

AI 大时代系列报告之一（基础篇）： 大模型与算力共振，奇点时刻到来

报告要点

大模型涌现，ChatGPT 开启 AI 未来时刻。人工智能是引领未来的新兴战略性新兴产业，自 19 世纪 50 年代诞生以来，相关理论和技术持续演进，近十余年得益于深度学习算法的突破、算力设备性能不断提升与海量数据的持续积累，人工智能得以大范围地从实验室研究走向产业实践。2017 年 Transformer 架构推出，创新提出注意力机制，有效提升了自然语言处理任务的精度和质量，计算效率和资源利用率亦明显改善，成为了当前自然语言领域的主流模型架构。2022 年 11 月，美国初创企业 OpenAI 推出 ChatGPT 聊天机器人，发布后访问量爆发式增长，发布 5 天用户数突破 100 万，2 个月过亿，用户增速创历史新高。ChatGPT 展示的连续对话、上下文理解、敢于质疑、用户意图思考捕捉的能力，让产业看到了人工智能的非凡力量，国内外厂商纷纷加入 AI 大模型军备竞赛，大模型参数规模与计算需求呈量级提升，同时积极探索大模型商业化落地应用，AI 奇点时刻正加速来临。

算力需求指数级增长，英伟达打造 AI 时代引擎。算力芯片包括 GPU、FPGA、ASIC、TPU 等。从 1950 年至今大模型参数与算力需求均呈现指数级增长，应用领域逐步从学术拓展至产业，全球企业持续增加对 AI 的投资，人工智能芯片搭载率有望持续增长，算力正成为新的生产力。目前 GPU 是 AI 算力的主要选择，中国人工智能芯片市场 GPU 占比接近 90%，其并行计算能力可大幅提升计算效率，可大幅缩短 AI 算法的训练和推理时间，成为 AI 时代的算力核心。据 JPR 统计，2023 年一季度英伟达 GPU 市占率达 84%，是 GPU 市场的主导者。英伟达 1999 年首次提出 GPU 概念，2006 年推出 CUDA 运算平台，2023 年 5 月发布超级计算机 DGX GH200，其算力规模达到 1Eflops，支持万亿参数 AI 大模型训练，为巨型人工智能模型提供线性可扩展性。英伟达基于其 GPU 与 CUDA 等生态体系的构建，已成为 AI 算力关键供应商。

中文大模型追赶，办公生产力应用加速推进。AI 大浪潮下，科技巨头、互联网大厂纷纷入局密集上线大模型，2023 年 3 月百度发布文心一言，4 月华为发布盘古大模型，阿里发布通义千问大模型，商汤科技公布日日新大模型体系，5 月科大讯飞发布星火大模型。多家上市公司亦开始布局，助力 AI 大模型产业化。AI 大模型能够理解更多形式的信息，降低编程门槛，弥合数字鸿沟，为各行各业带来新的变化。微软率先发布 Microsoft 365 Copilot，是一款基于 GPT-4 和 Microsoft Graph 的 AI 办公助手，将重复工作流程自动化，为用户提供了一种全新工作方式，提升工作效率，解锁生产力。国内多家厂商也推出大模型相关应用，重塑千行百业。

相关公司：1) AI 算力方向：服务器方向浪潮信息、紫光股份等；光模块方向中际旭创、天孚通信、新易盛等；2) AI 应用方向：OA 应用金山办公、万兴科技等。

风险提示： 1、AI 技术迭代不及预期风险，当前模型仍需持续迭代优化，若相关公司核心技术难以突破，将影响产品研发迭代；
2、若商业化落地不及预期，政策推进力度不及预期，经济发展不及预期，产业链各方合作不及预期，将影响产业链整体进展。

通信

评级： 看好

日期： 2023.07.11

分析师 周铃雅

登记编码：S0950521040002

☎：13661650971

✉：zhoulingya@wzq.com.cn

行业表现

2023/7/11



资料来源：聚源

相关研究

- 《光器件：从通信基石到智能之眼》(2022/9/4)
- 《通信模组：物联网信息之源，创新扩散与成本红利开启新阶段》(2022/6/7)
- 《电信运营商：5G 数字经济平台化引领者》(2022/3/2)

目录

1、大模型涌现，ChatGPT 引爆 AI 奇点	5
1.1、ChatGPT 加速人工智能发展奇点来临	5
1.2、大模型军备竞赛，开启 AI 未来时刻	10
2、算力质变，英伟达打造 AI 时代引擎	17
2.1、算力需求指数级增长，加速计算深度变革	17
2.2、英伟达 AI 时代卖水人，打造发展引擎	23
3、中文模型奋起直追，办公生产力率先变革	30
3.1、中文模型奋起直追，更强中文理解	30
3.2、生产力工具革新，释放创造力	38
4、相关公司	44
5、风险提示	45

图表目录

图表 1：ChatGPT 根据指示进行回复	5
图表 2：ChatGPT 可以实现的功能	5
图表 3：ChatGPT 用户访问量	6
图表 4：2023 年 6 月 ChatGPT 受众群体地区分布情况	6
图表 5：各应用程序达到 1 亿用户所需时间	7
图表 6：OpenAI 发展历程与股东情况	8
图表 7：ChatGPT 模型原理	9
图表 8：RLHF（人类反馈强化学习技术）训练模型技术原理	9
图表 9：AIGC 发展历程	10
图表 10：2023 年国内外发布的主要 AI 大模型梳理	11
图表 11：OpenAI 推出 ChatGPT Plus 订阅服务	12
图表 12：文本生成图像模型	13
图表 13：不同功能的语言模型	13
图表 14：Codex 语言模型示例	13
图表 15：OpenAI API 已在多个应用程序中应用	14
图表 16：GitHub Copilot 价格情况	14
图表 17：Keeper Tax 价格情况	15
图表 18：Viable 价格情况	15
图表 19：Duolingo 的数据与算法工具示例	16
图表 20：第一批 ChatGPT Plugins 合作伙伴	16

图表 21: 算力发展总体框架.....	17
图表 22: 1950-2022 年机器学习训练参数变化情况.....	18
图表 23: 1950-2022 年机器学习算力变化情况.....	18
图表 24: 2011-2022 年 GitHub AI 项目数量情况.....	19
图表 25: 2010-2021 年全球 AI 期刊论文发表数量.....	19
图表 26: 全球算力规模发展趋势.....	20
图表 27: 中国通用算力规模及预测.....	20
图表 28: 中国智能算力规模及预测.....	20
图表 29: 2020-2024 年全球人工智能、数字化转型支出及 GDP 增长趋势预测.....	21
图表 30: 全球人工智能服务器 GPU 芯片搭载率.....	21
图表 31: 全球人工智能服务器 ASIC 芯片搭载率.....	21
图表 32: 全球人工智能服务器 FPGA 芯片搭载率.....	22
图表 33: 2022 年中国人工智能芯片市场规模占比.....	22
图表 34: 英伟达建立了加速计算的良性循环.....	23
图表 35: 加速计算和人工智能将重塑计算机行业.....	23
图表 36: 英伟达推出超级计算机 DGX GH200.....	24
图表 37: 英伟达 Spectrum-X 以太网平台.....	24
图表 38: 英伟达 cuLitho 加速库合作客户.....	25
图表 39: 用于量子电路仿真的加速库 cuQuantum.....	26
图表 40: 英伟达 cuOpt 及客户.....	26
图表 41: 英伟达 CV-CUDA 平台.....	26
图表 42: NVIDIA Parabricks 医疗加速库.....	26
图表 43: 英伟达推出 DGX Cloud.....	27
图表 44: 英伟达 Nemo 语言模型.....	27
图表 45: 英伟达 Picasso 视觉语言模型.....	27
图表 46: NVIDIA 全新推理平台.....	28
图表 47: NVIDIA H100 NVL 性能情况.....	28
图表 48: NVIDIA SuperPOD 胖树拓扑网络结构.....	29
图表 49: 交换芯片技术演进路径.....	29
图表 50: 大模型的分类.....	30
图表 51: 百度发布“文心一言”.....	30
图表 52: “文心一言”文学创作能力.....	31
图表 53: “文心一言”数理逻辑推理能力.....	31
图表 54: “文心一言”中文理解能力.....	31
图表 55: “文心一言”多模态生成能力.....	31
图表 56: 百度人工智能全栈布局.....	31
图表 57: ERNIE 3.0 文心大模型架构示意图.....	32
图表 58: ERNIE 3.0 预训练数据集.....	32
图表 59: 全面的模型尺寸, 覆盖不同场景需求.....	33
图表 60: 百度面向特定任务的大模型.....	33
图表 61: 百度面向特定领域的大模型.....	33
图表 62: 飞桨重点产品示意图.....	34
图表 63: 百度文心一格 AI 艺术与创意辅助平台.....	34
图表 64: 百度文心百中产业级搜索系统.....	35
图表 65: 商汤科技推出日日新大模型体系.....	35
图表 66: 商汤 SenseChat 2.0 大语言模型能力全面升级.....	36
图表 67: 商汤 SenseMirage 3.0 文生图大模型全面升级.....	36

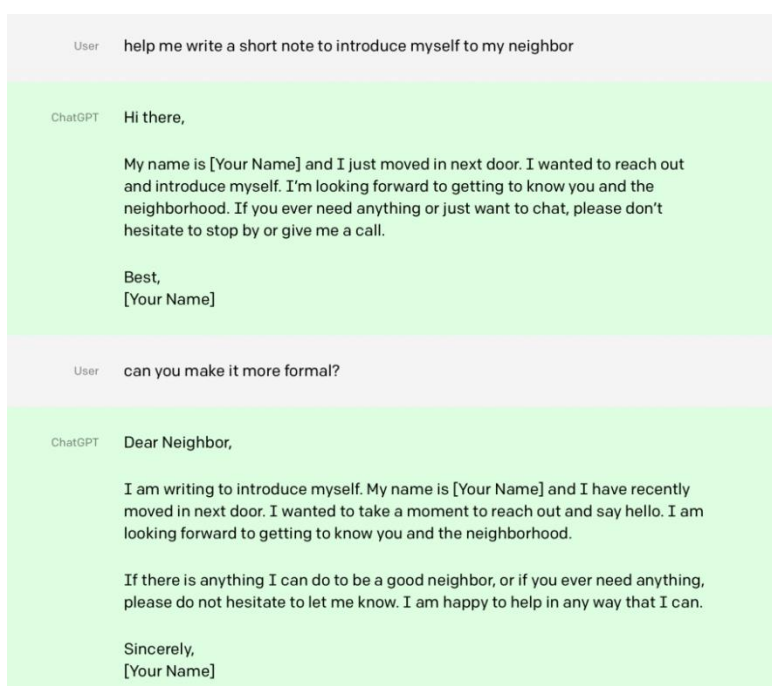
图表 68：华为云正式发布盘古大模型 3.0.....	37
图表 69：华为全栈核心技术.....	37
图表 70：微软发布 Microsoft 365 Copilot.....	38
图表 71：Microsoft 365 Copilot 技术原理.....	39
图表 72：Copilot 版 Word 使用示例.....	39
图表 73：Copilot 版 Excel 使用示例	39
图表 74：Copilot 版 PPT，将 Word 文档转换为幻灯片	40
图表 75：Copilot 版 PPT，为所有幻灯片生成演讲者备注	40
图表 76：Copilot 版 Outlook，自动管理和回复邮件	40
图表 77：Copilot 版 Teams，会议过程中提供会议信息	41
图表 78：Copilot 版 Teams，会议结束后总结用户错过的内容	41
图表 79：Copilot 版 Business Chat 使用示例，给予综合性答复	41
图表 80：Microsoft Security Copilot 的特点	42
图表 81：Microsoft Security Copilot 为微软各安全产品 AI 赋能	42
图表 82：Microsoft Security Copilot 基于 GPT-4 和微软安全专用模型.....	43
图表 83：Microsoft Security Copilot 集成了 GPT-4 和微软的安全能力.....	43
图表 84：赋能安全团队以机器的速度来解决问题.....	43
图表 85：Microsoft Security Copilot 使用案例	43
图表 86：相关公司业务情况.....	44

1、大模型涌现，ChatGPT 引爆 AI 奇点

1.1、ChatGPT 加速人工智能发展奇点来临

2022 年 11 月 OpenAI 发布 ChatGPT 聊天机器人，人工智能发展奇点加速来临。ChatGPT（Chat Generative Pre-trained Transformer）是由美国人工智能公司 OpenAI 开发的用于自然语言处理的大型预训练语言模型，由人工智能技术驱动的自然语言处理工具，于 2022 年 11 月 30 日发布，是一款全新的聊天机器人模型，可以根据用户的对话输入，产生出与其相关的回复，能够从文本输入中理解上下文因素，并生成有意义的句子回复，能够回答问题、承认错误、质疑不正确的前提和拒绝不适当的请求。

图表 1：ChatGPT 根据指示进行回复



资料来源：OpenAI，五矿证券研究所

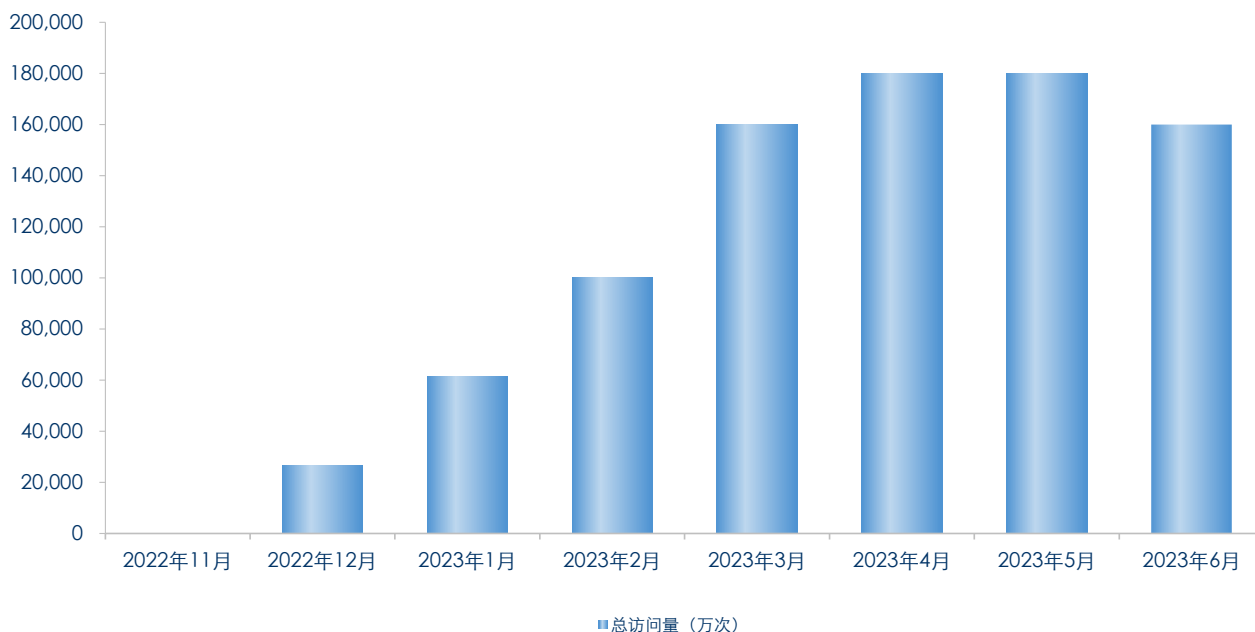
图表 2：ChatGPT 可以实现的功能



资料来源：通信产业网，五矿证券研究所

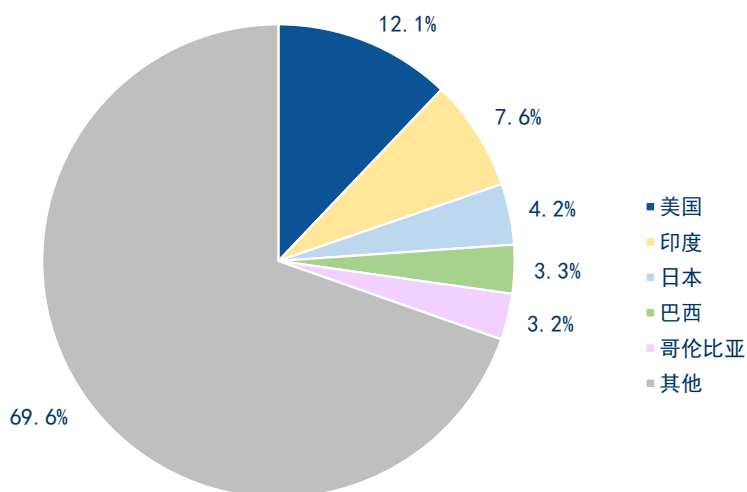
ChatGPT 发布后，Open AI 网站访问量爆发式增长。据 SimilarWeb 数据显示，过去 7 个月，网站 chat.openai.com 访问量快速增长，2023 年 6 月网站访问量达 16 亿次，平均访问时长约 8 分钟，每次访问页数 4.26 个。从受众群体地区来看，美国、印度、日本、巴西、哥伦比亚占比前五，分别为 12.12%、7.61%、4.17%、3.32%、3.16%。

图表 3: ChatGPT 用户访问量



资料来源: Similarweb, 五矿证券研究所

图表 4: 2023 年 6 月 ChatGPT 受众群体地区分布情况

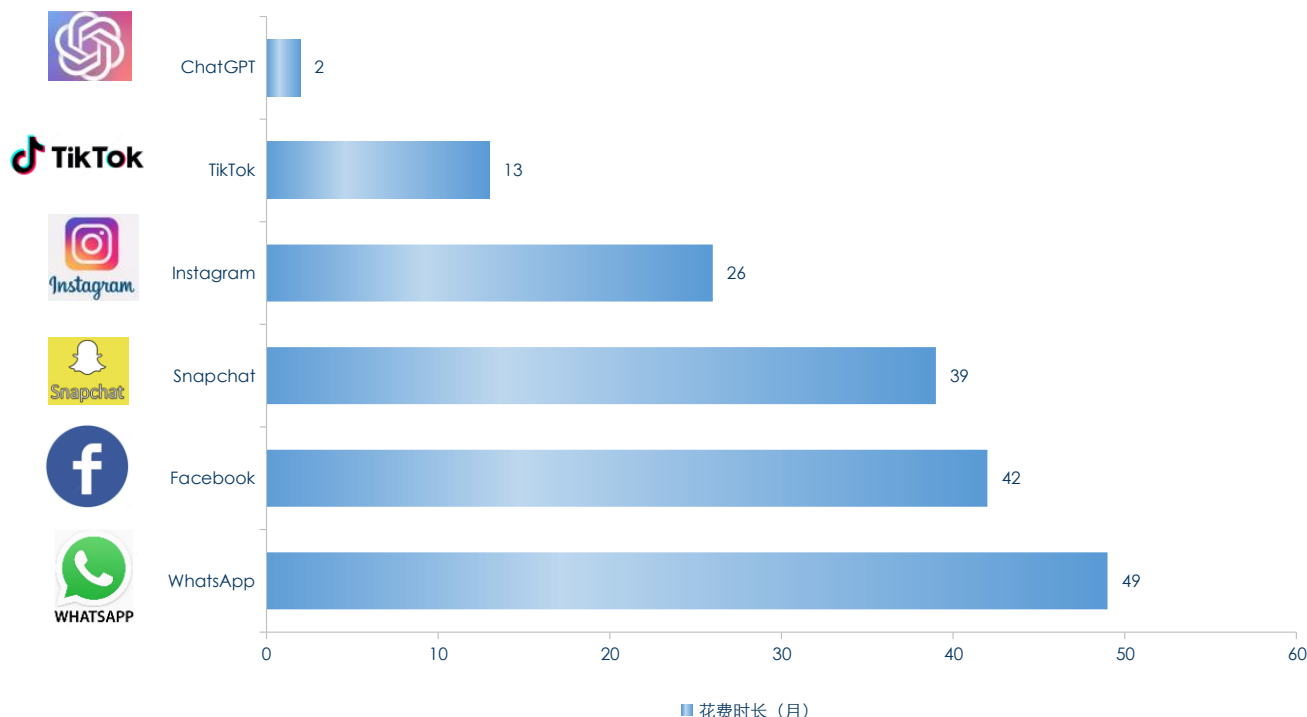


资料来源: Similarweb, 五矿证券研究所

ChatGPT 用户增速创历史新高，发布 5 天用户数突破 100 万，2 个月过亿。2022 年 11 月 30 日，OpenAI 推出 ChatGPT 聊天机器人产品。12 月 5 日，OpenAI 创始人 Sam Altman 发推称，ChatGPT 上线仅 5 天，用户数突破 100 万。至 2023 年 1 月末，ChatGPT 用户数突破 1 亿，距离发布时间仅 2 个月，成为史上用户数增长最快的消费者应用。

