基于开源技术的AIGC与大语言模型的应用

达 观 数 据 联 合 创 始 人　张 健

达观数据
DATA GRAND

达观数据
DATA GRAND

智能文本机器人专家

达观数据—智能文本机器人专家

达观专注文本自动化业务

达观数据专注于智能化文本处理软件系统的开发，广泛应用于各类办公业务的智能化无人化，大幅度提高企业效率与智能化水平
自主研发的智能文本处理（IDP）、光学字符识别（OCR）、机器人流程自动化（RPA）、搜索推荐与知识图谱（KG）等系统，成功赋能百业

连续开创多项行业第一

行业首家
国家级专精特新“小巨人”企业
文本智能处理领域唯一入围

市占率第一
文本智能化处理相关产品
位居市场第一

案例数量第一
全行业自动化
文本处理应用场景10000+

成熟度最高
全行业首家软件成熟度
最高等级认证CMMI5

原创技术
国内首家成功研发
IDP平台和智能审阅产品

产品最高认证
中国信通院
最高等级认证

服务网络遍及全国

达观数据
DATA GRAND

智能文本机器人专家

自然语言处理与大语言模型

达观数据
DATA GRAND

如今我们所有的工作都围绕文字在开展

撰写方案

回答问题

审核材料

翻译内容

等等

能否自动化来处理文字呢?
这样就省时省力多了

自然语言处理技术（Natural Language Processing，简称NLP）就是计算机学术界专门研究如何理解文字的专项技术

NLP从1950年代开始发展，历经70余年，无数科学家贡献了聪明才智

文字处理非常难 被誉为“人工智能皇冠上的明珠”

