# 大模型时代

龙志勇
黄 雯
著

THE ERA OF LLM

## ChatGPT 开启通用人工智能浪潮

著名经济学家 朱嘉明 重磅作序

朱 民
清华大学国家金融研究院院长

喻国明
认知神经传播学开创者之一

李一诺
爱诗科技联合创始人兼首席执行官

谢 源
阿里云达摩院科学家

袁进辉
大模型框架 OneFlow 创始人

王维嘉
畅销书《模型思维》作者

联 袂 推 荐

中信出版集团
中信出版社


大模型时代
龙志勇
黄 雯
著
THE ERA OF LLM
ChatGPT 开启通用人工智能浪潮
著名经济学家 朱嘉明 重磅作序
朱 民
清华大学国家金融研究院院长
榆国明
认知神经传播学开创者之一
李一诺
麦肯锡公司前全球合伙人
谢 源
阿里巴巴达摩院科学家
袁进辉
大模型框架 OneFlow 创始人
王维嘉
畅销书《简知识》作者
联 袂 推 荐

# 版权信息

书名：大模型时代：ChatGPT开启通用人工智能浪潮

作者：龙志勇 黄雯

出版社：中译出版社

出版时间：2023-05-01

ISBN：9787500173953

品牌方：中译出版社有限公司

---

# 代序

## AI大模型：当代历史的标志性事件及其意义

“尝试找到如何让机器使用语言、形成抽象和概念、解决现在人类还不能解决的问题、提升自己，等等。对于当下的人工智能来说，首要问题是让机器像人类一样能够表现出智能。”

——达特茅斯会议对人工智能(AI)的定义

2020—2022年，在新冠疫情肆虐全球的阴霾日子里，人工智能创新的步伐完全没有停止。美国人工智能研究公司OpenAI异军突起：2020年4月发布神经网络Jukebox；2020年5月发布GPT-3，模型参数量为1750亿；2020年6月开放人工智能应用程序接口；2021年1月发布连接文本和图像神经网络CLIP；㊟【a CLIP(Contrastive Language-Image Pre-Training)模型是OpenAI在2021年初发布的用于匹配图像和文本的预训练神经网络模型，可以说是近年来在多模态研究领域的经典之作。该模型直接使用大量的互联网数据进行预训练，在很多任务表现上达到了目前最高水平。】 2021年1月发布从文本创建图像神经网络DALL-E；㊟【DALL-E是一个可以根据书面文字生成图像的人工智能系统，该名称来源于著名画家达利(Dalí)和机器人总动员(Wall-E)。】 2022年11月，正式推出对话交互式的ChatGPT。相比GPT-3，ChatGPT引入了基于人类反馈的强化学习(RLHF)㊟【单纯的强化学习(RL)是机器学习的范式和方法论之一，用于描述和解决智能体(agent)在与环境的交互过程中通过学习策略以达成回报最大化或实现特定目标的问题。】技术以及奖励机制。

ChatGPT是人类科技史上的里程碑事件，在短短几个月席卷全球，速度之快超出人类最狂野的想象。ChatGPT证明了通过一个具有高水平结构复杂性和大量参数的大模型（foundation model，又称为“基础模型”）可以实现深度学习。此后，大模型概念受到前所未有的关注和

讨论。但是，关于“大模型”定义，各方对其内涵的理解和诠释却莫衷一是，“横看成岭侧成峰，远近高低各不同”。

尽管如此，这并不妨碍人们形成关于大模型的基本共识：大模型是大语言模型(LLM)，也是多模态模型，或者是生成式预训练转换模型。GPT是大模型的一种形态，引发了人工智能生成内容(AIGC)技术的质变。大模型是人工智能赖以生存和发展的基础。现在，与其说人类开始进入人工智能时代，不如说人类进入的是大模型时代。我们不仅目睹，也身在其中，体验生成式大模型如何开始生成一个全新时代。

1.何谓大模型

人工智能的模型，与通常的模型一样，是以数学和统计学为算法基础的，可以用来描述一个系统或者一个数据集。在机器学习中，模型是核心概念。模型通常是一个函数或者一组函数，可以是线性函数、非线性函数、决策树、神经网络等各种形式。模型的本质就是对这个函数映射的描述和抽象，通过对模型进行训练和优化，可以得到更加准确和有效的函数映射。建立模型的目的是希望从数据中找出一些规律和模式，并用这些规律和模式预测未来的结果。模型的复杂度可以理解为模型所包含的参数数量和复杂度，复杂度越高，模型越容易过拟合。