据英伟达在 ITF World 2023 演讲稿统计显示，海外版抖音 TikTok 发布后，大约用时 13 个月达到月活 1 亿用户；Instagram 用时 26 个月（2 年多）达到月活 1 亿用户；Facebook 用时 42 个月达到月活 1 亿用户。

图表 5：各应用程序达到 1 亿用户所需时间



资料来源：英伟达，五矿证券研究所

OpenAI 创始人背景深厚，与微软深度合作。2015 年 12 月，OpenAI 公司在美国旧金山成立，由人工智能领域的顶尖科学家和企业家联合创立的一家非盈利的人工智能研究公司。联合创始人包括美国创业孵化器 Y Combinator 总裁 Sam Altman（现 CEO）、马斯克（Elon Musk）、全球在线支付平台 PayPal 联合创始人 Peter Thiel、Ilya Sutskever（师从神经网络之父 Geoffrey Hinton，现首席科学家）等；

2019 年 3 月，OpenAI 创立了 OpenAI LP 公司，从非盈利公司转型至有限盈利公司；2019 年 7 月，微软向 OpenAI 注资 10 亿美元，同时微软 Azure 成为 OpenAI 的“独家”云计算服务提供商；

2023 年 1 月，微软宣布与 OpenAI 扩大合作关系，其旗下所有产品将全线整合 ChatGPT，除搜索引擎必应、Office 外，微软还将在云计算平台 Azure 中整合 ChatGPT，作为 OpenAI 的独家云提供商，Azure 将为研究、产品和 API 服务中的所有 OpenAI 工作负载提供支持。

图表 6: OpenAI 发展历程与股东情况

OpenAI: 宗旨——实现安全的通用人工智能(AGI)	
2015年12月11日	成立OpenAI • 非营利的人工智能研究公司 • 联合创始人: 萨姆·奥特曼 (Sam Altman)、彼得·泰尔 (Peter Thiel)、里德·霍夫曼 (Reid Hoffman) 和埃隆·马斯克 (Elon Musk) 等人 • 联合主席: Sam Altman、Elon Musk • 研究总监: Ilya Sutskever • CTO: Greg Brockman
2019年3月	成立OpenAI LP • OpenAI 的营利实体 • 从非盈利公司转型为有限盈利公司
2019年7月	微软首次注资OpenAI 10亿美元 
2023年1月	微软与Open AI扩大合作关系 将向Open AI开展多年、数十亿美元的投资 微软旗下所有产品将全线整合ChatGPT

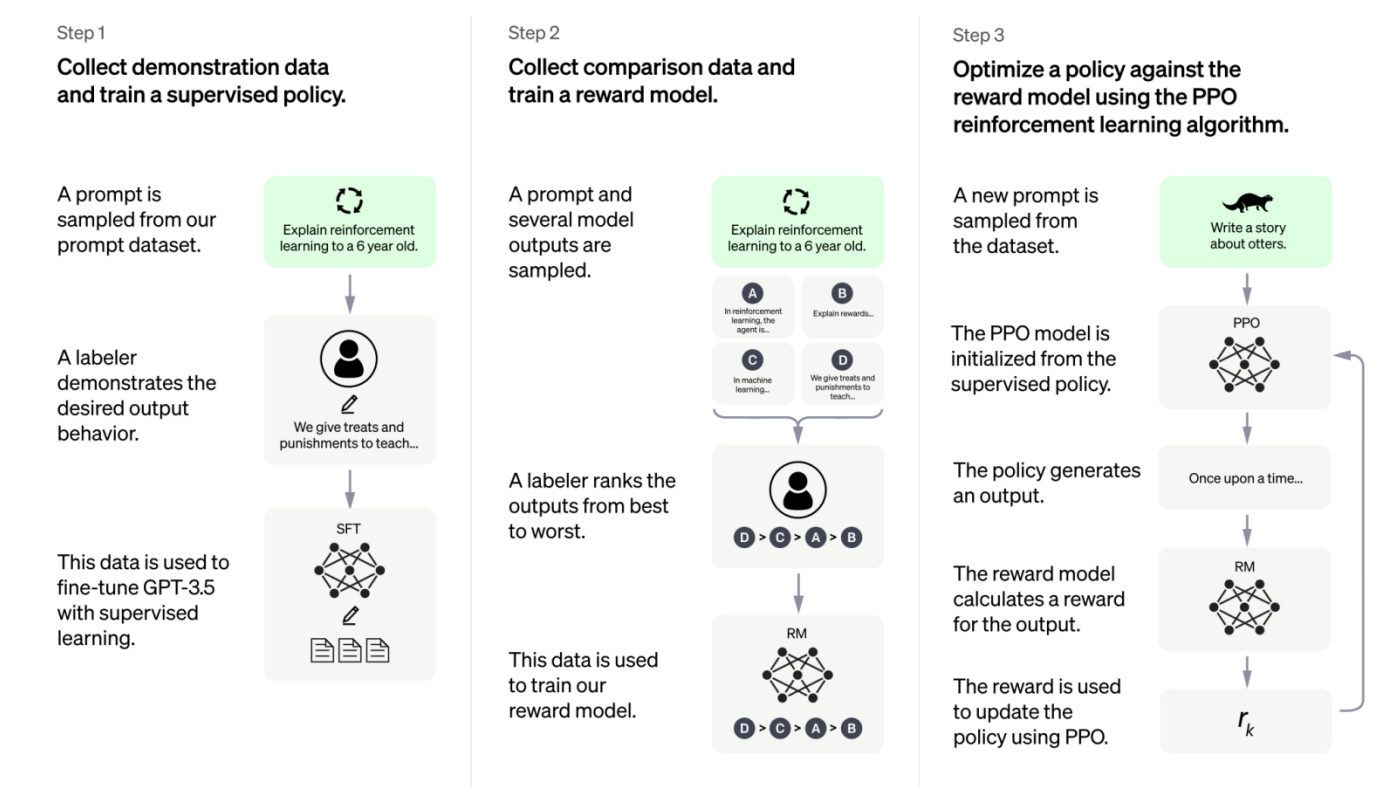
资料来源: OpenAI, 五矿证券研究所

ChatGPT 采用基于 GPT-3.5 架构的大型语言模型, 引入人类反馈强化学习技术 (RLHF) 训练模型。从技术原理看, ChatGPT 是基于 GPT-3.5 (Generative Pre-trained Transformer 3.5) 架构开发的对话 AI 模型, 是 InstructGPT 的兄弟模型, 采用“预训练+微调”的模型训练方式, 引入 RLHF 技术对 ChatGPT 进行训练, 利用强化学习方法从人类标注者的反馈中学习。

训练过程可分为三个步骤:

- 1) **训练监督学习模型:** 首先 ChatGPT 从 prompt 数据库中抽取若干问题并向模型解释强化学习机制, 随后人类标注者撰写期望的输出值, 对模型给予特定奖励或惩罚引导教育, 最后通过监督学习微调 GPT-3.5 模型;
- 2) **收集数据并训练奖励模型:** 从 prompt 数据库中取样, 并由人类标注者们对模型输出的多个结果进行投票, 按质量排序, 采用排序后的数据结果用于训练奖励模型;
- 3) **采用近端策略优化 (PRO) 强化学习微调模型:** 近端策略优化 (Proximal Policy Optimization) 是 2017 年 OpenAI 发布的强化学习算法, 首先通过监督学习生成初始 PRO 模型, 由奖励模型对回答打分后, 将反馈结果优化和迭代初始 PRO 模型, 通过多次优化迭代获得质量更高的模型。

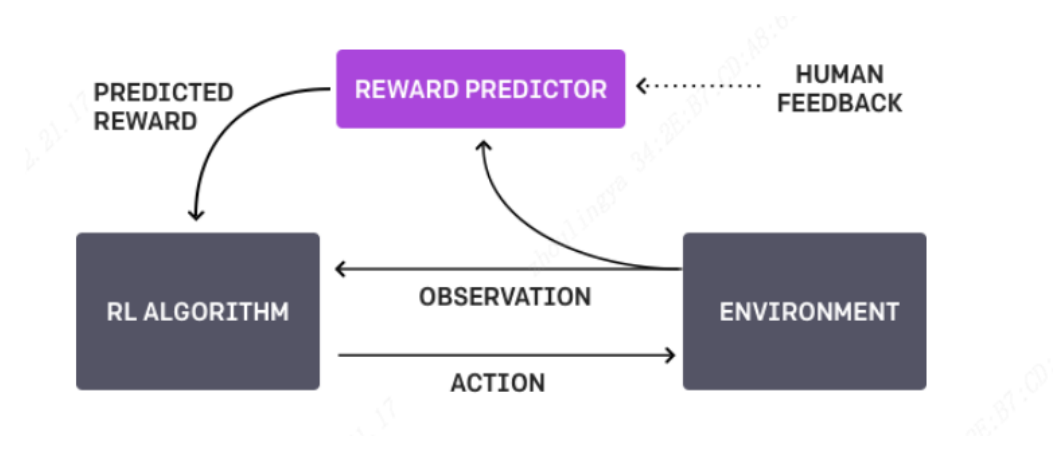
图表 7: ChatGPT 模型原理



资料来源: OpenAI, 五矿证券研究所

RLHF (Reinforcement Learning from Human Feedback) 是通过人类反馈强化学习技术优化语言模型。将人类的反馈纳入训练过程, 为机器提供了一种自然的、人性化的互动学习过程, 以更广泛的视角和更高的效率学习, 允许人类直接指导机器, 并允许机器掌握明显嵌入人类经验中的决策要素。

图表 8: RLHF (人类反馈强化学习技术) 训练模型技术原理



资料来源: Open AI, 五矿证券研究所

1.2、大模型军备竞赛，开启 AI 未来时刻

人工智能是引领未来的新兴战略性技术，是驱动新一轮科技革命和产业革命的重要力量。1950年艾伦·图灵（Alan Turing）提出了著名的“图灵测试”，给出了判定机器是否具有“智能”的实验方法，即机器是否能够模仿人类思维方式来“生成”内容继而与人交互，人们开始关注人类智能与机器之间的关系；直至20世纪50年代中期，由于计算机的出现与发展，人们开始了具有真正意义的人工智能的研究。随着人工智能越来越多地被应用于内容创作，人工智能生成内容（Artificial Intelligence Generated Content, AIGC）概念悄然升起。

AIGC 的发展可大致分为三个阶段：

早期萌芽阶段（1950s-1990s）：受限于当时的科技水平，AIGC 仅限于小范围实验。在1956年的达特茅斯会议上，“人工智能”的概念被首次提出，在之后的十余年内人工智能迎来了发展史上的第一个小高峰，研究者们纷纷涌入，取得了一批瞩目的成就。

沉淀积累阶段（1990s-2010s）：AIGC 从实验性向实用性转变。2006年，深度学习算法取得重大突破，同时图形处理器（GPU）、张量处理器（TPU）等算力设备性能不断提升，互联网使数据规模快速膨胀并为各类人工智能算法提供了海量训练数据，使人工智能发展取得了显著的进步，同时受限于算法瓶颈，应用仍然有限，效果有待提升。

快速发展阶段（2010s-至今）：生成内容百花齐放，效果逐渐逼真直至人类难以分辨。自2014年起，随着以生成式对抗网络（GAN）为代表的深度学习算法的提出和迭代更新，AIGC 迎来了新时代。

图表 9：AIGC 发展历程



资料来源：中国信息通信研究院《人工智能生成内容（AIGC）白皮书》，五矿证券研究所

国内外厂商纷纷加入 AI 大模型军备竞赛。在 AI 大浪潮下，国内外科技厂商纷纷入局，密集上线大模型，包括自然语言处理模型、计算机视觉大模型、多模态大模型等多种类型。谷歌、OpenAI 等国际科技巨头处于领先地位，中国阿里巴巴、华为、百度等公司均推出了自己的 AI 大模型产品。2023 年 3 月百度发布文心一言，4 月华为发布盘古大模型，阿里发布通义千问大模型，商汤科技公布“日日新 SenseNova”大模型体系，5 月科大讯飞发布星火大模型，此外多家上市公司也开始布局 AI 大模型领域，积极探索 AI 大模型技术的应用。

从模型参数量来看，模型参数量呈现上升趋势。多模态大模型参数量相对较大，多在千亿级以上，PaLM-E、Visual ChatGPT、GPT-4 的模型参数量分别为 5620，1750 和 1750 亿。国内大数据模型中，模型参数量达千亿的模型有文心大模型（2600 亿），盘古 NLP（千亿级）等。

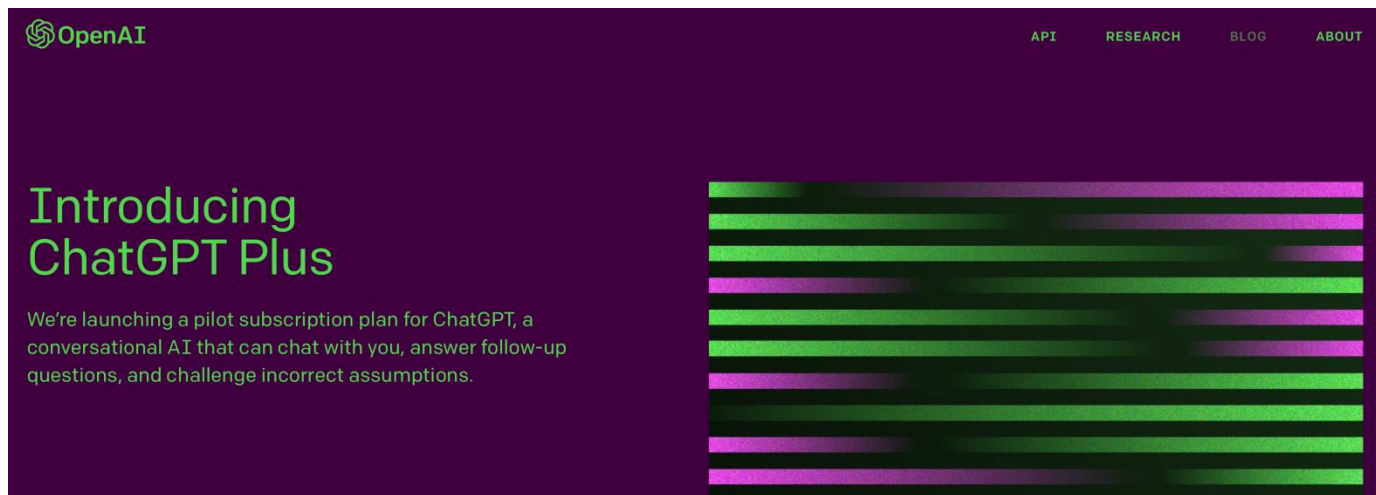
图表 10：2023 年国内外发布的主要 AI 大模型梳理

发布机构	发布时间	模型名称	模型参数量（亿）	模型类型	是否开源
Google	2023/2/10	Vit-22B	220	计算机视觉	否
复旦大学	2023/2/20	MOSS	160	多模态	是
Facebook	2023/2/24	LLaMA	650	自然语言处理	是
Google	2023/3/6	PaLM-E	5620	多模态	否
Microsoft	2023/3/8	Visual ChatGPT	1750	多模态	否
Together Computer	2023/3/11	GPT-NeoXT-Chat-Base-20B	200	自然语言处理	是
斯坦福大学	2023/3/13	Stanford Alpaca	70	自然语言处理	是
清华大学	2023/3/14	ChatGLM-6B	62	自然语言处理	是
OpenAI	2023/3/14	GPT-4	1750	多模态	否
Intel	2023/3/16	INT8 GPT-J 6B	60	自然语言处理	是
databricks	2023/3/24	Dolly	60	自然语言处理	是
Cerebras Systems	2023/3/28	Cerebras GPT	130	自然语言处理	是
HPC-AI Tech	2023/3/29	ColossalChat	130	自然语言处理	是
Lightning AI	2023/3/29	Lit-LLaMA	650	自然语言处理	是
Bloomberg	2023/3/30	BloombergGPT	500	自然语言处理	否
LM-SYS	2023/3/31	Vicuna-13B	130	自然语言处理	是
UC Berkeley	2023/4/3	Koala-13B	130	自然语言处理	否
Facebook	2023/4/5	Segment Anything Model	10	计算机视觉	是
腾讯	2023/4/6	LVDM	0	多模态	是
databricks	2023/4/12	dolly-v2	120	自然语言处理	是
Stability AI	2023/4/14	Stable Diffusion XL	23	计算机视觉	是
Meta	2023/4/14	DINOv2	11	计算机视觉	是
Microsoft	2023/4/17	LLaVA	130	多模态	是
Stability AI	2023/4/20	StableLM	1750	自然语言处理	是
Stability AI	2023/4/28	StableVicuna-13B	130	自然语言处理	是
上海人工智能实验室	2023/4/28	LLaMA Adapter V2	650	多模态	是
Google	2023/5/10	PaLM 2	3400	自然语言处理	是
Google	2023/5/17	DoReMi	80	自然语言处理	是
清华大学	2023/5/17	VisualGLM-6B	78	多模态	是
Meta	2023/5/22	LIMA	650	自然语言处理	是

资料来源：DataLearner，五矿证券研究所

ChatGPT 开启商业变现。当地时间 2023 年 2 月 1 日，美国人工智能公司 OpenAI 推出 ChatGPT 付费订阅版 ChatGPT Plus，每月收费 20 美元。订阅者可在免费服务基础上获得：1) 高峰时段免排队访问；2) 更快的相应时间；3) 优先使用新功能和优化等权益。2023 年 3 月 1 日，OpenAI 官方宣布正式开放 ChatGPT API。我们认为，付费订阅版 ChatGPT Plus 的推出，标志着 ChatGPT 正式商业化，开启产品向商业变现之路。

图表 11：OpenAI 推出 ChatGPT Plus 订阅服务



资料来源：OpenAI，五矿证券研究所

OpenAI 盈利模式：开放付费 API 接口，提供多种模型。目前 OpenAI 提供 DALL·E、GPT-3、Codex、Content filter 的 API 接口，用于执行图像模型、语言模型，每种模型又细分为多种子模型型号，每种型号有不同的功能和价位。用户可以根据自身业务需求选择对应的模型，API 接口根据模型类型、业务量等指标进行收费。图像模型根据图片分辨率的不同按张数收费，语言模型基于子模型型号按字符数收费，微调模型则包括训练和使用两部分的价格。

图像模型：基于 DALL·E 实现文本生成图像。通过使用 DALL·E 的 API，可实现 3 种与图像交互的方法：1) 基于文本提示从头开始创建图像；2) 基于新文本提示创建对现有图像的编辑；3) 创建现有图像的变体。生成的图像有 256x256、512x512、1024x1024 像素三种规格，尺寸较小的图像生成速度更快。用户只需简单输入一些文字描述，即可将文字描述的场景以图片的形式真实展现出来。据 Open AI 官网披露，目前已有 300 万人使用 DALL·E 模型扩展创造力，加快工作流程。

语言模型：不同功能和价格的语言模型。目前 OpenAI 提供三类语言模型，分别是 GPT-3、Codex、Content filter。其中 GPT-3 能够理解自然语言提示，并生成自然语言，拥有 1750 亿个参数；Codex 是 GPT-3 模型的后代，可以理解并生成代码，将自然语言转换成代码，并精通多种编程语言；Content filter 是检测文本是否敏感或不安全。

图表 12：文本生成图像模型

PROMPT	GENERATION
a white siamese cat	
a close up, studio photographic portrait of a white siamese cat that looks curious, backlit ears	

资料来源：OpenAI，五矿证券研究所

图表 13：不同功能的语言模型

模型名称	概况
GPT-3	能够理解和生成自然语言的模型
Codex	能够理解和生成代码的模型，同时可以实现将自然语言转换为代码
Content filter	检测文本是否敏感或不安全

