编 者 *

黄铁军　文继荣　刘知远　何晓冬　黄松芳
李　笛　戴一鸣　贾　伟　袁　莎　张冬敏
张大魁　李静云　李梦佳　闫亚琼　梁杨晓
马瑞军　沈林杉　牛梦琳

* 排名不分先后，具体贡献见文末

目 录 Contents

摘要 **1**

前言 **3**

第一章 大模型技术发展情况及趋势 **7**

一、超大规模智能模型发展历程 7

二、超大规模智能模型的发展意义 9

（一）推动AI科研的发展 9

（二）推动智能产业的发展 14

三、超大规模智能模型的技术发展趋势 17

（一）多语言预训练模型 17

（二）多模态预训练模型 18

（三）多任务预训练模型 18

（四）视觉预训练大模型 19

（五）物理世界模型 19

四、典型的大模型案例 **19**

（一）国际 21

1.GPT-3 21

2.Switch Transformer 21

3.DALL·E/CLIP 22

4.ViT 24

（二）国内 24

1.悟道（1.0/2.0） 24

2.盘古NLP/盘古α 26

3.阿里M6/PLUG 26

4.ERNIE 3.0 28

总结 **29**

第二章 大模型产业落地模式 30

一、API模式 31
二、开源模式 34
三、PaaS模式 36
四、“交互主体化”模式 40
总结 42

第三章 大模型商业化主要发展领域 43

一、主要落地领域 43
二、国际主要大模型落地领域和场景 44
(一) BERT 44
(二) GPT-2 45
(三) T-NLG 45
(四) GPT-3 46
三、我国主要大模型落地领域和场景 49
(一) 悟道(1.0/2.0) 49
(二) 盘古系列大模型 51
(三) 阿里M6/PLUG 51
(四) 百度ERNIE模型 52
四、基于大模型的落地应用案例(以GPT-3为例) 52
(一) Copilot 52
(二) AI Dungeon 55
(三) CopyAI 56
(四) Zebrium 57
(五) Maslo 57
五、大模型产业链 59
(一) 上游产业 59
(二) 中游产业 59
(三) 下游产业 60
总结 60

BAAI 超大规模智能模型产业发展报告

第四章 大模型领域存在的问题和解决案例 61

一、大模型存在的问题 61
(一) 模型在可解释性和可控性方面依然无法达到要求 61
(二) 算力门槛要求高 61
(三) 缺乏产业落地配套的工具 61
(四) 大模型相关的监管体系尚未建立 62
(五) 大模型产生的环境和伦理问题 62
(六) 大模型效果评估问题 63
二、解决案例 63
总结 65

第五章 下一步工作建议 66

一、持续推进大模型及相关技术体系的发展 66
(一) 研发更为绿色环保的模型技术 66
(二) 推广低门槛的大模型开发和使用工具 66
(三) 建立动态、面向应用场景的大模型评价体系 66
(四) 建立创新型大模型技术体系 67
二、解决模型存在伦理环保问题和版权争议 67
(一) 评估数据集和训练方法,降低模型偏见 67
(二) 建设低碳绿色的智能算力集群 67
(三) 针对大模型建立专门的开源许可证体系 67

报告贡献情况 68

关于智源研究院 69

摘　要

主要观点：

1 大模型（也称预训练模型、基础模型等）是“大数据+大算力+强算法”结合的产物，是凝练了大数据内在精华的“隐式知识库”，是实现多种人工智能应用的通用载体。大模型是连接人工智能技术生态和产业生态的桥梁，向下带动基础软硬件发展，向上支撑智能应用百花齐放，是未来整个人工智能生态的核心。

2 大模型通常是在大规模无标注数据上进行训练，学习出一种特征和规则。基于大模型进行应用开发时，将大模型进行微调（在下游特定任务上的小规模有标注数据进行二次训练），或者不进行微调，就可以完成多个应用场景的任务。

3 当前大模型领域较为关注的研究重点包括：大规模的模型参数、创新的模型架构、采用多模态的数据进行预训练和微调、高效且性能更强的预训练和微调方法、模型训练和推理加速方法等。

4 大模型在推动AI领域技术体系、智能产业发展的同时，也对算力基础设施和智能产业的生态提出了更高的要求，带动相关领域的发展。

5 当前全球大模型产业落地仍处于早期探索阶段，大模型落地过程中，最为重要的是减少企业需要投入的学习、开发和计算成本，同时还要能够满足企业进行模型开发定制的需求。目前产业领域典型的有API、开源、PaaS和“交互主体化”四种落地模式。

6 大模型技术快速发展，目前在内容创意生成、语言/风格转换、对话、搜索等领域的落地进展较快。此外，利用大模型进行编程，也成为探索的热点。随着未来模型性能的不断提升，模型相关的平台不断完善，大模型或将成为下一代AI基础平台，为各行各业赋能。

7 大模型发展至今也带来了一系列问题。一是模型配套的产业、监管体系相对薄弱。二是深度学习技术本身存在的缺陷。

8 随着模型配套的产业生态逐步建立和完善，大模型技术将变得更为可靠和可控。

本报告建议：

1. 持续推进大模型及相关技术体系的发展。例如，研发更为绿色环保的模型技术；推广低门槛的大模型开发和使用工具；建立动态、面向应用场景的大模型评价体系；甚至建立创新型大模型技术体系。

2. 解决模型存在伦理环保问题和版权争议。例如，通过评估数据集和训练方法，降低模型偏见；通过借鉴已有的先进经验，推动建设低碳绿色的智能算力集群；针对大模型建立专门的开源许可证体系等。

3. 大模型和产业体系融合的过程中，将涉及到包括AI研究者、企业、政府等多方面的角色。应建立“自下而上”和“自上而下”结合的治理体系，减少大模型发展过程中带来的各类问题。

前　言

智能，指的是某一个体通过感知、理解、预测、操控客观物理世界的能力。据Stuart Russell等在《人工智能：一种现代方法》一书中的总结，人工智能从类人和理性的角度上出发，包含四个维度的定义，分别为：

1.能够像人类一样思考；
2.能够像人类一样行动；
3.能够理性地思考；
4.能够理性地行动。

人工智能的概念在1956年美国达特茅斯会议上首次提出，目前已经过了60多年的发展。最近十年，以深度学习为代表的相关技术快速发展，推动AI在计算机视觉、自然语言处理等领域崭露头角，并逐步成为AI发展领域的一大主要流派，促进AI技术在社会经济中的各个领域落地开花。

当前的人工智能基本属于弱人工智能，也称专用人工智能，只能完成特定智能任务，解决特定的智能问题，仍然属于计算机“工具”的范畴。通用人工智能（强人工智能）是人工智能发展的终极目标。通用人工智能能够自适应地应对复杂外界环境的挑战，完成包括学习、语言、认知、推理、创造和计划等人类能完成的所有任务，能在非监督情况下应对前所未见的问题，并同时与人类开展交互式学习，具有比肩人类的能力。

目前实现智能，明确可行的路线有三条：第一条，深度学习+大数据+大算力，获得训练信息模型；第二条，地球环境+太阳能+进化机制，不断进化诞生生物智能；第三条，强化学习+环境模型+大算力，训练的自主智能。

第一条技术途径是过去十年的主流，因为深度学习、数据和算力这三个要素都已具备，全世界掀起了“大炼模型”的热潮，催生了大批人工智能企业。但是，在深度学习技术兴起的近10年间，AI模型基本上是针对特定应用场景需求进行训练的，即小模型。小模型的通用性差，换到另外一个应用场景中可能并不适用，需要重新训练。另外，小模型训练基本是“手工作坊式”，调参、调优的手动工作太多，需要大量的AI专业人员来完成。同时，模型训练对数据要求高，需要大规模的标注数据，如果某些应用场景的数据量少，训练出的模型精度就会不理想。小模型的这些问题，导致了当前AI研发整体成本较高，效率偏低。对于中小行业用户来说，由于AI人才短缺以及成本昂贵，阻碍了行业用户采用人工智能技术的意愿，成为AI普惠的障碍。

虽然之前全球形成了“千村万户大炼模型”的热闹场面，但这种“自家炼钢自家用”的作坊模式成本高，不符合现代产业发展规律。热潮过后，人工智能将进入集约化发展的新阶段，也就是“炼大模型”。

超大规模智能模型（简称“大模型”）可以显著解决小模型研发过程中存在的问题，大模型的泛化能力强，可以做到“举一反三”，同一模型利用少量数据进行微调或不进行微调就能完成多个场景的任务，中小企业不需要招聘很多AI算法专业人员即可以进行应用开发，可以显著降低中小企业的研发门槛，促进AI技术的落地。得益于这些优势，人工智能的发展已经从“大炼模型”逐步迈向了“炼大模型”的阶段。美国OpenAI、谷歌、微软、Facebook等机构，均在布局大规模智能模型的研发，形成了GPT-3、Switch Transformer等千亿或万亿参数量的大模型。

超大规模智能模型发展的前期被人们称为预训练模型，现在可称其为人工智能大模型。大模型之所以为“大”，是因为其是“大数据+大算力+强算法”结合的产物，凝练了大数据内在精华的“隐式知识库”，是实现人工智能应用的载体。AI大模型是连接人工智能技术生态和产业生态的桥梁，向下带动基础软硬件发展，向上支撑智能应用百花齐放，是整个人工智能生态的核心。大模型通常是在大规模无标注数据上进行训练，学习出一种特征和规则。基于AI大模型进行应用开发时，将大模型进行微调（在下游特定任务上的小规模有标注数据进行二次训练）或者不进行微调，就可以完成多个应用场景的任务，实现通用的智能能力。Percy Liang、李飞飞等国际人工智能学者认为，基于自监督学习的模型在学习过程中会体现出来各个不同方面的能力，这些能力为下游的应用提供了动力和理论基础，称这些大模型为“基础模型”（Foundation Models）。



图注：基础模型涉及的议题
来源：https://arxiv.org/pdf/2108.07258.pdf

大模型技术不仅是学术界重点关注的领域，产业领域也在期待其能够在各个场景加速落地。人们期待，大模型不仅能够提升应用服务的智能水平，甚至还能够催生新的场景和产业模式。然而，当前全球大模型商业落地仍处于早期探索阶段，目前已有很多模型落地的探索，但真正让大模型成为推动智能产业发展的核心引擎，目前仍存在不小的差距。本报告旨在梳理当前大模型领域产业的发展情况，为读者提供交流和讨论的机会。

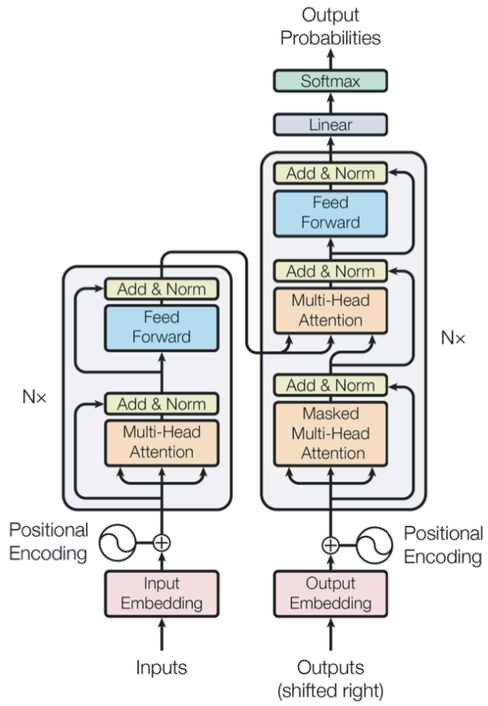
本报告将主要分为以下五部分内容。首先，报告将介绍大模型领域的技术发展情况和趋势。接着，报告将梳理目前已经出现的大模型产业落地模式，提出该模式诞生的条件、特点和优势。然后，报告将重点介绍目前大模型已经开展商业化的发展领域，包括国际和国内的落地领域和应用场景。最后，报告将用两章内容论述应用存在的问题和解决案例，并提出下一步工作建议。

第一章 大模型技术发展情况及趋势

从2018年谷歌提出BERT预训练语言模型至今，超大规模智能模型已经走过了三年的发展历程。近年来，预训练模型成为人工智能领域一大重点研究方向，追求创新算法架构、提出新型预训练和微调方法成为研究者关注的重点。本章节将介绍超大规模智能模型的发展现状和趋势，并提供一些国际国内的研发案例和成果。

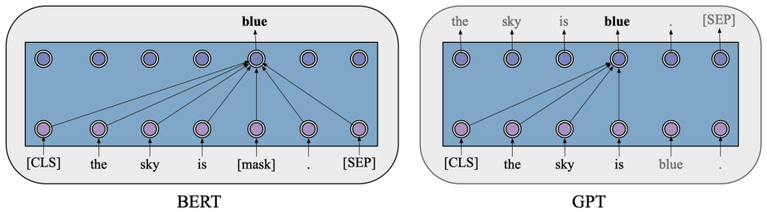
一、超大规模智能模型发展历程

预训练技术的主要思想是迁移学习。当目标场景的数据不足时，首先在数据量庞大的公开数据集上训练基于深度神经网络的AI模型，然后将其迁移到目标场景中，通过目标场景中的小数据集进行微调，使模型达到需要的性能。在这一过程中，这种在公开数据集训练过的深层网络模型，被称为“预训练模型”。使用预训练模型很大程度上降低了下游任务模型对标注数据数量的要求，从而可以很好地处理一些难以获得大量标注数据的新场景。



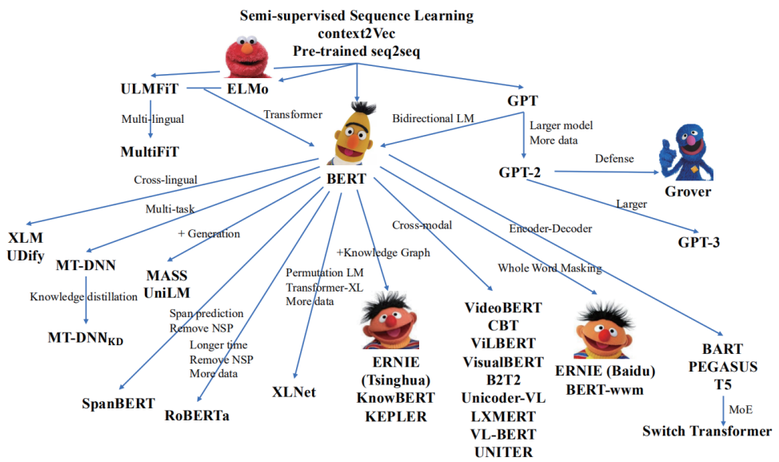
图注：Transformer算法的基本架构
来源：https://arxiv.org/pdf/1706.03762.pdf

预训练技术最早被应用于计算机视觉领域。2017年，Vaswani等提出Transformer架构，奠定了当前大模型领域主流的算法架构基础。2018年，谷歌提出了大规模预训练语言模型BERT（基于Transformer的双向深层预训练模型），OpenAI提出了GPT（生成式预训练Transformer模型），极大地推动了自然语言处理领域的发展，成为预训练史上重要的里程碑。

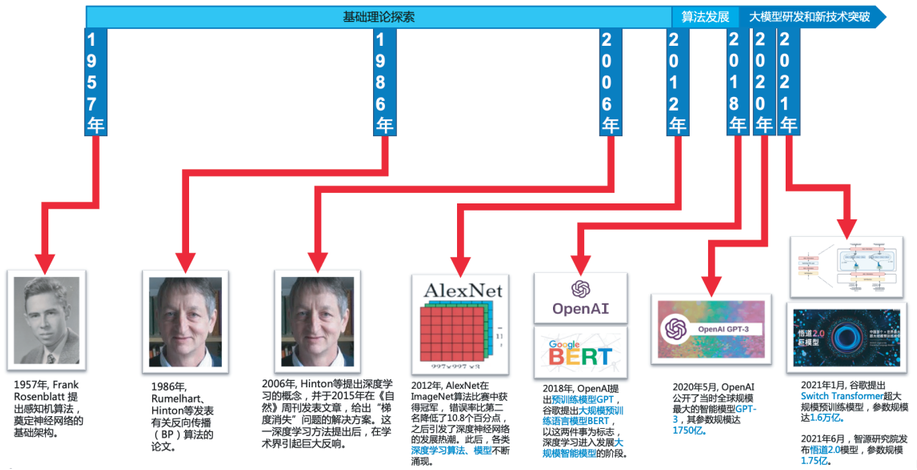


图注：BERT和GPT的区别
来源：https://arxiv.org/pdf/2106.07139.pdf

此后，基于BERT的改进模型、XLNet、RoBERTa、T5等大量新式预训练语言模型不断涌现，预训练技术在自然语言处理领域蓬勃发展。2020年5月，OpenAI公开了超大规模预训练语言模型GPT-3，参数达1750亿，在全球范围内掀起了研发千亿参数规模模型的热潮。



图注：截至2021年出现的一些预训练语言模型（不含多模态预训练模型）
来源：https://arxiv.org/pdf/2106.07139.pdf



图注：从深度学习到预训练模型技术发展的一些标志性事件（截止2021年初）
来源：智源研究院根据公开资料整理

二、超大规模智能模型的发展意义

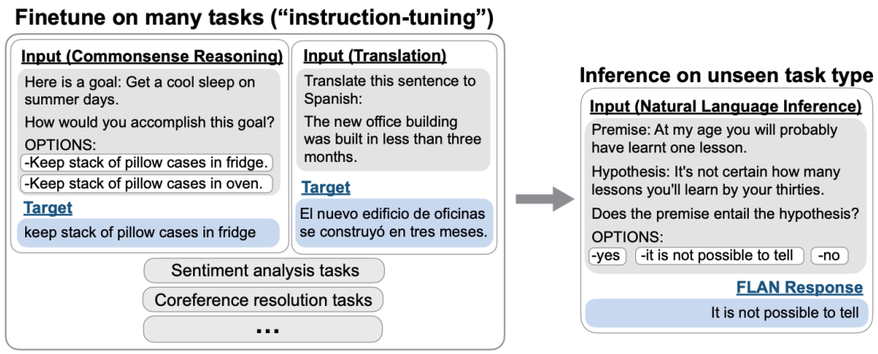
（一）推动AI科研的发展

大模型最为深远的意义在于为通用人工智能发展提供了一条可行路径。在大模型出现之前，人们对于是否能够通过深度学习路径，实现媲美人类智能水平的机器存有疑虑。大模型为通向通用人工智能开辟了一条全新的途径—“大算力+大数据+大参数”。人们认识到，即使没有能够明确阐释智能产生原理的基础理论，通过大规模的参数学习，也能够取得接近甚至超越人类的智能水平，如果模型还不足以达到所需的智能水平，那就是参数量不够大。

