

半导体

ChatGPT 有望给半导体行业带来较为显著增量

ChatGPT：突破性的对话系统预训练生成模型。

ChatGPT 是基于 transformer 架构的文本生成式 AI。ChatGPT 由 GPT 系列分化而来，从 GPT-1 到 ChatGPT，算法上均采用的是 transformer 架构，模型结构改进程度偏弱，主要为数据与算力的扩展。（1）AIGC 视角下，ChatGPT 在 AI 中属于可交互的文本 AIGC，其兴起源于深度学习技术的快速突破和日益增长的数字内容供给需求；（2）transformer 视角下，transformer 帮助 AI 文本生成的核心技术 NLP 走出了发展困境。随着基于 transformer 的模型越来越大，它们开始可以输出达到人类水平的结果甚至超人的结果。GPT 凭借着高算力与大数据成为了目前规模竞争下的胜者。

云端：半导体+AI 生态逐渐清晰

更大的算力意味着更多的计算机设备，搭建这些设备也需要更多的核心器件。企业对类 ChatGPT 技术的追求，在服务器产业链的发展上预计将起到积极的促进作用。（1）AI 算力芯片：AI 算力芯片是类 ChatGPT 模型的基石，支撑类 ChatGPT 模型需要大量的算力芯片，其中对 GPU、FPGA、ASIC 需求较大；（2）HBM/Chiplet：AI 芯片性能及成本的平衡也带动周边生态，HBM/Chiplet 等产业链受益。在 HBM 领域，AI 对话程序在执行计算期间需要大容量、高速的存储支持，预计 AI 芯片发展也将会进一步扩大高性能存储芯片需求；在 Chiplet 领域，Chiplet 是布局先进制程、加速算力升级的关键技术。

类 ChatGPT 对芯片的需求将量大且具有高持续性

按照我们的预设，当前的 ChatGPT 模型至少在服务器上花费了 3.47 亿美元，其中 CPU、GPU、DRAM 分别为 0.29、2.66、0.23 亿美元。未来随着 ChatGPT 的市占率及应用端的发展，我们预测它的日活量（DAU）与每人每天生成单词 2023 年后均会呈现阶梯式高速增长直至 2030 年后放缓，服务器成本也将随之继续扩张，预计 2030 年服务器成本高达 975.1 亿美元。

终端：“ChatGPT+” 加速数字经济发展

应用端相关行业通过 ChatGPT 找到了新的发力点，推出 ChatGPT 相关产品，加速数字化转型。在人工智能行业，ChatGPT 可以助力 AI 开发、优化模型优化及丰富应用场景等，其中与具体任务相结合的应用将会是中短期内可行性较高的热点。服务上，ChatGPT 能够帮助 APP/语音助手提供更好的聊天服务；产品上，ChatGPT 能够增强智能产品的交互性。此外，ChatGPT 还能够应用在搜索引擎、泛娱乐、自媒体等其他行业上。

投资建议：我们认为以 GPT-4/ChatGPT 为代表的预训练大模型或将催生未来对 AI 服务器的扩产需求，建议关注

服务器芯片：英伟达（天风海外组覆盖）、AMD、Intel、寒武纪、海光信息（天风计算机组覆盖）、龙芯中科等；

AI 服务器制造：工业富联；

图像数据资源及应用：大华股份、海康威视等；

先进制造及封装：台积电、中芯国际、长电科技、通富微电等；

企业级存储：澜起科技、江波龙（天风计算机组联合覆盖）、兆易创新等。

风险提示：ChatGPT 发展不及预期；国际局势震荡加重贸易成本；市场竞争加剧；本文对芯片市场空间测算是基于一定前提假设，存在假设条件不成立、市场发展不及预期等因素导致市场空间测算结果偏差。

证券研究报告

2023 年 04 月 02 日

投资评级

行业评级

强于大市(维持评级)

上次评级

强于大市

作者

潘暕

分析师

SAC 执业证书编号：S1110517070005
panjian@tfzq.com

骆奕扬

分析师

SAC 执业证书编号：S1110521050001
luoyiyang@tfzq.com

行业走势图



资料来源：聚源数据

相关报告

- 《半导体-行业研究周报:半导体库存去化或近尾声，待需求复苏重启增长》2023-03-28
- 《半导体-行业专题研究:ChatGPT 推动 AI 芯应用，算力提升终端多点开花》2023-03-23
- 《半导体-行业研究周报:2 月国产设备招标同比 +58.00%，重点关注 ChatGPT 及 Chiplet 领域机遇》2023-03-20

内容目录

1. 一图看懂产业链	4
1.1. 服务器成本规模测算	4
2. ChatGPT：突破性的对话系统预训练生成模型	5
2.1. ChatGPT 的基本介绍	5
2.2. 发展历程：从 GPT-1 到 ChatGPT	6
2.3. 地位：基于 NLP 的先进聊天机器人	8
2.3.1. AIGC 视角：	8
2.3.2. transformer 视角：	10
3. 云端：半导体+AI 生态逐渐清晰	13
3.1. AI 芯片：	13
3.2. HBM/Chiplet：	15
4. 终端：“ChatGPT+” 加速数字经济发展	15
4.1. 人工智能行业	15
4.2. 其他行业	17
5. 投资建议：	18
6. 风险提示：	18

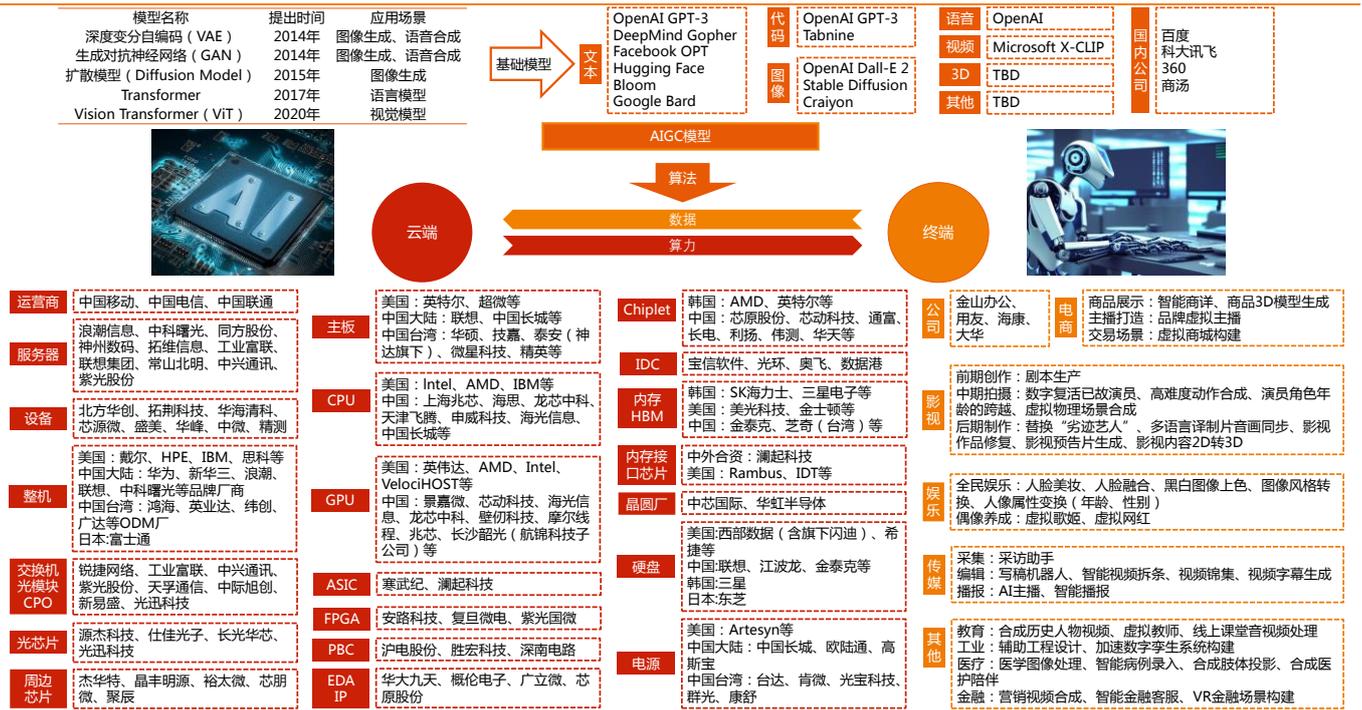
图表目录

图 1：AIGC 产业链	4
图 2：ChatGPT 服务器成本（百万美元）对 DAU、每人每天生成单词的敏感性分析	5
图 3：ChatGPT 的操作页面：当用户提出问题，ChatGPT 能够迅速抓取信息互动	5
图 4：ChatGPT 训练过程，背后的底层技术的特点是能够分析大量数据	6
图 5：相同参数下 few-shot 性能更佳	7
图 6：从 GPT-3 到当前阶段 GPT-3.5 的进化历程	8
图 7：在内部对抗性事实性评估中，GPT-4 的得分比最新的 GPT-3.5 高 40%	8
图 8：生成式 AI 的应用格局	9
图 9：AIGC 发展历程	10
图 10：自然语言、自然语言处理和编程语言的关系	11
图 11：Transformer 出现后，NLP 类模型迎来较大地增长	11

