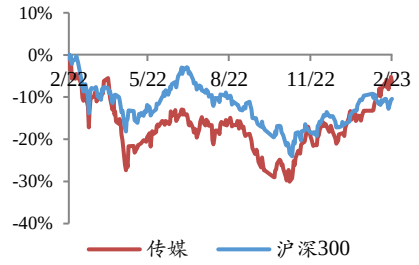
从 ChatGPT 等生成式 AI 的算力开销及商业化潜力，看微软和谷歌面临的挑战

行业评级：增持

报告日期： 2023-02-21

行业指数与沪深 300 走势比较



分析师：金荣

执业证书号：S0010521080002
邮箱：jinrong@hazq.com

相关报告

- 公司点评：快手-W(1024.HK)22Q4 前瞻：商业化及流量盘稳健，降本增效持续 2023-01-30
- 深度点评：东方甄选：借流量红利起步，有望靠品牌红利起飞 2023-01-19
- 公司点评：微盟集团（2013.HK）：SaaS 业务量价齐升，视频号起量充分获益 2023-01-19
- 公司深度：遥望科技(002291.SZ)：多引擎驱动收入增长，大中台彰显规模效应 2023-01-18
- 行业点评：3 年 1 月游戏版号下发，腾讯黎明觉醒和网易逆水寒均获批 2023-01-18
- 公司点评：携程集团-S(9961.HK)：“新十条”放宽出行限制，单季盈利表现不俗 2022-12-19

主要观点：

● 生成式 AI 对搜索引擎是否存在威胁？

类似 ChatGPT 的生成式 AI 在搜索领域实现替代仍然面临诸多挑战，生成式 AI 技术需要先达到一定程度的 “规模优势(包括预训练数据集规模,用户反馈量)”之后才有机会威胁到搜索引擎的生存地位。这种“规模优势”既意味着模型可以解决问题的领域在数量上足够庞大，又意味着同一个领域中模型可交付出的解决路径数量最够庞大。

这种“规模效应”的达成有 2 个制约因素：1）用户习惯的颠覆。基于当前技术迭代路径的“搜索引擎（包括 Google，Baidu，和 Bing 等）”已经教育了市场将近 25 年以上的时间，颠覆用户习惯需要极大的动能，这种动能一定是基于“替代方案”的效率要比“现存方案”优越数倍以上；2）生成式 AI 的模型进化的本质是依赖于对庞大的数据集的训练和微调，其背后的算力支撑是重要的技术驱动因素，而算力支撑取决于芯片技术(材料、设计、生产工艺)及“异构计算技术”的发展进程(包括计算开销的下降和计算交付结果精确程度的提高)。

● 生成式 AI 的算力需求

根据 Next Platform 对前期训练（不含微调）的估算，GPT-3 175B 的模型的每次训练成本在 875 万 - 1093.75 万美元之间，对应花费时间在 110.5 天-27.6 天，每 1 百万参数的训练价格是 50 美元-62.5 美元之间。根据 Cerebras AI model studio 的 GPT-3 模型训练服务（基于 4-node CS-2 cluster）的报价信息，GPT 70B （700 亿参数，14000 亿 Tokens，85 天训练时间）的训练价格在 250 万美元每次。

● 生成式 AI 的商业化潜力

类似 ChatGPT 的生成式 AI，在不久的将来，生成式 AI 有较大可能在“智能客服”和“搜索引擎”进行增值，并有较大可能以“插件”的形式赋能现有的“生产力工具链(工程软件/音视频制作工具等)”。

● 微软和谷歌共同和分别面临的挑战

无论对于微软还是谷歌而言，由于“生成式 AI”所带来的行业变革处于爆发的早期，行业的天花板较高，并且 AI 技术的上游硬件厂商也会微软等模型层和技术层厂商产生溢价，所以“赢者通吃”或“强者恒强”的局面并不会出现。

其共同面临的挑战包括：1）面临细分领域(电商,社交,游戏)的威胁或直接竞争，威胁包括细分领域数据集获得难度增大，竞争包括细分领域巨头直接下场竞争并更易于满足细分领域用户需求。2）算力开销驱动营业成本增加，生成式 AI 在发展早期的商业化绩效目标需要被理性的界定清楚。早期的发展，需要有持续的现金牛业务

支撑 AI 明星业务的研发及运维开支，同时也要避免业务间存在内部不配合和摩擦。3）如全球通胀及供应链扰动持续，生成式 AI 模型层和应用层等下游环节厂商所创造的价值不断转移到上游的硬件或者能源厂商。比如，高端 GPU/FPGA 的厂商较为集中，而可替代品有限，上游厂商有较强溢价权。

微软面临的挑战包括：1）因为是引领 GPT 技术发展的领先企业，所以在面临政府或公益组织对其生成式 AI 生产内容的法律及道德问题时也是首当其冲；2）在智能手机操作系统、娱乐、文化、快消、电商等领域的数据积累有限，在这些领域进行模型训练的学习曲线依然很陡峭；3）如果微软同时掌握了用户接入操作系统，办公软件，和生成式 AI 搜索引擎的入口，则将需要面临更多“反垄断”相关的问题。

谷歌面临的挑战包括：1）对话式 AI 的搜索方式如果走向普及，将威胁现有的点击付费的广告商业模式；2）对现有的搜索引擎技术有一定路径依赖，模型过于庞大和复杂，将生成式 AI 技术整合入搜索引擎需要更长时间的试错；3）缺少先发优势，Open AI 和微软合作更早更深远，而谷歌一直缺少对生成式 AI 技术落地的验证（Bard 并未达到预期）。

- **投资建议**

在模型层面和应用层面，建议关注在 AI 技术领域有较强人才优势，同时有成熟的商业化产品支撑其创新的科技型企业， 包括：百度集团-SW，阿里巴巴-SW，网易-S，Microsoft，Google 等

- **风险提示**

全球通胀和宏观经济下行，供应链扰动和贸易摩擦，导致 AI 计算领域上游硬件供需错配；生成式 AI 产品的商业化表现不及预期；生成式 AI 产品带来更多法律及道德风险

正文目录

1 关于 CHATGPT 要知道些什么？ 5

1.1 CHATGPT 的基本概念及原理 5
1.2 “生成式 AI”的算力成本 8
1.3 “生成式 AI”面向 B 端和 C 端的潜在场景 10

2 “生成式 AI”商业化潜力初探 12

2.1 “生成式 AI”在智能客服领域的潜在应用 12
2.2 “生成式 AI”在搜索引擎领域的潜在应用 15
2.3 “生成式 AI”作为生产力工具插件的潜在应用 17

3 从“生成式 AI”看微软和谷歌面临的挑战 18

3.1 微软和谷歌共同和分别面临的挑战 18
3.1 微软 AI 产品赋能回顾 19
3.2 谷歌 AI 产品赋能回顾 23

风险提示： 25

图表目录

图表 1 CHATGPT 工作原理简要流程图 5
图表 2 CHATGPT 工作流程 6
图表 3 CHATGPT 使用案例– 与机器人对话查找代码错误 7
图表 4 TRANSFORMER 模型结构 7
图表 5 CEREBRAS MODEL STUDIO 对不同 GPT 模型的训练成本报价 8
图表 6 GPT 各模型的训练成本 8
图表 7 GPT3 各类模型的训练的参数量 9
图表 8 英伟达 A100 核心参数 9
图表 9 生成式 AI(GENERATIVE AI) 面向 B 端的潜在场景 10
图表 10 生成式 AI(GENERATIVE AI) 面向 C 端的潜在场景 11
图表 11 全球“对话 AI(CONVERSATIONAL AI)”市场空间 13
图表 12 “智能对话机器人”在各领域全球市场空间（亿 USD） 13
图表 13 全球“智能对话机器人”智能客服领域市场空间(亿 USD) 14
图表 14 智能对话机器人–电商客服领域全球市场空间测算(亿 USD) 14
图表 15 谷歌服务(GOOGLE SERVICES)收入 (亿 USD) 15
图表 16 “生成式 AI”对搜索引擎的影响 16
图表 17 生成式 AI 与搜索引擎结合面临的挑战 16
图表 18 “生成式 AI”作为嵌入式插件的潜在应用领域 17
图表 19 微软和谷歌共同和分别面临的挑战 18
图表 20 “生成式 AI”的应用 – 新 BING (NEW BING) 搜索案例展示 19
图表 21 “生成式 AI”的应用 – 新 BING (NEW BING)可对话的领域举例 20
图表 22 微软具有代表性的 AI 赋能产品和项目 21
图表 23 微软 AI 技术突破时间线 22
图表 24 TRANSFORMER 模型中引入的 ATTENTION 机制示意图 23
图表 25 谷歌巴德(GOOGLE BARD)案例 – 规划旅行路线 24
图表 26 谷歌具有代表性的 AI 赋能产品和项目 24