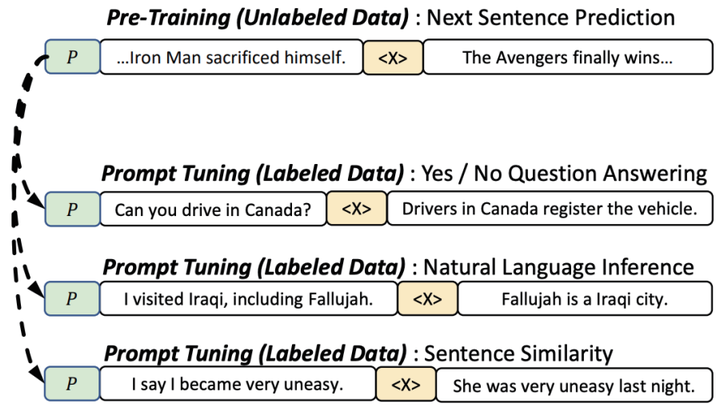
大模型在人工智能领域带来了巨大的学术贡献。从大模型技术诞生伊始，全球已经诞生了大量的相关研究。在Google Scholar中采用“Pre-trained model”（预训练模型）作为关键词，能够搜索到超过2.9w篇相关论文，如果加入限定词“BERT”，也有将近1.6w篇。

大模型的诞生在AI领域引起了研发更大规模模型的热潮。一些全球AI机构针对大模型开展“军备竞赛”。大模型的参数规模、性能和精度成为衡量AI机构技术实力的指标，全球科研机构通过模型的军备竞赛，在技术上“秀肌肉”。例如，国际上，谷歌、OpenAI、微软、Facebook等先后推出自研大模型，将参数规模推高了四个数量级，将数据从文本单模态推向图像、视频等多模态领域。国内的企业和科研机构也先后自研大模型，并搭配开放模型、API接口、平台等服务，极力成为“最强”的新标杆。

在研究领域的层面上，大模型技术已经开枝散叶，形成了多条技术路径。这包括基于预训练思路的架构创新（XLNet等）、基于原始模型的改进和提升（RoBERTa等）、以及在更大参数规模上的技术体系变革（GPT-3、Switch Transformer）等，逐渐形成了一个独立的AI技术范式——“预训练模型”范式。这一新型技术范式已经逐步成为推动算法革新、并行加速，以及模型小型化（剪枝、压缩、蒸馏）等领域发展的重要技术之一。例如，今年9月，谷歌的Quoc Le等研究者提出了指令微调（Instruction Tuning）技术，公开了了名为FLAN的预训练模型，拥有1370亿参数，性能超越GPT-3在零样本自然语言推理、阅读理解和开放域问答任务上的表现。此外，清华大学刘知远、黄民烈等提出“Pre-trained Prompt Tuning(PPT)”框架，在少样本任务上达到与全参数微调相当甚至更好的效果。

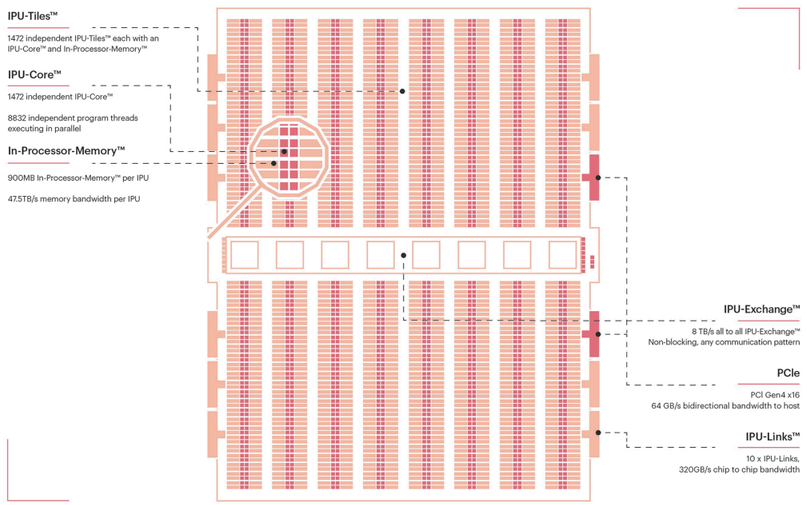


图注：指令微调的基本原理
来源：https://arxiv.org/pdf/2109.01652.pdf

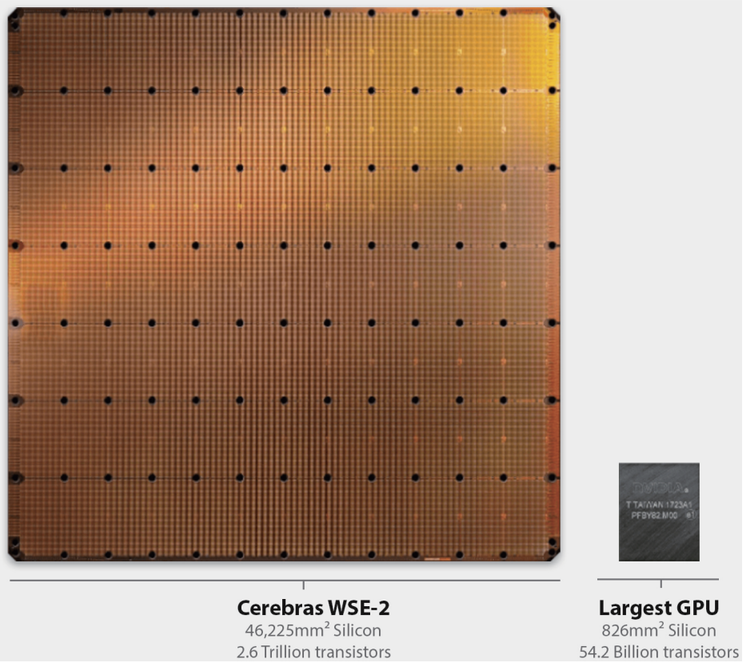


图注：PPT的应用案例
来源：https://arxiv.org/pdf/2109.04332.pdf

此外，大模型技术以点带面，推动芯片及IT基础设施等技术的发展。大模型参数巨大，即使是采用传统的GPU芯片进行并行加速，很多时候也难以达到预期的效果，因此很多芯片企业开始研发专用于大模型的加速芯片，如头部厂商英伟达（A100 GPU）、谷歌（TPUv4）、百度（昆仑），以及英国初创企业Graphcore（GC2 IPU）、美国Cerebras（WSE-2）等。



图注：Graphcore IPU架构图
来源：Graphcore官网



图注：Cerebras研发的芯片
来源：Cerebras官网

在IT基础设施方面，为了提供大模型所需的算力资源，服务器的通信、存储、资源管理等架构都需要迭代更新。为了抢占大模型技术的制高点，当前世界各地都在更新或兴建智能算力中心（或称“AI超算”），如美国的Perlmutter、Dojo，欧洲的Leonardo，国内各地的科研机构和企业兴建的算力基础设施等。这些智能算力中心的主要特征在于大量采用GPU等加速芯片进行AI计算，计算精度没有科学计算的要求高，能够承担大模型的训练工作等。

表1：全球已建或在建的AI超算（来源：智源研究院整理）

国家或地区	智能超算名称	建设目标	建设进展
美国	Frontier（前沿）	算力达 1.5E，即 1500P	美国政府投资建设，预计将于 2021 年内完工
	Perlmutter（帕尔穆特）	当前全球最强超算，算力达 4E，即 4000P	今年 5 月底投入运行，已确定了 20 余个科研应用领域，包括天体物理学、气候科学、材料学等
	Dojo（道场）	自研的 AI 训练芯片 D1，当前算力 1.1E	特斯拉投资建设，为研发自动驾驶相关 AI 模型提供算力支持
欧盟	意大利：莱奥纳多（Leonardo）	目标：FP16 位精度算力达到 10E，即 10000P 以上	将于 2021 年底完工
	在芬兰、卢森堡、西班牙等 8 个国家建设 P 级别算力的超算，如：芬兰 Lumi 超级计算机	形成统一的欧盟超算网络；芬兰 Lumi：552P，超越“日本富岳”；其他国家超算：算力达到 P 级别，均入围全球 Top50；	预计将于 2021 年内投入使用；其他国家项目正在稳步推进
英国	剑桥一号（Cambridge-1）	形成英国最强的 AI 超算	搭载英伟达药物研发软件“NVIDIA Clara for Computational Drug Discovery”，目前已投入新冠病毒药物和疫苗研制工作
日本	富岳（Fugaku）	已设定了 9 个对社会有重要意义的应用领域，如医学、药理学、材料学等，将探索利用 AI 技术进行新材料研发、药物研发、医疗诊断算法研究等。	已于 2021 年 3 月正式启用 在 2020 年全球超算性能 500 强榜单（Top500）、2020 年国际 AI 超算性能基准测试 HPL-AI 榜单上均位居榜首

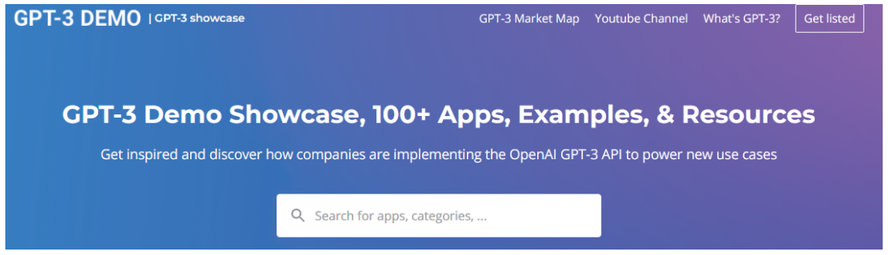
大模型改变的不仅仅是技术体系，还有AI科研范式。过去，大多数AI技术的研发过程是小团队主导的——通常由一个团队根据当前领域的发展情况，提出一个科学问题，通过研发一种新的算法架构/加速方法/训练方法/评价方法，解决这一问题。这样的研发模式虽然推动了AI技术的进步，但主要属于“点”的突破，很难与其他AI技术发展形成联动。在大模型时代，超大规模模型的研发更偏向完成一个大型工程项目。项目中需要来自算法、数据、硬件等方面的复合型人才，通过配合的方式，产出跨学科的研究成果。和航天工程一样，整个项目追求的目标并不是单个技术点的突破，而是通过整个大型工程项目的实施和发展，带动相关学科和技术的进步，这样一种科研模式让人们看到，在AI领域有时候不一定需要主攻技术理论，即使是集中力量在已有技术的基础上实现大规模的工程化扩展，也可以实现AI技术发展目标。

总体而言，大模型已经开启了AI科研的大规模工程化模式。许多科研机构和企业可以利用巨量的资金、人员和资源优势，通过汇聚大量科技人才的方式，研发大型工程化的模型，实现AI系统的总体突破。这样一来，即使在单点技术上没有足够的创新，这些大模型也可以凭借规模优势，成为AI领域的香饽饽。

（二）推动智能产业的发展

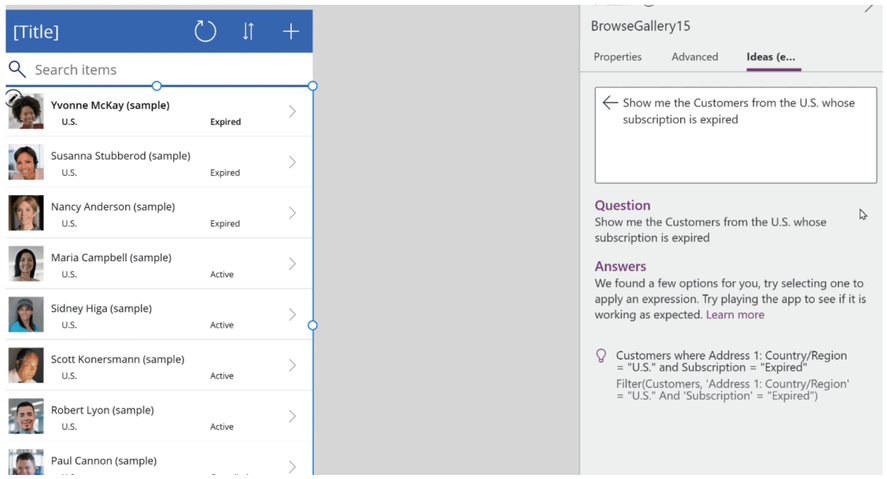
大模型在产业领域的意义主要包括：推动传统产业的智能化转型；催生基于智能模型的新产业；以及可能改变人类社会的生产和管理模式。

大模型最直接的产业意义在于对传统产业的智能化改造和升级，常见的应用包括更为智能的对话客服，更为高效的机器翻译和文本处理工具等。例如，GPT-3等大模型能够推动人工智能向高级智能应用领域迈进，如新闻文本生成、商业文本分析、法律文本分析等，具有较高的产业应用价值。同时，GPT-3可以进一步提升已有AI应用的智能化水平，推动产业快速发展，据OpenAI透露，截至今年3月，全球已有300多个APP采用了GPT-3技术。据gpt3demo.com统计，截至6月份，基于GPT-3的应用已经有150多个，涵盖智能搜索、翻译、对话、广告、营销等领域。OpenAI也认识到GPT-3的潜力，早早开放API供从业者进行试验，使其能够开发出智能代码补全等应用。



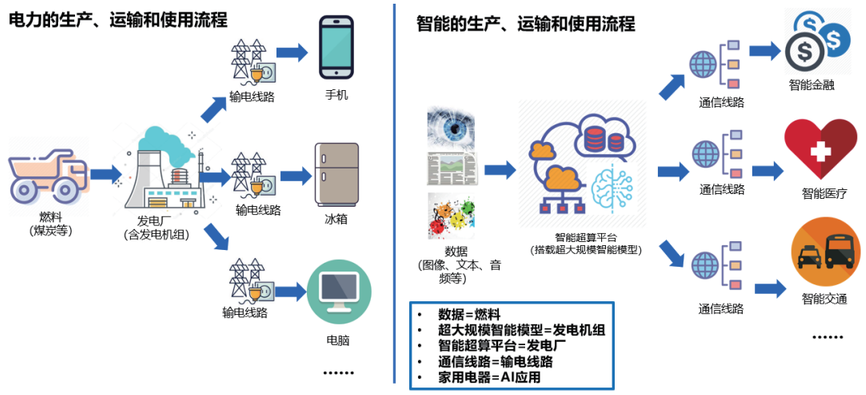
图注：gpt3demo.com页面
来源：gpt3demo官网，2021年8月

在代码生成领域，大模型能够根据数据科学家的需求自动生成分析检索代码，降低了数据分析的技术门槛，使企业业务人员可以直接进行大型分析和决策，如微软的Microsoft Power App数据分析平台中就集成了GPT-3模型，能够自动生成代码，执行数据检索、计算等操作。同时，大模型可能推动新型产业形态的诞生，如智能决策等。例如，去年11月，微软和OpenAI联合向美国政府建议，利用GPT-3等技术，实现美国技术出口管理工作的自动化。



图注：GPT-3能够在微软数据分析平台中根据自然语言指令生成对应的代码，加快数据检索效率
来源：微软官网

更为重要的是，大模型在提升社会生产力的同时，也将间接改变全社会的生产关系，围绕大模型形成智能能力（简称“智力”）的生产和消费模式。正如智源研究院理事长张宏江所言：“如果将AI比作电力，那么大模型则相当于‘发电机’，能将智能在更大的规模和范围普及。”大模型的智能能力在未来将成为一种公共基础资源，像电力或自来水一样随取随用。每个智能终端、每个APP，每个智能服务平台，都可以像接入电网一样，接入由IT基础设施组成的“智力”网络（俗称“智网”），让智能能力完成日常中方方面面的任务，帮助人类从枯燥重复的工作中解放出来，实现生产力的提升。



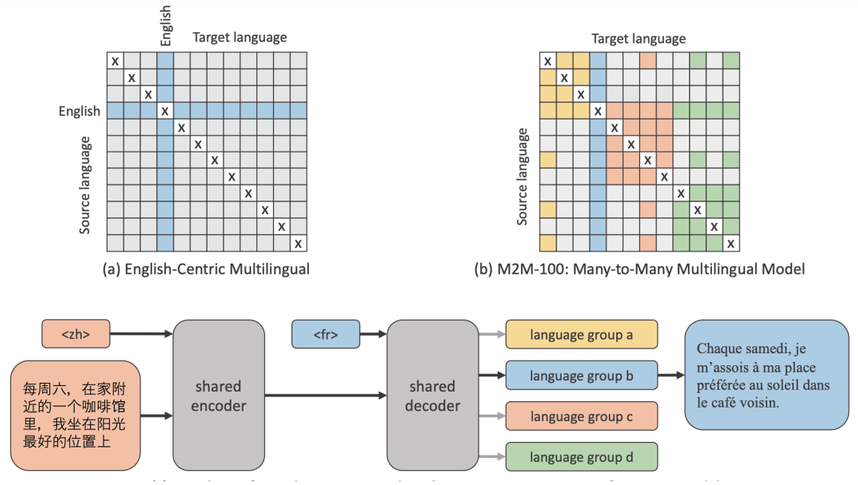
图注：电网和“智网”的类比
来源：智源研究院整理

基于大模型形成的大规模智能网络和云计算完全不同。云计算尽管也将算力比作随取随用的资源，但目前仍然做不到像电力一样的方便。这是因为云计算本质上需要云和用户进行双向的资源交换，用户需要经过繁琐的数据准备、计算过程定义等过程，将数据和算法转移到云端，再由云端完成任务后传回本地。而基于大模型的“智网”则不需要用户定义计算过程，只需要准备好数据，大模型强大的能力能够在少量微调（Few-shot）甚至不微调（Zero-shot）的情况下直接完成用户所需的工作，就像是家用电器只需要电力驱动电机，具体用途则由电器本身的属性定义，实现“动力”和“用途”的隔离。这一智能能力生产和消费的网络，未来或将成为人类社会产业发展的主流模式。