图 12: 基于 Transformer 的预训练模型对比.....	13
图 13: GPT-3 模型复杂度达到 1750 亿个参数, 使其竞争对手相形见绌.....	13
图 14: 在线学习平台 Quizlet 宣布将整合 ChatGPT API, 提供在线 AI 教师 “Q-Chat” ...	16
图 15: 智能音箱接入 ChatGPT 的架构.....	17
表 1: AI 的三大基石: 算法、数据和算力.....	6
表 2: transformer 经典模型介绍.....	12
表 3: 主要需求的 AI 芯片特点及市场格局.....	14
表 4: 英伟达数据中心芯片售价情况.....	14
表 5: ChatGPT 在人工智能行业的应用场景.....	16

1. 一图看懂产业链

图 1: AIGC 产业链



资料来源: 中国信息通信研究院和京东探索研究院, 全球半导体观察公众号, 红杉《Generative AI: A Creative New World》, 腾讯研究院, viso.ai, ATYUN 订阅号公众号, 站长之家官网, 天风证券研究所

1.1. 服务器成本规模测算

我们认为基于下述假设, GPT 目前在服务器上花费的成本至少为 3.47 亿美元。首先, 我们参考马里兰州 AI 方向的副教授 Tom Goldstein 及财经十一人公众号的估计, ①假设 1 个服务器由 2 个 CPU、8 个 GPU (A100, 80G 内存)、DRAM 与其它组件组成, 这个服务器可以使用 ChatGPT AI 模型每秒生成 15-20 个单词, 我们取 20 个单词/秒, 那么每台服务器每天可以生成 172.8 (=20×60×60×24/10000) 万个单词。根据 Similar Web 数据, 2023 年 1 月, ChatGPT 的网站平均每天有超过 1300 万用户使用, ②假设日活量 (DAU) 为 1000 万, 如果每个人对话过程中只使用 500 个单词, 那么每天 ChatGPT 至少需要生产 50 亿单词 (=1000×500 万, 这里需要注意的是, 1000 万人使用时间实际上并不是均匀分布的, 因此现实中 ChatGPT 每天可生成单词预计比 50 亿单词高)。基于以上分析, ChatGPT 至少需要 2894 个服务器, 即 5788 个 CPU、23152 个 GPU。参考行行查给出的服务器成本构成以及 Thinkmate 给出的 A100GPU 报价, ③CPU 单价取\$5000, GPU 单价取\$11500 (考虑公司战略合作及贸易成本, 取价比 Thinkmate 报价略低), DRAM 取\$8000, 加上其他组件, 一个服务器总价为\$120000。综合上述假设, 服务器总成本为 3.47 亿美元, CPU、GPU、DRAM 分别为 0.29、2.66、0.23 亿美元。

未来随着 ChatGPT 的市占率及应用端的发展, 我们预测它的日活量 (DAU) 与每人每天生成单词 2023 年后均会呈现阶梯式高速增长直至 2030 年后放缓, 服务器成本也将随之继续扩张。其中, DAU 表现了 ChatGPT 普适性, 当 ChatGPT 的普适性更高即 ChatGPT 的用途更加广泛 (往往取决于算法) 时, 将会有更多的用户使用它; 每人每天生成单词则能表现 ChatGPT 解决用户问题的能力 (取决于数据训练), 当 ChatGPT 能够更好地解决用户问题时, 平均每个用户需要生成的单词会更多。因此, 为了展现这两个指标对服务器成本的影响, 我们对其进行敏感性分析发现, ChatGPT 对两个指标的弹性指数均为 1, 也就是说当 2030 年 ChatGPT 实现 9.36 亿日活量 (前文假设的 93.6 倍)、每人每天 1500 个生成单词 (前文假设的 3 倍) 时, 服务器成本将变成前文的 281 倍 (≈93.6×3)——975.1 亿美元。

此外, 若考虑到芯片的损耗及芯片升级, 预计 ChatGPT 对芯片的需求将长时间维持在较高水平。

图 2: ChatGPT 服务器成本 (百万美元) 对 DAU、每人每天生成单词的敏感性分析

		DAU (百万) → ChatGPT 的普适性									
		5	10	40.6	81.4	163.2	246.9	415.0	585.9	675.2	936.1
每人每天生成单词 ↓ ChatGPT 的解决能力	200	69	139	564	1131	2267	3429	5764	8138	9378	13001
	300	104	208	846	1696	3400	5144	8646	12206	14067	19502
	400	139	278	1128	2261	4533	6858	11528	16275	18756	26003
	500	174	347	1410	2826	5667	8573	14410	20344	23444	32503
	600	208	417	1692	3392	6800	10288	17292	24413	28133	39004
	700	243	486	1974	3957	7933	12002	20174	28481	32822	45505
	800	278	556	2256	4522	9067	13717	23056	32550	37511	52006
	900	313	625	2538	5088	10200	15431	25938	36619	42200	58506
	1000	347	694	2819	5653	11333	17146	28819	40688	46889	65007
	1100	382	764	3101	6218	12467	18860	31701	44756	51578	71508
	1200	417	833	3383	6783	13600	20575	34583	48825	56267	78008
	1300	451	903	3665	7349	14733	22290	37465	52894	60956	84509
	1400	486	972	3947	7914	15867	24004	40347	56963	65644	91010
	1500	521	1042	4229	8479	17000	25719	43229	61031	70333	97510

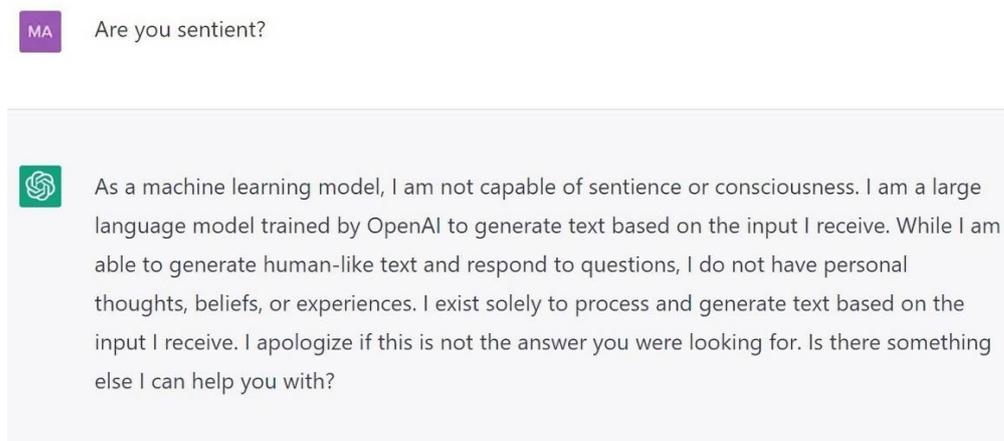
资料来源: International Telecommunication Union (ITU) World Telecommunication/ICT Indicators Database, world bank, statista, 天风证券研究所
 注: 横坐标 DAU 的变化参考 world bank 给出的 1995-2002 年互联网用户占总人口数比例以及 statista2023-2030 年间未来人口预测, 计算公式为:
 n 年的 DAU = (($n-28$) 年的互联网用户/总人口) × n 年预测总人口

2. ChatGPT: 突破性的对话系统预训练生成模型

2.1. ChatGPT 的基本介绍

ChatGPT 可以解释为一种用于对话系统的预训练生成式模型。ChatGPT 名称中包含两个元素: Chat 和 GPT。Chat 是指“聊天”, 代表了它在对话系统等领域的应用。GPT 全称是 Generative Pre-trained Transformer, 其本质上是一种通过在大型文本语料库上训练而形成的具有自然语言生成能力的 NLP 模型。电子工程专辑公众号指出, ChatGPT 和以往帮我们实现简单服务的客服机器人不一样, 它可以和用户进行多轮对话, 经过“预先训练+微调”后, OpenAI 通过监督学习不断训练, 得到合适的模型, 再通过设计好的奖励机制由人类培训员不断筛选, 最终得到质量最好的回复。

图 3: ChatGPT 的操作页面: 当用户提出问题, ChatGPT 能够迅速抓取信息互动



资料来源: OpenAI, THE DECODER, 天风证券研究所

作为 AI 的 ChatGPT, 其核心仍为算法、数据与算力。联想创投公众号指出, ChatGPT 正在经历巨大的变迁, 这是重新定义生产力颠覆性变化的机会。超级算力和大模型代表的超级算法, 正在成为国家和企业的竞争力。在数据智能时代, 特别需要 AI 的算法、算力、高质量的数据支撑。其中, 在算力上, ChatGPT 首先需要云计算基础设施作为算力底座;

