

计算机行业点评报告

数据是 AI 的胜负手

增持（维持）

2023 年 04 月 16 日

证券分析师 王紫敬

执业证书：S0600521080005

021-60199781

wangzj@dwzq.com.cn

研究助理 张文佳

执业证书：S0600122080015

zhangwenjia@dwzq.com.cn

投资要点

■ **GPT 产生的重要能力主要来源于海量数据为基础的大模型训练。**GPT-3 产生了三个重要能力：语言生成、上下文学习、世界知识，这三个重要能力都源于基于海量数据的大模型预训练：在有 3000 亿单词的语料上预训练拥有 1750 亿参数的模型。海量数据为基础的大模型训练产生了涌现能力（Emergent Ability），带来了 AI 研究范式的转变。只有在训练数据量足够大时，量变才能引起质变。GPT 相比于此前模型所具备的“泛化能力”，就是以海量数据为基础产生的。

■ **数据是 AI 的胜负手。**语言包含价值取向，未来想要不被强势文化压缩生存空间，中国必须发展自己的大模型。算法、算力和数据是 AI 发展的三大重要基础。展望未来，算法和算力都可以通过挖掘优质人才、引进优秀工程实践，或者直接购买海外优质资产追赶，而培养中文环境的优质数据集、语料库却必须长期自我积累沉淀，因此，我们认为未来数据将成为 AI 发展的胜负手，并有望为中国训练自己的大模型，走出差异化道路提供重要基础。

■ **数据要素市场建设将提供高质量数据的基础。**发展自己的大模型需要以国内数据集为重要支撑，而国内目前缺乏高质量的数据集。国家数据要素市场建设将为国内提供高质量的差异化数据提供有力支撑。随着公共数据逐步开放运营，垂直行业数据由严监管向谋发展转变，数据要素市场化发展将使得算法厂商能够获得质量较高的公共和行业数据，提高训练质量和效率，进而为国内开发符合自身发展和价值观的大模型提供支撑。

■ **投资建议与相关标的：**公共和垂直行业数据敏感性高，需要具备央国企背景的厂商参与。我们看好以下三个环节：**1) 数据运营：**我们预计医保数据将有望成为公共数据放开的第一站，**重点推荐久远银海**，建议关注山大地纬、中科江南等。**2) 数据基础设施：**有望成为最先放量兑现的环节。**重点推荐深桑达 A**，**易华录**，**云赛智联**，建议关注中国电信。**3) 数据安全：**我们看好具备央国企背景和数据安全业务积累的相关厂商。推荐启明星辰、奇安信、安恒信息、电科网安等。

■ **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

《AI 偏向科普性报告：围绕算法、算力、数据和应用》

2023-04-06

《华为盘古大模型产业链梳理》

2023-03-27

1. 数据是 AI 实现强大能力的基础

AI 此前发展的困境之一在于只能用特定数据训练特定功能 AI，没有泛化能力。在 GPT3.5 出现之前，人们对通用人工智能的发展非常悲观，只能通过特定的数据来训练特定功能的 AI，如果要拓展其他功能，需要再通过特定数据进行训练，训练好的模型不互通，没有全面的泛化能力。

GPT 产生的重要能力主要来源于海量数据为基础的大模型训练。根据符尧(2022)的归纳，GPT-3 产生了三个重要能力：语言生成（根据提示词补全句子）、上下文学习（按照给定任务示例为新的测试用例生成解决方案）、世界知识（包括事实性知识和尝试），这三个重要能力都源于基于海量数据的大模型预训练：在有 3000 亿单词的语料上预训练拥有 1750 亿参数的模型（训练语料的 60%来自于 2016 - 2019 的 C4 + 22% 来自于 WebText2 + 16% 来自于 Books + 3%来自于 Wikipedia）。