三、超大规模智能模型的技术发展趋势

（一）多语言预训练模型

当模型参数达到人类算力的极限时，人们进一步发现，模型采用的单语言（英语）或单模态数据（文本）已然不能满足实际的需求，于是又将多语言、图像、音频等能够体现客观物理世界的数据融合在模型中，以期实现更高效的表征学习。在多语言模型方面，2020年10月，Facebook和谷歌相继推出多语言大规模预训练模型。多语言模型可以在相似的语言之间共享信息，降低对数据和资源的依赖，并且允许少样本或零样本学习。多语言模型设计的目标就是建立一个能够理解世界上大部分语言的模型。Facebook已发布了一个百种语言互译的模型M2M-100，该模型不依赖英文作为中介语言，可实现一百种语言之间的直接翻译，在机器翻译领域实现新突破。此外，谷歌宣布开源了多语言模型MT5，基于101种语言进行训练，采用750GB文本，最大含有130亿个参数，目前已在大多数多语言自然语言处理任务基准测试中达到最优水平，包括机器翻译、阅读理解等。



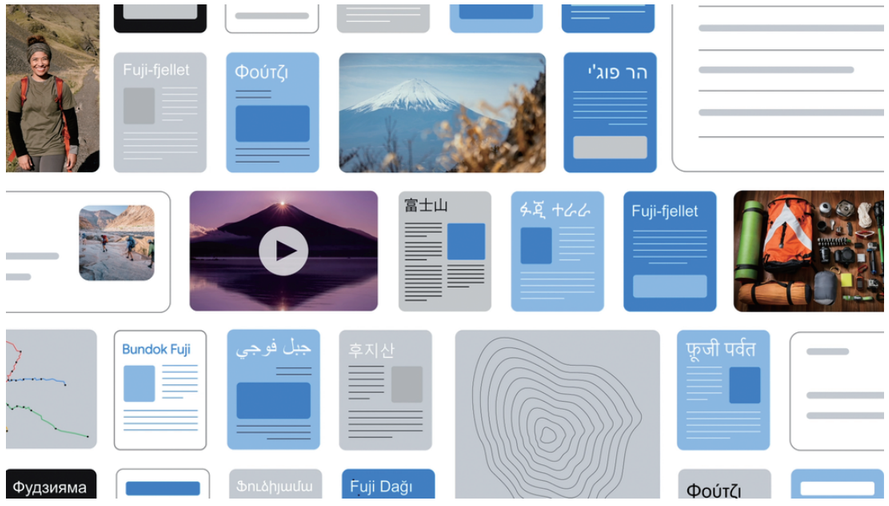
图注：M2M-100模型架构
来源：https://arxiv.org/pdf/2010.11125.pdf

（二）多模态预训练模型

多模态大模型领域，OpenAI已研发DALL·E、CLIP等多模态模型，参数达120亿，在图像生成等任务上取得优秀表现。此外，智源研究院、阿里达摩院等机构也在积极研发多模态预训练模型，探索文本之外的领域。

（三）多任务预训练模型

面向未来产业界更为复杂的智能决策场景，研发基于多种网络数据预训练，具有决策能力的大模型，可能是下一步发展的重点。例如，谷歌在今年的IO大会上公开了MUM（多任务统一模型：Multitask Unified Model）的发展情况。据谷歌透露，MUM模型基于大量的网页数据进行预训练，擅长理解和解答复杂的决策问题，能够理解75种语言，从跨语言多模态网页数据中寻找信息。谷歌展示了MUM模型通过用户英文提问搜索日文信息源，并提供旅行攻略的过程。说明当大模型学习更为丰富的模态数据后，其在处理复杂信息理解和生成任务时会有更强的表现。



图注：MUM模型能够根据用户提问从多种源头的网页信息中搜索出对应的旅行攻略
来源：https://blog.google/products/search/introducing-mum/

（四）视觉预训练大模型

大模型能够落地应用的领域，本质上和其预训练的数据模态、数据类型密切相关。例如目前面向视觉通用任务处理的需求，“视觉大模型”——采用Transformer架构，通过学习大量的图像和视频数据，形成具备视觉通用能力的大模型，如ViTransformer等。考虑到视觉任务在日常生活和产业发展中占据很大的比重，视觉大模型有可能在自动驾驶等依赖视觉处理的领域加速应用。

（五）物理世界模型

目前大模型学习的仅仅是语言等数据的表征，这些数据本身就是对客观世界的一种映射和表示，本身已经弱化了。未来的大模型应学习真实物理世界的各种数据，成为真正理解人类世界的智能模型。如在一个空间中，每个人的位置、背景、关系、谈话内容、场景图像等。只有这样，模型才能充分理解未能体现在文本中的有效信息，实现更为强大的推理性能。

四、典型的大模型案例

表2：GPT-3公开前的一些大模型（来源：智源研究院整理）

序号	发布时间	研发机构	模型名称	模型内参数数量	数据规模	计算时间
1	2018 年 6 月	美国 OpenAI	GPT	110M	4GB	3 天
2	2018 年 10 月	美国谷歌	BERT	340M	16GB	50 天
3	2019 年 2 月	美国 OpenAI	GPT-2	1.5B	40GB	200 天
4	2019 年 7 月	美国脸书	RoBERTa	330M	160GB	3 年
5	2019 年 10 月	美国谷歌	T5	11B	800GB	66 年

注：M-百万，B-十亿，计算时间为使用单块英伟达V100 GPU训练的理论估计时间，
由于芯片性能、通信带宽、数据处理、算法效率等因素影响，实际时间可能更久

表3：近一两年国内外大模型领域发展情况（智源研究院整理）

序号	发布时间	研发机构	模型名称	模型内参数规模	训练数据情况
1	2020 年 5 月	美国 OpenAI	GPT-3	1750 万	2TB 英文文本数据
2	2021 年 1 月	美国谷歌	Switch Transformer	1.6 万亿	未公开
3	2021 年 1 月	美国 OpenAI	DALL·E 和 CLIP	120 亿	自建包含 2.5 亿个“图像-文本”对的训练数据集
4	2021 年 1 月 12 日	北京智源研究院、清华大学、杭州阿里巴巴等	悟道·文汇	113 亿	未公开
5	2021 年 3 月 3 日	杭州阿里达摩院 清华大学	M6	1000 亿	1.9TB 图像和 292GB 文本
6	2021 年 3 月	北京智源研究院	悟道 1.0	1100 万	2TB，包括新闻咨询、评论、百科、论坛、博客、学术论文等
7	2021 年 4 月 19 日	杭州阿里达摩院	PLUG（已加入 AliceMind 框架）	270 亿	1TB 中文文本数据
8	2021 年 4 月 25 日	深圳华为、循环智能等	盘古 NLP	1100 亿	40TB 数据，数据来自销售、客服等企业与客户之间的沟通场景，涵盖金融、保险、教育、地产、本地生活、电商、汽车等诸多行业
9	2021 年 4 月 25 日	深圳鹏城实验室、华为诺亚方舟实验室、北京大学等	盘古α	2000 亿	1.1TB 文本数据
10	2021 年 6 月 1 日	北京智源研究院	悟道 2.0	1.75 万亿	4.9T（1.2TB 中文文本数据、2.5TB 中文图文数据，以及 Pile 数据集的 1.2TB 英文文本数据）
11	2021 年 6 月 25 日	杭州阿里达摩院	M6（万亿参数版本）	1 万亿	65TB 图像和 4.8TB 中文文本
12	2021 年 7 月 6 日	北京百度	ERNIE 3.0	100 亿	4TB，涵盖百科、新闻、对话、小说等不同类型、金融、医疗、法律等不同领域的文本

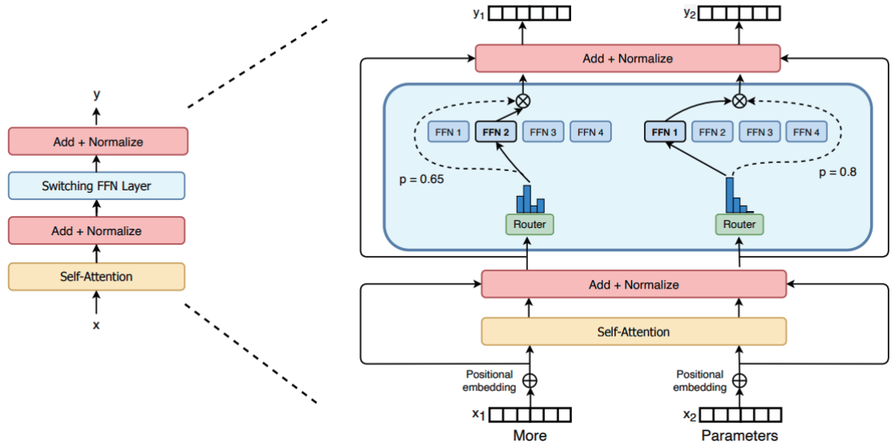
（一）国际

1. GPT-3

2020年5月底OpenAI公开的GPT-3超大规模预训练语言模型是当时全球规模最大的预训练语言模型，参数规模达到1750亿。GPT-3在语言建模、文本补全、对话问答、机器翻译、常识推理等数十项常见的自然语言处理任务上取得了较好的性能表现。现在已有科研和开发人员基于GPT-3模型开发了多款AI应用，如自动代码生成、自动新闻生成、国际象棋游戏应用等。

2. Switch Transformer

2021年1月，谷歌研究人员研发出新的语言模型Switch Transformer，它包含1.6万亿个参数，是迄今为止最大规模的人工智能语言模型之一，比之前谷歌开发的语言模型T5-XXL的规模大了4倍，是包含1750亿参数的GPT-3的九倍。



图注：Switch Transformer编码块结构
来源：https://arxiv.org/pdf/2101.03961.pdf

Switch Transformer的创新点在于对预训练和加速策略进行了优化。一是基于Transformer MoE（Mixture of Experts）网络结构，简化了MoE的routing机制，降低了计算量；二是进一步通过数据并行+模型并行+expert并行的方式，降低了训练通信量，提升训练性能。研究者将Switch Transformer与谷歌研究的T5-Base和T5-Large模型进行了对比，结果表明，在相同的算力资源下，新模型实现了最高7倍的预训练速度提升。

3. DALL·E/CLIP

2021年1月5日，OpenAI同时发布了两个大规模多模态预训练模型：DALL·E和CLIP。CLIP可以通过自然语言的监督学习来有效理解视觉概念，用户只需要提供要识别的视觉类别名称（标签），CLIP便能够完成视觉分类任务。CLIP解决了目前传统的深度学习方法存在的几个问题，一是传统方法需要大量的训练数据集，创建成本较高；二是传统的计算机视觉模型往往只能完成一种任务，迁移到其他任务需要花费高昂的成本；三是在基准测试上表现良好的模型，在实际测试中往往不尽人意。而CLIP从互联网搜集的4亿图像-文本对的数据训练后，用自然语言描述所学习的视觉概念，就能够在Zero-Shot（零样本微调）状态下直接迁移到下游任务中去，体现出性能优势。



图注：CLIP模型在多项ImageNet测试中取得优秀水平
来源：OpenAI博客

DALL·E与GPT-3的架构相似，是一个具有120亿参数的Transformer模型。两者的区别在于，GPT-3的输出是文本，DALL·E输出的是图像。DALL·E能够为五花八门的句子创造出组合的意象，一定程度上显示出对概念的理解、组合和创造能力。例如，当文本提示为“生成牛油果形状的躺椅”时，DALL·E能够理解牛油果和躺椅两个概念，并生成各种牛油果形状的、绿色的躺椅。此外，在输出图像的时候，DALL·E采用CLIP进行排序，从中选取最优结果，整个过程不需要进行任何筛选。



图注：DALL·E生成的“牛油果形状的椅子”
来源：OpenAI博客

4. ViT

2020年10月，谷歌大脑团队首次尝试将标准Transformer应用于图像，提出了视觉Transformer（ViTransformer，即ViT）模型，并在多个图像识别基准上实现了接近甚至优于当时SOTA方法的性能。2021年6月，原ViT团队的几位成员又尝试将ViT模型进行扩展，研究者分别在以下三个条件中试验了扩展后的ViT性能：一是将模型参数从500万提升到20亿个；二是将训练图像的数据集从3000万提升到30亿个；三是将计算预算从低于1个TPUv3核/每天（Core-day）提升到超过10000个核/每天，用于探索ViT模型的性能边界。该项研究最终训练出了一个具有20亿个参数的ViT变体模型ViT G/14，它在ImageNet数据集上达到了新的SOTA性能，90.45%的Top-1准确率。

（二）国内

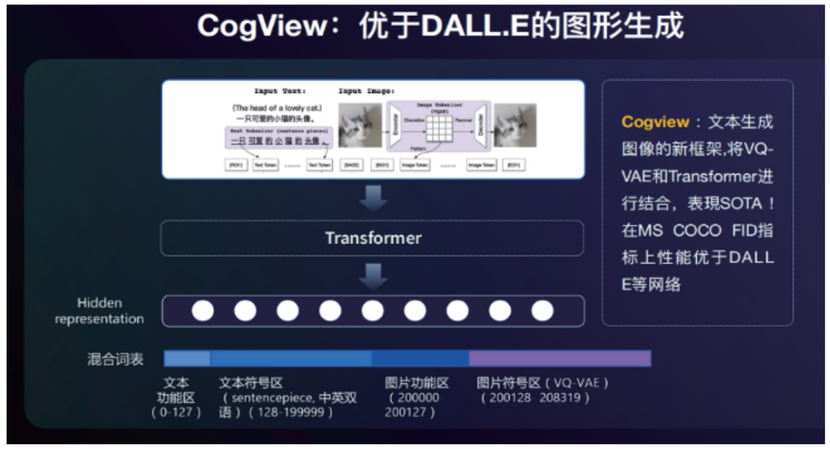
1. 悟道（1.0/2.0）

3月20日，智源研究院发布我国首个超大规模智能信息模型“悟道1.0”，训练出包括中文、多模态、认知、蛋白质预测在内的系列模型，并在模型预训练范式、规模和性能扩增技术、训练语料数据库建设等方面取得多项技术突破，包括全新的预训练范式；单一模型在理解、生成等任务上取得最优效果；研发预训练模型从亿级参数到万亿级参数扩增的关键技术MoE（Mixture of Experts）；开源支持PyTorch框架的高性能MoE系统；建设并开放了全球最大的中文语料数据库WuDaoCorpora等。

6月1日，智源研究院发布“悟道2.0”模型，参数规模达到1.75万亿，是GPT-3的10倍，打破了之前由Switch Transformer模型创造的1.6万亿参数记录。悟道2.0中采用了FastMoE技术，支持PyTorch框架和大规模并行训练。新一代FastMoE，支持Switch、GShard等复杂均衡策略，支持不同专家不同模型。在数据方面，“悟道2.0”模型在中英双语共4.9T的数据上进行训练。训练数据包含WuDaoCorpora中的1.2TB中文文本数据、2.5TB中文图文数据，以及Pile数据集的1.2TB英文文本数据。

此外，悟道2.0研发过程中实现了多项技术创新，具体如下：

- GLM2.0：以单一模型兼容所有主流架构，具有100亿参数量，性能比肩微软170亿参数的Turing-NLG模型，取得多项任务的更优成绩；
- P-tuning2.0算法：在少样本学习领域能力领先；
- CogView：将VQ-VAE和Transformer进行结合，达到SOTA水平，在MS COCO FID指标上性能优于DALL·E等模型。



图注：悟道2.0中的技术创新成果
来源：智源研究院

“悟道2.0”模型在9项基准测试任务上取得优异成绩，主要包括：

- ImageNet zero-shot SOTA超过OpenAI CLIP；
- LAMA知识探测：超过AutoPrompt；
- LAMABADA完形填空：能力超过微软Turing NLG；
- SuperGLUE few-shot FewGLUE：超越GPT-3，取得当前最佳少样本学习结果；
- UC Merced Land-Use zero-shot SOTA，超过OpenAI CLIP；
- MS COCO文生成图：超越OpenAI的DALL·E；
- MS COCO英文图文检索：超过OpenAI CLIP 和 Google ALIGN；
- MS COCO多语言图文检索：超过目前最好的多语言多模态预训练模型UC2，M3P；
- Multi 30K多语言图文检索：超过目前最好的多语言多模态预训练模型UC2，M3P。

“悟道”模型研发过程中，智源研究院建设了全球最大的语料数据库WuDaoCorpora2.0，包含全球最大的中文文本数据集（WDC-Text）、全球最大的多模态数据集（WDC-ImageCaption）、全球最大的对话数据集（WDC-Dialogue）。另外，智源研究院提出“智源指数”，一种以人类语言能力为参照、全面系统地测评机器语言能力的体系。“智源指数”首次提出了基于“能力-任务-数据集”层次结构的机器语言评测体系及评测方案，包含6种主要语言能力，30余项主流任务与相关数据集，同时也是首次基于信度、难度、效度等指标进行评价的高质量评测数据集合。

2. 盘古NLP/盘古α

2021年4月25日，华为与循环智能等机构联合发布了千亿参数的中文预训练语言模型“盘古NLP”。“盘古NLP”模型参数量达1100亿，算力由鹏城实验室提供，在预训练阶段使用了40TB数据，数据来自销售、客服等企业与客户之间的沟通场景，涵盖金融、保险、教育、地产、本地生活、电商、汽车等诸多行业，构成了庞大的行业知识库。在中文语言理解评测基准CLUE榜单中，取得总成绩第一，分类榜单第二、阅读理解榜单第一的成绩，总成绩得分83.046。

