芯跑资本
iRUN VENTURE CAPITAL
立德家国使命感 点燃产业满天星

芯跑资本丨月报资讯

2023 年 11 月，第 047 期：AIGC 带动产业升级与相关行业研究（一）

AIGC 带动产业升级与相关行业研究（一）

芯跑资本研究部 2023.11

前言：2022 年 11 月上线的 AIGC 应用 ChatGPT，凭借其在语义理解、文本创作、代码编写、逻辑推理、知识问答等领域的卓越表现，以及自然语言对话的低门槛交互方式，迅速获得大量用户，于 23 年 1 月突破 1 亿月活，打破此前消费级应用的增速记录。

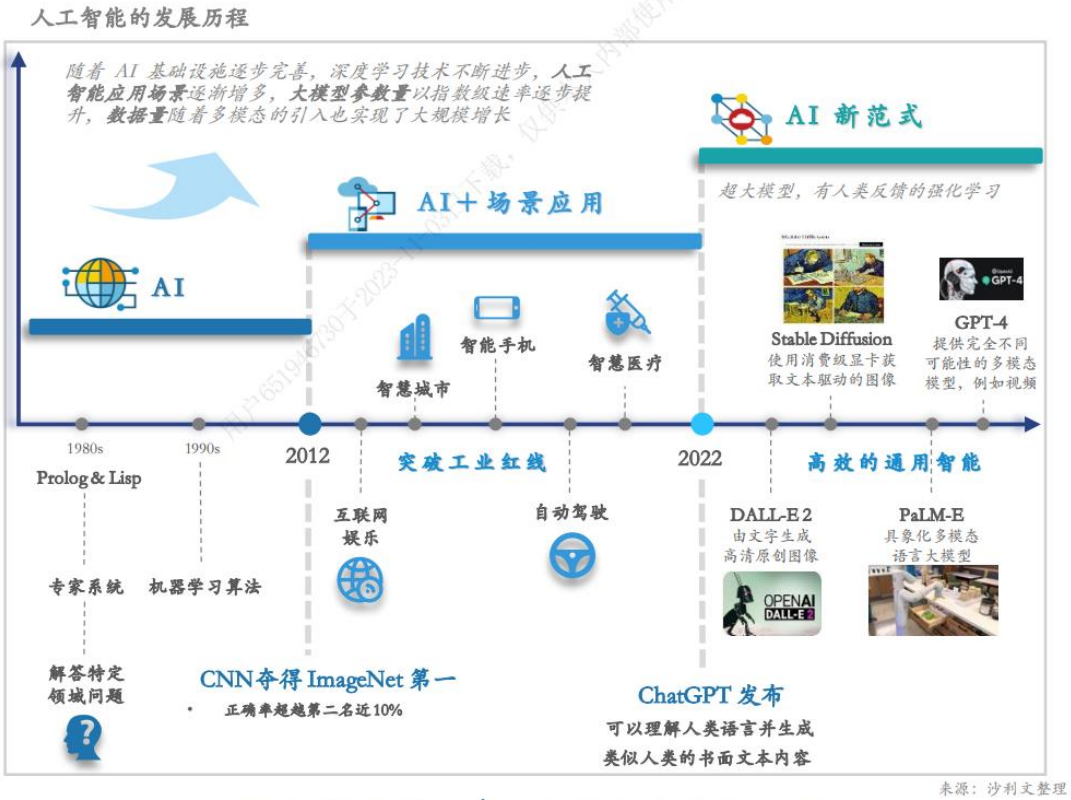
ChatGPT 等 AIGC 应用在多个领域的问题解决能力已超出一般人类水平，微软称其在 GPT-4（ChatGPT Plus 背后运行的大模型）中看到了 AGI（通用人工智能）的雏形。

大众的生活工作日常出现了 Midjourney 等新形态的各类 AIGC 应用，各行业的智能化升级也看到了新的可能性，“AI 产业”与“产业 AI”的想象空间进一步拓展。

一、 AI 的发展历程

1、近 20 年的标志性事件

AIGC (AI-Generated Content) 指利用人工智能技术 (生成式 AI 路径) 来生成内容的新型内容生产方式。AIGC 背后的人类对于通用人工智能（AGI）的追求。自从 1956 年的达特茅斯会议上，“人工智能”的概念被首次提出，六十多年以来，历经逻辑推理、专家系统、深度学习等技术的发展，人工智能迎来了里程碑式的进步。

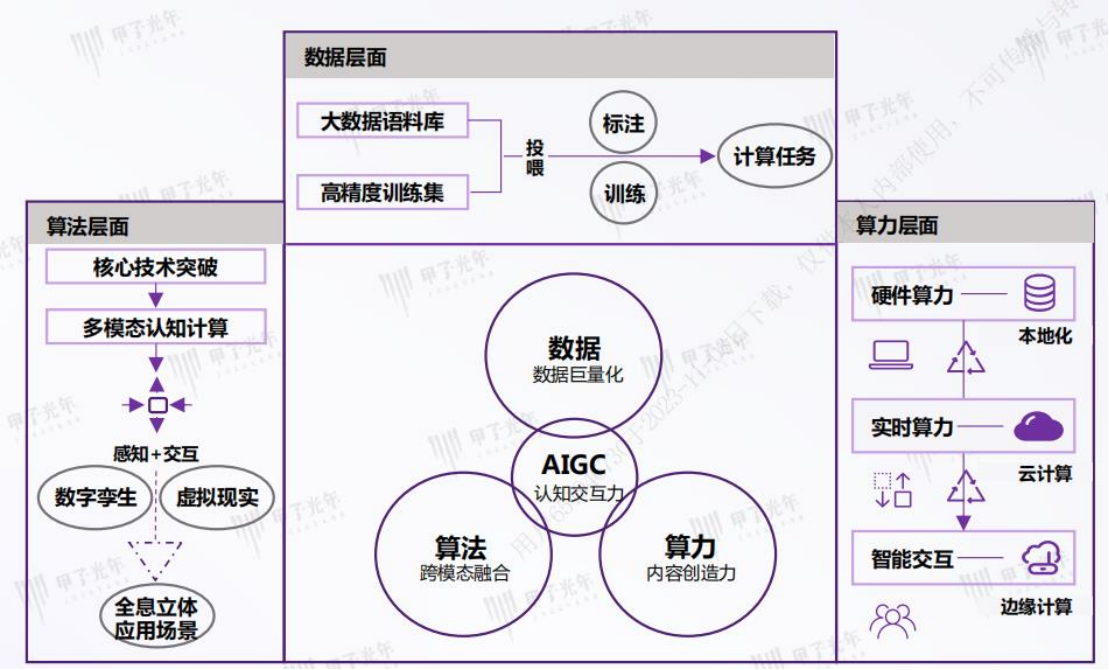


随着 AI 基础设施逐步完善，深度学习技术不断进步，人工智能应用场景逐渐增多，过去模型参数规模和数据量也实现了大幅度增长，为 NLP、CV 等领域带来更强大的表达能力和性能。人工智能发展历程中主要有两大里程碑：

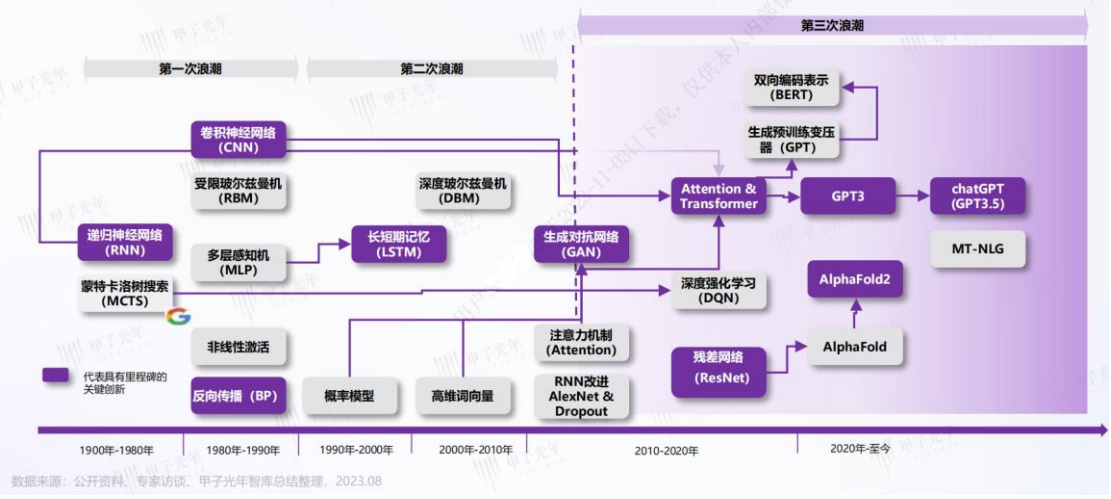
里程碑一：2012 年 CNN 获得 ImageNet 第一，标志着机器视觉识别能力开始逐渐超越人眼识别准确率，开启了人工智能革命。随着深度学习技术不断突破，诞生了一批“AI + 场景应用”的专属模型，但是整体研发成本比较高、研发时间比较长。

里程碑二：2022 年 ChatGPT 的出现，掀起了又一波人工智能发展热潮，以大模型 + RLHF 为核心的技术落地意味着人工智能开启 AI 新范式。人工智能相关产业开始基于强大的基模型进行发展，通过人类反馈和强化学习不断解锁基模型的能力，以解决海量开放式任务，带来了新的研究范式。