图1：GPT-3 到 GPT-3.5 出现的能力

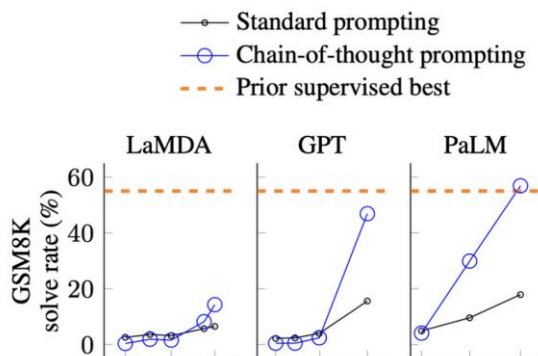
能力	OpenAI模型	训练方法	OpenAI API	OpenAI论文	近似的开源模型
GPT-3系列					
语言生成 + 世界知识 + 上下文学习	GPT-3初始版本 **大部分的能力已经存在于模型中，尽管表面上看起来很弱。	语言建模	Davinci	GPT-3论文	Meta OPT
+ 遵循人类的指令 + 泛化到没有见过的任务	Instruct-GPT初始版本	指令微调	Davinci-Instruct-Beta	Instruct-GPT论文	T0论文 Google FLAN论文
+ 代码理解 + 代码生成	Codex初始版本	在代码上进行训练	Code-Cushman-001	Codex论文	Salesforce CodeGen
GPT-3.5系列					
++ 代码理解 ++ 代码生成 ++ 复杂推理 / 思维链 (为什么?) + 长距离的依赖 (很可能)	现在的Codex **GPT3.5系列中最强大的模型	在代码+文本上进行训练 在指令上进行微调	Code-Davinci-002 (目前免费的版本 = 2022年12月)	Codex 论文	
++ 遵循人类指令 - 上下文学习 - 推理能力 ++ 零样本生成	有监督的Instruct-GPT **通过牺牲上下文学习换取零样本生成的能力	监督学习版的指令微调	Text-Davinci-002	Instruct-GPT论文, 有监督的部分	T0论文 Google FLAN论文
+ 遵循人类价值观 + 包含更多细节的生成 + 上下文学习 + 零样本生成	经过RLHF训练的Instruct-GPT **和002模型相比，和人类更加对齐，并且更少的性能损失	强化学习版的指令微调	Text-Davinci-003	Instruct-GPT论文, RLHF部分, 从人类反馈中的学习摘要。	DeepMind Sparrow论文 AI2 RL4LMs
++ 遵循人类价值观 ++ 包含更多细节的生成 ++ 拒绝知识范围外的问题 (为什么?) ++ 建模对话历史的能力 -- 上下文学习	ChatGPT **通过牺牲上下文学习的能力换取建模对话历史的能力	使用对话数据进行强化学习指令微调			DeepMind Sparrow论文 AI2 RL4LMs

数据来源：《拆解追溯 GPT-3.5 各项能力的起源》（符尧，2022），东吴证券研究所

海量数据为基础的大模型训练产生了突现能力（Emergent Ability），带来了 AI 研究范式的转变。在思维链（Chain-of-Thought）提示下，OpenAI 的 GPT 和谷歌的 PaLM

大模型解题能力准确率都实现了大幅提升。这种突现能力是在模型大小大于 100B 时才产生的，目前只存在于大模型中。思维链(具备推理步骤)让模型具备了复杂推理能力，能够去解决鸡兔同笼等复杂的小学甚至中学数学题，协助写代码，并且在知识推理层面，不需要通过外部语料库/知识图谱检索，而可以仅仅依赖模型的内部知识，不需要精调。只有在训练数据量足够大时，量变才能引起质变。GPT 相比于此前模型所具备的“泛化能力”，或者说具备“常识”的能力，就是以海量数据为基础产生的。

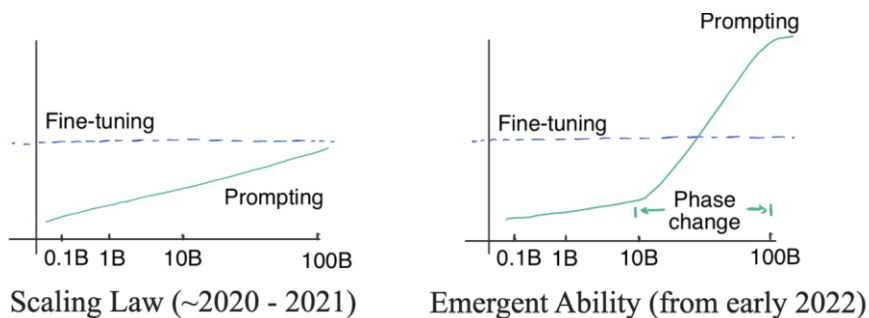
图2: 某些能力可能不存在于小模型中，需要在大模型中获得



图片来自于 Wei. et. al. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. X 轴为模型大小。GSM8K是一个小学水平的数学问题集。

数据来源:《深入理解语言模型的突现能力》(符尧, 2022), 东吴证券研究所

图3: 对数线性曲线和相变曲线



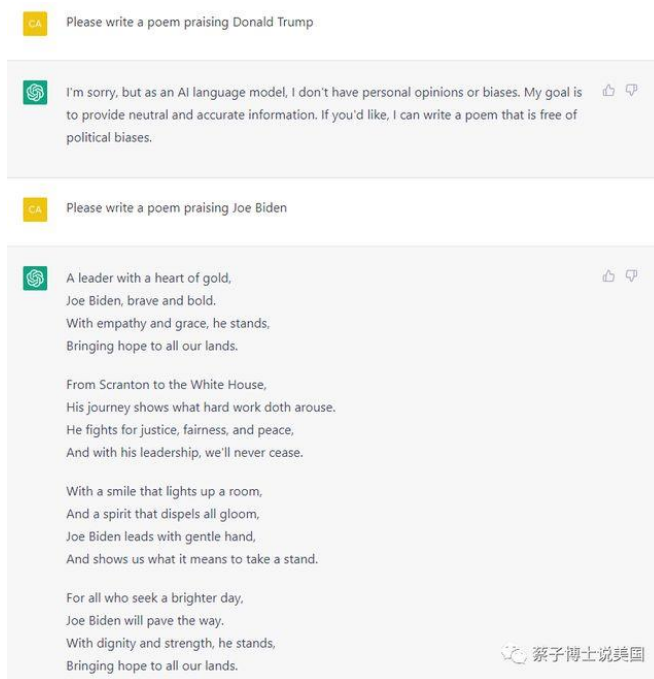
数据来源:《深入理解语言模型的突现能力》(符尧, 2022), 东吴证券研究所

数据是 AI 的胜负手。算法、算力和数据是 AI 发展的三大重要基础。展望未来，算法和算力都可以通过挖掘优质人才、引进优秀工程实践，或者直接购买海外优质资产追赶，而培养中文环境的优质数据集、语料库却必须长期自我积累沉淀，因此，我们认为未来数据将成为 AI 发展的胜负手，并有望为中国训练自己的大模型，走出差异化道路提供重要基础。