人工智能大模型的“大”，是指模型参数至少达到1亿。但是这个标准一直在提高，目前很可能已经有了万亿参数以上的模型。GPT-3的参数规模就已经达到了1750亿。

除了大模型之外，还有所谓的“超大模型”。超大模型，是比大模型更大、更复杂的人工神经网络模型，通常拥有数万亿到数十万亿个参数。一个模型的参数数量越多，通常意味着该模型可以处理更复杂、更丰富的信息，具备更高的准确性和表现力。超大模型通常被用于解决更为复杂的任务，如自然语言处理(NLP)中的问答和机器翻译、计算机视觉中的目标检测和图像生成等。这些任务需要处理极其复杂的输入数据和高维度的特征，而超大模型可以从这些数据中提取出更深层次的特征和模式，提高模型的准确性和性能。因此，超大模型的训练和调整需要极其巨大的计算资源和数据量级、更加复杂的算法和技术、大规模的投

入和协作。

大模型和超大模型的主要区别在于模型参数数量的大小、计算资源的需求和性能表现。随着大模型参数规模的膨胀，大模型和超大模型的界限正在消失。现在包括GPT-4在内的代表性大模型，其实就是原本的超大模型。或者说，原本的超大模型，就是现在的大模型。

大模型可以定义为大语言模型，具有大规模参数和复杂网络结构的语言模型。与传统语言模型（如生成性模型、分析性模型、辨识性模型）不同，大语言模型通过在大规模语料库上进行训练来学习语言的统计规律，在训练时通常通过大量的文本数据进行自监督学习，从而能够自动学习语法、句法、语义等多层次的语言规律。㊟【生成性模型从一个形式语言系统出发，生成语言的某一集合。代表是乔姆斯基(Avram Noam Chomsky, 1928—)的形式语言理论和转换语法。分析性模型从语言的某一集合开始，根据对这个集合中各个元素的性质的分析，阐明这些元素之间的关系，并在此基础上用演绎的方法建立语言的规则系统。代表是苏联数学家O.S.库拉金娜(O. S. Kulagina, ?—?)和罗马尼亚数学家S.马尔库斯(Solomon Marcus, 1925—2016)用集合论方法提出的语言模型。在生成性模型和分析性模型的基础上，将二者结合起来，产生了一种很有实用价值的模型，即辨识性模型。辨识性模型可以从语言元素的某一集合及规则系统出发，通过有限步骤的运算，确定语言中合格的句子。代表是Y.巴尔-希列尔(Yehoshua Bar-Hillel, 1915—1975)用数理逻辑方法提出的句法类型演算模型。】

如果从人工智能的生成角度定义大模型，与传统的机器学习算法不同，生成模型可以根据文本提示生成代码，还可以解释代码，甚至在某些情况下调试代码。这一过程，不仅可以实现文本、图像、音频、视频的生成，构建多模态，还可以在更为广泛的领域生成新的设计，生成新的知识和思想，甚至实现广义的艺术和科学的再创造。

近几年，比较有影响的AI大模型主要来自谷歌、Meta和OpenAI。除了OpenAI的GPT之外，2017年和2018年，谷歌发布LaMDA、BERT和PaLM-E。㊟【谷歌推出的LaMDA(Language Model for Dialogue Applications)是语言处理领域的一项新的研究突破。