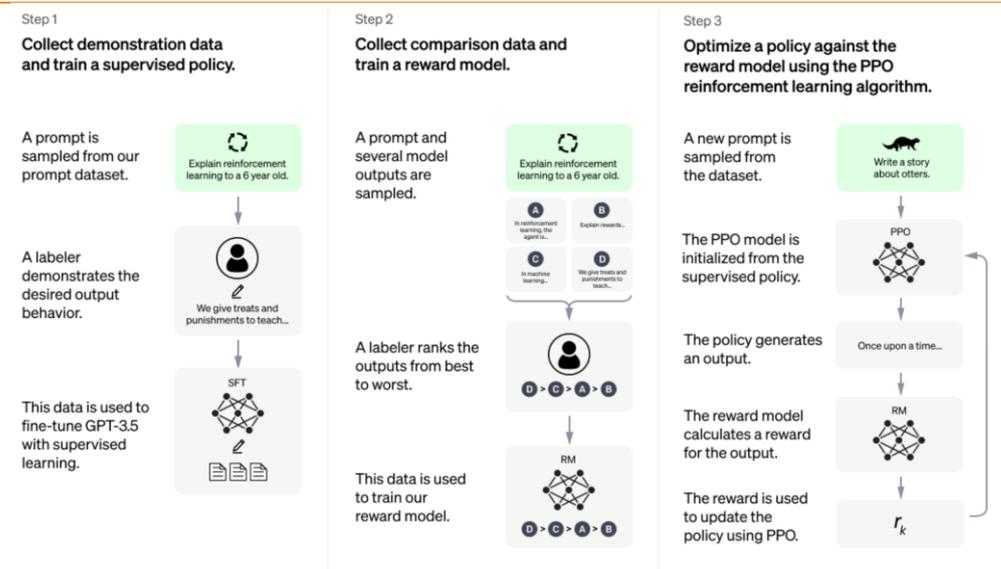
其次，云计算数据中心还需要包括服务器、芯片、光模块等硬件的支持，而 AI 芯片则是算力最重要的硬件基石之一。

表 1: AI 的三大基石：算法、数据和算力

名称	简介
算法	算法层指各类机器学习算法。如果根据训练方法来分类，机器学习算法也可以分成“无监督学习”、“监督学习”和“强化学习”等。按照解决问题的类型来分，机器学习算法包括计算机视觉算法（CV）、自然语言处理算法（NLP）、语音处理和识别算法（ASR）、智慧决策算法（DMS）等。
数据	数据层指的是人工智能为不同的行业提供解决方案时所采集和利用的数据。事实上，使用人工智能解决问题的步骤绝不仅仅包括搜集和整理数据。完整的流程和思路为：收集数据、数据准备、模型选择、训练、评估、参数调整、预测。
算力	算力层包括具备计算能力硬件和大数据基础设施。AI 发展历程中，历次算力层的发展都会显著推动算法层的进步，并促使技术的普及应用。21 世纪互联网大规模服务集群的出现、搜索和电商业带来的大数据积累、GPU 和异构/低功耗芯片兴起带来的运算力提升，促成了深度学习的诞生，促成了人工智能的这一波爆发。而 AI 芯片的出现进一步显著提高了数据处理速度：在 CPU 的基础上，出现了擅长并行计算的 GPU，以及拥有良好运行能效比、更适合深度学习模型的现场可编程门阵列（FPGA）和应用专用集成电路（ASIC）。

资料来源：CGGE 数字内容与技术服务平台公众号，天风证券研究所

图 4: ChatGPT 训练过程，背后的底层技术的特点是能够分析大量数据



资料来源：OpenAI, News 官网，天风证券研究所

2.2. 发展历程：从 GPT-1 到 ChatGPT

ChatGPT 是基于 GPT 系列演化而来的。ChatGPT 由 OpenAI 公司开发，OpenAI 是一家总部位于美国的人工智能研究机构，旨在不受产生财务回报需求的约束，以最有可能造福全人类的方式推进数字智能。OpenAI 开发了 GPT 模型的四代产品——GPT-1、GPT-2、GPT-3 和 GPT-4，分别发布于 2018 年、2019 年、2020 年、2023 年。

GPT 系列从 1 到 4，算法上均采用的是 transformer 架构（详见 2.3.2），模型结构改进程度偏弱，主要为数据与算力的扩展。

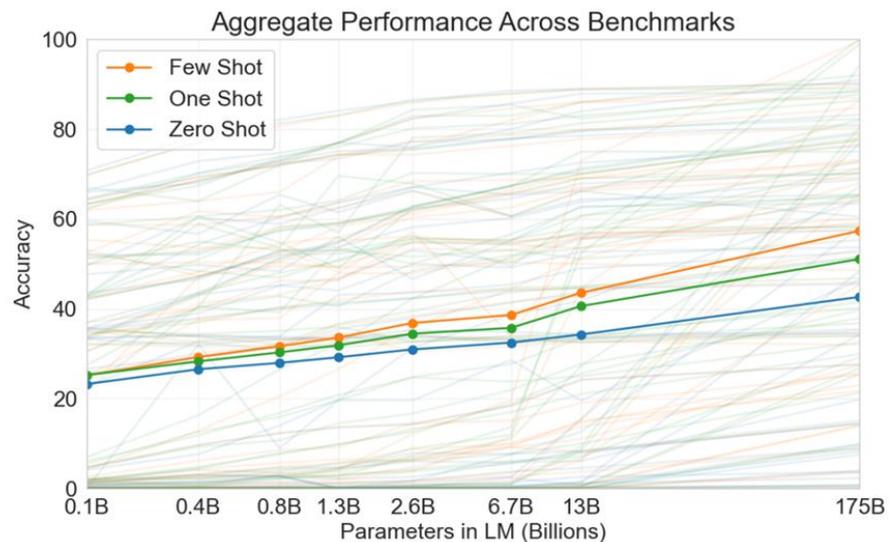
维普研究院公众号指出，GPT-1 主要有三点贡献：第一，它是最早一批提出在 NLP 任务上使用 pre-train + fine-tuning 范式的工作；第二，GPT 的实验证明了模型的精度和泛化能力会随着解码器层数增加而不断提升，而且目前还有提升空间；第三，预训练模型具有 zero-shot 的能力，并且能随着预训练的进行不断增强。

GPT-2 想通过 zero-shot，在迁移到其他任务上的时候不需要额外的标注数据，也不需要额外的模型训练。GPT-2 的核心思想是当模型的容量非常大且数据量足够丰富时，仅仅靠语言模型的学习便可以完成其他有监督学习的任务，不需要在下游任务微调。GPT-2

在较多任务上对比无监督算法取得了一定的提升，证明了 zero-shot 的能力，但在很多任务上与有监督微调的方法相比还有一些差距。

GPT-3 不再去追求极致的不需要任何样本就可以表现很好的模型，而是考虑像人类的学习方式那样，仅仅使用极少数样本就可以掌握某一个任务，因此就引出了 GPT-3 标题 Language Models are Few-Shot Learners。可以从 GPT-3 的分析结果得出一个重要的结论：当我们想要线性的提升一个任务的效果时，往往需要指数级的提升模型的规模和所需的数据量。

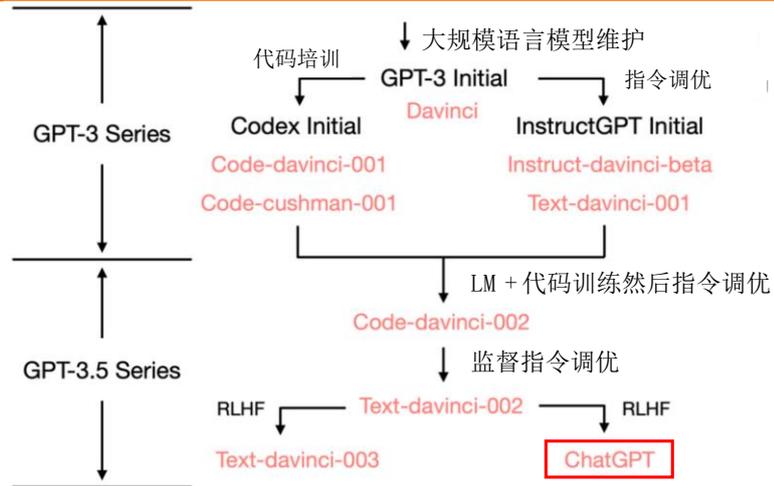
图 5：相同参数下 few-shot 性能更佳



资料来源：《Language Models are Few-Shot Learners》(Tom B. Brown, Benjamin Mann 等)，天风证券研究所

ChatGPT 可以看做是 GPT-3.5 版本之一，通过牺牲上下文学习的能力换取建模对话历史的能力。《How does GPT Obtain its Ability? Tracing Emergent Abilities of Language Models to their Sources》(Yao Fu, Hao Peng 等)中指出，code-davinci-002 和 text-davinci-002 是第一版的 GPT3.5 模型，一个用于代码，另一个用于文本。它们表现出了三种与初代 GPT-3 不同的重要能力：响应人类指令、泛化到没有见过的任务、利用思维链进行复杂推理。文章推测这些新的能力来源于指令微调和代码训练，具体来说：能够响应人类指令的能力是指令微调的直接产物；对没有见过的指令做出反馈的泛化能力是在指令数量超过一定程度之后自动出现的；使用思维链进行复杂推理的能力很可能是代码训练的副产物。指令微调不会为模型注入新的能力，其作用是解锁/激发这些能力，通过牺牲性能换取与人类的对齐。在 code-davinci-002 上进行指令微调后，模型可以生成更加符合人类期待的反馈——指令微调将 GPT-3.5 分化到不同的技能树，有些更擅长上下文学习，如 text-davinci-003，有些更擅长对话，如 ChatGPT。