达观数据
DATA GRAND

狗

动物，犬科，宠物，四条腿，嗅觉灵敏，有毛，会叫，由狼驯化

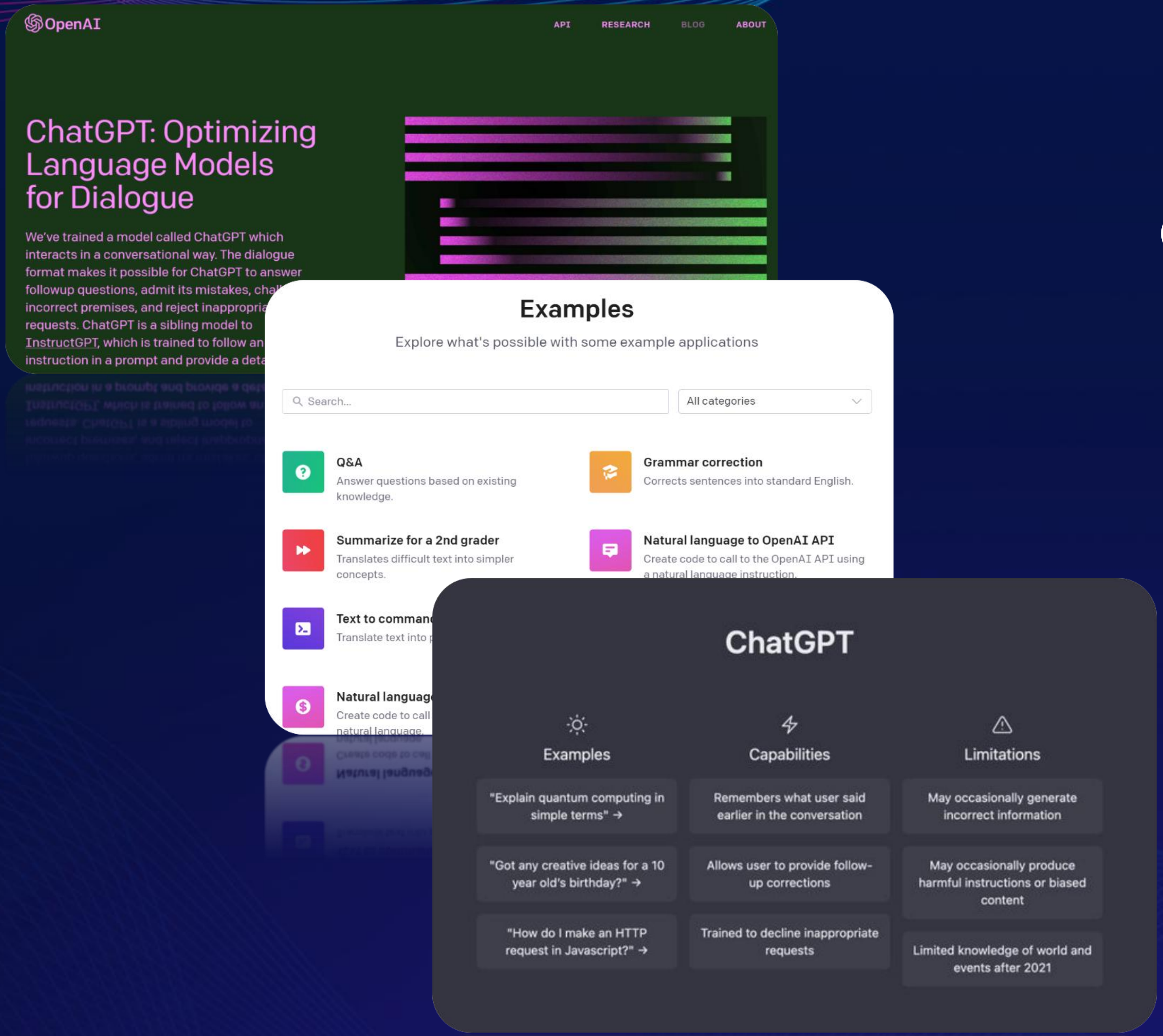
哈士奇，柴犬，京巴，德牧，边牧，阿拉斯加雪橇犬，贵宾，比熊

忠诚，友善，陪伴，看家护院，可爱，会咬人，狂犬病

狗眼看人低，狗皮膏药，狗改不了吃屎，狗仗人势，狐朋狗友

嗅觉2万类，听觉120万赫兹，晶状体2倍厚，5.3亿神经细胞，42颗牙齿，体温38度，色盲，前足5只脚趾、后足4只脚趾

等等数以万计的相关知识



ChatGPT的解决路径：一独三巨

- “一独三巨”：全球独树一帜的技术路线+巨大的参数规模+巨大的算力消耗+巨大的训练数据量
- 由OpenAI公司从2016年起历经7年，耗资20亿美元，经历若干次失败后，成功推出的划时代的人工智能系统
- 2022年11月30日发布，因其超越以往任何产品的卓越效果风靡全球，仅发布第5天注册用户数就超过100万，60天月活过亿

从GPT的名称中一窥究竟

GPT

- **Generative** 生成式：写作能力是其开发的出发点和核心能力
- **Pretrained** 预训练：能力来自于对大量文档的预先学习
- **Transformer** 转换器：对文字中知识的一种极为有效的提炼算法


Sam Altman
@sama

language models just being programmed to try to predict the next word is true, but it's not the dunk some people think it is.

animals, including us, are just programmed to try to survive and reproduce, and yet amazingly complex and beautiful stuff comes from it.

OpenAI创始人Sam Altman：简单的思考下一个词是什么的过程，这个简单的目标训练出了人类的智能

达观数据
DATA GRAND

大模型LLM与ChatGPT的关系

大模型LLM（Large Language Model）

对大规模语言资料的建模能形成强大的文字语义理解能力

- GPT是众多LLM大语言模型中的一种，也是目前被证明处理文字最成功的一种模型
- ChatGPT是面向对话（Chat）场景定向优化后的GPT子版本
- 未来针对其他应用场景还会出现更多的X-GPT系统，甚至更多其他的LLM语言模型

