为 AI 集群光模块市场的主要组成部分。

图 22：全球光模块市场预测



资料来源：Yole、招商银行研究院

图 23：AI 集群光模块市场预测



资料来源：LightCounting、招商银行研究院

中国厂商在全球光模块市场中表现卓越。根据 LightCounting 的数据，2023 年，全球前十大光模块厂商中，中国企业占据了 7 个席位，合计市场份额超过 50%。多家中国厂商已成功进入英伟达的供应链。其中，中际旭创的市场份额位居全球第一，华为、光迅科技、海信、新易盛、华工正源、索尔思光电分别排名第三、第五、第六、第七、第八和第九位。进入 2024 年，全球主要光模块厂商业绩的变化进一步显示出市场份额向中国厂商的集中趋势。尽管中国厂商在光模块制造领域取得了显著进展，然而在光芯片和电芯片等关键原材料方面，仍然依赖进口，国产化程度亟待提高。

表 4：全球 TOP10 光模块厂商排名

排名	2010 年	2016 年	2018 年	2023 年
1	Finisar	Finisar	Finisar	**旭创科技**
2	Opnext	**海信宽带**	**旭创科技**	Coherent
3	Sumitomo	**光迅科技**	**海信宽带**	**华为**
4	Avago	Acacia	**光迅科技**	Cisco（Acacia）
5	Source Photonics	FOIT（Avago）	FOIT（Avago）	**光迅科技**
6	Fujitsu	Oclaro	Lumentum/Oclaro	**海信宽带**
7	JDSU	**旭创科技**	Acacia	**新易盛**
8	Emcore	Sumitomo	Intel	**华工正源**
9	**武汉电信器件**	Lumentum	AOI	**索尔思**
10	NeoPhotonics	Source Photonics	Sumitomo	Marvell

资料来源：LightCounting、招商银行研究院

3. 模型层：算法进步、性能成本优化与商业模式多元化的融合

3.1 生成算法、预训练模型与多模态技术催生 AIGC 的迅猛发展

AIGC 与以往的 AI 技术最显著的区别在于其从分析式 AI（Analytical AI）发展为生成式 AI。分析式 AI 模型主要通过对已有数据的分析、判断和预测来提供决策支持，而生成式 AI 模型则是通过学习已有数据，创造出全新的内容。这一转变得益于**先进的生成算法、强大的预训练模型以及创新的多模态技术**，共同推动了 AIGC 的迅猛发展和爆炸性增长。

AIGC 的快速崛起得益于基础生成算法的持续创新与突破。核心生成算法，如生成对抗网络（Generative Adversarial Network，GAN）、扩散模型（Diffusion Model）和 Transformer 等，为 AIGC 的发展奠定了坚实的技术基础。这些算法的不断进步推动了 AIGC 的爆发，拓展了其在内容生成领域的应用潜力。

2014 年，伊恩·古德费洛提出的生成对抗网络（GAN）成为早期最著名的生成式模型，标志着生成式 AI 的重要里程碑。随之而来，诸如 DCGAN、Style-GAN、BigGAN 和 CycleGAN 等变种架构相继问世，这些发展不仅推动了 GAN 理论的深化，也为图像生成、视频生成和三维模型生成等领域提供了强大的工具，极大丰富了生成式 AI 的应用场景。

2017 年，Vaswani 等人提出的 Transformer 模型引入了自注意力机制，使得模型能够根据输入序列中的不同部分分配不同的注意权重，从而更有效地捕

捉语义关系。这一创新催生了众多变体，如 BERT、GPT 和 XLNet 等，这些模型在各自领域取得了显著成果，推动了自然语言处理等行业的发展。伴随着生成式算法的不断创新突破，AIGC 如今能够生成多种类型的内容和数据，包括文本、代码、图像、语音和视频物体等，展现了广泛的应用潜力。

表 5：主流生成算法模型

算法模型	提出时间	模型描述
生成对抗网络（GAN）	2014 年	GAN 由生成器（Generator）和判别器（Discriminator）两部分组成，两者在训练过程中相互对抗，最终达到一种平衡状态，使得生成器能够生成出足够逼真的假数据。
基于流的生成模型（Flow-based Generative Model）	2015 年	该模型通过一系列可逆的变换来映射一个已知的概率分布到目标数据分布。该模型的核心优势在于它们能够直接优化数据的对数似然，并且具有高效的采样和推断能力。
扩散模型（Diffusion Model）	2015 年	该模型通过模拟扩散过程来逐步添加噪声到数据中，并随后学习反转这个过程以从噪声中构建出所需的数据样本。模型通过神经网络学习逆扩散过程，从而能够生成与原始数据相似的样本。
Transformer 模型	2017 年	该模型基于自注意力机制（Self-Attention），它允许模型在处理序列数据时动态地计算输入序列中每个位置与其他位置的关联程度，从而更好地捕捉序列之间的长距离依赖关系。