在人工智能技术突破里程碑的背后，是数据、算法、算力三个维度的突破。



互联网以及移动互联网的发展产生了大量可以被标注的数据库，云计算的发展不断突破算力集群的上限，在此基础上了，跨模态的算法也得到了不断的突破。

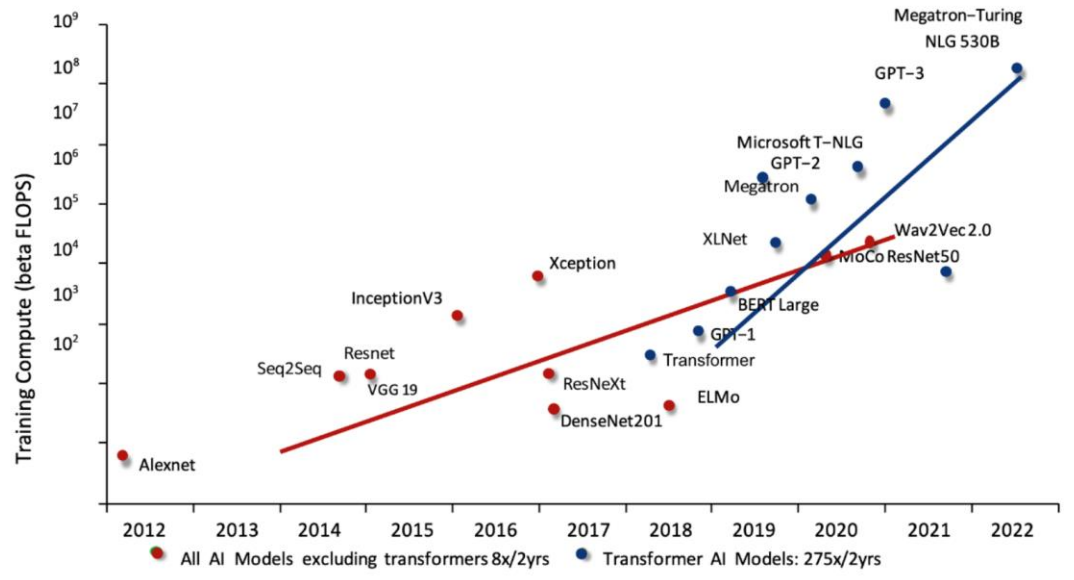


而火爆的 ChatGPT 则体现了人工智能的暴力美学：超大规模及足够多样性的数据、超大规模的模型、充分的训练过程，产生“涌现现象”。（**在复杂系统学科的定义中，当一个复杂系统由很多微小个体构成，这些微小个体凑到一起，相互作用，当数量足够多时，在宏观层面上展现出微观个体无法解释的特殊现象，就可以称之为“涌现现象”**）

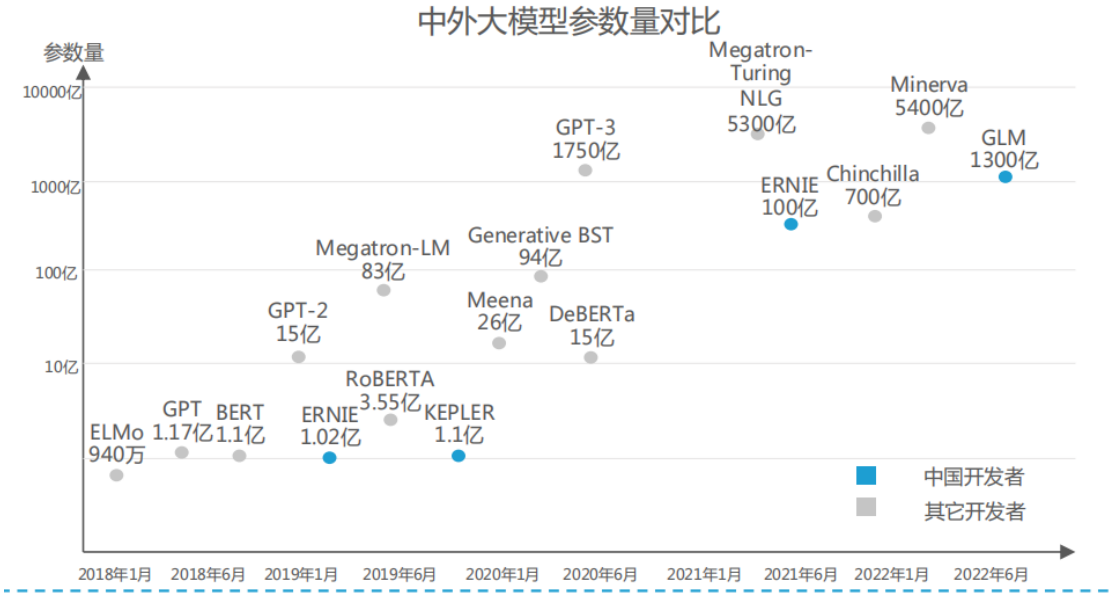
自 2018 年 OpenAI（未上市）发布了包含 1.17 亿参数的第一代 GPT（Generative Pre-trained Transformer）模型以来，**每一代 GPT 模型的迭代都伴随着参数量的飞跃。**一众中外的科技巨头们也不甘示弱，包括 Google、Meta、百度等纷纷发布了 PaLM 、LaMDA 、Llama、文心一言等为代表的大语言模型。

2020 年 1 月，OpenAI 团队论文《Scaling Laws for Neural Language Models》提出“缩放定律”（Scaling Laws），即大模型表现伴随模型参数量、数据集大小和计算量增长而增长，他们于 2023 年 5 月也再次强调，目前缩放定律仍未出现瓶颈。但我们也看到，谷歌在今年 5 月的 I/O 大会里发布的新一代 PaLM 大模型，PaLM2，就是通过算法上的改进达到训练数据增加为上一代 PaLM（7800 亿 tokens）的约 5 倍，达到 3.6 万亿个 tokens，但参数量为 3400 亿，小于 PaLM 的 5400 亿。

图表 ： AI 训练对算力的需求成倍上涨，尤其是 Transformer 相关模型



注：不同颜色代表不同模型种类
资料来源：英伟达官网、华泰研究



算法的发展及迭代极大地拉动了算力的需求，从 2022 年底，随着 ChatGPT 成功带来大规模参数通用大模型相继发布。这些大模型的训练需要千亿、甚至万亿级参数，以及上千 GB 的高质量数据，大模型的训练迭代将极大地拉动了智能算力的需求。

伴随 AI 技术升级和大模型成熟，AI 绘画与 ChatGPT 的成功破圈，生成式 AI 技术迎来发展拐点，行业关注度大幅提升。生成式 AI 是指基于大模型、生成对抗网络 GAN 等人工智能技术，通过已有数据寻找规律，并通过适当的泛化能力生成相关内容的技术，可生成如图像、文本、音频、视频等原创内容变体。例如，以 ChatGPT、Midjourney、文心一格、商汤商量、Codex 为代表的生成式 AI 应用拥有文本语言理解能力、涌现能力以及思维链推理能力，能够完成文学创作、新闻写作、数理逻辑推算、代码生成、图片生成等多项任务。目前，国内电商、游戏、文娱、设计等行业正在积极使用相关的生成式 AI 应用来提高自身工作效率，尤其以文生图应用为主。

资料来源:The information 官网

2、近 10 年的国内典型人工智能企业发展情况

AI 技术的发展给产业带来了非常多的可能性，在国内有两家非常有代表性的公司，分别是科大讯飞和商汤，他们的发展路径、成长逻辑和商业模式都能够给产业一些启示。

A、科大讯飞

科大讯飞股份有限公司于 2008 年 5 月于深交所上市，设立以来一直以智能语音技术产业化为主要发展方向，是以中文为主的智能语音核心技术提供商及应用方案提供商。智能语音技术简单来说就是让计算机、智能仪表、手机甚至家电和玩具等都能像人一样“能听会说”的技术。智能语音技术主要包括语音合成技术、语音识别技术、语音评测技术等。

科大讯飞的业务范围由 2010 年的“智慧城市+运营商”成长为业务横跨七大行业、开放平台超 587 项 AI 能力赋能千行百业的 AI 巨头：

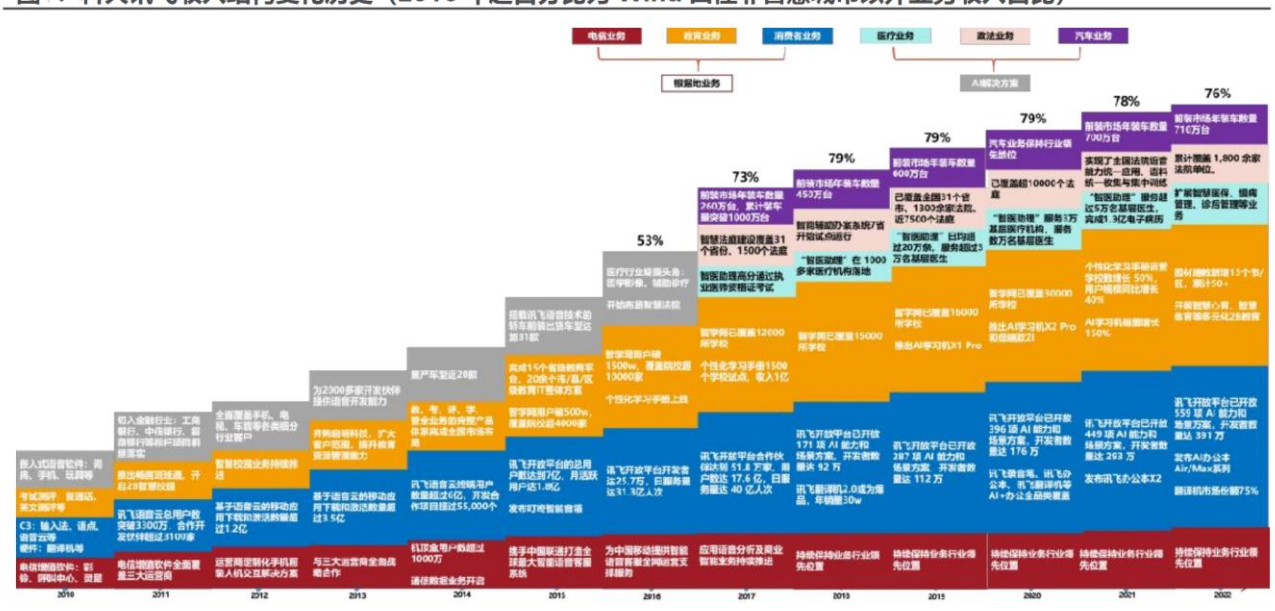
1）2010 年公司收入来源于智慧城市业务，和以电信运营商为主的语音支撑软件以及语音行业应用产品/系统；

2）2013 年公司收购启明科技，与前期积累的智慧教育资源形成互补作用，扩大智慧教育业务范围，同年讯飞嵌入式语音软件业务扩张服务范围，为 2000 多家开发伙伴提供语音技术；

