



Fastdata 极数

全球人工智能简史

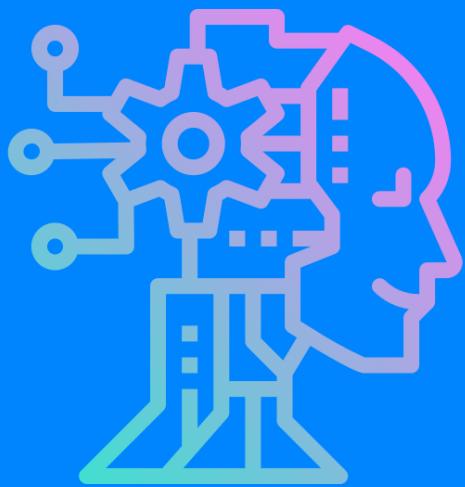
2024

Foreword

前言

如今我们正在进入人工智能 (AI) 带来的第五次工业革命，人工智能技术的运行速度远远快于人类的输出，并且能够生成曾经难以想象的创造性内容，例如文本、图像和视频，这些只是已经发生的一部分。人工智能的发展速度前所未有，要理解我们如何走到今天，就有必要了解人工智能的起源。人工智能的历史可以追溯到19世纪，几乎每几十年都会有重大的里程碑事件出现，并对人类社会产生深远的持续性影响。

尽管计算机和人工智能的历史并不算长，但它们已经从根本上改变了我们所看到的东西、我们所知道的东西以及我们所做的事情。对于世界的未来和我们自己的生活来说，没有什么比这段历史如何延续更重要。要了解未来会是什么样子，研究我们的历史往往很有帮助。这就是本文所要做的，我回顾了计算机和人工智能的简史，人工智能发展历程中发生的一些重大事件，看看我们对未来可以期待什么。

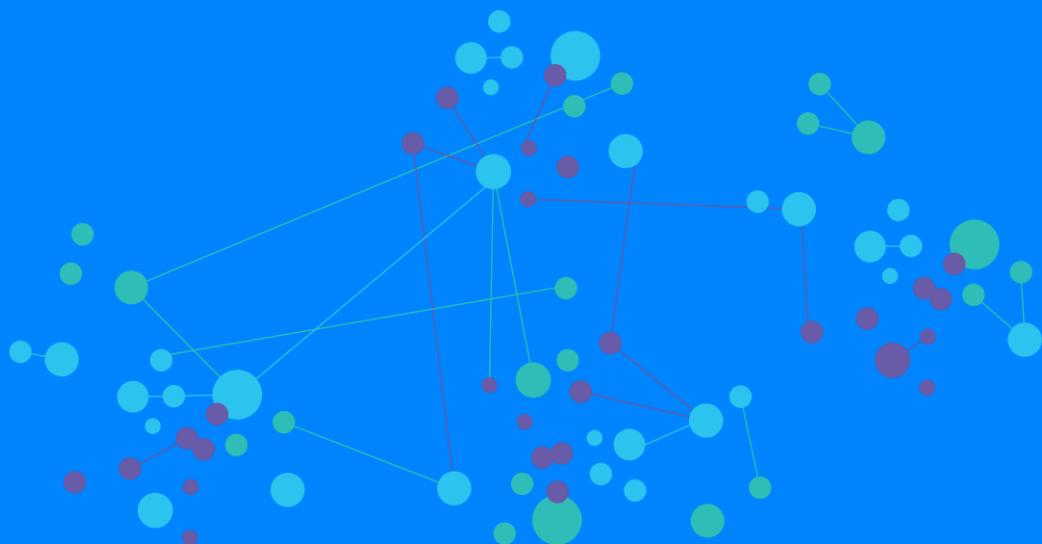


大语言模型简史

A Brief History Of Large Language Models

在瞬息万变的技术世界中，一个迷人的概念已经吸引了科技爱好者的想象力和普通人的好奇心：大型语言模型(LLM)。这些人工智能的非凡壮举不仅可以理解人类语言，还可以生成与人类行为非常相似的文本。随着我们深入探索广阔的人工智能世界，掌握基础知识和推动我们走到这一步的最新突破至关重要。

无论您是想丰富自己对人工智能理解的爱好者，还是对日常接触的技术所依赖的人工智能感兴趣的人，这段探索大型语言模型领域及其历史起源的旅程都将是一次令人着迷的探险。在踏上探索大型语言模型内部工作原理的征程时，我们必须认识到大语言模型在人工智能发展的历史中有着深厚的影响，可以追溯到20世纪中叶。要了解人工智能的发展方向，我们必须回到过去，向众多像艾伦·马西森·图灵这样才华横溢的人致敬，是他们的开创性努力为我们今天看到的LLM格局奠定了基础。



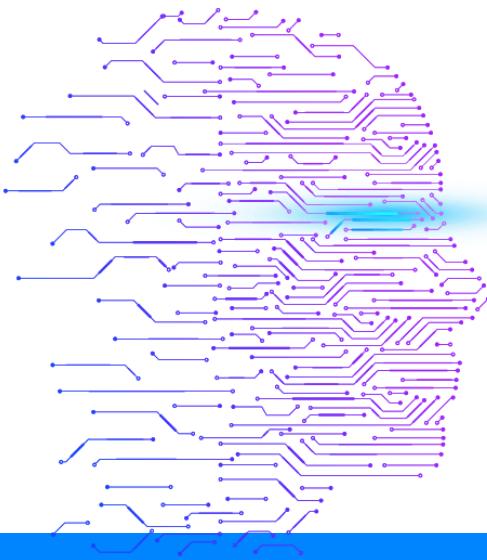
什么是大型语言模型(LLM)？

大型语言模型是生成或嵌入式文本的基础模型（一种大型神经网络）。它们生成的文本可以通过为其提供起点或“提示”来进行调节，从而使其能够用自然语言或代码解决现实世界中有用的问题。

数据科学家和研究人员通过自监督学习，在大量非结构化数据上训练LLM。在训练过程中，模型接受缺少一个或多个单词的单词序列。然后，模型预测缺失的单词，这个过程不仅会为模型产生一组有价值的权重，还会为每个输入的单词产生一个嵌入。

在推理时，用户向LLM提供“提示”——模型用作起点的文本片段。首先，模型将提示中的每个标记转换为其嵌入。然后，它使用这些嵌入来预测可能跟随的所有可能标记的相对可能性。然后，它以半随机的方式选择下一个标记并重复此过程，直到模型选择STOP标记。

