计算机行业深度报告

AI 偏向科普性报告：围绕算法、算力、数据和应用

2023 年 04 月 06 日

增持（维持）

证券分析师 王紫敬
执业证书：S0600521080005
021-60199781
wangzj@dwzq.com.cn
研究助理 王世杰
执业证书：S0600121070042
wangshijie@dwzq.com.cn

投资要点

- **大模型是 AI 开发的新范式，是人工智能迈向通用智能的里程碑：**大模型指通过在大规模宽泛的数据上进行训练后能适应一系列下游任务的模型，本质依旧是基于统计学的语言模型，只不过“突现能力”赋予其强大的推理能力。现有的大模型的框架在本质上是一致的，几乎所有参数规模超过千亿的大语言模型都采取 GPT 模式，但是不同类型的企业给予自己所在领域的优势，开发的大模型在功能上还是有所差异。技术对大模型的效果具有决定作用，因此未来竞争格局也依赖于技术突破。

- **算力是 AI 时代的“石油”：**大模型的训练和推理都会用到 AI 芯片的算力支持，在数据和算法相同情况下，算力是大模型发展的关键，是人工智能时代的“石油”。我们假设 GPT-3 训练时间为一个月，则需要 843 颗英伟达 A100 芯片。我们假设 GPT-3 每日日活为 5000 万，则需要约 16255 颗英伟达 A100 芯片。GPT-4 为多模态数据，我们预计算力需求量是 GPT-3 的 10 倍以上。中国大厂相继布局大模型，我们测算，仅十家头部厂商大模型 1 年内有望增加约 20 万片 A100 需求量。长期来看，则需求量有望超 200 万片，新增算力需求将使算力市场增长 2 倍以上。2021 年，中国加速卡市场中 Nvidia 占据超过 80%市场份额，国产 AI 芯片性能与海外仍有差距，国产大模型推出有望加快国产芯片发展。

- **数据资源是 AI 产业发展的重要驱动力之一：**数据集作为数据资源的核心组成部分，是指经过专业化设计、采集、清洗、标注和管理，生产出来的专供人工智能算法模型训练的数据。大规模语言模型性能强烈依赖于参数规模 N，数据集大小 D 和计算量 C，训练数据主要来自于维基百科、书籍、期刊、Reddit 社交新闻站点、Common Crawl 和其他数据集，GPT4 依靠大量多模态数据训练。未来 AI 模型的竞争力或体现在数据质量和稀缺性，发展数据要素市场，促进相关公共、企业、个人数据的进一步放开，将为国内 AI 发展提供重要支撑。

- **AI 赋能各行各业，未来是 AI 应用的星辰大海：**AI 堪比第四次技术革命，本轮最直接的应用在内容创作领域，打开产业的想象边界。我们应该去寻找在 AI 赋能下，应用功能显著改善、客户粘性显著提升，市场空间大幅提升的领域，主要有内容创作，办公软件，ERP，机器人以及芯片设计领域。当前部分大模型厂商已经开启产业化应用，但是算力依旧是限制 AI 大规模商业化落地的主要原因，一旦解决，直接受益 AI+的将是信息化行业，因此我们看好各行业信息化领域处于优势地位的龙头公司。

- **投资建议：**算法上，我们建议关注已经有先发优势的大模型公司：三六零、科大讯飞、同花顺等，此外还有一些实施企业，如软通动力、润和软件、汉得信息等；算力上，我们推荐景嘉微、中科曙光、神州数码，建议关注海光信息、寒武纪、四川长虹、拓维信息等；数据上，我们推荐各细分赛道的信息化龙头企业，如久远银海、容知日新、中控技术，建议关注国能日新、千方科技等；应用上，我们推荐在具备“杀手级”应用潜能的厂商金山办公、用友网络、恒生电子，建议关注广联达、石基信息等。

- **风险提示：**政策推进不及预期；行业竞争加剧

行业走势

相关研究

《华为盘古大模型产业链梳理》
2023-03-27
《数据安全，为数据要素市场发展保驾护航》
2023-03-24

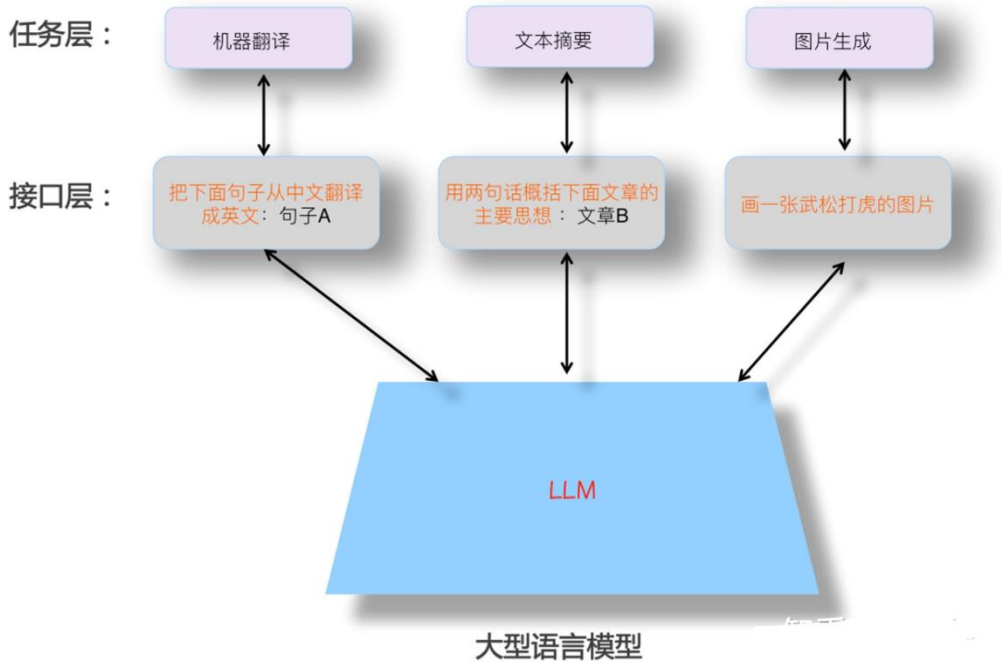
内容目录

图表目录

1. 算法：大模型——人工智能迈向通用智能的里程碑

大模型就是 Foundation Model（基础模型），指通过在大规模宽泛的数据上进行训练后能适应一系列下游任务的模型。大模型兼具“大规模”和“预训练”两种属性，面向实际任务建模前需在海量通用数据上进行预先训练，能大幅提升人工智能的泛化性、通用性、实用性，是人工智能迈向通用智能的里程碑技术。