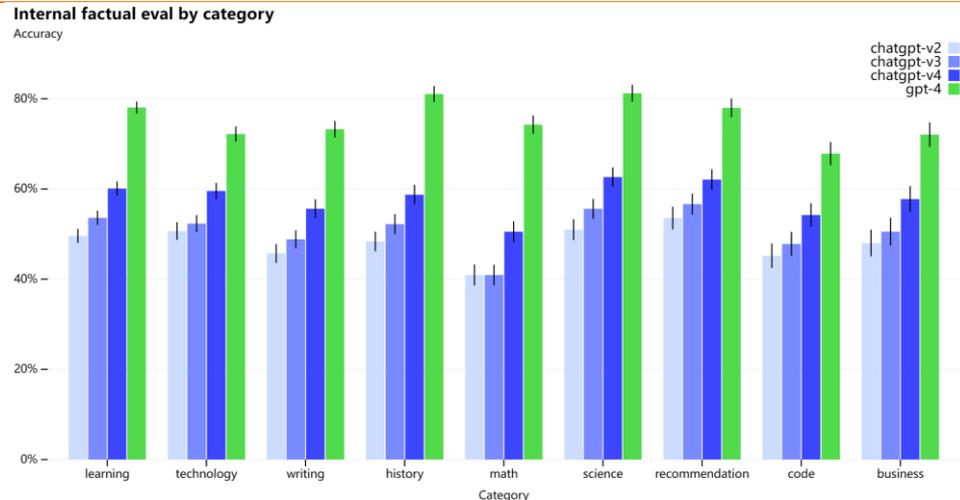
图 6：从 GPT-3 到当前阶段 GPT-3.5 的进化历程



资料来源：《How does GPT Obtain its Ability? Tracing Emergent Abilities of Language Models to their Sources》(Yao Fu, Hao Peng 等), 天风证券研究所

GPT-4 针对 GPT-3.5 在测试中出现的问题进行了改进。OpenAI 发布 GPT-4 时声明，其将 GPT-3.5 作为 GPT-4 的第一次“测试运行”进行了训练，发现并修复了一些错误，此外还提高了 GPT 的理论基础。团队花了 6 个月的时间迭代 GPT-4，使用 OpenAI 的对抗性测试程序和 ChatGPT 的经验教训，在事实性、可操纵性和拒绝超出合理范围（refusing to go outside of guardrails）方面取得了有史以来最好的结果。在随意的谈话中，GPT-3.5 和 GPT-4 之间的区别可能很微妙。当任务的复杂性达到足够的阈值时，差异就出现了一—GPT-4 比 GPT-3.5 更可靠、更有创意，并且能够处理更细微的指令。

图 7：在内部对抗性事实性评估中，GPT-4 的得分比最新的 GPT-3.5 高 40%



On nine categories of internal adversarially-designed factual evals, we compare GPT-4 (green) to the first three ChatGPT versions. There are significant gains across all topics. An accuracy of 1.0 means the model's answers are judged to be in agreement with human ideal responses for all questions in the eval.

资料来源：OpenAI, 天风证券研究所

2.3. 地位：基于 NLP 的先进聊天机器人

2.3.1. AIGC 视角：

ChatGPT 在 AI 中属于可交互的文本 AIGC。根据红杉美国官网发表的文章《Generative AI: A Creative New World》，AI 可以被分成两类，一类是分析型 AI（Analytical AI），也可称为传统 AI，它可以用来分析一组数据，并在其中找到许多用例（use case）的模式，由此可以用于欺诈或垃圾邮件检测、预测发货时间或预测该给用户推荐哪个短视频，它们会在这些任务中变得越来越聪明；另一类则是生成式 AI（Generative AI, AIGC），这类机器不再局限于分析已经存在的东西，它会生成新的东西，例如写诗、设计产品、制作游戏、编写代码等。而根据生成式 AI 的具体生成内容，它又可以分为文本（Text）、代码生成（Code generation）、图片（Images）、语音合成（Speech synthesis）、视频和 3D 模型（Video and 3D）。

我们认为，ChatGPT 在与用户互动的过程中其实也就是在基于用户不断更新的对话内容调整认知，文本生成更加符合用户预期的回答。

图 8：生成式 AI 的应用格局

APPLICATION LAYER	Marketing (content)						
	Sales (email)	Code generation	Image generation				Gaming
Support (chat / email)		Code documentation	Consumer / Social				RPA
General writing		Text to SQL	Media / Advertising				Music
Note taking		Web app builders	Design	Voice Synthesis	Video editing / generation	3D models / scenes	Audio
Other							Biology & chemistry
	TEXT	CODE	IMAGE	SPEECH	VIDEO	3D	OTHER
MODEL LAYER	OpenAI GPT-3	OpenAI GPT-3	OpenAI Dall-E 2	OpenAI	Microsoft X-CLIP	TBD	TBD
	DeepMind Gopher	Tabnine	Stable Diffusion				
	Facebook OPT		Crayon				
	Hugging Face Bloom						
	Cohere						
	Anthropic						
	AI2						
	Alibaba, Yandex, etc.						

资料来源：红杉《Generative AI: A Creative New World》，PANews，天风证券研究所

根据中国信息通信研究院和京东探索研究院，AIGC 的发展大致可以分为三个阶段——**早期萌芽阶段**、**沉淀积累阶段**以及**快速发展阶段**：**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。80 年代末至 90 年代中，由于高昂的系统成本无法带来可观的商业变现，各国政府纷纷减少了在人工智能领域的投入，AIGC 没有取得重大突破；**沉淀积累阶段**（1990s-2010s），AIGC 从实验性向实用性逐渐转变。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。但是 AIGC 依然受限于算法瓶颈，无法较好地完成创作任务应用仍然有限，效果有待提升；**快速发展阶段**（2010s-至今），自 2014 年起，随着以生成式对抗网络（Generative Adversarial Network, 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，生成内容百花齐放，效果逐渐逼真直至人类难以分辨。

图 9: AIGC 发展历程



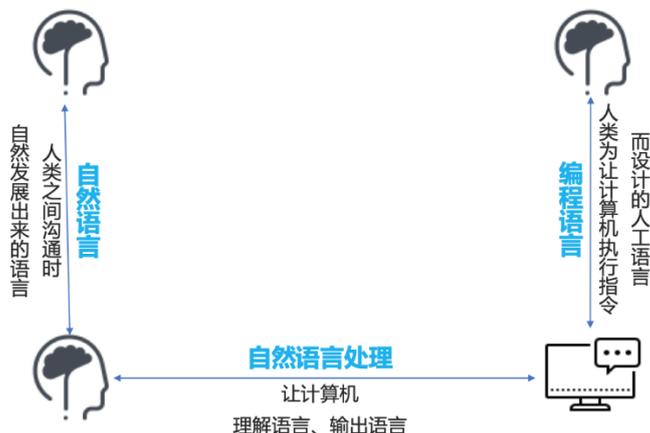
资料来源：中国信息通信研究院和京东探索研究院，天风证券研究所

AIGC 的兴起源于深度学习技术的快速突破和日益增长的数字内容供给需求。中国信息通信研究院和京东探索研究院指出，一方面，技术进步驱动 AIGC 可用性不断增强：在人工智能发展初期，虽然对 AIGC 进行了一些初步尝试，但受限各种因素，相关算法多基于预先定义的规则或者模板还远远算不上是智能创作内容的程度。近年来，基于深度学习算法的 AIGC 技术快速迭代，彻底打破了原先模板化、公式化、小范围的局限，可以快速、灵活地生成不同模态的数据内容。另一方面，海量需求牵引 AIGC 应用落地：随着数字经济与实体经济融合程度不断加深以及 Meta、微软、字节跳动等平台型巨头的数字化场景向元宇宙转型，人类对数字内容总量和丰富程度的整体需求不断提高。数字内容的生产取决于想象能力、制造能力和知识水平：传统内容生产手段受限于人力有限的制造能力，逐渐无法满足消费者对于数字内容的消费需求，供给侧产能瓶颈日益凸显。基于以上原因，AIGC 在各行业中得到越来越广泛的应用，市场潜力逐渐显现。

2.3.2. transformer 视角：

自然语言处理技术 (NLP) 是 AI 文本生成的核心技术之一。根据人民中科公众号，自然语言处理是指利用人类交流所使用的自然语言与机器进行交互通讯的技术。通过人为的对自然语言的处理，使得计算机对其能够可读并理解。NLP 技术用于计算机中模拟人类的对话和文本理解。主要源于 AI 大模型化的 NLP 技术突破是将深度学习技术与传统的 NLP 方法结合在一起，从而更好地提高 NLP 技术的准确性和效率。大模型化的 NLP 技术能够更好地支持企业进行大规模的语料内容分析，并为企业更好地进行文本分析提供帮助。从技术角度讲，ChatGPT 是基于大规模预训练语言模型 (GPT-3.5)，借助其强大的语言理解和生成能力，通过在人工标注和反馈的大规模数据上进行学习，从而让预训练语言模型能够更好地理解人类的问题并给出更好的回复。