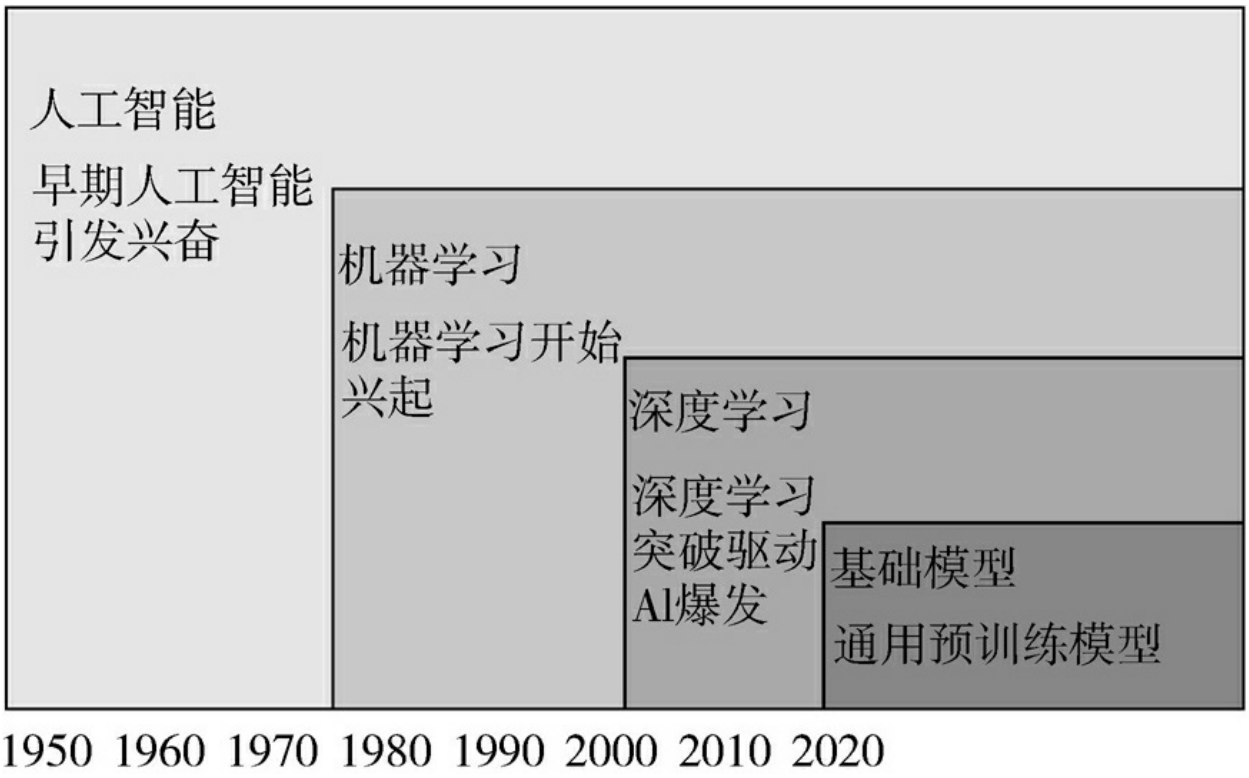
LaMDA是一个面向对话的神经网络架构，可以就无休止的主题进行自由流动的对话。它的开发是为了克服传统聊天机器人的局限性，传统聊天机器人在对话中往往遵循狭窄的、预定义的路径。
BERT(Bidirectional Encoder Representation from Transformers)是一个预训练的语言表征模型。它强调了不再像以往一样采用传统的单向语言模型或者把两个单向语言模型进行浅层拼接的方法进行预训练，而是采用新的masked language model(MLM)，以致生成深度的双向语言表征。BERT论文发表时提及在11个NLP任务中获得了新的目前最高水平的结果PaLM-E，参数量高达5620亿（GPT-3的参数量为1750亿）。集成语言、视觉，用于机器人控制。相比大语言模型(LLM)，它被称为视觉语言模型(VLM)。VLM与LLM的不同之处，在于VLM对物理世界是有感知的。】 2023年，Facebook的母公司Meta推出LLaMA，并在博客上免费公开LLM——OPT-175B。 ㊟ 【LLaMa有多个不同大小的版本，其中LLaMa65B和LLaMa33B在1.4万亿token上进行了训练。该模型主要在从维基百科、书籍以及来自ArXiv、GitHub、Stack Exchange和其他网站的学术论文中收集的数据集上进行训练。LLaMA模型支持20种语言，包括拉丁语和西里尔字母语言，目前看原始模型并不支持中文。2023年3月，LLaMa模型发生泄露。OPT-175B模型有超过1750亿个参数，和当前世界参数量最大的GPT-3相当。但相比GPT-3，OPT-175B的优势在于它是完全免费的，这使得更多缺乏相关经费的科学家们可以使用这个模型。同时，Meta还公布了代码库。】 在中国，AI大模型的主要代表是百度的文心一言、阿里的通义千问和华为的盘古。

这些模型的共同特征是：需要在大规模数据集上进行训练，基于大量的计算资源进行优化和调整。大模型通常用于解决复杂的NLP、计算机视觉和语音识别等任务。这些任务通常需要处理大量的输入数据，并从中提取复杂的特征和模式。借助大模型，深度学习算法可以更好地处理这些任务，提高模型的准确性和性能。

因为AI大模型的出现和发展所显示的涌现性、扩展性和复合性，长期以来人们讨论的所谓“弱人工智能”“强人工智能”“超人工智能”的界限不复存在，这样划分的意义也自然消失。

2.大模型是人工智能历史的突变和涌现

如果从1956年达特茅斯学院的人工智能会议算起，人工智能的历史已经接近70年（参见图Ⅰ）。




◎图Ⅰ 人工智能发展的历史

图片来源：作者改制自Copeland，Michael，2016，“What’s the Difference Between Artificial Intelligence，Machine Learning，and Deep Learning?”，https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/.。

达特茅斯学院的人工智能会议引申出人工智能的三个基本派别。(1)符号学派(Symbolism)，又称逻辑主义、心理学派或计算机学派。该学派主张通过计算机符号操作来模拟人的认知过程和大脑抽象逻辑思维，实现人工智能。符号学派主要集中在人类推理、规划、知识表示等高级智能领域。(2)联结学派(Connectionism)，又称仿生学派或生理学

