·专题:ChatGPT 与人工智能技术应用·

# 语言大模型的演进与启示

陶建华[1*] 聂 帅[2] 车飞虎[1]

1. 清华大学 自动化系,北京 100084
2. 启元实验室,北京 100190

[摘 要] 2022 年 11 月,OpenAI 推出对话人工智能大模型 ChatGPT,展现了令人惊艳的自然语言理解和生成能力,并具备了跨学科、多场景、多用途的通用性,在很多任务上的性能达到了人类专家的水平,引起了产业界和学术界的广泛关注。以 ChatGPT 为代表的大模型技术实现了人工智能技术从“量变”到“质变”的跨越,有望发展成为人工智能关键基础设施赋能百业,加速推进国民经济的高质量发展。本文首先回顾了大模型技术的演进历程,从技术、应用、生态等多个角度阐述大模型技术引发的新一轮人工智能变革,并指出大模型技术可能带来的风险和挑战,最后给出了我国大模型发展的一些启示与展望。

[关键词] ChatGPT;大模型;预训练;指令微调

## 1 大模型技术正引发人工智能新一轮变革

2006 年 Geoffrey Hinton 教授提出通过逐层无监督预训练的方式来缓解由于梯度消失而导致的深层网络难以训练的问题[1],引起了学术界和工业界对深度神经网络的重新关注。此后,深度学习在计算机视觉[2]、语音[3]、自然语言处理[4]等众多领域取得突破式的研究进展,开启了新一轮深度学习发展的浪潮。总结过去十多年的发展,深度学习技术大致经过了三次重大的研究范式转变。从开始的“监督学习+各自为政”,到“预训练模型+任务微调”,再到如今的“预训练大模型+提示生成”,经历了从专用到通用,从小数据到大数据,从小模型到大模型的发展历程。

大模型通常是指通过预先在海量数据上进行大规模预训练,然后通过指令微调以适应一系列下游任务的通用人工智能模型,被看作是一项人工智能技术迈向通用智能的里程碑式进展。传统上,人工智能模型往往依赖大量有标签数据的监督训练,而且一个模型一般只能解决一个任务,适用于单一场景,这使得人工智能的研发和应用成本高,场景适应能力弱,难以规模化应用。近十多年来,人工智能模型的参数量正在迅速变大,仅仅 2021 至 2022 年间,模型参数量增加了 10 倍以上,以 Transformer[5] 预训练为基础的大模型,在海量无标签数据上进行预训练学习,降低了对标注数据的要求,不仅使模型的性能相较于以往人工智能方法带来了突破性提升,而且随着数据量增大和模型的进一步变大,模型性能还会不断增强,甚至出现量变到质变的能力涌现现象。

从预训练机制下的模型发展角度来看,十年前在进行大规模应用时就可以发现这样一个重要的趋势,即深度学习技术总的发展趋势是通过各种技术

**陶建华** 清华大学自动化系教授,中国科学院大学人工智能学院教授,中国计算机学会会士,国家杰出青年科学基金获得者。在国内外学术期刊和会议上发表论文 400 余篇,担任国际主要期刊 *IEEE Transactions on Affective Computing* 指导委员会委员,*Speech Communication* 责任编辑,*Interspeech 2020*、*Affective Computing Intelligent Interaction*、*IEEE International Workshop on Machine Learning for Signal Processing*、中文口语语言处理国际会议等语音领域重要国际会议程序委员会主席等。研究方向包括语音识别与合成、人机交互、情感计算、多媒体信息处理。

收稿日期:2023-08-03;修回日期:2023-09-26

* 通信作者,Email: jhtao@tsinghua.edu.cn

手段使得我们能够利用更多数据训练层次更深规模更大的深度神经网络(图 1)。在“大数据＋大算力＋强算法”的加持下,AI 大模型实现了“暴力美学”,通过“提示＋指令微调＋人类反馈”方式,可以实现一个模型完成多种不同的任务,展现了令世人惊艳的自然语言生成能力和通用性,具备了跨学科、多场景、多用途的处理能力,支持多轮对话、语言翻译、信息检索、程序设计、诗词创作、数据分析等一系列功能,在部分应用上甚至已经可以媲美人类专家。ChatGPT 自 2022 年底发布以来,便在产业界和学术界引起了广泛关注,仅在 5 天内注册用户就超过了 100 万,产品上线两个月即在全球吸引了超过 1 亿用户注册使用,成为全球历史上用户数增长最快的应用。而后,微软于 2023 年 2 月 8 日基于其推出新一代 AI 驱动搜索引擎 New Bing,带来了全新的搜索体验,甚至对谷歌搜索的传统优势地位形成重要威胁。ChatGPT 同时还具备一定的逻辑推理能力,成功通过了谷歌初级代码工程师面试;此外,其所创作的学术论文甚至可以通过学术专家的评审。ChatGPT 等大模型技术可以在经济分析、社会安全等众多领域发挥重要作用,其具备从超大规模数据中学习知识,进行逻辑推理和利用强化学习自我优化的能力,同时能够完成多种类型的任务。因此,大模型被广泛认为很可能像 PC 时代的操作系统一样发展成为未来人工智能应用中的关键基础设施,引发人工智能新一轮变革,加速推进国民经济的高质量发展。