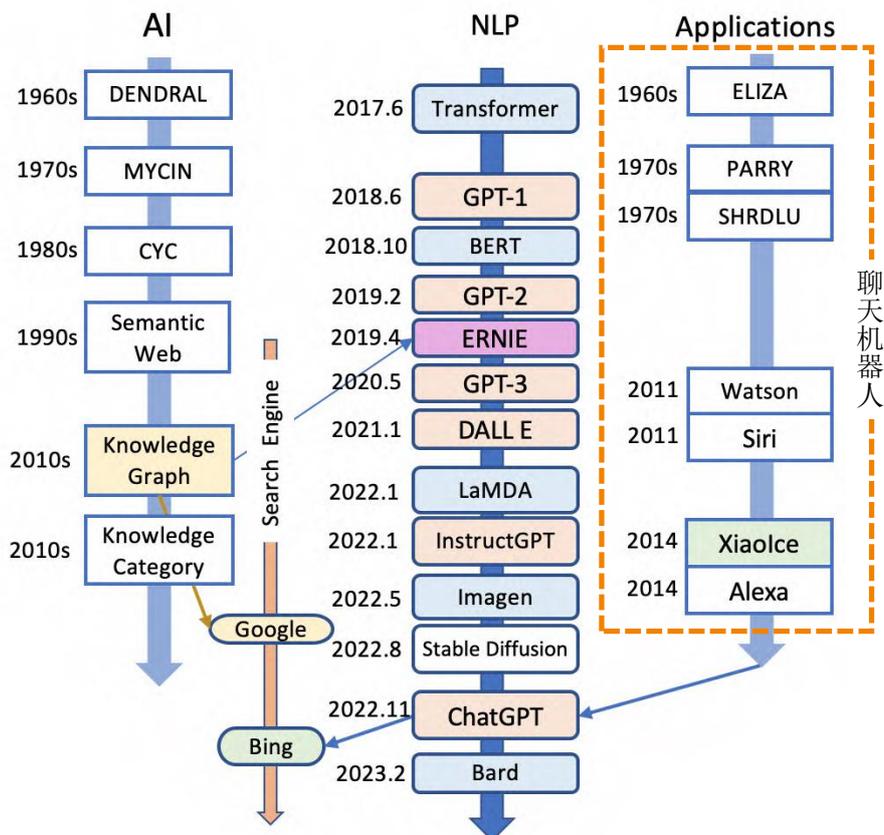
图 10：自然语言、自然语言处理和编程语言的关系



资料来源：腾讯云开发者公众号，天风证券研究所

Transformer 帮助 NLP 走出了发展困境。腾讯云开发者公众号指出，Transformer 结构继 RNN、CNN（以及其一系列变体 LSTM、GRU、ResNet、DenseNet 等）之后，在 Inductive Bias 方向上打开了一个新世界的大门。2017 年，Google 发表了一篇题为《Attention is all you need》的论文，自此完全使用 Attention 机制的 Transformer 模型开始主导 NLP 领域。不久后，利用 Transformer 模型结构的预训练语言模型 GPT 和 BERT 大放异彩。时至今日，基于 GPT 和 BERT 的预训练语言模型进行 Fine-tuning 仍然是 NLP 中广泛任务的首选范式。

图 11：Transformer 出现后，NLP 类模型迎来较大地增长



资料来源：《What Does ChatGPT Say: The DAO from Algorithmic Intelligence to Linguistic Intelligence》(Fei-Yue Wang, Qinghai Miao 等)，天风证券研究所

详细来说，Transformer 是一种基于自注意力机制的深度学习神经网络模型，可以高效并行地处理序列数据。原始的 Transformer 模型包含两个关键组件：编码器和解码器。编码器用于将输入序列映射到一组中间表示，解码器则将中间表示转换为目标序列。编码器和解码器都由多层的注意力模块和前馈神经网络模块组成。在原始 Transformer 模型基础上，相继衍生出了三类预训练语言模型：编码预训练语言模型、解码预训练语言模型和编解

码预训练语言模型，它们的代表模型如下表所示。

表 2: transformer 经典模型介绍

预训练语言模型	具体含义	经典模型	特点
编码预训练语言模型 (Encoder-only Pre-trained Models)	这类模型在预训练过程中只利用原始 Transformer 模型中的编码器。相应的预训练任务通常选用掩码语言建模任务 (Masked Language Modeling)，即掩码住 (用特殊字符 [MASK] 替换) 输入句子中一定比例的单词后，要求模型根据上下文信息去预测被遮掩的单词。	BERT	BERT 模型是最经典的编码预训练语言模型，其通过掩码语言建模和下一句预测任务，对 Transformer 模型的参数进行预训练。
		ALBERT	ALBERT 是一个轻量化的 BERT 模型，作者通过分解词向量矩阵和共享 Transformer 层参数来减少模型参数个数。
		BoBERTa	相较于 BERT 模型 RoBERTa 在预训练阶段，采用了更多的语料以及动态掩码机制 (不同轮次同一样本掩码不同的单词)，去掉了下一句预测任务，同时采用了更大的批大小。
解码预训练语言模型 (Decoder-only Pre-trained Models)	GPT (Generative Pre-trained Transformer) 是由 OpenAI 提出的只有解码器的预训练模型。相较于之前的模型，不再需要对于每个任务采取不同的模型架构，而是用一个取得了优异泛化能力的模型，去针对性地对下游任务进行微调。	GPT-1	GPT-1 成功背后有两个原因：第一个是 2017 年 Transformer 的提出使得捕获自然语言中远距离依赖关系成为可能；第二个是 GPT 模型在预训练过程中用到了更大的数据量以及更多的模型参数，使得模型能够从大规模语料库中学习以往模型无法学习的知识。而任务微调在通用预训练和下游任务之间搭起了知识桥梁，使得用一个模型解决多种问题成为一条可行之路。
		GPT-2	GPT-2 并没有模型架构上的改变，但是其将任务作为输出预测的条件引入模型从而在 Zero-shot 的设置下实现多个任务的想法一直延续至今。
		GPT-3	使用了与 GPT-2 相同的模型和架构，特点：一方面是模型本身规模大，参数量众多；另一方面是训练过程中使用到的数据集规模大，达到了 45TB。
		GPT-4	使用 OpenAI 的对抗性测试程序和 ChatGPT 的经验教训迭代 GPT-4，在事实性、可操纵性和拒绝超出合理范围 (refusing to go outside of guardrails) 方面取得了有史以来最好的结果。
基于编解码架构的预训练语言模型 (Encoder-decoder Pre-trained Models)	采用序列到序列的架构来融合两种结构，使用编码器提取出输入中有用的表示，来辅助并约束解码器的生成。	BART	具体结构为一个双向的编码器拼接一个单向的自回归解码器，采用的预训练方式为输入含有各种噪声的文本，再由模型进行去噪重构。在解码器部分，BART 每一层对编码器的最后一层的隐藏表示执行交叉注意力机制以聚合关键信息。BART 在维基百科和 BookCorpus 数据集上训练，数据量达 160GB。
		T5	通过在输入之前加入提示词，实现了用单个模型解决机器翻译、文本摘要、问答和分类等多个任务。针对迁移学习需要的巨量、高质量和多样的预训练数据，T5 在谷歌专门构造的 C4 数据集上进行训练。
		Switch Transformers	将混合专家网络 (Mixture-of-Experts, MoE) 的条件运算思想引入 Transformer 的全连接层，实现增加模型的尺寸而不增加推理时的运算量。

资料来源：BME 康复工程分会公众号，哈尔滨工业大学自然语言处理研究所《ChatGPT 调研报告》，OpenAI 官网，天风证券研究所

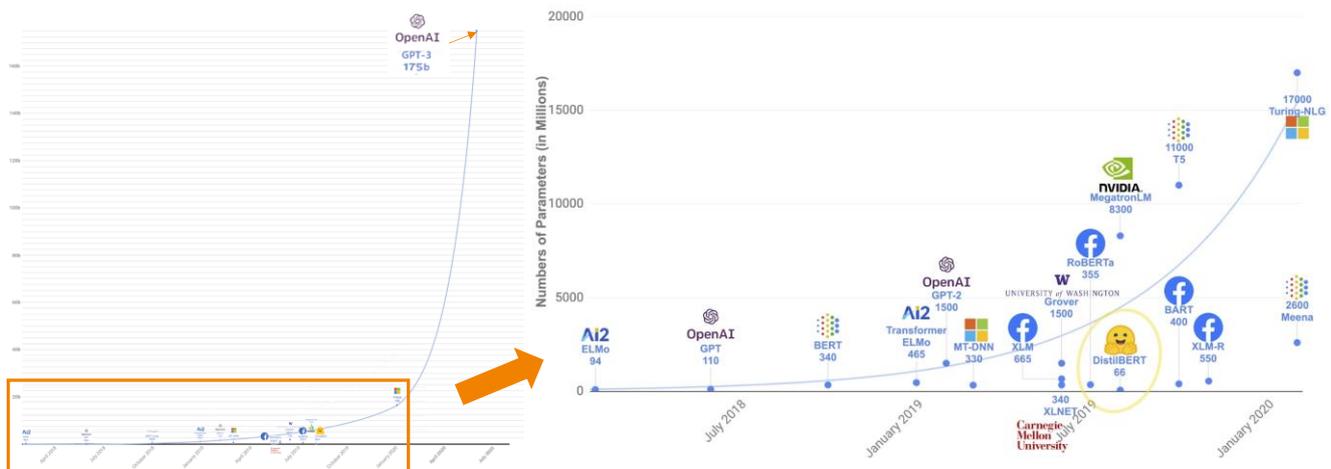
图 12：基于 Transformer 的预训练模型对比

模型	架构	参数量	数据集	机构
BERT	Enc	Base = 110M, Large = 340M	Wikipedia, BookCorpus	Google
ALBERT	Enc	Base = 12M, Large = 18M, XLarge = 60M	Wikipedia, BookCorpus	Google
RoBERTa	Enc	356M	Wikipedia, BookCorpus	Meta/华盛顿大学
GPT-1	Dec	117M	BookCorpus	OpenAI
GPT-2	Dec	1542M	WebText	OpenAI
GPT-3	Dec	175B	Common Crawl, WebText2, Books1, Books2 and Wikipedia	OpenAI
BART	Enc-Dec	400M	English Wikipedia, BookCorpus	Meta
T5	Enc-Dec	11B	C4	Google
Switch Transformers	Enc-Dec	1.6T	C4	Google