资料来源：OpenAI，五矿证券研究所

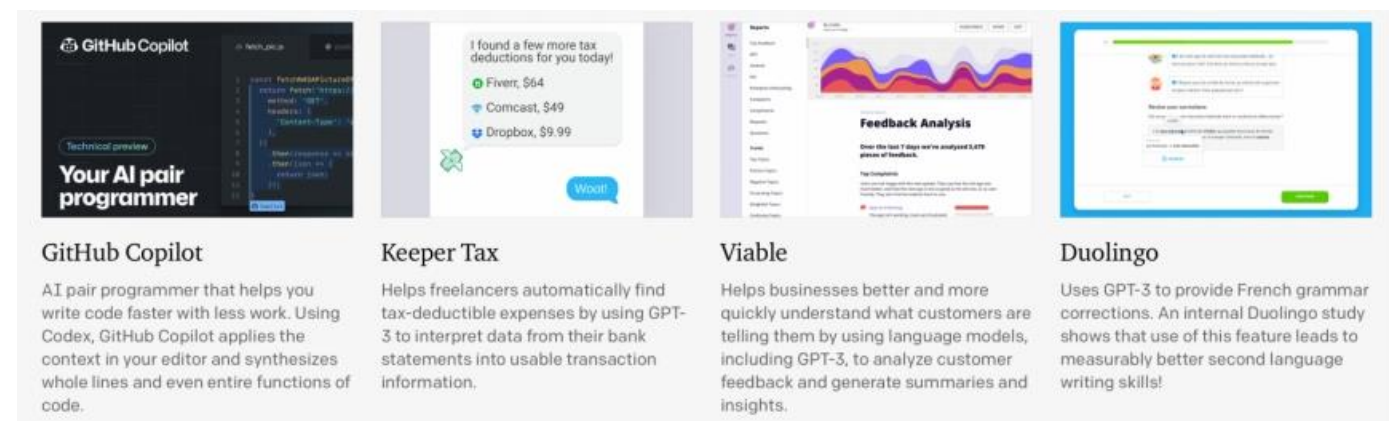
图表 14：Codex 语言模型示例

TEXT TO SQL TRANSLATION	CALLING AN API VIA NATURAL LANGUAGE	CODE CONTINUATION
<pre>Table customers, columns = [CustomerId, FirstName, LastName, Company, Address, City, State, Country, PostalCode, Phone, Fax, Email, SupportRepId, TotalSpend] Create a MySQL query for all customers in Texas who have spent over five thousand dollars. query = "SELECT * FROM customers WHERE State = 'TX' AND TotalSpend > 5000"</pre>		

资料来源：OpenAI，五矿证券研究所

Open API 已在多个应用程序中应用，从帮助人们学习新语言到解决复杂分类问题。GitHub Copilot 是一款 AI 编程器，由 OpenAI Codex 语言模型提供支持，帮助开发人员更快地编写代码；Keeper Tax 是一款帮助独立承包商和自由职业者纳税的 App，由 GPT-3 模型提供支持，可以实现提取文本并对交易分类；Viable 通过使用 GPT-3 语言模型，帮助公司从客户反馈中获得见解，智能读取用户评论；Duolingo 使用 GPT-3 语言模型进行语法修正。

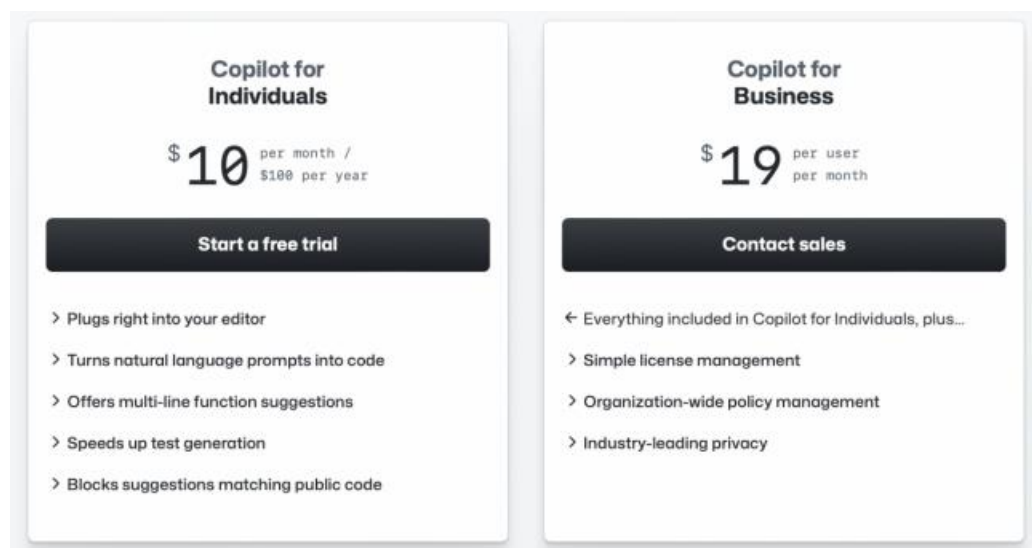
图表 15: OpenAI API 已在多个应用程序中应用



资料来源: OpenAI, 五矿证券研究所

GitHub Copilot: AI 编程工具软件，提高开发效率。2021 年 6 月 29 日，微软与 OpenAI 共同推出了一款 AI 编程工具 GitHub Copilot，2022 年 6 月 22 日，Copilot 正式上线，定价每月 10 美元或每年 100 美元，对学生用户和流行开源项目的维护者免费提供。GitHub Copilot 由 OpenAI Codex 提供支持，这是一个由 OpenAI 创建的生成预训练语言模型，可以根据上文提示为程序员自动编写下文代码。使开发人员更快地编写代码，专注于业务逻辑而不是样板，提高工作效率。

图表 16: GitHub Copilot 价格情况



资料来源: GitHub Copilot, 五矿证券研究所

Keeper Tax: AI 税务软件。通过定制版 GPT-3 模型来识别账单上的各种数据，以帮助用户找到可以免税的费用，连接银行自动发现隐藏的税收减免，通过搜索交易历史，发现容易错过的可抵扣税款的业务支出和其他税收减免。据 Keeper Tax 官网披露，Keeper 用户平均每年发现价值 1249 美元的额外税务冲销，目前 Keeper 已帮助超过 50000 名灵活雇员和创意经济工作者发现了价值 7000 万美元的税务冲销。

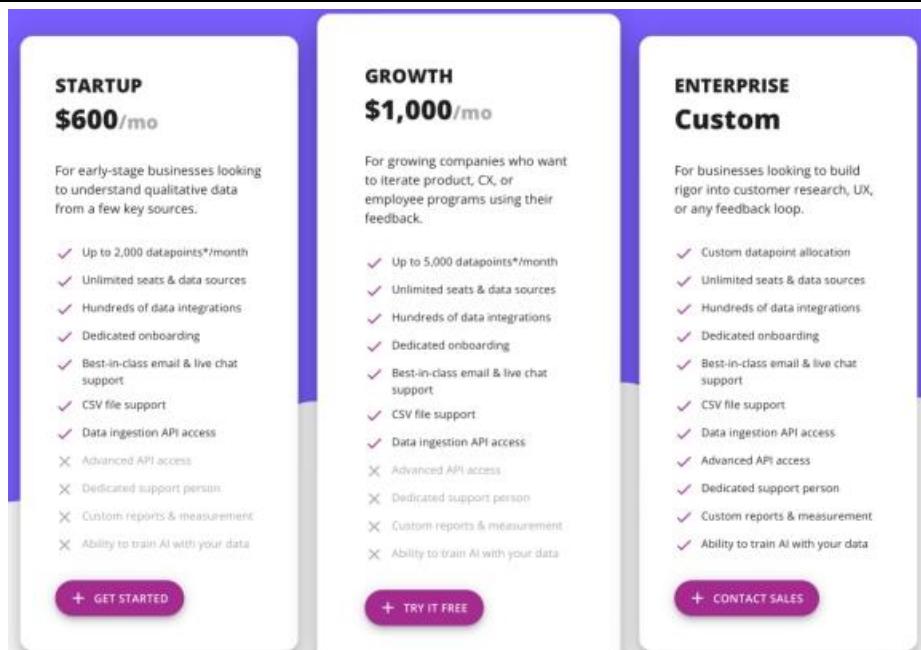
图表 17: Keeper Tax 价格情况



资料来源: Keeper Tax, 五矿证券研究所

Viable: AI 企业分析应用软件。Viable 是第一家使用人工智能研究分析初创公司,运用 OpenAI 的 GPT-3 语言模型,可以帮助公司从客户反馈中获得见解,智能读取用户评论,衡量其产品市场适合度。Viable 推出 3 款产品,对于初创企业,仅需了解几个关键数据来源的定性数据,每月收费 600 美元,每月最多 2000 个数据点;对于成长型企业,希望通过客户反馈迭代产品等,每月收费 1000 美元,每月最多 5000 个数据点;对于成熟企业,获得定制报告与测量,利用数据训练 AI,可选择定制服务。

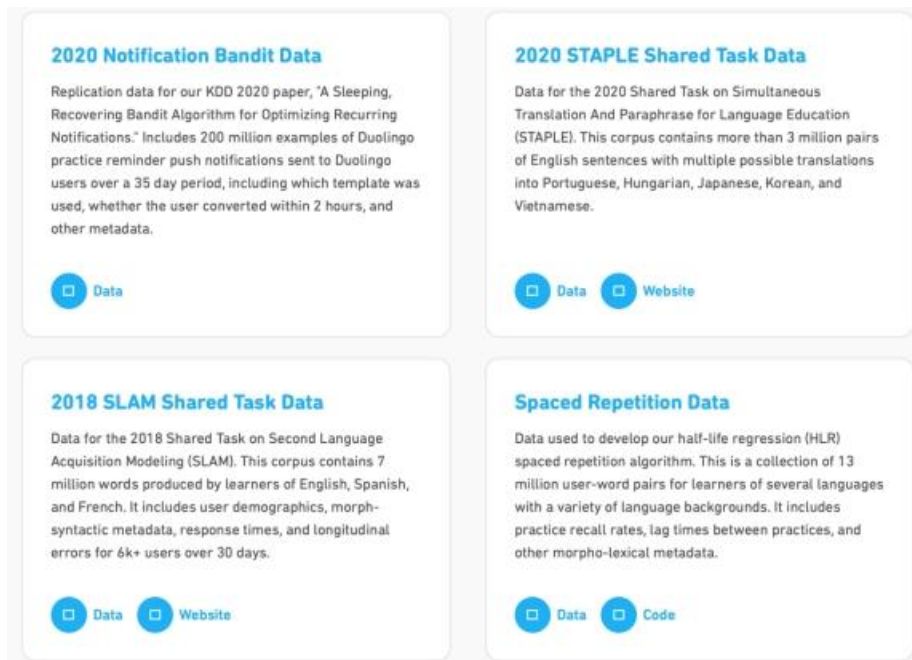
图表 18: Viable 价格情况



资料来源: Viable, 五矿证券研究所

Duolingo: AI 语言学习工具软件。 Duolingo 是一款语言学习工具软件，提供 40 多种语言课程，包括英语、日语、韩语、法语、粤语等大众语言，也包括威尔士语、纳瓦荷语等小语种和濒危语言。Duolingo 使用 OpenAI 的 GPT-3 进行语法修正。

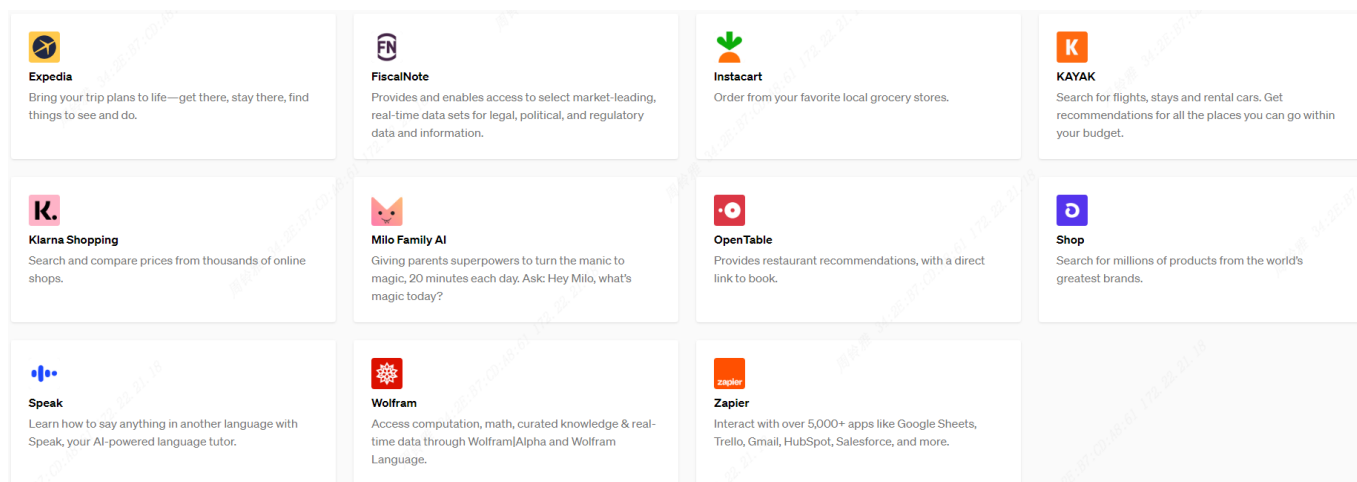
图表 19: Duolingo 的数据与算法工具示例



资料来源: Duolingo, 五矿证券研究所

推出 ChatGPT Plugins 插件功能，拓宽场景边界。当地时间 2023 年 3 月 23 日，OpenAI 推出 ChatGPT Plugins 插件功能，宣布 ChatGPT 中初步实现对插件的支持。ChatGPT Plugins 插件专门为大语言模型设计，以安全为核心原则，能够帮助 ChatGPT 访问最新的信息，运行计算，以及使用第三方服务，能够使 ChatGPT 参与开发者定义的 API 互动，增强 ChatGPT 的能力，拓宽场景边界。

图表 20: 第一批 ChatGPT Plugins 合作伙伴



资料来源: OpenAI, 五矿证券研究所

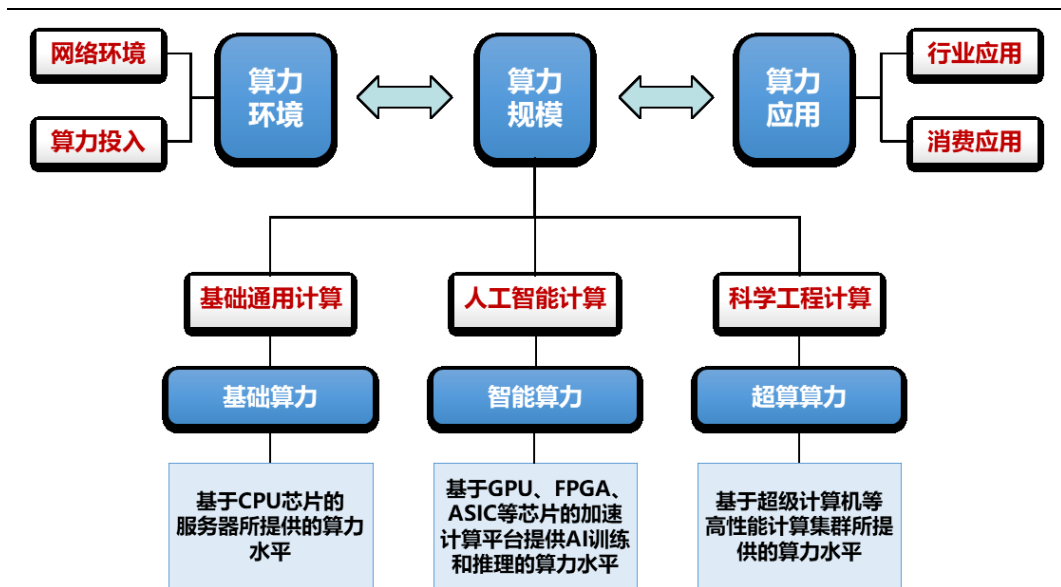
2、算力质变，英伟达打造 AI 时代引擎

2.1、算力需求指数级增长，加速计算深度变革

算力是设备通过处理数据，实现特定结果输出的计算能力。算力实现的核心是 CPU、GPU、FPGA、ASIC 等各类计算芯片，并由计算机、服务器、高性能计算集群和各类智能终端等承载，海量数据处理和各种数字化应用都离不开算力的加工和计算。算力数值越大代表综合计算能力越强，常用的计量单位是每秒执行的浮点数运算次数（Flops， $1\text{EFlops}=10^{18}\text{Flops}$ ）。据信通院测算， 1EFlops 约为 5 台天河 2A 超级计算机，或者 25 万台主流双路服务器，或者 200 万台主流笔记本的算力输出。

算力可分为基础算力、智能算力和超算算力三部分，分别提供基础通用计算、人工智能计算和科学工程计算。其中，基础通用算力主要基于 CPU 芯片的服务器所提供的计算能力；智能算力主要是基于 GPU、FPGA、ASIC 等芯片的加速计算平台提供人工智能训练和推理的计算能力；超算算力主要是基于超级计算机等高性能计算集群所提供的计算能力。

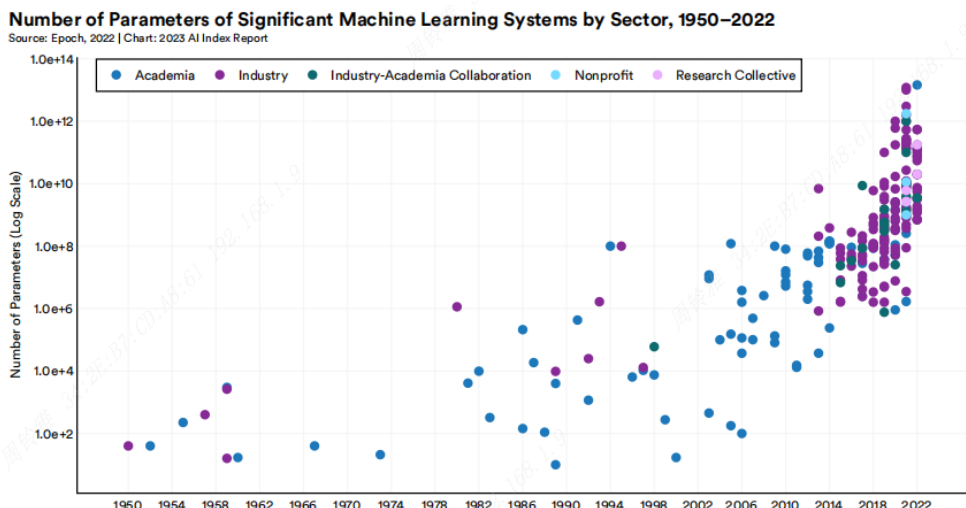
图表 21：算力发展总体框架



资料来源：中国信息通信研究院《中国算力发展指数白皮书》，五矿证券研究所

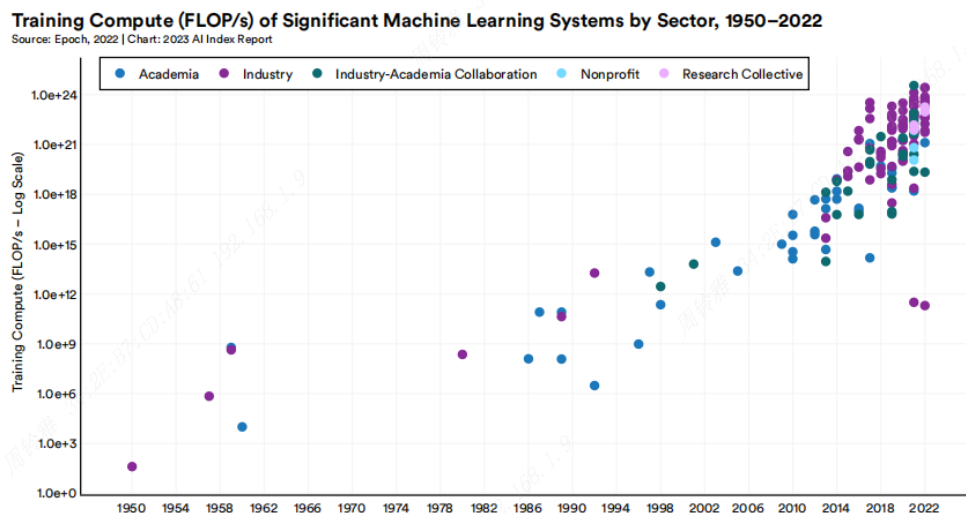
算力需求指数级增长，大模型参数指数级增长。经过大规模预训练的大模型，能够在各种任务中达到更高的准确性、降低应用的开发门槛、增强模型泛化能力等。随着海量数据的持续积累、人工智能算力多样化与算法的突破，大模型参数规模呈现指数级增长，先后经历了预训练模型、大规模预训练模型、超大规模预训练模型三个阶段，参数量实现百万亿级突破。与此同时，算力需求也呈现指数级增长。从行业分布上看，大模型的应用领域逐步从学术拓展至产业，2010 年后产业界对大模型的应用与算力需求显著增长，成为主导力量。