## 2 大模型技术的基础与演进

ChatGPT 起源于语言模型,本质上是通过深度学习模型从海量的无标注文本数据中学习语言知识和世界知识。语言模型是对词或句子依赖关系建模的过程,大致经历了四个发展阶段(图 2)。(1) 以 N-Gram[6] 为代表的统计语言模型,通过统计的方式建模相邻若干个词之间的依赖关系,计算量小,容易训练,可解释性强,但建模能力有限,缺乏对长期依赖的建模,不具备语言理解和推理的能力,存在稀疏和泛化性差的问题。(2) 以 Word2Vec[7] 为代表的神经网络语言模型,通过神经网络学习语言模型来实现语言中词组的稠密化向量表示。Word2Vec 可以将每个词向量化,且能够通过欧式空间中距离度量词与词之间的相近关系。Word2Vec 是许多自然语言处理任务中必不可少的一步,解决了自然语言中词组的表示问题。但 Word2Vec 仍是浅层表示,依然无法解决语言中的长距离依赖问题。(3) 以 Recurrent Neural Network based Language Model (RNNLM)为代表的循环神经网络语言模型[8,9],通过循环神经网络来实现自然语言的时序关系建模。具备一定的上下文语义建模的能力,能够抓住部分的长期依赖关系,但对于语句间的长距离建模能力依然有限,且只能抓住“从左到右”或者“从右到左”单向单调关系。(4) 以 Transformer[5] 为代表的注意力语言模型,自然语言在语义层面存在无向图一样的复杂关系。Transformer 是一种基于自注意力机制的深度神经网络模型,其自注意力模块可以学习文本中任意两个词或语义单元的依赖关系,即在处理每个位置的信息时,模型会考虑文本中其他所有位置上的信息,这种机制使得 Transformer 模型能够有效地处理长距离和高层语义单元的复杂依赖关系。Transformer 类的语言模型在技术路径上主要朝两个方向发展,一个是 Google 所推出的 BERT 系列[10-12],以 Transformer 编码器为基础,通过“完形填空(掩蔽预测)”的方式从海量无标注数据中学习语义关系,侧重于语言理解,以“任务微调”的方式

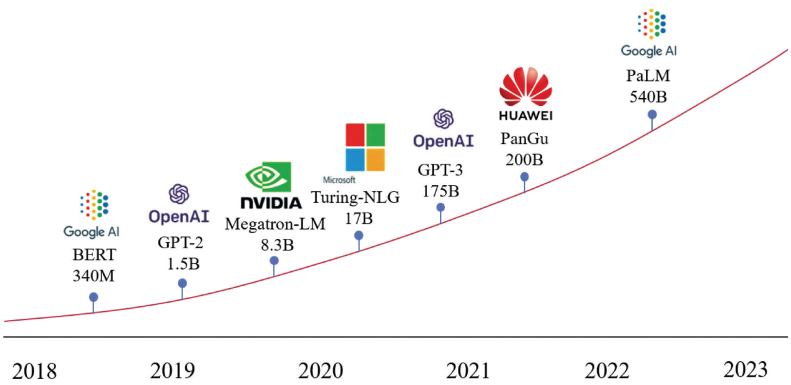


**图 1 深度学习模型的参数规模变化**

解决下游自然语言处理任务，这一类模型称之为 Encoder-Only 模型（图 3）。另一条路径是 OpenAI 所推出的 GPT 系列[13-15]，以 Transformer 解码器为基础，通过“预测未来”的方式从大量无标注数据中学习语言模型，侧重语言生成，以“指令微调”的方式激发其蕴含的知识解决下游自然语言处理任务，这一类模型为 Decoder-Only 模型（图 4）。研究表明，Decoder-Only 模型更容易适应下游任务，具有更强的通用能力。

此外，还有一类模型将同时利用 Transformer 的编码与解码能力，称为 Encoder-Decoder 模型，其中典型代表有 T5[16]（图 4）、GLM[17] 等模型。Encoder-Decoder 模型具有较强的序列学习和生成能力，特别是在实现输入序列到输出序列的结构映射方面，所以在机器翻译、文摘生成等任务上表现优异。

大模型带来了突破性的性能变化，但大模型并不是从 ChatGPT 才开始的，ChatGPT 是由 GPT 系列语言预训练模型演化而来的。GPT 模型使用多层 Transformer 的解码器作为模型，来预测下一个词的概率分布，通过在大规模无标注文本语料库上预训练得到。从 GPT-1 到 GPT-3 模型和数据规模都越来越大，能力也越来越强，再通过指令微调和人类反馈的强化学习方式使得 GPT 进化到更加符合人类要求的智能模型。

GPT-1[13] 主要采用 Transformer 的 decoder 作为模型，利用大量无标注数据进行无监督预训练，在不改变基座模型的情况下，仅通过对少量标注数据进行任务相关的输入变换，然后进行有监督微调就可以解决不同的下游任务，相对于之前每个任务需要重新训练一个模型的范式，GPT-1 模型具有重要的启发意义。到 GPT-2[14] 时，模型和参数规模变得更大，通过 Zero-shot Learning 的方式就解决不同的下游任务，虽然展现一定的通用性，但性能有限。到 GPT-3[15] 时，模型规模和数据规模达到千亿，它能够在没有任何梯度更新和微调情况下，仅通过提示词或少数样例可以非常好地完成指定的各种任务，甚至超过最好的专用模型。从这个角度来说，大量