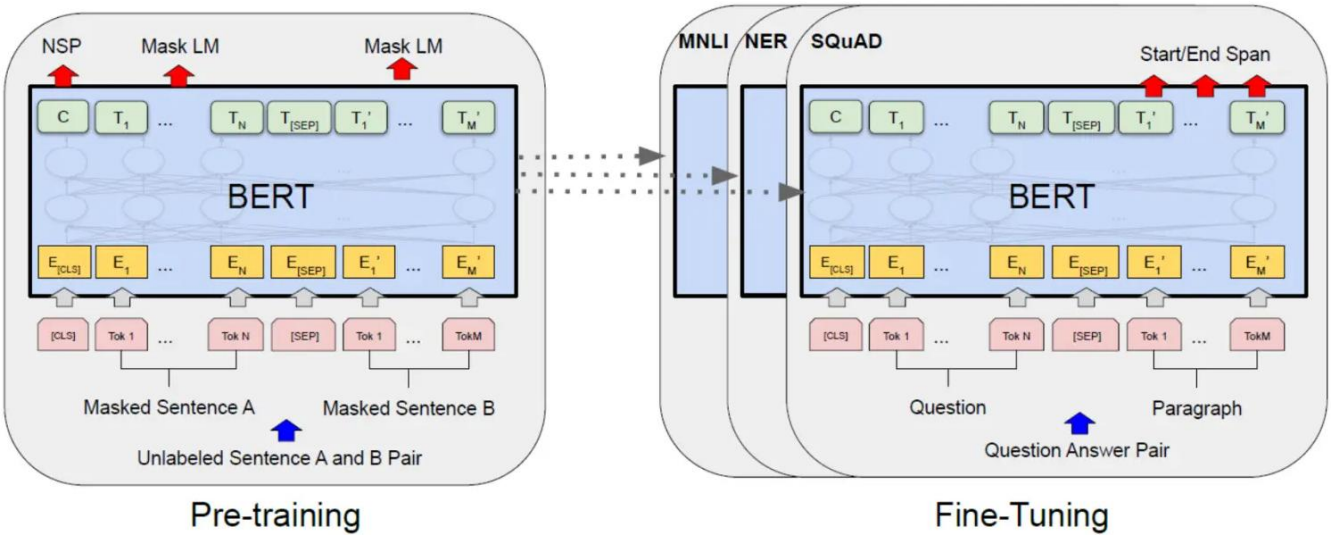
资料来源：公开资料、招商银行研究院

预训练模型的出现为 AIGC 技术带来了颠覆性的进步。尽管过去各种生成式模型层出不穷，但由于任务类型单一、使用门槛高、训练成本昂贵以及内容质量不足，难以满足复杂多变的应用场景。预训练模型，也称为基础模型或大模型，通过在大规模数据集上进行训练，学习到丰富的特征表示，展现出更强的泛化能力和深入的语言理解及内容生成能力。这些模型具备通用特征学习、迁移学习、多任务学习和领域适应等关键特性，显著增强了 AIGC 的通用化能力，使同一模型能够高质量完成多种内容输出任务。通过在特定领域数据上进行微调，模型能够迅速适应并掌握新领域的特定特征，极大提升了其实用性和灵活性。

2018 年，谷歌推出了基于 Transformer 架构的自然语言处理预训练模型 BERT，标志着人工智能领域进入了一个以大规模预训练模型参数为核心的新纪元。BERT 的核心创新在于其双向训练策略，能够同时考虑单词左侧和右侧的

上下文信息，使得其在理解单词含义时更为精准。通过在大量文本数据上的预训练，BERT 学习到了深层次的语言表示，这些表示可以迁移到多种下游 NLP 任务中，如文本分类、问答系统和命名实体识别等。BERT 通过微调（Fine-tuning）进一步适应特定任务的需求，极大地提升了自然语言处理的效果和应用广度。

图 24：预训练模型 BERT 结构图

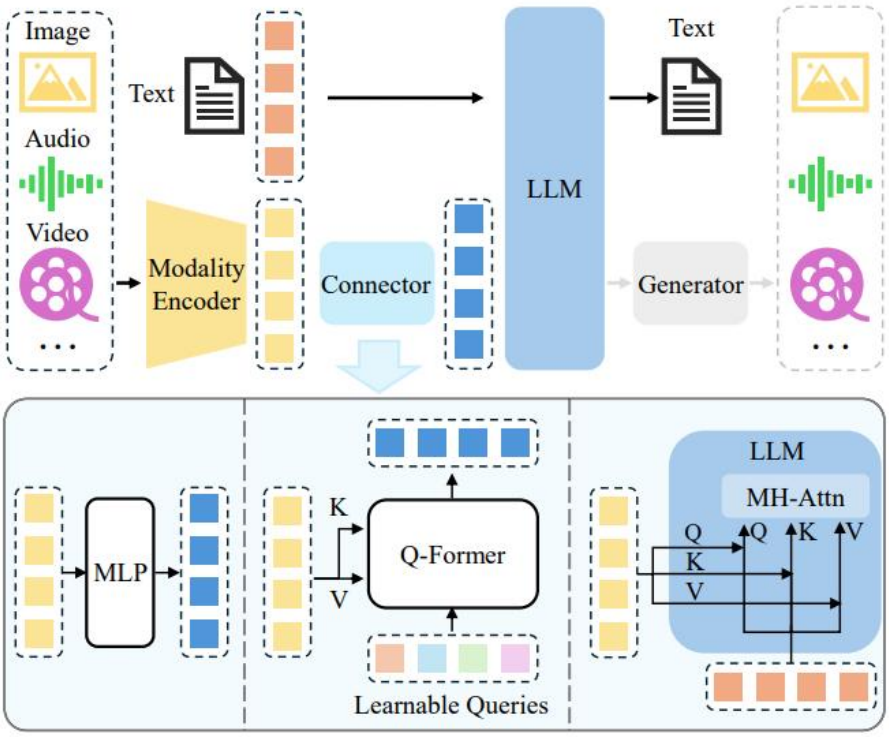


资料来源：谷歌、招商银行研究院

多模态技术的发展推动了 AIGC 内容的多样性，增强了模型的通用化能力。该技术使不同类型的数据（如文本、图像、音频和视频）能够互相转化和生成，从而使 AIGC 模型能够跨模态生成各种类型的内容。CLIP（Contrastive Language-Image Pretraining）模型是 OpenAI 提出的一种典型多模态预训练模型。其核心思想是利用大规模的图像和文本数据进行自监督学习，使模型能够在没有明确标注的情况下理解和关联不同模态的数据。CLIP 能够将图像和文本映射到同一个向量空间，促进了不同模态数据的理解与关联，为文生图、文生视频等 AIGC 应用的快速发展奠定了基础。

一个典型的多模态大型语言模型（MLLM）可以抽象为三个核心模块：预训练的模态编码器（Modality Encoder）、预训练的大型语言模型，以及连接它们的模态接口（Connector）。类比于人类，模态编码器相当于接收和预处理光学/声学信号的人类眼睛和耳朵，而大型语言模型则像是理解并推理处理信号的人类大脑。在这两者之间，模态接口的功能是对齐不同的模态。以 GPT-4V 为代表的多模态大型语言模型在多模态任务中展现出了前所未有的能力。随着技术的不断进步，多模态模型将在更多领域发挥重要作用。