GPT（Generative Pre-Training）家族的发展历程

GPT

- 无监督预训练+有监督微调
- 基础的文本生成能力

GPT-2

- 用Prompt形式统一多项NLP任务为生成任务
- 多任务训练+海量数据(40GB)

GPT-3

- 参数规模急剧扩张，训练样本提炼
- 支持48项特定类型的任务

InstructGPT

- 引入人类反馈强化学习（RLHF）效果明显提升
- 逻辑推理能力等提升

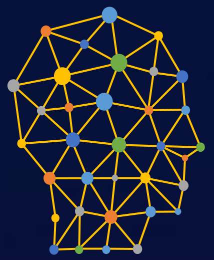
模型名字	发布时间	最大参数规模	训练数据	模型架构
GPT	2018 年	约 1.2 亿 (117M)	BooksCorpus	12 层 Transformer Decoder
GPT-2	2019 年	约 15 亿 (1542M)	WebText（约 40G 文本）	48 层 Transformer Decoder
GPT-3	2020 年	约 1750 亿	Common Crawl (filtered)、WebText2、Books1、Books2、Wikipedia（共约 5000 亿词）	96 层 Transformer Decoder
InstructGPT	2022 年	约 1750 亿	数万提示文本及其生成结果标注	GPT-3 + 基于人类反馈的强化学习 (RLHF)
ChatGPT	2022 年	与 GPT-3 相当	额外的标注数据（具体信息未知）	与 InstructGPT 类似

从GPT-3.5升级到GPT-4

- GPT3.5于2022年11月发布，GPT4于2023年4月发布，GPT5预计于2023年底前发布
- 技术的进步正在加速，每半年有一次大的技术飞跃
- 新的能力正在不断被开发出来，每次迭代让效果有巨大提升