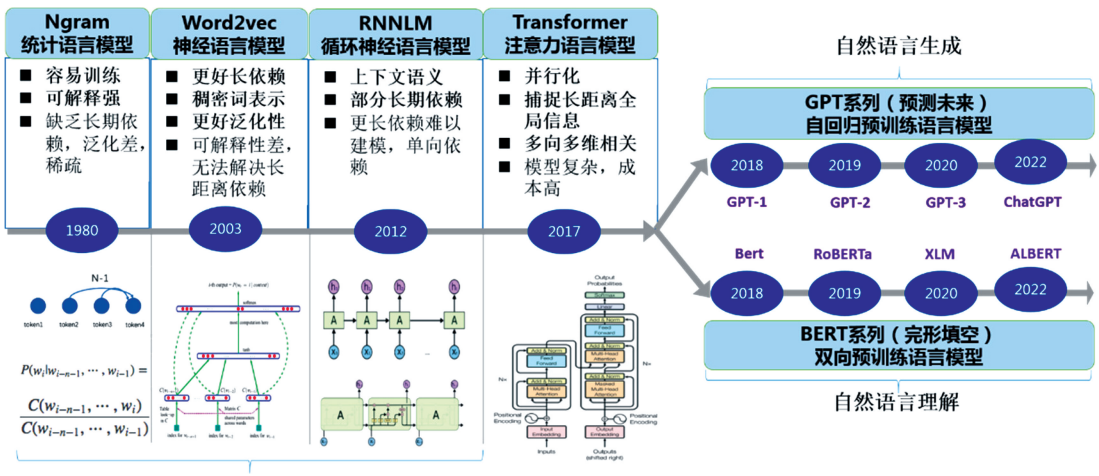


图 2　语言模型的演进路线

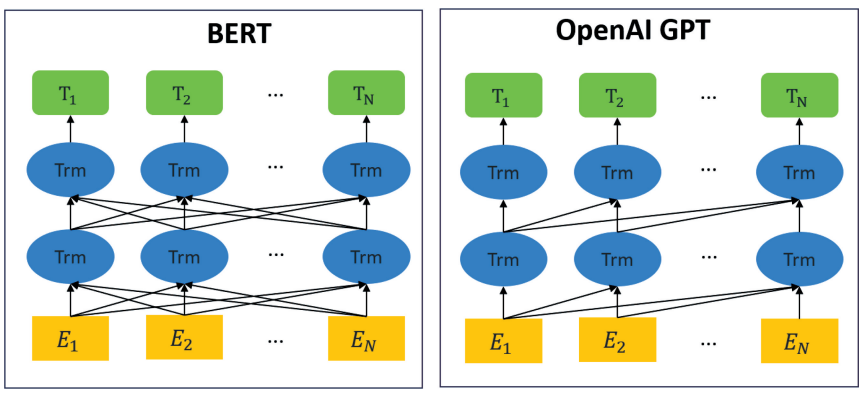


图 3　BERT 和 GPT 在网络结构上的区别[10]

信息提取能进一步提炼出很多超脱语义之上的信息，包括语义片段的信息，以及一些各个语义层次之间的关联信息。但其终究是个基于概率生成的语言模型，不可避免地会输出无用的、有害的信息，无法对齐人类的偏好。

为了克服 GPT 生成的内容无法与人类偏好对齐的问题，OpenAI 进一步提出 InstructGPT[18]。InstructGPT 采用有监督的指令微调[19]和人类反馈的强化学习（Reinforcement Learning from Human Feedback，RLHF）方法[20, 21]，使用近端策略优化（Proximal Policy Optimization，PPO）[22]强化学习模型实现模型的自我优化和更新，使得模型能够更好地遵循用户的意图，生成的内容和人类的要求和偏好进行有效对齐。实际上，ChatGPT 是在 GPT-3.5 基础上采用 InstructGPT 训练方式进一步微调出来的，而 GPT-3.5 在 GPT-3 的基础上加入了思维链[23]、代码[24]和多轮对话等数据进一步训练得到的，代码和思维链的训练使得模型具有更强的逻辑推理能力。这一系列的优化改进成就了 ChatGPT 惊艳的思维理解、多轮对话和通用能力[24]。

2023 年 3 月 OpenAI 进一步发布多模态 AI 大模型 GPT-4。相对于 ChatGPT，GPT-4[25]在事实性、可控性和拒绝越过防护栏方面获得了有史以来最好的结果，显著减轻了大模型的幻觉问题，且具有更长的上下文建模能力，允许用户定制模型的风格和行为，在很多任务上的表现甚至超过人类水平。此外，GPT-4 还具备强大的图像理解能力，能够直接以自然语言的方式对图像进行视觉问答。