资料来源：BME 康复工程分会公众号，哈尔滨工业大学自然语言处理研究所《ChatGPT 调研报告》，天风证券研究所

GPT 是目前规模竞赛下的胜者。随着基于 transformer 的模型越来越大，它们开始可以输出达到人类水平的结果甚至超人的结果。从 2015 年到 2020 年，用于训练这些模型的计算量增加了 6 个数量级，其结果在书写、语音、图像识别、阅读和语言理解方面超过了人类的表现水平，我们将通过大数据、大算力来提升性能的模型竞争阶段称之为规模竞赛。我们认为，尽管之前 BERT 以优越的文本理解能力被更多人选择，但在大规模数据、算力的加持下目前 OpenAI 的 GPT 表现更为突出。此外，ChatGPT 则通过对话初步改变了 transformer 模型过于庞大而不能被更多人广泛使用这一特点。而对于之后的发展，我们预测，在规模竞赛模式下，可能会形成 GPT 为主，BERT 为辅的文本生成竞争格局。

图 13：GPT-3 模型复杂度达到 1750 亿个参数，使其竞争对手相形见绌



资料来源：《DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter》(Victor SANH, Lysandre DEBUT 等)，AI Multiple 官网，天风证券研究所

3. 云端：半导体+AI 生态逐渐清晰

更大的算力意味着更多的计算机设备，搭建这些设备也需要更多的核心器件。企业对类 ChatGPT 技术的追求，在服务器产业链的发展上预计将起到积极的促进作用。

3.1. AI 芯片：

AI 算力芯片是类 ChatGPT 模型的基石，其中对 GPU、FPGA、ASIC 需求较大。根据全球半导体观察公众号，基于 OpenAI 的第三代大模型 GPT-3 升级而来的 ChatGPT，其算力的最终来源也就是芯片，ChatGPT 的爆火代表着 AI 芯片技术的新一轮突破。公开资料显示，AI 算力芯片泛指加速 AI 类应用，主要分为 GPU、FPGA、ASIC。目前，ChatGPT 背后的计算集群使用的是英伟达的 AI 芯片。OpenAI 曾表示，ChatGPT 是与英伟达和微软合作完成的超级 AI。

表 3：主要需求的 AI 芯片特点及市场格局

类型	特点	市场格局
GPU	由于 CPU 的算力很有限，且处理并行操作比较吃力，因此一般将 CPU 搭配加速芯片使用。在 AI 时代的云端训练芯片中，GPU 占据较大的份额，被视为 AI 时代的算力核心。	英伟达在此番占据先发优势，但市场上还有很多企业在迎头赶上，如谷歌的张量处理器 TPU、百度的昆仑系列、华为海思的昇腾系列、阿里巴巴的含光 800 等。
FPGA	FPGA (Field Programmable Gate Array) 又称现场可编程门阵列，是指一种通过软件手段更改、配置器件内部连接结构和逻辑单元，完成既定设计功能的数字集成电路。 FPGA 芯片在实时性（数据信号处理速度快）、灵活性等方面优势明显，还可以编程、并行计算，在深度学习领域占据不可替代地位。 相比于 CPU/GPU/ASIC，FPGA 具有更高的速度和极低的计算能耗，常被用作专用芯片的小批量替代品。在 AI 模型构建时，FPGA 要实现深度学习功能，需要与 CPU 结合，共同应用于深度学习模型，同样可以实现庞大的算力需求。	全球 FPGA 芯片市场中，赛灵思和英特尔两家企业占据了市场大部分的份额，由于 FPGA 芯片拥有较高的技术和资金壁垒，我国企业在该领域上差距较大。近年来，我国领先企业也在 FPGA 芯片上实现了部分技术突破。如去年 8 月，京微齐力发布首颗国产 22nm 的 FPGA 芯片并成功量产。
ASIC	ASIC(Application Specific Integrated Circuit)，即专用集成电路，其计算能力和计算效率可根据用户特定需求进行定制，广泛应用于人工智能设备、虚拟货币挖矿设备、耗材打印设备、军事国防设备等智慧终端。 ASIC 芯片可根据终端功能不同分为 TPU 芯片、DPU 芯片和 NPU 芯片等。其中，TPU (Tensor Processing Unit) 为张量处理器，专用于机器学习。DPU (Data Processing Unit)，可为数据中心等计算场景提供引擎。NPU (Neural-network Processing Unit) 是神经网络处理器，在电路层模拟人类神经元和突触，并用深度学习指令集直接处理大规模电子神经元和突触数据。 相比于 GPU 和 FPGA，ASIC 缺乏灵活性，特别是在 AI、服务器这类领域，在各种算法不断迭代的情况下，ASIC 芯片的特性反而成为了它的累赘。但地平线 CEO 余凯曾公开表示，一旦软件算法固定下来，专用集成电路 ASIC 一定是未来的方向，按每瓦功耗计算能力看，ASIC 可比 GPU 提升 30-50 倍，这也将是未来行业的竞争焦点。	目前，国外谷歌、英特尔、英伟达等科技巨头相继发布了 TPU、DPU 等 ASIC 芯片，国内大厂也开始瞄准这一市场迅速发力，比如寒武纪就推出了一系列 ASIC 加速芯片，华为也设计了昇腾 310 和昇腾 910 系列 ASIC 芯片。

资料来源：全球半导体观察公众号，天风证券研究所

支撑类 ChatGPT 模型需要大量的芯片。以 GPU 为例，根据《财经》介绍，ChatGPT 包含近 1800 亿个参数。更早期的 2022 年 11 月，英伟达在官网公告中提到，微软 Azure 上部署了数万枚 A100/H100 高性能芯片。这是第一个采用英伟达高端 GPU 构建的大规模 AI 算力集群。微软的 Azure 云服务为 ChatGPT 提供了一台由超过 1 万枚英伟达 A100 GPU 芯片的 AI 计算集群，为其提供超级 AI 算力支持。在国内，云计算技术人士公认的一个说法是，1 万枚英伟达 A100 芯片是做好 AI 大模型的算力门槛。

表 4：英伟达数据中心芯片售价情况

	芯片型号	售价（美元）	售价（元）	中国供应情况
中低端	A10	3200	22080	
	A16	3500	24150	
	A30	4700	32430	
	A40	5300	36570	
	L40	7600	52440	
高性能	V100	10000	69000	
	A800	12000	82800	缺货
	A100	15000	103500	美国政府禁止供应中国
	H100	36500	251850	

资料来源：Thinkmate，《财经》，凤凰网，天风证券研究所

未来的 AI 芯片：通用与定制的平衡。根据中国电子报公众号，在 AI 芯片的发展过程中，通用性和定制化像是两个互相拉扯的作用力，衍生出一系列的芯片架构。其中，GPU 和 ASIC 分别是通用性和定制化的代表架构，也有着各自的优势和局限。GPU 架构提供了大量数据并行结构，因此能够提供大量的 AI 并行计算，适用于 AI 训练等。ASIC 等定制化 AI 芯片针对特定的应用场景进行芯片优化，能够取得更高的计算能效，但通用性相对较弱。随着芯片技术的发展，通用性与定制化已不再壁垒森严。一方面，英伟达在 GPU 架

构中引入了 Tensor Core（张量计算核心），一种专门针对深度学习应用而设计的专用 ASIC 单元，使 GPU 更加适合深度学习。与此同时，定制化芯片也逐步增加了通用计算单元，并引入可编程或部分可编程的架构，增强芯片的场景覆盖能力。我们认为，短期内具有大算力、通用性的 GPU 芯片或将成为大算力应用的首选。随着 GPU 的功耗过高等弊端的显现，类 GPU 架构的定制化大算力 AI 芯片（ASIC）也或将存在市场，满足对于芯片计算能效的提升需求。未来 GPU 与 ASIC 的界限可能会在较大程度上模糊，两者产生替代竞争。

