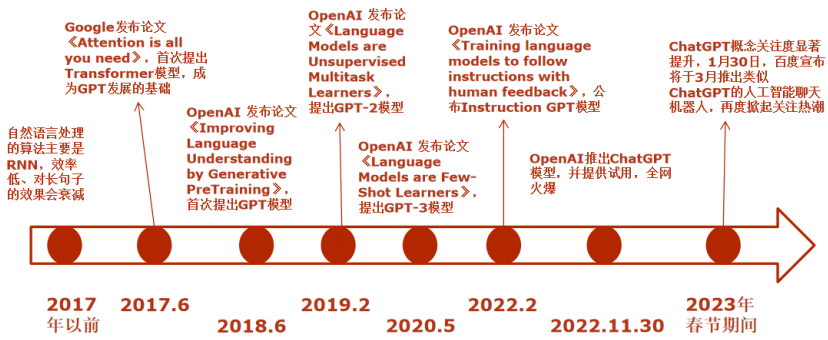
百度、三六零与科大讯飞：谁是 ChatGPT“国家队”？

2023 年 02 月 13 日

分析师：吕伟
执业证号：S0100521110003
电话：021-80508288
邮箱：lvwei_yj@mszq.com

➢ **从开源模型 GPT-2 迈向通用模型的 ChatGPT。**自 2017 年 6 月，Google 发布论文《Attentionisallyouneed》，首次提出 Transformer 模型，成为 GPT 发展的基础；2018 年-2020 年,OpenAI 基于 Transformer 模型发布多篇论文，并陆续提出 GPT-1、GPT-2、GPT-3 的三类语言模型，并在 2022 年 2 月发布论文《Training language models to follow instructions with human feedback》(使用人类反馈指令流来训练语言模型)，公布 InstructionGPT 模型，随后在 2022 年 11 月 30 日，OpenAI 推出 ChatGPT 模型，并提供试用。**仅仅不足 6 年时间，ChatGPT 走完从理论到现实的历程，其核心催化在于算法+数据+算力的共振。**

图1：ChatGPT 的发展历程



资料来源：openAI 官网，民生证券研究院整理

模型的进步是算法+算力的加持下，通过海量参数带来从量变到质变的升华。GPT 模型依托于 Transformer 解除了顺序关联和依赖性的前提，提出一个建设性的主张：先通过大量的无监督预训练(Unsupervisedpre-training)，再通过少量有监督微调（Supervisedfine-tunning），来修正其理解能力。整个算法模型包含三个步骤：1.人类反馈强化学习（RLHF）；2.收集参照参数并训练奖励模型；3.使用 PPO 算法进一步对 GPT 实现的内容进行强化学习加成，从人类偏好学习模型解决了强化学习对奖励机制保持一致的高度依赖。而复盘技术路径，算法模型在 2017 年时已被提出，从 GPT-1 到 ChatGPT 依然遵循 Transformer 的框架。**而真正带来升华的是在高性能算力加持下，通过优质数据的不断迭代演变而来。**

推荐

维持评级

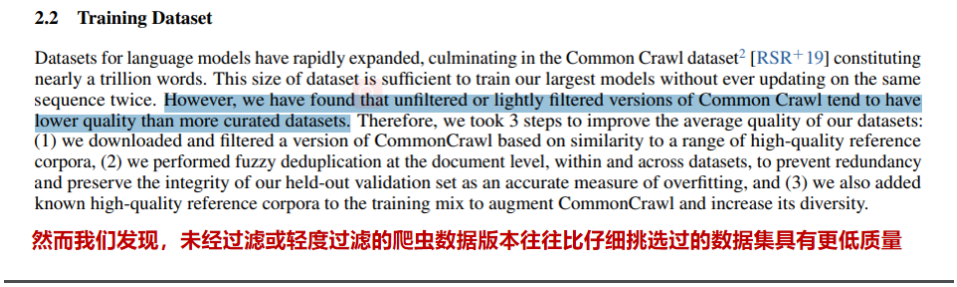
相关研究

1.计算机周报 20230212：ChatGPT 有望带来 5G 消息的业务重构-2023/02/12
2.计算机行业点评：计算机行业估值洼地：支付板块-2023/02/12
3.计算机行业事件点评：数字经济有望进入政策密集催化期-2023/02/07
4.计算机周报 20230205：预期差最大的主线：央企“搭台”，AI“唱戏”-2023/02/05
5.密码安全深度报告：密码：信创与数据安全皇冠上的“明珠”-2023/01/30