未来大模型在技术方面的演进一方面会朝着更高智能水平更加通用性方向发展，另一方面各种专业领域的大模型也将蓬勃发展，实现“大模型”与“小模型”协同并进的发展格局，同时大模型也将融入更多模态信息，实现多种模态的统一表示，大模型技术也将和知识图谱、搜索引擎、博弈对抗、脑认知等技术融合发展，相互促进。在产业应用方面，大模型正在变革社会生活和生产方式。ChatGPT 除了能作为人类助手之外，还将重塑生产力工具，变革信息获取方式。微软将 GPT-4 整合到整个 office 套件里，打通各个办公软件，实现协同办公。同时，将 GPT-4 整合进 Bing 搜索引擎中，以“GPT-4＋搜索”的方式实现了更加直接和智能的信息获取方式，解决了大模型知识陈旧难以更新的问题，显著提升了获取信息的准确性和可靠性，同时大模型还将与外部接口和服务融合，深入到各行各业。

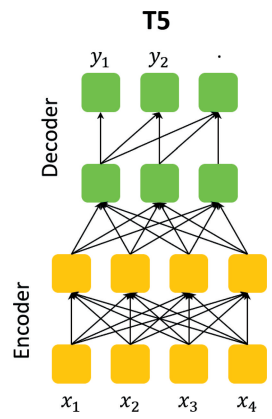


图 4 T5 模型网络结构[16]

## 3 大模型的相关技术

大模型之所以能在很多领域取得卓越表现依赖于一系列关键技术的突破和创新，其中比较重要的技术有模型预训练、适配微调、模型高效计算、推理加速等方面。

（1）模型预训练。高效的模型预训练技术为大模型奠定了坚实基础，主要包括：高效预训练策略、高质量预训练数据与高效模型架构。高效预训练策略是以低成本实现对语言大模型的预训练，目前主流的方法有设计高效的优化任务目标、热启动策略、渐进式训练策略与知识继承等方法。高质量预训练数据是指在构建预训练数据时，需通过质量过滤、冗余去除、隐私消除等方法对数据进行筛选。高效模型架构涉及到统一的序列建模[16]，即将多种自然语言处理任务统一到一个框架，提升模型的性能与泛化性，还设计提升 Transformer 模型架构的编解码效率、训练稳定性、显存利用率等方面。

（2）适配微调。由于语言大模型在通用领域中缺乏对特定任务的知识，因此需要引入适配微调帮助模型更好地适应特定需求，适配微调包括指令微调和参数高效微调[18]。指令微调让语言大模型在给定指令提示下给出特定回应，可帮助语言大模型获得人类语言指令遵循能力，包括指令理解、指令数据获取与指令对齐等内容。参数高效微调[26]包括：添加式方法，即在原模型的基础上仅微调引入的额外参数；指定式方法，即指定元模型的部分参数为可训练参数；重参数化方法，将模型参数重参数化到地位参数空间中，仅优化低维空间中的近似参数。适配微调不仅是连接通用预训练和特定下游任务的重要桥梁，还可以提升模型在特定领域任务上的表现，保护数据隐私、提高部署效率。

（3）模型高效计算。大模型的高效计算通过对模型计算、显存、内存与通信等方面的优化，提升训练吞吐量，实现在有限资源下最大化模型高效计算的目的。优化可从两个方面实现：设备内优化，如半精度浮点优化与混合精度浮点优化等方法[26]；多设备优化方法，即多种并行策略[27-30]，其中典型代表有：数据并行通过分配不同数据分区实现对大批量样本的并行处理，张量并行将庞大的模型张量分解为子运算后在不同设备上并行，流水线并行将模型不同层交给不同设备流水作业式执行。这些基本策略可根据实际需要进行灵活组合，形成混合并行模式。

（4）模型推理加速。大模型在推理阶段往往面临运算量巨大、显存占用过高等问题，因此可通过模型压缩提升模型推理速度[31]，主要包括模型稀疏化、模型蒸馏、模型参数共享与量化技术，量化技术通过减少模型中的数据精度来降低存储和计算需求。此外，还可采用多种压缩策略进行组合，如通过组合模型稀疏化与低比特量化等方法对大模型冗余信息进行将建，提升模型压缩率。

## 4　大模型技术的生态之争

随着大模型技术的发展，目前涌现出许多开源的通用大模型与领域大模型，这些开源模型为研究者和开发者提供了可靠的基础，降低了技术门槛。

LLaMA[32]是一个开源的大规模语言模型集合，其参数量范围从 70 亿到 650 亿不等。LLaMA 模型在数万亿量级的文本数据上进行预训练，展示了仅利用公开可获取的数据就可以训练出超越专有数据集的先进模型的可能性。LLaMA 模型采用类 GPT 的 Decoder-Only 架构，并进行了一些改进，如层规范化、SwiGLU 激活函数以及每层位置编码等。BLOOM[33]覆盖 46 种自然语言和 13 种编程语言，通过渐进式语料选择策略进行预训练，参数量高达 1 760 亿。相较原始 Transformer，BLOOM 进行了诸多架构创新，如采用基于距离的注意力机制等。此外，BLOOM 使用了精心设计的微调策略，使得模型可以更好地适配下游任务。Baichuan 由百川智能研发，是一款大规模中文英文预训练语言模型，其开源模型参数量有 70 亿、130 亿两种规格，训练数据达到 1.4 万亿词元量级，在多个中文英文基准测试上都获得领先水平。Baichuan 采用了类 LLaMA 的模型设计思路，如旋转嵌入、SwiGLU 激活函数等，以实现更好的语言建模效果。CPM-Bee[34]采用了 Transformer 的自回归结构，其参数量达到百亿级，在超过万亿词的高质量语料上进行预训练，展现出较强的中文和英文语言理解生成能力。文心一言采用有监督微调、人类反馈强化学习等技术，具备知识增强与检索增强等能力，还可以调用外部工具与服务。讯飞星火认知大模型具备多种自然语言处理、代码与数学能力。GLM[17]将自回归填空作为预训练任务，提升了大模型在长文本生成与序列任务处理的能力。