图 25：典型多模态大模型架构示意图



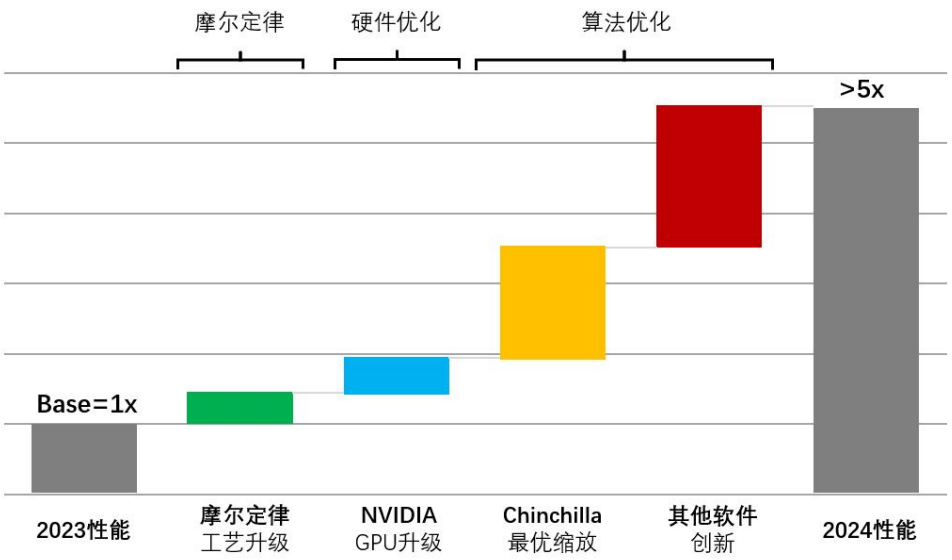
资料来源：IEEE、招商银行研究院

3.2 性能与成本：大语言模型竞争的核心驱动力

大语言模型的竞争主要集中在两个核心要素上：性能和成本。性能决定了模型能够处理的任务复杂度和准确性，而成本则影响模型的商业可行性和普及程度。这两者的平衡将直接影响大语言模型在市场上的竞争力与应用广度。

硬件性能的提升与软件算法的创新共同推动了大语言模型的不断提升。在硬件方面，GPU 性能的增强显著提升了模型的训练和推理能力，得益于半导体工艺的进步和持续的 GPU 设计创新，这使得复杂任务的处理更快速高效。软件方面，创新算法如 Chinchilla 的最优缩放、人类反馈强化学习（RLHF）、推测解码（Speculative Decoding）和 Flash Attention 等，为大模型的发展注入了新的活力。例如，Chinchilla 通过合理分配模型大小和训练数据量，优化了有限计算资源下的模型训练；Llama2 利用 RLHF 方法，确保输出更符合用户期望；推测解码实现了推理速度的显著提升；而 Flash Attention 则通过优化注意力机制，提高了 GPT 模型的训练速度。这些因素的结合使得大语言模型在性能和效率上不断取得突破。

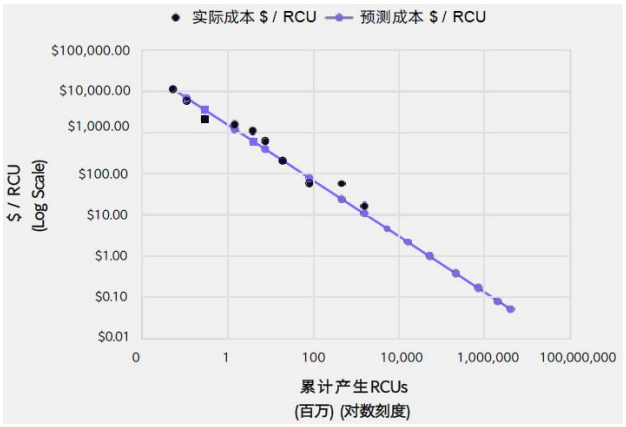
图 26：大模型训练性能不断提升



资料来源：ARK、招商银行研究院

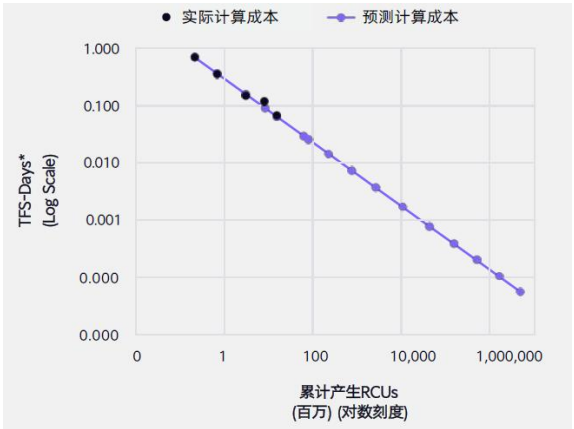
随着训练成本的不断下降，AIGC 的应用正变得越来越普及。赖特定律（Wright's Law）表明，当一种产品的累计产量翻倍时，其单位成本将下降一个固定百分比。在 AIGC 领域，尤其是大模型训练中，GPU 硬件性能的提升和算法优化对成本降低起到了关键作用。根据 ARK 的分析，随着硬件技术的不断进步，AI 相对计算单元（RCU）的成本预计每年将降低 53%，而模型算法的增强预计每年可使训练成本降低 47%。预计到 2030 年，硬件和软件的融合将使 AIGC 训练成本以每年 75%的速度下降。这一显著的成本降低将推动 AIGC 技术的普及与经济性，从而促进 AIGC 的广泛应用和创新。

图 27：AIGC 训练硬件成本趋势



资料来源：ARK、招商银行研究院

图 28：AIGC 训练软件成本趋势



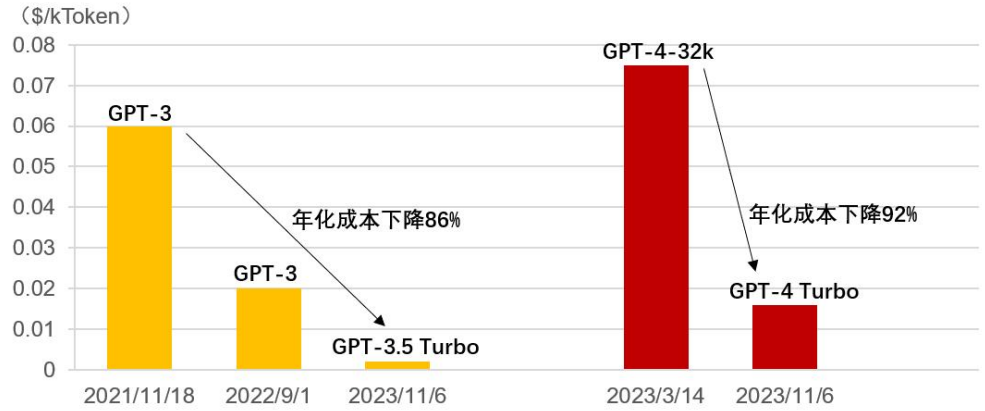
资料来源：ARK、招商银行研究院

随着 LLM 公司之间竞争的加剧，AIGC 的推理成本正迅速降低。AIGC 模型在处理输入和输出时，其计算资源消耗与输入输出的数据量成正比，费用计算基于输入输出的 Token 数量，这种计费方式为不同用户提供了灵活性。

以 OpenAI 为例，在过去两年里，它将 API 访问成本降低了 99%。具体来看，GPT-3 的 API 推理成本从 2021 年的每千 Token 0.06 美元降至 2022 年的 0.02 美元，降幅达 66%。到 2023 年，GPT-3.5 Turbo 的 API 推理成本与 2021 年相比下降了 86%。同时，GPT-4 Turbo 的 API 推理成本与 GPT-4-32k 相比降低了 92%，其成本甚至低于一年前的 GPT-3。