➢ **高质量的数据资源是推动 GPT 进化的重要抓手。**从 GPT-1 的 1.17 亿参数到 GPT-2 的 15 亿参数，再到 GPT-3 划时代的 1750 亿参数，OpenAI 依托筛选过的优质数据形成参数量的阶梯式上升，最终带来 GPT-3 乃至 ChatGPT 具备理解上下文、连贯性等诸多先进特征。

在提出 GPT-3 的论文《LanguageModelsareFew-ShotLearners》中，OpenAI 在收集近一万亿文字（参数）的数据库后，放弃直接使用海量数据训练模型，而是转向通过三种模式筛选优质数据进行训练，从而从万亿参数归纳出众人所熟知的 1750 亿参数，其核心原因在于**“未经过滤或轻度过滤的爬虫数据往往比筛选后数据集质量更低”。**

图2：放弃使用万亿数据集的原因



2.2 Training Dataset

Datasets for language models have rapidly expanded, culminating in the Common Crawl dataset[2] [RSR+19] constituting nearly a trillion words. This size of dataset is sufficient to train our largest models without ever updating on the same sequence twice. However, we have found that unfiltered or lightly filtered versions of Common Crawl tend to have lower quality than more curated datasets. Therefore, we took 3 steps to improve the average quality of our datasets: (1) we downloaded and filtered a version of CommonCrawl based on similarity to a range of high-quality reference corpora, (2) we performed fuzzy deduplication at the document level, within and across datasets, to prevent redundancy and preserve the integrity of our held-out validation set as an accurate measure of overfitting, and (3) we also added known high-quality reference corpora to the training mix to augment CommonCrawl and increase its diversity.

然而我们发现，未经过滤或轻度过滤的爬虫数据版本往往比仔细挑选过的数据集具有更低质量

资料来源：Tom B. Brown 等作者《LanguageModelsareFew-ShotLearners》，民生证券研究院

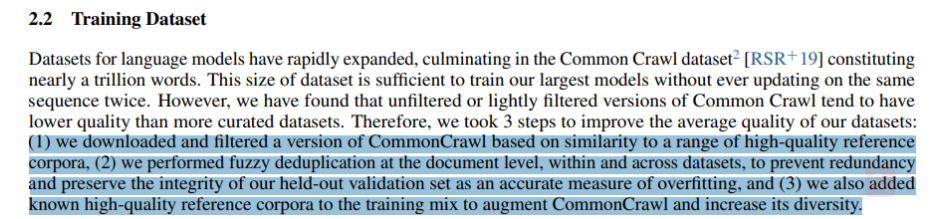
论文中所用的三种筛选数据模式如下：

1）根据与一系列高质量参考语料库的相似度比较，从而过滤出的爬虫数据；

2）通过对数据集内部和跨数据集的文档上执行重复数据的删除；

3）将已知的高质量参考语料库添加到训练组合中，以增强数据集的多样性。

图3：三种筛选数据模式



2.2 Training Dataset

Datasets for language models have rapidly expanded, culminating in the Common Crawl dataset[2] [RSR+19] constituting nearly a trillion words. This size of dataset is sufficient to train our largest models without ever updating on the same sequence twice. However, we have found that unfiltered or lightly filtered versions of Common Crawl tend to have lower quality than more curated datasets. Therefore, we took 3 steps to improve the average quality of our datasets: (1) we downloaded and filtered a version of CommonCrawl based on similarity to a range of high-quality reference corpora, (2) we performed fuzzy deduplication at the document level, within and across datasets, to prevent redundancy and preserve the integrity of our held-out validation set as an accurate measure of overfitting, and (3) we also added known high-quality reference corpora to the training mix to augment CommonCrawl and increase its diversity.