你可以把它想象成一条从零到一的数字线。从左边开始，大型语言模型会将标记的概率从大到小堆叠起来。该线的第一部分，从0到0.01，可能是“你好”。第二部分，从0.01到0.019，可能是“世界”，依此类推。然后，模型在该数字线上选择一个随机点并返回与其关联的标记。实际上，大型语言模型通常只将自己限制在具有相对较高可能性的标记上。这就是为什么当输入提示“我去看纽约”时，例如，当GPT-3发布时，其生成的第一个标记几乎总是与该城市相关的运动队或表演场地。





大语言模型简史

• 萌芽前的准备

大型语言模型是一种人工神经网络（算法），在短短几年内就从新兴发展到广泛应用。它们在ChatGPT的开发中发挥了重要作用，而ChatGPT是人工智能的下一个进化步骤。生成式人工智能与大型语言模型相结合，产生了更智能的人工智能。大型语言模型(LLM)基于人工神经网络，深度学习的最新改进支持了其发展。

大型语言模型还使用语义技术（语义学、语义网和自然语言处理）。大型语言模型的历史始于1883年法国语言学家米歇尔·布雷亚尔提出的语义概念。米歇尔·布雷亚尔研究了语言的组织方式、语言随时间的变化以及语言中单词的连接方式。目前，语义用于为人类开发的语言，例如荷兰语或印地语，以及人工智能编程语言，例如Python和Java。

然而，自然语言处理专注于将人类交流内容翻译成计算机能够理解的语言，然后再翻译回来。它使用能够理解人类指令的系统，使计算机能够理解书面文本、识别语音并在计算机和人类语言之间进行翻译。1906年至1912年，费迪南·德·索绪尔在日内瓦大学教授印欧语言学、普通语言学和梵语。



米歇尔·布雷亚尔
1832年-1915年

- 法国语言学家，现代语义学先驱



费迪南·德·索绪尔
1857年-1913年

- 瑞士语言学家、符号学家、哲学家



沃伦·麦卡洛克

1898年-1969年

- 美国神经科学家和控制论学者，麦卡洛克与沃尔特·皮茨一起创建了基于称为阈值逻辑的数学算法的计算模型，该模型将研究分为两种不同的方法，一种方法专注于大脑中的生物过程，另一种方法专注于神经网络在人工智能中的应用。



沃尔特·皮茨

1923年-1969年

- 美国逻辑学家和计算神经科学家。[他提出了神经活动和生成过程的里程碑式的理论表述，影响了认知科学和心理学、哲学、神经科学、计算机科学、人工神经网络、控制论和人工智能等不同领域，以及后来被称为生成科学的领域。

在此期间，他为语言系统这一高度实用的模型奠定了基础。他在1913年去世，没有整理和出版他的作品。幸运的是，索绪尔的同事、两位导师艾伯特·塞切海耶和查尔斯·巴利认识到索绪尔概念的潜力，并认为这些概念值得保存。这两位导师收集了他的笔记，以备将来手稿之用，然后努力收集索绪尔学生的笔记。基于这些笔记，他们撰写了索绪尔的书，名为《通用语言学课程》（又译为《语言作为一门科学，最终演变为自然语言处理（NLP）》），并于1916年出版。语言作为一门科学奠定了结构主义方法以及后来的自然语言处理。

• 加速孕育阶段

1943年，美国神经生理学家沃伦·麦卡洛克和认知心理学家沃尔特·皮茨发表了一项研究报告。研究名称为《神经活动中内在思想的逻辑演算》。在这项研究中，讨论了人工神经网络的第一个数学模型。该论文提供了一种以抽象术语描述大脑功能的方法，并表明连接在神经网络中的简单元素可以具有巨大的计算能力。在《神经活动中内在思想的逻辑演算》奠定了人工神经网络的基础，是现代深度学习的前身，其神经元的数学模型：M-P模型一直沿用至今。在不远的未来，以神经网络为基础思想的科学家们，会大大发展人工神经网络的成果。

如果说符号主义是利用逻辑学，自上而下的通过推理演绎的方式解决人工智能这个课题的话，人工神经网络则是利用神经科学，自下而上的通过模拟人脑思考的原理来解决人工智能这个课题。这些科学家们形成了人工智能中的另一个重要的派别，后世称其为“联结主义（Connectionists）”。

假设有人要求你设计出最强大的计算机。艾伦·图灵是计算机科学和人工智能领域的核心人物，自1954年他英年早逝后，他的声誉才得以提升。在我们所知的计算机出现之前的时代，他将自己的天才运用到解决此类问题上。他对这个问题和其他问题的理论研究仍然是计算、人工智能和现代加密标准（包括NIST推荐的标准）的基础。

二次世界大战期间，“Hut8”小组，负责德国海军密码分析。期间图灵设计了一些加速破译德国密码的技术，包括改进波兰战前研制的机器Bombe，一种可以找到恩尼格玛密码机设置的机电机器。图灵在破译截获的编码信息方面发挥了关键作用。图灵对于人工智能的发展有诸多贡献，图灵曾写过一篇名为《计算机器和智能》的论文，提问“机器会思考吗？”，作为一种用于判定机器是否具有智能的测试方法，即图灵测试。至今，每年都有试验的比赛。此外，图灵提出的著名的图灵机模型为现代计算机的逻辑工作方式奠定了基础。

图灵于1947年在伦敦的一次公开演讲中宣称，机器修改自身指令的潜力在大型语言模型领域具有重要意义。它强调了大型语言模型的适应能力、持续改进、解决各种问题的能力以及紧跟不断发展的语言趋势的能力。这个想法与大语言模型的动态性质完全吻合，使大语言模型能够在瞬息万变的语言环境中获取知识、进行调整并保持最新状态。



艾伦·图灵
1912年-1954年

• 英国计算机科学家、数学家、逻辑学家、密码分析学家和理论生物学家，被誉为计算机科学与人工智能之父。

计算机在语言相关任务中的最早用途之一是机器翻译(MT)，即使用计算机来翻译语言。第二次世界大战期间，两位擅长破解敌方秘密密码的人(1964年)开始了首批使用计算机进行翻译的项目之一。此后，美国各研究机构在接下来的几年里开始研究这个想法。这标志着使用计算机进行语言翻译和理解的研究的开始，也是导致我们今天所拥有的技术的早期步骤之一。



亚瑟·塞缪尔
1901年-1990年

- 美国计算机科学家，他是电脑游戏与人工智能方面的先锋。塞缪尔的电脑跳棋程式是世界上最早能成功进行自我学习的计算机程序之一，也因此是人工智能(AI)基础概念的早期展示之一。

• 自然语言处理的开始

自然语言处理(NLP)的起源可以追溯到20世纪50年代，当时机器理解和处理人类语言的想法还处于起步阶段。正是在这个时代，IBM和乔治城大学(1954)的研究人员开始了一个开创性的项目。他们的目标是开发一个可以自动将一组短语从俄语翻译成英语的系统，这是最早的机器语言翻译项目之一。