图1：大语言模型

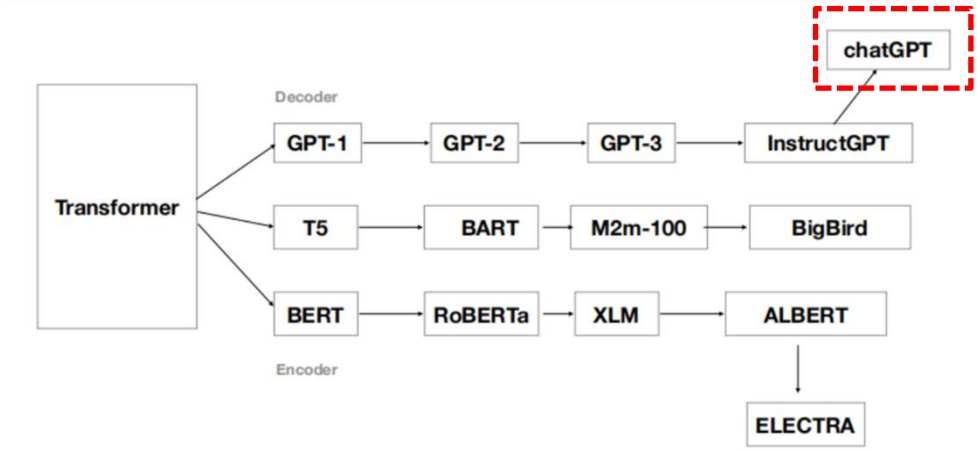


数据来源：人工智能前沿，东吴证券研究所

大模型的本质依旧是基于统计学的语言模型，“突现能力”赋予其强大的推理能力。通俗来讲，大模型的工作就是对词语进行概率分布的建模，利用已经说过的话预测下一个词出现的分布概率，而并不是人类意义上的“理解”。较过往统计模型不同的是，“突现能力”使得大模型拥有类似人类的复杂推理和知识推理能力，这代表更强的零样本学习能力、更强的泛化能力，

当前几乎所有参数规模超过千亿的大语言模型都采取 GPT 模式。近些年来，大型语言模型研究的发展主要有三条技术路线：Bert 模式、GPT 模式以及混合模式。Bert 模式适用于理解类、做理解类、某个场景的具体任务，专而轻，2019 年后基本上就没有什么标志性的新模型出现；混合模式大部分则是由国内采用；多数主流大语言模型走的还是 GPT 模式，2022 年底在 GPT-3.5 的基础上产生了 ChatGPT，GPT 技术路线愈发趋于繁荣。

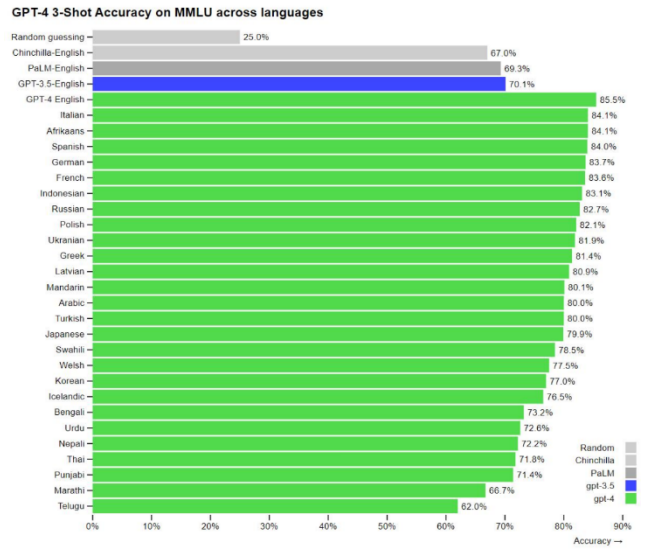
图2：Transformer 引领了大模型的爆发



数据来源：量子学派，东吴证券研究所

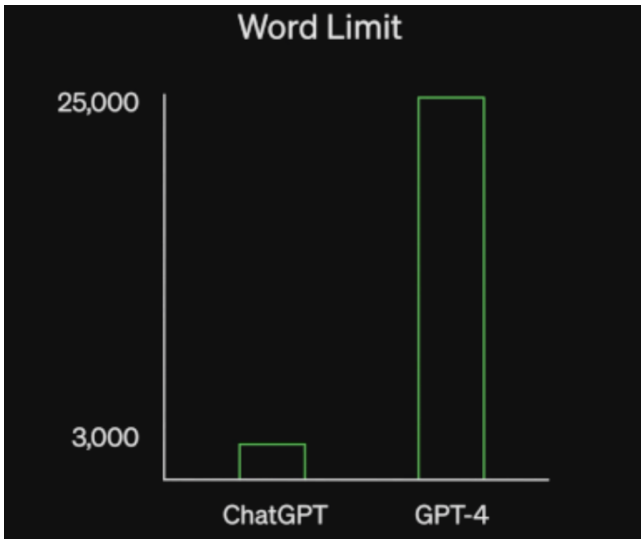
GPT4 作为人工智能领域最先进的语言模型，在如下四个方面有较大的改进。1）**多模态**：GPT4 可以接受文本和图像形式的 prompt，在人类给定由散布的文本和图像组成的输入的情况下生成相应的文本输出（自然语言、代码等）；2）**多语言**：在测试的 26 种语言的 24 种中，GPT-4 优于 GPT-3.5 和其他大语言模型（Chinchilla，PaLM）的英语语言性能；3）**“记忆力”**：GPT-4 的最大 token 数为 32,768，即 2^15，相当于大约 64,000 个单词或 50 页的文字，远超 GPT-3.5 和旧版 ChatGPT 的 4,096 个 token；4）**个性化**：GPT-4 比 GPT-3.5 更原生地集成了可控性，用户将能够将“具有固定冗长、语气和风格的经典 ChatGPT 个性”更改为更适合他们需要的东西。

图3：GPT-4 多语言性能表现优秀



数据来源：Accuracy，东吴证券研究所

图4：GPT-4 的文字输入限制



数据来源：Allmetaverse，东吴证券研究所

现有的大模型的框架在本质上是一致的，尚未出现技术上的“降维打击”。 GPT-4 虽然整体性能最为领先，但从技术上看，GPT-4 仍然是对自然语言处理增强学习、深度循环神经网络及其改进版本、大模型等已有技术的组合的创新，并且通过足够大量的数据进行支持，并非在大模型技术上有革命性突破。虽然国内 AI 大模型版本相对要滞后一些，但是并不存在不可弥补的鸿沟。