3）2015 年公司全赛道布局战略初见成效，2G 智慧教育完成 15 个省级教育平台、20 余个市级教育 IT 整体解决方案，2B 智学网覆盖院校超 4000 家，同时讯飞成功汽车业务前装车型达到 31 款，牢牢掌握行业市场占有率第一；

4）2023 年上半年，公司各赛道示范验证持续显现，已经在教育、医疗、智慧城市、企业数智化转型等重点赛道构建起可持续发展的“战略根据地”，并在 AI 学习机、智能办公本、智能录音笔、翻译机、智能助听器等 C 端智能硬件产品上形成了领先的品牌和可持续流水型收入的同时，讯飞开放平台为开发者提供 587 项 AI 能力及一站式人工智能解决方案，赋能千行百业。

图 ：科大讯飞收入结构变化历史（2016 年起百分比为 Wind 口径非智慧城市以外业务收入占比）



资料来源：wind，民生证券研究院

在企业发展过程中，科大讯飞不断积累其数据、算法、平台和算力的行业能力。

公司与中国科协、中国科学院合作，讯飞可以用各类科学资料进行训练。此外，在各个垂直行业，科大讯飞有针对性的数据库，如教育领域 G 端项目目前已累计在 50 多个市、区（县）级应用，B 端讯飞课后服务业务已覆盖超 300 区县、12,000 余所学校，平台市场占有率继续保持第一，拥有国内顶级的教育数据资源；医学领域拥有基于从人民卫生出版社的从本科到博士的教材，以及几百万份高水平的电子病历持续训

练迭代；司法领域，全国 26 个省市区，累计覆盖 1,800 余家法院单位接入讯飞 AI 司法平台，持续提供高质量案例数据；智慧城市领域，在江西新余，科大讯飞承建了全省第一个综合人口数据库，共汇聚约 4 亿条人口数据，基于该部分数据以及政务数据，在全省率先开展了普惠金融创新数字应用试点。在严格遵守适用法律法规前提下，在多年认知智能系统研发推广中积累了超过 50TB 的行业语料和每天超 10 亿人次用户交互的活跃应用，为训练实现达到专家水平的大模型提供了海量行业文本语料和用户反馈数据。

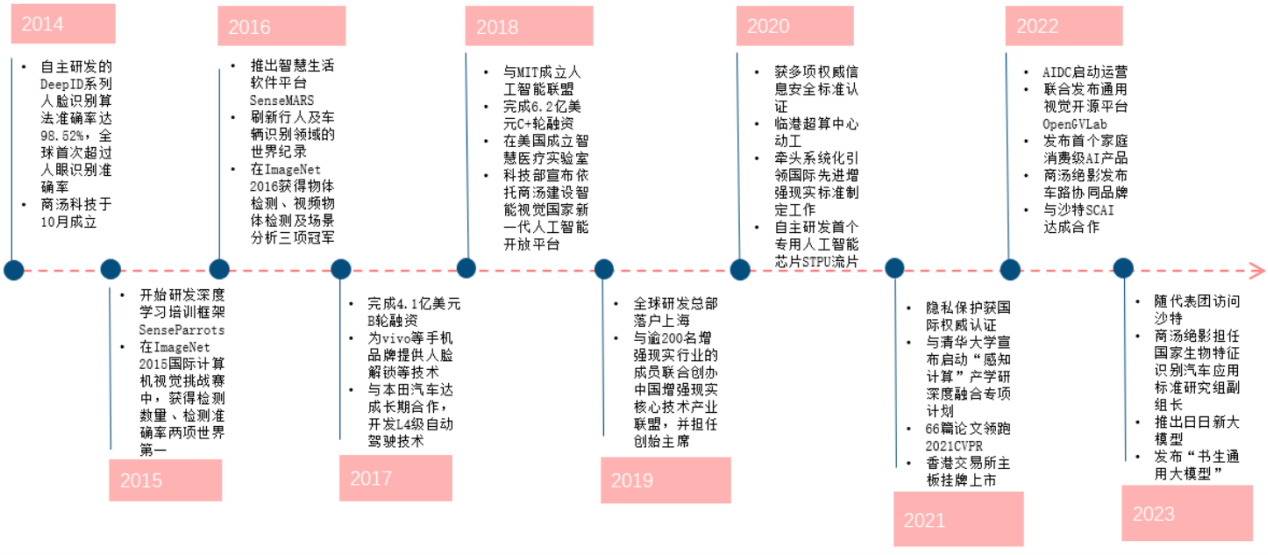
公司拥有三大国家级 AI 平台和顶级科学家团队，2014 年起实行讯飞超脑计划深耕 Transformer 等深度神经网络算法，积累了扎实的 AI 算法和技术资源，并在近期获得工信部认可，担任工信部人工智能关键技术和应用评测重点实验室大模型工作组副组长单位和国家人工智能标准化总体组大模型专题组联合组长单位。

公司与华为昇腾合作，实现从算力底层硬件基础设施到大模型再到大模型应用开发平台全生命周期覆盖，在深度优化下华为 GPU 可对标英伟达 A100。此外，公司于 2009 年开始算力基础设施建设，目前已建成 4 城 7 中心深度学习计算平台，算力不仅完全满足 AI 算法模型训练，还能够覆盖面向开放平台数百万开发者和其他行业伙伴提供相关 AI 服务的需求。

B、商汤

商汤科技于 2014 年正式成立，拥有深厚的学术积累并长期投入于原创技术研究，不断增强行业领先的多模态、多任务通用人工智能能力，涵盖多个关键技术领域。2015 年开始研发深度学习培训框架 SenseParrots，并在 ImageNet2015 国际计算机视觉挑战赛中获得检测数量、检测准确率两项世界第一。2016 年推出智慧生活软件平台 SenseME 和 SenseMARS，赋能超过 450 百万部智能手机及超过 200 个手机应用程序。2017 年开始研发城市方舟平台，并与本田汽车合作开发 L4 级自动驾驶技术。2018 年开始研发专用的人工智能芯片，2019 年开始研发人工智能传感器，2020 年完成首个专用人工智能芯片 STPU 流片，并开始上海临港 AIDC 的施工，2021 年于香港交易所主板挂牌上市。2022 年 AIDC 启动运营，联合发布通用视觉开源平台 OpenGVLab 及发布首个家庭消费级 AI 产品。2023 年商汤绝影担任国家生物特征识别汽车应用标准研究组副组长，商汤推出日日新大模型及书生通用大模型等。

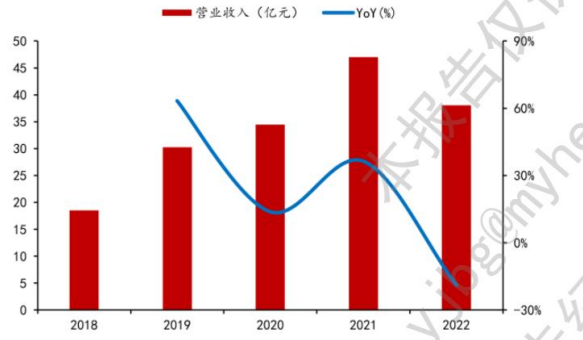
图 ：商汤发展历程



资料来源：公司招股说明书，公司公告，中信建投证券

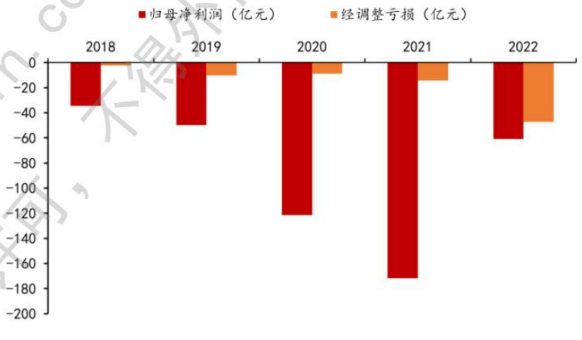
商汤科技在发展中，吸引了众多的资金关注，也是中国创投发展史上少有的吸金怪兽。在上市前，商汤从软银、阿里巴巴、IDG 等多家资方的 12 轮融资中，拿到 52.2 亿美元融资。但相对的，企业的“烧钱”速度也是惊人的，从 2018 年至 2022 的 5 年时间里，商汤分别亏损 34.33 亿元、49.68 亿元、121.58 亿元、171.77 亿元和 60.9 亿元，累计亏损达 438.26 亿元。

图表 ：商汤科技营业收入及增速



资料来源：公司公告，国联证券研究所整理

图表 ：商汤年度亏损和经调整亏损



资料来源：公司公告，国联证券研究所整理

商汤的市值也出现了较大的波动，从市值超 3000 亿，到目前的 500 亿，可以反应出资本市场对 AI 类公司发展预期的降低。

截至 2021 年 6 月 30 日，商汤已建立 23 个超级计算集群，拥有超过 20000 个 GPU，总算力高达每秒 1.17 百亿亿次浮点运算。AIDC 通过其大规模数据处理及高性能计算的能力，为研发提供支持。如今，作为 SenseCore 的算力基座，AIDC 基于 2.7 万块 GPU 的并行计算系统实现了 5.0 exaFLOPS 的算力输出，可支持最多 20 个千亿参数量超大模型（以千卡并行）同时训练。

图 ：商汤科技人工智能计算中心 AIDC

查看图表详情

数据来源：公司官网，中信建投证券

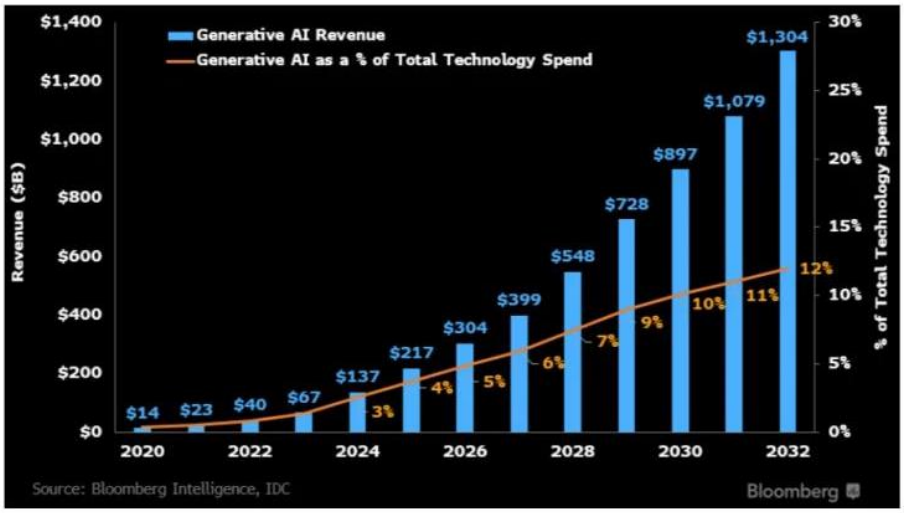
2023 年至今，公司投入了约 1 万张 GPU 在大语言模型研发上，2023 年 8 月推出的新模型 InternLM-123B 模型参数量达到 1230 亿，同时在全球 51 个知名评测集上测试成绩排名全球第二，超越 GPT-3.5 以及 Meta 最新发布的 LLaMA-2-70B。