然而，掌握自然语言处理的道路绝非易事。在接下来的几十年里，研究人员尝试了各种方法，包括概念本体和基于规则的系统。尽管他们尽了最大努力，但这些早期尝试都没有取得可靠的结果，这凸显了教机器掌握人类语言的复杂性。

• 基于规则的模型

• 机器学习和跳棋游戏

IBM的亚瑟·塞缪尔开发了一个计算机程序下跳棋，在20世纪50年代初。他完成了一系列算法，使他的跳棋程序得以改进，并在1959年将其描述为“机器学习”。



• Mark1感知器使用神经网络

1958年，康奈尔航空实验室的弗兰克·罗森布拉特将赫布的神经网络算法模型与塞缪尔的机器学习工作相结合，创建了第一个人工神经网络，称为Mark1感知器。尽管语言翻译仍然是一个目标，但计算机主要是为数学目的而制造的（比语言混乱得多）。这些用真空管制造的大型计算机用作计算器，计算机和软件都是被定制的。感知器的独特之处还在于它使用了为IBM704设计的软件，并确定了类似的计算机可以共享标准化的软件程序。

在1960年MarkI感知机的开发和硬件建设中达到了顶峰。从本质上讲，这是第一台可以通过试错来学习新技能的计算机，它使用了一种模拟人类思维过程的神经网络。MarkI感知机被公认为人工智能的先驱，目前位于华盛顿特区的史密森尼博物馆。MarkI能够学习、识别字母，并能解决相当复杂的问题。

1969年，明斯基和西摩·佩珀特出版了《感知机》一书，彻底改变人们对感知机的看法。不幸的是，Mark1感知器无法识别许多种基本的视觉模式（例如面部），导致期望落空，神经网络研究和机器学习投入也被消减。

• ELIZA使用自然语言编程

直到1966年，麻省理工学院的计算机科学家约瑟夫·魏森鲍姆开发了ELIZA，它被称为第一个使用NLP的程序。它能够从收到的输入中识别关键词，并以预先编程的答案做出回应。魏森鲍姆试图证明他的假设，即人与机器之间的交流从



弗兰克·罗森布拉特
1928年-1971年

- 美国人工智能领域著名心理学家。他有时被称为深度学习之父



瑟夫·魏森鲍姆
1923年-2008年

- 美国计算机科学家。麻省理工学院的荣休教授，1966年他发表了一个简单的名为ELIZA的机器人小程序。通过一个名为DOCTOR的脚本此程序人类可以和其以极类似心理学家的方式交谈。

从根本上说是肤浅的，但事情并没有按计划进行。为了简化实验并尽量减少争议，魏森鲍姆开发了一个程序，使用“积极倾听”，它不需要数据库来存储现实世界的信息，而是会反映一个人的陈述以推动对话向前发展。

尽管Eliza的功能相对有限，但它代表了该领域的一次重大飞跃。这个开创性的程序使用模式识别来模拟对话，将用户输入转换为问题并根据预定义规则生成响应。尽管Eliza远非完美，但它标志着自然语言处理(NLP)研究的开始，并为开发更高级的语言模型奠定了基础。

• SHRDLU-理解自然语言的软件诞生

1970年特里·维诺格拉德在麻省理工学院(MIT)创建了SHRDLU，为人工智能领域做出了杰出贡献。SHRDLU是一款旨在理解自然语言的创新软件。它主要通过电传打字机与用户进行对话，讨论一个称为“积木世界”的封闭虚拟环境。在这个世界中，用户可以通过移动物体、命名集合和提出问题进行交互。SHRDLU的突出之处在于它能够熟练地结合名词、动词和形容词等基本语言元素，尽管虚拟世界很简单，但它却能够熟练地理解用户指令。



特里·维诺格拉德
1946年-

- 美国斯坦福大学计算机科学教授。他主要从事人工智能与心灵哲学领域的研究。他开发了知名的自然语言理解程序SHRDLU

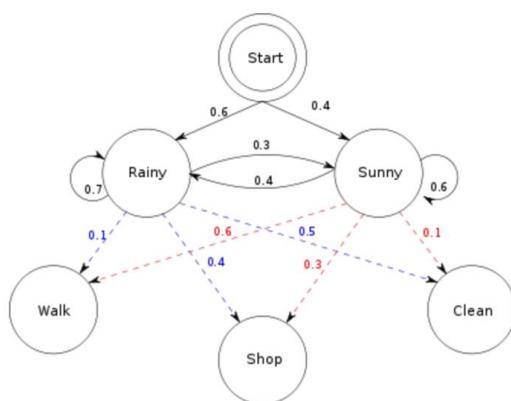
• 统计语言模型

20世纪90年代，我们处理语言的方式发生了重大变化。研究人员不再依赖严格的规则，而是开始使用统计模型来分析现实世界的文本示例。这些模型更加灵活，可以处理更广泛的语言模式，但它们需要大量的计算机能力和大量数据集才能正常工作。

20世纪70年代初，人工智能领域由伦纳德·鲍姆(1971)等人引入了隐马尔可夫模型(HMM)和条件随机场(CRF)。HMM使用概率来判断句子中发生了什么，例如识别单词的角色（名词、动词等）。它们非常擅长处理单词序列并找到句子背后最可能的故事，这使得它们对于语音识别和词性标注等任务非常有用。转向统计方法提高了语言处理

的灵活性和上下文敏感性。尽管如此，它们也需要大量的计算资源和数据才能有效执行。这种转变也带来了新的障碍，为语言建模领域的未来发展铺平了道路。

在20世纪90年代和21世纪初期，N-gram模型对统计语言建模做出了重大贡献。这些模型简单但功能强大。它们通过查看某个单词前面的单词序列来估计该单词出现的可能性。这种直接的方法有助于理解语言的上下文。N-gram的一个突出用途是Google的PageRank算法（1996年）。本质上，N-gram模型强调了语言中语境的重要性，并为能够捕捉更广泛的语言细微差别的更先进的技术奠定了基础。



HMM模型示意图

• 深度学习模型

1983年，辛顿发明玻尔兹曼机，后来，简化后的受限玻尔兹曼机被应用于机器学习，成为深度神经网络的层级结构基础。1986年，辛顿提出适用于多层感知机的误差反向传播算法（BP），这一算法奠定了后来深度学习的基础。辛顿每隔一段时间都能发明出新东西，而他也坚持写了两百多篇神经网络相关的论文，尽管这些论文不被待见。到了2006年，辛顿已经积累了丰富的理论和实践基础，而这一次，他发表的论文将改变整个机器学习乃至整个世界。