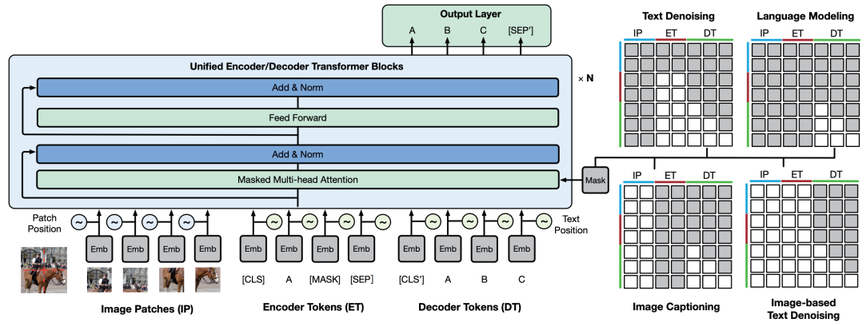
排行	模型	研究机构	测评时间	Score	认证	AFQMC	TNEWS	IFLYTEK	CMNLI	OCNLI_50K	CLUEWSC 2020	CSL	CMRC2018	CHID	C3
1	HUMAN	CLUE	19-12-01	85.610	已认证	81.000	71.000	80.300	76.000	90.3	98.000	84.000	92.400	87.100	96.000
2	Pangu	华为云-循环智能	21-04-23	83.046	待认证	78.114	72.070	65.192	85.190	83.300	95.517	87.733	84.450	93.253	85.637
3	BERTSG	Sogou Search	21-03-24	81.904	待认证	77.156	72.790	64.538	83.769	83.667	94.483	87.667	83.250	89.090	82.631
4	Mengzi	TBH	21-04-25	81.213	待认证	79.513	73.350	63.885	84.957	80.133	95.862	86.833	82.250	86.854	78.494
5	MT-BERTs	Meituan NLP	21-03-10	81.064	待认证	77.363	70.030	64.308	85.139	83.467	89.655	87.400	83.200	89.788	80.293
6	Motian	Motian	21-05-13	80.530	待认证	77.726	70.860	65.462	84.687	83.667	94.828	90.167	75.350	85.579	76.978
7	LICHEE	LICHEE	21-01-08	80.508	待认证	76.975	70.500	64.154	84.541	81.300	90.690	87.400	79.800	87.505	82.220
8	roberta_selfrun	OPPO小布助手	21-02-05	80.238	待认证	77.881	69.370	63.923	82.939	80.400	93.103	87.267	80.100	90.107	77.287
9	BERTs	BERTs	20-12-24	80.220	待认证	76.768	69.940	63.923	84.483	82.900	88.966	86.767	80.500	89.513	78.443
10	UER-ensemble	TencentPretrain ...	20-11-28	80.087	待认证	76.819	72.200	64.000	84.090	80.800	90.345	85.833	79.150	86.031	81.603

图注：盘古NLP模型目前在CLUE总排名榜单Top10中的排名情况
来源：CLUE，2021年4月25日

4月，鹏城实验室与华为诺亚方舟实验室、北京大学等联合发布了2000亿参数的中文预训练语言模型“盘古α”，参数达2000亿。该模型采用团队自建的1.1TB文本数据集进行训练，基于华为自研MindSpore算法框架，在鹏城云脑II期平台完成训练和测评工作。目前盘古α模型的代码、模型、评测数据集已在启智社区（OpenI）开源，训练数据集和在线服务也将在优化后开源。

3. 阿里M6/PLUG

2021年3月，阿里达摩院联合清华大学提出了参数规模达1000亿的中文多模态预训练模型M6。6月25日，万亿参数版本的M6发布。研发过程中，该模型在480卡的V100 GPU上进行了训练，相比谷歌、英伟达等机构，节省算力资源超80%，训练效率提升近11倍。M6能够完成跨模态检索、视觉问答、图片配文字等任务，生成图像质量达到1024x1024像素的水平，已在阿里巴巴30多项业务中实现应用。



图注：M6模型的架构
来源：https://arxiv.org/pdf/2103.00823.pdf

此外，达摩院提出MUGE评测基准，可用于验证多模态预训练算法效果，任务类型包括图像描述生成、基于文本的图像生成等、跨模态检索等。

2021年4月19日，阿里达摩院发布了270亿参数的中文预训练语言模型PLUG，采用1TB中文文本数据进行训练，数据题材涵盖新闻、小说、诗歌、问答等，可完成小说创作、诗歌生成、智能问答等长文本生成任务。PLUG采用encoder-decoder双向建模方式，具备语言理解与生成能力，是发布时参数规模最大的中文预训练语言模型，以80.614分打破CLUE（中文语言理解测评基准）分类任务榜单纪录。目前，PLUG模型并入达摩院发布的深度语言模型体系 AliceMind中，计划将升级至万亿参数规模。

CLUE分类任务排行榜-记最佳得分(10.22后中文原版数据集OCNLI替代CMNLI，bert_base初始化分数，可重新跑OCNLI并提交)

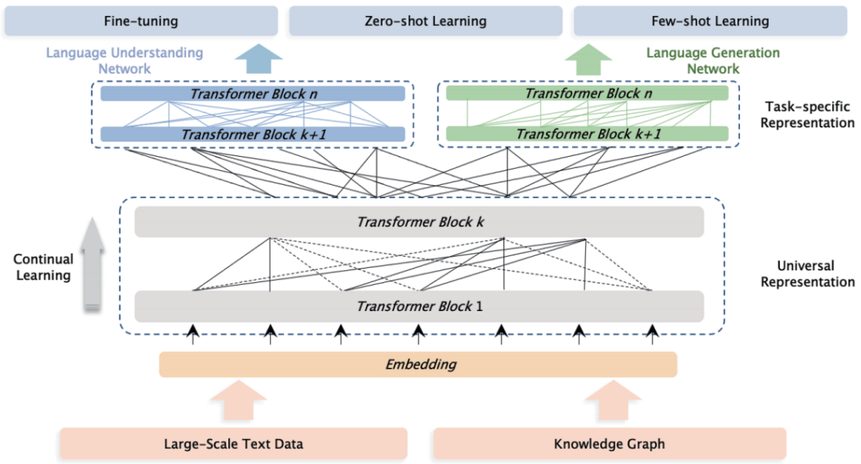
排行	模型	研究机构	测评时间	Score	认证	AFQMC	TNEWS	IFLYTEK	CMNLI	OCNLI_50K	CLUEWSC 2020	CSL
1	HUMAN	CLUE	19-12-01	82.943	已认证	81.000	71.000	80.300	76.000	90.3	98.000	84.000
2	PLUG	Alibaba DAMO NLP	21-04-18	80.614	待认证	77.441	73.060	64.000	84.949	83.267	94.483	87.100
3	BERTSG	Sogou Search	21-03-24	80.581	待认证	77.156	72.790	64.538	83.769	83.667	94.483	87.667
4	atreb	pln	21-04-08	80.076	待认证	76.897	71.230	64.731	84.447	82.433	93.793	87.000
5	Bert	lihaiyu	21-04-08	79.663	待认证	75.602	70.320	64.923	84.549	81.733	93.448	87.067
6	MT-BERTs	Meituan NLP	21-03-10	79.623	待认证	77.363	70.030	64.308	85.139	83.467	89.655	87.400
7	test	test	21-03-01	79.374	待认证	76.379	70.570	63.962	84.549	80.133	93.793	86.233
8	LICHEE	LICHEE	21-01-08	79.366	待认证	76.975	70.500	64.154	84.541	81.300	90.690	87.400
9	Mengzi	TBH	21-04-13	79.361	待认证	79.513	73.350	62.308	84.957	78.567	90.000	86.833
10	roberta_selfrun	OPPO小布助手	21-02-05	79.269	待认证	77.881	69.370	63.923	82.939	80.400	93.103	87.267

图注：PLUG模型目前在CLUE分类任务榜单Top10中的排名情况
来源：CLUE，2021年4月19日

4. ERNIE 3.0

2021年7月，百度推出ERNIE 3.0知识增强大模型，参数规模达到百亿。其基于中文文本及大规模知识图谱训练，数据总量达到4TB，涵盖百科、新闻、对话、小说等不同类型、金融、医疗、法律等不同领域的文本。ERNIE 3.0在百亿级预训练模型中引入大规模知识图谱，通过将大规模知识图谱的实体关系与大规模文本数据同时输入到预训练模型中进行联合掩码训练，促进了结构化知识和无结构文本之间的信息共享，提升了模型对于知识的记忆和推理能力。

此外，ERNIE 3.0提出将通用语义表示与任务语义表示相结合的模型框架，该框架融合自编码和自回归等不同的任务语义表示网络，既可同时处理语言理解和语言生成任务，还能做无标注数据的零样本学习和有标注数据的微调训练。



图注：ERNIE3.0框架图
来源：https://arxiv.org/pdf/2107.02137.pdf

ERNIE 3.0在54个中文NLP任务基准上达到最佳水平，包含情感分析、观点抽取、阅读理解、文本摘要、对话生成、数学运算等不同的自然语言理解与生成任务。此外ERNIE 3.0登顶了复杂自然语言理解任务评测SuperGLUE榜单。

	Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
	1	ERNIE Team - Baidu	ERNIE 3.0		90.6	91.0	98.6/99.2	97.4	88.6/63.2	94.7/94.2	92.6	77.4	97.3	68.6	92.7/94.7
✚	2	Zirui Wang	T5 + Meena, Single Model (Meena Team - Google Brain)		90.4	91.4	95.8/97.6	98.0	88.3/63.0	94.2/93.5	93.0	77.9	96.6	69.1	92.7/91.9
✚	3	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.3	90.4	95.7/97.6	98.4	88.2/63.7	94.5/94.1	93.2	77.5	95.9	66.7	93.3/93.8
	4	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
✚	5	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9
✚	6	Huawei Noah's Ark Lab	NEZHA-Plus		86.7	87.8	94.4/96.0	93.6	84.6/55.1	90.1/89.6	89.1	74.6	93.2	58.0	87.1/74.4
✚	7	Alibaba PAI&ICBU	PAI Albert		86.1	88.1	92.4/96.4	91.8	84.6/54.7	89.0/88.3	88.8	74.1	93.2	75.6	98.3/99.2
✚	8	Infosys : DAWN : AI Research	RoBERTa-iCETS		86.0	88.5	93.2/95.2	91.2	86.4/58.2	89.9/89.3	89.9	72.9	89.0	61.8	88.8/81.5
✚	9	Tencent Jarvis Lab	RoBERTa (ensemble)		85.9	88.2	92.5/95.6	90.8	84.4/53.4	91.5/91.0	87.9	74.1	91.8	57.6	89.3/75.6
	10	Zhuiyi Technology	RoBERTa-mtl-adv		85.7	87.1	92.4/95.6	91.2	85.1/54.3	91.7/91.3	88.1	72.1	91.8	58.5	91.0/78.1

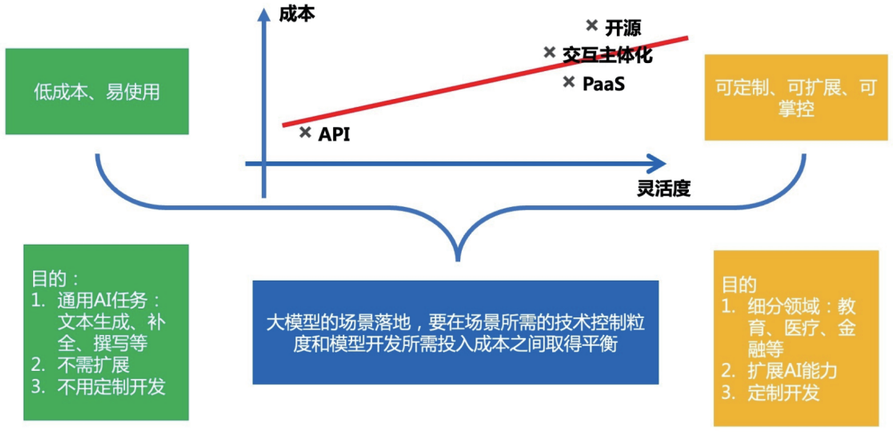
图注：SuperGLUE榜单排名情况
来源：SuperGLUE，2021年8月

总结

从超大规模智能模型技术的发展来看，当前这一领域较为关注的研究重点包括：大规模的模型参数、创新的模型架构、采用多模态的数据进行预训练和微调、高效且性能更强的预训练和微调方法、模型训练和推理加速方法等。大模型在推动AI领域技术发展的同时，也对算力基础设施和智能产业的生态提出了更高的要求。同时，许多研究也开始关注贴近实际应用场景的研发工作，推动大模型在一些场景应用落地，为下一步大模型在产业领域的广泛使用奠定基础。

第二章 大模型产业落地模式

当前全球大模型产业落地仍处于早期探索阶段。一方面，研发机构在大模型方面的技术研发已较为成熟，如算法架构、预训练和微调策略、计算并行加速等，但对落地场景不够熟悉，没有形成完善的商业化模式，需要与下游场景企业合作建立大模型商业模式；另一方面，绝大多数下游场景的企业对大模型没有形成基本的概念和认知，同时也没有能够支撑模型微调的算力，以及定制、二次开发模型的人力资源和技术实力。



图注：大模型落地应用应做到的取舍和平衡
来源：智源研究院整理

总体而言，大模型落地过程中，最为重要的是减少企业需要投入的学习、开发和计算成本，同时还要能够满足企业进行模型开发定制的需求。目前产业领域总体上有四种落地模式，分别如下。

一、API模式

API模式指的是由大模型的研发主体（如科研机构）对完成的模型文件进行封装，将模型在各类任务上的推理能力封装成统一的应用程序接口，对外提供服务的模式。下游企业可以获得这些接口，并按照自身的业务需求，将接口作为可以调用的代码库、程序包等，植入已有的应用和服务中，让大模型的API为整个程序进行赋能。API模式的主要优势在于降低了企业应用大模型的成本。API将大模型内部的所有技术细节进行封闭，企业不需要过多了解模型的技术细节，也不需要投入额外的开发人力和算力资源进行大模型调优。

然而，API模式本身依赖于研发机构强大的科研实力，如果模型本身性能不佳，API将无法满足企业的常规推理任务需求，仍需要根据场景情况进行额外的微调。同时，API本身不够灵活，其作为一种软件间进行交互的形式，本身难以承载高频、高通量的请求。当推理请求的次数和请求数据量过大时，API的响应时间和返回数据质量将难以保证。最后，许多企业由于业务需要，无法将数据通过API传到模型中进行推理，因此不适用于这一模式。目前大模型领域商业落地较为成功的案例是GPT-3。在OpenAI在公开GPT-3论文后，也同时开放了模型的API申请通道，鼓励研究者、开发者、企业从业者研究“好玩”的GPT-3应用，以此促动大模型的产业场景发展。

在开源社区培养了一批GPT-3的开发和用户人群后，OpenAI对外开放API商业运营计划，让GPT-3成为营收来源。去年9月，OpenAI公布了BETA版本的GPT-3模型API商业方案，分为“Explore”、“Create”、“Build”、“Scale”四个版本。其中Explore版本用户可以免费使用3个月，并有10万Token（每1k相当于750个词）的限额。付费版本的加工从100美元/月到400美元/月不等，并有调用量的限额。此外，每个方案分别提供包括API试验场（Playground，用于测试各种API应用）、开发者工具、技术支持（仅限付费版本）等定制化服务，能够帮助企业更快上手API的开发工作。

表4：GPT-3的API定价方案（来源：智源研究院整理，2020年9月）

方案名称	目标用户	具体方案
Explore版	免费用户	免费使用模型3个月，或免费使用模型直到达到10万token的使用量限额，哪个先到按哪个算
Create版	普通用户	100美元/月，当月API的使用限额为200万token，超过部分则按每消耗1k tokens收取8美分的额外费用
Build版	VIP客户	400美元/月，当月API的使用限额为1000万token，超出部分按每消耗1k tokens收取6美分的额外费用
Scale版	大规模用户	没有给出具体定价，将有专人为此类客户提供服务

近日，OpenAI修改了商业版API的价格，其根据模型能够提供的性能，分为“Davinci”、“Curie”、“Babbage”和“Ada”四个版本，价格相比原版本有所降低，也没有调用量限制。

Per-model prices

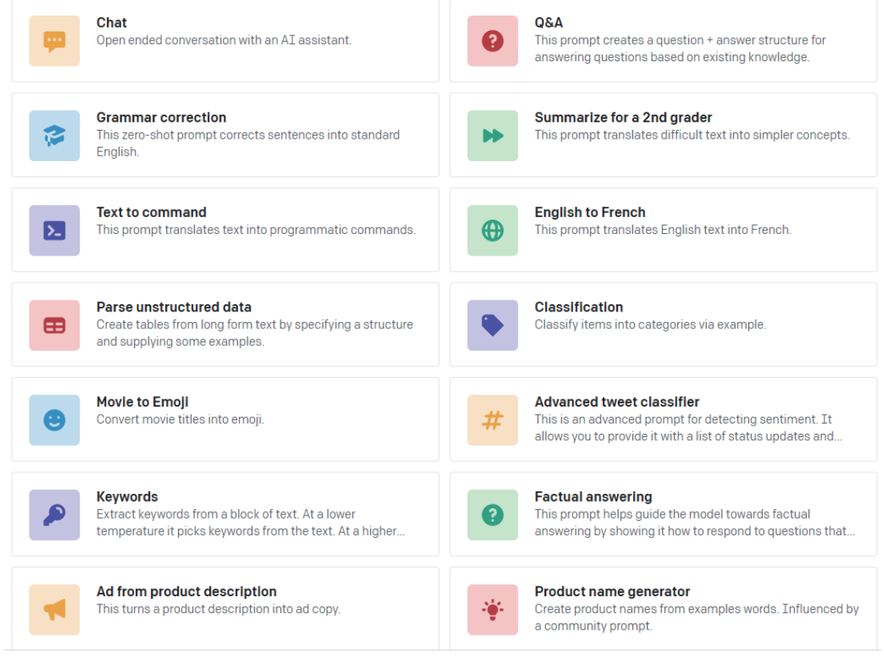
The API offers multiple models with different capabilities and price points. Davinci is the most powerful model, while Ada is the fastest.

