奇点将至，探他山之石

——从算力、算法、数据和应用看AIGC

2023年3月19日

目录

01 生成式AI：ChatGPT引燃市场，数字经济未来已至

02 数据：大模型训练的基础资源

03 算力：大模型发展带来高算力需求

04 算法：大模型算法助力AIGC突破

05 产业应用：各领域应用加速落地，商业化前景广阔

生成式AI：自然语言处理演变十余年，迎来变现阶段

AIGC(AI Generated Content)即生成式AI，多领域应用逐渐成熟。AIGC涉及无监督和半监督学习算法，截至目前其发展历程主要分为三个阶段：

- **统计机器学习方法阶段（2010年前）**：首先对数据进行手工标注，然后构建其重要特征，最后构建概率模型并进行参数优化，从而将概率最大的输出作为结果；
- **基于深度学习的神经网络模型（2010年-2017年）**：深度学习算法被引入，本质上是通过大量数据训练神经网络，主要表现形式为：CNN（卷积神经网络）、RNN(循环神经网络）等。相比统计学习方法，省去了复杂且手工的特征构建；
- **基于Transformer结构的预训练模型（2017年至今）**：利用大量无标注数据进行自监督学习，然后再使用少量的标注数据对下游任务进行微调（即迁移学习）。
- 在应用方面，按场景分类AIGC已经较为成熟地应用于文本和代码撰写、图像识别和生成，以GPT为首的AIGC模型也正在探索消费级AI技术的变现方式。展望未来，AIGC不仅会在现有应用领域持续进步，也将逐步拓展到视频和游戏领域，AIGC将会在更多的领域得到广泛应用，为各个行业和领域的发展和进步提供更多可能性。

表1：AI应用发展进程预测

	2020前	2020	2022	预计2025	预计2030	预计2050
文本	垃圾邮件检测 翻译 基础问答	基础文案撰写 生成草案	撰写更长文章 完善文稿	对科学论文等进行垂直微调	文章终稿超过人类平均水平	文章终稿超过专业作者水平
代码	单行自动完成	多行代码生产	更长代码 更高准确度	更多语言 深度提高	文本到产品（草稿）	文本到产品（终稿），超过大部分开发者
图像			艺术 Logo 摄影	产品设计、建筑等模型	产品设计、建筑等终稿	终稿超过大部分专业艺术家、设计师、摄影师水平
视频/3D/游戏				视频和3D制作的初稿	完善版本	AI创作平台 游戏和电影实现个性化定制
			开始尝试		**基本完成**	**黄金时期**

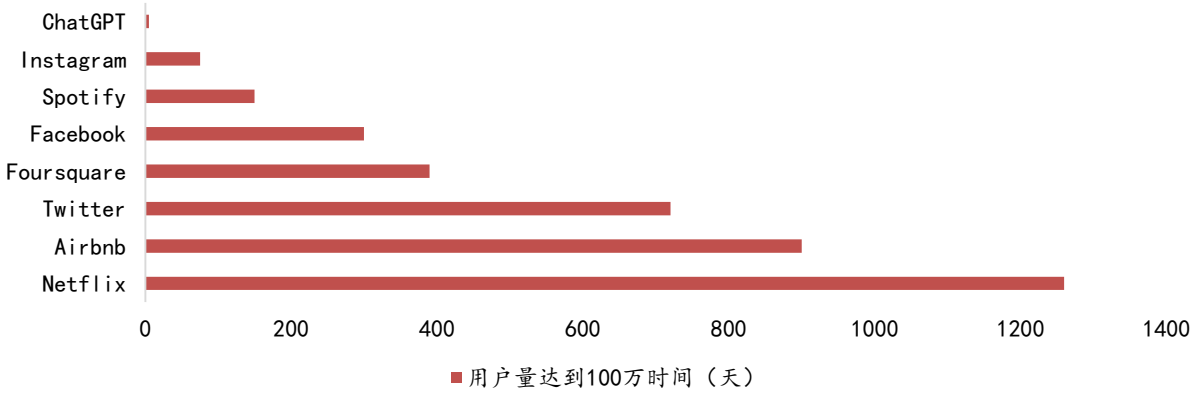
生成式AI： GPT模型迭代四大版本，进化速度不断提升

OpenAI创立于2015年12月，发布ChatGPT引燃AI行业热度。GPT系列是OpenAI打造的自然语言处理模型，采用以Transformer结构为核心的模型，其最大特点是使用了大量的未标注的语料进行无监督的预训练，然后在各种有监督的任务上进行微调。