当前大模型在金融、医疗和教育等领域已初步取得成效。在金融领域，度小满开源的千亿参数中文金融模型“轩辕”，通过在海量金融文本上进行预训练，获得了强大的金融术语理解和文本分析能力。马上金融推出的首个零售金融大模型“天镜”，可支持数字化银行的诸多智能化应用。在医疗方面，医联的 MedGPT 致力于构建全流程的智能化诊疗系统，形成了独特的医疗人工智能建模方法论。谷歌 DeepMind 研发的 Med-PaLM[35]在临床知识问题上达到专家水平。京东健康基于全流程医疗需求构建了“京医千询”医疗大模型和开放平台。医疗健康大模型正朝着覆盖全诊疗过程、多模态输入、人机交互的方向演进。在教育方面，网易有道推出的“子曰”教育大模型，是国内首个教育专用预训练语言模型，其强大的个性化建模和知识整合能力，将为在线教育提供有力帮助。华为云盘古大模型使用了全球 39 年的天气数据进行预训练，在保持训练精准度的同时大大提升了速度[36]。商汤遥感大模型在通用视觉大模型的基础上，在解译精度与时间上都实现了突破。

此外，在其他领域大模型技术具有广泛的应用场景，“大模型＋传媒”可以实现智能新闻写作，提升新闻的时效性；“大模型＋影视”可以拓宽创作素材，开拓创作思路，激发创作灵感，提升作品质量；“大模型＋营销”可以打造虚拟客服，助力产品营销；“大模型＋娱乐”可以加强人机互动，激发用户参与热情，增加互动的趣味性和娱乐性；“大模型＋军事”可以增强军事情报和决策能力，可以实现实时战场翻译，快速准确的威胁评估、作战任务规划和执行、战场感知、战术决策支持、改进态势感知等。总之，大模型的发展将给人类带来了非常强大的助推力，让数字世界和现实世界的共生变得更为便捷、更为有效。

大模型的通用性使其被认为是可以成为未来人工智能应用中的关键基础设施，就像 PC 时代的操作系统一样，赋能百业，加速推进国民经济的高质量

发展(图 5)。向上,大模型可带动上游软硬件计算平台的革新,形成高性能软硬件与大模型的协同发展,构建“大模型+软硬件+数据资源”上游发展生态;向下,大模型可以打造“大模型+应用场景”的下游应用生态,加速全产业的智能升级,对经济、社会和安全等领域的智能化升级中形成关键支撑。

大模型的生态之争是未来企业和国际间信息产业竞争的关键,在生态竞争中的胜出者,将形成强大的壁垒优势。大模型技术突破将带来产业竞争力提升、产业收入增加、用户数据积累,而这些又会反哺大模型技术研究,进入良性循环,并在“马太效应”的影响下占据数字化、智能化时代的主导权。目前,谷歌已将大模型技术融入其搜索引擎、操作系统、浏览器等核心产品中;微软也将大模型技术应用到其操作系统、办公软件和云平台等重要产品,试图抢先构建以大模型为核心的生态。我国在大模型的生态竞争上正处于奋力追赶阶段,然而作为人工智能产业大国,在数据、市场、应用场景等方面具有较大优势,抓紧布局大模型的应用生态,依然会使我国有机会抢占大模型的生态主导权。

## 5 大模型技术的风险与挑战

尽管以 ChatGPT 为代表的大模型技术取得了令人瞩目的成功,但短期来看,大模型技术仍面临着诸多挑战与风险。

首先,大模型的可信性无法得到保障。基于海量数据训练的大模型,其生成的内容非常符合语言规则,通顺流畅,人类几乎难以辨别,且生成的内容与人类的偏好对齐,极具欺骗性,但在事实性、时效性和数据准确性方面存在很多问题,不具备对其生成的内容提供可信性评估的能力[37, 38]。

其次,大模型的可解释性较差。大模型本质上是一个基于深度神经网络的黑盒语言模型,其能力来源的机理依然不清楚,难以解释。包括大模型的涌现能力[39]、规模定律[40]、知识表示、逻辑推理能力、泛化能力、情景学习能力[41, 42]等方面仍有待学术界展开进一步研究,以便为大模型的大规模实际应用提供理论保障。

再次,大模型的应用成本较高。大模型参数规模和数据规模都非常巨大,导致其训练和推理计算量大、功耗高、部署困难、应用成本高、还存在延迟问题,大大限制了其应用。提高推理速度降低大模型的使用成本是大规模应用的关键。

