

极客邦科技双数研究院

InfoQ^{LAB} 研究中心

大语言模型综合能力

测评报告2023



InfoQ^{LAB} 极客传媒

1

桌面研究

通过对行业公开信息进行桌面研究，资料包括但不限于专业机构学术论文、文章资料、论坛讨论、研究报告、相关厂商产品介绍、相关专家公开演讲内容等。

2

专家访谈

InfoQ 研究中心针对本次研究定向邀请了国内外的相关专家进行访谈。

3

InfoQ 分析

结合桌面研究和专家访谈进行观点沉淀和交流，并经由报告形式对外展示。

目录

CONTENTS

01 | 大模型发展背景

02 | 大模型产品特征和核心能力

03 | 大模型产品测评结果和特征

04 | 大模型产品未来发展展望

01

大模型发展背景



大语言模型发展经过三阶段在2023年进入爆发阶段

大语言模型诞生阶段

2017

- 谷歌推出用于处理自然语言任务的 Transformer 神经网络架构

2018

- OpenAI 发布 GPT-1

大语言模型探索阶段

2019

- OpenAI发布GPT-2并部分开源
- 谷歌推出BERT模型

2020

- 百度推出可以准确理解语义的 ERNINE2.0

2021

- OpenAI推出能实现文本生成图像的DALL-E模型
- FaceBook推出CLIP模型
- 华为正式发布盘古大模型
- OpenAI推出Codex

大语言模型爆发阶段

2022

- OpenAI 推出 ChatGPT-3.5

2023

- 微软基于ChatGPT发布New Bing
- FaceBook发布LLaMA-13B
- 谷歌发布Bard以应对ChatGPT
- 复旦团队发布MOSS
- OpenAI发布GPT-4并实现图像识别
- 百度文心一言发布
- 微软宣布将GPT-4接入Office全家桶
- 通义千问、盘古NLP、天工3.5、星火等国产大模型陆续发布
- 谷歌更新Bard并推出PaLM 2模型
- 微软宣布Windows系统全方位集成Copilot

国内外厂商齐发力，大语言模型产业规模可观

国外

基础模型

Google LaMDA T5 PaLM Imagen PaLM-E Flan Stanford University Stanford Alpaca	Google DeepMind Gopher Chinchilla Gato databricks Dolly 2.0	Meta LLaMA MMS OPT-175B LIMA-65B AI21 studio Jurassic-1 Jumbo	OpenAI GPT-4 DALL·E2 CodeX Claude	BigScience Bloom T0 BloomZ GPT-J 6B	stability.ai Stable Diffusion StableLM LMSYS ORG vicuna-13b
--	--	---	---	--	---

ChatBot

Bard BingChat ChatGPT Claude

其他应用

Notion AI Cedille AI Copilot Colab Copilot
--

国内

基础模型

百度 文心 通义 盘古 国家超级计算天津中心 天河天元大模型	二郎神 浪潮 源1.0 言犀 开放平台 基础模型 星火 科大讯飞 iFLYTEK	孟子 日日新 腾讯 混元 玉言 自研大模型 云从科技 CLOUDWALK
--	---	---

ChatBot

ChatGLM ChatJD 从容 MOSS 商汤 SenseChat 天工 文心一言 360智脑

其他应用

钉钉 斜杠 WPS AI 万兴科技 学而思网校 MathGPT 受益一生的能力 雪湖·海若 METASOTA 写作猫	出门问问 序列猴子 EMOTIBOT FRIDAY 有道 youdao 子曰 达观数据 曹植 DATA GRAND 知乎 知海图AI 小冰
---	--

大语言模型研发的关键影响要素

大语言模型产品研发需要同时具备三大要素，分别为数据资源要素、算法和模型要素、资金和资源要素。InfoQ研究中心分析目前市场中的产品特征，数据资源、资金和资源两要素为大模型研发的基础要素，即必要不充分要素。