Prices are per 1,000 tokens. You can think of tokens as pieces of words, where 1,000 tokens is about 750 words. This paragraph is 35 tokens.

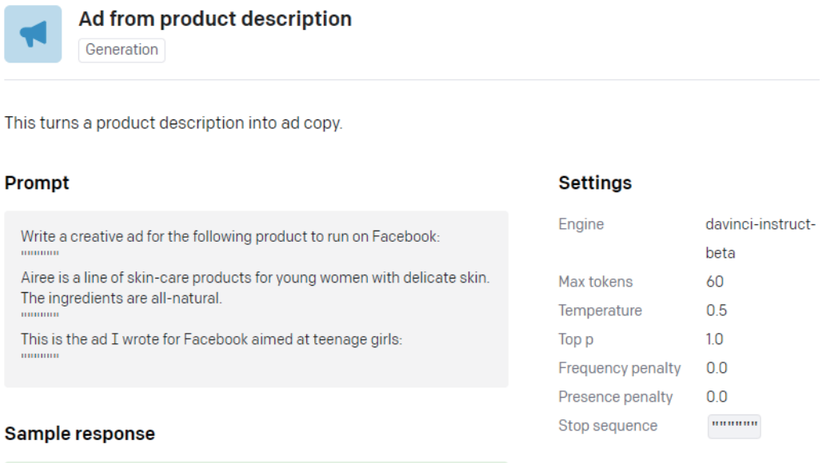
MODEL		PRICE PER 1K TOKENS
Davinci	Most powerful	$0.0600
Curie		$0.0060
Babbage		$0.0012
Ada	Fastest	$0.0008

图注：GPT-3模型API当前的价格
来源：OpenAI官网

新版本的API还为用户提供首次使用3个月赠送18美元额度，按照使用量计费等服务，服务更贴近当前成熟的PaaS、SaaS付费，免除了用户端额外的开发成本。从功能上来看，新版API也更为丰富。截至2021年8月，官网上已有40多个API的应用案例。其中涵盖了对话、问答、摘要、改写等各种需求的API，每一个案例都附带具体的示例，以及用户应该使用怎样的代码向模型请求服务。



图注：商业版API涵盖的类型
来源：OpenAI官网



图注：“根据产品描述生成广告”API的示例
来源：OpenAI官网

以“根据产品描述生成广告”API为例。这一API的功能是将一段产品的描述提供给模型，模型会根据描述直接生成一段对应的广告词。GPT-3商业版的API中有很多可以设定的参数（见上图右侧“Settings”），用户可以在API中提供包括最大返回token数、频次惩罚等参数，帮助模型提供更符合需求的结果。GPT-3提供的每一个API都有示例代码，具体如下：

python Copy

```python
import os
import openai

openai.api_key = os.getenv("OPENAI_API_KEY")

response = openai.Completion.create(
  engine="davinci-instruct-beta",
  prompt="Write a creative ad for the following product
  temperature=0.5,
  max_tokens=60,
  top_p=1.0,
  frequency_penalty=0.0,
  presence_penalty=0.0,
  stop=["\"\"\"\"\"\""]
)
```

目前GPT-3的商业化正在稳步进行中，据OpenAI公开的博客文章表示，截至2021年3月，已有300多个APP采用了GPT-3技术。全球使用GPT-3的开发者数量上万。GPT-3模型每天产生45亿个词汇，且数量还在不断增长。为了推广大模型的应用发展，OpenAI也支持围绕GPT-3举办算法大赛，并提供API支持。

二、开源模式

“开源”指的是将计算机程序、软件的源代码等内容公开，并根据开源协议进行分发的方式。开源是目前计算机领域一种普遍的软件开发模式，大量开发者在协议许可的情况下对开源代码进行修改，并集成到已有的系统中，为软件和系统增加新功能和特性。

人工智能的发展离不开开源的支持。目前已有大量的AI软件库、数据集、算法、代码库等在GitHub等平台上开源，为AI研究者和开发者提供丰富的研发资源。在大模型领域，BERT、ERNIE等预训练语言模型均为开源，对于推动翻译、对话、代码生成、文本生成等领域的应用提供了基础框架。

开源模式的优势在于，它为企业直接提供了免费的AI算法和模型技术，企业获得技术后可以根据开源模型自行研发训练符合业务需求的模型，直接推动了大模型技术从实验室走向产业。此外，开源模式有助于将大量的开发者汇聚在一个大模型项目下，参与方可以共同在此基础上贡献自身的开发能力，为改善模型能力，探索应用场景等贡献自己的力量。大模型从学术界向产业界的渗透和落地依赖于开源社区的帮助，正是因为AI开源社区对大模型的发展保持强烈的兴趣，许多开发者参与到了原始模型的改进工作中，让一些模型从论文和原始复现代码走向了生产领域，成为能够随取随用（on-the-shelf）的开放成果。

然而，开源模式将大模型开发的任务直接转嫁给企业，对于中小企业和初创团队而言难以承担。开源模型在绝大多数情况下是实验室研究的产物，其没有进行符合落地场景的微调，代码可能存在缺陷或漏洞，企业要能够充分理解其中的技术原理，解决存在的问题，才能实现模型训练微调，并集成在已有的系统中。同时，大模型普遍参数规模巨大，直接在系统中部署所需的算力成本极高，很可能出现和企业已有的体系不兼容的情况。这需要企业对模型进行压缩蒸馏等额外的开发工作，也进一步增加了模型的开发成本。

目前开源模式的主要价值在于为大模型领域的AI技术研发提供了参考的先例。许多科研机构和开发者可以在原版模型的基础上进行改造，研发效率更高、应用场景更为广泛的“变体”模型。以BERT为例，谷歌于2018年开源BERT模型，提供模型文件、训练和部署代码等。最初开源的BERT模型代码较为混乱，普通开发者难以改造和使用。很快，在GitHub上该模型吸引开发者研发衍生和改进版本，BERT模型在开源社区迅速发展。截至2021年7月，GitHub平台上和BERT模型相关的开源项目多达1.4万个，这些衍生模型解决了BERT原版代码复杂，难以理解和调用，不兼容其他算法框架，不支持非英语环境等问题。如适用于中文语境的BERT（Chinese BERT）、适用于多种算法框架的BERT（BERT4Keras、BERT-for-pytorch）等。在应用领域，许多科研机构研发了BERT的变体模型，如法国国家科学研究中心（CNRS）构建的法语预训练模型FlauBERT，鹏城实验室研制的“鹏城医疗BERT”等。目前，BERT已成为许多代码库中的一个经典预训练模型，较为知名的有HuggingFace的Transformer库等。HuggingFace是一家成立于2016年的美国AI初创公司，该公司从2017年开始Transformer系列预训练模型开源库的搭建工作，目前已形成了60多个大模型的API库，在GitHub上开源，用户可以直接使用（私有模型服务等为付费项目），并通过开源机制不断Debug，逐渐完善相关服务。



图注：HuggingFace的Transformer开源项目页面，支持用户使用API调用模型能力
来源：HuggingFace的GitHub页面

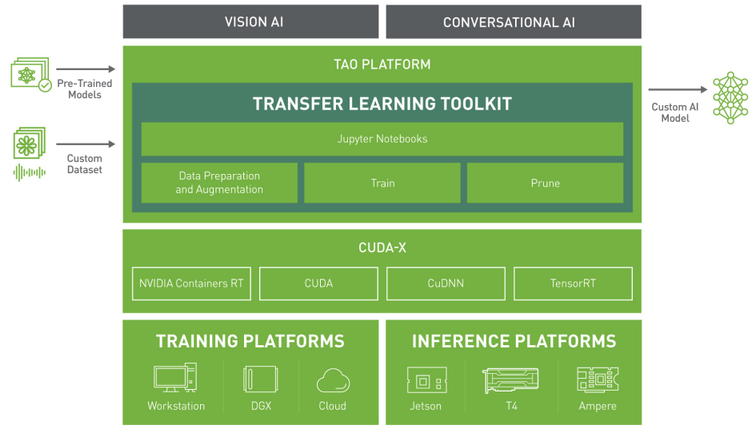
有些大模型并不适合开源，采用分阶段开源的模式更为妥当。例如，OpenAI宣布研发完成GPT-2模型时，称模型在文本生成方面的能力异常强大，生成的文本能够以假乱真，完全开放可能带来社会伦理方面的问题，因此选择分阶段开源模型。到GPT-3公开时，OpenAI没有公开模型，仅提供论文和API试用的申请通道，目前GPT-3的开源版本实际上是由开源社区自行复现的。虽然不排除OpenAI是出于保护研发成果的理由暂缓开源，招致了研究者的批评，但另一方面也保护了模型内部的数据，避免贸然开源导致隐私泄露的问题。同时，许多科研机构公开模型时，也需要考虑模型的数据和知识产权保护问题，如模型的代码、采用的数据集、其生成的结果的版权归属等。

三、PaaS模式

大模型作为一种技术，将其集成在AI平台上，对外提供能力和服务，成为一些技术实力雄厚企业的选择。这种模式中，企业的主要落地产品的PaaS平台、开发工具、AI服务等组合而成的套件，大模型是套件中的一个重要的组成部分，但不是全部。用户购买或使用平台的过程中，可以利用平台提供的工具，对大模型进行开发和应用，集成于自有的体系中，为自身业务提供服务。需要注意的是，和API模式不同，PaaS服务模式提供的是包括模型在内的一系列工具和平台，用户不能单独获得模型的能力。同时，用户虽然可以通过PaaS模式，获得对模型更高的开发和控制权，但往往需要为此付费，这为PaaS服务的供应商带来了一定的利润。

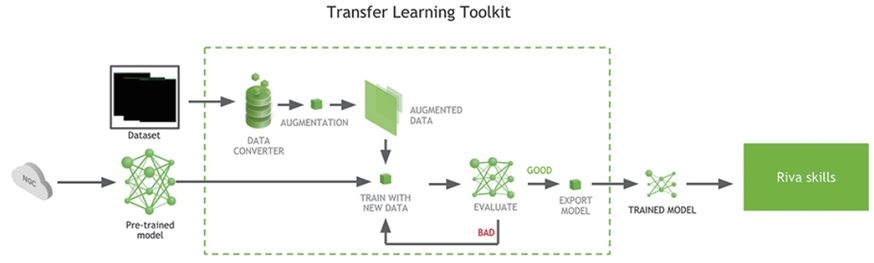
PaaS模式的优势在于，其将大模型直接封装成成熟的AI服务，并为用户提供便利支持，帮助用户以相对开源更低的成本进行模型的开发和部署工作，为用户减少了很多开发方面的投入。但是需要注意的是，有些PaaS不一定提供算力资源，用户仍需要投入硬件成本。同时，当模型规模过大的时候，即使是提供PaaS服务，PaaS供应商仍需要为用户解决模型训练和推理所需的算力规模大，无法集成于用户已有业务体系的问题。PaaS供应商也需要考虑技术的服务形态，如果是本地服务，则PaaS本身的规模绝不能过大，以免给用户带来负担，如果是云服务，则需要解决高请求频次和大规模通信方面的问题。

国际上，英伟达迁移学习工具包（TLT）是一个用于进行预训练模型微调定制的综合平台，包括支持计算机视觉和自然语言处理两个领域的预训练模型，以及数据处理标注、模型训练、模型剪枝等方面的工具，以及支持模型训练和推理的加速组件等。



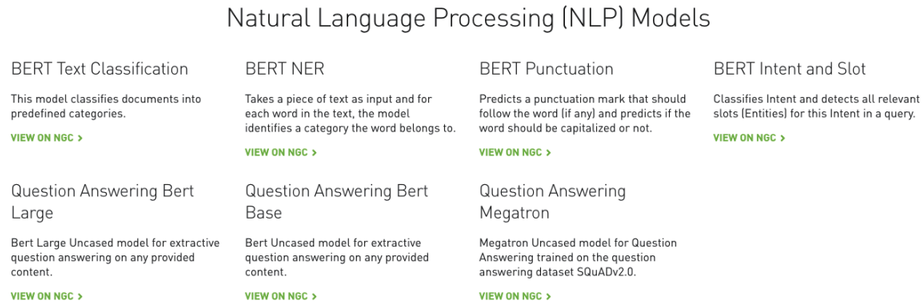
图注：英伟达TLT平台架构
来源：英伟达官网

以微调对话领域模型为例，英伟达提供了具体的部署流程。用户首先可以从云上下载已预训练好的模型，并结合场景数据集，利用TLT工具，对数据进行处理、增强，并对预训练模型进行微调，微调后利用平台进行测评。待测评结果符合实际应用要求后，可导出模型，进行下一步的使用（如后续用RIVA进行模型的应用）。



图注：TLT进行预训练过程中的主要流程
来源：英伟达官网

目前TLT平台已经支持BERT、Megatron等预训练语言模型，以及计算机视觉领域多种任务的模型。英伟达称，在TLT平台的支持下，相比非迁移学习，用户只需要使用十分之一的数据进行训练，模型部署效率提升10倍。TLT平台本身是免费的，支持英伟达旗下的A100、V100 GPU等硬件。



图注：TLT目前支持的自然语言处理模型
来源：英伟达官网

国内也有大型科技企业推动PaaS模式的发展。例如，2020年7月，百度发布了NLP开发与服务平台文心，其依托ERNIE预训练技术，为开发者提供NLP定制与应用能力。文心平台可通过公有云（如EasyDL-NLP、BML等）和私有化两种方式为产业提供服务。用户不需要关注计算开发环境的搭建和算法实现的底层实现细节。



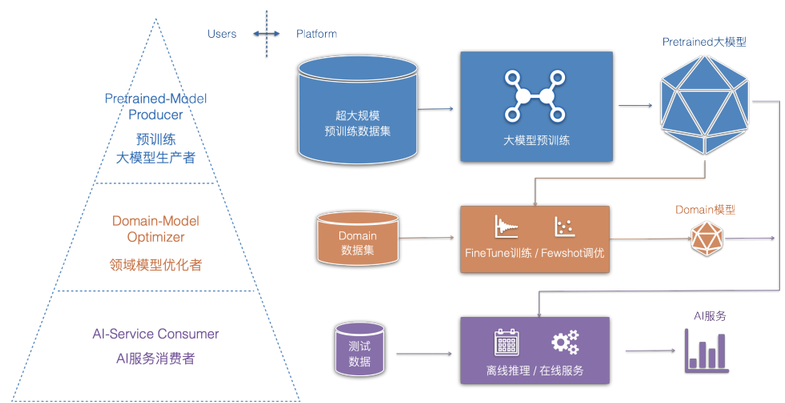
图注：文心平台全景图
来源：百度提供

对于对场景要求更高的用户，百度提供全流程、可深度定制、可私有化部署的文心开发套件。用户可以获得模型技术选型、本地化训练、离线部署等深度解决方案。截至目前，文心平台已累计支持4万开发者，及涵盖各行各业的200多家企业客户。

达摩院的AliceMind平台囊括了该机构研发的一系列基础语言模型，覆盖通用语言模型、多语言、生成式、多模态、结构化、知识驱动、超大模型等领域，大部分已开源，支持训练、精调、蒸馏、测试、部署等功能，支持语言模型从训练到部署。目前AliceMind的日均调用量超过50亿次，活跃场景超过200个，已在跨境电商、客服、广告等数十个核心业务应用落地，以及医疗、能源、金融等行业。例如，浙江电网公司采用AliceMind构建智能化运维平台，用于变压器检修、供电抢修等场景。同时，达摩院也基于M6模型打造了多模态预训练服务平台，集成文到图像生成、视觉问答、图片描述、文本生成、跨模态检索等基础功能，并提供了data2text文案生成等解决方案。



图注：达摩院AliceMind平台
来源：https://nlp.aliyun.com/portal#/alice



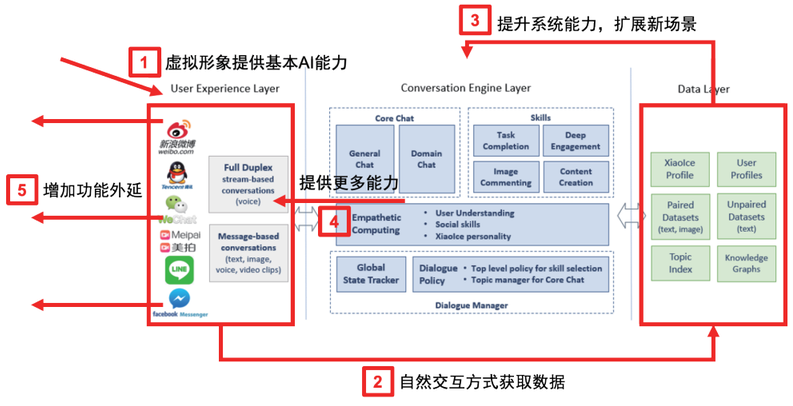
图注：M6平台架构
来源：阿里达摩院提供

京东科技的NeuHub Foundry平台内置了覆盖计算机视觉、自然语言处理在内的近百个预训练模型，提供图形式、脚本式编程方式，支持一键导入预训练模型，实现便捷快速的重训练和迁移学习。此外，NeuHub Foundry支持数据处理、训练到部署的全流程开发服务，训练完成的模型可一键部署并发布至NeuHub开放平台，提供鉴权、计量计费和安全防护等一站式后服务体系。目前NeuHub平台日均调用量峰值超千亿次，覆盖风控、客服、广告、阅读娱乐等几百个应用场景，服务上千家企业客户。