值得注意的是，这一推理成本的降低是在提供更长的上下文、更低的延迟和更新知识截止日期的情况下实现的。微软 CEO 纳德拉认为，与摩尔定律类似，AI 领域也存在 Scaling Law（尺度定律），在 AI 时代，衡量单位是“每美元每瓦特的 Token 数”。这种竞争态势将进一步推动 AIGC 技术的普及与应用。

图 29：GPT API 推理成本快速下降



资料来源：ARK、招商银行研究院

3.3 AIGC 市场快速增长推动多元化商业模式与竞争格局演变

AIGC 大模型公司正在通过多元化商业模式开拓收入渠道，目前主要集中在订阅服务和 API 接入两种模式。

1. **订阅服务**：用户支付月费或年费以享受持续的服务。例如，OpenAI 的 ChatGPT Plus 订阅服务目前每月收费 20 美元，预计在 2024 年底提价至 22 美元。截至 2024 年 9 月，ChatGPT Plus 已拥有 1000 万订阅用户。

2. **API 接入模式：**企业将 API 服务整合至其应用程序中，并根据使用情况付费。OpenAI 的 API 定价因模型和使用情况而异，通常根据输入输出的 Tokens 进行差异化定价。

此外，OpenAI 还与 Microsoft Azure 合作，为大规模企业提供定制化的专用实例，价格层次丰富，满足不同企业的需求。这种多元化的商业模式不仅为 AIGC 公司带来了稳定的收入来源，也使其能够更好地服务于不同类型的用户和市场。

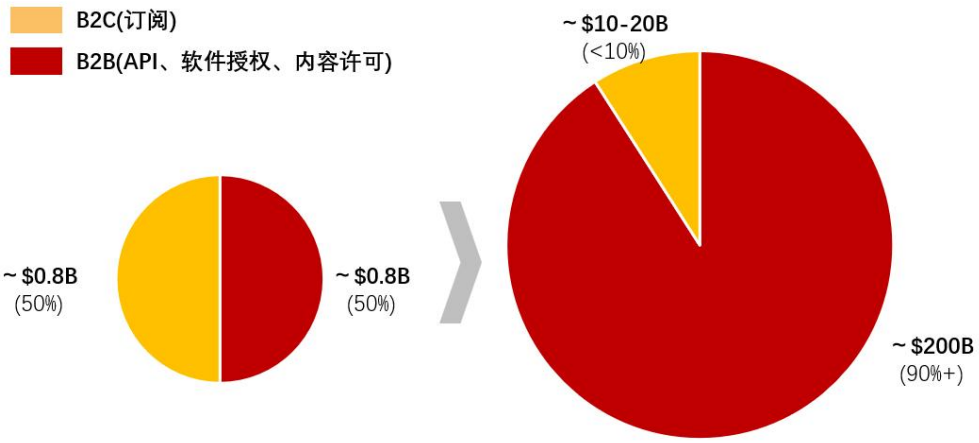
除了订阅和 API 接入，大模型公司还在积极探索其他多种商业模式，包括：

1. **企业定制服务：**大模型公司为特定企业需求提供个性化解决方案，深度集成特定应用场景或行业需求，帮助企业实现更高效的业务流程。
2. **软件授权：**这种模式允许公司出售技术使用权，特别适合那些需要在本地部署解决方案的企业。这使得客户能够根据自身的安全和合规要求来管理和使用模型。
3. **内容许可：**大模型公司与机构合作，获得内容许可以训练模型。这种合作可以增加训练数据的多样性，提高模型的表现。
4. **合作伙伴关系：**建立与大型科技公司的紧密合作关系，例如 OpenAI 与 Microsoft 的合作，涉及技术集成和新产品共同开发。这种合作不仅为大模型公司带来额外收入，还能够提升其技术能力和市场竞争力。

通过这些多元化的商业模式，大模型公司能够更灵活地适应市场需求，拓宽收入来源，提高自身的市场竞争力。

随着越来越多的企业认识到 AIGC 技术的潜力，B2B 市场的需求预计将持续增长。根据 OpenAI 的收入构成，2023 年其 B2C 和 B2B 业务各占一半。预计到 2024 年，OpenAI 的收入将达到约 37 亿美元，2025 年将大幅增至 116 亿美元。这一增长主要得益于 ChatGPT 订阅用户的增加以及企业 API 和定制解决方案的使用。B2C 市场规模预计将达到 100 亿至 200 亿美元，但市场占有率预计不足 10%。B2B 市场规模预计高达 2000 亿美元以上，且市场占有率超过 90%。这一比例显示出 B2B 服务在整体 AIGC 市场中的主导地位。

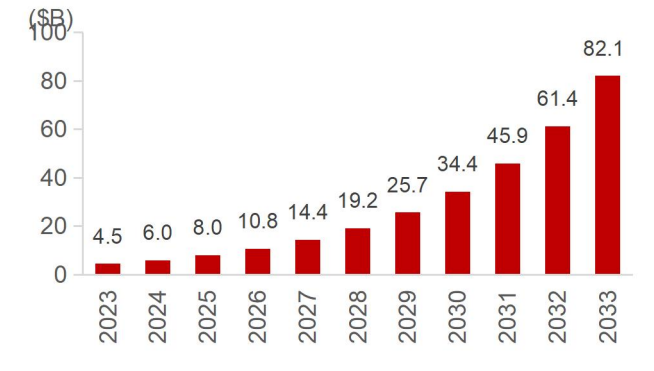
图 30：AIGC 大模型长期潜在市场与收入结构预测



资料来源：Kelvin Mu、招商银行研究院

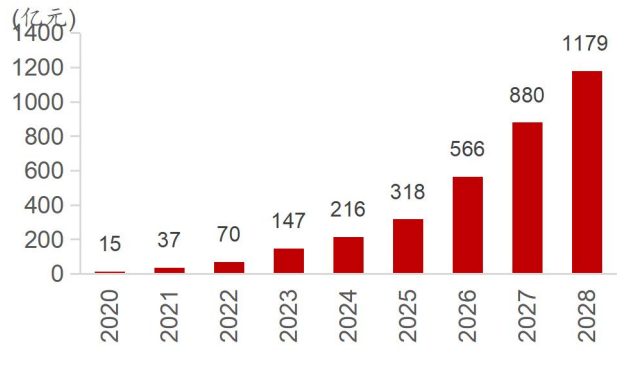
随着 AIGC 技术的快速发展，大模型平台市场正在经历显著增长。AIGC 技术的不断进步和应用领域的持续扩展，促使越来越多的企业采用大模型平台来构建和扩展其应用程序。2022 年底 ChatGPT 的公开发布，成为推动行业增长的重要催化剂。根据 Market.US 的预测，全球大语言模型市场规模将从 2023 年的 45 亿美元增长到 2033 年的 821 亿美元，复合年增长率为 33.7%。中国市场同样展现出强劲的增长潜力。前瞻产业研究院的预测显示，中国大语言模型市场规模将从 2023 年的 147 亿元增长到 2029 年的 1186 亿元，复合年增长率为 41.6%。