虽然数据、资金资源为大语言模型研发设置了高门槛，但对于实力雄厚的大型企业仍然是挑战较小的。算法和模型是目前区分大语言模型研发能力的核心要素。算法和模型影响的模型丰富度、模型准确性、能力涌现等都成为评价大语言模型优劣的核心指标。

算法和模型

- 模型训练的形式
- 涌现出的思维链
- 基于人类反馈的学习模型

数据资源

- 模型参数规模
- 训练模型次数
- 训练使用的数据集



基础要素



核心要素

资金和资源

- 顶级开发者构建的大规模研发团队
- 昂贵的GPU背后的大规模资金投入



基础要素

大语言模型训练之需要足够“大”

百亿参数是入场券

GPT-3和LaMDA的数据显示，在模型参数规模不超过100亿-680亿时，大模型的很多能力（如计算能力）几乎为零。

大量计算触发炼丹机制

根据NVIDIA 研究论文里的附录章节显示，一次迭代的计算量约为4.5 ExaFLOPS，而完整训练需要9500次迭代，完整训练的计算量即为430 ZettaFLOPS（相当于单片A100跑43.3年的计算量）。

大量且丰富的数据集

常见的数据集包括GSM8k、USSE、MMLU、HumanEval等。



O1. 模型参数规模

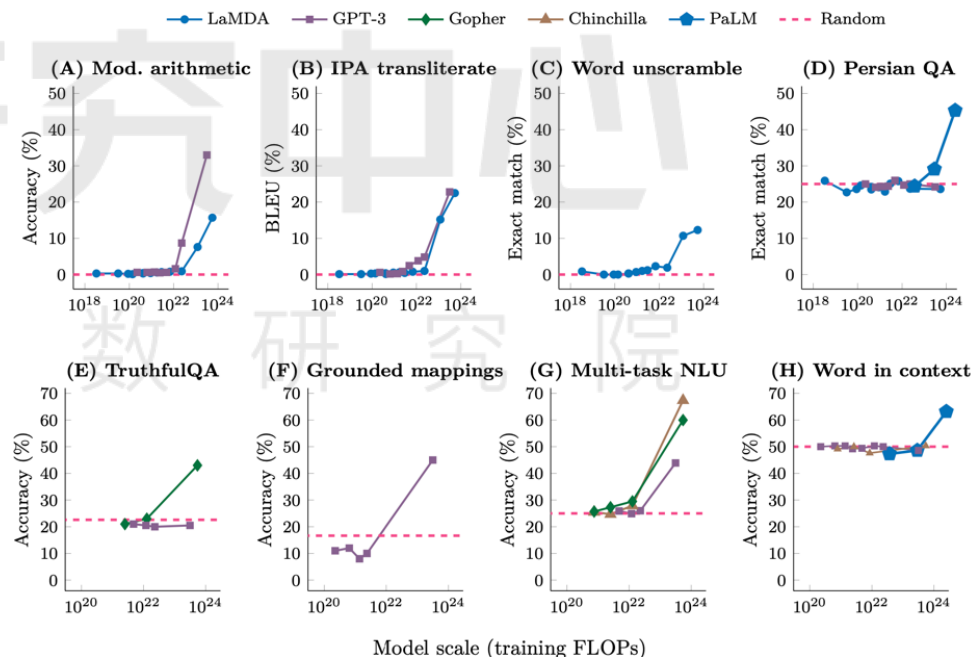


O2. 模型计算量



O3. 训练使用数据集

- 模型在参数规模达到一定程度后，性能首先得到急剧提升，同时涌现许多新的能力。特别是任务所训练的模型适用于更多以前未经训练的任务。涌现对大型模型应用的影响非常重要，只有通过这种能力，我们才能高效地实现模型的泛化，并实现模型的迁移。



数据来源：Sparks of Artificial General Intelligence Early experiments with GPT-4

大模型训练参数规模量级最高或达5万亿以上

- 国内大模型出现大量参数规模大于100亿的模型
- 百度研发的Ernie和华为研发的盘古目前是有数据的国内大模型参数规模的领先者
- 国际领先的大模型GPT-4据推测参数规模量级可达5万亿以上