1 关于 ChatGPT 要知道些什么？

1.1 ChatGPT 的基本概念及原理

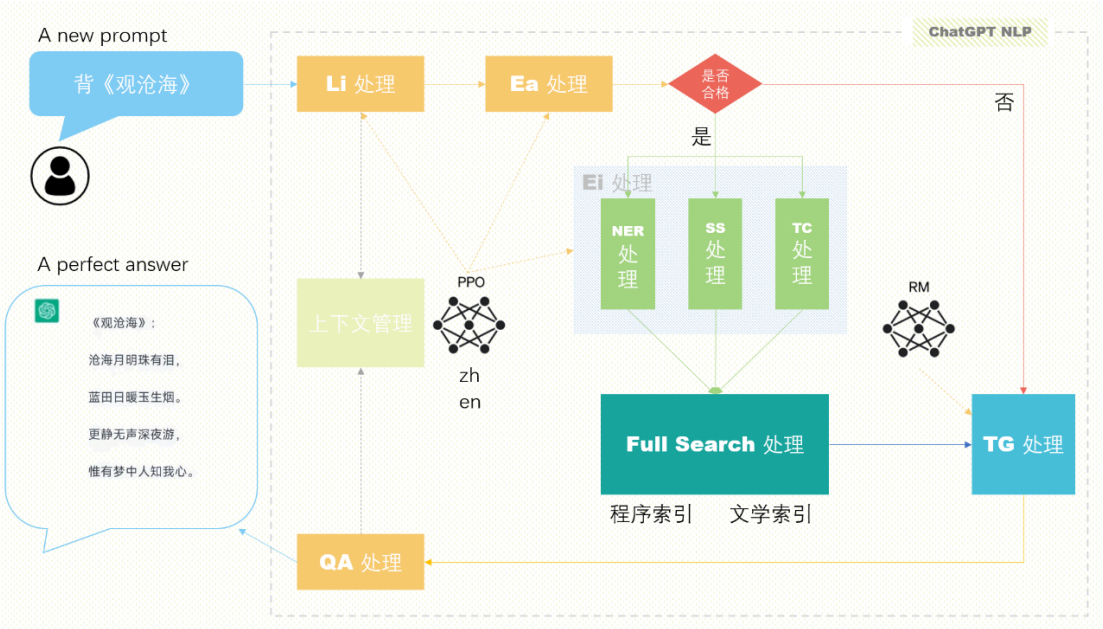
ChatGPT 是 Open AI 公司发布的“语言模型”，该“模型”采用大规模的自然语言(natural language model)算法，这个“模型”拥有和用户之间的交互界面，用户可以将问题提交给“模型”，然后“模型”做出回复，就好比是在和一个智能的机器人进行文字对话。

AI 应用分成分析型 AI 应用和生成式 AI 应用。其中，分析型 AI 应用已经被广泛使用，包括短视频推送、人脸识别、搜索分类、语音助手和等；生成式 AI 还并未被广泛应用，生成式 AI 的应用包括书写文字段落，生成图片，和生成代码等。

ChatGPT 是一种生成式 AI。ChatGPT 的工作原理可以简单的类比为，机器通过阅读人类说出前半句话是什么，然后预测人类想要得到的后半句话应该是什么，可以类比为诗词填空，出题人(即，人类)给出诗词的前半句，然后让答题者(即，机器)填写后半句。达成这个目标，需要用人类已经存在的语料信息、图片信息或代码信息等作为数据集，对模型进行训练，所以模型的认知边界将不断的趋近但无法超越人类自身的认知边界。

简要的来说，GPT 模型的实现有 2 个关键环节：第一，需要先用大量的数据对 AI 模型进行预训练，得到预训练模型；第二，在这个基础上进行一些人为的标注，进行微调，提高其预测的准确度和可靠性。比如，要生成一个可以正确地帮助人类规划旅行路线的模型，模型的开发人员会使用规模极大的数据集（包含酒店信息、旅行路线请求信息、地图信息、天气信息等）去进行模型的预训练，得到一个通用的预训练模型；然后在这个通用模型的基础上，使用外包的人工标注团队，开展人工标注，对一些细节进行标注，比如将已经停业的酒店标注出来，确保最终 ChatGPT 在规划旅行计划时不包括这些已经停业的酒店。

图表 1 ChatGPT 工作原理简要流程图



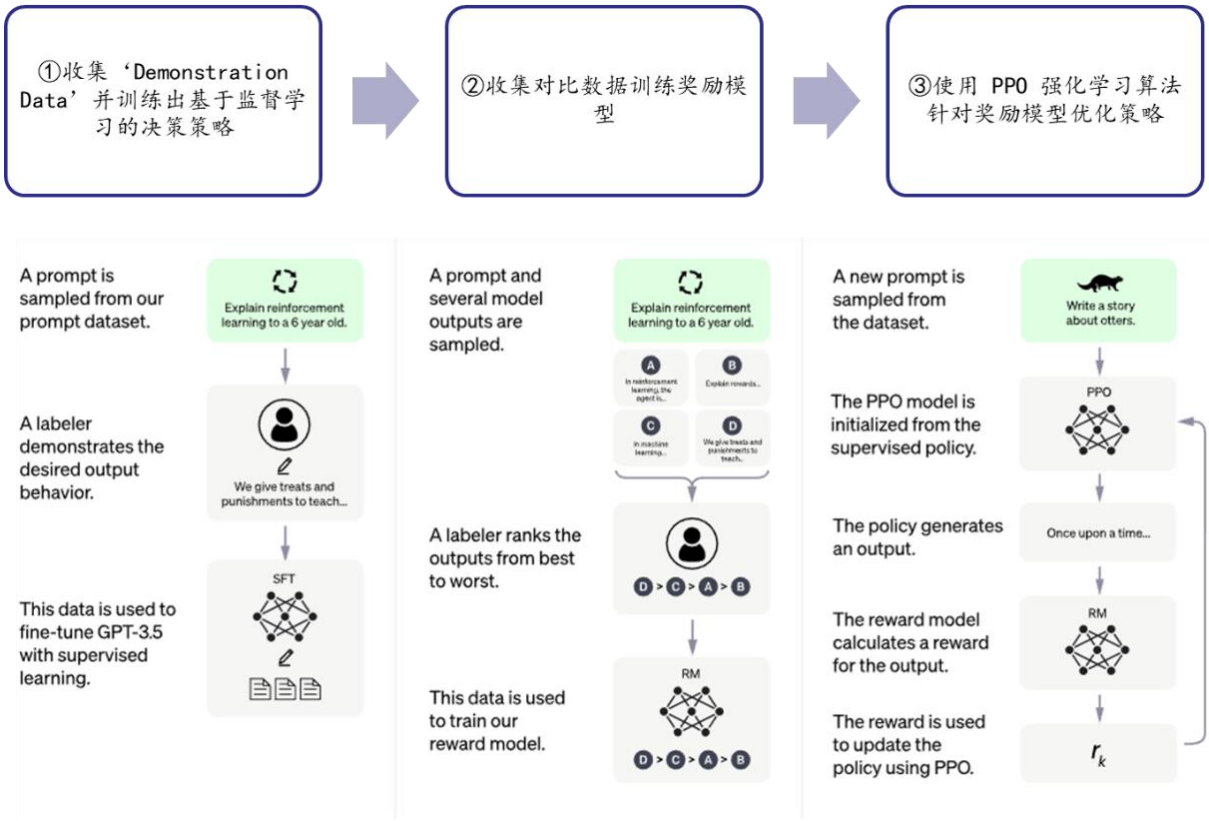
资料来源：InfoQ，华安证券研究所

根据 Open AI 发布的信息，ChatGPT 的训练流程具体包括 3 部分：1）收集“展示数据(Demonstration Data)”并训练出基于监督学习的决策策略；2）收集对比数据训练奖励模型；3）使用 PPO 强化学习算法针对奖励模型优化策略。ChatGPT 目前基于 GPT-3 和 GPT-4 技术之间，GPT-3 于 2020 年发布，GPT-4 有望于 2023 年上半年发布。ChatGPT 的开发者采用了监督学习(Supervised Learning)和强化学习(Reinforcement Learning)算法去微调（fine-tune）ChatGPT 模型；其中，开发者采用了具有人类反馈的强化学习算法(Reinforcement Learning from Human Feedback，RLHF)，最小化了模型所输出回答和真实情况的偏差度。

从技术的起源来看，ChatGPT(全称：Chat Generative Pre-Trained transformer)，采用了生成式的预训练的“Transformer”模型，而 Transformer 模型于 2017 年由谷歌的研究人员在论文《Attention is all you need》中发布，这个模型促成了自然语言学习(NLP)领域中 GPT 和 BERT 这 2 大模型的发展。这些年，在自然语言学习(NLP)领域，Transformer 模型逐步替代 RNN(循环神经网络)和 CNN(卷积神经网络)；比如，相比于 RNN 模型，Transformer 模型引入了自我注意力(Self-attention)机制，结合算法优化，可以实现并行运算，大量节约训练时间。

ChatGPT 的回答包括不限于以下几个方面：1）直接回答问题；2）做文字的摘要；3）写代码和查代码错误；4）语言翻译；5）撰写演讲稿和故事等。比如，一个经典的使用案例就是用户可以在和 ChatGPT 模型的对话框中，复制一段程序代码并向 ChatGPT 提问如何修改以下这段代码以确保这段代码可以正常工作，然后 ChatGPT 会进行回应，将指出代码的出现的问题所在。