大语言模型行业应用开源情况

实现人工智能的 三要素



智能系统

算法 模型调优的算法细节 操作步骤 优化技巧 尝试

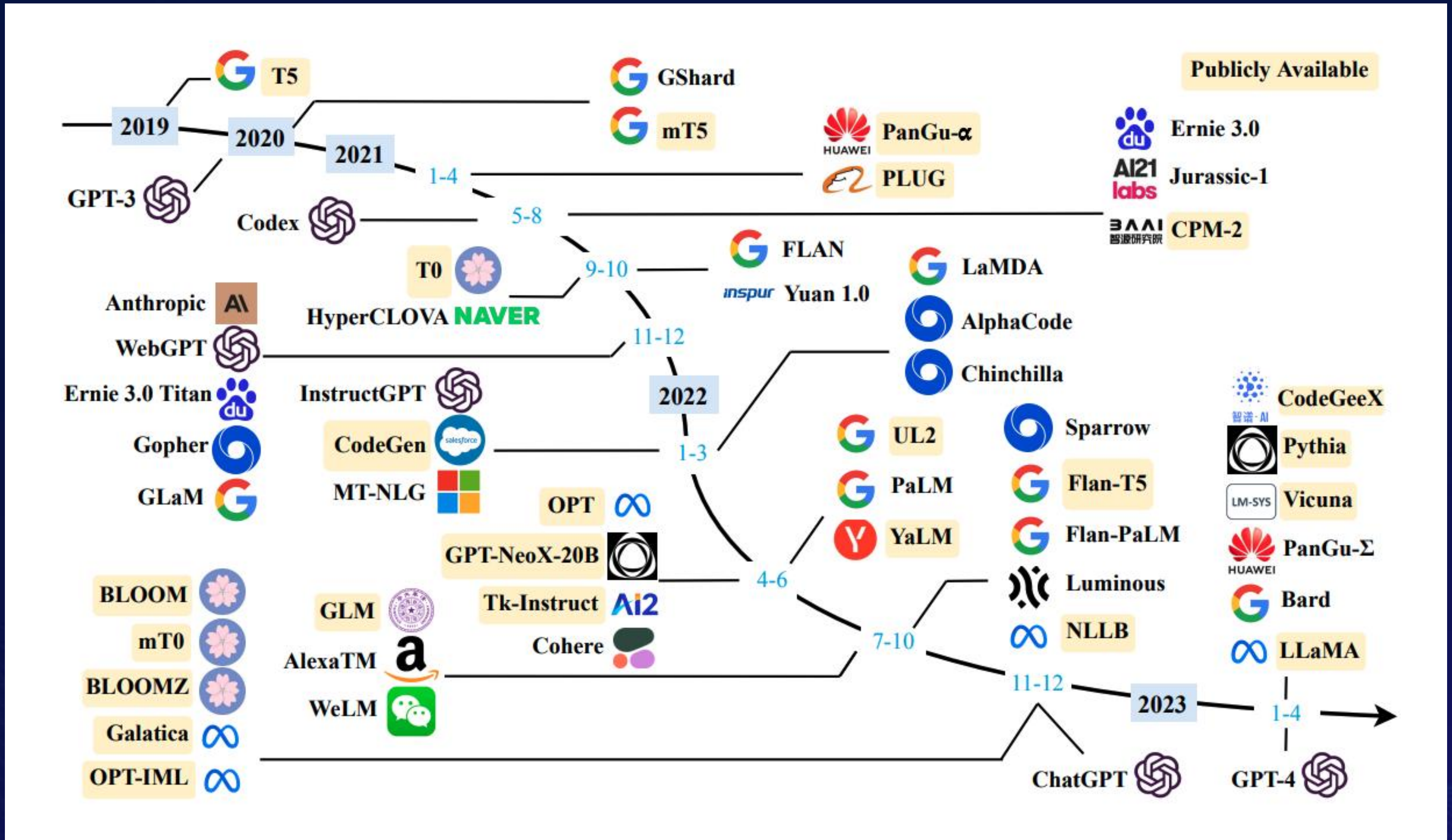
数据 足够丰富 多样 优质的高质量文本资料 稀缺

算力 组建多机多卡的高性能运算集群 工程挑战

人才

大模型开源生态-算法

达观数据
DATA GRAND



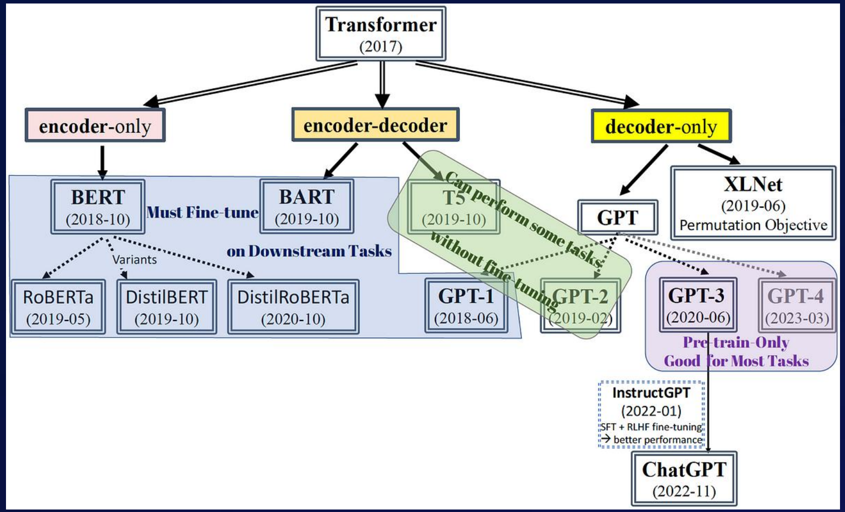
大模型技术蓬勃发展

- 商用模型：除了openAI以外，谷歌、Meta等也在加速布局AI大模型。谷歌宣布正式开放Bard的公测，Meta宣布将推出“Meta人工智能大型语言模型”系统“LLaMA”
- 开源模型：标黄的模型为开源模型

大模型开源生态-算法

达观数据
DATA GRAND

基础模型结构



开源模型底座

开源模型

预训练模型

- GPT-Neo
- GPT-J
- CPM-Bee
- T5
- LLaMA
- BLOOM
- ... ...