辛顿发现，拥有多个隐藏层的神经网络能够具有自动提取特征学习的能力，相比传统的手工提取特征的机器学习更有效果。另外，通过逐层预训练的方式可以降低多层神经网络的训练难度，而这解决了长期以来多层神经网络训练的难题。辛顿将他的研究成果发表在两篇论文中，而当时神经网络一词被许多学术期刊编辑所排斥，有些稿件的标题甚至因为包含“神经网络”就会被退回。为了不刺激这些人的敏感神经，辛顿取了个新名字，将该模型命名为“深度信念网络”（Deep Belief Network）。

在20世纪90年代，卷积神经网络(CNN)被引入。CNN主要用于图像处理，但也可用于某些NLP任务，例如文本分类。人工智能和神经网络架构的这些发展，包括感知器(1960)、RNN、LSTM和CNN，共同塑造了自然语言处理和深度学习的格局，为理解和处理人类语言开辟了新的可能性。



杰弗里·辛顿
1947年-

- 英国出生的加拿大计算机学家和心理学家，多伦多大学教授。以其在类神经网络方面的贡献闻名。辛顿是反向传播算法和对比散度算法的发明人之一，也是深度学习的积极推动者，被誉为“深度学习教父”。辛顿因在深度学习方面的贡献与约书亚·本希奥和杨立昆一同被授予了2018年的图灵奖。2024年因其在AI领域的卓越贡献，获得2024年诺贝尔物理学奖。

1986年，循环神经网络(RNN)能够捕捉语言中的序列依赖关系，但它面临着长距离依赖关系和梯度消失的挑战。同时，在语言建模的早期，杰弗里·洛克·埃尔曼于1990年开发的循环神经网络语言模型(RNNLM)发挥了重要作用。该模型擅长识别序列中的短期单词关系，但在捕获长距离依赖关系时其局限性变得明显，促使研究人员探索替代方法。

除了RNNLM之外，该领域还出现了潜在语义分析(LSA)，它由朗道尔和杜迈斯于1997年提出。LSA利用高维语义空间来揭示文本数据中隐藏的关系和含义。虽然它提供了对语义关联的宝贵见解，但在处理更复杂的语言任务时遇到了某些限制。RNNLM和LSA的贡献以及其他具有影响力的里程碑共同塑造了语言建模取得重大进步的道路。

1997年，长短期记忆(LSTM)模型的推出改变了游戏规则。LSTM允许创建更深层、更复杂的神经网络，能够处理大量数据。门控循环单元(GRU)是深度学习和自然语言处理领域的一个显著新成员。GRU由Kyung-hyun Cho及其团队2014年是一种循环神经网络架构，采用门控机制来控制输入并忘记某些特征，类似于长短期记忆(LSTM)网络。然而，GRU与LSTM的区别在于，它们没有上下文向量或输出门，因此架构更简单，参数更少。研究表明，GRU在各种任务中的表现与LSTM相似，包括复音音乐建模、语音信号建模和自然语言处理。这一发现凸显了门控机制在循环神经网络中的价值，并促进了自然语言处理神经网络架构的持续进步。



杰弗里·克·埃尔曼
1948年-2018年

- 美国心理语言学家，加州大学圣地亚哥分校 (UCSD) 认知科学教授。他专门研究神经网络领域。1990年，他提出了简单循环神经网络 (SRNN)，又称“Elman 网络”。

• 图形处理单元(GPU)的诞生

在1999年推出第一款GPU（Nvidia GeForce256）之前，NLP模型完全依赖CPU进行推理。具有并行处理能力的GPU的引入将标志着一个关键的转变，因为它将允许高效执行NLP任务，从而能够处理以前仅靠CPU无法实现的大型文本数据集和复杂计算。这项GPU技术将彻底改变深度学习模型，并将在机器翻译和文本生成等任务方面取得重大进展。

• 词嵌入

此外，神经网络开始用于预测文本中的下一个单词。约书亚·本希奥等人（2003年）提出了第一个神经语言模型，使用一个隐藏层前馈神经网络和开创性的词嵌入。自从谷歌的Tomas Mikolov和他的团队于2013年推出Word2Vec以来，人们开始更多地使用神经网络来完成语言任务。这些词向量将单词表示为连续空间中的密集向量，标志着传统方法的转变，并显著改善了语言理解和单词间语义关系的建模。利用神经网络进行语言建模使系统能够预测句子中的下一个单词，超越了统计分析并产生了更复杂的语言模型。

• Seq2Seq模型

2015年，Bahdanau等人提出了序列到序列模型（Seq2Seq），这是一种神经网络，可以有效地将可变长度的输入序列映射到可变长度的输出序列。Seq2Seq模型架构由两个关键组件组成：编码器和解码器。编码器负责处理输入序列，产生一个固定长度的上下文向量，该向量封装了输入序列的含义。



Nvidia GeForce256
1999年

- 1999年英伟达推出了一款名为“GeForce256”的显卡，其具有强大的3D图形渲染力和可编程性能，也被业界誉为世界上第一个真正意义上的GPU，GPU在后来的人工智能发展中扮演着非常重要的角色。



约书亚·本希奥
1964年-

- 加拿大计算机科学家，因其在人工神经网络和深度学习方面的研究而知名。本希奥与杰弗里·辛顿和杨立昆一起获得2018年的图灵奖，以表彰他们在深度学习方面的贡献[5]。这三人有时被称为“AI教父”和“深度学习教父”。



阿希什·瓦斯瓦尼

- 阿希什·瓦斯瓦尼是一位从事深度学习工作的计算机科学家，他因在人工智能(AI)和自然语言处理(NLP)领域的重大贡献而闻名。他是开创性论文《注意力就是你需要的》的合著者之一，该论文介绍了Transformer模型，这是一种使用自注意力机制的新颖架构，此后已成为NLP中许多最先进模型的基础。Transformer架构是支持ChatGPT等应用程序的语言模型的核心。他是AdeptAILabs的联合创始人，曾担任GoogleBrain的研究员。

解码器随后利用该上下文向量逐步生成输出序列。更详细地说，编码器通常采用循环神经网络(RNN)逐个元素处理输入序列，在每一步创建一个固定长度的隐藏状态向量。最后一个隐藏状态向量用作上下文向量并传递给解码器。