图表 2 ChatGPT 工作流程



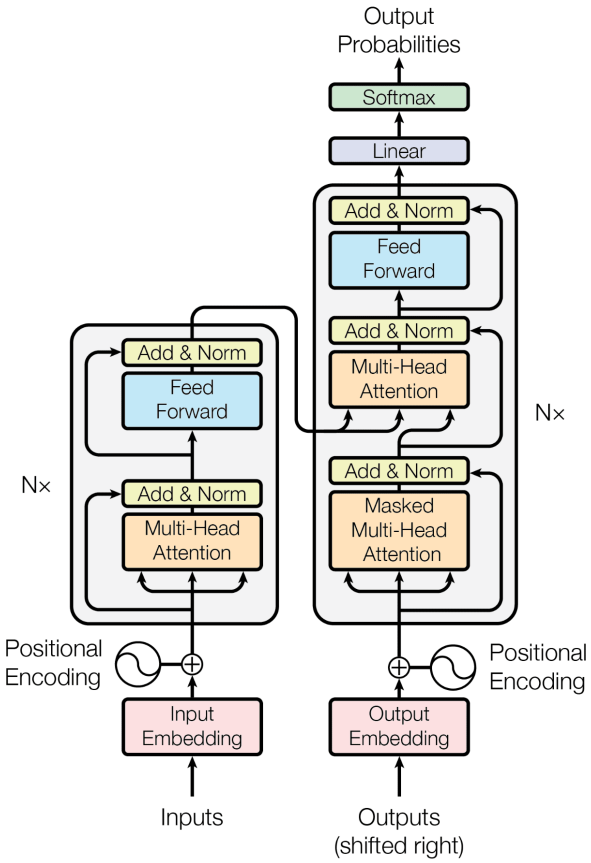
资料来源：Open AI 官网，华安证券研究所整理

图表 3 ChatGPT 使用案例– 与机器人对话查找代码错误



资料来源：Open AI 官网，华安证券研究所

图表 4 Transformer 模型结构



资料来源：Google Research，华安证券研究所

1.2 “生成式 AI”的算力成本

根据 Next Platform 发布的信息，Cerebras AI model studio 的 GPT-3 模型训练服务（基于 4-node CS-2 cluster）的报价信息如下。已知的 GPT 70B（700 亿参数，14000 亿 Tokens，85 天训练时间）的训练价格在 250 万美元。而 GPT-3 模型最多有 175 B 参数(即，1750 亿参数)，根据 Next Platform 对前期训练（不含微调）的估算，**GPT-3 175B 的模型的训练成本在 875 万 –1093.75 万美元之间，对应花费时间在 110.5 天-27.6 天，每 1 百万参数的训练价格是 50 美元-62.5 美元之间。**其中，GPT-3 175 B 如果是基于 4-node CS-2 cluster 设备配置则成本是 875 万美元，如果是基于 16-node CS-2 cluster 设备配置则成本是 1093.75 万美元。

图表 5 Cerebras Model Studio 对不同 GPT 模型的训练成本报价

Model	Parameters	Tokens to Train to Chinchilla Point (B)	Cerebras Model Studio CS-2 Day to Train	Cerebras Model Studio Price to Train
GPT3-XL	1.3	26	0.4	$2,500
GPT-J	6	120	8	$45,000
GPT-3 6.7B	6.7	134	11	$40,000
T-5 11B	11	34*	9	$60,000
GPT-3 13B	13	260	39	$150,000
GPT NeoX	20	400	47	$525,000
GPT 70B	70	1,400	85	$2,500,000
GPT 175B	175	3,500	Contact For Quote	Contact For Quote

* - T5 tokens to train from the original T5 paper. Chinchilla scaling laws not applicable.

资料来源：Next Platform，华安证券研究所

图表 6 GPT 各模型的训练成本

模型类别	参数量（十亿,Bn）	Tokens（十亿,Bn）	服务器配置	训练日期时长（天，days）	训练成本(USD)	每 1 百万参数训练成本(USD)
GPT-3XL	1.3	26	4*CS-2	0.4	2,500	1.9
GPT-J	6	120	4*CS-2	8	45,000	7.5
GPT-3 6.7B	6.7	134	4*CS-2	11	40,000	6.0
T-5 11B	11	*34*	4*CS-2	9	60,000	5.5
GPT-3 13B	13	260	4*CS-2	39	150,000	11.5
GPT NeoX	20	400	4*CS-2	47	525,000	26.3
GPT NeoX	*20*	*400*	*16*CS-2*	*11.1*	*656,250*	*32.8*
GPT 70B	70	1400	4*CS-2	85	2,500,000	35.7
GPT 70B	*70*	*1400*	*16*CS-2*	*21.3*	*3,125,000*	*44.6*
GPT 175B	175	3500	4*CS-2	110.5	8,750,000	50.0
GPT 175B	*175*	*3500*	*16*CS-2*	*27.6*	*10,937,500*	*62.5*

资料来源：Next Platform，华安证券研究所整理

图表 7 GPT3 各类模型的训练的参数量

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Fig-2: Details of variants of the GPT-3 model

资料来源：Sigmoid，华安证券研究所整理

图表 8 英伟达 A100 核心参数

	Peak Performance
Transistor Count	54 billion
Die Size	826 mm^2
FP64 CUDA Cores	3,456
FP32 CUDA Cores	6,912
Tensor Cores	432
Streaming Multiprocessors	108
FP64	9.7 teraFLOPS
FP64 Tensor Core	19.5 teraFLOPS
FP32	19.5 teraFLOPS
TF32 Tensor Core	156 teraFLOPS | 312 teraFLOPS*
BFLOAT16 Tensor Core	312 teraFLOPS | 624 teraFLOPS*
FP16 Tensor Core	312 teraFLOPS | 624 teraFLOPS*
INT8 Tensor Core	624 TOPS | 1,248 TOPS*
INT4 Tensor Core	1,248 TOPS | 2,496 TOPS*
GPU Memory	40 GB
GPU Memory Bandwidth	1.6 TB/s
Interconnect	NVLink 600 GB/s PCIe Gen4 64 GB/s
Multi-Instance GPUs	Various Instance sizes with up to 7MIGs @5GB
Form Factor	4/8 SXM GPUs in HGX A100
Max Power	400W (SXM)

资料来源：Serve the home，华安证券研究所

1.3 “生成式 AI”面向 B 端和 C 端的潜在场景

ChatGPT 是一种较为高级形态的生成式 AI(Genrative AI)。它可以通过与用户对话的方式提供人工智能程度较高的应答。基于生成式 AI 现在的发展情况，及 ChatGPT 在这个领域的进步，未来类似 ChatGPT 的智能对话机器人有望为产品或服务的价值链所有环节增值。在一个产品涉及到的市场营销及销售，产品运营/生产运营，研发/设计，和职能等众多领域，只要涉及到与人打交道的环节，智能对话机器人未来都将有望胜任这些工作。基于 NLP 技术的大规模模型，将帮助智能对话机器人应用实现从“AI 辅助”到“AI 替代”的跨越。

生成式 AI 可以较好的赋能 B 端和 C 端。对于 B 端而言，生成式 AI 的意义在于对于人力资源的替代。比如，在市场营销环节中，应用生成式 AI 的智能对话机器人可以结合用户需求，生成高度定制化的社交媒体广告推送内容，这些内容形态将不限于文字、视频和图片；在产品交付给客户之后，机器人可以作为客服，接收用户的反馈，并为客户提供产品使用的指导或者解决临时问题。同时，机器人可以很好的作为“内容创作者”出现，帮助用户编辑图片、编辑影视动画桥段、游戏关卡、或甚至生成文学小说等内容。

图表 9 生成式 AI(Generative AI) 面向 B 端的潜在场景

消费者	生成式AI(Generative AI)				生产者
	市场营销及销售	**产品运营/生产运营**	**研发/设计/内容创作**	**职能(HR,风控)**	
	广告。生成高度定制化的社交媒体广告推送内容(包括文字、视频、图片等)；	**客服。**分析客户的提交的临时情况或突发情况的反馈，分析问题原因，提供及时的应答及技术解决方案；	**蓝图设计。**研发人员通过输入用户需求，让模型自动设计出解决方案及工作流程；	**人员培训。**结合用户需求和产品及服务的特点，开展各个部门职员的培训。参与者可以通过与模型对话的方式获得培训；	
	需求匹配。接收客户关于当前需求的提问，并为其匹配对应的消费产品及服务；	**用户指导。**用户可以通过提问的形式获得关于产品如何使用的指导；	**设计技术细节/撰写代码。**通过输入具体的用户需求和约束条件，让模型生成产品参数、电路图、工程图纸或者代码；	**招聘。**可以很好的与应聘者进行沟通，解答应聘者的疑问，并从应聘者回答的方式分析其是否符合招聘要求；	
	客户关系维护。和消费者进行互动，定期获得消费者的反馈及需求，给消费者提供短期折扣或者促销方案。	**生产问题解决。**针对产品制造过程中出现的质量问题，工艺效率问题，结合产品的特点，提供解决方案。	**内容创作。**通过用户提交的需求，创作出对应的内容(文字,语音，视频，图片)	**识别风险。**通过监督各个部门运行过程中出现的问题，及时识别出风险项，并为风险程度划定优先级。	