最后,小数据环境下 AI 大模型的能力迁移问题。更大模型更大数据会使得模型涌现更强的能力,但针对特定领域,在小数据环境下实现 AI 大模型的能力迁移,能够显著扩大 AI 大模型的应用范围降低应用成本。大模型在更为复杂场景下的鲁棒性和泛化能力方面也值得探索,大模型并不能适用所有场景,它本质上还是依赖训练数据所能覆盖的场景。在场景规模数据不大的情况下,不得不去依赖对它进行微调。但是,通过不同细分领域划分成不同类别进行数据有效的筛选、标注,以及相应的微调技术就能使得大模型具有较好的在不同小场景、场景较复杂情况下适用的能力,提升它的可靠性。

大模型除了存在诸多技术挑战外,还存在一系列的技术风险。大模型具有强大的自然语言理解和生成能力,与语音合成、图像视频生成等技术结合可以产生人类难以辨别的音视频内容,有可能会成为制造虚假信息、恶意引导行为、舆论攻击、甚至危害国家安全的工具[43, 44]。此外,作为深度学习模型,大模型存在一定的安全风险,目前针对大模型的安全漏洞攻击技术包括:数据投毒攻击、对抗样本攻击、模型窃取攻击、后门攻击等。大模型的安全漏洞被攻击者利用有可能会使和大模型关联的所有业务面临整体失效的风险,潜在威胁着以大模型为基础

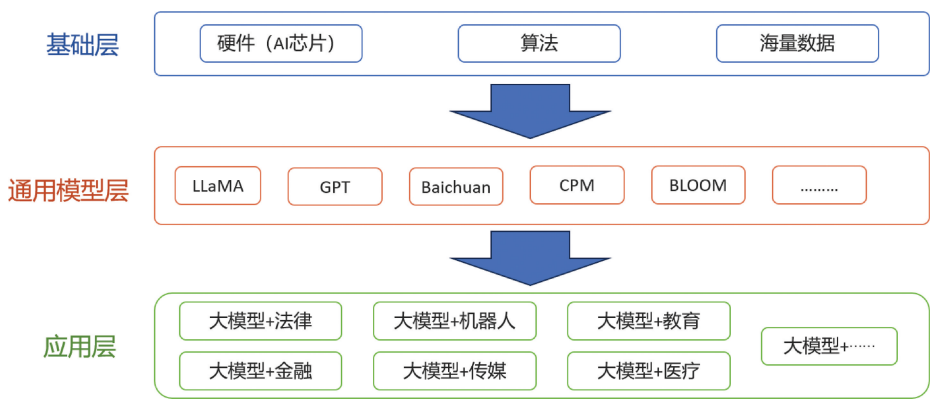


图 5 大模型的生态架构

构建的应用生态。因此，在发展大模型的同时，需要着重加强大模型的安全防范能力，以保障整个大模型的应用生态的安全。最后，大模型利用海量的互联网数据进行训练，大量个人、企业甚至是国家的隐私数据被编码进大模型的参数中，因此大模型还存在数据隐私问题，研究表明通过一定的提示甚至可以从大模型中窃取个人隐私数据，这也是我们应该关注的大模型技术的风险之一。

## 6　大模型技术发展的启示

从 2018 年起，全球众多机构围绕大模型的研制开展了激烈竞争，“大炼大模型”成为当前人工智能领域的科技前沿。然而，在激烈的竞争中，为何是 OpenAI 的 ChatGPT 大模型脱颖而出？OpenAI 所处的良好创新生态，及其在技术、数据和计算硬件上的优势，共同催生了这一成果。良好的创新生态和对潜在颠覆性技术的长期、大力和稳定投入是基础。优良的创新文化吸引了一批高水平的人工智能专家专注于大模型技术的研发，使得 OpenAI 在多项基础理论与关键技术上，成为布局最早、研究最深的机构，最终形成了突破。开源生态为 OpenAI 大模型的发展提供了强有力的多方支撑。OpenAI 在发展早期，对所有的大模型进行了开源，通过众多高水平志愿者共同对该公司提供了源源不断的技术支撑。对大规模数据的长期优化和建设是其性能形成领先的保障。OpenAI 在研发过程中，使用了超过 45 T 的海量数据，包涵书籍、杂志、百科、论坛、代码等众多领域和类型，并对其进行了精心地筛选、去重、清洗、整理，丰富多样的数据是其通用性的基础。同时 OpenAI 还投入了大量人力物力和资金，收集和标注了高质量的指令微调数据和人类反馈数据，其领先的性能与其在数据上长期坚持投入有很大关系。强大的计算硬件是保证，ChatGPT 的研发，需要数以万计的高性能计算芯片，而当前美国垄断了高性能计算芯片，受美国芯片禁运限制的国家和机构开展类似研究将面临巨大挑战。为了抢占 ChatGPT 后新的技术制高点，创新生态营造、基础技术投入、数据平台建设、计算硬件变革是发力重点。