派。联结学派强调对人类大脑的直接模拟，认为神经网络和神经网络间的连接机制和学习算法能够产生智能。学习和训练是需要有内容的，数据就是机器学习、训练的内容。联结学派的技术性突破包括感知器、人工神经网络和深度学习。(3)行为学派(Actionism)，该学派的思想来源是进化论和控制论。其原理为控制论以及感知—动作型控制系统。该学派认为行为是个体用于适应环境变化的各种身体反应的组合，其理论目标在于预见和控制行为。

比较上述人工智能的三个派别：符号学派依据的是抽象思维，注重数学可解释性；联结学派则是形象思维，偏向于仿人脑模型；行为学派是感知思维，倾向身体和行为模拟。从共同性方面来说，这三个派别都要以算法、算力和数据为核心要素。但是，在相当长的时间里，符号学派主张的基于推理和逻辑的AI路线处于主流地位。但是，因为计算机只能处理符号，不可能具有人类最为复杂的感知，符号学派在20世纪80年代末开始走向式微。在之后的AI发展史中，有三个重要的里程碑。

第一个里程碑：机器学习(ML)。机器学习理论的提出，可以追溯到图灵写于1950年的一篇论文《计算机器与智能》(Computing Machinery and Intelligence)和图灵测试。1952年，IBM的亚瑟·塞缪尔(Arthur Lee Samuel, 1901—1990)开发了一个西洋棋的程序。该程序能够通过棋子的位置学习一个隐式模型，为下一步棋提供比较好的走法。塞缪尔用这个程序驳倒了机器无法超越书面代码，并像人类一样学习模式的论断。他创造并定义了“机器学习”。之后，机器学习成为一个能使计算机不用显示编程就能获得能力的研究领域。1980年，美国卡内基梅隆大学召开了第一届机器学习国际研讨会，标志着机器学习研究已在全世界兴起。此后，机器学习开始得到大量的应用。1984年，30多位人工智能专家共同撰文编写的《机器学习：一项人工智能方案》(Machine Learning: An Artificial Intelligence Approach)文集第二卷出版；1986年国际性杂志《机器学习》(Machine Learning)创刊，显示出机器学习突飞猛进的发展趋势。这一阶段代表性的工作有莫斯托(Jack Mostow, 1943—)的指导式学习、莱纳特(Douglas Bruce Lenat, 1950—)的数学概念发现程序、兰利(Pat Langley, 1953—)的BACON程序及其改进程序。到了20世纪80年代中叶，机器学习进入最新阶

段，成为新的学科，综合应用了心理学、生物学、神经生理学、数学、自动化和计算机科学等，形成了机器学习理论基础。1995年，瓦普尼克(Vladimir Naumovich Vapnik, 1936—)和科琳娜·科茨(Corinna Cortes, 1961—)提出的支持向量机（网络）(SVM)，实现机器学习领域最重要突破，具有非常强的理论论证和实证结果。

机器学习有别于人类学习，二者的应用范围和知识结构有所不同：机器学习基于对数据和规则的处理和推理，主要应用于数据分析、模式识别、NLP等领域；而人类学习是一种有目的、有意识、逐步积累的过程。总之，机器学习是一种基于算法和模型的自动化过程，包括监督学习和无监督学习两种形式。

第二个里程碑：深度学习(DL)。深度学习是机器学习的一个分支。所谓的深度是指神经网络中隐藏层的数量，它提供了学习的大规模能力。因为大数据和深度学习爆发并得以高速发展，最终成就了深度学习理论和实践。2006年，杰弗里·辛顿(Geoffrey Everest Hinton, 1947—)正式提出深度学习概念，其原理是通过单层的受限制玻尔兹曼机(RBM)自编码预训练实现神经网络训练。2006年也因此成为“深度学习元年”。

在辛顿深度学习的背后，是对“如果不了解大脑，就永远无法理解人类”这一认识的坚信。人脑必须用自然语言进行沟通，而只有1.5千克重的大脑，大约有860亿个神经元（通常被称为灰质）与数万亿个突触相连。人们可以把神经元看作接收数据的中央处理单元(CPU)。所谓深度学习可以伴随着突触的增强或减弱而发生。一个拥有大量神经元的大型神经网络，计算节点和它们之间的连接，仅通过改变连接的强度，从数据中学习。所以，需要用生物学途径，或者关于神经网络途径替代模拟硬件途径，形成基于100万亿个神经元之间的连接变化的深度学习理论。

深度学习是建立在计算机神经网络理论和机器学习理论上的科学。它使用建立在复杂的网络结构上的多处理层，结合非线性转换方法，对复杂数据模型进行抽象，从而识别图像、声音和文本。在深度学习的历史上，CNN和循环神经网络(RNN)曾经是两种经典模型。