2023 年至今，公司投入了约 2000 张 GPU 在文生图模型研发上，2023 年 7 月推出的秒画 3.0 版本模型参数量达到 70 亿，并且在 COCObenchmark 上超过了 Imagen 与 DALL E2.出图效果国内领先，全球跻身前三。

二、 AIGC 带动产业升级及行业结构

1、AIGC 带来的产业升级机会

随着 Google 的 Bard 和 OpenAI 的 ChatGPT 等消费者生成式 AI 程序的涌入，生成式 AI 市场即将爆发，在未来 10 年内从 2019 年的 400 亿美元市场规模增长到 1.3 万亿美元。根据彭博资讯 (BI) 的最新报告，到 2022 年。BI 研究发现，在近期培训基础设施的推动下，增长可能以 42% 的复合年增长率扩大，并在中长期逐渐转向大型语言模型 (LLM)、数字广告、专业软件和服务的推理设备。



BI 估计，到 2032 年，生成式 AI 的影响力将从占 IT 硬件、软件服务、广告支出和游戏市场支出总额的不到 1% 扩大到 10%。增量收入的最大驱动力将是生成式 AI 基础设施作为用于培训法学硕士的服务（到 2032 年将达到 2470 亿美元），其次是由技术驱动的数字广告（1920 亿美元）和专门的生成人工智能助理软件（890 亿美元）。在硬件方面，收入将由人工智能服务器（1320 亿美元）、人工智能存储（930 亿美元）、计算机视觉人工智能产品（610 亿美元）和对话式人工智能设备（1080 亿美元）驱动。

Bloomberg

Bloomberg Intelligence Interactive Calculator: Generative AI Market Opportunity

($ million, unless otherwise specified)

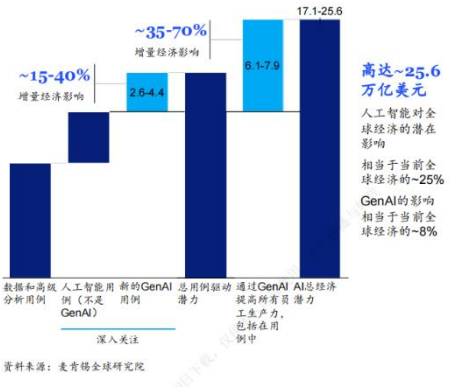
Generative AI Revenue Projections	2022	2027E	2032E	2022-32E CAGR
Hardware	**$37,973**	**$223,615**	**$641,737**	**33%**
Devices (Inference)	$4,128	$82,965	$168,233	45%
Computer Vision AI Products	$1,032	$22,124	$60,564	50%
Conversational AI Products	$3,096	$60,841	$107,669	43%
Infrastructure (Training)	$33,845	$140,650	$473,505	30%
AI Server	$22,563	$49,641	$133,817	19%
AI Storage	$9,025	$33,094	$92,642	26%
Generative AI Infrastructure as a Service	$2,256	$57,915	$247,046	60%
Software	**$1,493**	**$58,826**	**$279,899**	**69%**
Specialized Generative AI Assistants	$447	$20,864	$89,035	70%
Coding, DevOps and Generative AI Workflows	$213	$12,617	$50,430	73%
Generative AI Workload Infrastructure Software	$439	$13,468	$71,645	66%
Generative AI Drug Discovery Software	$14	$4,042	$28,343	113%
Generative AI Based Cybersecurity Spending	$9	$3,165	$13,946	109%
Generative AI Education Spending	$370	$4,669	$26,500	53%
Generative AI Based Gaming Spending	**$190**	**$20,668**	**$69,414**	**80%**
Generative AI Driven Ad Spending	**$57**	**$64,358**	**$192,492**	**125%**
Generative AI Focused IT Services	**$83**	**$21,690**	**$85,871**	**100%**
Generative AI Based Business Services	**$38**	**$10,188**	**$34,138**	**97%**
Total	**$39,834**	**$399,345**	**$1,303,551**	**42%**

Source: Bloomberg Intelligence, IDC, eMarketer, Statista

而根据麦肯锡预测，人工智能整体将为全球经济带来高达 25.6 万亿美元的正面经济影响，而其中来自 GenAI 的贡献高达 7.9 万亿美元。这既包括 GenAI 用例本身为企业带来的直接收入增加及成本优化，也包括了 GenAI 推动全行业生产效率提升所带来的经济价值。

在 GenAI 的推动下，科技 在诸多工作上有望实现相当于人类中位数水平的表现，且发展速度在加快，科技的各项能力表现达到人类工作表现前 25%水准的时间比先前预估已有所提前。例如，麦肯锡全球研究院之前的分析认为，科技在自然语言理解方面最早会在 2027 年到达人类中位数水准，但根据最新分析，这一时间已提前到 2023 年。整体而言，将 GenAI 与现有技术结合，可通过三种方式对全球产业产生巨大影响：提高企业生产效率、推动企业产品创新以及改变行业竞争格局。

图 ：人工智能技术对全球经济的潜在影响在17.1-25.6万亿美元之间，相当于~25%的生产率增长



资料来源：麦肯锡全球研究院

图 ：GenAI用例在不同行业和部门中具有不同规模的影响

GenAI在不同行业和部门中的产值[1]

	占行业总收入的比例	总产值，十亿美元	营销与客户销售	客户运营	产品与软件开发	软件开发	供应链风险与战略运营	风险与战略法务	与公司融资	公司IT[2]	人才与组织
			760-1,200	340-470	230-420	580-1,200	280-530	180-260	120-260	40-50	60-90
高科技	4.8-9.3	240-460									
银行业	2.8-4.7	200-340									
药品与医疗产品	2.6-4.5	60-110									
通信	2.3-3.7	60-100									
教育	2.2-4.0	120-230									
医疗	1.8-3.2	150-260									
保险	1.8-2.8	50-70									
媒体与娱乐	1.5-2.6	60-110									
高端制造[3]	1.4-2.4	170-290									
消费品	1.4-2.3	160-270									
高端电子与半导体	1.3-2.3	100-170									
旅行、交通与物流	1.2-2.0	180-300									
零售[4]	1.2-1.9	240-390									
地产	1.0-1.7	110-180									
能源	1.0-1.6	150-240									
行政与专业服务	0.9-1.4	150-250									
化学	0.8-1.3	80-140									
建筑	0.7-1.2	90-150									
基础材料	0.7-1.2	120-200									
农业	0.6-1.0	40-70									
公共与社会部门	0.5-0.9	70-110									
		2,600-4,400									

低影响 高影响

注：由于四舍五入，数字之和可能不等于100%
1. 不包括实施成本（例如培训、许可证）
2. 不包括软件工程
3. 包括航空航天、国防和汽车制造
4. 包括汽车零售

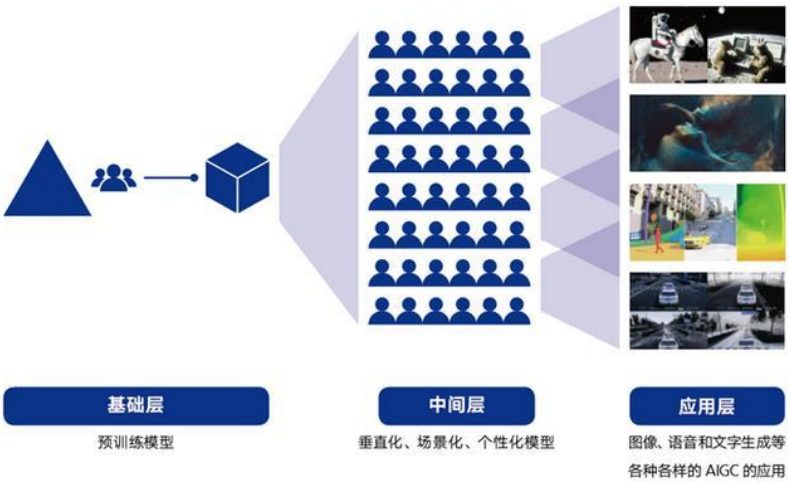
资料来源：比较行业服务(CIS)、IHS Markit；牛津经济杂志；麦肯锡公司和业务职能数据库；麦肯锡制造和供应链360；麦肯锡销售导航；麦肯锡数据库Ignite；麦肯锡分析

2、AIGC 的行业结构

根据腾讯研究院发布的《AIGC 发展趋势报告 2023：迎接人工智能的下一个时 代》，整个 AI 生成内容链条第一层是基础层，也是由预训练模型为基础搭建的 AIGC 技 术基础设施层。

第一层是基础层，以预训练模型为基础搭建的 AIGC 技术基础设施层。在国外，以 OpenAI、Stability.ai 为代表，通过受控 API、开源等方式输出模型能力。

第二层是中间层，是在预训练模型基础上，通过专门的调试和训练，快速抽取形成垂直化、场景化、定制化的小模型和应用工具层，可以实现工业流水线式部署，同时兼具按需使用、高效经济的优势。比如，知名的二次元画风生成模型 Novel-AI，以及各种风格的角色生成器等，就是基于 Stable Diffusion 开源进行的二次开发。随着 AIGC 模型加速成为新的技术平台，模型即服务（Model-as-a-Service，MaaS）开始成为现实，预计将对商业领域产生巨大影响。