资料来源：Open AI，McKinsey，华安证券研究所整理

对于 C 端而言，应用生成式 AI 技术的机器人可以作为用户的个人智能助理出现，其硬件载体将不局限于智能终端，智能驾驶平台和智能家居。机器人在 C 端的增值环节，有 2 个不同层面：1）执行层；2）决策层。执行层任务将更多的以“接受用户指令-并提供’siri 式’的用户反馈”为核心；决策层的任务将基于生成式 AI 的“自学习能力”，为用户解决一些进阶的需求，包括基于用户的日程安排，帮用户安排任务的优先级。

图表 10 生成式 AI(Generative AI) 面向 C 端的潜在场景

	执行层	决策层
智能终端	类似用户个人助理，接受指令并帮助用户完成更高阶的需求，比如帮助用户制定出日程安排表，并帮用户订好机票并规划好路线	具有一定自学习能力，可以主动帮助用户解决工作中遇到的浅层次问题，比如项目优先级及时间安排冲突的重新调整等
智能驾驶平台	作为驾驶员助手，接受指令并帮助驾驶员调节座舱功能，选择规划路径，打通对话 AI 系统和智能驾驶系统	可以自学习用户日常的指令，按照用户不同时间段的出行需求及路况，规划路线并实施自动驾驶
智能家居	接受指令，协助用户调节智能家居功能，作为指令中转枢纽，打通多种智能家居电子设备，同时可以监控家庭内的卫生安全并及时向用户反馈	具备自学习能力，可以根据外界环境及边界条件的变化，为用户调整生活空间的舒适度，并通过与用户进行互动进行调整

资料来源：Open AI，McKinsey，华安证券研究所整理

2 “生成式 AI”商业化潜力初探

“生成式 AI (generative AI)”市场化空间较为广阔，基于现行的 NLP 算法发展程度及数据集规模，在不久的将来，生成式 AI 有较大可能在“智能客服”和“搜索引擎”进行增值，并有较大可能以“插件”的形式赋能现有的“生产力工具链(工程软件/音视频制作工具等)”。

在为客服领域增值的过程中，有希望在人工客服的全链路中（包括问题识别、潜在解决方案交付、反馈优化、和风险识别等）实现对人工客服的替代。可实现替代的领域包括电商零售、医疗健康、金融服务、和电信等领域的客服环节。

在为搜索引擎增值的过程中，生成式 AI 在中短期之内，并不会颠覆现如今的搜索引擎地位，但是会逐步成为搜索引擎进步的主要技术驱动因素，相比于传统搜索引擎技术，基于“生成式 AI”技术的搜索反馈结果，将基于其“自学习”能力适应不同用户的需求变化，其最终结果交付将实现从“信息的匹配和中转”到“内容创造和问题解决”的跃迁。

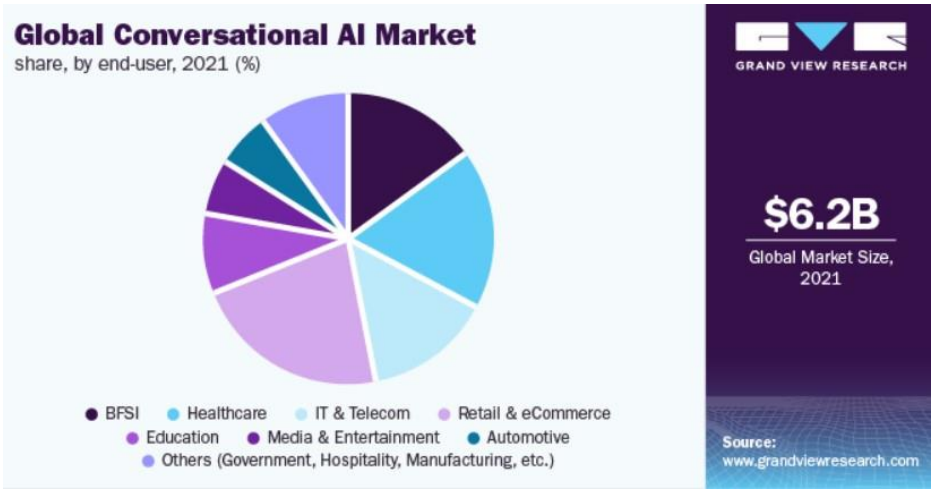
在赋能“生产力工具链”的过程中，生成式 AI 未来有希望通过“插件”的形式，作为工具链的一部分，结合使用者的请求生成工程图纸、代码、图形等，提高工具的易用度，降低人的工具执行成本。这些工具涉及到的领域包括但不限于，工程设计（包括 Solidworks，CATIA,和 AUTOCAD 等），代码设计/运算(Matlab，TensorFlow 等)，音视频制作（Adobe Photoshop 等），游戏/图形设计(Unity, UE4 等)，操作系统（IOS，Linux，智能车操作系统），和元宇宙仿真（Omniverse）。

2.1 “生成式 AI”在智能客服领域的潜在应用

根据 Grand View Research，全球“可对话 AI(Conversational AI)”2021 年市场空间为 62 亿美元，其中，BFSI(银行保险等金融服务)，医疗，零售和电商，和电信领域的市场空间为 42.5 亿美元左右，占 68.5%左右；根据 Grand View Research 预测，2030 年“可对话 AI”的市场空间将达到约 413.9 亿美元，对应 2022 至 2030 年复合增长率(CAGR)为 23.6%。该领域的核心竞争者包括：谷歌，微软，亚马逊，IBM，甲骨文，和 SAP 等。市场增长的主要驱动因素包括各领域对于应用 AI 技术替代人力这一需求的提升，和持续下降的 AI 对话程序的研发成本。

根据 Reuters 报道，目前智能对话机器人领域的佼佼者 Open AI 公司（核心产品是 ChatGPT），2024 年的收入有望达到 10 亿美元，目前的估值约为 200 亿美元。

图表 11 全球“对话 AI(Conversational AI)”市场空间



资料来源：Grand View Research，华安证券研究所

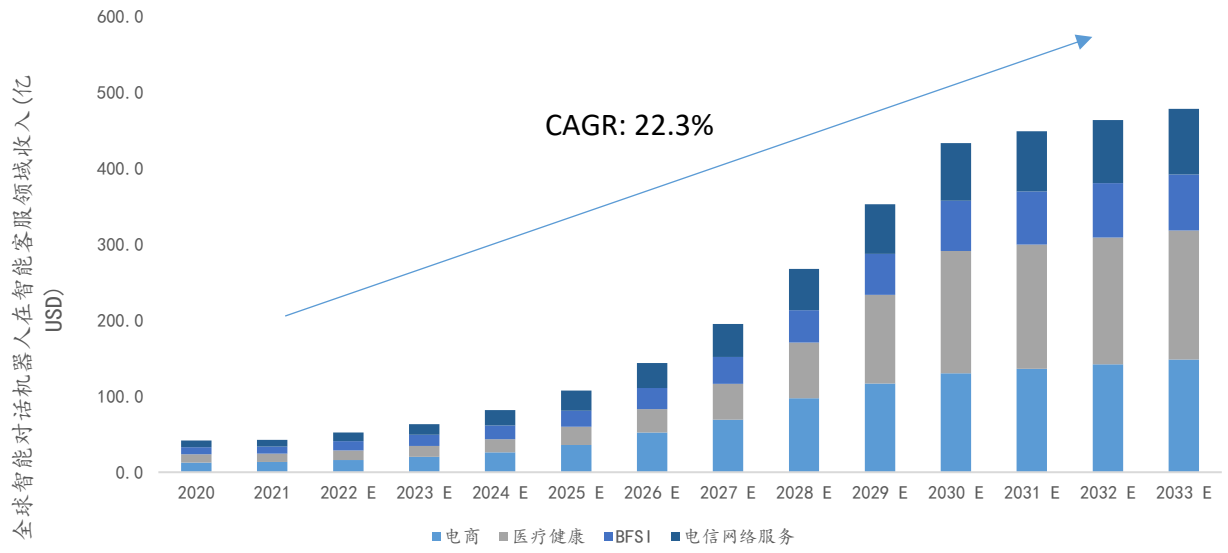
我们判断，从现在起至未来几年，作为“生成式 AI”重要应用的“对话 AI(Conversational AI)”的商业化模式中较为清晰且可行的，是在各个领域对于人工客服的替代。基于以下两点考虑：1、全球主要发达经济体人口增长乏力，劳动力数量减少，用工成本攀升，有强烈的使用 AI 对话机器人替代人工客服的需求；2、“智能对话机器人”相比“人工客服”创造更多价值，即，机器人可以完成更多人工客服无法胜任的任务，并且工作效率高，解决问题出错率较低。3、“智能对话机器人”研发成本的显著下降，部署该机器人的实践经验的可复制性不断提高。由于电商、医疗健康、BFSI、和电信网络服务的客服中产生的问题及解答，较为结构化并依赖劳动力密集产出（其中，医疗健康领域不包括医生看诊环节，仅包括挂号预约、初步咨询、取药、和护理服务沟通等专业性较低环节），所以这 4 个应用领域有望成为“可对话 AI”可以进行“人力资源替代”的主要领域。我们从这 4 个主要领域入手，基于对未来驱动因素的假设，进行了市场空间测算。根据 Grand View Research，2021 年这 4 个主要领域市场空间约为 42.5 亿美元左右；基于此，根据我们的测算，**2033 年这 4 个主要领域的市场空间可以达到 478 亿美元。**