第一，抓紧推动大模型技术研发的同时，鼓励交叉原始创新，强化数据基础优势。集中国家资源抓紧投入大模型技术的研发，缩短与美国的差距，力争实现反超，是一件不可回避的迫切任务。但也要充分认识到 OpenAI 等公司在创新机制上给我们带来的重要启示，通过政策、评估和体制上，建设具有中国特色的高效原始创新模式与机制。同时，在认识到大模型带来重要机遇的同时，也要充分认识到大模型依然存在一系列关键技术挑战，着力推动人工智能与脑科学、认知科学的交叉创新研究，力争从人工智能的可解释性、可信性、高可靠性和低功耗性等方面形成重要突破。此外，还需要充分发挥我国数字基础设施建设好、数字应用数据多的优势，统筹建设全国一体化大数据中心，加大投入数据治理与优化，为大模型的进一步突破奠定基础。

第二，构建技术生态体系，发展应用生态平台。AI 大模型是实现通用人工智能的有效途径，其成功是建立在深度学习、强化学习和分布式训练等技术基础之上的。在技术生态体系建设上，建议制定 AI 大模型发展纲要，在包括算法模型、芯片和分布式训练架构方面启动国家重点研发计划，在 AI 大模型核心环节和相关技术上进行知识产权布局。在应用生态上，建议组建包括由芯片、云计算、互联网、数据、应用等上下游企业组成的产业发展联盟，鼓励相关企业基于 AI 大模型进行数字化转型升级，支持产学研三方协同的 AI 大模型研发模式。

第三，突破计算硬件瓶颈，抢占全新赛道先机。大模型对大规模算力的依赖，是我国面临迫切的“卡脖子”问题。除了我们在传统算力方面要加强投入、研发之外，也需要在一些新的赛道上，比如光电计算等方面，加强相应部署，争取能够在算力上不被“卡脖子”。我国应在人工智能芯片等方面，通过政府引导，扶持关键企业，勇于攻坚克难。另一方面，在光电计算与量子芯片等“新赛道”加强投入，颠覆现有技术路线和产业生态，出奇抢占先机，重塑产业格局。

第四，加强人工智能安全技术和伦理治理机制建设。大模型和其他 AI 技术均存在大量的漏洞和被攻击风险，迫切需要加大力度进行 AI 安全检测与防御技术的研发与部署，包括加强针对大模型的数据隐私窃取和保护的技术研发和制度建设等。加强大模型生成内容的技术审核与规范构建，构建人工智能生成内容的知识产权保护机制。强化科技伦理教育，建构用户使用规范，特别是针对未成年人的使用进行规范教育。

## 参　考　文　献

[1] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science, 2006, 313(5786): 504—507.

[2] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84—90.

[3] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Processing Magazine, 2012, 29(6): 82—97.

[4] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality// Proceedings of the 26th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2013: 3111—3119.

[5] Vaswani A, Shazeer N, Parmar N, et al. Attention is all You need// Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2017: 6000—6010.

[6] Jelinek F. Continuous speech recognition by statistical methods. Proceedings of the IEEE, 1976, 64(4): 532—556.

[7] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model. Journal of Machine Learning Research, 2003, 3: 1137—1155.

[8] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model// Proceedings of the 11th Interspeech. Chiba: International Speech Communication Association, 2010: 1045—1048.

[9] Sundermeyer M, Schlüter R, Ney H. LSTM neural networks for language modeling// Proceedings of the 13th Interspeech. Portland: International Speech Communication Association, 2012: 194—197.

[10] Devlin J, Chang MW, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. (2018-10-11)/[2023-08-02]. https://arxiv.org/pdf/1810.04805.pdf.

[11] Yang ZL, Dai ZH, Yang YM, et al. XLNet: generalized autoregressive pretraining for language understanding. (2019-06-19)/[2023-08-02]. https://arxiv.org/pdf/1906.08237.pdf.

[12] Liu YH, Ott M, Goyal N, et al. RoBERTa: a robustly optimized BERT pretraining approach. (2019-07-26)/[2023-08-02]. https://arxiv.org/pdf/1907.11692.pdf.

[13] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training. (2018-06-11)/[2023-08-02]. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[14] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners. OpenAI blog, 2019, 1(8): 9.

[15] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. Advances in Neural Information Processing Systems. 2020: 1877—1901.

[16] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text Transformer. Journal of Machine Learning Research, 2020, 21: 5485—5551.

[17] Du ZX, Qian YJ, Liu X, et al. GLM: general language model pretraining with autoregressive blank infilling. (2021-03-18)/[2023-08-02]. https://arxiv.org/pdf/2103.10360.pdf.

[18] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems. 2022: 27730—27744.

[19] Wei J, Bosma M, Zhao VY, et al. Finetuned language models are zero-shot learners. (2022-02-08)/[2023-08-02]. https://arxiv.org/pdf/2109.01652.pdf.

[20] Knox WB, Stone P. Augmenting reinforcement learning with human feedback// Proceedings of the 28th International Conference on Machine Learning. Washington: International Machine Learning Society. 2011: 3.

[21] Knox WB, Stone P. Interactively shaping agents via human reinforcement: the TAMER framework// Proceedings of the fifth International Conference on Knowledge Capture. New York: Association for Computing Machinery, 2009: 9—16.

[22] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms. (2017-07-20)/[2023-08-02]. https://arxiv.org/pdf/1707.06347.pdf.