资料来源：Tom B. Brown 等作者《LanguageModelsareFew-ShotLearners》，民生证券研究院

根据 OpenAI 的设计，在筛选出的优质数据下，最终训练出的 GPT-3 成本极其高昂。**即使在团队明确发现失误的前提下，依然无法承担二次训练的代价，其本质原因在于优质数据的来源是 OpenAI 通过大量前期的工作筛选而成。**通过梳理，筛选后的数据主要分为：**1）过滤后的爬虫数据、2）WebText2 的数据集、3）一号图书馆数据、4）二号图书馆数据、5）英文版的维基百科**等五种。而将五类数据映射至国内，我们发现在互联网高歌猛进的建设中，我国天然具备五类数据的优质土壤。

图4：筛选后的不同种类数据在训练中的情况

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

Table 2.2: Datasets used to train GPT-3. "Weight in training mix" refers to the fraction of examples during training that are drawn from a given dataset, which we intentionally do not make proportional to the size of the dataset. As a result, when we train for 300 billion tokens, some datasets are seen up to 3.4 times during training while other datasets are seen less than once.

资料来源：Tom B. Brown 等作者《LanguageModelsareFew-ShotLearners》，民生证券研究院

➢ **以百度、360 和科大讯飞为代表的通用模型中国队，同时具备模型+算力+数据的天然属性。在模型上，**无论是 GPT-3、亦或是 ChatGPT，其底层的技术仍未跳出 2017 年 Transformer 模型的框架。1）三六零：根据 2 月 7 日，公司在互动平台上的回答，公司的搜索引擎团队及人工智能研究院从 2020 年开始一直在包括类 ChatGPT、文本生成图像等技术在内的 AIGC 技术上有持续性的研发及算力投入，目前公司的类 ChatGPT 技术的各项指标已实现强于 GPT-2 的水平，并在中文语境下实际效果强于 ChatGPT2；2) 百度公众号宣布将在 3 月上线类 ChatGPT 应用“文心一言”；3）在 NLP 所在的认知智能领域，科大讯飞主导承建了认知智能全国重点实验室（科技部首批 20 家标杆全国重点实验室之一），多年来始终保持关键核心技术处于世界前沿水平，并在去年获得 CommonsenseQA2.0、OpenBookQA 等 12 项认知智能领域权威评测的第一；4) 浪潮信息发布的源 1.0，作为人工智能巨量模型，单体模型参数量达到 2457 亿，超越美国 OpenAI 组织研发的 GPT-3 模型，成为全球最大规模的中文语料 AI 巨量模型。

结合 OpenAI 以非盈利的模式下，仅仅在一年多便从 GPT-2 升级到 GPT-3，我们判断百度、三六零以及科大讯飞为代表中国队，在模型上的差距有望在一定时间内实现追平。

➢ **在算力上，**OpenAI 的算力依托于微软为其推出的超级计算机，根据微软表示，最新与 OpenAI 和合作研发的这款超级计算机居于世界 Top5 之列。而根据 2022 年 6 月 1 日新华网的报道，2022 年上半年的全球超级计算机 500 强榜单中，中国共有 173 台超算上榜，上榜总数蝉联第一。同时，百度智能云落地新一代高性能 AI 计算集群，成为领先的 AI 原生云算力底座。研究人员可基于全新发布的实例组建上千节点规模的超高性能计算集群，成倍缩短超大 AI 模型的训练时间。经过百度内部 NLP 研究团队的验证，在这个网络环境下的超大规模集群上提交千亿模型训练作业时，同等机器规模下整体训练效率是普通 GPU 集群的 3.87 倍。**我们认为，即使国内厂商在单颗芯片的算力上无法达到欧美水准，但通过组建多个算力集群叠加多员工迭代的加持上，将进一步抹平算力上的差距。以三六零为例，截至 2022 年半年报，公司货币资金超 200 亿，2021 年研发投入超 30 亿，具备充足资金储备面对算力竞赛的格局。**

➢ **三大巨头具备国内海量优质数据的优势。**通过对 GPT-3 的五类数据分类，以百度、360 和科大讯飞为代表的国内厂商，天然具备优质数据的储存。**如百度和三六零**同时具备类似 CommonCrawl(filtered)的数据，对标 Reddit 的百度知道和 360 问答，以及对标维基百科的百度百科和 360 百科。两者更是国内搜索引擎第一与第二的龙头厂商，根据 2 月 7 日三六零在互动平台的回答，目前 360 搜索是