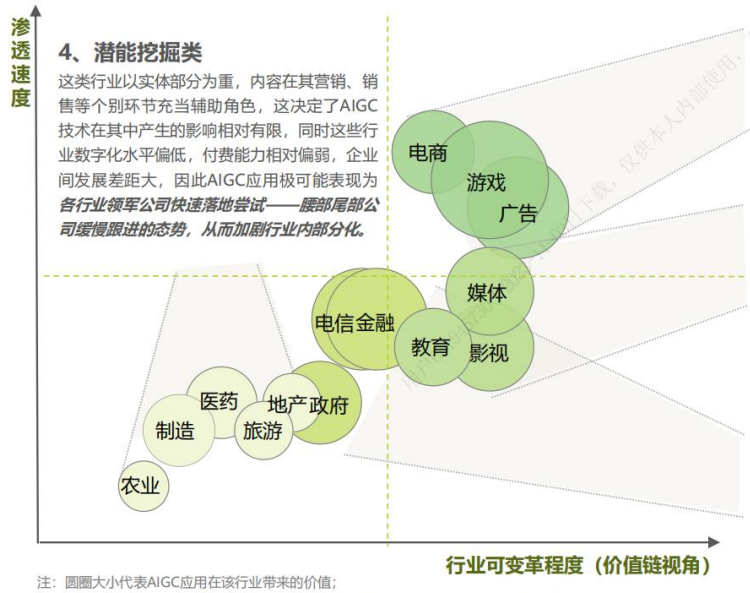


第三层是应用层，依托底层模型和中间层的垂直模型，各厂商进一步开放面向 C 端和 B 端用户的各种各样的 AIGC 产品和服务，满足海量用户的内容创建和消费需求。例如群聊机器人、文本生成软件、头像生成软件等 AIGC 消费工具。

目前，从提供预训练的 AI 大模型的基础设施层公司到专注打造垂直领域内 AIGC 工具的中间层公司、再到直接面对消费者和终端用户提供产品和服务的应用层公司，美国围绕 AIGC 生长出繁荣的生态，技术创新

引发的应用创新浪潮迭起；中国也有望凭借领先的 AIGC 技术赋能千行百业。

总体而言，AIGC 主要影响内容创作与人机交互，因此价值链线上化程度越高，内容在价值链中占比越高，AIGC 对其颠覆效应越明显；另一方面，行业自身的数据、知识、监管要求等特点也会深刻影响到 AIGC 技术的渗透速度。比如电商、游戏、广告、影视传媒等以内容生产为价值核心的行业，以及电商、金融等研发设计、营销等环节在行业价值链中地位较高的行业，能够快速看到 AIGC 应用对原有生产工具的替代和业务流程的变革。



1、快速颠覆类

电商、游戏和广告行业线上化程度高，且内容质量直接决定其价值创造，这两大特征使得AIGC应用在其中能够产生最大化的价值，并能够迅速渗透至核心生产环节。据统计，AIGC相关应用已经帮助游戏行业在研发制作环节节约**50-70%**人力或时间成本。

行业中大小公司均有机会抢占先机，甚至受到个人开发者的冲击，行业可能面临洗牌。

2、匀速增长类

这类行业也以内容为价值核心，但与第1类区别在于其内容生产更多环节在实体环境完成，具有更强的专业性、灵活性，在这些行业中AIGC可变革上限略低，且会更加考验AIGC技术的成熟度，当前应用多属于单点尝新而非刚性替代，未来市场空间大，但实现大规模行业渗透需要更长时间。

行业格局预计产生局部波动，行业价值链各环节地位排序面临重构。

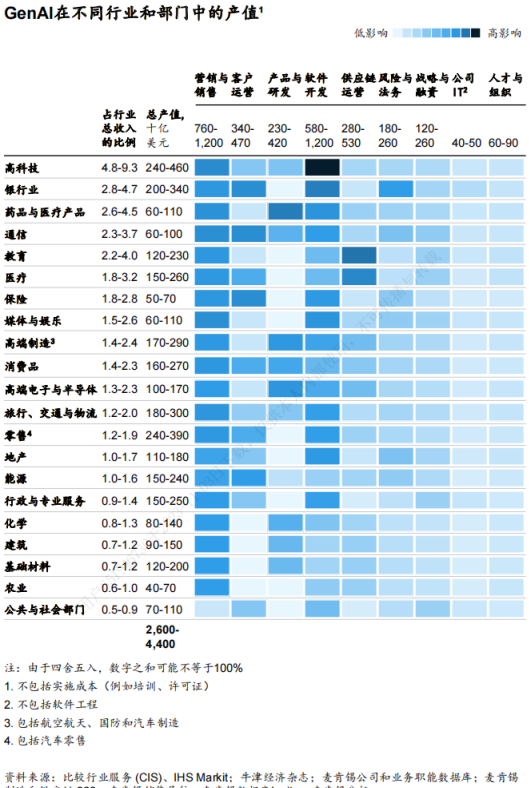
3、稳中求变类

金融、电信和政府行业价值链中涉及大量内容生产和人与人交互的环节，数字化转型付费意愿和付费能力强。但同时因为其具有业务流程与组织架构的可变性低，对数据安全极为敏感等特质，因此对AIGC应用的态度最为保守。***解决大模型私有化部署的ROI问题是打开这类市场的关键。***

注：圆圈大小代表AIGC应用在该行业带来的价值；
行业可变革程度：在行业价值链中涉及到内容生产和人机交互的环节占比，以及AIGC对相关环节的影响程度；
渗透速度：由各行业政策监管、数字化特别是数据基础建设水平、数据安全要求、行业创新能力等指标构成
来源：艾瑞咨询研究院自主研究绘制。

图3：GenAI用例在不同行业和部门中具有不同规模的影响

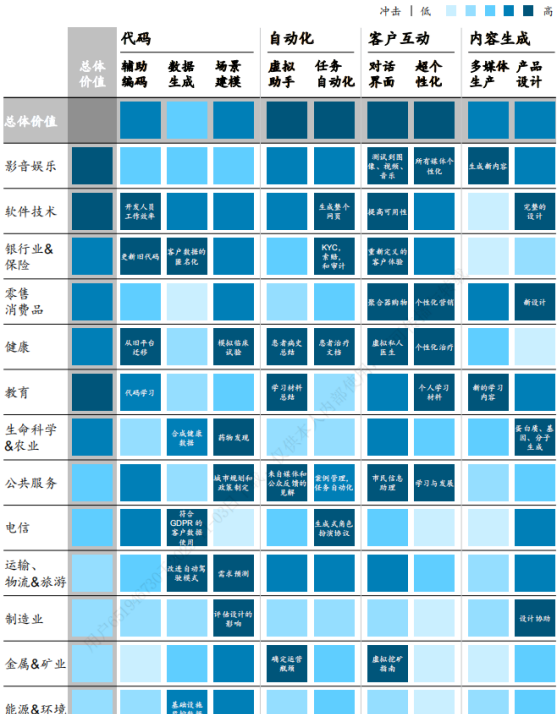
GenAI在不同行业和部门中的产值[1]



注：由于四舍五入，数字之和可能不等于100%
1. 不包括实施成本（例如培训、许可证）
2. 不包括软件工程
3. 包括航空航天、国防和汽车制造
4. 包括汽车零售

资料来源：比较行业服务 (CIS)、IHS Markit；牛津经济杂志；麦肯锡公司和业务职能数据库；麦肯锡制造和供应链 360；麦肯锡销售导航；麦肯锡数据库Ignite；麦肯锡分析

图 ：产品创新：我们预计软件和技术以及媒体和娱乐领域的影响最大



1. 每个行业都包含行业特定的软件 2.与行业无关的软件 3.包括航空航天和国防、汽车和装配、化学品、半导体、基础材料 4.包括石油和天然气、电力

资料来源：麦肯锡分析

根据麦肯锡研究预测，AIGC 对营销与销售、软件开发、客户运营与产品研发三大部门影响最大，其中在营销与销售领域，AIGC 的价值达 7,600 亿至 1.2 万亿美元。从行业来看，AIGC 产生价值最大的三个行业为高科技、零售及银行业。其中在高科技行业，AIGC 的价值将达 2,400 亿至 4,600 亿美元。从所占行业收入比例来看，高科技、银行业和药品及医疗产品行业位居前三，其中在高科技行业，AIGC 所带来的价值占行业收入的比例达 4.8%~9.3%。

能够从 AIGC 应用的这一能力中受益最深的行业当属影音娱乐、软件开发与教育。例如，AIGC 能协助影音娱乐企业生成突破想象力的多媒体内容，软件开发企业可利用 AIGC 设计全新的产品形态，学校则能够通过 AIGC 针对每个学生特点设计作业及课后辅导。

3、典型企业在 AIGC 的布局

（1）OpenAI

OpenAI 成立于 2015 年，由伊隆·马斯克、Sam Altman、Greg Brockman 等人发起，旨在推动人工智能技术的发展和应用。在成立初期，OpenAI 的研究重点主要集中在机器学习和自然语言处理领域。

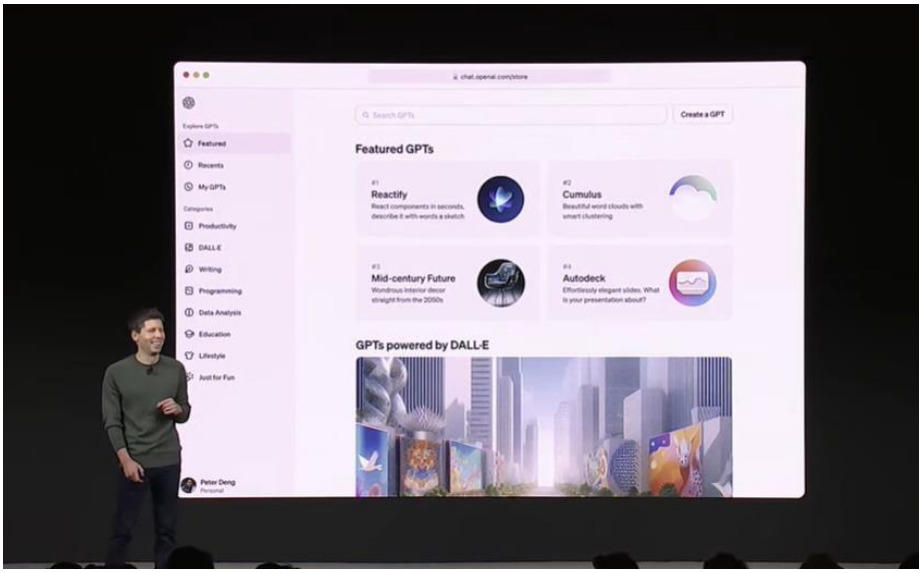
2016 年，AlphaGo（由 DeepMind 开发）在围棋比赛中击败了世界冠军李世石，引起了广泛关注。AlphaGo 的成功突显了人工智能在复杂任务上的能力，也激发了 OpenAI 对人工智能的进一步研究和探索。