图表 12 “智能对话机器人”在各领域全球市场空间（亿 USD）

(单位：亿 USD)	2020	2021	2022 E	2023 E	2024 E	2025 E	2026 E	2027 E	2028 E	2029 E	2030 E	2031 E	2032 E	2033 E
电商	12.8	13.4	16.1	20.1	26.0	36.0	52.2	69.3	97.3	116.6	130.3	136.1	142.2	148.4
医疗健康	10.9	11.0	12.7	14.3	17.6	24.1	30.7	47.1	73.5	116.7	160.6	163.6	166.6	169.7
BFSI	9.3	9.4	12.2	15.0	17.9	20.8	27.9	35.2	42.6	54.3	66.3	69.7	71.8	73.9
电信网络服务	8.6	8.7	11.2	13.8	20.1	26.5	33.1	43.6	54.3	65.1	76.1	79.4	82.7	86.0
总计	**41.6**	**42.5**	**52.1**	**63.1**	**81.6**	**107.4**	**143.9**	**195.1**	**267.6**	**352.6**	**433.3**	**448.8**	**463.3**	**478.0**

资料来源：世界银行, Grand view research, Statista，华安证券研究所测算

图表 13 全球“智能对话机器人”智能客服领域市场空间(亿 USD)



资料来源：世界银行，Grand View Research，华安证券研究所测算

图表 14 智能对话机器人-电商客服领域全球市场空间测算(亿 USD)

智能对话机器人-电商领域全球市场空间测算

核心假设：

年平均网购次数（次/人·年)	18
每年与客服沟通次数/每年网购次数	80%
客服效率(次/(人·年))	18000
客服年薪(USD/年)	9000

	2020	2021	2022 E	2023 E	2024 E	2025 E	2026 E	2027 E	2033 E
全球电商购物人数（亿人)	**20.5**	**21.4**	**22.3**	**23.2**	**24.1**	**25.0**	**25.9**	**26.7**	**31.7**
yoy	*N/A*	*4.4%*	*4.2%*	*4.0%*	*3.80%*	*3.70%*	*3.60%*	*3.30%*	*2.80%*
客服服务人次需求（亿次)	295.2	308.2	321.1	333.9	346.6	359.4	372.3	384.6	456.6
客服人员需求（亿人)	0.016	0.017	0.018	0.019	0.019	0.020	0.021	0.021	0.025
yoy	*N/A*	*4.4%*	*4.2%*	*4.0%*	*3.8%*	*3.7%*	*3.6%*	*3.3%*	*2.8%*
全球电商客服支出(亿 USD)	*147.6*	*154.1*	*160.5*	*166.9*	*173.3*	*179.7*	*186.2*	*192.3*	*228.3*
全球“对话 AI” 电商订阅收入（亿 USD)	**12.81**	**13.43**	**16.08**	**20.06**	**26.02**	**35.97**	**52.16**	**69.27**	**148.43**
“对话 AI “订阅收入/电商客服支出	8%	8%	10%	12%	15%	20%	28%	36%	65%

资料来源：世界银行, Grand view research, Statista，华安证券研究所测算

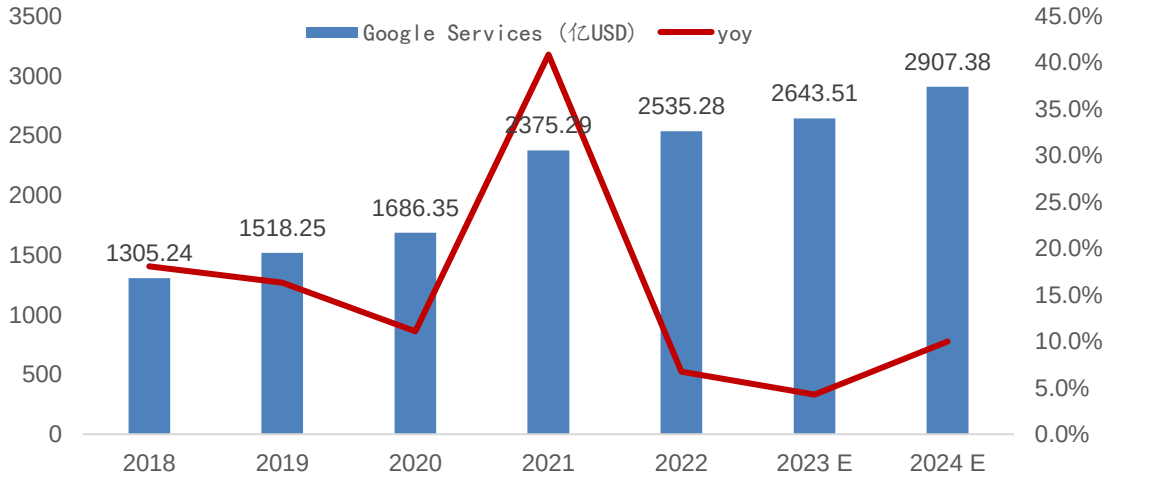
2.2 “生成式 AI”在搜索引擎领域的潜在应用

在互联网搜索领域，目前谷歌占据绝对的领军地位。谷歌主要的业务板块是谷歌服务(Google Services)，2022 年谷歌服务收入 2535.28 亿美元。谷歌服务依赖谷歌在搜索引擎领域长期积累的技术和商业化优势，包括搜索，Youtube，google play 智能手机平台，广告，浏览器，邮箱，云盘等。

在考虑未来“生成式 AI”在搜索领域可以实现的收入时，考虑范围应该不局限于“搜索引擎”本身，还应该包括基于“搜索引擎”技术所拓展出的外延部分，比如 Youtube，Google Map 和 Google Play 等；这些搜索引擎外延部分的商业化表现，本质是基于平台是否可以基于用户的搜索请求或者使用习惯，将产品及服务较精准的分发给用户，然后用户并为此付费。所以，谷歌服务部分的收入天花板，可以用来作为“标尺”去衡量未来智能对话机器人在“搜索领域”的潜在收入中最易于理解的部分。

“生成式 AI(Generative AI)”未来有较大技术进步空间和较难估量的商业化潜力。基于以下 3 点原因，在中长期，“生成式 AI”的市场天花板应该不限于当前搜索引擎所达到的成就：1）它为用户增值的目标从“信息的搜索和呈现”跨越到了“独立解决问题”；2）它在“解决问题”的过程中，存在较大的商业化空间，可以将定制化的“产品”或“服务”分发给终端用户；3）它具备一定自学习能力，基于不同用户的使用习惯，有希望为用户提供定制化的体验。

图表 15 谷歌服务(Google Services)收入 (亿 USD)



资料来源：Bloomberg，华安证券研究所

但是生成式 AI 在搜索领域实现替代仍然面临诸多挑战，生成式 AI 技术需要先达到一定程度的 “规模优势(包括预训练数据集规模,用户反馈量)”之后才有机会威胁到搜索引擎的生存地位。这种“规模优势”既意味着模型可以解决问题的领域在数量上足够庞大，又意味着同一个领域中模型可交付出的解决路径数量最够庞大。这种“规模效应”的达成有 2 个制约因素：1）用户习惯的颠覆。基于当前技术迭代路径的“搜索引擎（包括 Google，Baidu，Bing 等）”已经教育了市场将近 25 年以上的时间，颠覆用户习惯需要极大的动能，这种动能一定是基于“替代方案”的效率要比“现存方案”优越数倍以上；2）生成式 AI 的模型进化的本质是依赖于对庞大的数据集的训练和微调，其背后的算力支撑是重要的技术驱动因素，而算力支撑取决于芯片技术(材料、设计、工艺制程)及“异构计算技术”的发展进程(包括计算开销的下降和计算交付结果精确程度的提高)。

图表 16 “生成式 AI”对搜索引擎的影响

输入	过程		输出	商业化路径
文字	内容分发者	理解搜索语义	交付搜索结果	广告
图片		内容分类	广告推送/排名	
语音		内容排名		
		内容匹配		



输入	过程		输出	商业化路径
复杂表述的文字段落	内容创造者	归纳/演绎	AI生成图片，视频，声音等	打通生产力工具链，订阅付费
代码/伪代码		自学习进化	AI生成文字	定制化广告
视频/复杂声音信号		交付质量控制	AI生成工程设计，代码	
电路图/机械图纸				