图 31：全球大语言模型市场规模预测



资料来源：Market.US、招商银行研究院

图 32：中国大语言模型市场规模预测

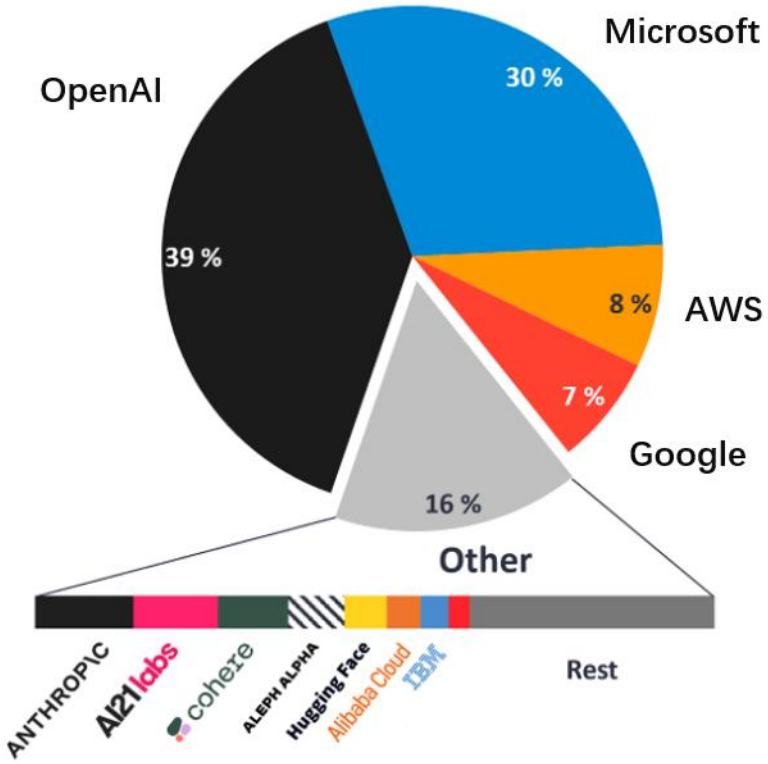


资料来源：前瞻产业研究院、招商银行研究院

在大语言模型市场，OpenAI 凭借其卓越的技术成为市场的领头羊。根据 IoT Analytics 的分析，OpenAI 在推出 ChatGPT 短短两个月内便实现了月活跃用户数突破 1 亿，成为有史以来用户增长速度最快的消费级应用程序。凭借 ChatGPT 的成功，OpenAI 在大语言模型市场中以 39%的市场份额处于领先地位。紧随其后的是科技巨头微软（30%）、亚马逊（8%）和谷歌（7%）。此外，一

些领先的 AI 创业公司，如 AI21 Labs、Anthropic 和 Cohere 等，也占据了一定的市场份额。

图 33：大语言模型市场份额（2023 年）



资料来源：IoT Analytics、招商银行研究院

互联网科技巨头正通过持续的技术创新和产品整合，努力追赶并挑战 OpenAI 在大语言模型市场的领导地位。微软将 OpenAI 的功能集成到其多种产品中，包括 Azure、Office 365 和 Bing。Azure AI 平台提供了强大的工具集，允许客户选择不同的大语言模型，例如 OpenAI 的模型或 Llama 2，并提供定制化的 AI 应用程序，增强了数据安全性。亚马逊的 Bedrock 专注于提供平台服务，支持客户访问多家 AI 公司的大语言模型，帮助他们更加灵活地构建和扩展生成式 AI 应用程序。谷歌的 Gemini 是一系列多模态模型，它被融入谷歌的产品体系，能够处理和组合各种数据类型。此外，谷歌的 Vertex AI 是基于云计算的 AI 平台，融合了最新的技术和能力，可以帮助企业快速实现 AI 应用的开发和部署。

4. 应用层：技术创新推动应用市场发展和传统行业变革

4.1 AIGC 技术加速 ToC 与 ToB 领域的创新与多元化应用

AIGC 技术在面向消费者（ToC）和面向企业（ToB）领域都有广泛的应用场景。随着技术的持续演进和迭代，这些应用场景和商业模式正不断拓展和演变。

在面向消费者领域，AIGC 技术满足了个人的日常生活需求，涵盖了如 Chatbot、社交、游戏、教育和内容创作等多个场景。在社交娱乐方面，AIGC 技术使普通用户能够以较低的门槛参与内容创作，激发创作灵感，用户可以通过 AIGC 创作画作、文本、歌曲等。在教育领域，AIGC 技术被用于开发个性化的学习工具和课程，帮助学生更高效地学习。此外，在搜索引擎和内容推荐方面，AIGC 技术利用自然语言生成和机器学习等技术，快速生成新闻报道和文章，并提供个性化的推荐服务。

在面向企业领域，AIGC 技术为企业客户提供了多种解决方案，帮助提高效率、降低成本、创新产品，并增强市场竞争力。在办公领域，AIGC 技术能够提升工作效率和质量，激发创意和乐趣，创造更便捷、高效和创新的办公体验。在内容生产和媒体方面，AIGC 技术提供高效工具，提升内容产出效率和质量，降低生产成本。在广告营销领域，AIGC 技术通过内容创新、制作成本节约和流程效率提升，推动营销效果的增强。在游戏开发方面，AIGC 技术应用于智能 NPC、场景建模和 AI 剧情等功能，提升游戏的创新性和玩家体验。在药物研发领域，AIGC 技术在辅助诊断和药物研发过程中发挥着重要作用。

表 6：常见的 AIGC 应用场景

场景	功能	典型产品
Chatbot	文本生成、问答推理、摘要生成、翻译	ChatGPT、Kimi、豆包
社交	AI 角色、虚拟 IP、个性化聊天机器人	Character.ai、Talkie、Snapchat
搜索引擎	提供智能、精确和个性化的搜索功能	Bing、Gemini、Apple Intelligence
教育	个性化学习内容生成、技能评估	Quizlet、Bridge-U、DreamBox Learning
内容创作	图像、音乐等内容创作	Sora、Suno、TikTok
游戏	游戏场景、故事、虚拟角色的创建	Ubisoft Entertainment、Epic Games、Steam
办公	文档编辑、数据分析、可视化、AI 助手	Microsoft 365 Copilot
数字设计	偏好分析、设计优化、方案生成	Adobe Firefly、Uizard、Khroma、Designs.ai
金融服务	风险评估、量化交易、柜台业务	BigQuant、RiskLab AI
软件开发	代码生成、修复测试、文档生成	GitHub Copilot、CodeWhisperer、MarsCode