图表 22: 1950-2022 年机器学习训练参数变化情况



资料来源:《Artificial Intelligence Index Report 2023》, 五矿证券研究所

图表 23: 1950-2022 年机器学习算力变化情况

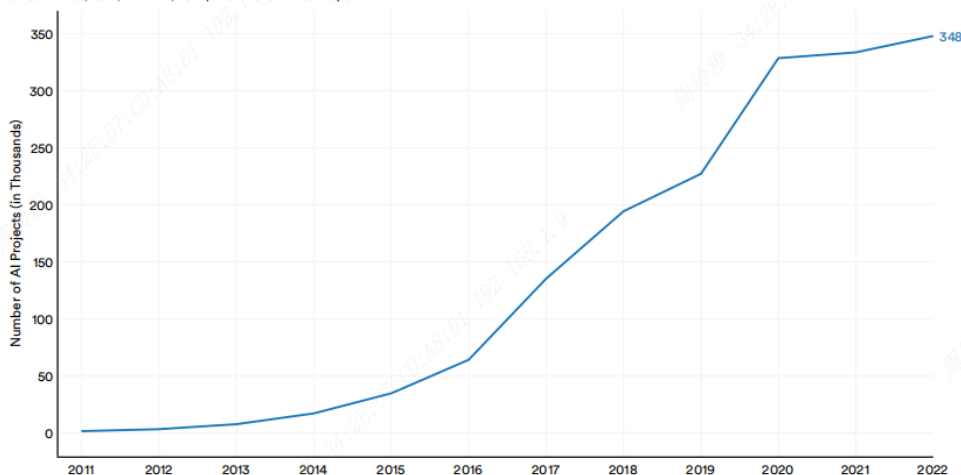


资料来源:《Artificial Intelligence Index Report 2023》, 五矿证券研究所

AI 期刊论文与开源项目快速增长。ChatGPT 引发了新一轮的 AI 浪潮,越来越多的研究机构与公司加大对 AI 的研发投入。据斯坦福大学发布的《2023 年 AI 指数报告》统计,人工智能论文的总数自 2010 年呈现翻倍增长,从 2010 年的 20 万篇增长到 2021 年的近 50 万篇(49601),模式识别、机器学习和计算机视觉是人工智能领域研究的热门话题,从国家分布看,中国保持领先地位,2021 年占比为 39.8%,其次是欧盟和英国(15.1%)与美国(10%)。AI 开源项目方面,据 GitHub 统计,自 2011 年以来人工智能相关的 GitHub 项目稳步增长,从 2011 年的 1536 个增长至 2022 年 34.79 万个,从国家分布看,2022 年印度软件开发人员占比 24.2%,其次是欧盟和英国(17.3%)与美国(14%)。

图表 24：2011-2022 年 GitHub AI 项目数量情况

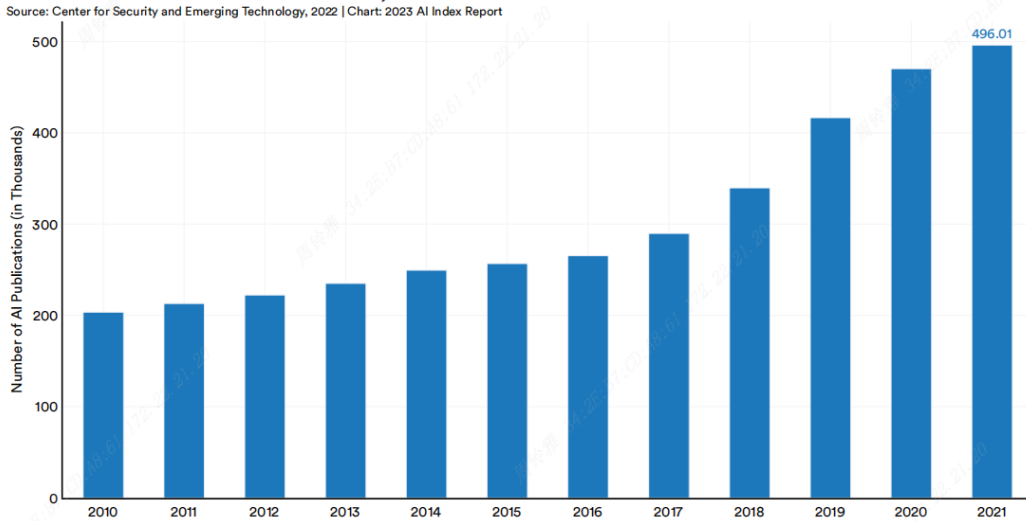
Number of GitHub AI Projects, 2011-22
Source: GitHub, 2022; OECD AI, 2022 | Chart: 2023 AI Index Report



资料来源：《Artificial Intelligence Index Report 2023》，GitHub，五矿证券研究所

图表 25：2010-2021 年全球 AI 期刊论文发表数量

Number of AI Publications in the World, 2010-21
Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report

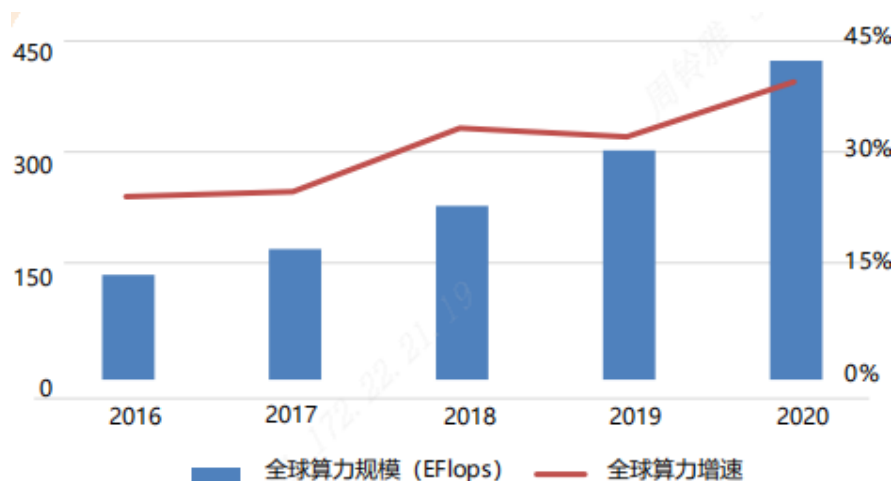


资料来源：《Artificial Intelligence Index Report 2023》，五矿证券研究所

全球算力规模不断增长。据 IDC 统计，2020 年全球算力总规模达到 429EFlops，同比增长 39%，其中基础算力规模（FP32²，FP32 为单精度浮点数，采用 32 位二进制来表达一个数字，常用于多媒体和图形处理计算）为 313EFlops³，智能算力规模（换算为 FP32）为 107EFlops⁴，超算算力规模（换算为 FP32）为 9EFlops⁵。据 IDC 预测，未来五年全球算力规模将以超过 50% 的速度增长，到 2025 年整体规模将达到 3300Eflops。

全球主要国家和地区纷纷加快算力布局。算力水平方面，据 IDC 统计 2021 年美国、中国、欧洲、日本在全球算力规模中的份额分别为 36%、31%、11% 和 6%。

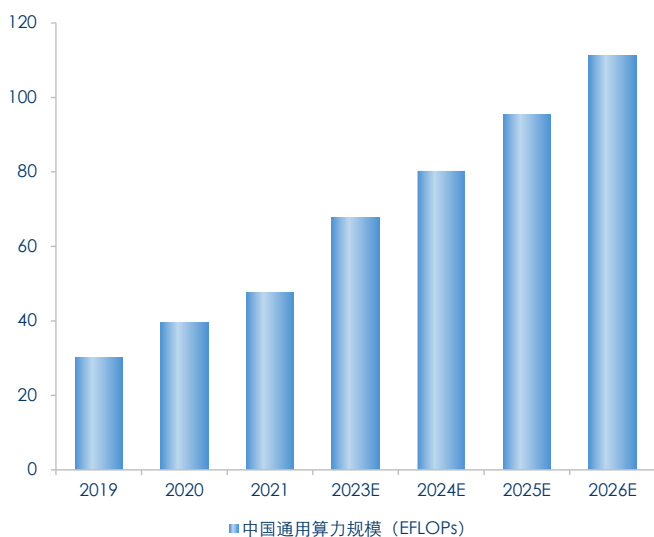
图表 26：全球算力规模发展趋势



资料来源：中国信息通信研究院《中国算力发展指数白皮书》，IDC，Gartner，TOP500，五矿证券研究所

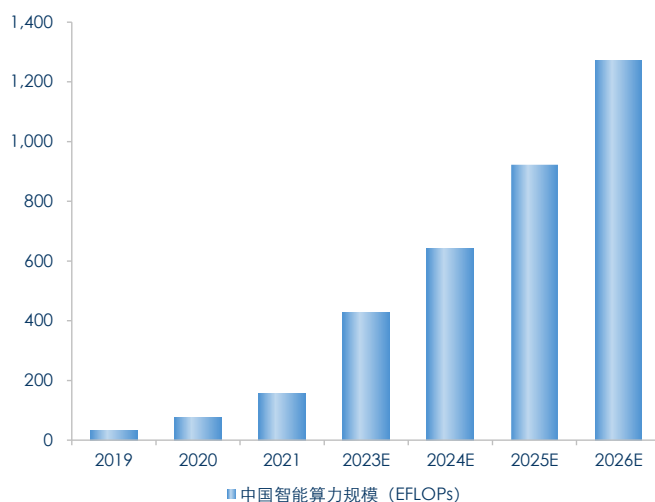
中国智能算力规模持续扩大。据 IDC 统计与预测,2021 年中国智能算力规模达 155.2EFLOPS，2022 年智能算力规模将达到 268.0 EFLOPS，预计到 2026 年智能算力规模将进入每秒十万亿亿次浮点计算（ZFLOPS）级别，达到 1271.4EFLOPS。

图表 27：中国通用算力规模及预测



资料来源：IDC，五矿证券研究所

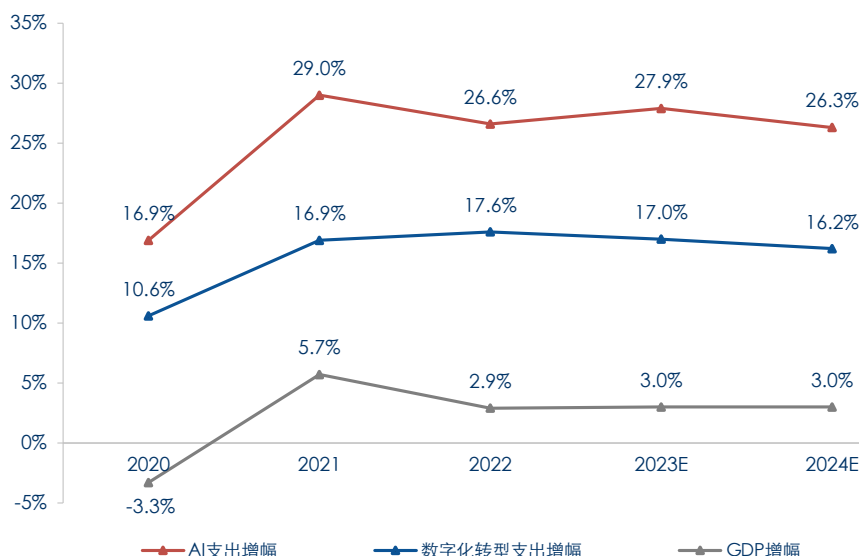
图表 28：中国智能算力规模及预测



资料来源：IDC，五矿证券研究所

全球企业持续增加对 AI 的投资。伴随数字经济的持续发展，为满足企业内部发展需求和外部市场需求，企业一直大力投资数字化转型相关技术。据 IDC 统计，全球企业在包括软件、硬件和服务在内的人工智能（AI）市场的技术投资从 2019 年的 612.4 亿美元增长至 2021 年的 924 亿美元，有望到 2025 年突破 2000 亿美元，增幅高于企业数字化转型（DX）支出整体增幅。

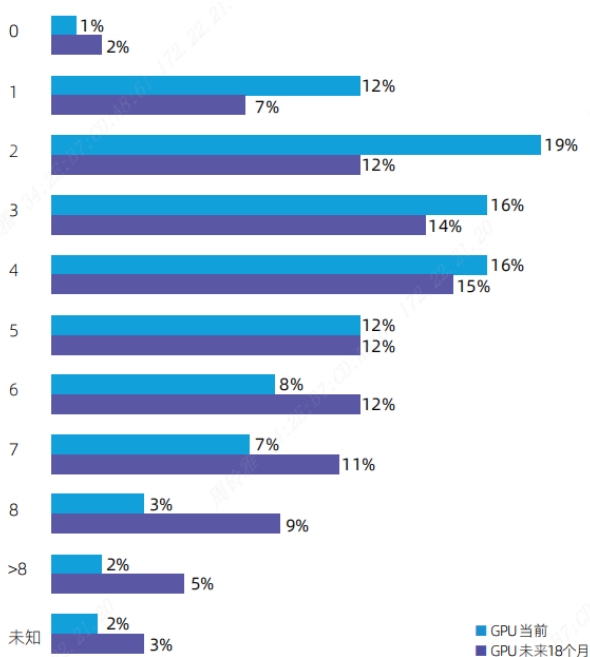
图表 29：2020-2024 年全球人工智能、数字化转型支出及 GDP 增长趋势预测



资料来源：IDC，世界银行，五矿证券研究所

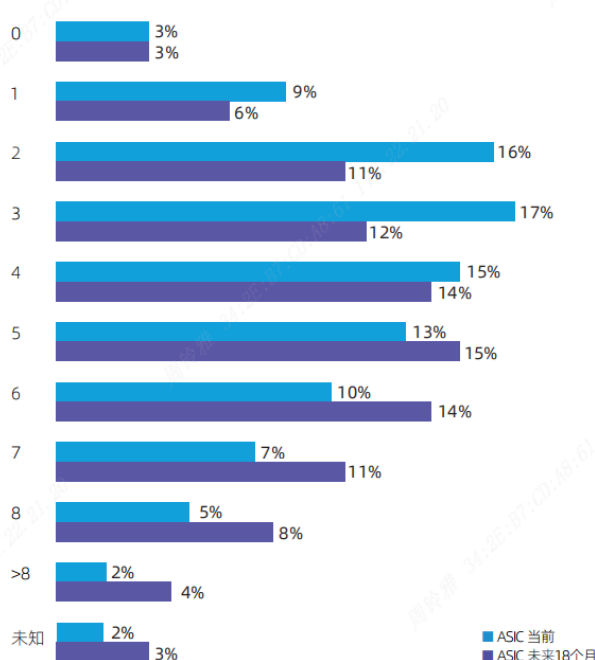
全球人工智能芯片搭载率将持续增长，低能耗为大势所趋。算力是实现 AI 产业化的核心力量，AI 产业技术不断提升，产业商业化应用加速落地，推动全球人工智能芯片市场高速增长，据 IDC 预测，至 2025 年全球人工智能芯片市场规模将达 726 亿美元。人工智能算力规模的快速增长将刺激更大的人工智能芯片需求，未来 18 个月，人工智能服务器 GPU、ASIC 和 FPGA 的芯片搭载率均会上升。目前中国市场主要以 GPU 为主，实现数据中心计算加速，市场占有率达 89%，GPU 芯片多用于图形图像处理、复杂的数学计算等场景。此外 ASIC，FPGA，NPU 等非 GPU 芯片市场占有率超过 10%，其中，NPU 较以往具有明显增长，NPU 芯片设计逻辑相对简单，常用于边侧和端侧的模型推理。

图表 30：全球人工智能服务器 GPU 芯片搭载率



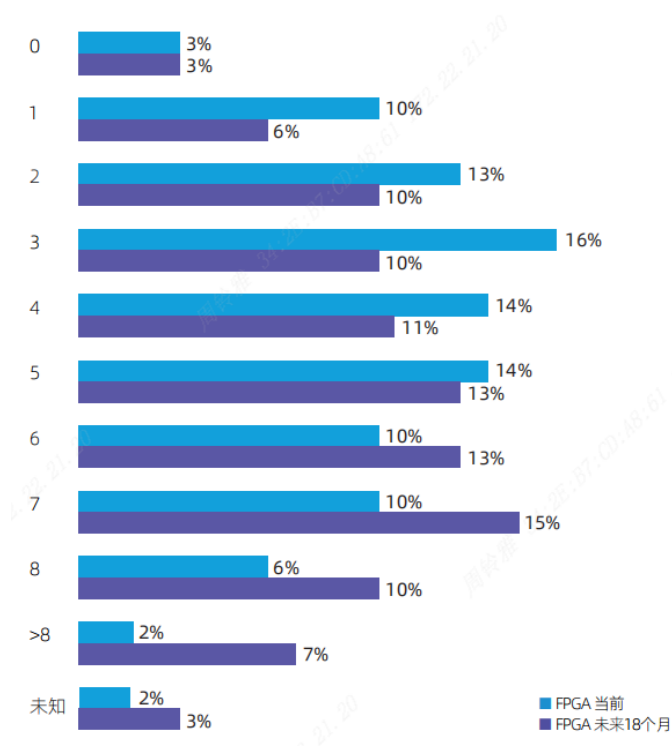
资料来源：IDC，五矿证券研究所

图表 31：全球人工智能服务器 ASIC 芯片搭载率



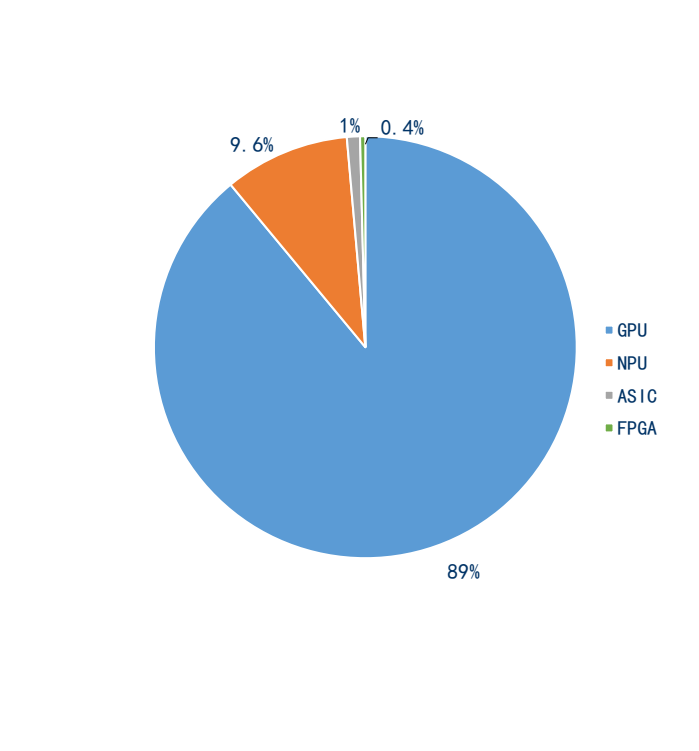
资料来源：IDC，五矿证券研究所

图表 32：全球人工智能服务器 FPGA 芯片搭载率



资料来源：IDC，五矿证券研究所

图表 33：2022 年中国人工智能芯片市场规模占比

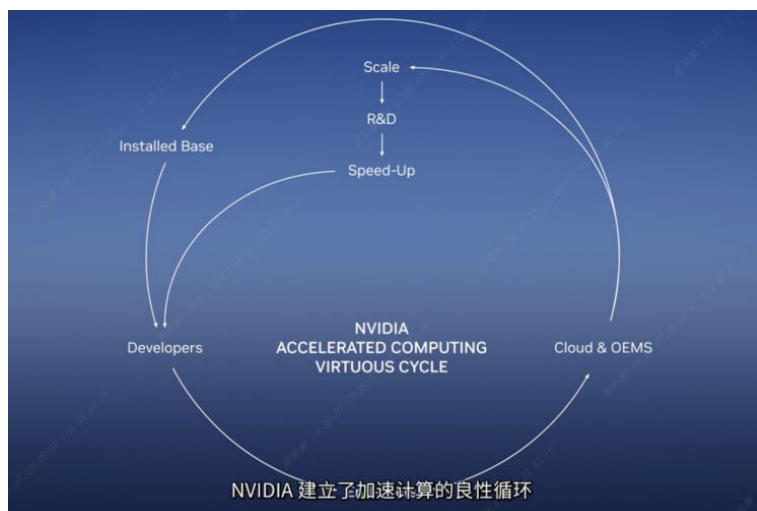


资料来源：IDC，五矿证券研究所

2.2、英伟达 AI 时代卖水人，打造发展引擎

AI 的 iPhone 时刻。2023 年 3 月 21 日，英伟达举行了 GTC 2023（GPU Technology Conoference）主题演讲，CEO 黄仁勋表示“我们正处于 AI 的 iPhone 时刻”。在会上黄仁勋介绍了英伟达在 AI 领域的最新进展，包括全新的量子计算系统、云服务平台、视觉图像系统等，并发布了四款 AI 推理芯片、三个大模型云服务、超级计算机，以及针对场景优化的应用 100 个、更新功能的工业元宇宙 Omniverse。计算技术发展日新月异，曲速引擎是加速计算，动力来源是 AI，随着计算量的激增，生成式 AI 使得企业需重新思考其商业模式与产品。目前英伟达生态系统涵盖 400 万开发者、4 万家公司、1.4 万家初创公司。