资料来源：Marketing AI Institute，华安证券研究所整理

图表 17 生成式 AI 与搜索引擎结合面临的挑战

生成式 AI 与搜索引擎结合面临的挑战			
供给侧		需求侧	
数据集	训练过程	用户认知	终端用户增值
- 样本规模	- 算力开销	- 搜索习惯固化	- 增值显著性
- 样本有效性	- 算法效率	- 较高学习成本	
- 样本获取合法性			

资料来源：Open AI，华安证券研究所整理

2.3 “生成式 AI”作为生产力工具插件的潜在应用

生成式 AI 有较大的可能会以插件的形式，为现有的生产力工具链赋能。其中，工程设计、音视频制作和游戏设计等领域工具，是搭建未来基于 XR 的“元宇宙世界”的重要工具。生成式 AI，可以简化用户进行操作的难度，快速将用户的想法反馈成设计结果。比如，在面向 B 端的工业元宇宙的场景中，在搭建虚拟数字化的工业生产线时，用户在未来有可能会通过向生成式 AI 提出请求，然后 AI 自动生成对应的对应设备的 3D 模型和生产工艺流程。在面向 C 端的游戏场景中，生成式 AI 有希望帮助游戏开发人员降低研发难度，即便不会写代码的游戏策划人员，也可以通过 AI 请求，让 AI 生成对应的游戏组件甚至是场景关卡。在商业化的过程中，生成式 AI 的插件有希望以订阅制付费的形式，作为现有生产力工具的一部分，向用户收费。

图表 18 “生成式 AI”作为嵌入式插件的潜在应用领域



资料来源：华安证券研究所整理

3 从“生成式 AI”看微软和谷歌面临的挑战

3.1 微软和谷歌共同和分别面临的挑战

无论对于微软还是谷歌而言，由于“生成式 AI”所带来的行业变革处于爆发的早期，行业的天花板较高，并且 AI 技术的上游硬件厂商也有一定可能会对微软等模型层和技术层厂商产生溢价，所以“赢者通吃”或“强者恒强”的局面并不会出现。

图表 19 微软和谷歌共同和分别面临的挑战

	微软 Microsoft	谷歌 Google
共同面临的挑战	面临细分领域(电商,社交,游戏)的威胁或直接竞争，威胁包括细分领域数据集获得难度增大，竞争包括细分领域巨头直接下场竞争并更易于满足细分领域用户需求。	
	算力开销驱动营业成本增加，生成式 AI 在发展早期的商业化绩效目标需要被理性的界定清楚。同时，早期的发展，需要有持续的现金牛业务支撑 AI 明星业务的研发及运维开支，同时也要避免业务间存在内部不配合和摩擦。	
	如全球通胀及供应链扰动持续，生成式 AI 模型层和应用层等下游环节厂商所创造的价值不断转移到上游的硬件或者能源厂商。比如，高端 GPU/FPGA 的厂商较为集中，而可替代品有限，上游厂商有较强溢价权。	
分别面临的挑战	因为是引领 GPT 技术发展的领先企业，所以在面临政府或公益组织对其生成式 AI 生产内容的法律及道德问题时也是首当其冲。	对话式 AI 的搜索方式如果走向普及，将威胁现有的点击付费的广告商业模式。
	在智能手机操作系统、娱乐、文化、快消、电商等领域的数据积累有限，在这些领域进行模型训练的学习曲线依然很陡峭。	对现有的搜索引擎技术有一定路径依赖，模型过于庞大和复杂，将生成式 AI 技术整合入搜索引擎需要更长时间的试错。
	如果微软同时掌握了用户接入操作系统，办公软件，和生成式 AI 搜索引擎的入口，则将需要面临更多“反垄断”相关的问题。	缺少先发优势，Open AI 和微软合作更早更深远；而一直缺少对生成式 AI 技术落地的验证(Bard 并未达到预期)。

资料来源：Open AI，Google AI，华安证券研究所整理

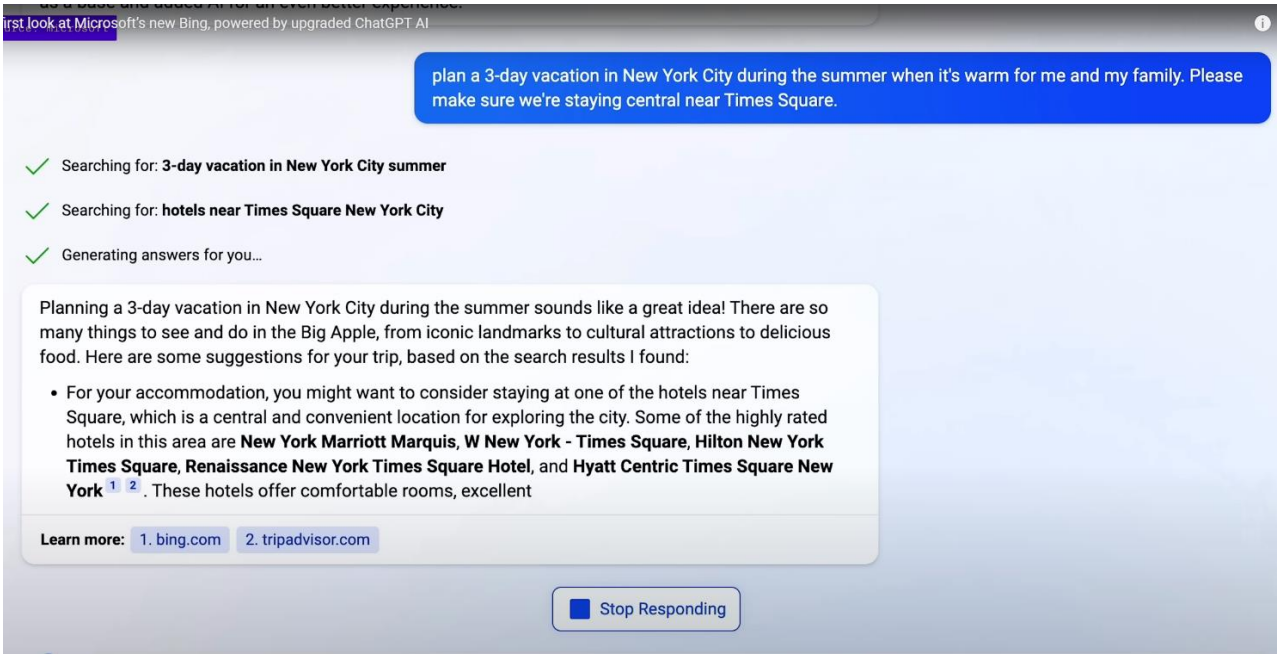
3.1 微软 AI 产品赋能回顾

2023 年 2 月初，微软公布新 Bing (New Bing)搜索引擎，该版本搜索引擎集成了 ChatGPT 技术，公布不到 48 小时，申请用户量已经过百万。用户需要排队注册申请并等待获得使用新 Bing 测试版的资格。

根据微软官方的解释，新 Bing 可以扮演类似研究助理(research assistant)，个人计划员(personal planner)，和创意合作伙伴(creative partner)的角色。和常规的搜索引擎相比，新 Bing 的搜索结果将不再是简单的提供给用户一个链接列表，而是给用户一个概括的答案，解决用户的具体问题，并且提供可靠的信息来源。用户可以按照思考和沟通的方式与新 Bing 对话。同时，新 Bing 也可以作为创意工具，帮助用户写诗，写故事或者写分享关于项目的想法。根据 The Verge 发布的新 Bing 测试使用体验，用户可以直接向 Bing 提问如何帮助自己规划一个在纽约市 3 日的旅行，同时确保自己可以待在纽约时代广场附近；而新 Bing 的回复可以做到将酒店的选择按照一定优先级为用户规划出来。总之，基于 ChatGPT 技术的新 Bing，有如下拟人化的对话特点：