参数量和数据量是决定了模型效果。通常认为，参数量大于 1000 亿时，模型才有可能形成“突现能力”，这种现象在 GPT3 后开始更加显著。过往的 NLP 模型是按照具体任务和具体数据来训练的，所以数据质量越好，模型效果越好。而从 Transformer 开始，除了数据质量外，数据数量的重要性也愈发重要。因此参数量和数据量决定了模型最终的效果，最直观的效果指标就是准确度。此外，IDC 搭起了大模型评估框架 V1.0 以充分评估大模型技术能力、功能丰富度与底层深度学习平台开发能力，以及对各行业赋能的实际效果。

图5：大模型评估框架 V1.0

分类	一级	二级
产品能力	模型能力	功能丰富度
		模型性能
	平台工具能力	功能丰富度
		平台成熟度
		易上手程度
	开放性	开放可体验的能力数
		对用户数据隐私保护、数据安全措施
应用能力	应用广度	覆盖的行业数
	应用深度	客户业务流程关键环节渗透度
生态能力	应用生态	基于大模型进行产品开发的开发者数
		基于大模型工具与平台开发者创建的模型或应用数目

数据来源：IDC，东吴证券研究所

当前大模型的商业模式是“通用大模型+产业模型”。底层 AI 大模型的研发具有极高的研发门槛，面临高昂的成本投入，不利于人工智能技术在千行百业的推广。而具有数据、算力、算法综合优势的企业可以将模型的复杂生产过程封装起来，通过低门槛、高效率的生产平台，向千行百业提供大模型服务。各个行业的企业只需要通过生产平台提出在实际 AI 应用中的具体需求，生产大模型的少数企业就能够根据应用场景进一步对大模型开发训练，帮助应用方实现大模型的精调，以达到各行业对于 AI 模型的直接应用。

图6：大模型的投入成本

项目	成本
智算集群建设成本	一台搭载 A800 的服务器成本超过 40 万元，服务器采购成本通常是数据中心建设成本的 30%，一个智算集群的建设成本超过 **30 亿元**。
模型训练成本	大模型一次完整的模型训练成本为 1000 万-1 亿元人民币级别，如果进行 10 次完整的模型训练，成本便高达**数亿元**，再加上数据采集、人工标注、模型训练等一系列软性成本。
运营成本	数据中心内的模型训练需要消耗网络带宽、电力等资源，成本也以**亿元**为单位计算。

数据来源：财经十一人，东吴证券研究所

不同类型的企业在发展大模型拥有的优势也不尽相同。1）一是以阿里巴巴、华为、腾讯及百度为代表的基础云厂商，既具备做出通用 ChatGPT 的能力，也有着足够的数据和算力。2）二是以科大讯飞为代表的 AI 算法领先企业，被视为计算机板块中最有可能做出通用 ChatGPT 的公司。3）拥有天然的场景应用及配套数据优势的互联网平台。例如，国内最大的在线问答社区——知乎，以问答类任务为主模式与 GPT 天然契合。4）拥有高价值内容数据的企业也具备做好大模型的核心要素，可以大幅提升对人类意图的理解，从而提升回答信息的准确性。

各大厂商大模型百花齐放，核心差异在于细节。以国内厂商为例：1）百度由于多年在 AI 领域的深耕，其文心大模型涵盖基础大模型、任务大模型、行业大模型的三级体系，打造大模型总量约 40 个。2）腾讯混元应用方向则主要是腾讯自身生态的降本增效，其中广告类应用表现出色。3）阿里更重技术，通义大模型基于阿里云、达摩院打造的硬件优势，可将大模型所需算力压缩到极致；另外其底层技术优势还有利于构建 AI 的统一底层。4）华为的优势则在于其训练出业界首个 2000 亿参数以中文为核心的预训练生成语言模型，包括 NLP、CV、多模态、科学计算大模型，目前已实现医学、气象、时尚等多个 AI 场景落地。5）中科院的紫东太初是全球首个视觉-文本-语音三模态预训练模型，同时具备跨模态理解与跨模态生成能力。

图7：百度文心一言

数据来源：文心一言官网，东吴证券研究所

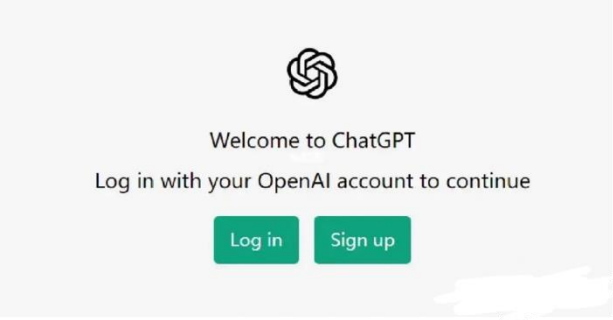
图8：华为盘古

数据来源：华为开发者大会 2021，东吴证券研究所

当前全球人工智能创新链基本形成了中美两国主导、东亚北美西欧协同引领的格局。美国是人工智能发展领域的前沿国家，其拥有一系列具备充足技术和资金资源的公司和实验室，各巨头科技公司均有相关的技术资源。其代表性模型有 ChatGPT、Claude、BarT、

BlenderBot3、Megatron-Turing 等；中国虽然在大模型上差距尚存，但研究和开发都非常活跃，正在加速追赶，也开发出了一些比较有代表性的模型，如百度文心一言、阿里通用、腾讯混元、华为盘古、中科院紫东太初等；东亚、北美、西欧等地区国家协同引领大模型发展，各有成果问世，如俄罗斯的 YaLM、英国的 Gopher、韩国的 HyperCLOVA、以色列的 Jurassic-1 Jumbo 等。

图9：OpenAI-ChatGPT



数据来源： OpenAI，东吴证券研究所

图10：Google-BERT

数据来源：Google，东吴证券研究所

如果仍维持 Transformer 的模型架构基础，未来行业将是寡头垄断的竞争格局。一方面，现有的大模型已经开始训练，在模型训练上有绝对的优势，模型的效果也会更好；另一方面，随着大模型版本的迭代，每一代大模型的算力、训练成本也有迹可循，对资金的需求也会持续扩大，没有雄厚资金支持的企业会逐渐掉队。