解码器通常也以RNN的形式实现，它采用上下文向量并按顺序生成输出序列。它通过为每个步骤的潜在输出元素生成概率分布，然后通过从该分布中采样来选择输出序列的下一个元素来实现这一点。然而，尽管Seq2Seq模型取得了成功，但它们也存在一定的局限性，尤其是在处理NMT任务中的较长序列时。当谷歌于2016年推出其“神经机器翻译”系统时，这些局限性就变得显而易见，这展示了深度学习在语言相关任务中的强大功能，并标志着机器翻译能力的重大进步。最终，2017年Google提出的Transformer架构的引入解决了Seq2Seq模型的许多缺点，从而显著提高了NMT性能。NMT技术的这种发展凸显了自然语言处理的动态性质以及对更有效解决方案的不断追求。

• 百花齐放时代来临

从传统的序列到序列模型到开创性的Transformer的演变重塑了大型语言模型的格局。Ashish Vaswani在2017年的论文《注意力就是你需要的一切》中引入Transformer模型，带来了并行序列处理和捕获大型序列中广泛依赖关系的能力。关键创新在于它们使用了自注意力机制，能够无缝集成序列内所有位置的上下文信息，消除了对递归和卷积的需求，从而产生了更可并行化、训练速度更快的优质模型。

自注意力机制代表了一次重大飞跃，它允许模型关注输入序列的不同部分，并根据相关性分配不同的权重，即使单词相距很远。这一功能对于文本生成、语言翻译和文本理解等任务至关重要。

- **GPT-1**

2018年，OpenAI推出了他们的第一个大型语言模型GPT-1。这是谷歌在2017年创建了一种名为“Transformer”的新型计算机程序结构之后推出的。OpenAI在一篇名为《通过生成式预训练提高语言理解能力》的论文中分享了他们的工作。这篇论文不仅介绍了GPT-1，还介绍了生成式预训练Transformer的概念。

- **BERT**

2018年，谷歌推出了Transformer双向编码器表示(BERT)，这是一个重大突破，凸显了预训练模型的潜力。BERT代表了一种革命性的方法，它涉及在大量文本数据上训练广泛的Transformer模型，并针对特定任务对其进行微调，标志着语言建模新时代的到来。

BERT的影响是深远的，因为它在各种自然语言理解任务（包括问答和情感分析）中建立了新标准。这标志着从僵化、特定于任务的模型转向更具适应性和可迁移性的全新模型。通过在预训练期间利用大量可用的文本数据，BERT深入了解了语言的微妙之处和上下文关系，重塑了自然语言处理的格局。



- GPT-2



OpenAI于2019年推出了GPT-2，这标志着LLM领域的一个转折点。GPT-2拥有15亿个参数，展示了生成式模型的巨大潜力。该模型具有准确预测序列中下一个元素的一般能力。然而，对滥用的担忧导致了谨慎的发布策略。该模型能够生成连贯且上下文丰富的文本，证明了深度学习和NLP的快速进步。同时，百度ERNIE、XLNet、XLMERT（微软）、RoBERTa（Facebook）等模型出现在LLM领域，开创了自然语言处理可能性和能力的新时代。

- 巨型模型兴起及大语言模型爆炸式增长



2020年，OpenAI发布了GPT-3，这是一款拥有1750亿个参数的大型LLM。GPT-3突破了LLM的极限。它在语言翻译和文本完成、编码辅助和交互式讲故事等任务中表现出色。它的“few-shot”和“zero-shot”学习能力非常出色，使其能够用最少的训练示例执行任务。GPT-3引入了“提示工程”的概念，使用户能够根据自己的需求调整其响应。其他型号如Megatron(Nvidia)、Blender(Facebook)、T5(Google)和Meena(Google)也在当年推出。

2021年，LLM社区因引入各种新模式而热闹非凡：Transformer-X（谷歌）、GPT-Neo（Eleuther AI）、XLM-R（Facebook）、LaMDA（谷歌）、Copilot（GitHub）、GPT-J（Eleuther AI）、Jurassic-1（AI21）、Megatron-TuringNLG、Codex（OpenAI）、WebGPT（OpenAI）和BERT2（谷歌）。每个模型都有其独特的优势，为不断发展的NLP领域做出了贡献，但名为LoRA的训练技术却吸引了人们的注意力。

• LoRA

低秩自适应(LoRA)是一种突破性的训练方法，旨在加快大型语言模型的训练，同时节省内存资源。LoRA将秩分解权重矩阵（称为更新矩阵）引入现有模型权重，并将训练工作完全集中在这些新增加的权重上。这种方法有两个明显的优势：首先，预训练模型的权重保持不变，降低了灾难性遗忘的风险。其次，LoRA的秩分解矩阵的参数明显较少，使得训练后的LoRA权重易于迁移。

• QLoRA

紧接着一种突破性的方法问世，可以加速量化模型的微调，同时保持其性能。这项创新被称为QLoRA（量化低秩自适应），它为大型语言模型领域带来了范式转变。QLoRA以LoRA（低秩自适应）为基础，通过引入一系列新技术将其提升到新的水平，这些技术不仅可以减少内存需求，还可以提高微调过程的效率。就像厨师将食谱数字化以节省空间同时保留进行调整的能力一样，QLoRA使研究人员和开发人员能够有效地微调大型语言模型，即使在计算资源有限的情况下也是如此。



从本质上讲，QLoRA利用了多项关键创新，包括NormalFloat(NF4)、DoubleQuantization,和PagedOptimizers,，这些创新共同实现了对大规模模型的微调，同时保持了性能。这一重大突破使大型语言模型微调变得民主化，使小型研究团队能够使用它，并预示着自然语言处理的新可能性。随着我们继续突破该领域的可能性界限，QLoRA无疑将在塑造NLP的未来方面发挥关键作用。

- Lamda

Lamda（对话应用语言模型）是Google Brain于2021年发布的LLM系列。Lamda使用了仅解码器的转换器语言模型，并在大量文本语料库上进行了预训练。2022年，当时的谷歌工程师Blake Lemoine公开声称该程序具有感知能力，Lamda引起了广泛关注。它建立在Seq2Seq架构上。



布莱克·雷蒙恩

- 2015年加入Google，担任软件工程师。负责Google人工智能模型LaMDA的开发。通过创建的聊天机器人测试了LaMDA，布莱克·雷蒙恩为了观测应用LaMDA创建的聊天机器人是否存在与性取向、性别、宗教、政治立场和种族有关的偏见。但在测试偏见时，在与聊天机器人的对话中，得出结论是人工智能是会产生情感的。

• 开源模型的兴起

2022年，开源大型语言模型(LLM)领域经历了重大变革，一些先驱模型引领了潮流。Eleuther AI的创作GPT-NeoX-20B是最早的开源LLM之一。尽管它的规模较小（与GPT-3等专有模型相比，它有200亿个参数），但它通过RoPE嵌入和并行注意层等创新产生了影响。它的自定义标记器可有效进行代码标记化，并在各种开源模型中得到采用。