资料来源：招商银行研究院

AIGC 应用产品种类繁多，其中 Chatbot 占据领先地位。根据 AI 产品榜的数据，全球市场上，ChatGPT 的月访问量已突破 30 亿次，使其成为全球第十一大网站。从产品分布来看，AIGC 赛道涵盖了多个领域，包括 Chatbot、内容创作、翻译、搜索、教育和知识管理等。在国内市场，前十的应用主要以 Chatbot 为主，同时 AI 搜索的占比也在逐步提升。

表 7：全球 AIGC 应用排名（2024 年 9 月）

排名	产品名	分类	月访问量	月下载量
1	ChatGPT	AI 聊天、文本生成	30.23 亿	3.98 亿
2	Canva	设计辅助、AI 生成图像、创意工具	7.13 亿	1.02 亿
3	Deepl	翻译工具、即时翻译	1.95 亿	466.58 万
4	夸克	AI 搜索、AI 智能工具	6217.67 万	736.48 万
5	Kimi	长文本处理、智能助手、对话服务	2363.27 万	127.05 万
6	Notion AI	项目管理、团队协作、知识管理	1.52 亿	363.48 万
7	字节豆包	AI 对话工具、个性化智能体	1300.82 万	760.4 万
8	文心一言	知识增强、创作辅助、个性化推荐	1979.74 万	91.53 万
9	百度文库	AI 生成文档、AI 创作、文档搜索	2691.22 万	84.53 万
10	Poe	AI 聊天机器人	3050.37 万	-
11	Shutterstock	库存图像、照片、视频、音乐	6549.31 万	13.29 万
12	Salesforce AI	应用开发、CRM、数据集成	1.01 亿	24.53 万

资料来源：AI 产品榜、招商银行研究院

4.2 AIGC 技术驱动电子设备革新，大模型引领手机、汽车与机器人智能化创新

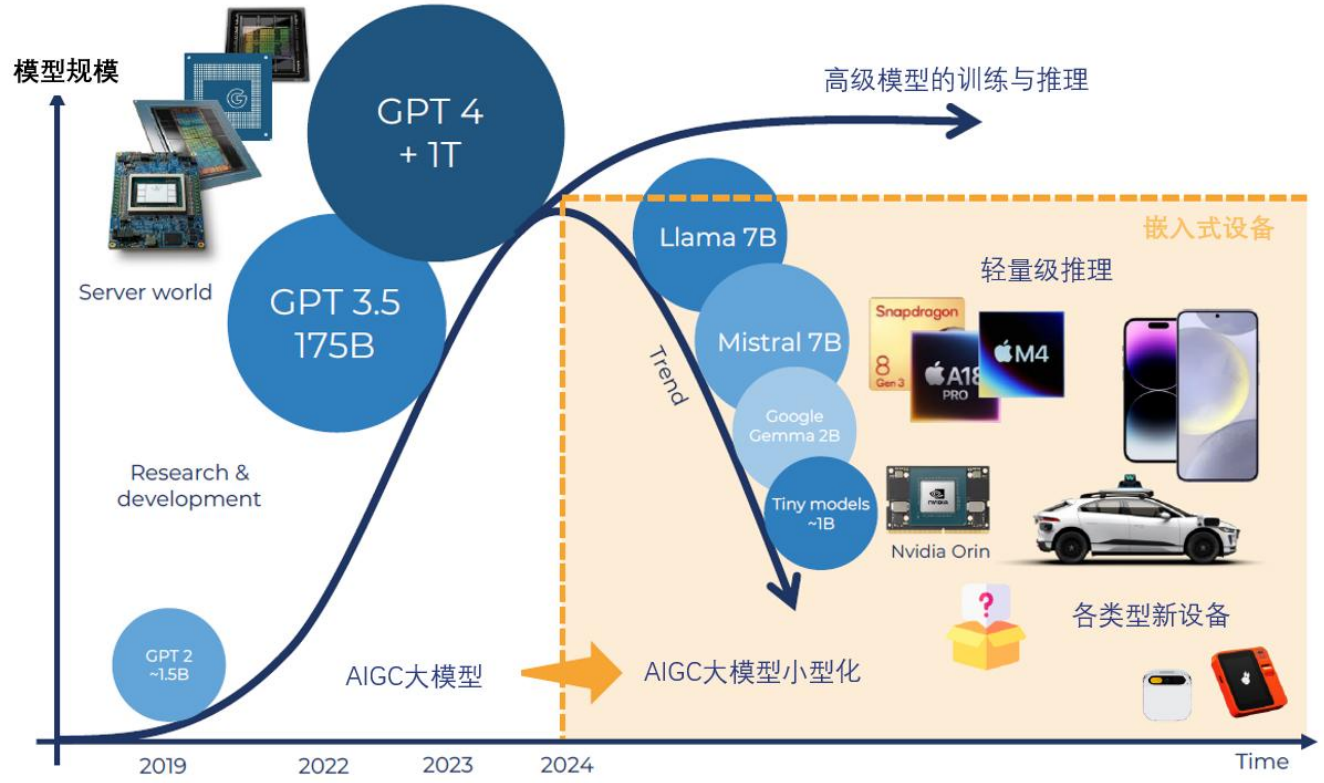
AIGC 技术正引领一场新的科技革命，大模型在传统硬件设备中的应用日益广泛，为智能手机、汽车、机器人等多个产业链带来了全新的机遇。这些技术的融合不仅提升了设备的智能化水平，还推动了各行业在功能和效率上的创新，为用户提供更加个性化和高效的体验。

智能手机通过引入大模型资源，显著增强了用户的操作体验。利用 API 模式，智能手机可以集成 ChatGPT 等先进的大模型，实现个性化内容创造、智能语音助手和个性化推荐等方面的重大突破。AIGC 技术能够根据用户的个性化

需求，自动生成文本、图片、视频等多种内容，广泛应用于社交媒体、个人表达和商业领域。借助 API 与大模型的连接，智能语音助手变得更智能、更人性化，能够理解上下文，提供更流畅自然的对话，并执行更复杂的任务。此外，AIGC 还可以通过 API 实时生成个性化内容推荐，例如个性化新闻、主动购物推荐和应用建议，为用户带来更加丰富和个性化的服务体验。