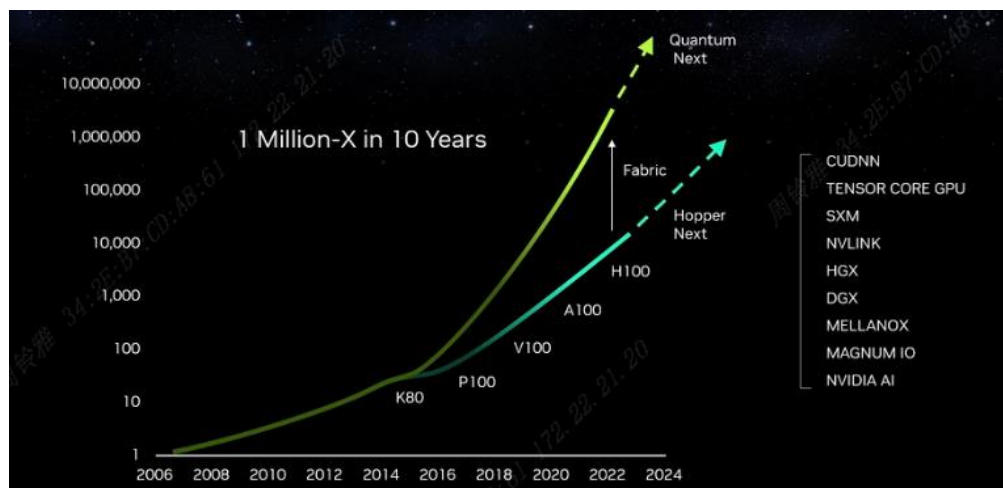
图表 34：英伟达建立了加速计算的良性循环



资料来源：英伟达官网，《英伟达 GTC 2023 主题演讲》，五矿证券研究所

加速计算和人工智能将重塑计算机行业，新计算时代临界点正到来。2023 年 5 月 29 日，英伟达 CEO 黄仁勋在 Computex 2023 上发表主题演讲，表示“加速计算已经带来了全新的计算时代”，在 AI 发展中 GPU 担任的运算器角色变得愈发重要，加速计算有望成为关键引擎。

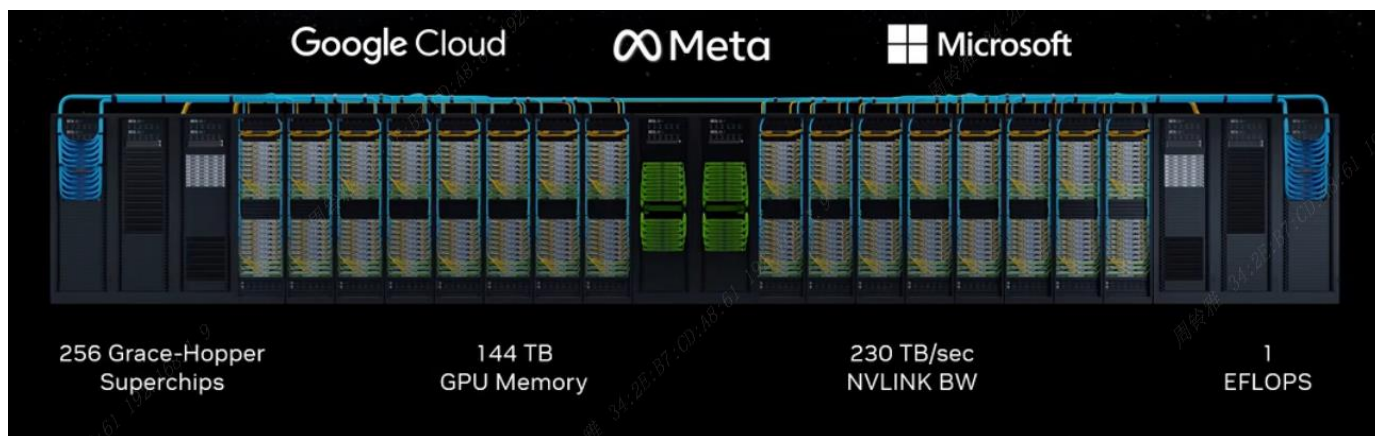
图表 35：加速计算和人工智能将重塑计算机行业



资料来源：英伟达官网，五矿证券研究所

英伟达推出超级计算机 **DGX GH200**，**E 级 AI 超算**。2023 年 5 月，在 COMPUTE X 大会上英伟达推出超级计算机 DGX GH200, DGX GH200 通过 NVLink 互连技术及 NVLink Switch System，将 256 个 Grace Hopper 超级晶片连接成一个超大型、1-Exaflop、144TB GPU 的巨型模型，用于驱动生成式人工智能、推荐系统和数据分析，为巨型人工智能模型提供线性可扩展性，谷歌云、Meta 和微软将会是首批使用 DGX GH200 的公司。同时 NVIDIA 的 Grace Hopper 超级晶片已开始量产。

图表 36：英伟达推出超级计算机 DGX GH200



资料来源：英伟达，五矿证券研究所

英伟达推出全新 **Spectrum-X 加速以太网平台**。2023 年 5 月 29 日，NVIDIA 推出 NVIDIA Spectrum-X 网络平台，该平台是一个加速网络平台，用以提高基于以太网 AI 云的性能与效率。Spectrum-X 将 NVIDIA Spectrum-4 以太网交换机与 NVIDIA BlueField-3 DPU 紧密结合，取得 1.7 倍整体 AI 性能和能效提升，同时可在多租户环境中提供一致、可预测的性能，Spectrum-X 还提供 NVIDIA 加速软件和软件开发套件（SDK），使开发人员能够构建软件定义的云原生 AI 应用。

图表 37：英伟达 Spectrum-X 以太网平台



资料来源：英伟达官网，五矿证券研究所

英伟达发布多款加速库助力加速计算。加速库是加速计算的核心，通过加速库连接应用，再连接至各行各业，形成网络中的网络。经过 30 年的开发，已有数千款应用被英伟达的库加速，涉及科学和工业多个领域。目前 NVIDIA GPU 兼容 CUDA，为开发者提供了庞大的安装基础和广泛的覆盖范围，吸引了众多终端用户。同时英伟达建立了加速计算的良性循环，涵盖了 300 个加速库和 400 个 AI 模型。

英伟达在 2023 年 GTC 大会重点介绍了几款加速库：

- **cuQuantum 量子加速库：**用于量子电路仿真的加速库。cuQuantum 由英伟达与 Quantum Machines 合作开发，将 NVIDIA GPU 连接到量子计算机以极快的速度进行纠错，对量子计算机进行加速，可供研究人员推进量子编程模型，系统架构和算法。
- **cuOpt 运筹规划加速库：**运筹学优化 API。NVIDIA cuOpt 使用进化算法和加速计算每秒分析 300 亿次动作，打破了世界纪录。cuOpt 可以优化物流，每年有 4000 亿个包裹被投递到 3770 亿个站点。德勤、Capgemini、Softserve、埃森哲和 Quantiphi 正在使用 NVIDIA cuOpt 来帮助客户优化运营。
- **CV-CUDA 视频处理加速库：**用于计算机视觉的云规模加速库。CV-CUDA 包括 30 个计算机视觉算子，可用于检测、分割和分类，VPF 是一个 Python 视频编解码加速库，腾讯使用 CV CUDA 和 VPF 每天处理 30 万个视频，Microsoft 使用 CV CUDA 和 VPF 来处理视觉搜索。
- **Parabricks 医疗加速库：**用于云端或仪器设备内的端到端基因组分析的 AI 加速库。NVIDIA Parabricks 将会在 PacBio、OxfordNanopore、Ultima、Singular、BioNano 和 Nanostring 的第四代 NVIDIA 加速基因组学设备上运行。
- **cuLitho 计算光刻加速库：**通过计算技术大幅优化芯片制造流程。cuLitho 历时 4 年，利用 GPU 技术实现计算光刻，可以使传统光刻技术提速 40 倍以上，为 2m 及更先进芯片的生产提供助力。如果在 GPU 上运行 cuLitho，只需 8 小时即可处理完一个掩模板。台积电可通过在 500 个 DGX H100 系统上使用 cuLitho 加速，将功率从 35MW 降至 5MW，从而替代用于计算光刻的 4 万台 CPU 服务器。

图表 38：英伟达 cuLitho 加速库合作客户



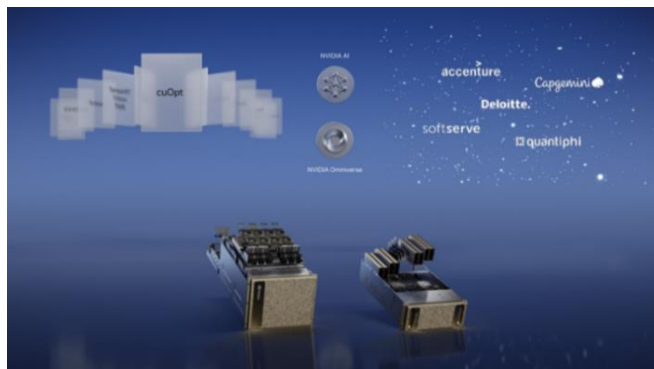
资料来源：英伟达官网，《英伟达 GTC 2023 主题演讲》，五矿证券研究所

图表 39：用于量子电路仿真的加速库 cuQuantum



资料来源：英伟达官网，《英伟达 GTC 2023 主题演讲》，五矿证券研究所

图表 40：英伟达 cuOpt 及客户



资料来源：英伟达官网，《英伟达 GTC 2023 主题演讲》，五矿证券研究所

图表 41：英伟达 CV-CUDA 平台



资料来源：英伟达官网，《英伟达 GTC 2023 主题演讲》，五矿证券研究所

图表 42：NVIDIA Parabricks 医疗加速库



资料来源：英伟达官网，《英伟达 GTC 2023 主题演讲》，五矿证券研究所

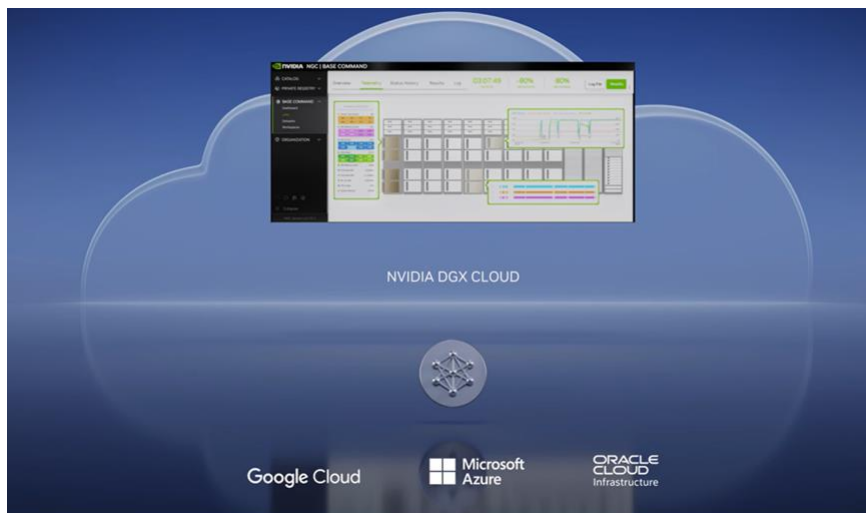
推出 DGX AI Cloud 云服务，重塑大模型格局，加速 AI 生态发展。英伟达宣布推出 NVIDIA DGX Cloud，这是一项 AI 超级计算服务，能让企业立即访问为生成式 AI 和其他突破性应用程序训练高级模型所需的基础设施和软件。企业可按月租用 DGX Cloud 集群，每月租金为 3.7 万美元，通过浏览器立即访问英伟达的 DGX AI 超级计算机和 AI 软件，为生成式 AI 和其他开创性应用训练先进模型，可快速轻松地扩展大型多节点训练工作负载的开发，而无需等待通常需求量很大的加速计算资源。

此外，英伟达还成立了 NVIDIA AI Foundations，为有专门要求的客户构建和改进相关 AI 模型训练，从而降低 AI 数据的生成时间和提升客户的工作效率。

NVIDIA AI Foundations 包括语言、视觉和生物模型制作服务。其中语言模型 NVIDIA Nemo 用于构建定制的语言文本转文本；视觉模型 PICASSO 用于视觉语言模型制作；生物学模型 BioNeMo 可帮助研究人员使用专有数据创建、微调和提供自定义模型。

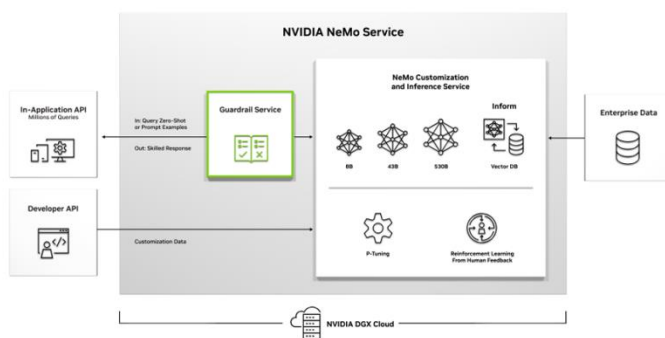
英伟达将与甲骨文、微软和谷歌等云服务提供商合作，托管 DGX Cloud 基础设施。

图表 43：英伟达推出 DGX Cloud



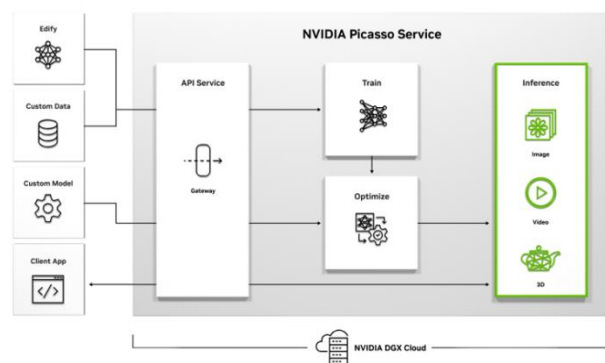
资料来源：英伟达官网，《英伟达 GTC 2023 主题演讲》，五矿证券研究所

图表 44：英伟达 Nemo 语言模型



资料来源：英伟达官网，《英伟达 GTC 2023 主题演讲》，五矿证券研究所

图表 45：英伟达 Picasso 视觉语言模型



资料来源：英伟达官网，《英伟达 GTC 2023 主题演讲》，五矿证券研究所

推出 GPU 推理平台，为 AIGC 设计专用算力。英伟达推出全新 GPU 推理平台，4 种配置分别用于加速 AI 视频、图像生成、大型语言模型部署和推荐系统，为不同种类的 AIGC 产业提供了对应的硬件产品。

- **L4:** 针对 AI 视频设计的通用 GPU，可提供比 CPU 高 120 倍的 AI 视频性能，能效提高 99%；
- **L40:** 用于图像生成，针对图形和 AI 支持的 2D、视频和 3D 图像生成进行了优化，推理性能是英伟达最受欢迎的云推理 GPU T4 的 10 倍；
- **H100NVL:** 针对 ChatGPT 等大型语言模型的大规模部署，配备双 GPU NVLink，将两张拥有 94 GB HBM3 显存的 PCIe H100 GPU 拼接在一起，可处理拥有 1750 亿参数的 GPT-3 大模型；
- **Grace Hooper:** 适用于推荐系统和大型语言模型的 AI 数据库，图推荐模型、向量数据库和图神经网络的理想选择，通过 900GB/s 的高速一致性芯片到芯片接口连接英伟达 Grace CPU 和 Hopper GPU。

图表 46：NVIDIA 全新推理平台



资料来源：英伟达官网，《英伟达 GTC 2023 主题演讲》，五矿证券研究所

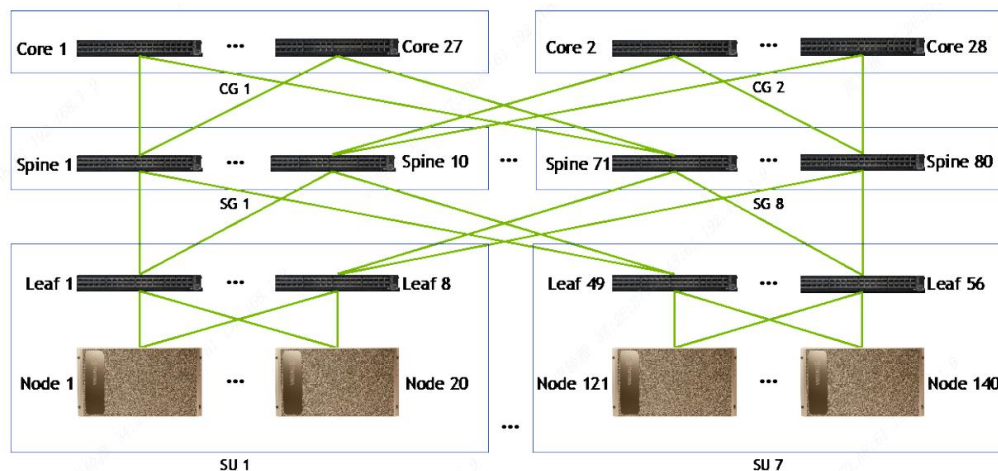
图表 47：NVIDIA H100NVL 性能情况

	H100 NVL		
	H100 PCIe	H100 NVL	
FP16 Tensor Core	1,513 TFLOPS*	3,958 TFLOPS*	2.6X
FP8 Tensor Core	3,026 TFLOPS*	7,916 TFLOPS*	2.6X
GPU Memory	80GB HBM2e	188GB HBM3	2.4X
GPU Memory Bandwidth	2TB/s	7.8TB/s	3.8X
Interconnect	NVLink 600GB/s PCIe Gen5 128GB/s	NVLink Bridge 600GB/s PCIe Gen5 128GB/s	

资料来源：英伟达官网，《英伟达 GTC 2023 主题演讲》，五矿证券研究所

胖树网络架构成为人工智能数据中心的重要需求之一。NVIDIA DGX SuperPOD 是 NVIDIA 推出的一个分布式集群的参考架构，最新一代是基于 NVIDIA DGXA100 和 NVIDIA HDR 200Gb/s ConnectX-6 HCAs 组建，旨在帮助 AI 研究人员快速搭建一套强大、灵活、高效的系统。DGX SuperPOD 采用模块化的设计，支持不同规模大小的设计，一个标准的 SuperPOD 由 140 台 DGX A100 和三层 Infiniband 交换机通过胖树结构全互联起来。每台 DGXA100 配有 8 个 200Gb/s 的高速计算网，并配有 2 个 200Gb/s 的高速存储网，采用计算和存储网络分离的方案。人工智能数据中心内部数据流量较大，高速率、低延时的要求推动胖树架构在数据中心网络中部署。