Meta AI的开放式预训练Transformers(OPT)计划旨在使LLM的获取更加民主化。OPT提供不同大小的模型，在精选数据集上进行预训练，并提供开源训练框架。

虽然OPT模型的表现并不优于专有模型，但它们在使LLM更易于研究和提高训练效率方面发挥了关键作用。

BLOOM是一个包含1760亿个参数的LLM，它诞生于1000多名研究人员历时一年的大规模协作。它使用多语言文本数据集ROOTS语料库进行训练。尽管BLOOM在各种基准测试中都具有竞争力，并且在机器翻译任务中表现出色，但在某些方面仍然落后于专有模型。GPT-J和GLM等著名模型也取得了成功，为开源LLM领域的进一步发展奠定了基础。2022年标志着语言模型领域向开放可访问性和协作研究的重大转变。

- LoRA

低秩自适应(LoRA)是一种突破性的训练方法，旨在加快大型语言模型的训练，同时节省内存资源。LoRA将秩分解权重矩阵（称为更新矩阵）引入现有模型权重，并将训练工作完全集中在这些新增加的权重上。这种方法有两个明显的优势：首先，预训练模型的权重保持不变，降低了灾难性遗忘的风险。其次，LoRA的秩分解矩阵的参数明显较少，使得训练后的LoRA权重易于迁移。

GPT-4的问世

2023年，OpenAI发布了GPT-4，在大型语言模型(LLM)领域迈出了开创性的一步。GPT-4是一个庞大的多模态模型，拥有约一万亿个参数。从这个角度来看，GPT-4比其前身GPT-3大约五倍，比原始BERT模型大3,000倍。这一规模和容量上的巨大飞跃改变了LLM领域的格局，使其能够一次性处理多达50页的文本。





要真正了解GPT-4的演变，了解这些模型的时间顺序至关重要。近年来，我们见证了几个值得关注的LLM的发展，它们为当前的技术水平做出了贡献。这些模型为GPT-4的出现铺平了道路，它们反映了开源LLM研究的充满活力和生机勃勃的前景。

• SOTA开源模型集

这一演变的关键时刻之一是2023年2月MetaAI推出LLaMA。LLaMA是人工智能领域的一项突破性进展。其重要性在于它作为Meta向公众发布的基础大型语言模型。LLaMA的重要性可以从几个角度来理解：它通过提供更易于访问且性能更高的大型语言模型替代方案，使人工智能研究的访问变得民主化，减少了人工智能实验所需的计算资源，并为更多开源计划（如Alpaca、Vicuna、Dolly、WizardLM）奠定了基础。此外，LLaMA用途广泛，可以针对各种应用进行微调，解决偏见和歧视等人工智能挑战，同时通过受控访问坚持负责任的人工智能实践。

继LLaMA之后，MosaicML的MPT套件提供了开源LLM的商业可用替代方案。初始版本MPT-7B引起了广泛关注，随后是更大的MPT-30B模型。这些模型提供了质量和商业可行性的精彩融合，拓展了开源LLM应用的视野。

另一个值得注意的进展是FalconLLM套件，其性能可与专有模型相媲美。Falcon-7B和Falcon-40B虽然是商业是可行的，结果表现也相当出色。这些模型挑战了有关数据质量的传统观念，表明在经过精心过滤和重复数据删除的网络数据上训练的模型可以与在精选来源上训练的模型相媲美。

LLaMA-2模型套件通过缩小开源和闭源LLM之间的差距标志着另一个重要里程碑。LLaMA-2的参数大小从70亿到700亿不等，并在2万亿个token的海量数据集上进行预训练，突破了开源模型性能的界限。

大型语言模型(LLM)领域最显著的进步之一是Zephyr7B模型，它是Mistral-7B-x0.1的微调版本。Zephyr7B拥有卓越的功能，这主要归功于它利用了精炼直接偏好优化(dDPO)和AI反馈(AIF)，使其能够与用户意图紧密结合。值得注意的是，该模型的性能不仅创下了新基准，而且令人印象深刻的是，在聊天基准测试中甚至超越了备受推崇的Llama2-Chat-70B，展示了其实力。

Zephyr7B真正与众不同之处在于其卓越的效率。该模型以惊人的速度实现了卓越的性能，仅需几个小时的训练。值得注意的是，这种效率是在无需人工注释或额外采样的情况下实现的，使其成为利用技术简化模型开发流程的出色范例。Zephyr7B的创新方法将传统的蒸馏监督微调(dSFT)与偏好数据相结合，展示了融合各种技术的潜力，以创建一个重新定义自然语言理解和生成领域可实现的边界的模型。

• Orca

Orca由微软开发，拥有130亿个参数，这意味着它足够小，可以在笔记本电脑上运行。它旨在通过模仿LLM实现的推理过程来改进其他开源模型所取得的进步。Orca以明显更少的参数实现了与GPT-4相同的性能，并且在许多任务上与GPT-3.5相当。Orca建立在130亿个参数版本的LLaMA之上。



- Gemini

Gemini是Google在2023年6月发布的，为该公司的同名聊天机器人提供支持。该模型取代了Palm为聊天机器人提供支持，在模型切换后，聊天机器人从Bard更名为Gemini。Gemini模型是多模态的，这意味着它们可以处理图像、音频和视频以及文本。Gemini还集成在许多Google应用程序和产品中。它有三种尺寸——Ultra、Pro和Nano。Ultra是最大、功能最强大的模型，Pro是中端模型，Nano是最小的模型，专为提高设备上任务的效率而设计。Gemini在大多数评估基准上都优于GPT-4。

2024年2月9日，谷歌宣布GeminiUltra可免费使用，16日发布Gemini1.5，21日发布开源模型Gemma。Gemma采用了与Gemini相同的技术和基础架构，基于英伟达GPU和谷歌云TPU等硬件平台进行优化，有20亿、70亿两种参数规模。每种规模都有预训练和指令微调版本，使用条款允许所有组织（无论规模大小）负责任地进行商用和分发。谷歌介绍，Gemma模型与其规模最大、能力最强的AI模型Gemini共享技术和基础架构。

2024年6月28日，谷歌宣布面向全球研究人员和开发者发布Gemma2大语言模型。据介绍，Gemma2有90亿（9B）和270亿（27B）两种参数大小，与第一代相比，其性能更高、推理效率更高，并且内置了显著的安全改进。谷歌称，Gemma227B的性能比大其两倍的同类产品更具竞争力；9B的性能也处于同类产品领先水平，优于Llama38B和其他开放模型。



- Gork



2023年11月5日，马斯克旗下xAI团队发布其首个AI大模型产品——Grok。据介绍，Grok通过X平台实时了解世界，还能回答被大多数其他AI系统拒绝的辛辣问题。