OpenAI于2022年11月先后推出了GPT-3.5和ChatGPT，GPT-3.5使用了更新的语料进行预训练，而ChatGPT是基于GPT-3.5的对话机器人，能够根据用户的输入生成流畅、有逻辑的回答，以及完成撰写论文报告、翻译文字、编写代码等文本生成任务，并且能根据聊天的上下文进行互动。

ChatGPT发布后爆火，仅用5天时间用户量便破百万，推出2个月后用户量破亿，成为史上用户增长速度最快的消费级应用程序。3月14日，OpenAI进一步推出GPT-4.0，相比当前ChatGPT使用的GPT-3.5，增加了输入图像的功能；扩写能力增强，能处理超过25000个单词的文本；更具创造力，并且能够处理更细微的指令。**GPT模型迭代的参数量及训练量均呈指数级增长，使得AI从实验技术成长为稳定生产力。**

图1：ChatGPT仅发布5天便达到百万用户

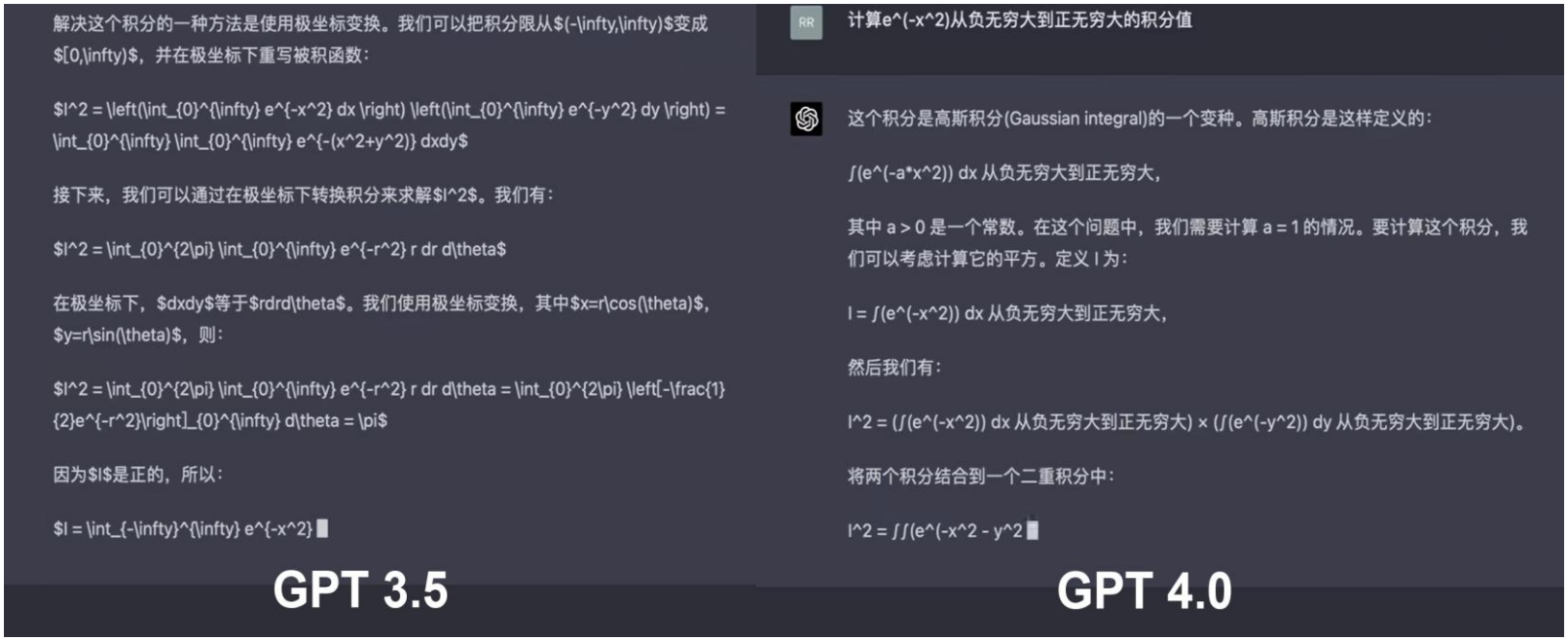


生成式AI： AI产品全面开花，生产力将达新高度

GPT模型稳定进步，AI已是成熟生产工具。从GPT-1到最新发布的GPT-4模型，其应用已经不仅局限于问答、阅读理解等文本处理，虽然目前GPT-4在现实场景中的能力可能不如人类，但**在各种专业和学术考试上表现出明显超越人类水平的能力，GPT-4在模拟律师考试中，分数排在前10%；相比之下，GPT-3.5的得分则在倒数10%附近。**

随着算力、算法、数据量的演进，行业内不断出现高质量的AI产品，微软New Bing、AI绘画、智能驾驶等等，体现出AI未来在多个领域的应用潜力。ChatGPT版Office、百度“文心一言”两大产品正式推出，或将AI的生产力推向新的高度。

图2：GPT 4.0 数学能力大幅提升



生成式AI： AI进化加速，数字经济未来已至

AI行业星辰大海，数字经济未来已至。从2018到2023年，四代GPT模型高速进步，从简单的问答、阅读理解、文本总结，到在众多测试中获得“人类级别表现”评级，此外近期AI衍生产品的层出不穷，显现出背后AI行业的星辰大海。