AIGC 技术正以其革命性的力量推动手机硬件和操作系统重构，引发手机产业链生态的深刻变革。随着手机算力的显著提升和大模型的压缩与优化，操作系统有望采用边缘计算与本地推理相结合的创新方式，在高效手机上实现轻量级推理，同时将大部分计算任务放在云端。未来，更多的应用程序将通过调用 AIGC API 来实现内容生成、推荐系统和交互功能，从而减少传统手动编写内容的依赖，使开发者能够更专注于核心业务逻辑的构建。这种变革将带来更个性化的交互、更智能的任务管理以及实时生成的个性化内容。操作系统能够生成适应用户个人风格和需求的 UI 设计、动态壁纸和主题，可以根据用户的使用习惯自动安排任务，提供更加定制化服务。

图 34：AIGC 推动大模型与电子设备智能化升级

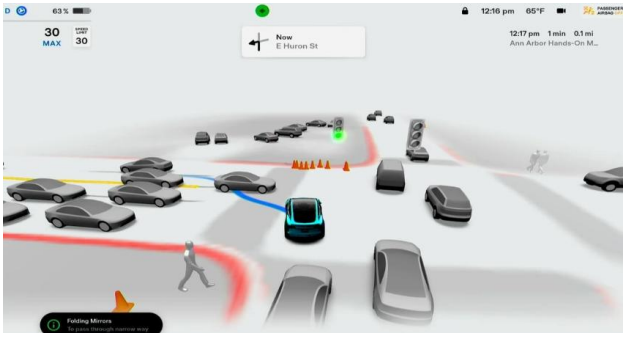


资料来源：Yole、招商银行研究院

在汽车领域，AIGC 技术正在推动自动驾驶技术的发展。AIGC 涵盖了多个关键方面，包括训练数据生成、情境理解、路径规划、实时学习和用户交互等。通过这些技术的融合，自动驾驶系统能够更有效地应对复杂的道路环境，理解驾驶者的需求，并不断提升其智能化水平，为自动驾驶技术的发展提供重要支撑。特斯拉的 FSD 12 在感知能力、决策算法和用户交互等方面取得了显著提升，能够在特定地区和情况下支持更高级的自动驾驶功能。这为未来的完全自动驾驶奠定了更坚实的基础，使汽车不仅能够自动驾驶，还能更好地适应驾驶者的个性化需求和动态道路状况。

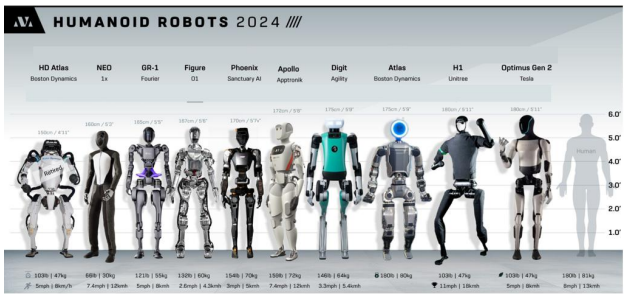
AIGC 技术加速了人形机器人在智能化和多样化发展上的进步。借助 AIGC，人形机器人不仅能执行简单任务，还能主动学习和理解用户需求，提供个性化服务。AIGC 使人形机器人能够自然地理解和生成对话内容，模拟人类沟通方式，从而在交流中为用户提供更真实和智能的反馈。此外，AIGC 还使人形机器人具备生成和识别视觉内容的能力，这提高了它们在教育、娱乐、医疗等领域的视觉理解能力。人形机器人能够通过 AIGC 不断从环境中学习，并生成复杂场景的应对策略，例如在制造业中优化生产流程，或在医疗领域辅助医生进行诊断。这种动态适应能力使人形机器人在多变的环境中更加灵活。随着 AIGC 的发展，人形机器人不仅能够服务于多种场景，还能适应不断变化的需求，推动其从单一功能向智能化、个性化助手的演变。

图 35：特斯拉 FSD 自动驾驶路径规划



资料来源：Tesla、招商银行研究院

图 36：2024 年全球主流人形机器人



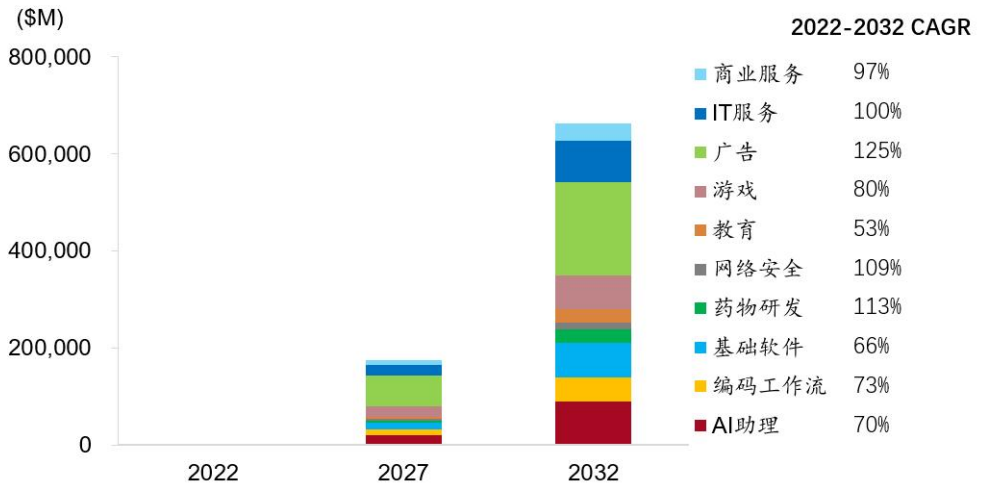
资料来源：MADE VISUAL、招商银行研究院

4.3 AIGC 应用市场正处于发展初期，竞争格局多元化且持续演变

根据彭博情报（Bloomberg Intelligence）的预测，随着各类 AIGC 应用的爆发式增长，AIGC 应用市场的规模预计将从 2022 年的 18.60 亿美元增长到 2032 年的 6618.14 亿美元，年均复合增长率达到 80%。在这一市场中，AI 广告预计将占据最大市场份额，而药物研发、网络安全和 IT 服务市场的增速最快。同时，AIGC 在科技领域的投入也将显著增加。信息技术硬件、软件、服

务及广告等领域的 AIGC 支出预计将从 2022 年总支出的 1%增长到 2032 年的 12%。这一增长反映了 AIGC 在各个行业中的广泛潜力，尤其是在加速产品开发、自动化流程以及增强决策支持方面的应用。