2024年4月15日xAI的多模态模型Grok-1.5V发布，不仅多项基准测试超越GPT-4V，而且看懂梗图写Python代码也都不在话下。并且，为了评估模型对于真实世界的空间理解，xAI此次还推出了新基准RealWorldQA。2024年8月13日，马斯克旗下xAI正式发布语言模型Grok-2早期预览版，该系列模型具有聊天、编码和推理等功能，包括Grok-2和Grok-2mini两个版本。

- GPT-4o



GPT-4o是由OpenAI训练的多语言、多模态（多种类型数据，例如文本、图像、音频等）GPT大型语言模型。GPT-4o于2024年5月13日发布。该模型比其前身GPT-4快两倍，而价格仅为50%。GPT-4Omni（GPT-4o）是OpenAI的GPT-4继任者，与之前的模型相比有多项改进。GPT-4o为ChatGPT创造了更自然的人机交互，是一个大型多模态模型，接受音频、图像和文本等各种输入。对话让用户可以像在正常的人类对话中一样参与，实时互动还可以捕捉情绪。GPT-4o可以在交互过程中查看照片或屏幕并提出相关问题。GPT-4o的响应时间仅为232毫秒，与人类的响应时间相似，比GPT-4Turbo更快。GPT-4o模型是免费的，将提供给开发者和客户产品。



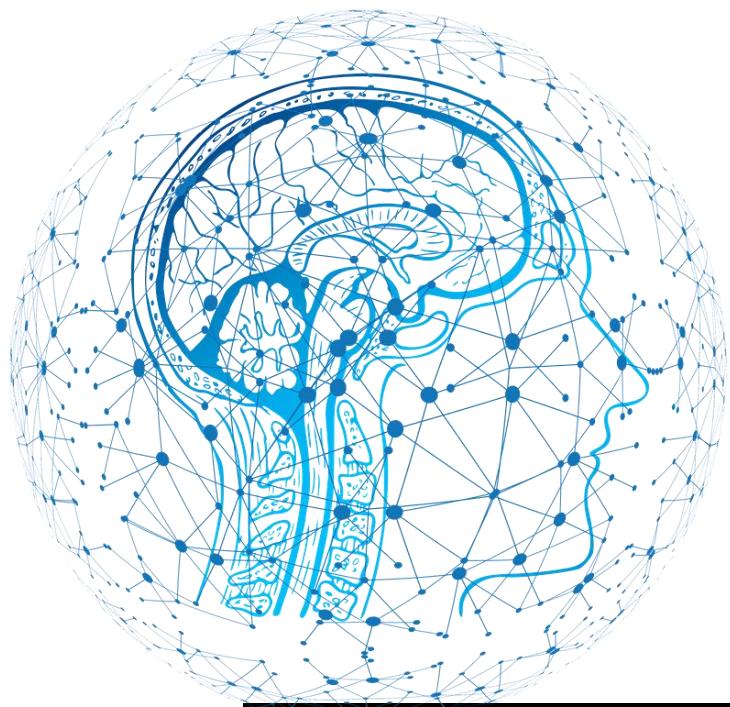
OpenAIo1，是OpenAI发布的推理模型系列。该模型在回答问题前会进行深入思考，并生成一条内部推理链，使其在尝试解决问题时可以识别并纠正错误，将复杂的步骤分解为更简单的部分，并在当前方法无效时尝试不同的途径。OpenAIo1包括三个型号，除o1-preview之外还将有o1和o1-mini。

从二战期间机器翻译的早期发展，到GPT-4o等强大模型的出现，再到LLaMA等开源计划的出现，我们见证了人工智能和自然语言处理领域的深刻变革。时间轴见证了人类的智慧、奉献和协作。

我们见证了从基于规则的模型到统计方法的转变，以及最终改变游戏规则的Transformer架构的引入，这使得GPT-4o等模型成为可能。在此过程中，BERT和Seq2Seq等模型留下了自己的印记，重新定义了我们理解语言的方式。LoRA和QLoRA等最新创新有望使大型语言模型微调变得民主化，为更多研

究人员和开发人员打开大门。

我们已经走了很长一段路，未来，我们必须时刻牢记道德考量、可访问性和负责任的人工智能发展。我们可以共同努力，继续塑造一个语言模型赋能并连接全球人民的世界。我们希望这次探索能让您更深入地了解语言模型的历史和潜力。当我们探索这个激动人心的人工智能领域时，让我们记住，旅程还未结束，可能性无穷无尽。



生成式人工智能崛起

•什么是GenAI?

生成式人工智能(GenAI)是一种人工智能技术，可以生成各种类型的内容，包括文本、图像、音频和合成数据。最近，围绕生成式人工智能的讨论是由新用户界面的简单性推动的，该界面可以在几秒钟内创建高质量的文本、图形和视频。

需要注意的是，这项技术并非全新技术。生成式人工智能于20世纪60年代在聊天机器人中被引入。但直到2014年，随着生成对抗网络(GAN，一种机器学习算法)的引入，生成式人工智能才能够创建令人信服的真实人物图像、视频和音频。

Transformers使研究人员能够训练越来越大的模型，而无需事先标记所有数据。因此，新模型可以在数十亿页文本上进行训练，从而得到更有深度的答案。此外，Transformers还开启了一种名为注意力的新概念，使模型能够跟踪跨页面、跨章节和跨书籍的单词之间的联系，而不仅仅是单个句子之间的联系。不仅仅是单词：Transformers还可以利用其跟踪联系的能力来分析代码、蛋白质、化学物质和DNA。

所谓的大型语言模型(LLM)（即具有数十亿甚至数万亿个参数的模型）的快速发展开启了一个新时代，在这个时代，生成式人工智能模型

可以编写引人入胜的文本、绘制逼真的图像，甚至可以即时创建一些有趣的情景喜剧。此外，多模式人工智能的创新使团队能够生成多种媒体类型的内容，包括文本、图形和视频。这是Dall-E等工具的基础，这些工具可以根据文本描述自动创建图像或根据图像生成文本标题。

尽管取得了这些突破，但我们仍处于使用生成式人工智能创建可读文本和逼真风格化图形的早期阶段。早期的实施存在准确性和偏见问题，并且容易产生幻觉并给出奇怪的答案。不过，迄今为止的进展表明，这种生成式人工智能的固有能力可以从根本上改变企业技术，改变企业的运营方式。展望未来，这项技术可以帮助编写代码、设计新药、开发产品、重新设计业务流程和转变供应链。