2020年，马斯克预言五年内人工智能将比人类更聪明，当前AI迭代进化的速度越来越快，虽然GPT还未通过图灵测试，距离真正的“智能”还有距离，但我们认为，AI达到人类水平、乃至超越人类的时代即将到来。

表2：历代GPT学习目标及表现情况

模型	发布时间	参数量	预训练数据量	学习目标	模型表现
GPT-1	2018年6月	1.17亿	约5GB	无监督语言模型（Pre-training）有监督fine-tune	在9/12任务中获得“先进”表现：问答、阅读理解、文本总结
GPT-2	2019年2月	15亿	40GB	多任务 零次学习Zero Short Task Transfer	在7/8任务中超过“先进”表现 随着模型参数变多，模型的表现呈现log-linear上升，没有到达瓶颈
GPT-3	2020年5月	1,750亿	45TB	语境学习 小样本学习	在小样本学习、单样本学习、零样本学习中表现突出
GPT-4	2023年3月	待公布		基于规则的奖励模型(RBRM)	**在GLUE, SuperGLUE, SQuAD等测试中获得“人类级别表现” 拥有图像处理能力**

生成式AI：算力、算法、数据三位一体

数据，通过算力，最后产生了算法或者应用。AIGC是人工智能、大数据、云计算、5G等多个技术领域的整合，是一种跨领域的合作发展模式。在AIGC行业中，算力、算法、数据是三个核心概念，它们共同构成了这个领域的基础设施。未来随着技术的进步和应用场景的不断拓展，这三个概念将继续发挥重要作用，推动整个行业的创新和发展。

- **算力（Computing Power）：**算力是指计算设备执行算法、处理数据的能力，包括CPU、GPU、FPGA、ASIC等。云计算技术和5G通信技术的发展使得算力的分布和调度更加灵活，有助于满足各种场景下对高性能计算的需求。
- **算法（Algorithm）：**算法是一系列解决问题、实现特定功能的有序指令和步骤。在AIGC行业中，算法是模型的基础，用于实现数据分析、人工智能模型训练等功能。
- **数据（Data）：**在AIGC行业中，数据是支撑决策和优化的基础，是算法发挥作用的前提。大数据技术可以对海量数据进行有效处理、分析和存储，而人工智能技术可以通过对数据进一步学习，实现各种智能化应用，如图像识别、自然语言处理等。