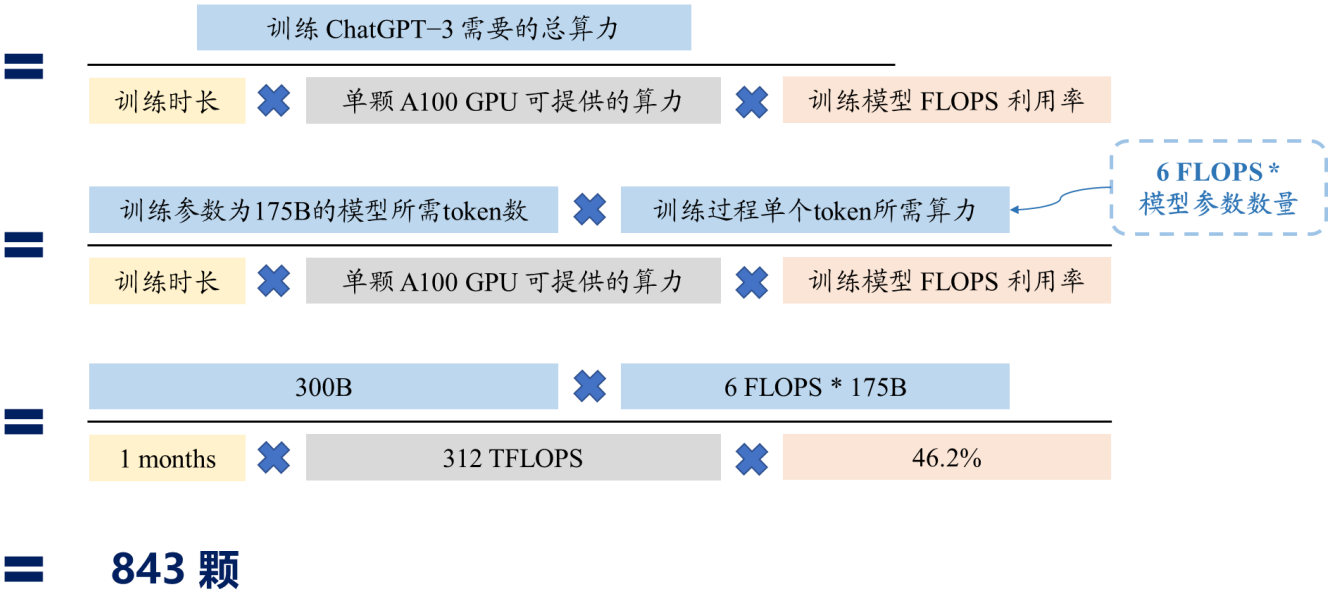
一旦大模型技术出现突破，行业竞争格局有望一家独大。由于当前大模型的技术是公用的，没有哪家存在明显的技术领先，因此各行各业厂商纷纷入局，希望分得一杯羹，因此出现了大模型百花齐放的竞争格局。一旦出现技术突破，大模型的准确度以及智能化出现了“碾压”的优势，行业需求会迅速向 NO.1 集中，有望形成一家独大的竞争格局。

2. 算力：AI 训练的基础设施

大模型算力成本主要分为初始训练成本和后续运营成本。

初始训练：根据 openAI 官网数据，每个 token（token 是服务端生成的一串字符串，以作客户端进行请求的一个令牌）的训练成本通常约为 6N FLOPS（FLOPS 指每秒浮点运算次数，理解为计算速度,可以用来衡量硬件的性能），其中 N 是 LLM（大型语言模型）的参数数量。1750 亿参数模型的 GPT-3 是在 3000 亿 token 上进行训练的。根据 openAI 官网数据，在训练过程中，模型的 FLOPS 利用率为 46.2%。我们假设**训练时间为 1 个月**，采用英伟达 A100 进行训练计算（峰值计算能力为 312 TFLOPS FP16/FP32），则**测算结果为需要 843 颗英伟达 A100 芯片。**

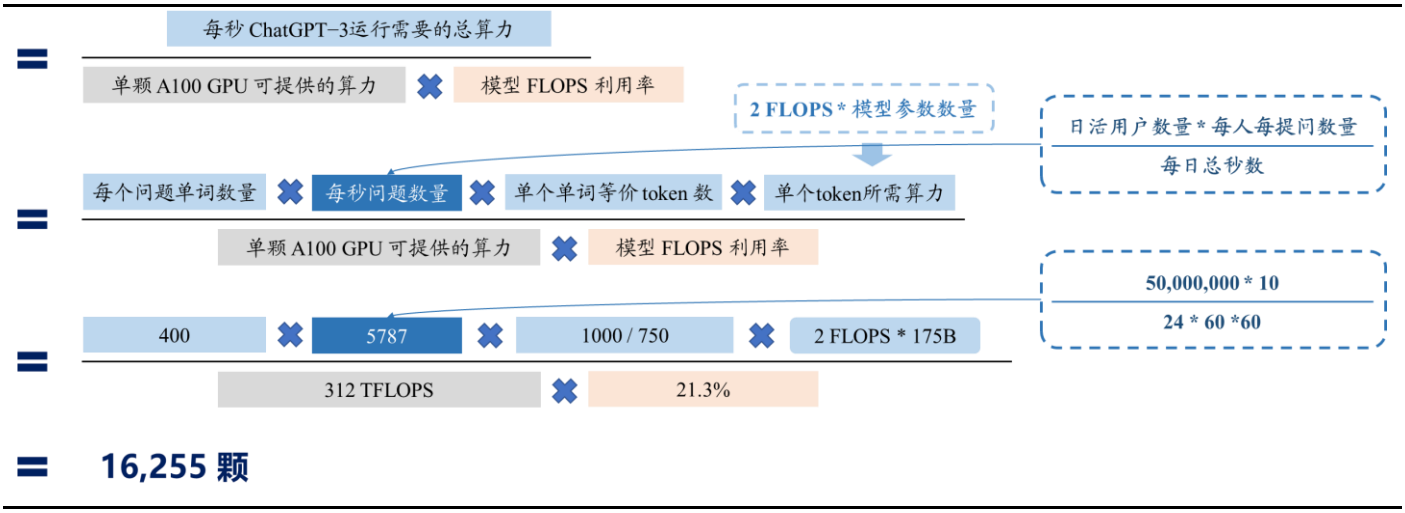
图11：用时 1 个月训练 ChatGPT-3 需要英伟达 A100 芯片数量



数据来源：CSDN，东吴证券研究所测算

运营（推理）成本：运营阶段所需算力量与使用者数量紧密相关。根据 openAI 官网数据，每个 token 的推理成本通常约为 2N FLOPS，其中 N 是 LLM 的参数数量。根据 openAI 官网数据，在训练过程中，模型的 FLOPS 利用率为 21.3%。同样采用英伟达 A100 进行推理计算（峰值计算能力为 312 TFLOPS FP16/FP32）。我们假设 GPT-3 每日 5000 万活跃用户，每个用户提 10 个问题，每个问题回答 400 字，则**测算结果为需要 16255 颗英伟达 A100 芯片。**

图12：维持 ChatGPT-3 每日 5000 万月活运营需要英伟达 A100 芯片数量



数据来源：CSDN，东吴证券研究所测算

GPT-4 为多模态大模型，对算力要求相比 GPT-3 会提升 10 倍。GPT-4 的收费是 8k context 为$0.03/1k token，是 GPT-3.5-turbo 收费的 15 倍（$0.002 / 1K tokens），因此我们推断 GPT-4 的参数量是 GPT-3 的 10 倍以上，预计 GPT-4 的算力需求是 GPT-3 的 10 倍以上。