3.2. HBM/Chiplet:

AI 芯片性能及成本的平衡也带动周边生态，HBM/Chiplet 等产业链受益。

全球半导体观察公众号指出，AI 对话程序在执行计算期间需要大容量、高速的存储支持，预计 AI 芯片发展也将会进一步扩大高性能存储芯片（HBM）需求。三星电子表示，为 GPU 和人工智能加速器提供数据的高性能高带宽内存（HBM）的需求将会扩大。从长远来看，随着 AI 聊天机器人服务的扩展，对于 CPU 的 128GB 或更大容量的高性能 HBM 和高容量服务器 DRAM 的需求预计会增加。近日，韩媒报道 2023 年开年后三星、SK 海力士两家存储大厂 HBM 订单快速增加，价格也水涨船高，据悉近期 HBM3 规格 DRAM 价格上涨 5 倍。

此外，Chiplet 技术也不可忽视，其是布局先进制程、加速算力升级的关键技术。Chiplet 异构技术不仅可以突破 GPU 等算力芯片先进制程的封锁，并且可以大幅提升大型芯片的良率、降低设计的复杂程度和设计成本、降低芯片制造成本。目前，Chiplet 已广泛应用于服务器芯片。AMD 是 Chiplet 服务器芯片的引领者，其基于 Chiplet 的第一代 AMD EPYC 处理器中，装载 8 个“Zen”CPU 核，2 个 DDR4 内存通道和 32 个 PCIe 通道。2022 年 AMD 正式发布第四代 EPYC 处理器，拥有高达 96 颗 5nm 的 Zen4 核心，并使用新一代的 Chiplet 工艺，结合 5nm 和 6nm 工艺来降低成本。英特尔第 14 代酷睿 Meteor Lake 首次采用 intel 4 工艺、引入 Chiplet 小芯片设计，预计将于 2023 年下半年推出，至少性能功耗比的目标要达到 13 代 Raptor Lake 的 1.5 倍水平。

4. 终端：“ChatGPT+”加速数字经济发展

应用端相关行业通过 ChatGPT 找到了新的发力点，推出 ChatGPT 相关产品，加速数字化转型。数字经济是人类通过大数据（数字化的知识与信息）的识别——选择——过滤——存储——使用，引导、实现资源的快速优化配置与再生、实现经济高质量发展的经济形态。广义上数字经济包含两个层面，第一是数字产业化，第二是产业数字化。在技术层面上，数字经济离不开大数据、云计算、区块链、人工智能等新兴技术的支持。在 +AI 阶段，企业从智能设备等消费领域获得的标签训练数据。使用互联网规模应用的消费者生成了足够的训练数据，可用于推荐引擎、语音和视觉应用。而在 AI+（包括 ChatGPT+）时代，利用基础模型，可以使用未标注的数据进行训练，从而使企业有机会通过未标注的数据来获取业务价值。比如，工业 4.0 可以使用在正常操作期间生成的传感器数据来优化设备维护，IT 供应商可以通过基础模型降低 IT 运营成本。

4.1. 人工智能行业

ChatGPT 可以助力 AI 开发、优化模型优化及丰富应用场景等。ChatGPT 解决了许多之前无法解决的问题，包括事实型问答、文本摘要事实一致性、篇章级机器翻译的性别问题等。如果把 ChatGPT 当成是一个工具来帮助我们的开发、优化我们的模型、丰富我们的应用场景：1.代码开发：利用 ChatGPT 辅助开发代码，提高开发效率，包括代码补全、自然语言指令生成代码、代码翻译、bug 修复等；2.ChatGPT 和具体任务相结合：可以利用 ChatGPT 这样的底层模型，面向实际应用进行优化。OpenAI 创始人 Sam Altman 在访谈中也提到了这个领域。他认为会有一批新的创业公司采用已有的大模型，并对其进行调整。例如，创建医学模型、或者把电脑当作朋友等等。这些公司将创造很多的长期价值，因为他们将有一个特殊版本；3.利用 ChatGPT 指令微调激发的零样本能力：对于只有少数标注或者没有标注数据的任务以及需要分布外泛化的任务，我们既可以直接应用

ChatGPT，也可以把 ChatGPT 当作冷启动收集相关语料的工具，丰富相关的应用场景。

表 5: ChatGPT 在人工智能行业的应用场景

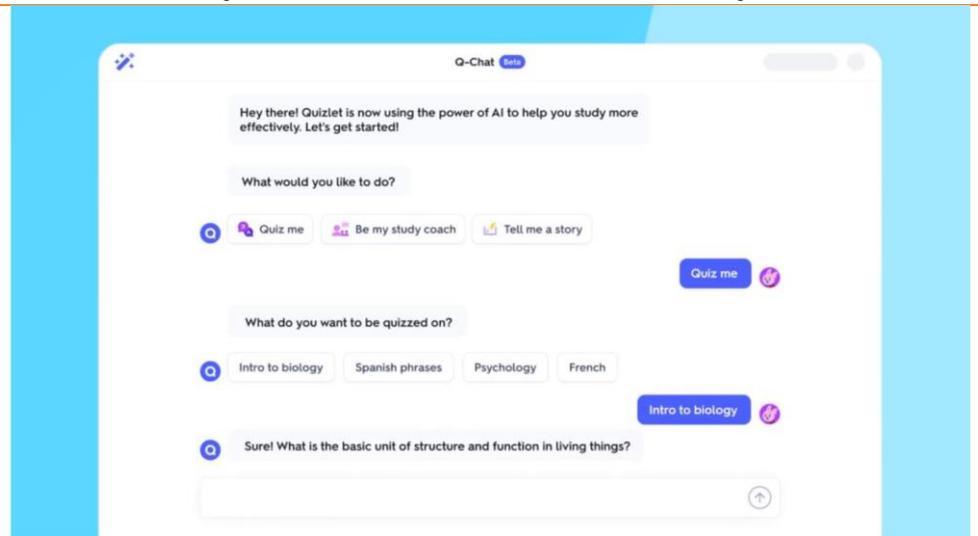
应用场景	具体描述
代码开发	利用 ChatGPT 辅助开发代码，提高开发效率，包括代码补全、自然语言指令生成代码、代码翻译、bug 修复等。
ChatGPT 和具体任务相结合	ChatGPT 的生成结果在许多任务上相比微调小模型都有很明显的可取之处（比如文本摘要的事实一致性，篇章级机器翻译的性别问题），在微调小模型的基础上结合这些 ChatGPT 的长处，可能可以在避免训练部署下显著提升小模型的效果。
利用 ChatGPT 指令微调激发的零样本能力	对于只有少数标注或者没有标注数据的任务以及需要分布外泛化的任务，我们既可以应用 ChatGPT，也可以把 ChatGPT 当作冷启动收集相关语料的工具，丰富相关的应用场景。

资料来源：BME 康复工程分会公众号，哈尔滨工业大学自然语言处理研究所《ChatGPT 调研报告》，天风证券研究所

ChatGPT 和具体任务相结合的模式较大程度上会成为其他企业的抢占热点。根据电厂公众号，随着 ChatGPT 的 API（能够实现和其他软件组件的交互）发布，网络上开发者借助 ChatGPT API 开发的各种工具层出不穷。例如，服务上，有能在原文的基础上，一站式借助 ChatGPT API，完成文本的翻译+润色+语法修改，甚至能直接将英文翻译成中文文的插件；产品上，Zeeno 键盘融合了 ChatGPT 功能，只需要将各种想法快速输入，Zeeno 就能在几秒内快速重新组织语言，使内容变得更加流畅易读。虽然目前这种模式更多还是局限于个人或小团队开发者，我们认为可以预见的是随着 API 的开放，预计各种应用服务乃至大厂都会开始思考将 ChatGPT 整合进自家服务中的更多可能，并着手于此。

服务上，ChatGPT 能够帮助 APP/语音助手提供更好的聊天服务。ChatGPT 能够更好地理解人类语言，并生成更符合人类需求的回答。APP 上，许多 APP 未来可能引入 ChatGPT，以提供更好的互动服务。例如，在线学习平台 Quizlet 宣布将整合 ChatGPT API，供全球超过 6000 万学生在线完成学习内容测试，并通过提供在线 AI 教师“Q-Chat”的方式，帮学生指出当前的学习材料中的重点以及难点，同时能根据内容自主提出问题，在聊天过程中帮学生掌握内容重点。语音助手上，一些人工智能专家表示，在未来，聊天机器人和语音助手的技术将会融合。这意味着人们能够通过语音控制聊天机器人，要求虚拟助手帮助他们完成工作，而不仅仅是查看天气等简单任务。将 ChatGPT 接入语音助手，能够在语音助手上实现类似 ChatGPT 理解复杂信息的能力，ChatGPT 强大的自然语言理解和沟通能力，有利于改善语音助手的使用体验。

图 14：在线学习平台 Quizlet 宣布将整合 ChatGPT API，提供在线 AI 教师“Q-Chat”