表3：AIGC行业三大核心概念

核心概念	描述	应用及关联技术
算力（Computing Power）	衡量计算设备执行算法、处理数据的能力，关系到系统的运行效率和任务完成速度。	数据中心、分布式计算、云计算、边缘计算、高性能计算（HPC）
算法（Algorithm）	解决问题、实现特定功能的有序指令和步骤，是计算机程序的基础，用于实现各种功能。	机器学习（ML）、深度学习（DL）、自然语言处理（NLP）、计算机视觉（CV）、推荐系统等
数据（Data）	对现实世界的描述和反映，以数字、文字、图像等形式表现，是支撑决策和优化的基础。	数据挖掘、数据分析、数据仓库、数据可视化、数据安全、隐私保护等

目录

01 生成式AI：ChatGPT引燃市场，数字经济未来已至

02 数据：大模型训练的基础资源

03 算力：大模型发展带来高算力需求

04 算法：大模型算法助力AIGC突破

05 产业应用：各领域应用加速落地，商业化前景广阔

数据：大模型训练的基础资源，需求不断扩大

数据是训练大模型的基础资源，以GPT系列模型为例，对比三代模型间使用的数据集，训练所需的数据集在质量和数量方面均不断提升。**随着人工智能模型迭代发展，高质量数据集的需求将进一步增长。**

表4：GPT系列模型训练使用数据集概要

模型	数据集概要
GPT-1	BooksCorpus（7000不同的未发表的书籍，包括冒险、幻想、浪漫等题材，数据集中包含大量连续文本）
GPT-2	在Reddit上爬取的外链，构建了WebText数据集，包含了这4500万个链接的文字子集，移除了所有的Wikipedia文档，因为它是很多下游任务的数据源，这是为了避免数据集重叠而影响评估
GPT-3	使用Common Crawl数据集(几乎包含整个互联网的数据)，进行了3步过滤操作，增加了一些高质量数据集，最终采用混合数据集输入。数据集大小合计将近5千亿tokens

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

数据：大模型训练的基础资源，需求不断扩大

从自然数据源简单收集取得的原料数据并不能直接用于有监督的深度学习算法训练，必须经过专业化的采集、加工，形成相应的工程化训练数据集后才能供深度学习算法等训练使用。

目前，带有监督学习的算法对于训练数据的需求远大于现有的标注效率和投入预算，基础数据服务将持续释放其对于算法模型的基础支撑价值。

表5：数据服务商部分公司概况

公司	主营业务	公司优势
海天瑞声	AI训练数据的研发设计、生产及销售业务	1.拥有的成品训练数据集数量大，在产品领域覆盖方面比较完善 2.已取得专利授权28项，计算机软件著作权159项，对比同业公司在专利技术储备方面具备一定优势 3.公司的产品和服务已获得字节跳动、阿里巴巴、腾讯、百度、科大讯飞、海康威视、微软、亚马逊、三星、中国科学院、清华大学等国内外客户的认可，市场认可度较高
澳鹏（Appen）	数据采集和标注解决方案	1.覆盖超过235个语种/方言，语言覆盖面具有优势 2、成立于1996年，经营历史较长，规模较大，拥有人工智能辅助数据注释平台，在全球170多个国家与100多万名专业承包合作 3.客户包括亚马逊、微软、谷歌等全球大型科技公司，产品质量得到认可
标贝科技	智能语音交互和AI数据服务	1.拥有语音合成模型和算法，可覆盖音乐类训练数据。拥有TOBI标注体系，通过自主研发的TTS评测系统，提供高质量的数据服务。 2.已与微软、百度、阿里、腾讯、京东、滴滴、字节跳动等国内外百余家企业客户建立合作，服务项目累计超过1000项