表1：GPT-4 和 GPT-3.5-turbo 收费标准

	收费标准
GPT-4	$0.03/1k token
GPT-3.5-turbo	$0.002 / 1K tokens

数据来源：OpenAI 官网，东吴证券研究所

国产大模型有望带动国内新增 A100 出货量超 200 万颗，使得中国算力市场空间增加 2 倍以上。我们假设国内百度，华为，阿里，腾讯，字节等前 10 位头部大厂都会发布自己的大模型。**短期来看，**考虑到时间紧迫性，参考 GPT-3 的算力需求，仅十家头部厂商大模型 1 年内有望增加约 20 万片 A100 需求量。**长期来看，**如果后续迭代为多模态大模型或者活跃用户量大幅提升，则需求量有望超 200 万片。根据 IDC 数据，2021 年，中国加速卡数量出货超过 80 万片，新增算力需求将使算力市场增长 2 倍以上。

表2：国产大模型带动算力需求测算（短期为 2023 年，长期为 2024-2025 年）

国内自研大模型厂商数量（家）	10 家
每家短期需要 GPU 数	20,000 片
每家长期需要 GPU 数	200,000 片
短期总需求数	200,000 片
长期总需求数	2,000,000 片
2021 年中国加速卡出货量	800,000 片

数据来源：IDC，东吴证券研究所测算

加速卡国产化率较低，美国制裁加速。根据 IDC 数据，2021 年，**中国加速卡市场中 Nvidia 占据超过 80%市场份额。**2022 年 10 月 7 日，美国商务部工业和安全局（BIS）发布一套新的、范围广泛的出口管制措施，对向中国出口先进人工智能（AI）和超级计算芯片制造、生产设备以及所需的某些工具实施新限制。英伟达的 A100 和 H100 被列入出口管制清单。

英伟达推出中国特供版 A800，算力与 A100 基本一致。2022 年 11 月 8 日，英伟达推出 A800 GPU，将是面向中国客户的 A100 GPU 的替代产品。A800 符合美国政府关于减少出口管制的明确测试，并且不能通过编程来超过它。A800 GPU 在算力上与 A100 保持一致，但增加了 40GB 显存的 PCIe 版本，但在 NVLink 互联速度上，A800 相较于 A100 下降了 200GB/s 的速度。同时，A800 80GB SXM 版本目前已经不支持 16 块 GPU 的成套系统，上限被限制在 8 块。**总的来看，A800 能够满足国内市场需求，是 A100 的平替版本。**

图13：A800 和 A100 性能对比

公司	型号	INT8 Tensor Core	FP16 Tensor Core	TF32算力	FP64 Tensor Core	最大功耗 TDP	显存带宽 GB/s	NVLink	服务器选项 *
A100	80GB PCIe	624 /1248 TOPS	312 /624 TFLOPS	156 /312 TFLOPS	19.5 TFLOPS	300W	1935 GB/s	600 GB/s	1~8
	80GB SXM	624 /1248 TOPS	312 /624 TFLOPS	156 /312 TFLOPS	19.5 TFLOPS	400W	2039 GB/s	600 GB/s	4/8/16
A800	40GB PCIe	624 /1248 TOPS	312 /624 TFLOPS	156 /312 TFLOPS	19.5 TFLOPS	250W	1555 GB/s	400 GB/s	1~8
	80GB PCIe	624 /1248 TOPS	312 /624 TFLOPS	156 /312 TFLOPS	19.5 TFLOPS	300W	1935 GB/s	400 GB/s	1~8
	80GB SXM	624 /1248 TOPS	312 /624 TFLOPS	156 /312 TFLOPS	19.5 TFLOPS	400W	2039 GB/s	400 GB/s	4/8

* 合作伙伴和NVIDIA认证系统的GPU数量

数据来源：英伟达官网，东吴证券研究所

国产 AI 芯片性能与海外仍有差距，国产大模型推出有望加快国产芯片发展。国产 AI 芯片厂商主要有寒武纪，景嘉微，沐曦，燧原等，但其产品性能距离海外仍有差距。以国产寒武纪为例，MLU370 性能为 FP32 24TFLOPS，仅为英伟达 A100 的 10%不到。要达到同等算力要求，国产芯片片数需求量会更大，但大量 AI 芯片并行运行会对控制能力有较高要求，难以满足。但发展自己的 AI 芯片产业迫在眉睫，各家厂商正在快速追赶。

图14：国产 AI 芯片产品算力对比

公司	型号	场景	生产工艺	INT4算力	INT8算力	INT8 Tensor Core	FP16算力	FP16 Tensor Core	FP32算力	TF32算力	FP64算力	FP64 Tensor Core	Tensor 性能	最大功耗 TDP	最大功耗 TBP	显存带宽 GB/s
华为昇腾	昇腾310	推理	12nm FFC		16 TOPS		8 TOPS							8W		
	昇腾910	训练	N7+		640 TOPS		320 TFLOPS							310W		
寒武纪	MLU370-S4	推理	7nm		192 TOPS		96 TOPS		18 TFLOPS					75W		307.2 GB/s
	MLU370-X4	训练+推理	7nm		256 TOPS		96 TFLOPS		24 TFLOPS					150W		307.2 GB/s
	MLU370-X8	训练+推理	7nm		256 TOPS		96 TFLOPS		24 TFLOPS					250W		614.4 GB/s
	MLU290-M5	训练	7nm		512 TOPS		256 TOPS (INT16)		64 TOPS (CINT32)					350W		1228 GB/s
	MLU270-S4	推理		256 TOPS	128 TOPS		64 TOPS (INT16)							70w		102 GB/s
	MLU270-F4	推理		256 TOPS	128 TOPS		64 TOPS (INT16)							150w		102 GB/s
景嘉微	JM9100								512G FLOPS					5-15W		25.6GB/s
	JM92系列								1.2T FLOPS					15-30W		128GB/s
	M9系列								1.5T Flops					<30W		128GB/s
	JM7201		28nm CMOS											5~15W		
	JM7500		28nm CMOS											5-15W		
海光	深算一号		7nm FinFET				24.5 TFLOPS					10 TFLOPS		350 W		1024 GB/s
	深算二号		7nm FinFET				50 TFLOPS			25 TFLOPS		20 TFLOPS				
燧原科技	云燧T20	训练			256 TOPS		128 TFLOPS		32 TFLOPS	128 TFLOPS				300W		1.6TB/s
	云燧T21	训练			256 TOPS		128 TFLOPS		32 TFLOPS	128 TFLOPS				300W		1.6TB/s
	云燧21	推理			256 TOPS		128 TFLOPS		32 TFLOPS	128 TFLOPS				150W		819 GB/s
英伟达	V100 PCIe	训练+推理							14 TFLOPS		7 TFLOPS		112 TFLOPS	250W		900 GB/s
	V100 SXM2	训练+推理							15.7 TFLOPS		7.8 TFLOPS		125 TFLOPS	300W		900 GB/s
	V100S PCIe	训练+推理							16.4 TFLOPS		8.2 TFLOPS		130 TFLOPS	250W		1134 GB/s
	A100 80GB P	训练+推理				624 TOPS| 1248 TOPS	312 TFLOPS | 624 TFLOPS		19.5 TFLOPS	LOPS | 312 TFLOPS	9.7 TFLOPS	19.5 TFLOPS		300W		1935 GB/s
	A100 80GB S	训练+推理				624 TOPS| 1248 TOPS	312 TFLOPS | 624 TFLOPS		19.5 TFLOPS	LOPS | 312 TFLOPS	9.7 TFLOPS	19.5 TFLOPS		400W*		2039 GB/s
	H100 SXM	训练+推理				3958 TOPS		1979 teraFLOPS	67 teraFLOPS	989 teraFLOPS	34 teraFLOPS	67 teraFLOPS		700W		3.35TB/s
	H100 PCIe	训练+推理				3026 TOPS		1513 teraFLOPS	51 teraFLOPS	756teraFLOPS	26 teraFLOPS	51 teraFLOPS		300-350W		2TB/s
AMD	MI250	训练	TSMC 6nm Finf	362.1 TOPs	362.1 TOPs		362.1 TFLOPs		45.3 TFLOPs		45.3 TFLOPs				500W | 560W Peak	100 GB/s
	MI250X	训练	TSMC 6nm Finf	383 TOPs	383 TOPs		383 TFLOPs		47.9 TFLOPs		47.9 TFLOPs			500W | 560W Peak		100 GB/s