指令微调模型

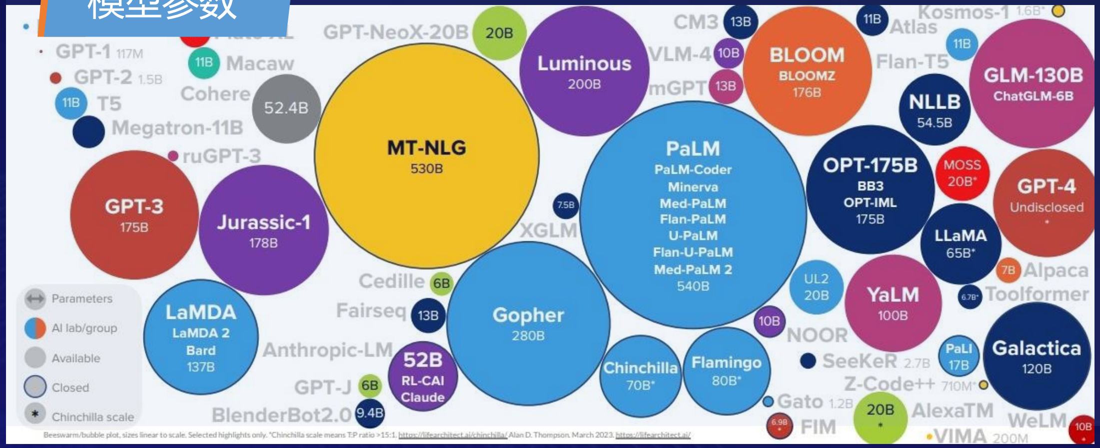
- BLOOMZ
- FLAN-T5
- Galactica
- T0
- ... ...

模型训练框架

NVIDIA
deepspeed
Alpa
BMTrain
Colossal-AI

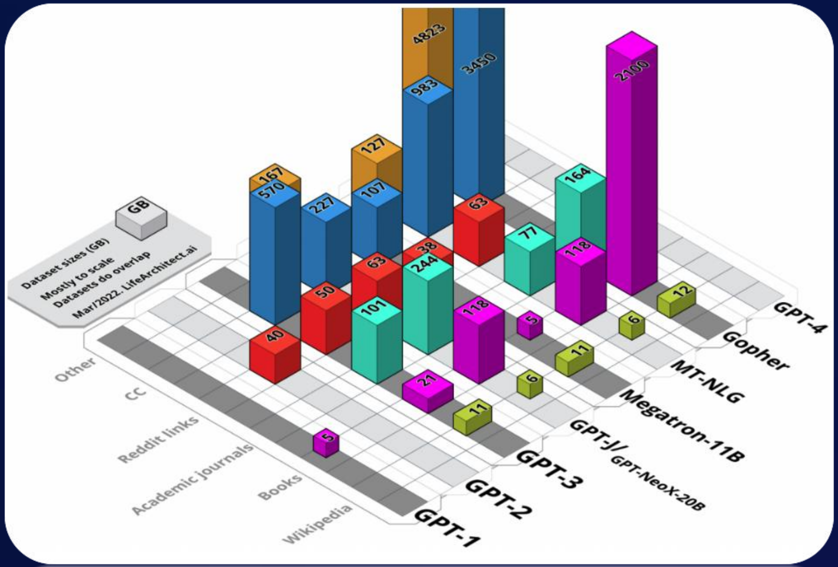
- Megatron
- DeepSpeed
- Alpa
- Colossal-AI
- BMTrain
- ... ...