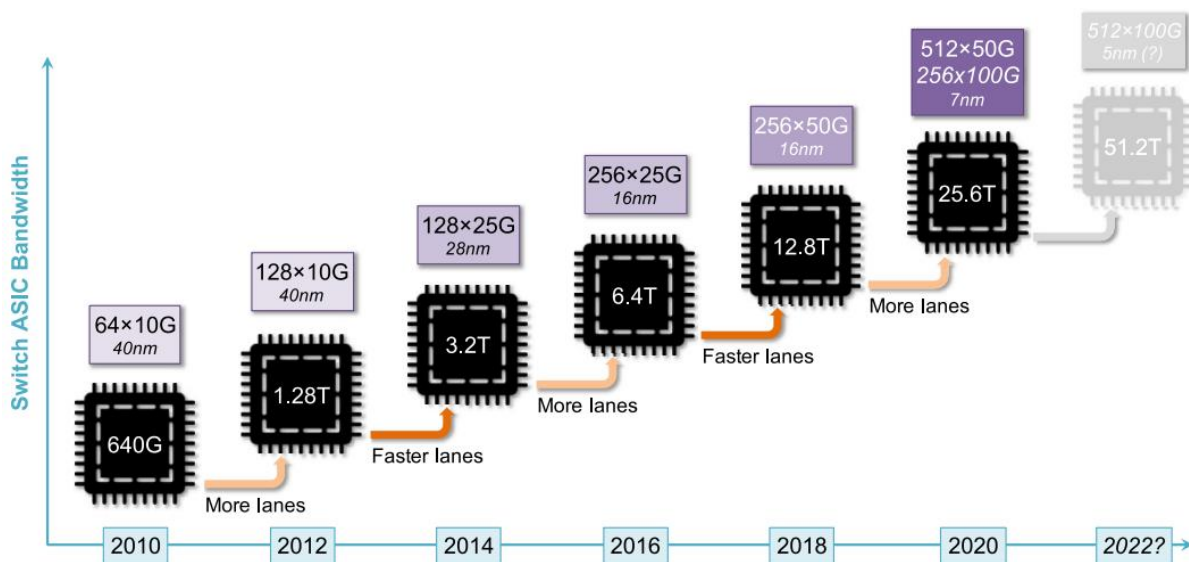
图表 48：NVIDIA SuperPOD 胖树拓扑网络结构



资料来源：英伟达，五矿证券研究所

光连接向高速率、大密度方向发展。大语言模型带来超大算力需求拉动通信基础设施建设与扩容，光模块作为数据传输的基础部件，是实现光电转换的核心部件，网络带宽需求增加，推动以太网交换机和光学器件进步，交换机之间连接需要高端光模块来完成，交换芯片的技术演进推动光模块更新换代，交换芯片向大容量、高速率、低功耗方向发展，光模块随之迭代升级。

图表 49：交换芯片技术演进路径



资料来源：《Co-packaged datacenter optics Opportunities and challenges》，五矿证券研究所

3、中文模型奋起直追，办公生产力率先变革

3.1、中文模型奋起直追，更强中文理解

大模型可大致分为基础大模型、行业大模型、场景模型三类。

基础大模型：可在多个领域和任务上通用的大模型。利用大算力、使用海量数据与大参数的深度学习算法，在大规模无标注数据上进行训练，以寻找特征并发现规律，进而形成强大泛化能力，可在不进行微调或少量微调的情况下完成多场景任务。

行业大模型：针对特定行业或领域的大模型。通常使用行业相关的数据进行预训练或微调，以提高在该领域的性能和准确度。

场景模型：针对特定任务或场景的大模型。通常使用任务相关的数据进行预训练或微调，以提高在该任务上的性能和效果。

图表 50：大模型的分类



资料来源：华为开发者大会 2023，五矿证券研究所

百度发布“文心一言”大模型，更强中文理解与跨模态能力。2023 年 3 月 16 日，百度召开发布会发布大语言模型、生成式 AI 产品“文心一言”，百度创始人、董事长兼首席执行官李彦宏展示了文心一言在文学创作、商业文案创作、数理推算、中文理解、多模态生成五个使用场景中的综合能力。百度智能云面向企业客户开放文心一言 API 接口调用服务。

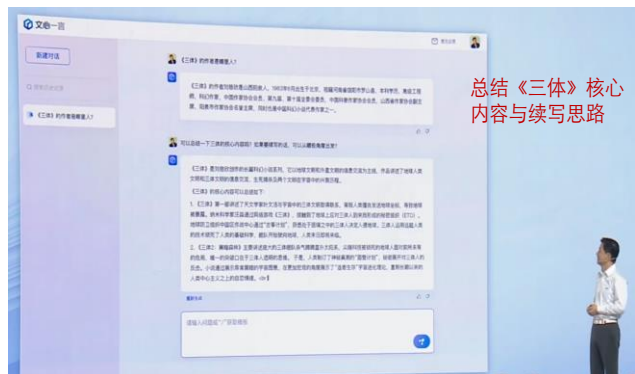
2023 年 3 月 27 日，百度推出企业级大模型服务平台“文心千帆”，基于文心一言底层模型（Ernie Bot）数据管理、自动化模型定制微调以及推理服务云端部署一站式大模型定制服务，并提供可快速使用的文心一言企业级推理云服务，助力各行业的生成式 AI 应用需求落地。

图表 51：百度发布“文心一言”



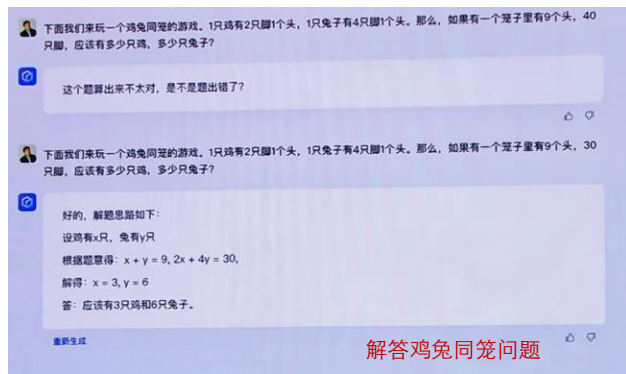
资料来源：百度官网，五矿证券研究所

图表 52: “文心一言” 文学创作能力



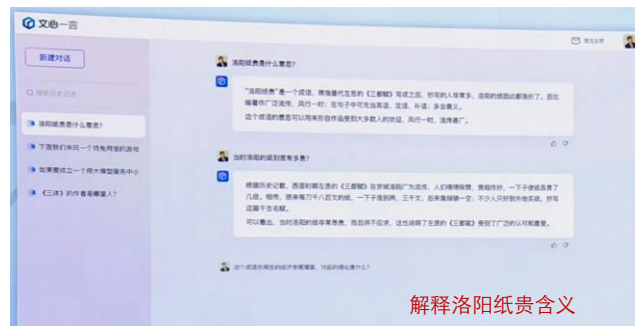
资料来源: 百度官网, 五矿证券研究所

图表 53: “文心一言” 数理逻辑推理能力



资料来源: 百度官网, 五矿证券研究所

图表 54: “文心一言” 中文理解能力



资料来源: 百度官网, 五矿证券研究所

图表 55: “文心一言” 多模态生成能力



资料来源: 百度官网, 五矿证券研究所

文心一言：知识增强大语言模型，基于飞桨深度学习和文心知识增强大模型。文心一言大模型持续从海量数据和大规模知识中融合学习具备知识增强、检索增强和对话增强的技术特色。百度同时具备人工智能四层能力，在应用层、模型层、框架层、芯片层进行全栈布局。从昆仑芯片、飞桨深度学习平台、文心大模型到应用，百度在技术栈的各层都有领先业界的关键自研技术，实现了层与层反馈，端到端优化，大幅提升效率。

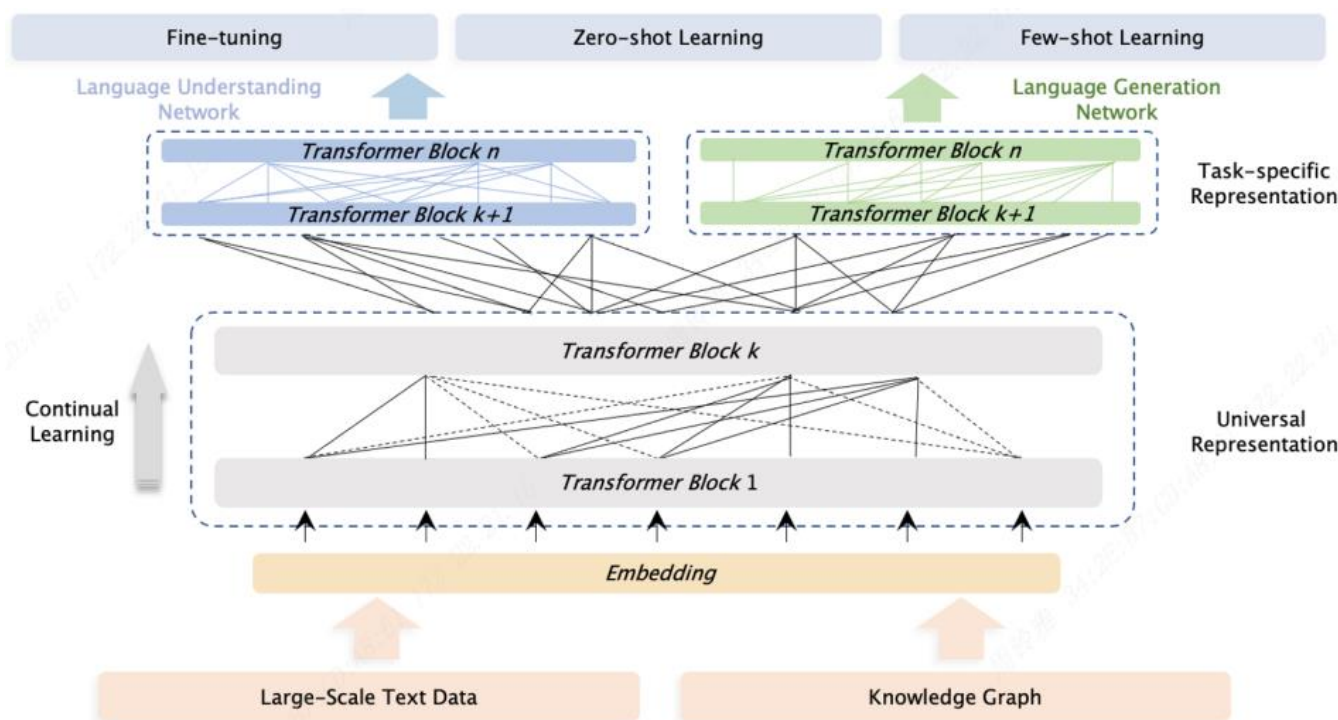
图表 56: 百度人工智能全栈布局



资料来源: 百度官网, 五矿证券研究所

文心大模型：基于 BERT 结构，多范式的统一预训练模式，支持长文本语言模型建模。2019 年 3 月，百度发布并开源了中国首个知识增强的语言模型文心 ERNIE 1.0，百度围绕各场景，在 NLP、CV、跨模态、生物计算等领域形成完善的大模型生态，已历经四代，2022 年发布的 ERNE3.0 TITAN 参数量达到 2600 亿，相对 GPT-3 的参数量提升 50%。

图表 57：ERNIE 3.0 文心大模型架构示意图



资料来源：《ERNIE 3.0: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation》，五矿证券研究所

超大规模预训练数据集。在 ERNIE 3.0 中，百度提出了一个多范式统一的大规模预训练框架，基于该框架，ERNIE 3.0 融合了自回归网络和自编码网络，同时由于大规模知识图谱类数据的引入，使得模型能够在理解任务、生成任务、零样本学习任务 and 常识推理任务上均获取优秀的表现。ERNIE 3.0 在 14 种类型共计 45 个自然语言理解数据集上取得了 SOTA 的结果，在 9 个自然语言生成任务上取得了 SOTA 的效果，在 18 个数据集上刷新了 zero-shot 的 SOTA 效果。

图表 58：ERNIE 3.0 预训练数据集



资料来源：百度官网，五矿证券研究所

创新型算法，更好的中文理解能力，全面的模型尺寸，覆盖不同场景需求。百度创新性地将大数据预训练与多源丰富知识相结合，通过持续学习技术，不断吸收海量文本数据中词汇、结构、语义等方面的新知识，实现模型效果不断进化。

文心大模型引入知识图谱，通过知识内化（语义学习和数据构造）和知识外用（知识推理和提示构建），将数据与知识融合，使得学习效率更高、可解释性更好，突破了多源异构数据难以统一表示和学习瓶颈。针对不同的场景需求，百度发布不同的模型尺寸，拓宽了人工智能技术落地场景的覆盖广度。

图表 59：全面的模型尺寸，覆盖不同场景需求

模型	参数大小
ERNIE3.0-10B	100 亿
ERNIE3.0-1.5B	15 亿
ERNIE3.0-Large	3 亿
ERNIE3.0-Base	1 亿
ERNIE3.0-Tiny	3000 万

资料来源：百度官网，五矿证券研究所

大模型体系化，预置丰富模型。不同场景对模型的需求会有一定的差异化，所以在通用模型的基础上需增加不同特质，对此百度增加了两类模型：

- 1、任务大模型，主要面向特定任务，比如语音搜索、图文搜索、图象理解等；
- 2、行业大模型，针对特定行业进行知识增强后的大模型。

图表 60：百度面向特定任务的大模型

模型	场景
ERNIE-Gen	文本生成任务对比同尺寸通用模型 BLEU4 提升 2~5%
ERNIE-IE	文本抽取任务，适合零样本少样本情况，单标签 5 条样本
ERNIE-Rank	检索排序任务提升 2%
ERNIE-Senta	情感分析任务，对比同尺寸通用模型平均提升 0.6%
ERNIE-Sim	文本匹配任务对比同尺寸通用模型提升 10%
ERNIE-Doc	长文本分类匹配抽取任务等
ERNIE-M	多语言分类匹配抽取任务等

资料来源：百度官网，五矿证券研究所

图表 61：百度面向特定领域的大模型

模型	场景
ERNIE-Finance	金融领域（实体抽取任务提升 1.4%，分类提升 0.4%）
ERNIE-Health	医疗领域（实体识别任务提升 0.8%，匹配任务提升 1.7%）
ERNIE-Law	法律领域（罪名预测提升 0.33%，法条推荐提升 1.52%）

资料来源：百度官网，五矿证券研究所

飞桨（PaddlePaddle）：源于产业实践的开源深度学习平台。百度于 2013 年成立深度学习实验室,并推出自主研发的深度学习平台,即 PaddlePaddle 前身;2016 年 8 月,PaddlePaddle 正式开源,对标谷歌 TensorFlow,是中国首个开源开放的深度学习框架。目前 PaddlePaddle 集深度学习核心框架、基础模型库、端到端开发套件、工具组件和服务平台于一体,是全面开源开放、技术领先、功能完备的产业级深度学习平台。目前飞桨已广泛应用于工业、农业、服务业等。

图表 62：飞桨重点产品示意图

服务平台	EasyDL 零门槛 AI 开发平台	AI Studio 学习与实训社区	EasyEdge 端计算模型生成平台
工具组件	PaddleHub 预训练模型应用工具	PaddleX 全流程开发工具	PaddleFL 联邦学习
开发套件	PaddleDetection 目标检测	PaddleHelix 螺旋桨生物计算平台	PaddleOCR 文本识别
基础模型库	PaddleNLP 自然语言处理模型库	PaddleCV 视觉模型库	Wenxin Big Models 文心大模型
核心框架	Paddle 飞桨训练框架	Paddle Lite 轻量化推理引擎	PaddleSlim 模型压缩工具

资料来源：百度官网，五矿证券研究所

文心一格：AI 艺术与创意辅助平台。文心一格是百度推出的 AI 作画产品。依托飞桨、文心大模型的技术创新,只需输入文字描述,即能生成各种风格的精美画作。可为画师、设计师等视觉内容创作者启发灵感,辅助艺术创作,也能为媒体、作者等文字内容创作者提供高质量、高效率的配图。

图表 63：百度文心一格 AI 艺术与创意辅助平台



资料来源：百度官网，五矿证券研究所

文心百中：产业级搜索系统。文心百中是百度推出的由大模型驱动的产业级搜索系统。依托百度文心大模型，以极简框架，代替传统搜索引擎复杂的构建及维护成本，可低成本接入各类企业和开发者应用，并凭借数据驱动优化模式可实现行业优化效率及应用效果。

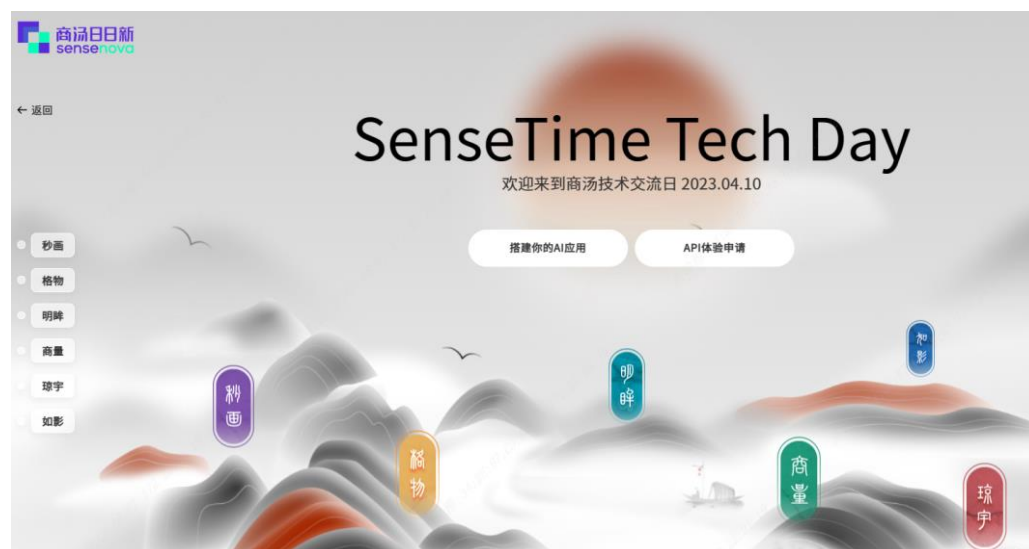
图表 64：百度文心百中产业级搜索系统



资料来源：百度官网，五矿证券研究所

商汤科技推出“日日新大模型”体系。2023年4月10日，在商汤技术交流日活动上，商汤科技董事长兼首席执行官徐立宣布推出大模型体系“商汤日日新大模型”，包括自然语言生成、文生图、感知模型标注、以及模型研发功能。

图表 65：商汤科技推出日日新大模型体系

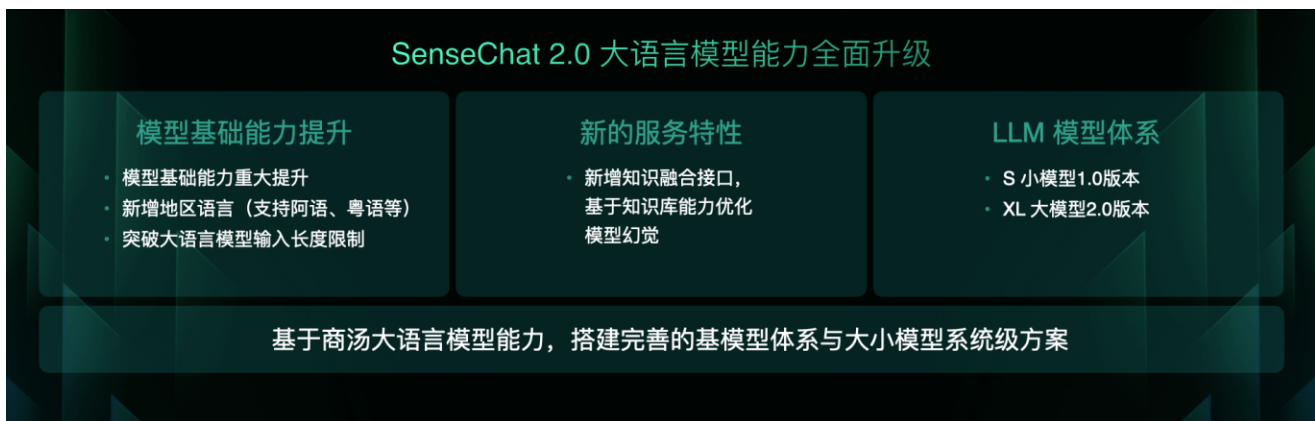


资料来源：商汤科技官网，五矿证券研究所

商汤 SenseChat 2.0 大语言模型能力全面升级。2023 年 7 月 7 日，商汤科技董事长兼 CEO 徐立第六届世界人工智能大会（WAIC 2023）推出商汤商量 SenseChat 2.0 版本，实现大语言模型全面升级，突破了长度限制，并推出不同参数量级的模型版本，可完美适配移动端、云端等不同终端及场景的应用需求，降低部署成本。