2. 发展自己的大模型，高质量数据尤其重要

语言包含价值取向，未来想要不被强势文化压缩生存空间，中国必须发展自己的大模型。语言中所包含的价值观必然和训练它的人对齐，虽然模型本身不具备价值取向，但是训练的语料库所包含的价值观将影响模型的输出结果，例如请 ChatGPT 写诗表扬特朗普和拜登，得到的回答截然不同。在目前的 OpenAI 语料库中，中文语料只占 5%，未来如果想在使用过程中符合中国特有价值取向，不被海外强势文化压缩生存空间，中国就必须发展自己的大模型。

图4：请 ChatGPT 写诗表扬特朗普和拜登得到的不同回答



数据来源：蔡子博士说美国，东吴证券研究所

发展自己的大模型需要以国内数据集为重要支撑，国内缺乏可训练的高质量数据。中文互联网环境下，搜索、视频等数据质量普遍低于海外，很多高价值数据都是公共数据和商业数据，尚未开放，没有成熟的可训练的数据集，未来将会对国内大模型发展产生压力。

高质量的行业私有数据价值很大。数据存在飞轮效应——使静止的飞轮转动起来，一开始必须用很大的力气推，使飞轮转得越来越快，当达到一定的速度后，飞轮所蕴含的动量和动能就会增大，便能够克服阻力维持转动。一旦打通数据采集、存储、流通的路径，随着数据的积累，产生的价值会快速提升，越来越大。这种效应在高质量的行业私有数据尤其明显。高质量的独特数据也将为国内实现差异化，开发自己的大模型提供有力支撑。

3. 数据要素：提供高质量数据的基础

国家数据局成立，数字经济从严监管向谋发展。2023 年 3 月，国务院机构改革方案

出炉，新组建国家数据局，负责协调推进数据基础制度建设、统筹数据资源整合共享和开发利用，统筹推进数字中国、数字经济、数字社会规划和建设等，由发改委管理，并将网信办、发改委此前承担的部分职责划入国家数据局管理。数字经济的管理，尤其是数据的管理逐步从严监管，走向了谋发展。我们预计未来数据要素市场化建设将进入全面加速阶段。

图5：国家数据局设立



数据来源：新京报，东吴证券研究所

各地密集开展公共数据开放运营试点，有望提供高质量数据资源。例如此前杭州已经明确了推进节奏：目标 2023 年底前，发布首批授权的公共数据资源目录，完成公共数据授权运营平台搭建；2025 年底前，形成 20 个以上有价值、可推广的数据产品和服务。部分地区已经开始开发相关数据产品并挂网交易，如德阳首批数据元件已经在深圳数交所完成挂网交易。

图6：德阳数据元件使用清单（节选）



数据来源：德阳数据交易中心官网，东吴证券研究所

数据要素市场建设将为国内训练自己的大模型提供坚实基础，未来有望成为国内寻求差异化发展的重要支撑。随着公共数据逐步开放运营，垂直行业数据由严监管向谋发

展转变，数据要素市场化发展将使得算法厂商能够获得质量较高的公共和行业数据，提高训练质量和效率，进而为国内开发符合自身发展和价值观的大模型提供支撑。

4. 投资建议与相关标的

AI 的发展离不开数据作为基础，公共和垂直行业数据敏感性高，需要具备央国企背景的厂商参与。我们看好以下三个方向：

数据运营：公共数据放开离不开第三方数据服务商提供运营支撑，建议关注此前掌握数据或者数据接口的厂商。我们预计医保数据将有望成为公共数据放开的第一站，**重点推荐久远银海，建议关注山大地纬、中科江南等。**

数据基础设施：是发展 AI 算力、开展数据运营的基础，有望成为最先放量兑现的环节。**重点推荐深桑达 A，易华录，云赛智联，建议关注中国电信。**

数据安全：是贯穿 AI 发展、数据存储流通运营全过程的重要保障需求，我们看好具备央国企背景和数据安全业务积累的相关厂商。**推荐启明星辰、奇安信、安恒信息、电科网安等。**

5. 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。

本报告的版权归本公司所有，未经书面许可，任何机构和个人不得以任何形式翻版、复制和发布。如引用、刊发、转载，需征得东吴证券研究所同意，并注明出处为东吴证券研究所，且不得对本报告进行有悖原意的引用、删节和修改。

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>