*400W TDP（适用于标准配置）。HGX A100-80 GB 自定义散热解决方案 (CTS) SKU 可支持高达 500W 的 TDP

数据来源：各公司官网，东吴证券研究所

3. 数据：AI 发展的驱动力

数据资源是 AI 产业发展的重要驱动力之一。数据集作为数据资源的核心组成部分，是指经过专业化设计、采集、清洗、标注和管理，生产出来的专供人工智能算法模型训

练的数据。人工智能应用的数据越多，其获得的结果就越准确。联想集团首席技术官芮勇认为，大模型的特点可以概括为“一大三多”：“‘一大’是指参数规模大，是千亿参数级别的超大型人工智能模型；‘三多’是指利用多来源、多模态、多任务的互联网海量数据进行训练。

大规模语言模型性能强烈依赖于参数规模 N，数据集大小 D 和计算量 C。OpenAI 在 2020 年曾经提出大模型缩放规律，计算量增加 10 倍，模型规模要增加 5 倍，训练数据增加 2 倍。尽管后来 DeepMind 重现定义了最优模型训练的参数规模和训练数据量之间的关系，说明数据规模和参数量同等重要，我们仍然可以定性地认为，大模型的性能提升需要依靠持续扩大的数据集实现。互联网提供的海量数据是 AI 近期能够取得突破性进展的重要基础。

大模型的训练数据主要来自于维基百科、书籍、期刊、Reddit 社交新闻站点、Common Crawl 和其他数据集。OpenAI 虽没有直接公开 ChatGPT 的相关训练数据来源和细节，但可以从近些年业界公布过的其他大模型的训练数据推测出 ChatGPT 的训练数据来源，近几年大模型训练采用的数据来源基本类似。国内大模型的数据来源和自身优势业务有较强相关性，如百度文心一言大模型的来源主要基于互联网公开数据，包括网页、搜索、图片、语音日均调用数据，以及知识图谱等。

图15：大模型训练数据来源统计（表中数字单位为 GB）

	Wikipedia	Books	Journals	Reddit links	CC	Other	Total
GPT-1		4.6					4.6
GPT-2				40			40
GPT-3	11.4	21	101	50	570		753
The Pile v1	6	118	244	63	227	167	825
Megatron-11B	11.4	4.6		38	107		161
MT-NLG	6.4	118	77	63	983	127	1374
Gopher	12.5	2100	164.4		3450	4823	10550

数据来源：Alan D. Thompson，东吴证券研究所

GPT4 依靠大量多模态数据训练。GPT4 是一个大规模的多模态模型，相比于此前的语言生成模型，数据方面最大的改进之一就是突破纯文字的模态，增加了图像模态的输入，具有强大的图像理解能力，即在预练习阶段输入任意顺序的文本和图画，图画经过 Vision Encoder 向量化、文本经过普通 transformer 向量化，两者组成多模的句向量，练习目标仍为 next-word generation。根据腾讯云开发者推测，GPT4 训练数据中还额外增加了包含正误数学问题、强弱推理、矛盾一致陈述及各种意识形态的数据，数据量可能是 GPT3.5（45TB 数据）的 190 倍。

未来 AI 模型的竞争力或体现在数据质量和稀缺性：根据 Google 的研究，数据质量

在高风险的人工智能领域具有更高的重要性，但人们往往只关注于模型，而忽略数据质量，在所有 AI 相关领域几乎都是如此。我们认为 GPT-4 更多依赖模型效率和数据质量的提升来实现改进，未来在细分垂直行业的优化也将基于行业特定数据展开。

高质量、稀缺的数据放开对 AI 发展至关重要。发展国内自己的大模型需要国内的高质量、稀缺数据。然而，根据发改委高技术司，我国政府数据资源占全国数据资源的比重超过 3/4，开放的规模却不足美国的 10%，个人和企业可以利用的规模更是不及美国的 7%，但这类数据的开放共享程度不高，全国开放数据集规模仅约为美国的 11%，数据有待进一步开放汇集，为开发更符合国内需求的大模型提供基础。**发展数据要素市场，促进相关公共、企业、个人数据的进一步放开，将为国内 AI 发展提供重要支撑。**

我们认为可以主要关注两个方面：能够采集、处理细分行业稀缺数据的厂商：久远银海、容知日新、国能日新、千方科技、中控技术、千方科技、用友网络等，以及具有专业数据处理服务能力的通用第三方厂商：海天瑞声等。

图16：数据采集示意图

数据来源：行行查，东吴证券研究所

表3：计算机各行业数据要素相关厂商

行业	相关公司
电信	中国移动，电信，联通，思特奇
广播电视	广电网络
能源	国能日新，恒实科技，朗新科技，远光软件，国网信通，朗新科技，海联讯，金现代，普联软件
金融	中科江南，税友股份，宇信科技，长亮科技，神州信息，新大陆，广电运通，ST 御银，证通电子，京北方，同花顺，银江技术，银之杰，新国都，浩云科技，高伟达，四