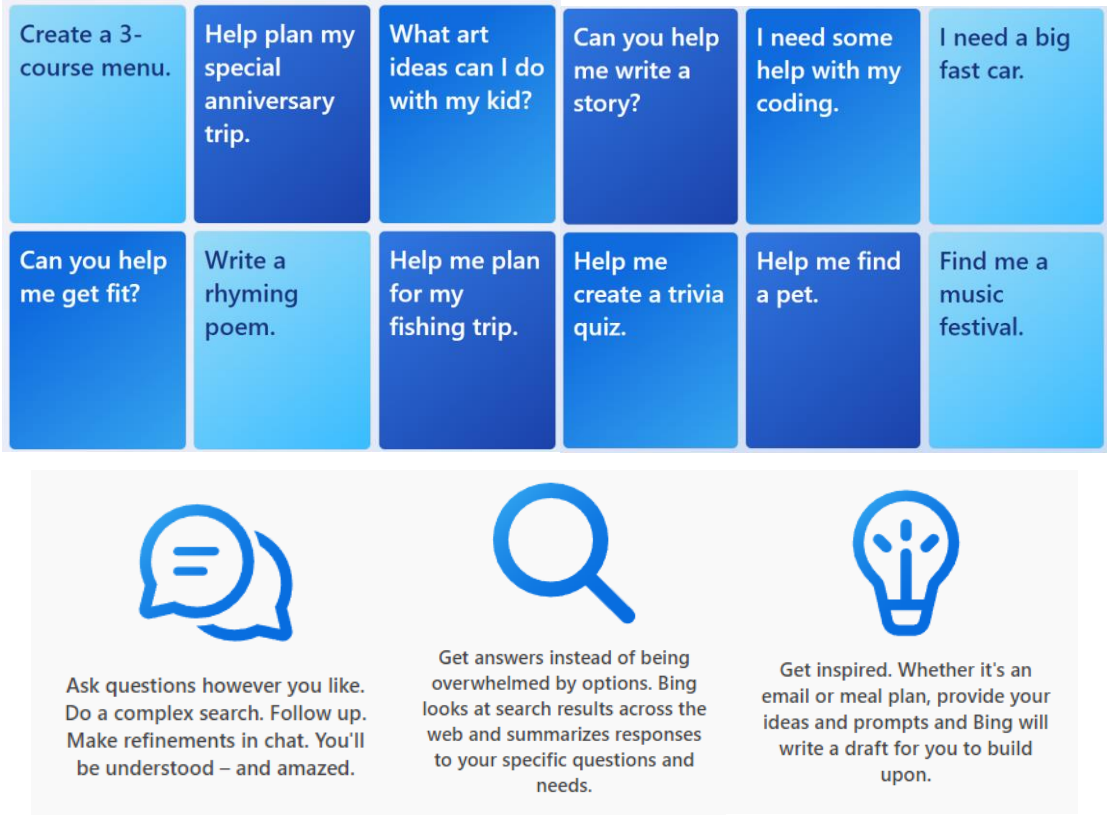
- 可以解决具体问题并给出凝练的回答
- 可以接受用户的追问，得到定制化的回应
- 可以承认自身的错误并进行一定程度的调整

图表 20 “生成式 AI”的应用 – 新 Bing (New Bing) 搜索案例展示



资料来源：The Verge，华安证券研究所整理

图表 21 “生成式 AI”的应用 – 新 Bing (New Bing)可对话的领域举例



资料来源：Bing，华安证券研究所

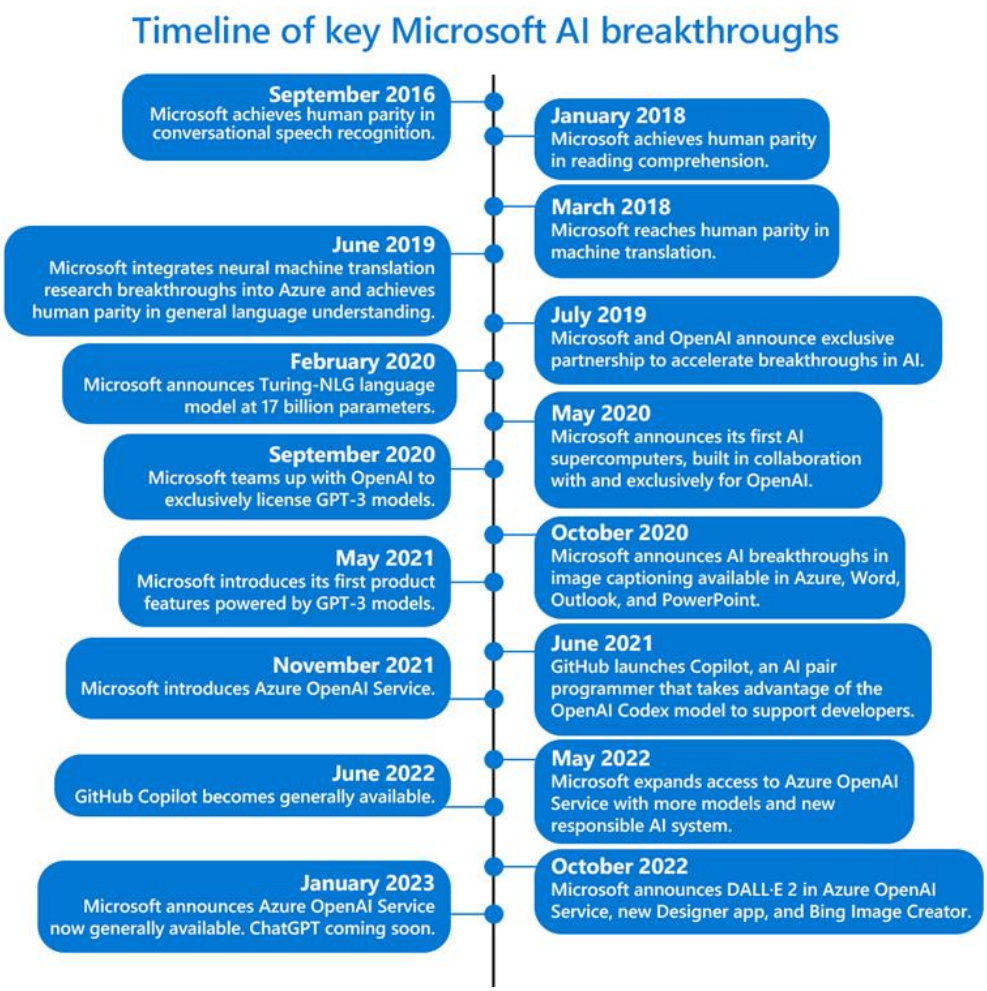
从生成 AI 技术走向成熟到尝试触达用户，微软对于生成式 AI（generative AI）加大投资和整合进自身产品线的举措，不仅是公司层面的转折事件，更是整个科技行业的转折事件。从微软过去在新兴 AI 技术的落地上，已经证明了自身的技术和产品实力。在近些年的 AI 技术落地的实践中，微软在 Office 组件，Azure 云服务，GitHub Copilot, Viva 和 Team Premium 中落地了 AI 相关技术，为产品终端用户提升了工作效率。由于微软已经通过较多的用例积累验证了其在 AI 领域中研发和应用的竞争优势，只要微软可以在未来几年中持续平衡好研发支出和营业利润，微软仍然有较大概率在接下来的由“生成式 AI”所引领的行业变革中，持续引领行业去拓展“生成式 AI”的技术及商业化版图的边界。

图表 22 微软具有代表性的 AI 赋能产品和项目

公司	产品	AI赋能过程
微软		New Bing。整合ChatGPT-4模型进入Bing，赋予搜索引擎类ChatGPT式的体验。
		AI-Powered Office Suite.将Office Suite(word,excel等)加入ChatGPT对话工具，可以让用户根据提交的请求指令，生成对应文本。
		Azure Open AI Services.将Open AI模型整合到云服务领域，提供Azure OpenAI Service，更多的业务可以通过高级的AI模型(包括GPT 3.5，Codex，和 DALL·E 2来部署。
		GitHub Copilot.基于云端的AI编程工具，由Github和OpenAI合作。可以将自然语言的prompts转换成代码编写建议。
		AI-Powered Microsoft Viva.将GPT技术整合到Viva，帮助销售人员生成邮件内容，涵盖针对客户请求的定制化的销售数据等等，节约销售人员时间提升工作效率。
		AI-Powered Team Premium.将GPT3.5集成到现有的Team Premium中，将提供智能摘要，自动会议笔记，实时字幕翻译等功能。

资料来源：微软官网，华安证券研究所整理

图表 23 微软 AI 技术突破时间线



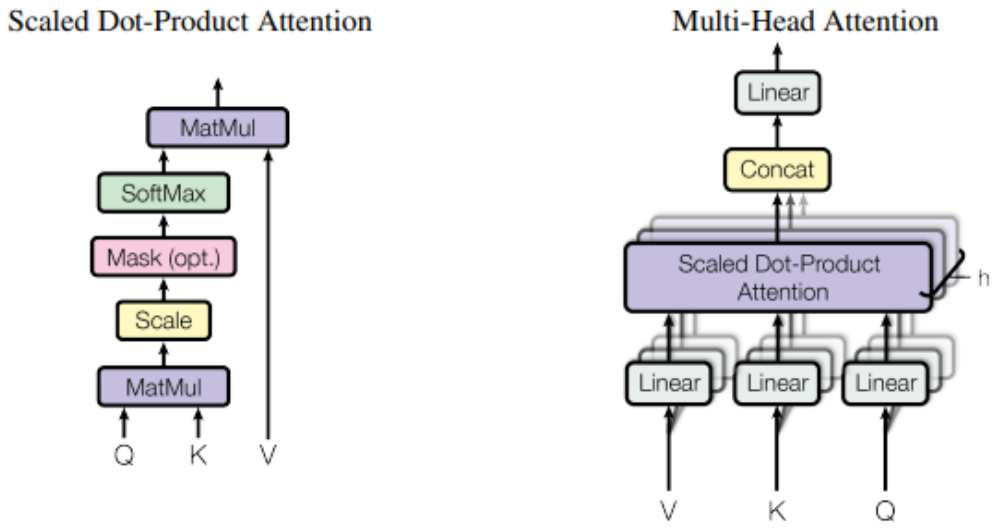
资料来源：微软官网，华安证券研究所

3.2 谷歌 AI 产品赋能回顾

谷歌于 2023 年 2 月 7 日发布了一项 AI 技术，命名为 Bard（巴德）。Bard 的直接竞品为 Open AI 旗下的语言模型 ChatGPT。谷歌巴德是谷歌的一项实验性的 AI 对话技术，功能和 ChatGPT 类似，具备拟人化的对话能力和一定的问题解决能力和内容创造能力。但谷歌发布巴德之后，并没有达到用户的期待。在巴德发布演示中，巴德对于“关于韦伯望远镜拍摄太阳系以外行星的问题”的回答产生了事实性的错误，从发布之后的用户评价来看，巴德的表现低于用户的期待。在发布之后不久，谷歌母公司 Alphabet 主席 John Hennessy 称，谷歌之前在犹豫是否要将 Bard 应用于产品中，因为 Bard 还没有真的准备好。