与此同时，商汤秒画 SenseMirage 3.0 文生图大模型亦全面升级，模型参数从 2023 年 4 月的 10 亿提升至 7 月的 70 亿，可实现专业摄影级的图片细节刻画。

图表 66：商汤 SenseChat 2.0 大语言模型能力全面升级



资料来源：商汤科技，五矿证券研究所

图表 67：商汤 SenseMirage 3.0 文生图大模型全面升级



资料来源：商汤科技，五矿证券研究所

华为云正式发布盘古大模型 3.0，面向行业的大模型系列。2023 年 7 月 7 日，华为开发者大会 2023 在东莞召开，华为常务董事、华为云 CEO 张平安正式发布了华为云盘古大模型 3.0，首次明确定位“为行业而生”。盘古大模型 3.0 分为 L0 基础大模型、L1 行业大模型、L2 场景模型三层架构，将重塑千行百业。盘古大模型强调在细分场景的落地应用，主要解决商业环境中低成本大规模定制的问题，用 AI 赋能千行百业。

在煤矿领域，盘古矿山大模型已经在全国 8 个矿井规模使用，一个大模型可以覆盖煤矿的采、掘、机、运、通、洗选等业务流程下的 1000 多个细分场景，让更多煤矿工人能够在地面上作业，极大地减少安全事故；

在气象领域，原来预测一个台风未来 10 天的路径，需要在 3000 台服务器的高性能计算机集群上花费 5 小时进行仿真，现在基于预训练的盘古气象大模型，10 秒内就可以获得更精确的预测结果，速度比传统数值预报提速 10000 倍以上。7 月 6 日，该研究成果登上国际顶级学术期刊《自然》杂志；

在交通运输行业，盘古铁路大模型能精准识别现网运行的 67 种货车、430 多种故障，综合准确性高达 99.8%。通过盘古铁路大模型的“慧眼”，无故障图片的筛除率为 95%，实现效率提升。

图表 68：华为云正式发布盘古大模型 3.0



资料来源：华为开发者大会 2023，五矿证券研究所

盘古大模型在性能、深度、架构及数据增强方面均进行了升级。华为拥有算力、框架、平台全栈核心技术，盘古大模型实现了以鲲鹏和昇腾为基础的 AI 算力云平台，以及异构计算架构 CANN、全场景 AI 框架昇思 MindSpore、AI 开发生产线 ModelArts 的全栈自主创新。单集群 2000P Flops 算力、千卡训练 30 天长稳率达到 90% 的昇腾 AI 云服务在华为云的乌兰察布和贵安 AI 算力中心同时上线。

图表 69：华为全栈核心技术

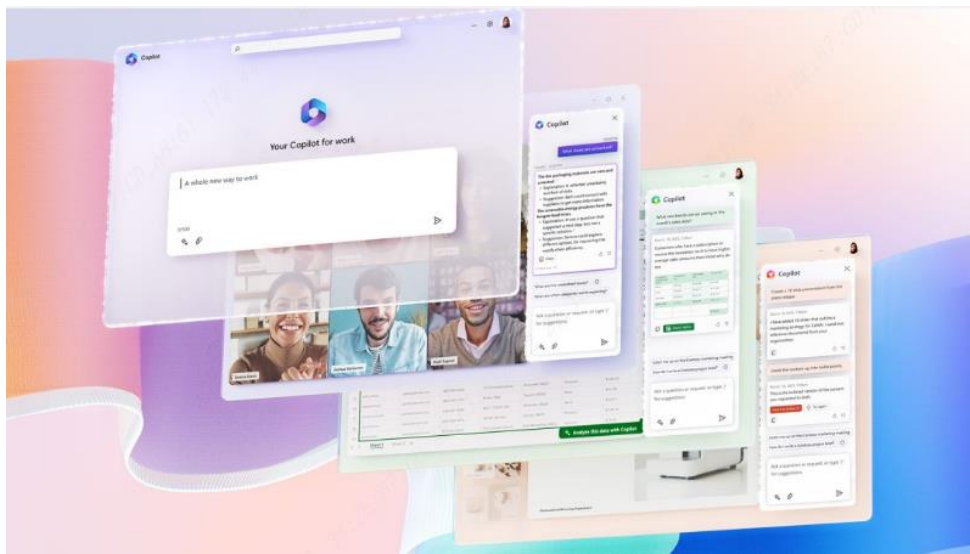


资料来源：华为开发者大会 2023，五矿证券研究所

3.2、生产力工具革新，释放创造力

微软发布 **Microsoft 365 Copilot**，提升工作效率，革新工作方式。2023 年 3 月 16 日，微软发布了集成了 GPT4 功能的 Microsoft 365 Copilot，提供了包括文字和图片内容生成，归纳，数据分析、辅助决策等一系列新功能。Copilot 一款基于 GPT-4 和 Microsoft Graph 的 AI 办公助手，为用户提供了一种全新的工作方式。Copilot 通过两种方式集成至 Microsoft 365 中，第一种是嵌入至 Microsoft 365 应用程序中，如 Word、Excel、PowerPoint、Outlook、Teams 等，第二种是邮件、联系人、在线会议等软件数据接入大语言模型，借助 Power Platform 中的 Copilot，将重复的工作流程自动化，解锁生产力。

图表 70：微软发布 Microsoft 365 Copilot



资料来源：微软官网，五矿证券研究所

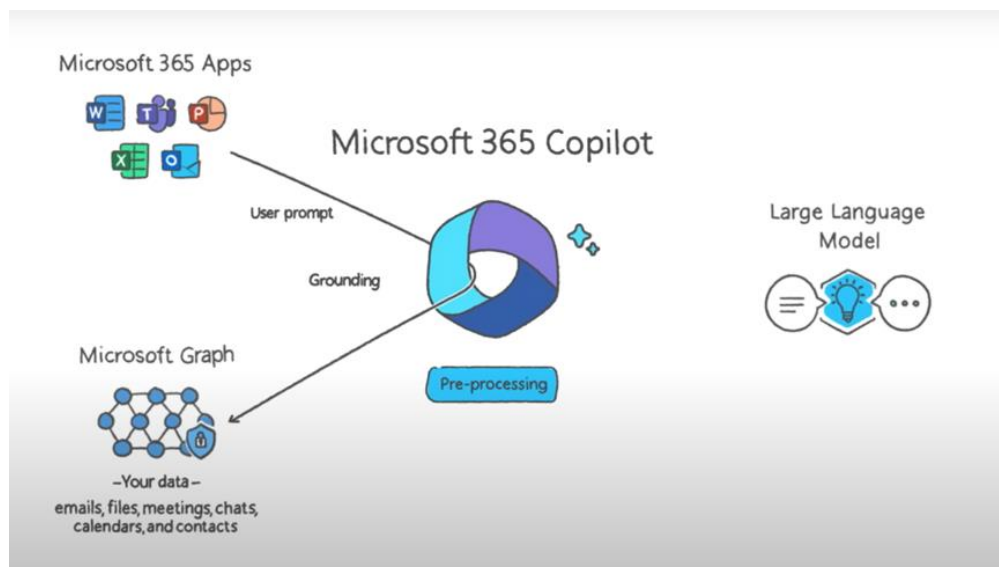
Microsoft 365 Copilot 技术原理：LLM+Microsoft Graph+Microsoft 365 Apps。 Microsoft 365 Copilot 由“Copilot 系统”提供支持，利用了三种基础技术：

- **Microsoft 365 应用：**即 Word、Excel、PowerPoint、Outlook、Teams 等；
- **Microsoft Graph：**是 Microsoft 365 中通往数据和智能的网关，它提供统一的 可编程模型，可用于访问 Microsoft 365、ows 10 和企业移动性 + 安全性中的海量数据，相当于数据传输的通道，连结大模型和 Microsoft 365 应用；
- **大型语言模型 LLM：**使用大量数据训练的深度学习模型，可以生成自然语言文本 或理解语言文本的含义。

用户输入指令后，Copilot 将该指令信息发送到 Microsoft Graph，进行上下文检索和修正提升，并将修正后的信息发送到 GPT-4，之后将回复传回 Microsoft Graph，完成安全和合规性检查后，最终传回到 Microsoft 365 应用。

Copilot 能扫描 Microsoft Graph 中存储的电邮、会议、聊天内容和日历中标注的日期和事件等所有数据，根据这些内容作出回应。这些数据可以帮助支持 Copilot 的大语言模型改进回答，让回答更具体、对用户个人更有针对性。

图表 71: Microsoft 365 Copilot 技术原理



资料来源: 微软官网, 五矿证券研究所

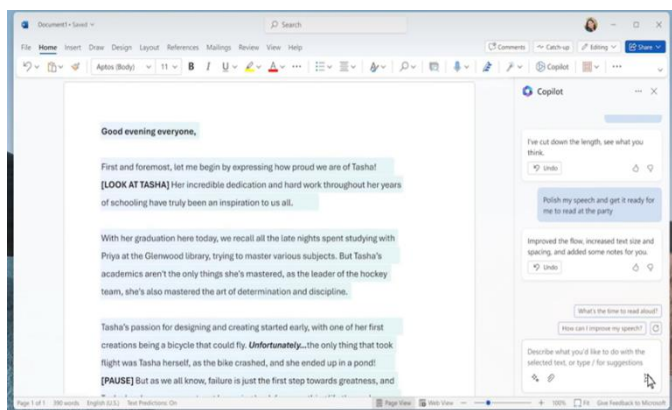
Word: 通过 Copilot 可实现文稿的快速生成和修改。

- 帮助用户写作、编辑、摘要和创作文档, 能让用户集中精力发挥创造力, 快速撰写具有说服力的提案;
- 可以根据简单的笔记和其他内部文件草拟提案, 生成初稿, 并按照用户的需求调整格式、插入图片和提取视觉素材;
- 可以生成摘要和提供建议以改进文档。

Excel: 通过 Copilot 可实现海量数据可视化, 帮助用户分析和探索数据。

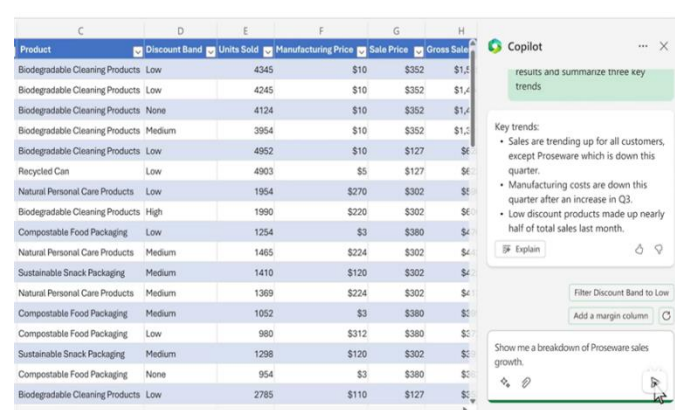
- 用户可以让 Copilot 找出关键趋势、提出跟进问题, 并将数据可视化;
- Copilot 还能创建简单的假设模型, 并将操作过程逐步分解, 以使用户进行修改。

图表 72: Copilot 版 Word 使用示例



资料来源: 微软官网, 五矿证券研究所

图表 73: Copilot 版 Excel 使用示例

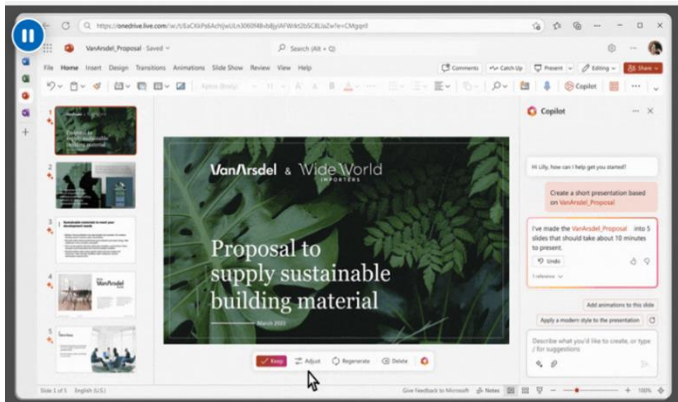


资料来源: 微软官网, 五矿证券研究所

PowerPoint: 通过 Copilot 可基于文稿生成幻灯片, 通过便捷指令调用编辑功能。

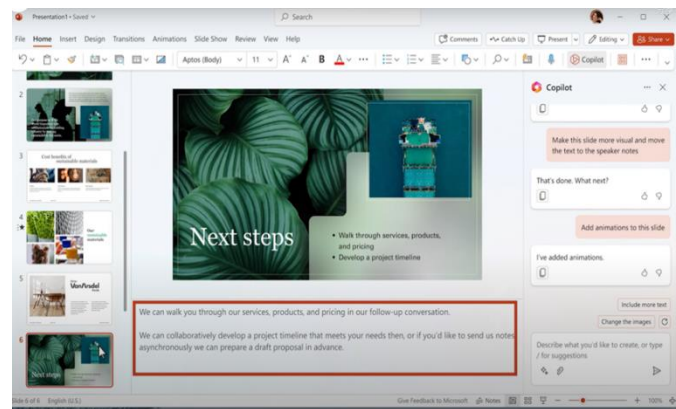
- 可将 Word 文档转换为演示文稿, 亦可将 PowerPoint 转换为 Word 文档;
- 用户只需给出简短的指示, Copilot 即能迅速完成幻灯片制作, 让用户进行审查;
- 能为文字过多幻灯片添加动画效果, 以实现专业设计效果;
- 可以为所有幻灯片生成演讲者备注, 帮助用户成为更出色的演讲者。

图表 74: Copilot 版 PPT, 将 Word 文档转换为幻灯片



资料来源: 微软官网, 五矿证券研究所

图表 75: Copilot 版 PPT, 为所有幻灯片生成演讲者备注

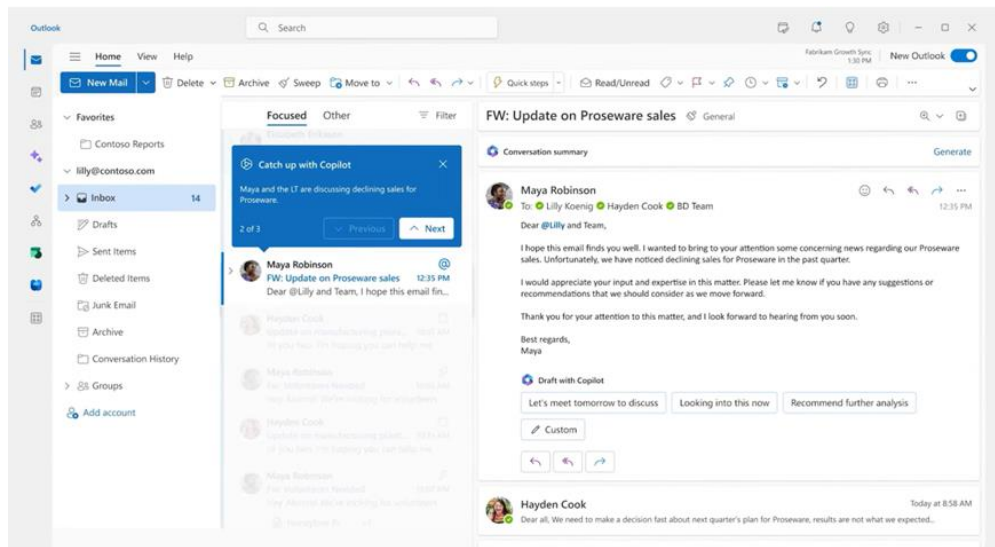


资料来源: 微软官网, 五矿证券研究所

Outlook: 通过 Copilot 辅助整理、归纳、起草电子邮件。

- 可自动管理和回复邮件, 用户可通过简单命令指引对邮件进行分类、总结、提取和回复等操作;
- 可根据收件人或主题生成合适的语气或建议, 帮助用户节省处理邮件时间。

图表 76: Copilot 版 Outlook, 自动管理和回复邮件

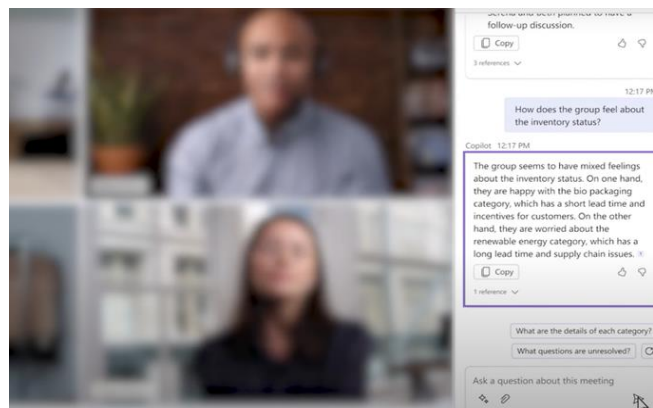


资料来源: 微软官网, 五矿证券研究所

Teams: 通过 Copilot 总结会议要点、追踪待办事项。

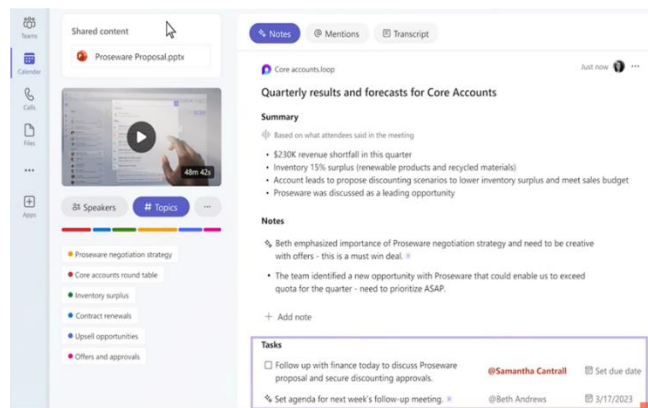
- 可帮助用户参与和组织会议，提供即将进行的会议信息、相关项目更新、组织变化等；
- 可让用户跟踪会议，从而解决时间冲突问题；
- 会议结束后，可总结用户错过的内容，包括共享的内容、概括笔记和与用户或其他被提及的待办事项，并在 Teams 中发出通知；
- 用户可向 Copilot 提出澄清问题以获取详细的回答和有用的背景信息。

图表 77: Copilot 版 Teams，会议过程中提供会议信息



资料来源：微软官网，五矿证券研究所

图表 78: Copilot 版 Teams，会议结束后总结用户错过的内容

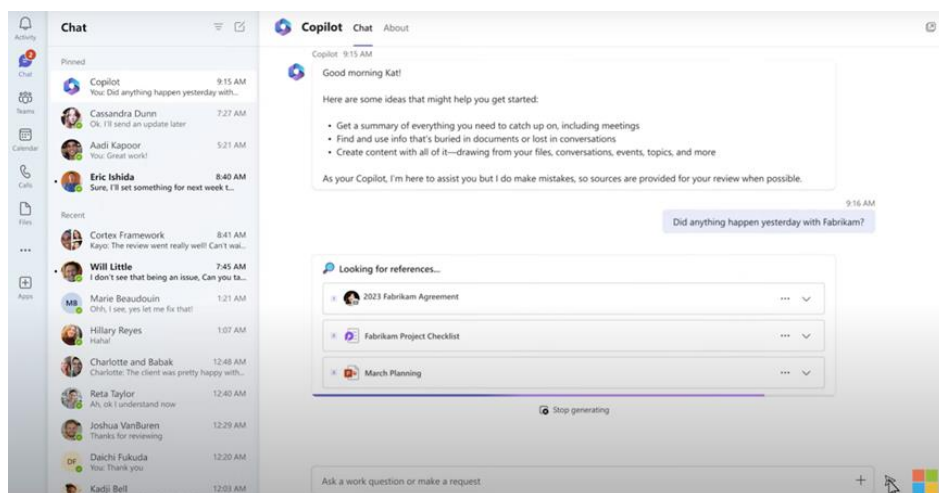


资料来源：微软官网，五矿证券研究所

Business Chat: 通过 Copilot 实现全局信息视角下的交互式回答。

- Business Chat 融合了大模型能力、Microsoft365 应用和用户在日历、邮件、聊天、文档、会议、联系人中的数据，能够基于全局信息视角给予综合性答复；
- 能够与用户进行商业相关的对话，根据用户的日常活动生成提醒和状态更新，让用户随时了解工作进展。