[23] Wei J, Wang XZ, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. (2022-01-28)/[2023-08-02]. https://arxiv.org/pdf/2201.11903.pdf.

[24] Chen M, Tworek J, Jun H, et al. Evaluating large language models trained on code. (2021-07-07)/[2023-08-02]. https://arxiv.org/pdf/2107.03374.pdf.

[25] OpenAI. Gpt-4 technical report. (2023-03-15)/[2023-08-02]. https://arxiv.org/pdf/2303.08774.pdf.

[26] Ding N, Qin YJ, Yang G, et al. Delta tuning: a comprehensive study of parameter efficient methods for pre-trained language models. (2022-03-14)/[2023-08-02]. https://arxiv.org/pdf/2203.06904.pdf.

[27] Saharia C, Chan W, Saxena S, et al. Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems, 2022, 35: 36479—36494.

[28] Darema F, George DA, Norton VA, et al. A single-program-multiple-data computational model for EPEX/FORTRAN. Parallel Computing, 1988, 7(1): 11—24.

[29] Shoeybi M, Patwary M, Puri R, et al. Megatron-LM: training multi-billion parameter language models using model parallelism. (2019-09-17)/[2023-08-02]. https://arxiv.org/pdf/1909.08053.pdf.

[30] Huang YP, Cheng YL, Bapna A, et al. GPipe: efficient training of giant neural networks using pipeline parallelism. (2018-11-16)/[2023-08-02]. https://arxiv.org/pdf/1811.06965.pdf.

[31] Liang TL, Glossner J, Wang L, et al. Pruning and quantization for deep neural network acceleration: a survey. Neurocomputing, 2021, 461: 370—403.
[32] Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient foundation language models. (2023-02-27)/[2023-08-02]. https://arxiv.org/pdf/2302.13971.pdf.
[33] Scao TL, Fan A, Akiki C, et al. BLOOM: a 176B-parameter open-access multilingual language model. (2022-11-09)/[2023-08-02]. https://arxiv.org/pdf/2211.05100.pdf.
[34] Zhang ZY, Han X, Zhou H, et al. CPM: a large-scale generative Chinese pre-trained language model. AI Open, 2021, 2: 93—99.
[35] Singhal K, Tu T, Gottweis J, et al. Towards expert-level medical question answering with large language models. (2023-05-16)/[2023-08-02]. https://arxiv.org/pdf/2305.09617.pdf.
[36] Bi KF, Xie LX, Zhang HH, et al. Accurate medium-range global weather forecasting with 3D neural networks. Nature, 2023, 619(7970): 533—538.
[37] 车万翔，窦志成，冯岩松，等. 大模型时代的自然语言处理：挑战、机遇与发展. 中国科学：信息科学，2023，53(9)：1645—1687.
[38] Liu P, Yuan W, Fu J, et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 2023, 55(9): 1—35.
[39] Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models. (2022-06-15)/[2023-08-02]. https://arxiv.org/pdf/2206.07682.pdf.
[40] Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models. (2020-01-23)/[2023-08-02]. https://arxiv.org/pdf/2001.08361.pdf.
[41] Dai DM, Sun YT, Dong L, et al. Why can GPT learn In-context? language models implicitly perform gradient descent as meta-optimizers. (2022-12-20)/[2023-08-02]. https://arxiv.org/pdf/2212.10559.pdf.
[42] Akyürek E, Schuurmans D, Andreas J, et al. What learning algorithm is in-context learning? Investigations with linear models. (2022-11-28)/[2023-08-02]. https://arxiv.org/pdf/2211.15661.pdf.
[43] 陶建华，傅睿博，易江燕，等. 语音伪造与鉴伪的发展与挑战. 信息安全学报，2020，5(2)：28—38.
[44] 陶建华. 加强深度合成算法安全科研攻关 推进深度合成服务综合治理. (2023-01-11)/[2023-08-02]. https://mp.weixin.qq.com/s/3tE3mxkodLKX70ZvTezxhg.

## The Evolution and Inspiration of Large Language Model Technology

Jianhua Tao[1*]　Shuai Nie[2]　Feihu Che[1]
1. *Department of Automation, Tsinghua University, Beijing* 100084
2. *Qiyuan Laboratory, Beijing* 100190

**Abstract**　In November 2022, OpenAI launched ChatGPT, a large model of conversational AI, which has demonstrated amazing natural language understanding and generation capabilities, and has cross-disciplinary, multi-scene and multi-purpose versatility, with performance in many tasks reaching the level of human experts, attracting widespread attention from industry and academia. The large model technology represented by ChatGPT has realized the leap from "quantitative change" to "qualitative change" in AI technology, and is expected to develop into a key infrastructure of AI to empower all industries and accelerate the high-quality development of national economy. This paper firstly reviews the evolution of large model technology, explains the new round of AI changes caused by large model technology from the perspectives of technology, application and ecology, and points out the possible risks and challenges brought by large model technology, and finally gives some insights and prospects for the development of big model in China.

**Keywords**　ChatGPT; large models; pre-training; instruction fine-tuning

（责任编辑　崔国增　姜钧译）

* Corresponding Author, Email: jhtao@tsinghua.edu.cn