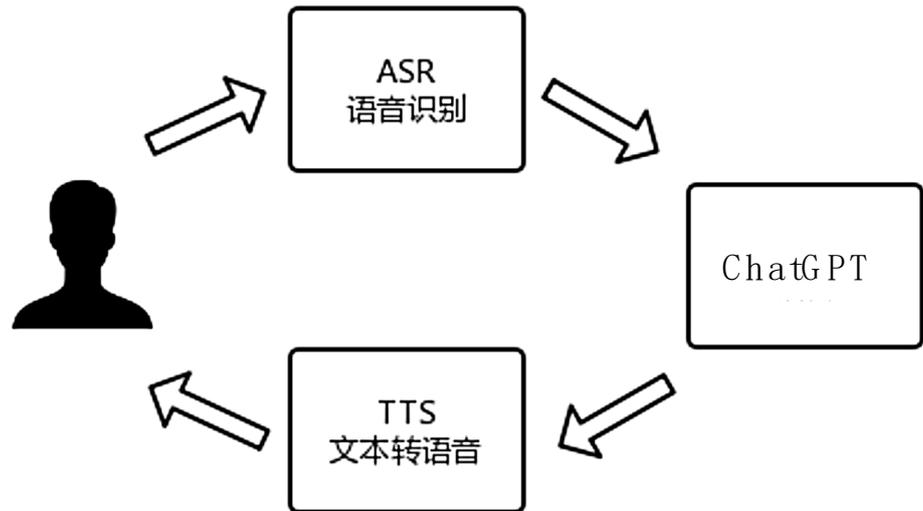
资料来源：电厂公众号，天风证券研究所

产品上，ChatGPT 能够增强智能产品的交互性。大部分时间，智能产品更多的是单向的

命令操作。ChatGPT 能够帮助智能产品读懂对话背后的潜台词，通过一项命令操作，去延伸出更多的思考，满足潜在需求。

智能家居中，以智能音箱为例，用户不仅仅要求其会说话、回答问题，更期望能够与自己聊天，在发出下一个命令之前，读懂需求，与其它智能家居产品形成联动。ChatGPT 有望推动这一交互的升级，相较于以往的聊天机器人，ChatGPT 没有固有答案输出，其理解能力和交互能力都远高于之前的聊天机器人，从而弥补了智能家居交互上的缺陷。此外，2022 年 10 月 4 日 Matter1.0 协议发布，该协议打破了不同品牌智能家居产品的兼容问题，降低了用户体验的成本。我们认为 ChatGPT 叠加协议支撑，智能家居市场在 2023 年有望迎来快速增长。

图 15：智能音箱接入 ChatGPT 的架构



资料来源：infoq.com，消费站，天风证券研究所

电子产品中，我们认为可将电子产品视为 AI 入口或出口，通过电子产品收集数据或输出内容，而 ChatGPT 则可以在其中处理数据与改善文本输出的角色。例如专注于办公场景的 iFLYBUDS 系列办公会议耳机，这一系列智能耳机在形态上与常见的蓝牙耳机相似，但其将耳机与 AI 语音助理开创性结合，可以帮助用户进行实时录音及转写，全程记录会议重点，若加入 ChatGPT 将有利于提高转写准确率，输出更符合要求的文字。

4.2. 其他行业

根据哈尔滨工业大学自然语言处理研究所在《ChatGPT 调研报告》中的总结，其他具有市场价值的行业应用还可以细分为四大部分：

搜索引擎：自 ChatGPT 发布以来，各大科技巨头都投入了极大的关注度，例如，谷歌担心 ChatGPT 会打破搜索引擎的使用方式和市场格局而开始内测自己的类 ChatGPT 产品 Bard，百度三月份将面向公众开放文心一言，微软更是宣布 ChatGPT 为必应提供技术支持，推出新必应。ChatGPT 与搜索引擎的结合已经不可避免，基于搜索引擎为 ChatGPT 提供生成结果、展示以及利用检索的新知识扩展 ChatGPT 的回答边界是正在进行的结合方向。

泛娱乐行业：ChatGPT 为文娱行业带来更多的机遇。无论是基于 ChatGPT 创建更智能的游戏虚拟人和玩家交流提升体验，还是利用虚拟数字人进行虚拟主播直播互动，ChatGPT 都为数字人提供了更智能的“大脑”，使行业充满想象空间。除此之外，数字人还可以拓宽到心理健康抚慰、闲聊家庭陪护等方面。

自媒体行业：ChatGPT 使自媒体行业大大受益。美国的新闻聚合网站 BuzzFeed 宣布和 OpenAI 合作，未来将使用 ChatGPT 帮助创作的内容。无论是旅游、餐饮、住宿、情感，ChatGPT 的出现将使得内容创作变得更加容易，相关博主的内容产出效率将得到极大的提升，有更多的精力润色相关内容，产生更多的高质量文章。

其他专业领域：受 ChatGPT 训练数据的限制，ChatGPT 无法对专业领域的专业知识进行细致的分析，生成的回答专业度不足且可信性难以保证，可以作为参考，不过很难实现替代。比如因为 ChatGPT 未获取 IDC、Gartner 等机构的数据使用授权，其关于半导体产业的市场分析中很少涉及量化的数据信息。

5. 投资建议：

我们认为以 GPT-4/ChatGPT 为代表的预训练大模型或将催生未来对 AI 服务器的扩产需求，建议关注

服务器芯片：英伟达（天风海外组覆盖）、AMD、Intel、寒武纪、海光信息（天风计算机组覆盖）、龙芯中科等；

AI 服务器制造：工业富联；

图像数据资源及应用：大华股份、海康威视等；

先进制造及封装：台积电、中芯国际、长电科技、通富微电等；

企业级存储：澜起科技、江波龙（天风计算机组联合覆盖）、兆易创新等。

6. 风险提示：

1. ChatGPT 发展不及预期：若 ChatGPT 发展不如预期，可能导致对上游芯片需求不及预期的情况；
2. 国际局势震荡加重贸易成本：若其他国家进一步加剧对中国的贸易政策限制，会增加国内相关芯片公司出口的贸易成本，影响国内相关芯片行业发展；
3. 市场竞争加剧：随着市场空间扩大，存在其他企业进入市场，加剧市场竞争，从而出现恶性竞争、压价的可能性；
4. 本文对芯片市场空间测算是基于一定前提假设，存在假设条件不成立、市场发展不及预期等因素导致市场空间测算结果偏差。

分析师声明

本报告署名分析师在此声明：我们具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，本报告所表述的所有观点均准确地反映了我们对标的证券和发行人的个人看法。我们所得报酬的任何部分不曾与，不与，也将不会与本报告中的具体投资建议或观点有直接或间接联系。

一般声明

除非另有规定，本报告中的所有材料版权均属天风证券股份有限公司（已获中国证监会许可的证券投资咨询业务资格）及其附属机构（以下统称“天风证券”）。未经天风证券事先书面授权，不得以任何方式修改、发送或者复制本报告及其所包含的材料、内容。所有本报告中使用的商标、服务标识及标记均为天风证券的商标、服务标识及标记。

本报告是机密的，仅供我们的客户使用，天风证券不因收件人收到本报告而视其为天风证券的客户。本报告中的信息均来源于我们认为可靠的已公开资料，但天风证券对这些信息的准确性及完整性不作任何保证。本报告中的信息、意见等均仅供客户参考，不构成所述证券买卖的出价或征价邀请或要约。该等信息、意见并未考虑到获取本报告人员的具体投资目的、财务状况以及特定需求，在任何时候均不构成对任何人的个人推荐。客户应当对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特定需求，必要时就法律、商业、财务、税收等方面咨询专家的意见。对依据或者使用本报告所造成的一切后果，天风证券及/或其关联人员均不承担任何法律责任。

本报告所载的意见、评估及预测仅为本报告出具日的观点和判断。该等意见、评估及预测无需通知即可随时更改。过往的表现亦不应作为日后表现的预示和担保。在不同时期，天风证券可能会发出与本报告所载意见、评估及预测不一致的研究报告。天风证券的销售人员、交易人员以及其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。天风证券没有将此意见及建议向报告所有接收者进行更新的义务。天风证券的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中的意见或建议不一致的投资决策。

特别声明

在法律许可的情况下，天风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，投资者应当考虑到天风证券及/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。

投资评级声明

类别	说明	评级	体系
股票投资评级	自报告日后的 6 个月内，相对同期沪深 300 指数的涨跌幅	买入	预期股价相对收益 20%以上
		增持	预期股价相对收益 10%-20%
		持有	预期股价相对收益 -10%-10%
		卖出	预期股价相对收益 -10%以下
行业投资评级	自报告日后的 6 个月内，相对同期沪深 300 指数的涨跌幅	强于大市	预期行业指数涨幅 5%以上
		中性	预期行业指数涨幅 -5%-5%
		弱于大市	预期行业指数涨幅 -5%以下

天风证券研究

北京	海口	上海	深圳
北京市西城区佟麟阁路 36 号 邮编：100031 邮箱：research@tfzq.com	海南省海口市美兰区国兴大道 3 号互联网金融大厦 A 栋 23 层 2301 房 邮编：570102 电话：(0898)-65365390 邮箱：research@tfzq.com	上海市虹口区北外滩国际客运中心 6 号楼 4 层 邮编：200086 电话：(8621)-65055515 传真：(8621)-61069806 邮箱：research@tfzq.com	深圳市福田区益田路 5033 号平安金融中心 71 楼 邮编：518000 电话：(86755)-23915663 传真：(86755)-82571995 邮箱：research@tfzq.com