模型参数



大模型开源生态-数据

达观数据
DATA GRAND



Pile数据集的分布情况

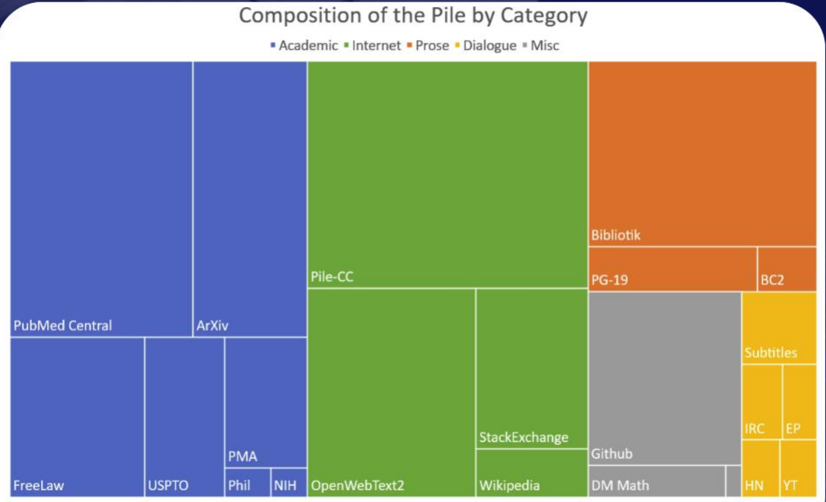


Figure 1: Treemap of Pile components by effective size.

通用大模型的数据集研究

- 高质量的文字资料是主要的训练数据，如教科书、论文、百科、报告、小说、新闻、优质问答等，低质量的文字反而对模型结果有伤害
- 特定类型的数据能训练获得独特的能力，例如使用大量的程序代码被证明可以成功获得思维链（CoT能力， Chain-of-Thought ）
- 结合历史数据增长率和数据使用率，有研究预计互联网上可用的数据资源尤其是高质量语言数据2026年可能会耗尽
- 根据大模型的构建阶段，需要获取春预料、指令学习和偏好对齐等数据

纯语料

指令学习数据

偏好对齐数据

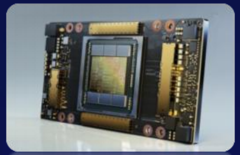
大模型开源生态-算力

达观数据
DATA GRAND

算力类型

- CPU
  整数运算

- GPU
  向量运算
  - 消费级GPU
    nVidia 3090 4090


  - 高性能GPU
    nVidia A100 H100


- nVidia 的GPU提供的矩阵向量运算能力 为各类AI的应用提供必不可少的底层基础
- 美国政府于2022年8月底发布出口禁令，禁止向中国出口A100及其他性能更高的GPU，为此nVidia推出了性能阉割后的中国特供版A800

美国禁止英伟达和AMD向中国出口高端GPU和AI计算芯片

发布日期：2022-09-01　来源：盖世汽车　浏览量：1051　扫描到手机　关注公众号

8月份，美国限制了最先进的芯片设计软件的出口，以遏制中国在本地生产尖端芯片的努力。两家领先的美国半导体设备制造商Lam Research和KLA的首席执行官也在财报会议上证实，美国已进一步限制出口可用于制造比14纳米更先进的芯片的机器。

目前，芯片制造商英特尔、台积电和三星正在竞相利用最新的芯片生产技术生产3纳米芯片。更小纳米尺寸的芯片功能更强大，但开发起来也更具挑战性。

根据乔治敦大学（Georgetown University）安全与新兴技术中心6月份的一份研究报告，在记录中可以确定的97款AI芯片中，几乎都是由英伟达、AMD、英特尔和Microsemi设计的。