中国搜索引擎的 Top2，市场份额为 35%。海量数据存于自身，天然具备数据清洗和数据迭代的核心功能。而科大讯飞 AI 训练模型依托自身在医疗领域和教育领域的领军地位，通过教育领域的成绩单和题库，以及医疗领域大量的处方单和病例等专业数据支撑自身专业 AI 布局，形成专用领域数据闭环的功能。

表1：OpenAI 的五类数据在国内的类似模式

OpenAI 筛选后的优质数据类型	说明	国内对应类型
CommonCrawl(filtered)	过滤后的爬虫数据	百度爬虫和 360 搜索蜘蛛等
WebText2	来自从 Reddit 的大于 4500 万个网页的文本的筛选	知乎、豆瓣、百度知道、360 问答等
Books1	图书库	科大讯飞教育+医疗等专业数据、中国知网、万方、中国期刊网等
Books2	图书库	科大讯飞教育+医疗等专业数据、中国知网、万方、中国期刊网等
Wikipedia	英文版的维基百科	百度百科、360 百科等

资料来源：OpenAI，民生证券研究院整理

投资建议：百度和 360 作为国内前二的两大搜索引擎，具备海量通用数据之外，着重布局国家与科技巨头算力军备竞赛环节，**均构建算法+数据+算力三大核心能力，或将成为国内通用算法的领军企业。**科大讯飞通过在 NLP 方面长期的技术优势构建讯飞开放平台，提供超过 500 项 AI 产品及方案，并链接 500 万＋合作伙伴共建人工智能生态；结合自身在教育、医疗、翻译、金融和司法等专业领域的数据积累，**有望形成通用+专项模型的共振。**在 ChatGPT 带来业务逻辑质变重估的趋势下，充分看好三者在通用模型的核心竞争力，维持“推荐”评级：三六零、科大讯。由于 2022 年疫情反复带来订单交付延期和费用率上升等原因，对应调整相关公司盈利预测，三六零：预计 2022-2024 年归母净利润为-24.49/5.01/11.70 亿元，23-24 年对应 PE 为 104X/45X；科大讯飞：预计 2022-2024 年归母净利润为 5.57/17.32/27.59 亿元，23-24 年对应 PE 为 64X/40X。建议重点关注：百度集团-SW。

风险提示：技术落地不及预期，竞争格局加剧。

重点公司盈利预测、估值与评级

代码	简称	股价	EPS（元）			PE（倍）			评级
		(元)	2021A	2022E	2023E	2021A	2022E	2023E	
601360	三六零	9.37	0.13	-0.34	0.07	58	/	104	推荐
002230	科大讯飞	47.91	0.67	0.24	0.75	72	200	64	推荐
9888.HK	百度集团	121.8	3.51	2.63	5.48	32	46	22	未评级

资料来源：Wind，民生证券研究院预测;

（注：股价为 2022 年 2 月 10 日收盘价；未覆盖公司数据采用 wind 一致预；百度股价基于 2 月 10 日 1:0.8675 换算而成）

分析师承诺

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并登记为注册分析师，基于认真审慎的工作态度、专业严谨的研究方法与分析逻辑得出研究结论，独立、客观地出具本报告，并对本报告的内容和观点负责。本报告清晰准确地反映了研究人员的研究观点，结论不受任何第三方的授意、影响，研究人员不曾因、不因、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到任何形式的补偿。

评级说明

投资建议评级标准	评级		说明
以报告发布日后的 12 个月内公司股价（或行业指数）相对同期基准指数的涨跌幅为基准。其中：A 股以沪深 300 指数为基准；新三板以三板成指或三板做市指数为基准；港股以恒生指数为基准；美股以纳斯达克综合指数或标普 500 指数为基准。	公司评级	推荐	相对基准指数涨幅 15%以上
		谨慎推荐	相对基准指数涨幅 5%～15%之间
		中性	相对基准指数涨幅-5%～5%之间
		回避	相对基准指数跌幅 5%以上
	行业评级	推荐	相对基准指数涨幅 5%以上
		中性	相对基准指数涨幅-5%～5%之间
		回避	相对基准指数跌幅 5%以上

免责声明

民生证券研究院：

上海：上海市浦东新区浦明路 8 号财富金融广场 1 幢 5F；200120

北京：北京市东城区建国门内大街 28 号民生金融中心 A 座 18 层；100005

深圳：广东省深圳市福田区益田路 6001 号太平金融大厦 32 层 05 单元；518026