国内未公布参数规模

- 自研大模型（字节）
- 1+N认知智能大模型（科大讯飞）
- 二郎神模型（IDEA研究院）
- 自研AI大模型（光年之外）
- 自研大模型（燧原科技）
- 超拟人大模型（聆心智能）
- 自研大模型（香依科技）
- 魔力写作（竹间智能）
- 自研大模型（MiniMax）
- 蛋白质大模型（浙江大学杭州国际科创中心）

国内模型参数规模 <100亿

- 书生3.5（商汤科技）
- 孟子（澜舟科技）
- DriveGPT（毫末智行）
- ChatGLM（清华大学）

国内模型参数规模 >100亿

- ERNIE 3.0（百度）
- 盘古（华为）
- MOSS（复旦大学）
- 遵义（阿里）
- 言犀（京东）
- 混元（腾讯）
- 伏羲（网易）
- 源1.0（浪潮信息）
- 行业精灵（云从科技）
- 八卦炉（达摩院）
- 元语大模型（莫塔社区）
- 曹植大模型（达观数据）
- 紫东太初（中科院自动化研究所）
- 自研大模型（西湖星辰）
- 悟道2.0（智源研究院）

国际模型参数规模

- GPT-4（OpenAI）
未公开，推测为超过50000亿
- PaLM（Google）5400亿
- BERT（Google）4810亿
- GPT-3.5（OpenAI）1750亿
- LaMDA（Google）1370亿
- Galatica（Meta）1200亿
- LLaMDA（Meta）650亿
- Chinchilla（DeepMind）700亿
- Claude（Anthropic）520亿
- Mineva（Google）5400亿

资料来源：民生证券研究院和wiki百科

算法和训练模型水平主导大语言模型的能力表现

模型训练技术（举例）

Prompt-tuning

使用自然语言提示（prompt）的方法，以指导模型生成特定的输出。这种方法的目的是通过对模型进行定向训练，使其在特定任务上表现出更好的性能。

Instruction-tuning

通过为模型提供任务相关的指令来指导模型学习的方法。这种方法的目的是使模型更好地理解任务的要求，并提高其生成能力和上下文理解能力。

Chain of Thought

通过分解训练过程为较小的相互关联的任务来训练模型的方法。这种方法的目的是使模型能够理解和维护文本中的思维链，从而生成连贯的、上下文相关的响应。

Human Feedback

通过人类给予反馈对模型形成奖励机制，帮助模型进行强化学习的训练。这种方法可以在预训练模型和产品投入市场后持续获得反馈，帮助模型增强判断力。

训练方式
工程化

基础模型

训练方式直接决定大模型产出的效率，根据已经公开的论文解读，现有优秀模型训练方式呈现高度工程化特征。

工程化训练方式主要呈现三个特征：

- 1、详细而严格的规则：对于如何处理数据和什么是高质量数据等给出详细和严格的执行和判断的方法论；
- 2、明确定义标注意图：如详细说明标注原因，并要求如果标注人员不能完全理解，则迅速跳出流程；
- 3、团队培训和考核机制完善：通过李克特评分等方式，持续保证团队处在目标水准以上。

自研闭源元模型

自研开源元模型

在开源模型基础上微调的模型

自研闭源元模型：典型代表包括Open AI的GPT3.5、GPT 4等，国内厂商百度的原模型ERNIE3.0、华为的元模型PanGu- Σ 等。

自研开源元模型：典型代表包括Open AI的GPT2、Google的BERT等。

在开源模型基础上微调的模型：典型代表包括清华大学的ChatGLM-6B、商汤科技和华中科技大学开源中文语言模型骆驼 Luotuo等。

欢迎访问：电子书学习和下载网站 (<https://www.shgis.com>)

文档名称：大语言模型综合评测报告2023.pdf

请登录 <https://shgis.com/post/1785.html> 下载完整文档。

手机端请扫码查看：