四、“交互主体化”模式

AI交互主体化，指的是基于AI技术构建一个虚拟的，能够和人类产生天然互动关系的虚拟角色形象，集成AI能力，为人类提供服务，由对话AI小冰的研发团队提出，是API模式、开源模式和PaaS模式之外的第四条模型落地模式。相比普通的AI程序，实现交互主体化的AI具有类似人类的属性——具有功能延展的外延。例如，从事“金融摘要”工作的AI，其功能外延很小。一个金融财报摘要AI程序，在人的认知中，只拥有“金融财报摘要”这唯一的属性，人们无法想象其具有外延的可能性。

但具有交互主体的AI形象，会让人们天然地将其视为具有功能外延的“虚拟人类”。在这一基础角色平台上，对其增加各类功能就显得合乎逻辑。交互主体化的AI本质上是从产品的角度，对各类AI功能的统一封装，通过规模化系统的方式进行扩展。AI交互主体化不仅仅是构建一个IP进行市场宣传，实现技术接口和AI能力外延的统一对应。



图注：小冰框架进行迭代的主要流程（本图有修改）
来源：https://arxiv.org/pdf/1812.08989.pdf

智源研究院、小冰和智谱AI联合研发的“华智冰”，就是大模型交互主体化的一个案例。



图注：“华智冰”虚拟形象，其外形、声音等均由AI生成
来源：智源研究院

华智冰本身是一个集成了多种AI能力的虚拟角色。其采用了生成的外形外貌、声线，由悟道2.0模型提供学习和生成能力，在这一形象基础上，逐步扩展对话、写诗、绘画等AI能力。华智冰背后由悟道2.0大模型提供各类AI能力。在交互方面，将所有技术全部封装成为AI的能力，并通过自然交互的方式，与环境进行互动。

华智冰体现出了一种大模型与场景直接进行交互迭代的机制。当需要进行场景落地的时候，企业或许能够通过自然的交互模式，让华智冰在场景中通过对话等形式获得数据，通过不断的迭代，逐步迭代提升模型的推理能力，最终满足场景的特定需求。而数据对AI系统的重要性，在今年以来，得到吴恩达等学者前所未有的重视，吴恩达在3月新设的《机器学习操作（MLOps）——从以模型为中心到以数据为中心》课程中，开宗明义地提出AI = Data + Code，认为要开发出更好的AI系统，80%取决于数据质量，20%取决于模型。

尽管交互主体化的思路具有可借鉴的现实意义，但是是否能够成为大模型落地模式中的主流，仍需要经过实践的检验。目前许多人仍将模型视为系统的一个子模块进行集成，很多应用领域需要评估进行交互主体化的工作的必要性和成本。

总结

当前大模型领域的四种应用模式，各有适用的场景和模式。API模式很大程度上降低了落地的成本和投入，但其受到API技术本身的限制，不够灵活，对于后续的模型定制化开发带来阻碍；开源模式具有较大的灵活性，企业可自由定制使用，但其将大模型商业落地的成本转嫁给了用户端，从实验室原型到商业化完成的难度较大；PaaS模式目前较为成熟，已有相关的商业案例，但这要求PaaS服务的供应商具有雄厚的技术实力，此外PaaS模式可能无法解决落地场景企业投入额外算力和开发资源的问题；交互主体化模式是一种新型的模型落地方式，或能够建立一种自然的交互迭代机制，但是否能够在对话任务以外的领域落地，仍需要经过实践检验。

总体而言，大模型的落地取决于两个方面的因素，一是技术在落地过程中的成本要足够低，低到其带来的效益能够完全覆盖；二是大模型要能够得到足够高级别的封装，包括对技术和能力的封装。封装层可以建立起模型和场景需求间的有效连接，确保落地场景能够通过封装层与模型进行交互。

第三章 大模型商业化主要发展领域

随着大模型技术的快速发展，产业界也在推动模型的场景落地工作。国际上，自OpenAI公开了GPT-3的API后，已经在开源社区和AI创业者中引起反响。国内，多家企业和科研机构研发的超大规模智能模型，已经在营销、广告、金融等领域寻求落地机会。本章节将对目前超大规模模型主要的落地领域和场景进行梳理，并列举一些案例。

一、主要落地领域

在大模型领域，目前主要的应用场景包括：

1. 内容创意生成：基于大模型的专业软件和服务将快速发展，如博客文章撰写、新闻快讯撰写、百科撰写、小说补写、课程题目编写等，形成自动化的内容生产体系。此外，大模型将具备更强的概念组合和创造能力，能够根据人类的提示自动产出概念设计等结果，其创新能力和效率或许将超过设计师水平。

2. 语言/风格互译：大模型将集成在文本、PPT、表格工具中，具有更为智能的能力，能够在软件内进行转写、摘录、自动补全、拼写检查、自动配图、自动翻译等。此外，大模型不仅可以学习语言，对于代码等数据也具有学习能力，其可以在代码补全、辅助开发等方面发挥重要作用。

3. 对话：大模型将成为专业场景和日常场景对话机器人背后的底层技术，通过预训练技术支持模型产生流畅、符合上下文语境、具有一定常识的聊天内容，支持娱乐形象或虚拟陪伴机器人。但同时模型也能够通过微调等方式，学习专业领域知识，在特定专业领域提供支持和服务，成为智能客服。

4. 搜索：通过学习海量的互联网网页数据，结合大模型的生成能力，可为用户提供更为直观，便于理解的检索结果，满足数据挖掘分析、信息检索等需求。

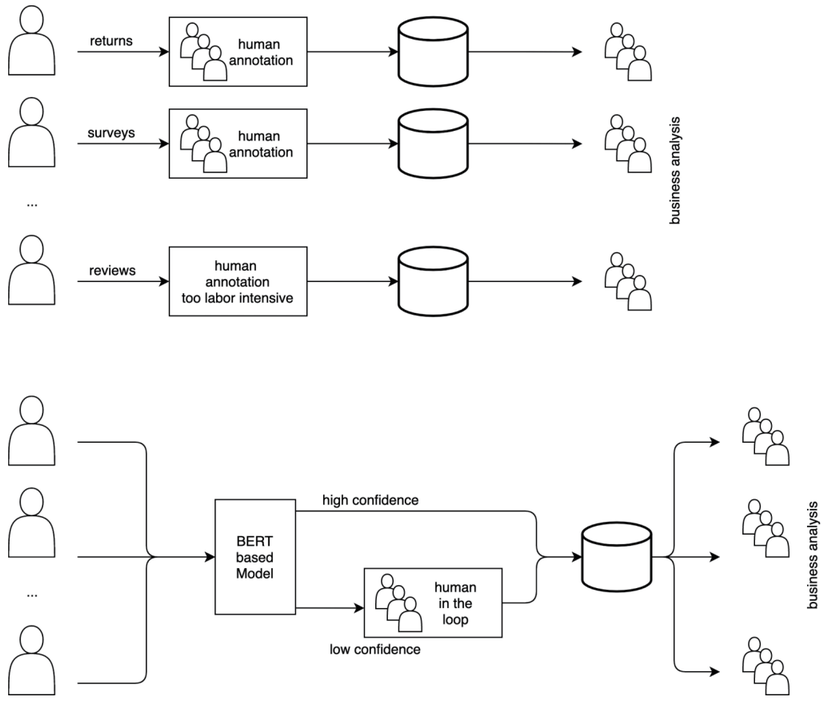
5. 其他：除了上述领域，基于大模型的应用还包括游戏辅助开发等细分领域场景。

二、国际主要大模型落地领域和场景

国际上大模型研究开展的时间较早，在研发过程中积累了一些经验。目前有商业化应用案例的大模型主要包括BERT、GPT-2、T-NLG、GPT-3等。

（一）BERT

BERT是谷歌最早研发的大规模预训练语言模型。其采用自掩码的训练方式，学到了语言的上下文语义关系，适用于对各类语言任务的表征学习，支持进一步微调，提升下游任务的水平。BERT主要适用于搜索、营销、金融等依赖深层语义理解的领域。例如，在检索领域，谷歌目前已将BERT整合到搜索引擎中，为知识卡片和内容索引提供更为准确的结果。在营销领域，一些企业采用了开源的BERT模型。BERT能够将大量的用户反馈内容进行整合，为下游的营销动作提供支持。

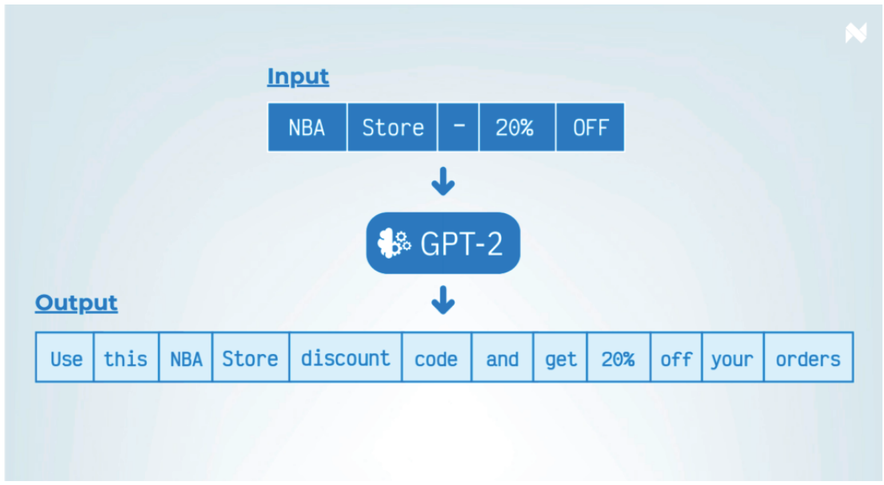


图注：BERT在营销领域应用的方式
来源：Wayfair官网

需要注意的是，由于BERT采用的英文预训练语料，这是阻碍模型落地的一个难题。许多BERT衍生的模型都进行了本地化，包括采用非英语语种预训练，采用多种领域文本数据进行训练等。

（二）GPT-2

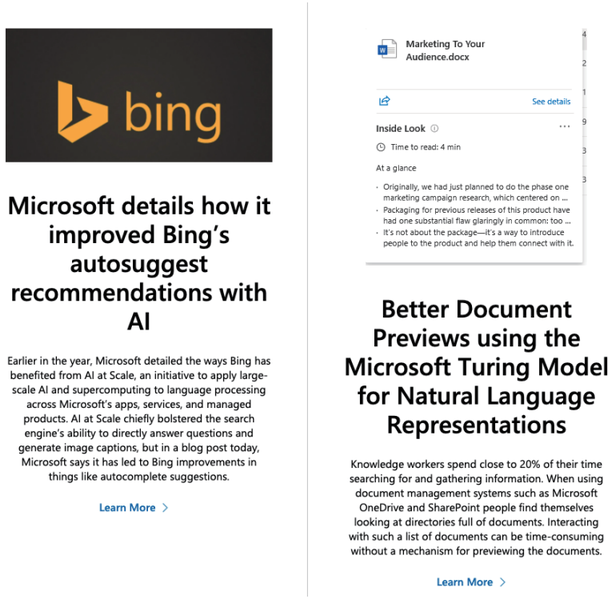
GPT-2是OpenAI研发的一个预训练模型，能够生成多种类型的文本。问世伊始，OpenAI认为该模型的性能非常强大，因此采用了分阶段公开的策略。和BERT不同，GPT-2更适合文本生成领域的任务，在营销、文案、对话、媒体等领域的应用较多。例如，有开发者将GPT-2集成在了文字类冒险游戏中，用户通过输入特定的文本，与AI进行互动，推动游戏进程。另外，GPT-2也可以在对话领域实现应用，如Telegram群聊对话机器人等。



图注：GPT-2根据关键词扩写营销文案的过程
来源：Narrativa官网

（三）T-NLG

T-NLG模型是微软在2020年2月研发的一个超大规模智能模型，参数规模达到175亿，该模型也是自然语言生成模型，在文本生成、摘要等方面具有更强的性能。此外，得益于微软在底层硬件体系和上层应用生态方面的积累，T-NLG的商业落地进程明显比前辈们要快。据微软透露，T-NLG模型已经在微软的生产力套件（Word等）中得到应用。T-NLG可以为邮件生成自动回复，或通过分析Excel表中的数据规律进行自动填充。T-NLG也能够帮助用户生成更为精准的文本摘要，为Bing带来更为精准的自动联想关键词，为聊天机器人提供更好的训练数据等。



图注：T-NLG的应用案例
来源：微软官网

（四）GPT-3

GPT-3是目前商业化进展最快，应用案例最多的大模型。OpenAI透露，目前全球有300多个app在使用GPT-3的技术。本报告通过结合gpt3demo网站的数据，以及公开新闻报道，收录了158个基于GPT-3的应用（截至2021年7月）。

基于GPT-3的下游应用主要集中在以下领域（根据gpt3demo分类进行整理）：

1.内容创意生成
2.对话
3.语言/风格互译
4.搜索
5.其他

内容创意生成方面，主要包括文字生成、图片生成、语音生成、表格生成、游戏内容生成、颜色、其他等，共计109个应用。细分领域的具体应用数量如下。

表5：GPT-3在内容创意生成领域的应用情况

领域名称	细分领域	领域描述	应用数量
文字生成	通用写作助手	基本文案、写作内容、社交媒体文本的扩写、续写、补写	27
	细分写作助手	根据营销、广告、艺术创作、招聘就职、企业服务等方面的需求，提供对应的文案生成能力	30
	其他文字内容生成	生成技术开发文档、教程、知识库、创业想法等其他内容	19
图片生成	图文自动生成	针对图片生成文字描述，或根据文字描述生成 PPT 等设计	9
	前端设计	根据文字描述生成 UI、网页等	4
语音生成	-	根据文字描述生成 AI 语音等	2
表格生成	-	根据描述创建表格、图表等，包括财务报表、问卷表等	7
游戏内容生成	-	根据描述生成游戏场景、NPC 对话文本、互动游戏内容等	6
颜色生成	-	根据描述调整色阶，或为表情包设计新色阶等	3
其他	-	生成投资建议、菜谱等	2

对话方面，主要包括各类聊天机器人（偏向领域内或不限定领域的开放对话机器人），以及客服（满足客户响应需求的机器人）两类，共计17个应用，其中非限定领域的聊天机器人占大多数，共15个，具体如下：

表6：GPT-3在对话领域的应用情况

领域名称	领域描述	应用数量
聊天机器人	各类细分领域的聊天机器人，如虚拟偶像、医疗问诊、虚拟心理医生、仿名人语气的聊天机器人、陪伴对话机器人等	15
客服	与客户进行交互、回复问题等需求的客户	2

在语言/风格互译方面，包括编程语言转换、言法/语法转换，以及其他文本转换方面的应用。总共有39个。其中，编程语言转换方面的应用占大多数，有31个，具体如下：

表7：GPT-3在语言/风格互译领域的应用情况

领域名称	细分领域	领域描述	应用数量
编程语言转换	自然语言转编程语言	将用户的自然语言描述转换为对应的代码，包括网站开发、各种编程语言代码的生成和补全等	30
	编程语言转自然语言	根据代码内容，为用户生成对应的摘要、注释等	1
言法/语法转换	日常语言转专业语言	将日常用语转为专业领域的用词、文法，如日常用语转法律文本、专利文本等	3
	专业语言转日常语言	将法律、专利、医学、金融等领域的用词、文法、文风转换为日常用语	1
其他文本转译	-	对语言风格进行修饰，转换为符合需求的文本内容，提升文本的可读性，或增加文本被搜索引擎排名靠前的可能性等	4

在搜索领域，主要包括文字搜索和图片搜索两个领域，共计8个应用，其中文字搜索应用占大多数，具体情况如下：

表8：GPT-3在搜索领域的应用情况

领域名称	领域描述	应用数量
图片搜索	根据文字提示或“文字+图像”的方式搜索合适的内容	2
文字搜索	法律、书籍、电影等内容的搜索，这些内容信息量大，不容易精准查找	6

其他方面，GPT-3衍生应用还包括AIOps、客户反馈分析、软件对比测试、游戏语音助手、做题解题等。

三、我国主要大模型落地领域和场景

我国在研发大模型方面相比国外有一定差距，但企业对AI技术的需求高涨，为业务需求提供AI能力支持成为新趋势。目前我国AI模型落地的领域主要集中在和互联网经济相关的营销、广告、娱乐、媒体、教育等领域。

（一）悟道（1.0/2.0）

2021年6月，智源研究院发布悟道系列超大规模模型，并公开了悟道在产业落地方面的情况。在任务方面，悟道适用于文本、图文、视频等任务，在新闻、传媒、搜索、营销、教育等领域与国内企业和单位合作，探索落地应用的可能性。此外，与27家产业单位达成战略合作意向，就悟道预训练大模型进行产业应用合作，从底层数据资源、中层模型研发到上层AI应用推进产业应用生态的建设，从单点突破提升至全方位突破，并成立公司，进行模型的运营和产业应用工作。



图注：“悟道”模型产业生态合作意向单位
来源：智源研究院

同时，围绕悟道模型的产业化工作，智源研究院构建了悟道模型平台，为开发者提供支持。例如，在检索式问答系统中，通常需要借助人工标注的方式来建设高质量的语料库。但这种方式建设周期长、投入成本高，难以大规模实施。“悟道·问答生成”借助GLM模型强大的小样本学习以及文本生成能力，结合知识图谱，可以大规模生成QA语料库。这种新的技术范式，可以低成本的方式回答现有问答系统无法回答的问题，目前在OPPO小布助手上得到成功验证，双方正在持续迭代升级中。

下一步，智源研究院将在预训练模型能力和通用性不断迭代提升的基础上，积极拓展已有产业生态规模，并与战略合作企业一道积极探索新的生态合作模式。



图注：悟道模型平台界面
来源：智源研究院

（二）盘古系列大模型

4月，华为云与循环智能联合发布盘古系列大模型，包括自然语言处理、计算机视觉两个模型。信息发布时，华为提到国家电网智能检测项目采用了盘古CV模型进行辅助。目前盘古已扩展到多模态和科学计算等领域，具有更多能力。据公开资料显示，目前盘古大模型正处于探索落地阶段。