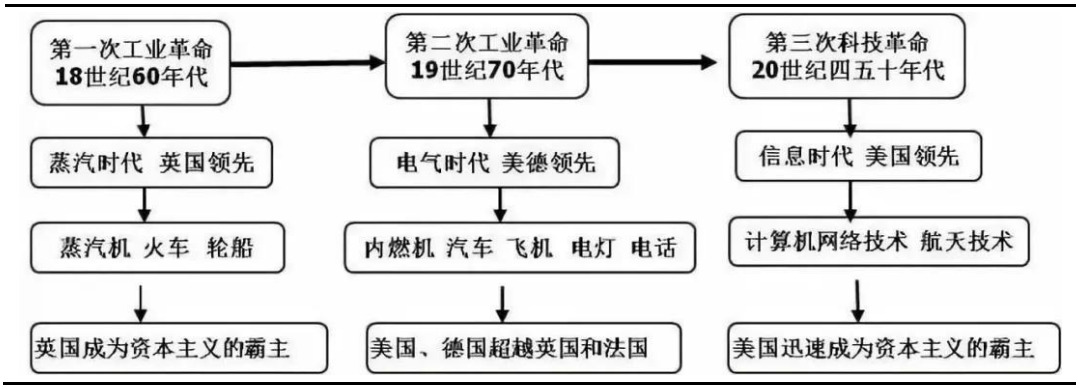
	方精创，古鳌科技，天阳科技，恒生电子，金证股份，顶点软件
公路水路运输	中远海科，千方科技，金溢科技，鸿泉物联，皖通科技，锐明技术，德赛西威，盛视科技，有棵树，运达科技，万集科技，天迈科技，通行宝，微创光电，多伦科技，诺力股份，道通科技
铁路	世纪瑞尔，唐源电气，思维列控
民航	航天信息
邮政	湘邮科技
水利	和达科技
应急管理	辰安科技
卫生健康	久远银海，国新健康，创业汇康，卫宁健康，山大地纬，万达信息，思创医惠，朗玛信息，荣科科技，和仁科技
社会保障	美亚柏科
国防科技	旋极信息，华如科技，佳缘科技，能科科技
冶金	上海钢联
地理信息	航天宏图，中科星图，四维图新，超图软件
酒店	石基信息
建筑	广联达，立方数科，恒华科技，盈建科，宏景科技，品茗科技
教育	竞业达，新开普，佳发教育，鸥玛软件，科大讯飞
煤炭	梅安森，北路智控，龙软科技

数据来源：Wind，东吴证券研究所

4. 应用：AI 的星辰大海

AI 时代已经来临，最大的市场将是被 AI 赋能的下游应用市场。如果说 AI 是第四次工业革命，那么正如前三次工业革命，最大的市场将是被 AI 赋能的下游应用市场。本轮革命性的产品 ChatGPT 将极大地提升内容生产力，率先落地于 AIGC 领域，打开其产业的想象边界。文本生成、代码生成、图像生成以及智能客服将是能直接赋予给下游行业的能力，打开其产业想象的边界。

图17：三次工业革命带来下游应用技术爆发



数据来源：维基百科，东吴证券研究所

我们应该去寻找“杀手级”的下游应用市场。所谓“杀手级”应用市场，即在 AI 赋能下，该应用功能显著改善，客户粘性显著提升，最后体现为客户付费率和付费单价显著提升，市场空间大幅提升。我们认为根据美国产业发展现状来看，当前来看已经涌现的“杀手级”应用领域主要有内容创作，办公软件，ERP，机器人以及芯片设计领域。

最直接的应用在内容创作领域。ChatGPT 的功能核心是基于文本的理解和分析，与内容创作行业趋同。ChatGPT 可用于创建新闻文章、博客文章甚至小说等内容，它可以生成原创且连贯的内容，为内容创作者节省时间和资源。整体生成式 AI 已用于创建图像，视频，3D 对象，Skyboxes 等。这大大节省了创作时间，同时带来了多样的创作风格。

图18：GPT-4 画出了《三体》中的罗辑

数据来源：行者慎思，东吴证券研究所

图19：AI 生成不同的 3D 建筑风格

数据来源：设计癖，东吴证券研究所

在办公软件领域是划时代的生产力的解放。我们认为办公场景是当前所能看到的 AI 最大级别应用场景，Microsoft365 Copilot 将会带来需求的刚性，是人工智能杀手级应用。在 Word 中，Copilot 可以帮助打草稿、排版、修改；在 Excel 中，Copilot 可以帮助用户修改样式，并进行数据分析、预测、可视化等；在 PPT 中，Copilot 能够基于文本自动绘制 PPT，根据 PPT 生成讲稿。Copilot 协助完成 Office 套件中大量执行类工作，并提供低阶创意参考，极大提升使用者效率，节省重复性时间 Microsoft365 Copilot 的推出标志着人类与计算机交互方式的重大进步，这将彻底改变我们的工作方式，开启新一波生产力增长。

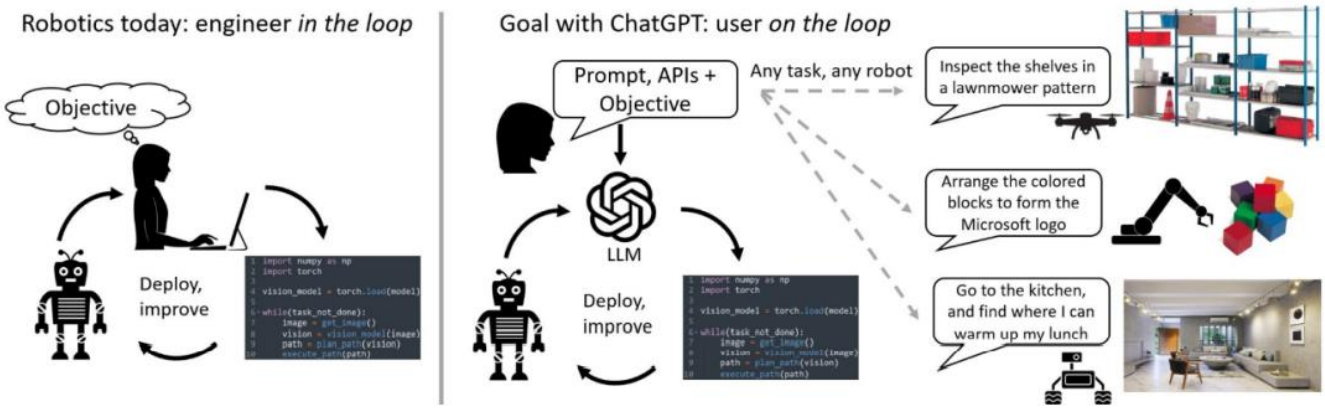
图20：Microsoft365 Copilot

数据来源：微软发布会，东吴证券研究所

与 ERP 的结合有望重构企业管理。ERP 能够学习企业管理“通用数据”，又能学习企业管理“私有数据”，让 ERP 在做到贯彻领先企业管理理念的同时，越来越个性化，低成本满足企业对 ERP 定制化的需求。ERP 使用流程繁琐，使用 ChatGPT 和直接询问，获得想要的信息，可降低使用者门槛。ERP 往需要很多繁重的人工操作，例如手动输入数据、生成内容和标注笔记等，使用 ChatGPT 可以自动抓取，减少人数。ERP 是企业管理核心软件，AI 赋能后有望进一步增强客户粘性，为客户创造更多价值。