2018 年，OpenAI 发布了第一个基于生成对抗网络（GAN）的语言模型 GPT（Generative Pre-trained Transformer）。GPT 在自然语言处理领域取得了巨大的成功，能够生成逼真的文本和回答问题。随后，OpenAI 陆续发布了 GPT-2 和 GPT-3，这些模型规模更大、能力更强，被广泛应用于文本生成、对话系统和其他领域。

OpenAI 与其他公司和组织展开了合作，共同推动人工智能的研究和发展。2018 年，OpenAI 与微软合作，共同开发人工智能技术，并获得了微软的投资。OpenAI 还与其他公司合作，包括与 GPT-3 的商业合作伙伴，通过 API 提供 GPT-3 的服务。

2019 年，OpenAI 宣布转型为一家公司，以更好地推动人工智能的发展。公司成立了 OpenAI LP，作为非营利实体，以及 OpenAI Inc.，作为营利实体，以支持 OpenAI 的研究和商业化目标。

2021 年，OpenAI 发布了 GPT-3.5 Turbo 模型，这是在 GPT-3 基础上进行优化和改进的新版本。GPT-3.5 Turbo 模型在保持高质量的同时，提供了更高的性能和更低的价格，使更多人能够使用和受益于该技术。2023 年 3 月发布 GPT-4.0，GPT4 相较于 GPT3.5，模型规模、模型能力、模型输入、模型训练，都有了质的提升。**2023 年 11 月发布 GPT-4 Turbo、GPTs 和 GPT 应用商店。**

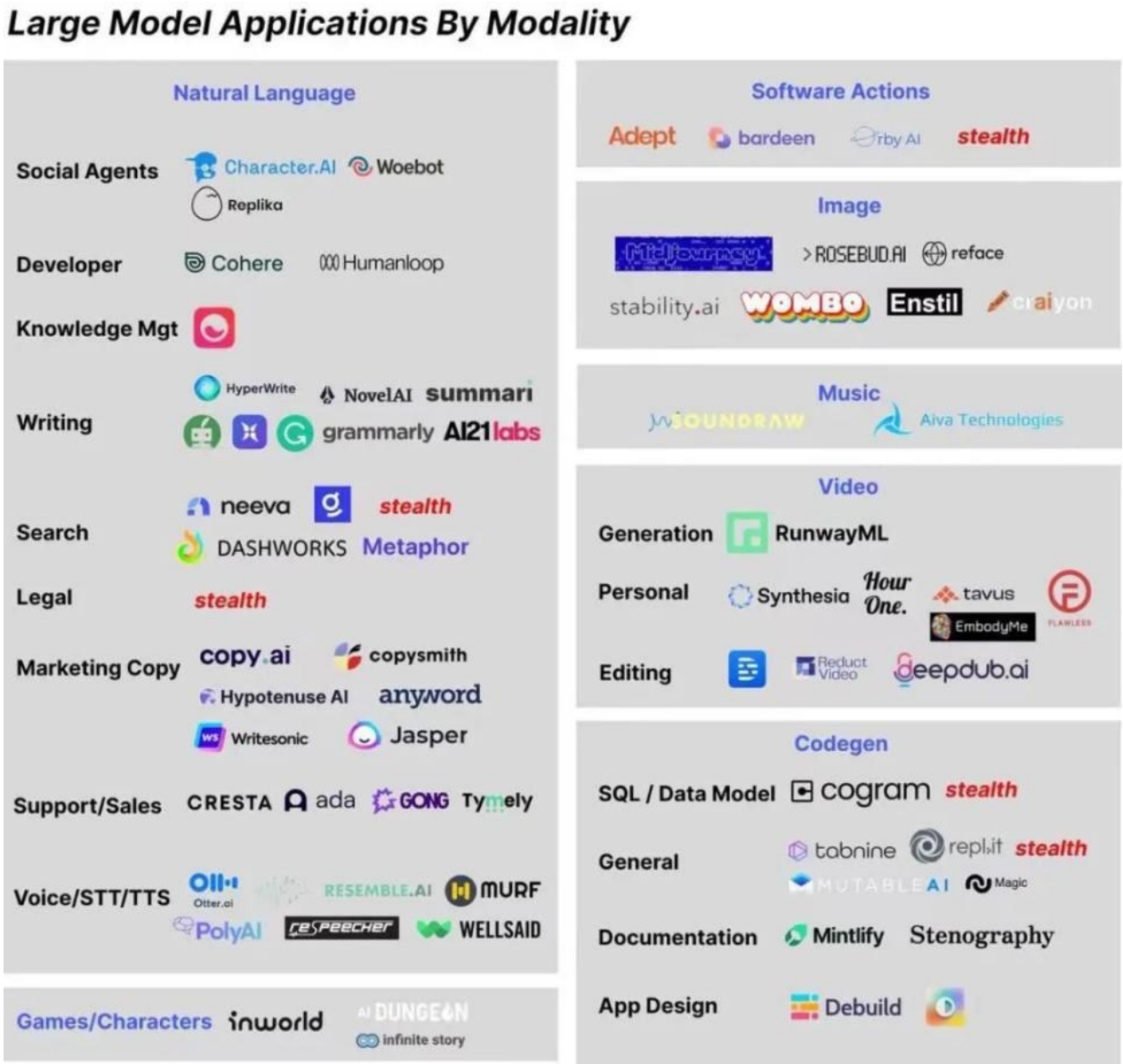


截至 2023 年 11 月，已经有 200 万开发者正在使用 OpenAI 的 API（应用程序接口），在全球各地提供多种多样的服务；92%的财富 500 强公司正在使用 OpenAI 的产品搭建服务，而 ChatGPT 的周活用户数也达到 1 亿人。

“GPT Store”是下一次重大飞跃，类似于 Apple 的 App Store 引起的革命。它是一个数字市场，开发者和创作者可以在其中发布和获利他们自己建

立的定制 AI 代理。通过利用 GPT 框架，“GPT Store”不仅仅是一个应用程序仓库，还是一个培育创新者和用户繁荣社区的动态平台。

基于 ChatGPT 生态的应用很多，GPT-3 开放 api 很久，细分赛道很多。大致可以根据生成内容不同分为两类：机器编程语言生成、人类自然语言生成。前者主要有代码和软件行为的生成等，后者主要有新闻撰写、文案创作、聊天机器人等。



1） 代码生成：Github Copilot

代表公司是微软出品的 Github Copilot，编程中的副驾驶。**该产品基于 OpenAI 专门用 GPT-3 为编程场景定制的 AI 模型 Codex**。使用者文字输入代码逻辑，它能快速理解，根据海量开源代码生成造好的轮子供开发者使用。提高一家科技公司 10% 的 coding 效率能带来很大收益，微软内部已进行推广使用。

相比低代码工具，Copilot 的目标群体是代码工作者。未来的低代码可能是两者结合：低代码 UI 界面实现代码框架搭建，代码子模块通过 Copilot 自动生成。

正如 Copilot 的 slogan：Don't fly solo，没有 Copilot 的帮助 coder 的工作会变得繁冗，没有 coder 的指引 Copilot 生成的内容可能会出现纰漏。也有用户报告了一些侵犯代码版权、或代码泄露的案例，当前技术进步快于版权法规产生了一定的空白。

2）软件行为生成：Adept.ai

Adept.ai 是一家明星创业公司。创始团队中有两人是 Transformer 模型论文作者，CEO 是谷歌大脑中大模型的技术负责人，已经获得 Greylock 等公司 6500 万美元的 A 轮融资。

他们的主要产品是大模型 ACT-1，让算法理解人类语言并使机器自动执行任务。目前产品形态是个 chrome 插件，用户输入一句话，能实现单击、输入、滚动屏幕行文。在展示 demo 中，一位客服让浏览器中自动记录下与某位顾客的电话，正在考虑买 100 个产品。这个任务需要点击 10 次以上，但通过 ACT-1 一句话就能完成。

（2）阿里巴巴

阿里 AI 大模型"通义千问"于 2023 年 4 月的阿里云峰会重磅发布。阿里巴巴集团 CEO 张勇表示，基础大模型的核心是能够支撑各行各业，阿里希望能够为客户与合作伙伴提供面向千行百业的专属大模型。阿里巴巴表示，所有产品未来都要接入大模型进行全面的升级，所有行业和所有服务都值得重新做一遍。

来源：机器之心微信公众平台，国金证券研究所

阿里 AI 大模型“通义千问”前身系阿里达摩院 M6 项目，阿里达摩 院于 2020 年 6 月发布 3 亿参数基础模型，21 年 1 月模型参数规模达百亿，同年 5 月达万亿参数，同年 10 月达 10 万亿，成为全球首个 10 万亿参数多模态大模型，并落地应用于天猫虚拟主播等 40 多个细分场景。22 年 9 月达摩院发布“通义”大模型系列，打造业界首个 AI 底座，且兼顾大小模型的层次化建构体系。

M6 项目于 2020 年启动，同年 6 月推出 3 亿参数的基础模型，2021 年 1 月，模型参数规模达百亿，成为世界最大的中文多模态模型。 2021 年 5 月，达摩院发布万亿参数模型 M6 并正式投入使用，追上谷歌发展脚步。M6 在多模态 GreenAI、文到图生成、商业化领域并肩世界一流水平， 与英伟达、谷歌相比，M6 仅用 480 卡 V100 32G GPU 就实现了万亿模型，节省算 力资源超 80%，训练效率提升近 11 倍。2021 年 10 月，M6 进一步升级成为全球首个 10 万亿参数的多模 态大模型，并应用于天猫虚拟主播等 40 多个创造相关场景中；在绿色低碳方面，相 比 GPT-3，M6 实现了同等参数规模下，能耗仅为 1%。 2022 年 9 月，达摩院发布“通义”大模型系列，打造业界首个 AI 统一底 座，并构建了大小模型协同的层次化人工智能体系，其中，统一底座 M6-OFA 模型 在不引入新增结构情况下，可同时处理 10 余项单模态和跨模态任务，通义大模型的 出现将为 AI 从感知智能迈向知识驱动的认知智能提供先进基础设施。