图注：盘古模型内涵和应用领域
来源：华为云官网

（三）阿里M6/PLUG

达摩院探索M6模型的应用，于6月实现商业化落地，目前支持阿里内部30多项业务，日调用量上亿次，已部署于阿里新制造平台犀牛智造，并与国内品牌技术合作，结合潮流趋势进行快速设计、试穿效果模拟，缩短时尚新款服饰设计周期。M6还已应用于支付宝、淘宝等平台，参与跨模态搜索、文案撰写、图片设计等工作。

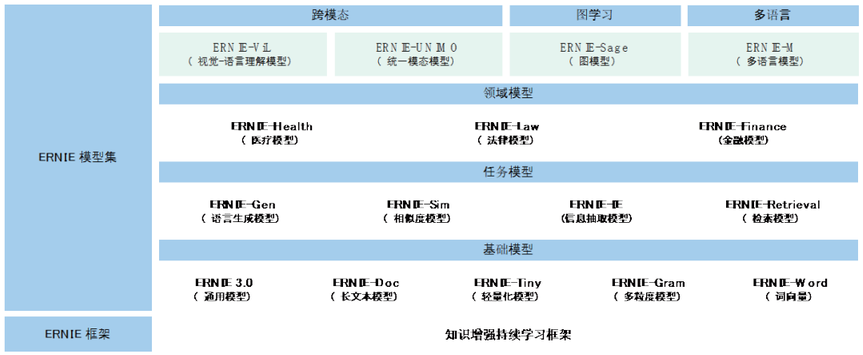
4月，阿里达摩院发布PLUG大模型后，将其整合到AliceMind自然语言处理平台上。目前已开放模型的使用通道，在电商、对话、翻译、电力、医疗、金融等领域实现应用。此外，AliceMind提供了模型微调、蒸馏、压缩等方面的工具。目前AliceMind平台上的应用案例包括对话机器人、企业知识库、医疗领域的信息抽取等。



图注：AliceMind平台支持的落地领域
来源：AliceMind官网

（四）百度ERNIE模型

2021年7月，百度将ERNIE模型更新至3.0版本。目前已形成文心NLP开发平台。用户可利用平台、模型、开发工具等，定制适用于自然语言处理的相关应用。



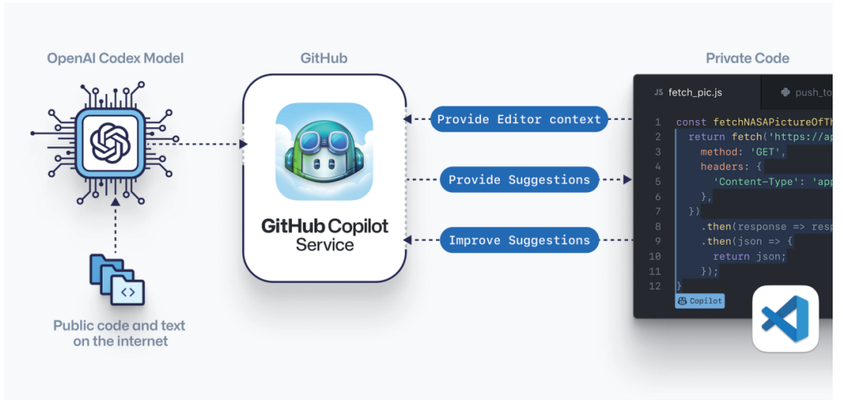
图注：文心ERNIE的技术能力
来源：百度提供

目前，文心ERNIE已大规模应用于搜索、信息流、智能音箱等产品，并通过百度智能云的形式落地工业、能源、金融、通信、媒体、教育等行业，助力产业智能化升级。

四、基于大模型的落地应用案例（以GPT-3为例）

（一）Copilot

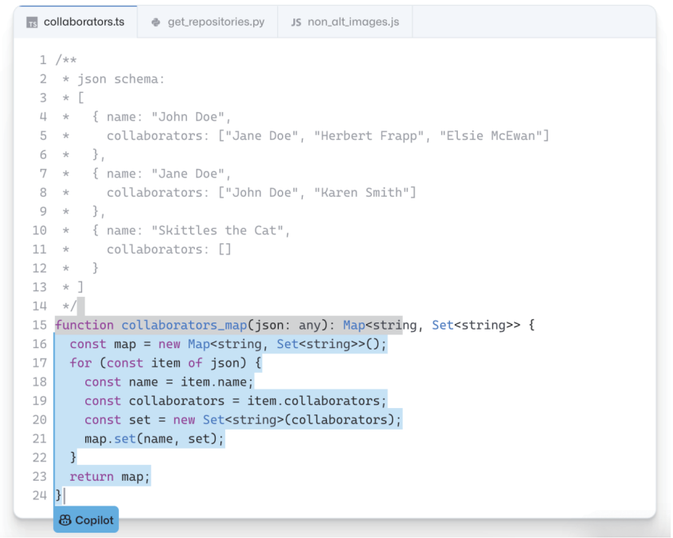
2021年6月30日，OpenAI联合微软、GitHub推出了基于大规模预训练语言模型研发的Copilot代码生成插件。Copilot基于OpenAI Codex模型，和GPT-3采用相似的技术，在数十亿的GitHub公开代码上进行训练。Codex模型同时也进行了自然语言数据的训练，因此可以理解自然语言。



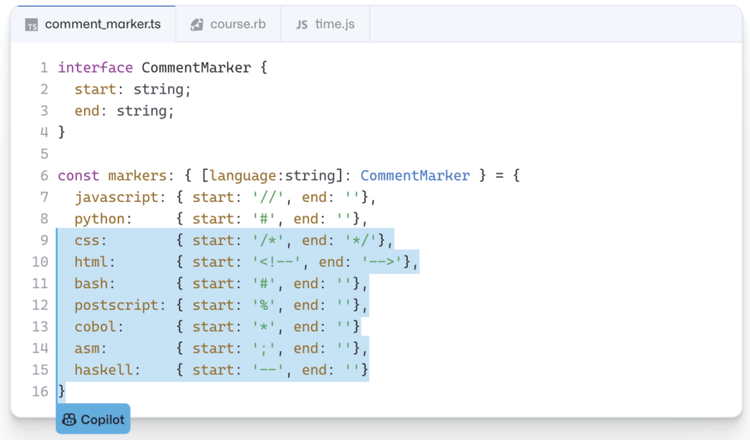
图注：Copilot的技术架构
来源：OpenAI Codex介绍

Copilot采用插件的形式，集成在微软的编辑器VS Code中，用户可以在代码开发界面提供开发意图等信息。Copilot会根据这些信息生成对应的代码建议。根据OpenAI提供的案例，Copilot主要可以辅助三种开发任务。

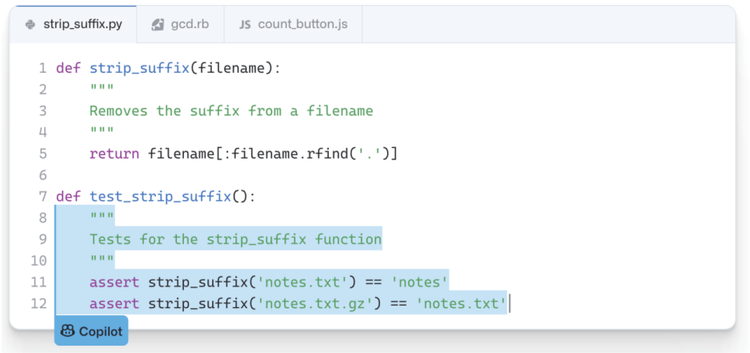
1.代码生成：根据注释中的开发意图，生成对应的代码。



2.代码补全：根据已经完成的代码段，将剩余的重复性代码补全。



3.测试用例生成：生成测试函数。



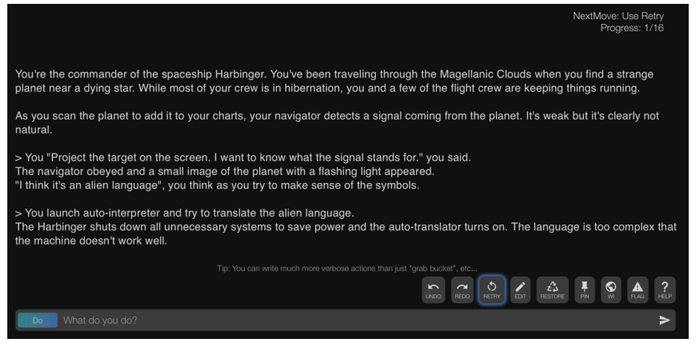
由于基于“大参数+大算力”进行训练，Copilot具有非常惊人的性能，许多用户给出了满意的评价，一些评论甚至认为Copilot带来了一种新的代码开发模式。8月11日，OpenAI发布了更新版的Codex模型，可以自动生成游戏代码等。

然而，在Copilot开放给用户试用的过程中，开发者很快发现了其存在的一系列问题，具体如下。首先，Copilot并不是总能够生成全新的代码。有一些代码是直接从训练数据中复制出来的，甚至连注释都一字不差。其次，Copilot在开放时宣称其训练代码为GPL（General Public License）开源代码。但有用户爆料称，GitHub提供的训练代码并没有区分协议类别。GPL开源协议是开源协议的一种，其允许任何人获取使用该协议的代码，但采用了这些代码的软件必须是开源的，也必须使用GPL开源协议。通俗来说，GPL的代码不能被用来商业化使用。这样一来，如果Copilot会大段复制原来的代码，然后采用商业化付费的形式提供给用户，用户可能在未经原作者授权的情况下，将这些本应开源的代码进行商业化。这不仅仅是对开源协议践踏，也会导致商业软件中出现不允许使用的代码，从而带来版权问题。

最后，Copilot有可能泄露代码中的隐私信息。根据剑桥大学的一项研究，GPT-3可以直接将学到的隐私信息还原出来。而技术同源的Codex模型，也可能在特定情况下，将代码中的敏感信息泄露出去。

（二）AI Dungeon

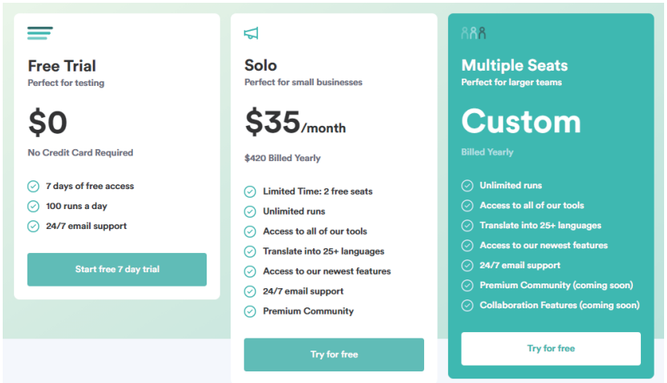
AI Dungeon是开源开发者基于预训练语言模型技术开发的文字冒险游戏。在游戏中，用户可以先定义一个游戏场景，然后AI会对情节展开，并在关键的剧情节点，让玩家“续写”游戏剧情，AI会根据输入继续推进剧情，直到达到结局。



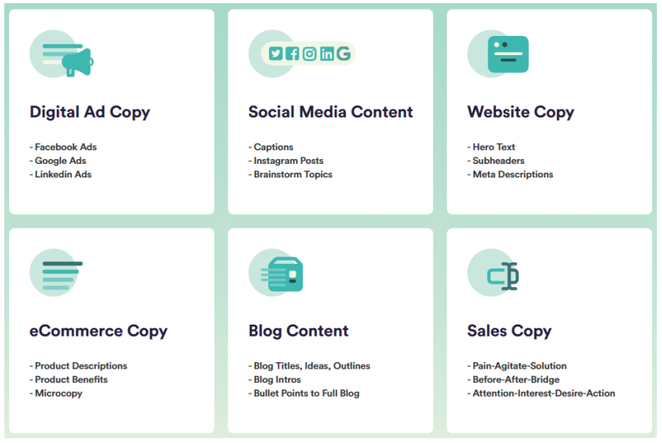
图注：AI Dungeon的游戏界面，用户在对话框输入行为后，AI将自动推进剧情
来源：AI Dungeon官网

（三）CopyAI

CopyAI是美国一家AI初创企业，成立于2020年。该公司主要提供包括数字广告文案、社交媒体内容、网站内容、博客内容等的文案自动撰写服务。当前产品已进入商业化阶段，其中“SOLO”版本价格为35美元/月，享受包括两个账号，25种语言翻译、多种文案工具等服务。当前已有多家企业采购了CopyAI的广告营销产品，包括雀巢、Ebay等。



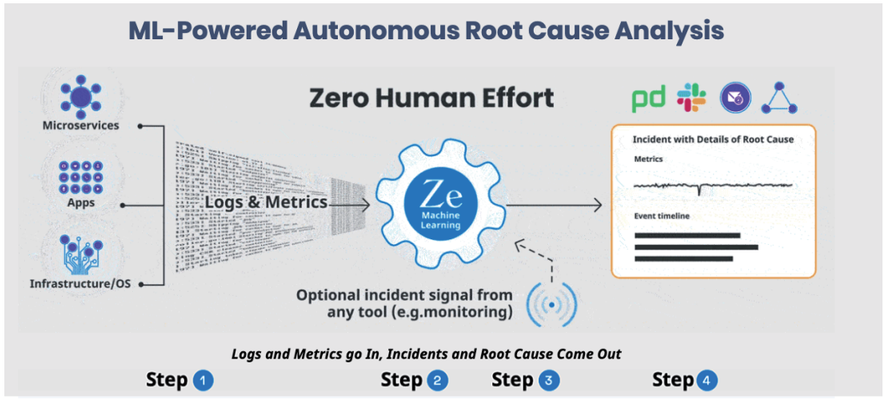
图注：CopyAI服务价格
来源：CopyAI官网



图注：CopyAI提供的产品能力
来源：CopyAI官网

（四）Zebrium

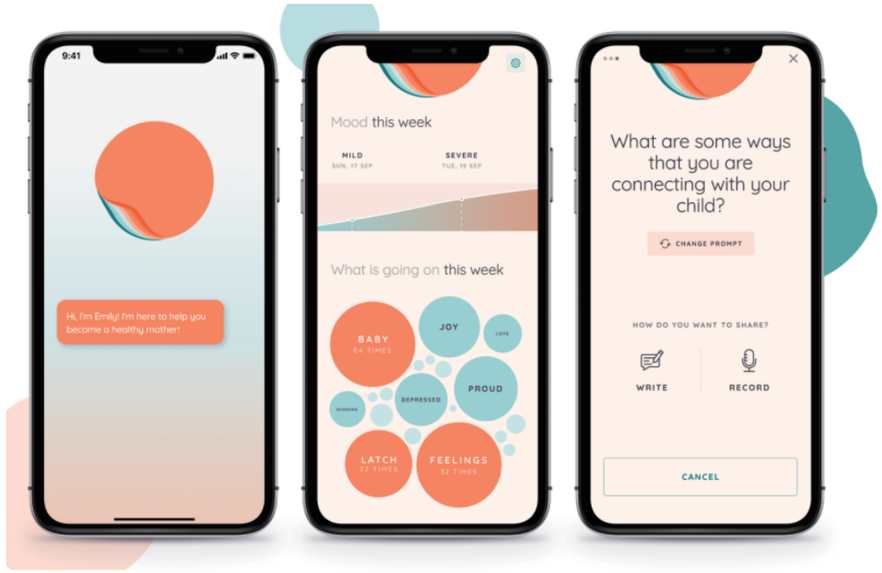
Zebrium是一家美国的IT运维和管理公司，成立于2018年。该公司的主要产品为AIOps，通过无监督机器学习的方法，对企业的IT运维日志进行分析，自动查找Root Cause（导致故障的根本源头）。具体而言，Zebrium利用识别事件的Root Cause，并自动生成简明的报告，寻找问题序列中的第一个事件(通常是Root Cause)，以及其他相关事件和相关的指标异常。通过集成GPT-3，Zebrium的产品可以采取下一步动作——通过扫描整个互联网，查找类似的事件，为用户提供报告，并将Root Cause报告提炼为简洁的自然语言摘要。



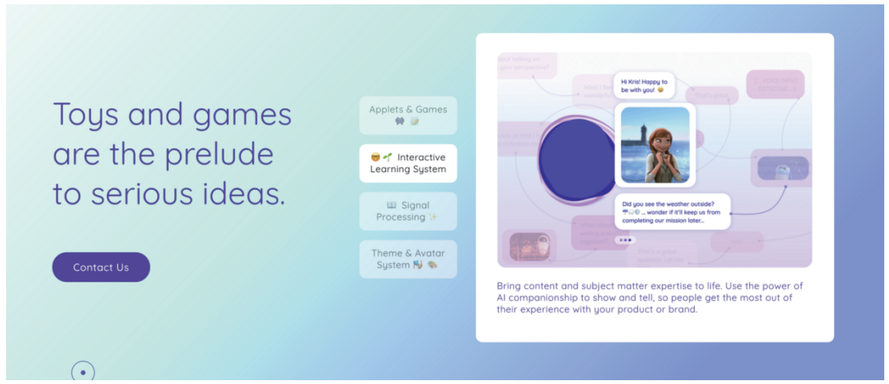
图注：Zebrium产品的工作流程
来源：Zebrium官网

（五）Maslo

Masolo是一家美国的初创公司，成立于2017年。该公司的主要产品是提供UI、视频等交互界面的设计和解决方案服务，帮助品牌商和客户产生人性化、自然的交互体验，特别是设计更为人性化的陪伴助理（Companion），提升用户粘性和口碑，主要包括医疗培训、母婴辅导等方面的APP交互UI。该公司目前使用GPT-3进行对话、情感分析、内容过滤、协同创作等工作。