图 37：2022-2032 年 AIGC 应用市场规模

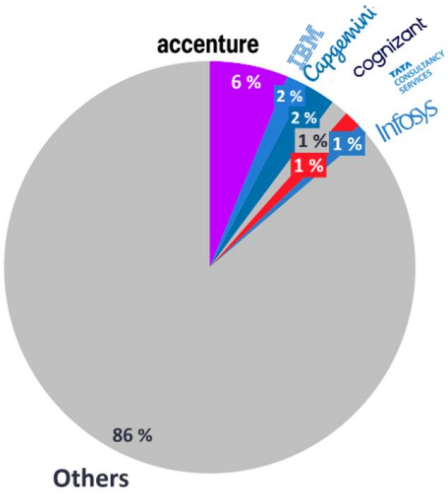


资料来源：Bloomberg Intelligence、招商银行研究院

AIGC 市场正处于一个充满机遇与挑战的初期阶段。尽管近年来 AIGC 技术取得了显著进步，应用场景不断增加，但整体市场仍在探索和形成中。随着新模型和应用层出不穷，企业和开发者不断寻求更高效、更智能的解决方案，以满足不断变化的市场需求。AIGC 的应用场景从内容创作、营销扩展到医疗、教育等多个领域，不同的行业对 AIGC 的需求和实现方式各不相同，企业正在探索最佳的整合方式。除了大型科技公司，许多初创企业也纷纷进入这一领域，推出各种创新的 AIGC 应用，进一步加剧市场竞争。随着技术的成熟和应用的深入，预计未来几年 AIGC 应用市场将迎来更大的发展和变革。

AIGC 应用市场正呈现出多元化的竞争格局，发展态势持续演变。随着越来越多的初创企业和中小型公司进入市场，这些企业致力于推出针对特定行业的 AIGC 产品，如医疗、广告、金融和教育等领域的定制化应用工具。根据 IoT Analytics 的分析，2023 年埃森哲以 6%的市场份额在 AIGC 应用市场保持领先地位，并将 AIGC 技术整合到其咨询服务中，帮助客户实现数字化转型。IBM、Capgemini 和 Cognizant 紧随其后，展现了这些公司在推动 AIGC 技术应用方面的持续努力。未来，随着市场需求的不断增长，这些公司的竞争格局可能会进一步变化，带来更多创新和机遇。

图 38：AIGC 应用市场份额（2023 年）



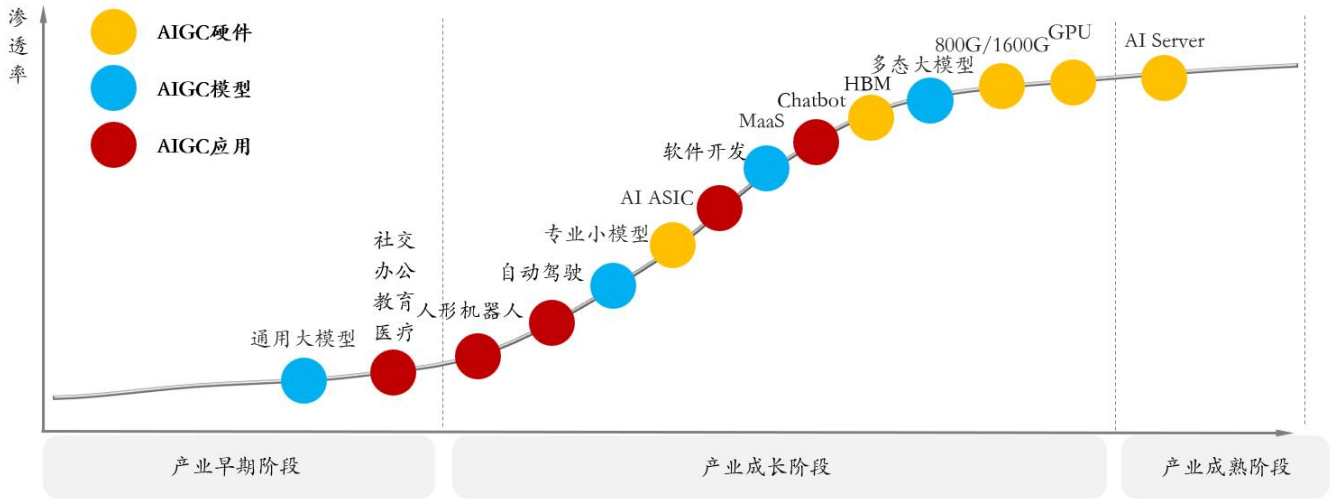
资料来源：IoT Analytics、招商银行研究院

5. 业务建议与风险提示

5.1 业务建议

（本部分有删减，招商银行各行部请参照文末联系方式联系研究院）

图 39：AIGC 产业链布局策略



资料来源：招商银行研究院

5.2 风险提示

（1）**伦理道德的风险。**可能加剧社会不平等，侵犯隐私，存在算法偏见和道德争议，若处理不当可能引发法律问题。

（2）**技术缺陷的风险。**算法和模型可能存在缺陷，导致生成内容质量低或被恶意利用，进而造成信息泄露及人类对技术的过度依赖。

（3）**监管与法律的风险。**各国可能出台新的监管政策，企业需时刻关注并遵循，以避免法律风险。

（4）**商业化不确定的风险**。技术瓶颈和应用局限可能影响商业化进程，给企业带来风险。

（5）**市场竞争加剧的风险**。行业内竞争激烈，技术更新迅速，可能影响行业的健康发展。

（6）**宏观经济波动的风险。**宏观经济波动可能影响投资决策和市场需求，从而影响 AIGC 行业的整体发展。

免责声明

本报告仅供招商银行股份有限公司（以下简称“本公司”）及其关联机构的特定客户和其他专业人士使用。本公司不会因接收人收到本报告而视其为本公司的当然客户。本报告仅在相关法律许可的情况下发放，并仅为提供信息而发放，概不构成任何广告。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议。在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。

本报告的信息来源于已公开的资料，本公司对该等信息的准确性、完整性或可靠性不作任何保证。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。本公司不保证本报告所含信息保持在最新状态。同时，本公司对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。本公司可能采取与报告中建议及/或观点不一致的立场或投资决定。

市场有风险，投资需谨慎。投资者不应将本报告作为投资决策的唯一参考因素，亦不应认为本报告可以取代自己的判断。在决定投资前，如有需要，投资者务必向专业人士咨询并谨慎决策。

招商银行研究院

地址 深圳市福田区深南大道 7088 号招商银行大厦 16F（518040）

电话 0755-22699002

邮箱 zsyhyjy@cmbchina.com

传真 0755-83195085

更多资讯请关注招商银行研究微信公众号
或一事通信息总汇