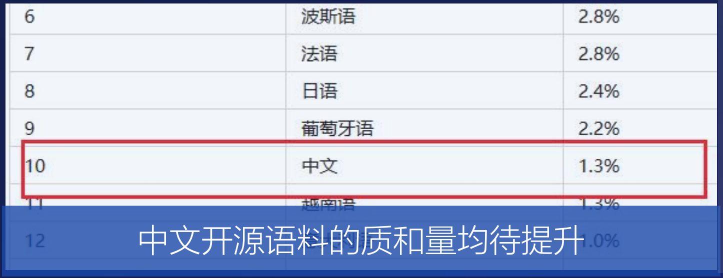
GPU型号	英伟达A100 80GB SXM	英伟达H100 SXM	英伟达A800 80GB SXM
制程工艺	7纳米	4纳米	7纳米
FP32下算力	19.5 TFLOPS	67 TFLOPS	19.5 TFLOPS
显存带宽	2039 GB/s	3.35 TB/s	2,039 GB/s
互联	600 GB/S	900GB/s	400GB/s
可用性	受限	受限	特供

国内大模型开源技术发展的挑战

局限性

相关技术对国外依赖度较高

国内的开源模型比较依赖国外的框架和与训练基座



6	波斯语	2.8%
7	法语	2.8%
8	日语	2.4%
9	葡萄牙语	2.2%
10	中文	1.3%
11	越南语	1.3%
12	[illegible]	[illegible]

中文开源语料的质和量均待提升

中文环境的语料积累不足，特别是对模型构建有效的高质量语料规模较小

基础硬件卡脖子，信创硬件性能待提升

英伟达高端GPU对中国供应受到限制。国产算力生态尚不完善

解决方式

机构合作研发

- 高校、机构和开源社区展开深度合作
- 对相关技术进行统一规划倡导，支撑学界和工业界研究
- 加快大模型分布式训练研究

共享优质数据

- 对数据进行安全性处理后，充分共享到社区和机构
- 对语料的形式、形态和处理标准形成统一标准

优质算力共享

- 建立算力联盟，促进优质算力共享
- 构建更大规模的显卡集群，容纳更大的模型
- 完善信创硬件生态，鼓励开源社区的信创支撑

达观”曹植“大语言模型

达观数据
DATA GRAND

达观研发 曹植® 大语言模型

垂直、专用、国产的 GPT 模型

AIGC
AI GENERATED CONTENT

智能写作
INTELLIGENT WRITING

语义分析
SEMANTIC ANALYSIS

达观“曹植”模型 强调 “垂直+专用+国产”