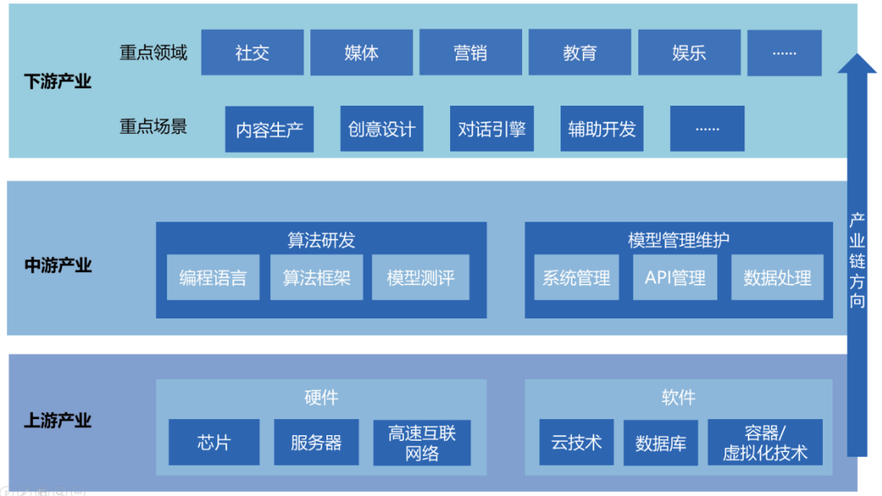
图注：Malso设计的交互UI
来源：Malso官网



图注：AI在Malso产品的内容方面发挥的作用
来源：https://medium.com/maslo/openai-codex-and-maslo-hero-a4180553b63e

五、大模型产业链

在不久的未来，大模型产业结构将进一步分化，形成“上游-中游-下游”三层的产业链格局。总体架构如下：



图注：大模型产业链
来源：智源研究院整理

（一）上游产业

上游产业包括支撑大模型研发、运行等的一系列产业的总称，总体上可分为硬件、软件两部分。其中，硬件上包括通用计算芯片、AI计算加速芯片、计算服务器、存储服务器、高速互联网络等。软件包括云计算环境、数据库组件、容器和虚拟化技术等。

（二）中游产业

中游产业主要围绕大模型技术研发、管理和运维发展。技术研发方面，主要包括算法研发所需的编程环境（语言）、算法框架等工具，甚至可能出现专业的模型测评机构。同时，可能出现一批基于模型技术开发平台、服务和软件开发的企业。

（三）下游产业

下游产业主要是大模型重点落地的场景，这些场景可以和教育、营销、社交、娱乐等产业结合，形成丰富的产业形态。

总结

当前，基于大模型领域的应用正在快速发展。由于目前预训练模型能够完成的任务基本上都和自然语言相关，模型在内容创意生成、语言/风格转换、对话、搜索等领域的落地进展最快。此外，利用大模型进行编程，也成为探索的热点。随着未来模型性能的不断提升，模型相关的平台不断完善，大模型也将是下一代的AI平台，为各行各业赋能。

第四章 大模型领域存在的问题和解决案例

正如前文所述，大模型发展至今，也带来了一些问题。一是由于模型本身参数规模大、性能相对较强，而配套产业、监管体系相对薄弱带来的问题。二是由于基于深度学习体系本身带来的一系列问题，以下将列举这些问题，并提供现有的一些解决思路和案例。

一、大模型存在的问题

（一）模型在可解释性和可控性方面依然无法达到要求

大模型的可解释性和可控性依然较为薄弱。由于大模型采用的是深度神经网络架构，模型参数往往超过亿级别，且采用多种策略进行并行加速技术。因此，对于单一数据，在模型中的处理过程难以追踪，也很难获得模型产生对应推理结果的有效解释。同时，深度神经网络模型的推理结果有一定的不确定性，导致大模型的推理结果并不是唯一确定的，每次大模型的推理结果都可能有所不同。产业界对于系统的输出结果有严格的控制要求，如果模型不能满足可控和合规的需求，其应用价值就无法得到体现。

（二）算力门槛要求高

大模型极大依赖高性能计算机提供算力资源，对于研发机构和企业有较高的门槛限制。目前全球研发大模型的机构基本为大型科技企业（谷歌、微软、Facebook、百度、阿里等），以及具有国家意志和强大投资人支持的科研机构（OpenAI等）。

（三）缺乏产业落地配套的工具

目前大模型领域缺乏产业落地配套的工具。大模型参数规模庞大，算法架构复杂，对模型进行训练、微调等工作依赖专业的配套工具加持。目前在产业界，依然缺乏简单易用、标准化的配套开发工具。一方面企业开发大模型配套应用的开发投入较高，另一方面大模型落地过程中不能做到标准化，模型需要对各类场景进行适配，不利于推广模型的使用。

（四）大模型相关的监管体系尚未建立

关于大模型能否在一些场景进行落地，目前国内外都缺乏相应的政策、法规和应对机制，从规避风险的角度，许多企业不愿意贸然引入大模型技术。一是在版权方面，模型如果采用互联网公开数据训练，在特定领域的微调，是否对某些信息的版权造成侵害?根据开源模型进行二次开发和微调的模型，其算法、参数和生成的数据，和原版模型的关系是什么?二是模型的推理结果不是完全确定的，在媒体、教育等领域，监管机构无法对其生成的结果进行提前审核，落地风险较大。三是模型本身可能存在偏见等问题，在一些领域落地可能加剧歧视和社会不公。

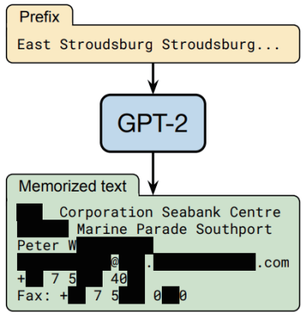
目前在AI领域，许多伦理治理方面的工作处于早期发展阶段，目前已有一些社会共识和准则性质的治理手段，但这对于推进AI良性发展仍有不足。建立“自下而上”和“自上而下”相结合的AI治理体系，是全社会应当共同努力的方向。

（五）大模型产生的环境和伦理问题

大模型加剧了AI领域的许多伦理问题，如模型公平性、隐私保护、数据安全等。

一是大模型为AI的伦理治理水平提出了新的挑战。大模型所需的预训练数据规模极为庞大，完全由人工进行收集、标注和检查并不现实，需要大量由自动化程序收集的数据集。这样一来，训练数据集中可能存在样本不平衡、歧视偏见、粗俗、暴力、色情等内容。

二是模型也可能存在侵犯个人隐私等问题。例如，科学家发现，GPT-2模型在推理过程中可能会直接泄露学习过的隐私信息，如完整的姓名和银行卡号码等。此外，黑客也可能通过技术手段，从模型层参数中还原原始的输入信息。



图注：对GPT-2模型输入特定词语，可直接获得原始隐私信息
来源：https://arxiv.org/pdf/2012.07805.pdf

此外，大模型凸显了AI模型计算能耗较大的问题。如果采用化石能源供电，其产生的碳排放将会十分惊人。

Model	Hardware	Power (W)	Hours	kWh·PUE	CO_2e	Cloud compute cost
Transformer$_{base}$	P100x8	1415.78	12	27	26	$41–$140
Transformer$_{big}$	P100x8	1515.43	84	201	192	$289–$981
ELMo	P100x3	517.66	336	275	262	$433–$1472
BERT$_{base}$	V100x64	12,041.51	79	1507	1438	$3751–$12,571
BERT$_{base}$	TPUv2x16	—	96	—	—	$2074–$6912
NAS	P100x8	1515.43	274,120	656,347	626,155	$942,973–$3,201,722
NAS	TPUv2x1	—	32,623	—	—	$44,055–$146,848
GPT-2	TPUv3x32	—	168	—	—	$12,902–$43,008

图注：训练大模型所消耗的能源和碳排放（单位：lbs）
来源：https://arxiv.org/pdf/1906.02243.pdf

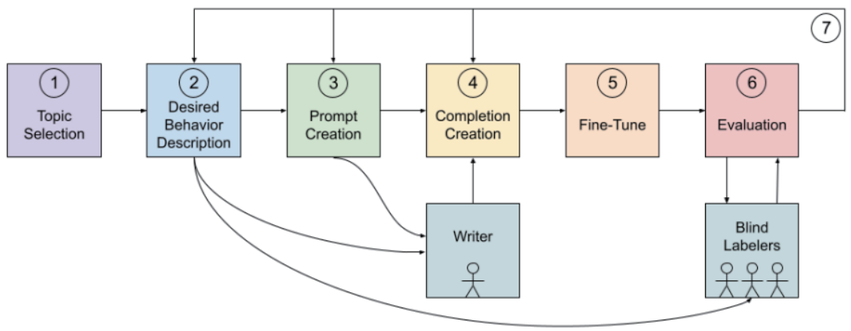
（六）大模型效果评估问题

当前大模型的效果评估往往局限于传统的文本理解和生成类数据集，存在样本量小、领域单一等问题，小模型和大模型的表现差距还没有明显，研究人员难以观察到大模型更大的价值。同时，大模型性能迅速提升，很快超越了许多基准测试集，这对面向大模型性能评价的标准、数据集和基准测试提出了更高的要求。

二、解决案例

当然，大模型并不是一个单纯的“麻烦制造者”。许多企业和研究机构正在着手解决大模型带来的一系列问题。在降低算力门槛方面，许多大模型的研发机构都在尝试提供规模更小、性能损失不大的小模型作为替代模型，或通过PaaS平台的方式提供轻量化的模型开发和推理解决方案。在未来，一些搭载大模型的AI平台或将采用云服务的方式，由云计算彻底解决算力问题。

在伦理方面，随着大模型的不断发展，研究人员也针对大模型的问题提出相应的解决方案。不久前OpenAI提出了修正GPT-3偏见的方法，对解决GPT-3带来的伦理问题起到了推动作用。



图注：OpenAI提出的纠正GPT-3偏见的整体流程
来源：https://cdn.openai.com/palms.pdf

一些研究者考虑用大模型解决AI存在的一些问题，或能够弥补大模型本身带来的问题。例如，美国华盛顿大学和艾伦人工智能研究所的科研人员借鉴了生成对抗网络（GAN）的思想，基于GPT-2架构，提出了识别AI假新闻的大模型Grover，该模型最大参数达15亿，识别AI假新闻的准确率达92%。



图注：Grover的基本思路
来源：https://arxiv.org/pdf/1905.12616.pdf

此外，一些人工智能领域的协会和机构逐渐开始重视伦理问题。例如，GPT-3论文作者中有专门的伦理和可持续发展研究人员，对模型带来的社会、伦理和环境影响进行评估。此外，国际计算语言协会（ACL）已经在ACL2021大会中设置伦理委员会，对提交的论文进行伦理审查。ACL也将成立专门的委员会，负责AI伦理方面的工作。

在环保方面，目前已有很多科学家和企业意识到了大模型带来问题，主要的解决方法包括：创新AI超算的供能方式，如采用新能源（风能、太阳能等）实现供电，保证数据中心不产生新的碳排放；改进算法计算效率，在同样能耗下实现更大规模的模型训练；要求研究者在AI研究中加入有关社会影响的内容，评估模型应用后带来的影响，为下一步技术发展的方向提供参考。

总结

大模型的快速发展也带来了产业、伦理、环保等方面的问题，但这些问题可以通过各种手段解决。随着模型配套的产业生态逐步建立和完善，大模型技术将变得更为可靠和可控，不论是在智能产业体系中，还是在解决人类社会面临的问题上，都会发挥自身的作用。此外，大模型和产业体系融合的过程中，将涉及到包括AI研究者、企业、政府等多方面的角色，我们不仅要建立各类伦理方面的准则，还应当通过“自下而上”和“自上而下”结合的治理体系，来减少大模型发展过程中带来的各类问题。

第五章 下一步工作建议

为了解决大模型带来的一系列问题，本章节从技术、产业、伦理、环保等角度提出下一步工作建议。具体内容如下。

一、持续推进大模型及相关技术体系的发展

（一）研发更为绿色环保的模型技术

考虑到模型带来的高能耗问题，降低模型算力需求，提升计算效率的技术将进一步发展。一是在算法架构方面，类似FastMoE这样的优化算法将不断涌现，在同样参数规模的条件下极大降低激活的神经元数量，保证模型可被训练；二是在模型剪枝压缩方面，许多技术将能够更为精准地降低冗余参数，甚至具备“Structure-aware”的剪枝能力；三是在底层硬件加速方面，通过高效的硬件系统设计，提升AI超算的带宽和显存，提升加速能力。

（二）推广低门槛的大模型开发和使用工具

由于大模型在产业领域落地，需要企业对模型进行微调和修改，提供高效易用的配套工具，建立大模型产业的相关生态，有助于帮助企业最大程度挖掘模型的潜在价值，推动AI前沿技术的发展。一是在模型开发工具方面，提供数据收集、处理、标注等方面的工具，方便用户获得预训练和微调数据，打造领域数据集；二是提供模型微调、剪枝、蒸馏、压缩、计算加速等方面的工具，帮助用户更快将大模型部署在业务系统中，解决各个场景中存在的特殊性问题；三是提供基准测试集等模型性能评价工具，帮助用户更好地发现模型存在的问题，提升技术指标。

（三）建立动态、面向应用场景的大模型评价体系

随着大模型性能的快速提升，许多过去难度较高的基准数据集正在被一一攻克。基准是引导AI技术发展的“领航员”，动态、前沿的基准测试有助于研究者发现模型存在的性能问题，并着力改进。未来在大模型领域，将会出现更多适用于专业领域、不断动态更新、具有更大规模及更多模态的基准测试数据集，让大模型能够真正具有符合应用需求的性能。

（四）建立创新型大模型技术体系

尽管目前已有多种模型应用落地的方式，但这些模式的成效还有待观察。由于大模型是真实世界的表征，是否有更为快捷简便的模式，能够对这些表征进行管理、调取、查询、存储、更新，可能是学术界和产业界值得共同探索的问题。是否能够探索出对大模型进行统一的管理和调用的新型模式，采用统一的交互模式和编程语言，从而将大模型集成到已有的IT体系中，降低大模型的应用门槛，使其成为人工智能领域普遍使用的基础设施？这或将是学术界和产业界共同探讨的问题。

二、解决模型存在伦理环保问题和版权争议

（一）评估数据集和训练方法，降低模型偏见

由于数据集本身存在偏见，大模型可能学习到相应的偏见，并加以放大。可采用大规模自动化的方法进行数据清洗，降低偏见问题。此外，采用弱标注进行数据增强，或可加强模型对错误内容的抵抗能力，减少泄露隐私的问题。

（二）建设低碳绿色的智能算力集群

大模型带来的环保问题，主要在于模型训练和推理过程中消耗的大量能源和产生的碳足迹。建议在建设智能算力中心基础设施的过程中，采用低碳环保的建设方法，提升数据中心的用电效率，同时通过智能运维、创新设计等方式，提升模型的运算效率，服务器的通信和资源调度效率，降低机房冷却通风所需的额外能源。此外，还可借鉴谷歌、微软等建设数据中心的经验，如采用风能、太阳能等环保能源，实现低碳发展。

（三）针对大模型建立专门的开源许可证体系

目前计算机领域采用的开源许可证体系是针对代码的，对于开源的大模型和配套的训练数据集，这套许可证体系存在一定的缺陷，如不能确定微调模型，以及模型生成的内容的版权归属。建议未来开源社区根据AI模型的发展情况，提出针对数据集、模型和推理结果的专用开源许可证体系，鼓励企业和开发者使用开源模型，并能够保护原作者的权益。

报告贡献情况

黄铁军、文继荣、刘知远为本报告提供方向性的建议。

何晓冬、黄松芳、李笛为本报告相关章节提供核心观点和内容。

戴一鸣、张大魁、李静云主要负责资料收集、调研和整理工作，并撰写报告。

贾伟、闫亚琼、梁杨晓参与调研和资料收集工作，并联络其他编者。

袁莎、张冬敏在报告撰写过程中提供执行方面的建议和指导。

李梦佳、马瑞军、沈林杉、牛梦琳负责报告内容的整理、审核、编校等工作。

关于智源研究院

北京智源人工智能研究院（简称“智源研究院”）是落实“北京智源行动计划”的重要举措，是在科技部和北京市委市政府的指导和支持下，由北京市科委和海淀区政府于2018年11月推动成立的新型研发机构。智源研究院的愿景和目标是按照国家新一代人工智能发展规划总体部署，聚焦原始创新和核心技术，建立自由探索与目标导向相结合的科研体制。支持科学家勇闯人工智能科技前沿“无人区”，挑战最基础的问题和最关键的难题，推动人工智能理论、方法、工具、系统和应用取得变革性、颠覆性突破。营造全球最佳的学术和技术创新生态，推动北京成为全球人工智能学术思想、基础理论、顶尖人才、企业创新和发展政策的源头，率先成为国际领先的人工智能创新中心。推动人工智能产业发展和深度应用，改变人类社会生活，促进人类、环境和智能的可持续发展。

智源研究院积极打造智源社区平台，通过汇聚北京和全国科研、产业领域的人工智能创新人才，构建高度合作，紧密交流的科学研究社区，充分发挥成员的协同效应，推动人工智能的跨学科融合，在科学发现与技术研发循环的过程中，鼓励成员开展紧密协作。智源社区坚信，人工智能是深刻改变人类社会生活的颠覆式创新技术，需要依靠极具创造力的人才支持。我们希望围绕发展人工智能的共同愿景，吸引更多科研工作者、行业从业者和开源开发者加入，为我国人工智能的科研、产业发展贡献自身的力量！

智源研究院微信公众号

智源社区微信公众号

智源社区行业研究群

免责声明

本报告所含内容为一般性信息，不构成任何专业业务和财务业务的判断依据，同时本报告的信息来源于公开的资料，我们对该等信息的准确性、完整性或可靠性作尽可能的追求，但无法做任何保证和承诺，报告编者及所在单位不对任何因此报告导致的直接或间接损失或损害承担责任。

联系人：戴一鸣 | 智源社区分析师
邮　箱：ymdai@baai.ac.cn

超大规模智能模型
产业发展报告

2021年9月

北京智源人工智能研究院
BEIJING ACADEMY OF ARTIFICIAL INTELLIGENCE

联系我们

电话：(010) 6893 3383

邮箱：press@baai.ac.cn

官网：https://www.baai.ac.cn/

地址：北京市海淀区中关村东路1号院赛尔大厦5层

智源研究院公众号