图表 ：通义大模型架构基础框架

来源：机器之心微信公众平台，国金证券研究所

通用模型层主要包含通义-M6、通义-AliceMind、通义-视觉三种通用模型。1）通义 -M6 是国际首个参数规模达到 10 万亿的全球最大预训练模型。2）通义-AliceMind 作为开源深度语言模型体系，形成了从文本 PLUG 到多模态 mPLUG 再到模块化统 一模型演化趋势。3）通义-视觉可在电商行业实现图像搜索和万物识别等场景应用， 并在文生图以及交通和自动驾驶领域发挥作用。

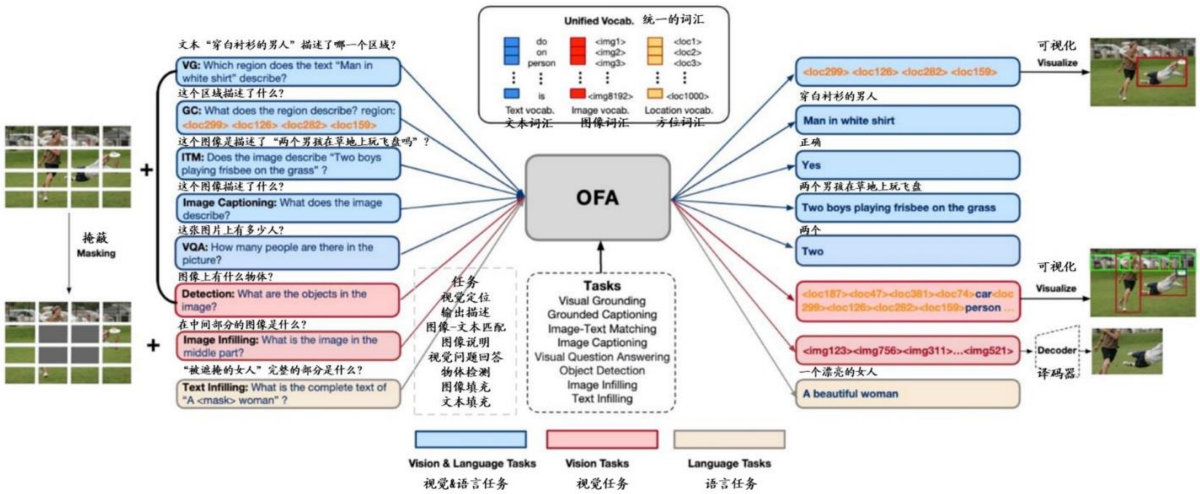
通义大模型在国内率先构建 AI 统一底座，在业界首次实现模态表示、任务表示、模型结 构的统一，统一学习范式 OFA 是通义大模型背后的核心技术支撑。

架构统一：M6-OFA 采用了 Transformer Encoder-Decoder + ResNet Blocks 架构， ResNet Blocks 用于提取图像特征，Transformer Encoder 负责多模态特征的交互， Transformer Decoder 采用自回归方式输出结果。无需增加任何任务特定的模型层， 即可实现预训练与微调的相同学习模式。

模态统一：M6-OFA 构建了一个涵盖不同模态的通用词表，以便模型使用该词表表 示不同任务的输出结果。其中 BPE 编码的自然语言 token 用于表示文本类任务或图 文类任务的数据；图片中连续的横纵坐标编码为离散化 token，用于表示视觉定位、 物体检测的数据；图片中的像素点信息编码为离散化 token，用于表示图片生成、图 片补全等任务的数据。

任务统一：通过设计不同的 instruction，M6-OFA 将涉及多模态和单模态（即 NLP 和 CV）的所有任务都统一建模成序列到序列（seq2seq）任务。M6-OFA 覆盖了 5 项 多模态任务，视觉定位、定位字幕、图文匹配、图像字幕和视觉问答；2 项视觉任务， 检测和图像填补和 1 项文本任务，即文本填补。

图表 ：M6-OFA 实现模态统一



来源：机器之心微信公众平台，国金证券研究所

通义千问赋能天猫精灵有望成为智能居家生态入口的不二选择。阿里大模型通义千问有望 赋能旗下智能音箱天猫精灵打造居家场景智能生态入口，与萤石网络等智能家居厂商优势 互补，通过“人机自然交互、信息上传云端，联动控制反馈”的机制，开展智能家居生态共建：

图表 ：通义千问赋能天猫精灵有望打通智能居家生态，实现居家五大场景智能化落地

来源：萤石网络官网，国金证券研究所

阿里大模型赋能电商场景，借助平台生态入口打造全域智能电商。从文案的宣发设计到最 终的客户关系维护，再到新一轮产品的宣传设计，阿里大模型赋能下的智能电商有望实现 全流程闭环，其 5 大革新趋势展望如下：

营销文案智能生成（宣传设计）：通过人机交互的方式对商品宣发需求进行自然语言 描述，应用阿里妈妈发布的“AI 智能文案”产品，结合淘宝、天猫的海量优质内容与推 荐算法，可基于商品本身自动生成高品质的营销文案。

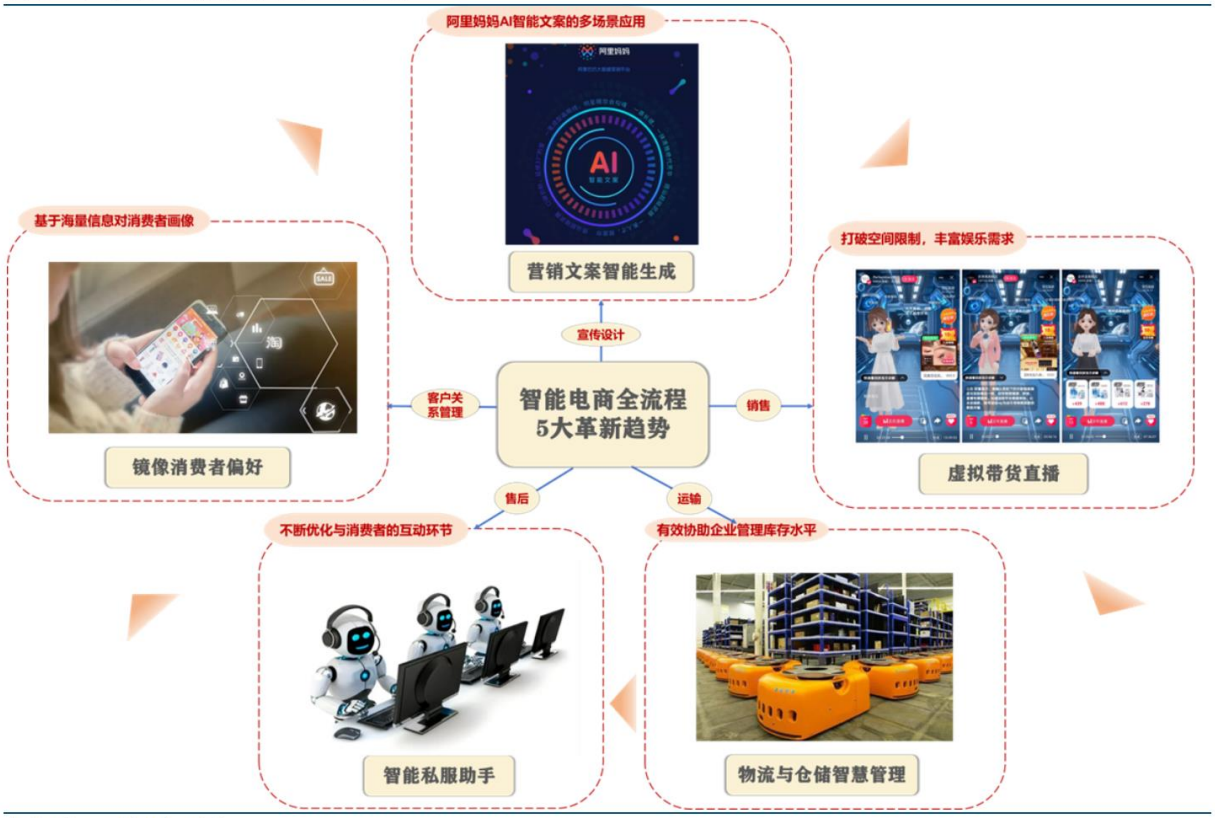
虚拟带货直播（销售）：使用虚拟形象进行直播，实现内容和营销上的创新，打造个 性化专属虚拟数字人 IP，通过企业的虚拟形象与消费者进行实时互动，更好的促进 企业传递品牌价值，创造收益。

物流与仓储智慧管理（运输）：根据库存实时变化、区域订单体量、退（换）货数量 等指标，协助企业进行合理的物流调度与规划。此外，通过分析消费者购买偏好、竞 争对手销售数据、消费者评论点赞等数据，帮助企业将库存维持在最合适的水平。

智能私服助手（售后）：作为阿里赋能的电商平台智能私服助手，区别于以往的智能 客服，私服助手提供更智能、更便捷的定制化专属服务，避免了“1 客服对接 n 客户”的低效场景，大幅提升了售后环节调整商品和评价反馈的效率。

镜像消费者偏好（客户关系管理）：人工智能的相关技术就像是一面镜子，对于海量 消费者的喜好、反馈等信息进行汇总、统计，然后进行画像，无论是商品的改进，还 是服务的优化，都将变得有迹可循。