达观数据
DATA GRAND

垂直

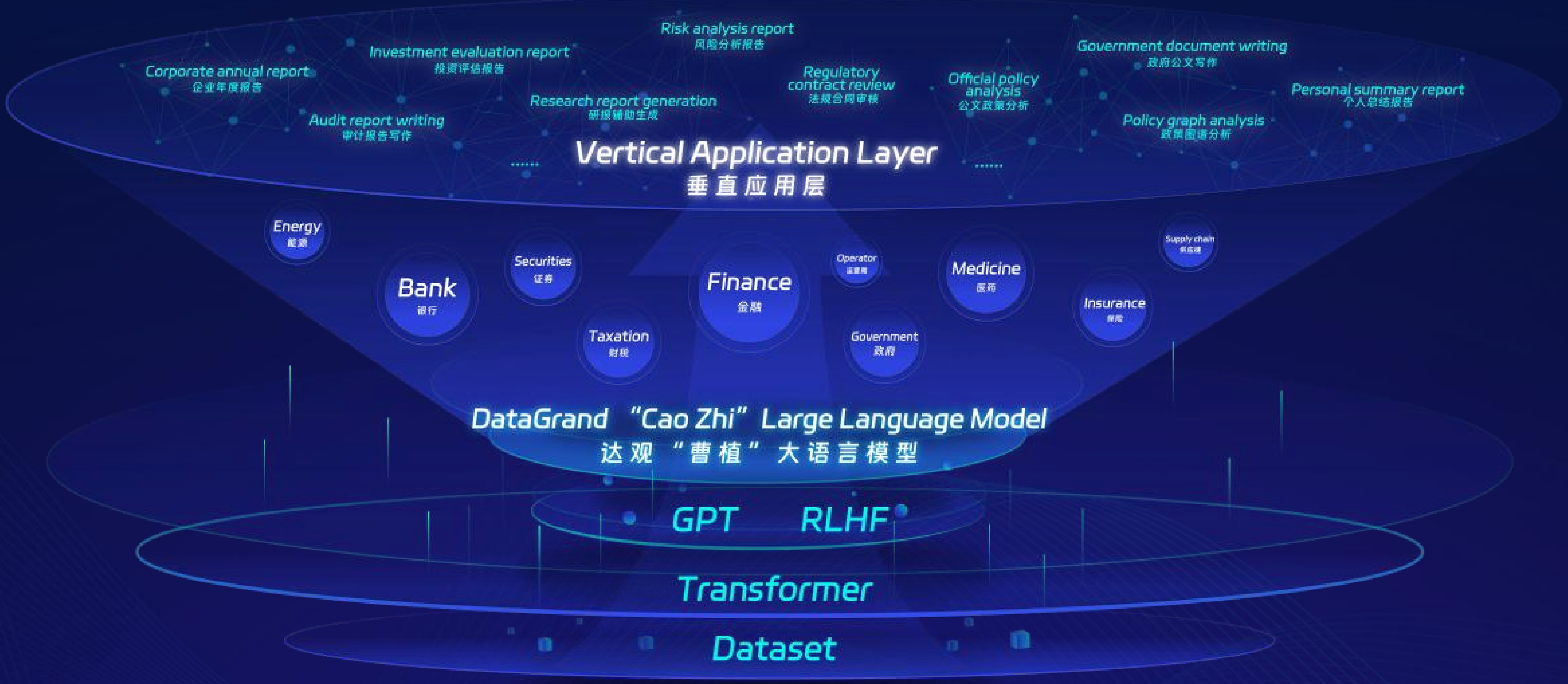
针对金融、工业、财税、政务、能源等垂直行业来开发特定应用

专用

系统可以为每个客户量身定制、私有化部署，确保数据安全私密

国产

坚持原创自主，训练数据和算法模型自主可控

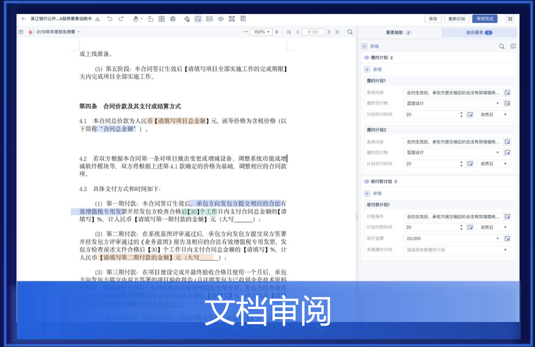


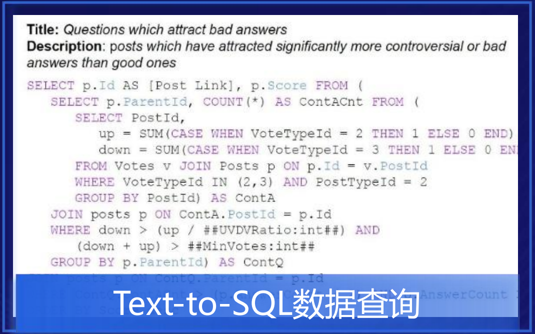
达观数据
DATA GRAND

智能文本机器人专家

针对垂直场景下的各类文本处理需求开发相应自动化系统

曹植的应用





达观数据
DATA GRAND

智能文本处理技术一定将创造巨大价值

[宋] 毕昇

- 如同一千年前的活字印刷术那样，将大幅提升社会效率
- 10年内AI能代替90%以上的日常文字处理工作
- 更多创新的智能应用在未来几年将层出不穷

[现代] 科技

400-175-9889