图表 79: Copilot 版 Business Chat 使用示例，给予综合性答复



资料来源：微软官网，五矿证券研究所

微软发布 **Microsoft Security Copilot**，进入网络安全领域。2023 年 3 月 28 日，微软宣布推出人工智能产品 Microsoft Security Copilot，将 AI 技术应用于网络安全领域，为安全专业人员提供一个有效工具。

Security Copilot 可深度集成到微软的产品和工作流程中，Security Copilot 与微软的三大安全产品 Microsoft Sentinel、Microsoft Defender、Microsoft Intune 进行了集成，提供事件安全、应用程序防护、云端到本地等多维度防护。此外开创了交互式防护体验，用户可以通过自然语言询问安全事件响应、威胁追踪和安全报告，帮助企业、个人用户缓解诸多网络安全业务痛点。

图表 80: Microsoft Security Copilot 的特点



资料来源：微软官网，五矿证券研究所

图表 81: Microsoft Security Copilot 为微软各安全产品 AI 赋能



资料来源：微软官网，五矿证券研究所

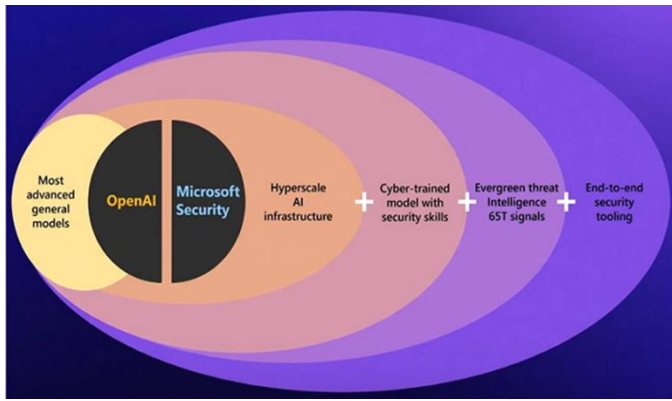
Microsoft Security Copilot 的主要功能：

- 1、简化安全流程复杂性：**用户借助 Security Copilot 可实现“分钟级”自动网络安全事件评估和响应，通过基于自然语言的调查，提供关键的分步指导，可以汇总所有安全、调查事件，简化安全流程，让用户直观地看到最重要的安全信息，并对其做出安全防护对策；
- 2、深度分析、优先级排序：**Security Copilot 发现安全问题后，会借助微软网络安全威胁库深度分析攻击行为，并预测黑客下一步的行动以及抵御方法；
- 3、回答专业的网络安全知识：**Security Copilot 通过专业的网络安全数据训练，可为用户提供专业、详细的网络安全知识，例如，介绍攻击病毒的历史，如何日常防护设备、数据资产等，为制定防护计划提供帮助；
- 4、自我学习能力：**Security Copilot 具备自我学习能力，用户的每一次反馈都将加速 Security Copilot 对安全事件响应、网络攻击的分析效率和准确率。

Microsoft Security Copilot 技术原理：GPT-4+Microsoft 安全专用模型。

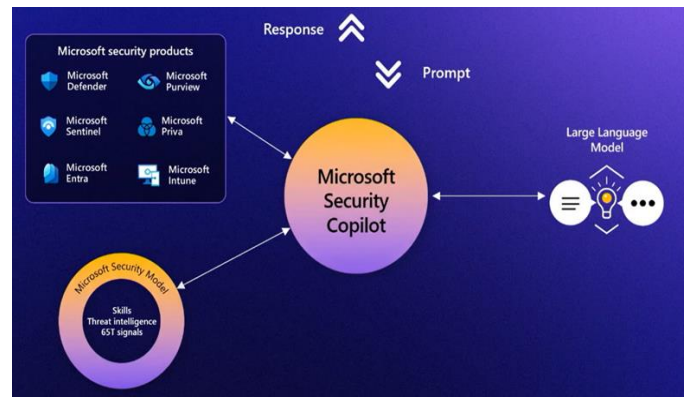
Security Copilot 是将 OpenAI 大语言模型 GPT-4 与 Microsoft 的安全特定模型相结合所开发出的生成式 AI 安全产品，在 Microsoft Azure 的超大规模基础架构上运行。Security Copilot 结合了微软拥有的 65 万亿个网络安全威胁的安全模型库，通过收集到的信号让专业的安全研究人员追踪威胁，极大提升了安全范围。

图表 82: Microsoft Security Copilot 基于 GPT-4 和微软安全专用模型



资料来源: 微软官网, 五矿证券研究所

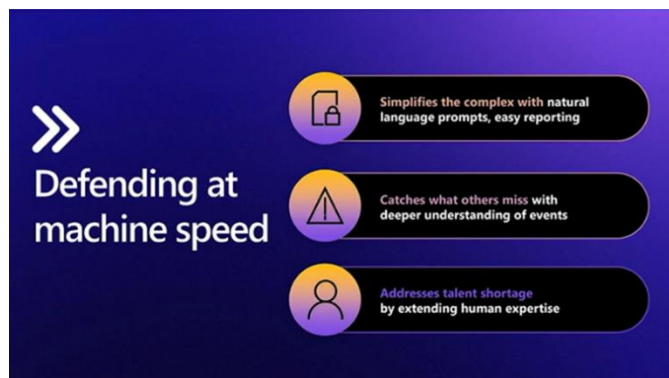
图表 83: Microsoft Security Copilot 集成了 GPT-4 和微软的安全能力



资料来源: 微软官网, 五矿证券研究所

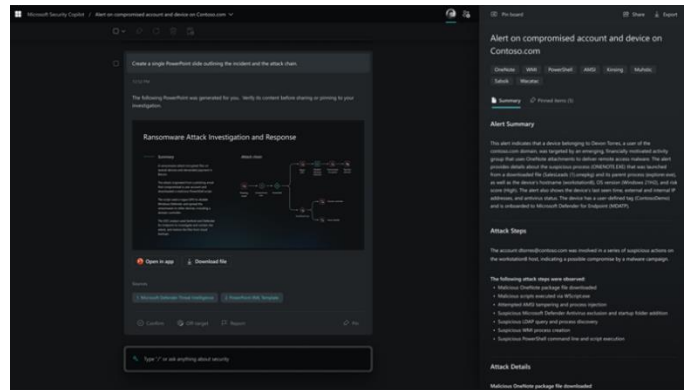
接受自然语言输入, 协助安全研究人员提升能力。Security Copilot 接受自然语言输入, 大幅降低使用门槛, 安全研究人员可以通过输入文本来要求它提供特定漏洞的摘要, 或者通过输入文件、URL 或代码片段来完成分析。Security copilot 可以实时分析进攻威胁、处理大量信息、生成解决方案, 并以可视化的形式输出。以勒索软件为例, 以往仅了解勒索软件事件就要花费大量的时间, 而解决相应事件需花费更多时间。Security Copilot 可以在几秒内获取重要信息、分析进攻数据, 并展示可视化图像、安全事件过程及相应的解决方案。大大提升了安全威胁分析及处理的效率。

图表 84: 赋能安全团队以机器的速度来解决问题



资料来源: 微软官网, 五矿证券研究所

图表 85: Microsoft Security Copilot 使用案例



资料来源: 微软官网, 五矿证券研究所

4、相关公司

- 1) **AI 算力方向**：服务器方向**浪潮信息**、**紫光股份**等；光模块方向**中际旭创**、**天孚通信**、**新易盛**等；
- 2) **AI 应用方向**：OA 应用**金山办公**、**万兴科技**等。

图表 86：相关公司业务情况

行业	股票代码	公司名称	相关业务
AI 算力方向	服务器	000977.SZ 浪潮信息	全球领先 IT 基础设施产品、方案和服务提供商，是中国最早的 IT 品牌之一。业务覆盖计算、存储、网络三大关键领域，在算力方面，公司在人工智能计算领域完成了在硬件、平台、管理和框架等四个层面的业务布局。 浪潮信息人工智能服务器的市占率位居全球第一，在中国市场份额连续五年保持 50%以上。2021 年 9 月，浪潮发布参数量为 2457 亿的中文语言大模型“源 1.0”，2022 年，浪潮深入推进“源 1.0”开源开放计划，持续加强开发者社区建设及行业应用落地。
		000938.SZ 紫光股份	紫光股份是紫光集团“从芯到云”战略中云战略主要承载者。公司控股子公司新华三推出多款 AI 服务器产品，持续保持技术领先。新华三 AI 服务器在训练任务中表现优异，在国际权威 AI 基准评测组织 MLPerf 公布了最新 AI 训练（Training V3.0）测试榜单斩获 7 项冠军。 新华三智能边缘服务器 2022 年市场份额跃居中国第二，同比增长 18%。
	光模块	300308.SZ 中际旭创	全球领先的光模块供应商，2021 年中际旭创和 II-VI 并列全球第一。 集高端光通信收发模块的研发、设计、封装、测试和销售于一体，为云数据中心客户提供 100G、200G、400G 和 800G 等高速光模块，为电信设备商客户提供 5G 前传、中传和回传光模块以及应用于骨干网和核心网传输光模块等高端整体解决方案。
		300394.SZ 天孚通信	全球光器件核心部件领域的领先企业，光器件一站式解决方案提供商。 主要产品包括光组件、光器件等，在精密陶瓷、工程塑料、复合金属、光学玻璃等基础材料领域积累沉淀了多项全球领先的工艺技术。产品核心技术能力矩阵布局完整，高速光引擎、800G 光器件、车载激光雷达用光器件、保偏光器件等研发项目顺利进展。
		300502.SZ 新易盛	全系列光通信应用光模块厂商。是国内少数批量交付运用于数据中心市场的 100G、200G、400G 高速光模块，已成功推出 800G 光模块产品系列组合、基于硅光解决方案的 400G 光模块产品及 400GZR/ZR+相干光模块。
	OA 应用	688111.SH 金山办公	国产办公软件龙头。公司主要产品包括 WPS Office 办公软件、金山文档等办公能力产品矩阵以及金山数字办公平台解决方案。 金山办公推出国内协同办公赛道首个类 ChatGPT 式应用，2023 年 7 月 6 日金山办公正式推出基于大语言模型的智能办公助手 WPS AI，WPS AI 官网同步上线，WPS AI 已接入金山办公多个办公产品组件。WPS AI 具备内容生成、公式生成等功能，用户输入相关主题即可一键生成文字内容、制作 PPT，并可分析表格数据生成报告、自动生成复杂公式等。
		300624.SZ 万兴科技	国内领先的数字创意软件产品和服务的提供商。在视频创意、绘图创意、文档创意、实用工具等多个细分领域拥有十余年的产品研发和运营经验，业务范围遍及全球 200 多个国家和地区，全球累计用户逾 15 亿。 推出万兴喵影、万兴优转、万兴录演等视频创意软件，Wondershare Filmstock、万兴喵库视频创意素材平台，万兴爱画、万兴播爆等 AIGC 应用。

资料来源：，公司公告，公司官网，五矿证券研究所

请仔细

更多免费行业报告

并购家

www.ipoipo.cn



5、风险提示

- 1、AI 技术迭代不及预期风险,当前模型仍需持续迭代优化,若相关公司核心技术难以突破,将影响产品研发迭代;
- 2、若商业化落地不及预期,政策推进力度不及预期,经济发展不及预期,中美贸易摩擦加剧,产业链各方合作不及预期,将影响产业链整体进展。

分析师声明

作者在中国证券业协会登记为证券投资咨询(分析师)，以勤勉的职业态度，独立、客观地出具本报告。作者保证：(i) 本报告所采用的数据均来自合规渠道；(ii) 本报告分析逻辑基于作者的职业理解，并清晰准确地反映了作者的研究观点；(iii) 本报告结论不受任何第三方的授意或影响；(iv) 不存在任何利益冲突；(v) 英文版翻译若与中文版有所歧义，以中文版报告为准；特此声明。

投资评级说明

投资建议的评级标准		评级	说明
报告中投资建议所涉及的评级分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后 6 到 12 个月内的相对市场表现，也即以报告发布日后的 6 到 12 个月内的公司股价(或行业指数)相对同期相关证券市场代表性指数的涨跌幅作为基准。其中：A 股市场以沪深 300 指数为基准；香港市场以恒生指数为基准；美国市场以纳斯达克综合指数或标普 500 指数为基准。	股票评级	买入	预期个股相对同期相关证券市场代表性指数的回报在 20%及以上；
		增持	预期个股相对同期相关证券市场代表性指数的回报介于 5%~20%之间；
		持有	预期个股相对同期相关证券市场代表性指数的回报介于-10%~5%之间；
		卖出	预期个股相对同期相关证券市场代表性指数的回报在-10%及以下；
		无评级	预期对于个股未来 6 个月市场表现与基准指数相比无明确观点。
	行业评级	看好	预期行业整体回报高于基准指数整体水平 10%以上；
		中性	预期行业整体回报介于基准指数整体水平-10%~10%之间；
		看淡	预期行业整体回报低于基准指数整体水平-10%以下。

一般声明

五矿证券有限公司（以下简称“本公司”）具有中国证监会批复的证券投资咨询业务资格。本公司不会因接收人收到本报告即视其为客户，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。本报告的版权仅为本公司所有，未经本公司书面许可，任何机构和个人不得以任何形式对本研究报告的任何部分以任何方式制作任何形式的翻版、复制或再次分发给任何其他人。如引用须联络五矿证券研究所获得许可后，再注明出处为五矿证券研究所，且不得对本报告进行有悖原意的删节和修改。在刊载或者转发本证券研究报告或者摘要的同时，也应注明本报告的发布人和发布日期及提示使用证券研究报告的风险。若未经授权刊载或者转发本报告的，本公司将保留向其追究法律责任的权利。若本公司以外的其他机构（以下简称“该机构”）发送本报告，则由该机构独自为此发送行为负责。

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，本报告所指的证券或投资标的的价格、价值及投资收入或将产生波动；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。

本报告的作者是基于独立、客观、公正和审慎的原则制作本研究报告。本报告的信息均来源于公开资料，本公司对这些信息的准确性和完整性不作任何保证，也不保证所包含信息和建议不发生任何变更。本公司已力求报告内容的客观、公正，但文中的观点、结论和建议仅供参考，不包含作者对证券价格涨跌或市场走势的确定性判断。在任何情况下，报告中的信息或意见不构成对任何人的投资建议，投资者据此做出的任何投资决策与本公司和作者无关。在任何情况下，本公司、本公司员工或者关联机构不承诺投资者一定获利，不与投资者分享投资收益，也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。本公司及作者在自身所知情范围内，与本报告中所评价或推荐的证券不存在法律法规要求披露或采取限制、静默措施的利益冲突。

五矿证券版权所有。保留一切权利。

特别声明

在法律许可的情况下，五矿证券可能会持有本报告中提及公司所发行的证券并进行交易，也可能为这些公司提供或争取提供投资银行、财务顾问和金融产品等各种金融服务。因此，投资者应当考虑到五矿证券及/或其相关人员可能存在影响本报告观点客观性的潜在利益冲突，投资者请勿将本报告视为投资或其他决定的唯一参考依据。

联系我们

上海	深圳	北京
地址：上海浦东新区陆家嘴街道富城路 99 号震旦国际大厦 30 楼 邮编：200120	地址：深圳市南山区滨海大道 3165 号五矿金融大厦 23 层 邮编：518035	地址：北京市海淀区首体南路 9 号 4 楼 603 室 邮编：100037

Analyst Certification

The research analyst is primarily responsible for the content of this report, in whole or in part. The analyst has the Securities Investment Advisory Certification granted by the Securities Association of China. Besides, the analyst independently and objectively issues this report holding a diligent attitude. We hereby declare that (1) all the data used herein is gathered from legitimate sources; (2) the research is based on analyst's professional understanding, and accurately reflects his/her views; (3) the analyst has not been placed under any undue influence or intervention from a third party in compiling this report; (4) there is no conflict of interest; (5) in case of ambiguity due to the translation of the report, the original version in Chinese shall prevail.

Investment Rating Definitions

The rating criteria of investment recommendations The ratings contained herein are classified into company ratings and sector ratings (unless otherwise stated). The rating criteria is the relative market performance between 6 and 12 months after the report's date of issue, i.e. based on the range of rise and fall of the company's stock price (or industry index) compared to the benchmark index. Specifically, the CSI 300 Index is the benchmark index of the A-share market. The Hang Seng Index is the benchmark index of the HK market. The NASDAQ Composite Index or the S&P 500 Index is the benchmark index of the U.S. market.	Ratings		Definitions
	Company Ratings	BUY	Stock return is expected to outperform the benchmark index by more than 20%;
		ACCUMULATE	Stock relative performance is expected to range between 5% and 20%;
		HOLD	Stock relative performance is expected to range between -10% and 5%;
		SELL	Stock return is expected to underperform the benchmark index by more than 10%;
		NOT RATED	No clear view of the stock relative performance over the next 6 months.
	Sector Ratings	POSITIVE	Overall sector return is expected to outperform the benchmark index by more than 10%;
		NEUTRAL	Overall sector expected relative performance ranges between -10% and 10%;
		CAUTIOUS	Overall sector return is expected to underperform the benchmark index by more than 10%.

General Disclaimer

Minmetals Securities Co., Ltd. (or "the company") is licensed to carry on securities investment advisory business by the China Securities Regulatory Commission. The Company will not deem any person as its client notwithstanding his/her receipt of this report. The report is issued only under permit of relevant laws and regulations, solely for the purpose of providing information. The report should not be used or considered as an offer or the solicitation of an offer to sell, buy or subscribe for securities or other financial instruments. The information presented in the report is under the copyright of the company. Without the written permission of the company, none of the institutions or individuals shall duplicate, copy, or redistribute any part of this report, in any form, to any other institutions or individuals. The party who quotes the report should contact the company directly to request permission, specify the source as Equity Research Department of Minmetals Securities, and should not make any change to the information in a manner contrary to the original intention. The party who re-publishes or forwards the research report or part of the report shall indicate the issuer, the date of issue, and the risk of using the report. Otherwise, the company will reserve its right to taking legal action. If any other institution (or "this institution") redistributes this report, this institution will be solely responsible for its redistribution. The information, opinions, and inferences herein only reflect the judgment of the company on the date of issue. Prices, values as well as the returns of securities or the underlying assets herein may fluctuate. At different periods, the company may issue reports with inconsistent information, opinions, and inferences, and does not guarantee the information contained herein is kept up to date. Meanwhile, the information contained herein is subject to change without any prior notice. Investors should pay attention to the updates or modifications. The analyst wrote the report based on principles of independence, objectivity, fairness, and prudence. Information contained herein was obtained from publicly available sources. However, the company makes no warranty of accuracy or completeness of information, and does not guarantee the information and recommendations contained do not change. The company strives to be objective and fair in the report's content. However, opinions, conclusions, and recommendations herein are only for reference, and do not contain any certain judgments about the changes in the stock price or the market. Under no circumstance shall the information contained or opinions expressed herein form investment recommendations to anyone. The company or analysts have no responsibility for any investment decision based on this report. Neither the company, nor its employees, or affiliates shall guarantee any certain return, share any profits with investors, and be liable to any investors for any losses caused by use of the content herein. The company and its analysts, to the extent of their awareness, have no conflict of interest which is required to be disclosed, or taken restrictive or silent measures by the laws with the stock evaluated or recommended in this report.

Minmetals Securities Co. Ltd. 2019. All rights reserved.

Special Disclaimer

Permitted by laws, Minmetals Securities Co., Ltd. may hold and trade the securities of companies mentioned herein, and may provide or seek to provide investment banking, financial consulting, financial products, and other financial services for these companies. Therefore, investors should be aware that Minmetals Securities Co., Ltd. or other related parties may have potential conflicts of interest which may affect the objectivity of the report. Investors should not make investment decisions solely based on this report.

Contact us

Shanghai

Address:
30F, Aurora Plaza, No. 99 Fucheng Road, Pudong New District, Shanghai
Postcode: 200120

Shenzhen

Address: 23F, Minmetals Financial Center, 3165 Binhai Avenue, Nanshan District, Shenzhen
Postcode: 518035

Beijing

Address: Room 603, 4F, No. 9 Shoutinan Road, Haidian District, Beijing
Postcode: 100037