图表 .：智能电商全流程五大革新趋势



来源：腾云，国金证券研究所

阿里大模型有望赋能钉钉实现 AI 智能生成群聊摘要、辅助内容创作、自动总结会议、拍 照生成应用等功能，助力办公场景的智能化发展。

智能生成群聊摘要：通义千问赋能钉钉基于群聊此前聊天内容自动生成聊天摘要，帮 助用户快速了解上下文，无需手动爬楼。

辅助内容创作：在钉钉文档里，用户通过自然语言输入需求，可实现包括文案文本与 创意图片在内的内容创作。

自动总结会议：钉钉视频会议可在会议中生成实时字幕；新入会成员可通过智能摘要 快速了解之前内容；视频会议也支持在会后自动生成重点摘要与待办事项。

拍照生成应用：通过拍照生成低代码业务应用，上传一张功能草图，无需代码即可生 成一款应用，应用开发的门槛被再一次降低。

图表 ：阿里大模型赋能钉钉实现四大智能办公场景

来源：钉钉微信公众平台，国金证券研究所

阿里入局 AI 大模型竞争的核心优势不仅在于算力算法层面的 优越性与 C 端生活场景数据的丰富性，而且在于能够精准把控 C 端用户的生态入口。

（3）英伟达

人工智能的开发需要算力、算法与数据，三者缺一不可。而英伟达的 GPU 正如同 AI 时代的原油，如今正逐步成为人工智能行业最重要的基础设施之一。1999 年 10 月，英伟达推出了 GeForce 256，并在营销中表示“这是世界上第一款 GPU”，**现如今，英伟达已是全球 GPU 市场领导者，拥有绝对领先优势。**

2019 年，NVIDIA GPU 部署在了四大云提供商 (AWS、谷歌、阿里巴巴和 Azure) 97.4% 的 AI 加速器实例中（ 用于提高处理速度的硬件)。Cambrian AI Research 分析师 Karl Freund 表示，该公司在人工智能算法训练市场上占据了“近 100%”的份额。与此同时，几乎所有的人工智能里程碑都发生在 NVIDIA 的硬件上，吴恩达的 YouTube 寻猫器、DeepMind 的棋牌游戏冠军 AlphaGo、OpenAI 的语言预测模型 GPT-3 都在 NVIDIA 硬件上运行。可以说英伟达的 GPU 是人工智能研究人员的立足之地。

随着头部互联网及科技企业纷纷推出大语言模型，国内对算力资源的需求将呈现井喷。面对如此火热的市场需求，英伟达创始人在今年的 GTC 大会上再一次抢占先机，发布了可为 ChatGPT 提速 10 倍的专用 GPU 芯片 NVIDIA H100 NVL，**体现出了其在 AIGC 群雄逐鹿的市场中趁胜追击的巨大野心。**

在 AIGC 领域，英伟达也做了大量的基础工作，为开发者提供了一个更为高效的开发环境。


NVIDIA AI 模型 和"代工厂
NVIDIA AI 平台软件
NVIDIA 加速型基础架构
NVIDIA NeMo
提供先进的大型语言基础模型、定制工具和大规模部署功能，为您开启高度个性化的企业 AI 应用之旅。NVIDIA NeMo™ 由 NVIDIA DGX™ Cloud 提供支持，是模型制作服务套件 NVIDIA AI Foundations 的一部分，该套件旨在促进企业级生成式 AI 的发展并实现跨用例定制。
详细了解 NeMo
NVIDIA BioNeMo
借助 NVIDIA BioNeMo™ 研究人员和开发者可以使用生成式 AI 模型快速生成蛋白质和分子的结构与功能，从而加快新候选药物的研发。BioNeMo 是 NVIDIA AI Foundations 的一部分，由 NVIDIA DGX™ Cloud 提供支持。
详细了解 BioNeMo
NVIDIA Picasso
NVIDIA Picasso 让生成式 AI 更上一层楼。企业、软件创作者和服务提供商可以在其模型上运行优化推理，使用专有数据训练先进的生成模型，或者以预训练模型为基础，利用文本或图像提示生成图像、视频 3D 和 360 HDRi 内容。Picasso 由 NVIDIA DGX™ Cloud 提供支持，是 NVIDIA AI Foundations 的一部分，并通过 Cloud API 与生成式 AI 服务无缝集成。
详细了解 Picasso
NVIDIA ACE
ACE 可帮助中间件、工具和游戏开发者在软件或游戏中构建和部署自定义语音、对话和动画 AI 模型。
详细了解 ACE
生成式 AI 模型
NVIDIA 提供先进的社区以及由 NVIDIA 构建的基础模型（包括 GPT、T5 和 Llama），帮助加快生成式 AI 的采用过程。用户可以从 Hugging Face 或 NGC 目录下载这些模型，并在浏览器中使用 AI Playground 直接对其进行测试。
体验 NGC 的模型
NVIDIA AI 模型 和"代工厂
NVIDIA AI 平台软件
NVIDIA 加速型基础架构
NVIDIA AI Enterprise
对于运用 AI 开展业务的企业，NVIDIA AI Enterprise 提供了一个安全可靠、适用于开发和部署的生产级端到端软件平台。该平台包含 100 多个框架、预训练模型和开源开发工具（例如 NeMo、Triton™、TensorRT™），以及生成式 AI 参考应用和企业支持，可有效简化采用过程。
NVIDIA AI Enterprise 可部署在任意位置。企业组织可以灵活地在云、数据中心、工作站和边缘运行其支持 NVIDIA AI 的解决方案，实现一次开发，随处部署。
详细了解 NVIDIA AI Enterprise
NVIDIA NeMo
NVIDIA NeMo 可以帮助企业组织从头开始构建自定义大型语言模型（LLM），定制预训练模型，并进行大规模部署。NeMo 随 NVIDIA AI Enterprise 提供，包含训练和推理框架、护栏工具包、数据管护工具，以及预训练模型。
详细了解 NeMo
NVIDIA Triton 推理服务器
此软件跨各种工作负载提供了标准化的 AI 模型部署和执行方法。借助强大的优化性能，您可以在单 GPU、多 GPU 和多节点配置上实现出色的推理性能。NVIDIA AI Enterprise 随附的 NVIDIA Triton 管理服务可自动部署多个 Triton 推理服务器实例，从而实现性能更出色、利用率更高的大规模推理。
详细了解 Triton
NVIDIA TensorRT-LLM
这个开源库可以针对 NVIDIA GPU 上的生产部署优化新型 LLM 的模型推理性能。通过使用 TensorRT-LLM 试验新型 LLM，开发者无需深入掌握 C++ 或 CUDA 知识即可确保性能如飞。
TensorRT-LLM 基于 FasterTransformer 项目构建，不仅具有出色的灵活性，而且能与 NVIDIA Triton 推理服务器紧密配合，使先进的 LLM 发挥出更优异的端到端性能。
申请抢先体验 TensorRT-LLM
NVIDIA AI 模型 和"代工厂
NVIDIA AI 平台软件
NVIDIA 加速型基础架构
NVIDIA DGX
NVIDIA DGX 将 AI 软件、专用硬件和专业知识集成到一款综合的 AI 开发解决方案中，适用于从云到本地数据中心等各种环境。NVIDIA DGX Cloud 为多节点训练提供了一个全栈无服务器 AI 平台，其中包含企业级开发者套件、出色的基础架构，以及直接联系 NVIDIA AI 专家的机会。该平台以可预测的总体价格提供，便于企业立即上手使用。
探索云端 AI 超级计算机
NVIDIA 认证系统
为了提高运营效率并交付先进的产品和服务，企业需要依靠具有出色性能、可靠性和可扩展性的计算基础架构。NVIDIA 认证系统™ 为企业提供了安全可靠、针对各种加速的现代化工作负载（桌面、数据中心和边缘）而优化的硬件解决方案，确保企业可以放心部署。
寻找认证服务器和工作站
优势
更快推出解决方案
借助专为多节点训练而设计，并与领先云服务提供商合作提供的基础架构，打造并租赁您自己的 AI 卓越中心。
生产就绪
放心部署安全可靠、针对生成式 AI 工作负载而优化的加速型基础架构。
强劲性能
利用专为训练和部署 LLM 而优化的强大生成式 AI 加速器创造价值。

摘自英伟达官网

Graphcore 的联合创始人兼首席执行官奈杰尔•图恩曾表示：“英伟达在隐藏 GPU 复杂性方面做得非常出色。它之所以有效，是因为他们创建了软件库、框架和优化，使复杂性得以隐藏。这是 NVIDIA 在那里承担的一项非常繁重的工作。”

最近，英伟达公布了 2023 财年第三财季的财报。财报显示，第三财季，该公司的营收达到创纪录的 181.2 亿美元，同比增长 206%，环比增长 34%。其中，数据中心业务营收达到创纪录的 145.1 亿美元，环比增长 41%，同比增长 279%，是英伟达营收和利润快速增长的主要推动力。

同时，英伟达发布了人工智能（AI）芯片 HGX H200，该 GPU 利用 Hopper 架构来加速人工智能应用。H200 是去年发布的 H100 GPU 的后续产品。H100 是英伟达首款基于 Hopper 架构打造的 GPU，此前是英伟达最强大的人工智能 GPU 芯片。英伟达称，H200 是首款提供 HBM3e 内存的 GPU。得益于 HBM3e，H200 可提供 141GB 内存和每秒 4.8TB 的带宽，可加速生成式人工智能和大型语言模型(LLM)，同时可以处理人工智能和超级计算工作负载所需的大量数据。公司还表示，H200 的推出将带来进一步的性能飞跃，包括在 Llama 2（一个具有 700 亿参数的 LLM）上的推理速度比 H100 快了近一倍。预计在未来的软件更新中，H200 的性能将进一步领先和提高。英伟达表示，全球系统制造商和云服务提供商将从 2024 年第二季度开始使用 H200。

英伟达（NVIDIA）首席执行官黄仁勋表示，**全球正处于人工智能（AI）浪潮的开端，他对此充满信心。他认为，数据中心的增长势头将延续至 2025 年，并强调公司正在扩大芯片供应链以满足这一增长需求。**

参考资料：
1、 麦肯锡_中国金融业 CEO 季刊：捕捉生成式 AI 新机遇；
2、 甲子光年_2023AIGC 市场研究报告：ChatGPT 的技术演进、变革风向与 AIGC 投资机会分析；
3、 甲子光年_中国 AIGC 产业算力发展报告：AIGC 爆发，算力服务机会 or 变革；
4、 艾瑞咨询_2023 年中国 AIGC 产业全景报告：日就月将，学有缉熙于光明；
5、 弗若斯特沙利文_2023AI 大模型市场研究报告：迈向通用人工智能，大模型拉开新时代序幕；
6、 中信证券_技术跃迁专题研究系列之五：AIGC 应用端主题-千帆竞发，大海星辰。