ChatGPT 解决了机器人的痛点。ChatGPT 开启了一种新的机器人范式，允许潜在的非技术型用户参与到回路之中，ChatGPT 可以为机器人场景生成代码。在没有任何微调的情况下，利用 LLM 的知识来控制不同的机器人动作，以完成各种任务。ChatGPT 大大改善了机器人对指令的理解，并且不同于以前单一、明确的任务，机器人可以执行复合型的任务。

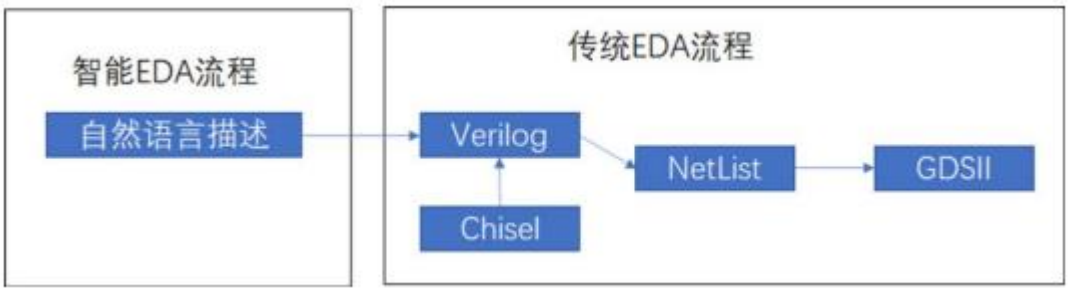
图21：ChatGPT 改善了机器人对环境的适应性



数据来源：每日经济新闻，东吴证券研究所

看好 ChatGPT 在芯片设计领域的应用。传统的芯片设计强烈依赖模板而忽视了大量可以复用的优秀数据，同时数据量大导致 ChatGPT 泛化性更好。此外芯片硬件模块相对单一，有一些成熟范式，芯片设计代码复杂但人工不足，这些都与 ChatGPT 有很好的互补。AI 使得芯片开发成本降低、周期缩短，具备足够多训练数据和 AI 能力的芯片设计公司竞争优势可能会扩大。

图22：智能 EDA 和传统 EDA 流程图



数据来源：机器之心专栏，东吴证券研究所

我们认为算力是限制 ChatGPT 大规模商业化落地的主要原因。ChatGPT-4 访问被持续限流，本质上是算力成本承压。OpenAI 对于 Plus 付费用户的 GPT-4 访问阈值在较短的时间内连续下降了 4 次，背后是其日活和周活用户数的持续攀升，大规模的用户访问使得 GPT 的算力成本进一步增长。随着应用端逐渐丰富，对算力的需求提出了更多的需求，预计未来的算力需求缺口将会持续扩大。纵使 ChatGPT 前尚处于发展的早期探索阶段，也存在一些如算法模型不完善、理解能力不足、回答问题不够灵活等突出问题，但是其目前已经成功跑出了商业模式，却由于算力需求缺口不得不进行访问限制或者降低精度。随着 GPT 生态的建立、相关应用的爆发，算力的需求将持续扩大，算力需求缺口将会持续扩大，成为 ChatGPT 大规模商业化的限制。

AI 时代已经来临，AI+万物将赋能千行百业，未来各信息化赛道都会探索出各自的人工智能应用场景。这其中，我们更加看好各行业信息化领域处于优势地位的龙头公司，他们不仅具备了较高的市场份额，同时在资源集聚、行业 Knowhow 积累和行业壁垒上

都实现了比较优势，这些公司包括海康威视、金山办公、恒生电子、广联达、深信服、中科创达、用友网络、科大讯飞、三六零、同花顺、石基信息等。

5. 投资建议与相关标的

算法上，我们建议关注已经有先发优势的大模型公司：三六零、科大讯飞、同花顺等，此外还有一些实施企业，如软通动力、润和软件、汉得信息等；

算力上，我们推荐景嘉微、中科曙光、神州数码，建议关注海光信息、寒武纪、四川长虹、拓维信息等；

数据上，我们推荐各细分赛道的信息化龙头企业，如久远银海、容知日新、中控技术，建议关注国能日新、千方科技等；

应用上，我们推荐在具备“杀手级”应用潜能的厂商金山办公、用友网络、恒生电子，建议关注广联达、石基信息等。

6. 风险提示

政策推进不及预期。相关政策推进受到多种因素影响，节奏和力度可能不及预期。

行业竞争加剧。行业市场空间广阔，可能吸引更多公司参与行业竞争。

免责声明

东吴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。

本研究报告仅供东吴证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议，本公司不对任何人因使用本报告中的内容所导致的损失负任何责任。在法律许可的情况下，东吴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。

市场有风险，投资需谨慎。本报告是基于本公司分析师认为可靠且已公开的信息，本公司力求但不保证这些信息的准确性和完整性，也不保证文中观点或陈述不会发生任何变更，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。

东吴证券投资评级标准:

公司投资评级:

买入：预期未来 6 个月个股涨跌幅相对大盘在 15%以上;

增持：预期未来 6 个月个股涨跌幅相对大盘介于 5%与 15%之间;

中性：预期未来 6 个月个股涨跌幅相对大盘介于-5%与 5%之间;

减持：预期未来 6 个月个股涨跌幅相对大盘介于-15%与-5%之间;

卖出：预期未来 6 个月个股涨跌幅相对大盘在-15%以下。

行业投资评级:

增持： 预期未来 6 个月内，行业指数相对强于大盘 5%以上;

中性： 预期未来 6 个月内，行业指数相对大盘-5%与 5%;

减持： 预期未来 6 个月内，行业指数相对弱于大盘 5%以上。

东吴证券研究所

苏州工业园区星阳街 5 号

邮政编码：215021

传真：（0512）62938527

公司网址： http://www.dwzq.com.cn