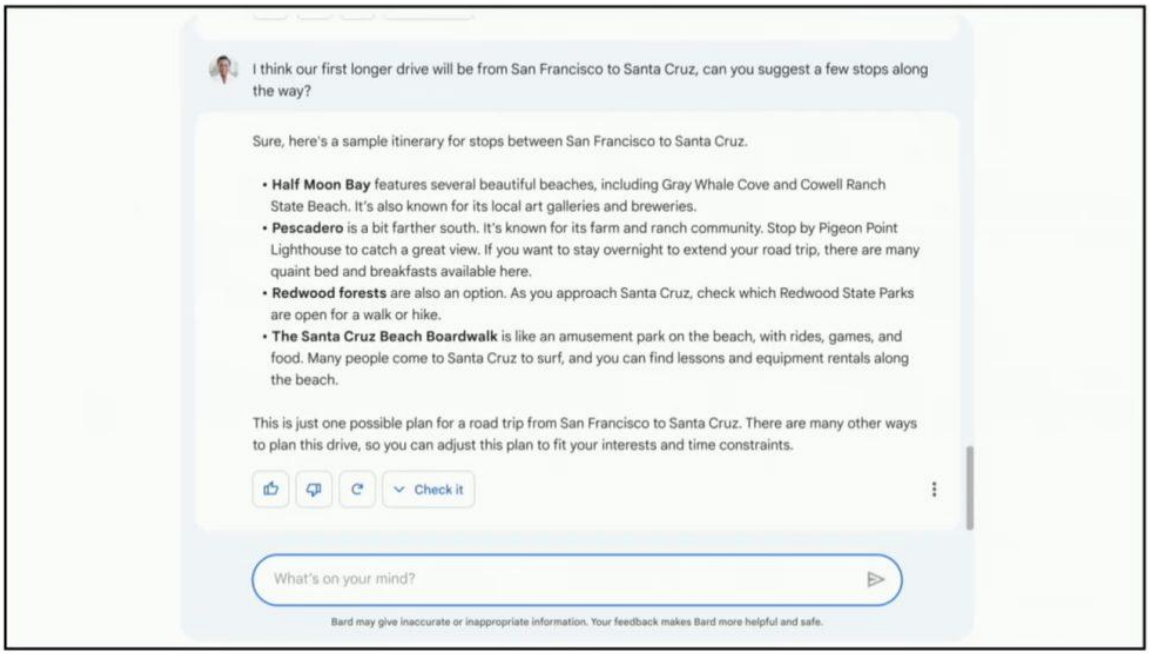
即使巴德是一项新发布的技术，但是其应用的核心模型-LaMDA(Launguage Model for Dialogue Applications，对话应用语言模型)是谷歌于 2 年之前发布。LaMDA 模型是基于 Transformer 模型发展出来；该模型于 2017 年由谷歌的研发人员在论文《Attention is all you need》中发布；ChatGPT 也是在 Transformer 模型的基础上发展起来的。Transformer 模型是基于注意力机制(attention mechanism)，它是一种模仿认知注意力的机制，可以对神经网络输入数据中不同部分给予不同的关注权重，将重点集中在最重要的小部分数据上。如果从技术起源的视角来看，谷歌于 2017 年发布 Transformer 模型这一事件，可以理解为这波“生成式 AI”技术浪潮的开始。

图表 24 Transformer 模型中引入的 Attention 机制示意图



资料来源：Google Research，华安证券研究所整理

图表 25 谷歌巴德(Google Bard)案例 – 规划旅行路线



资料来源：9 to 5 Google，华安证券研究所

谷歌近年来除了搜索引擎之外的一些具有代表性的产品或项目也在不断的促进 AI 相关的技术的落地，其中一部分应用了自然语言处理相关的技术，这些产品或项目包括 Cloud, Translate, Assistant, Lens, Imagen,Parti, MusicLM 等；这些技术的落地帮助用户满足语言翻译，智能手机扫描识别，通过文字生成图片和音乐等需求。

图表 26 谷歌具有代表性的 AI 赋能产品和项目

公司	产品	AI赋能过程
	谷歌翻译 Google Translate	采用GNMT(谷歌神经机器翻译系统），于2016年推出，支持100种以上语言的翻译。
	谷歌云 Google Cloud	为谷歌云的自然语言(NL)API更新了基于LLM的内容分类模型，帮助机器更好的理解人类语言。
	谷歌智慧镜头 Google Lens	应用计算机视觉，机器学习和谷歌knowledge graph，通过手机摄像头识别出现实世界的物品，标签和文字等。
	谷歌助手 Google Assistant	使用自然语言处理技术去理解人类的意图，解决用户提交的请求。
	Imagen模型	Imagen是谷歌的一个文本转图片的Diffusion模型，具备深度的语言理解能力。
	MusicLM模型	MusicLM是谷歌的文字生成音乐的模型。

资料来源：Google 官网，华安证券研究所整理

风险提示：

全球通胀和宏观经济下行，供应链扰动和贸易摩擦，导致 AI 计算领域上游硬件供需错配；生成式 AI 产品的商业化表现不及预期；生成式 AI 产品带来更多法律及道德风险

分析师与研究助理简介

分析师：金荣，香港中文大学经济学硕士，天津大学数学与应用数学学士，曾就职于申万宏源证券研究所及头部互联网公司，金融及产业复合背景，善于结合产业及投资视角进行卖方研究。2015 年水晶球第三名及 2017 年新财富第四名核心成员。执业证书编号：S0010521080002

重要声明

分析师声明

本报告署名分析师具有中国证券业协会授予的证券投资咨询执业资格，以勤勉的执业态度、专业审慎的研究方法，使用合法合规的信息，独立、客观地出具本报告，本报告所采用的数据和信息均来自市场公开信息，本人对这些信息的准确性或完整性不做任何保证，也不保证所包含的信息和建议不会发生任何变更。报告中的信息和意见仅供参考。本人过去不曾与、现在不与、未来也将不会因本报告中的具体推荐意见或观点而直接或间接收任何形式的补偿，分析结论不受任何第三方的授意或影响，特此声明。

免责声明

华安证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。本报告由华安证券股份有限公司在中华人民共和国（不包括香港、澳门、台湾）提供。本报告中的信息均来源于合规渠道，华安证券研究所力求准确、可靠，但对这些信息的准确性及完整性均不做任何保证。在任何情况下，本报告中的信息或表述的意见均不构成对任何人的投资建议。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。投资者务必注意，其据此做出的任何投资决策与本公司、本公司员工或者关联机构无关。华安证券及其所属关联机构可能会持有报告中提到的公司所发行的证券并进行交易，还可能为这些公司提供投资银行服务或其他服务。

本报告仅向特定客户传送，未经华安证券研究所书面授权，本研究报告的任何部分均不得以任何方式制作任何形式的拷贝、复印件或复制品，或再次分发给任何其他人，或以任何侵犯本公司版权的其他方式使用。如欲引用或转载本文内容，务必联络华安证券研究所并获得许可，并需注明出处为华安证券研究所，且不得对本文进行有悖原意的引用和删改。如未经本公司授权，私自转载或者转发本报告，所引起的一切后果及法律责任由私自转载或转发者承担。本公司并保留追究其法律责任的权利。

投资评级说明

以本报告发布之日起 6 个月内，证券（或行业指数）相对于同期相关证券市场代表性指数的涨跌幅作为基准，A 股以沪深 300 指数为基准；新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为基准；香港市场以恒生指数为基准；美国市场以纳斯达克指数或标普 500 指数为基准。定义如下：

行业评级体系

增持—未来 6 个月的投资收益率领先市场基准指数 5%以上；

中性—未来 6 个月的投资收益率与市场基准指数的变动幅度相差-5%至 5%；

减持—未来 6 个月的投资收益率落后市场基准指数 5%以上；

公司评级体系

买入—未来 6-12 个月的投资收益率领先市场基准指数 15%以上；

增持—未来 6-12 个月的投资收益率领先市场基准指数 5%至 15%；

中性—未来 6-12 个月的投资收益率与市场基准指数的变动幅度相差-5%至 5%；

减持—未来 6-12 个月的投资收益率落后市场基准指数 5%至 15%；

卖出—未来 6-12 个月的投资收益率落后市场基准指数 15%以上；

无评级—因无法获取必要的资料，或者公司面临无法预见结果的重大不确定性事件，或者其他原因，致使无法给出明确的投资评级。