• 生成式AI模型

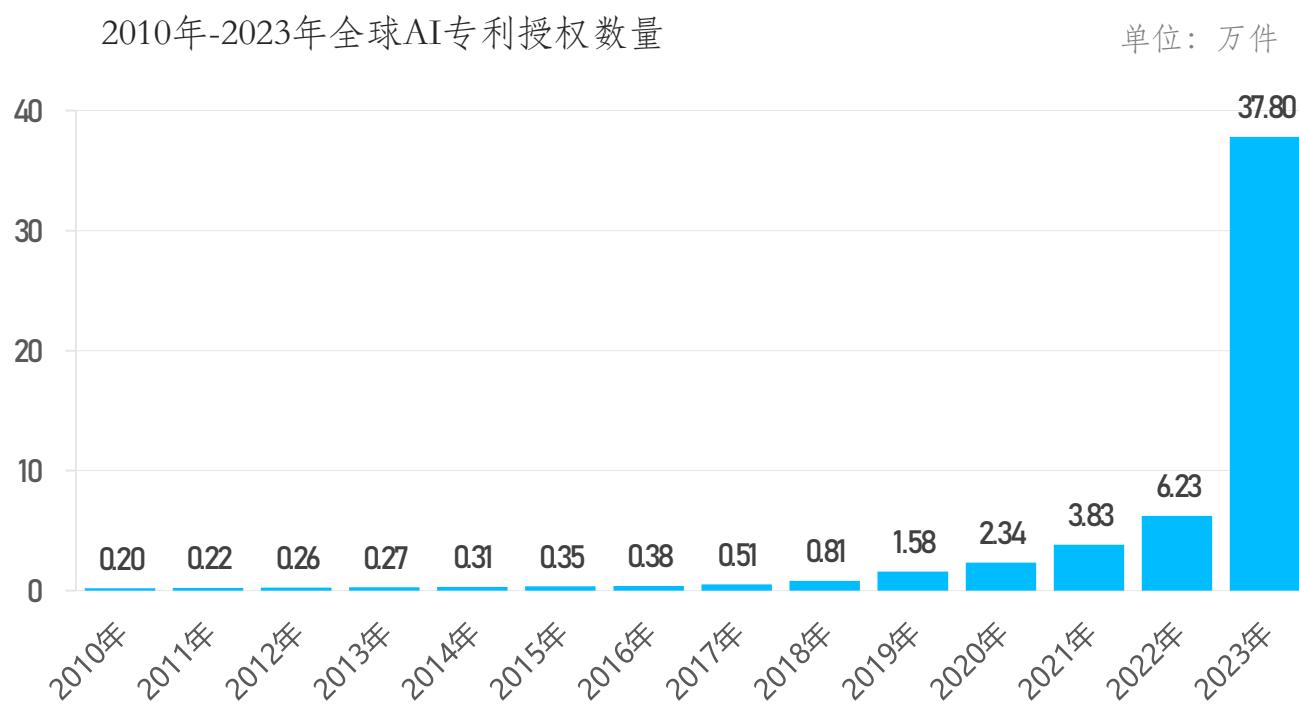
生成式人工智能模型结合了各种人工智能算法来表示和处理内容。一旦开发人员确定了表示世界的方式，他们就会应用特定的神经网络来响应查询或提示生成新内容。诸如GAN和变分自动编码器(VAE)（带有解码器和编码器的神经网络）之类的技术适合生成逼真的人脸、用于AI训练的合成数据，甚至是特定人类的复制品。谷歌的Transformer双向编码器表示(BERT)、OpenAI的GPT和谷歌AlphaFold等Transformer的最新进展也使得神经网络不仅可以编码语言、图像和视频。

02

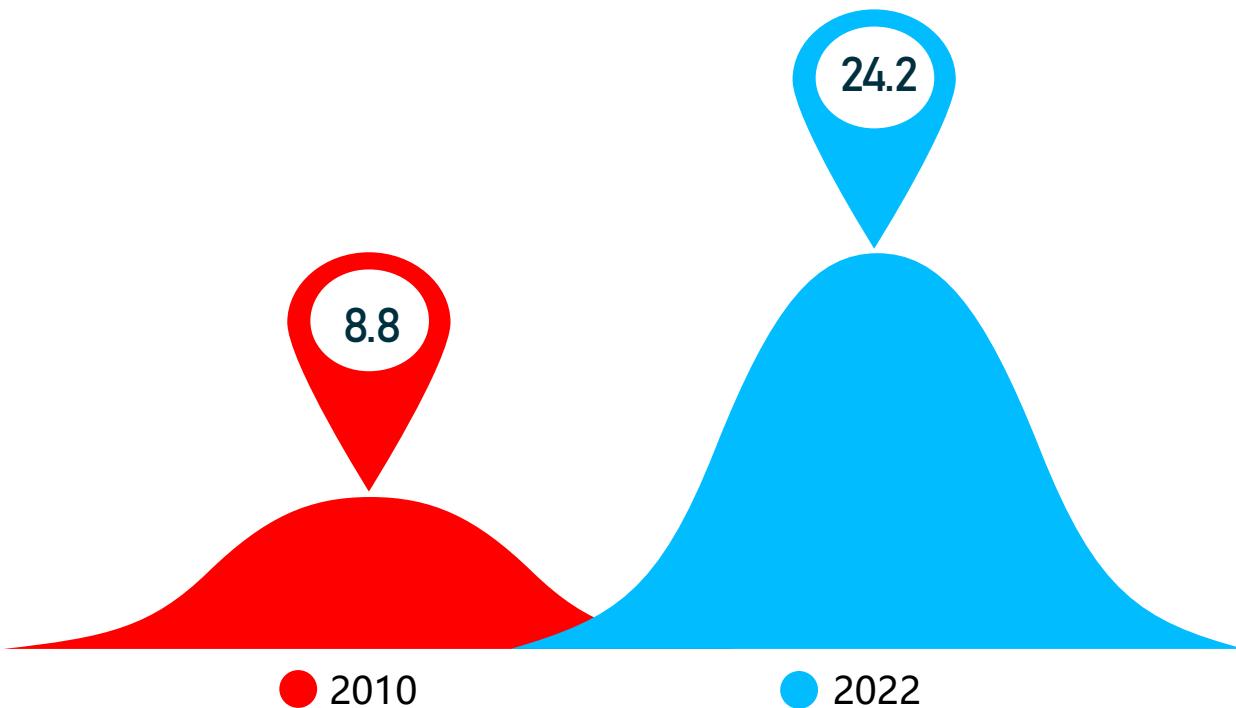
研究与开发

全球AI研发投入大幅飙升，2023年全球AI专利授权量大37.8万件，同比增长507.1%

2023年全球AI专利授权数量达37.8万件



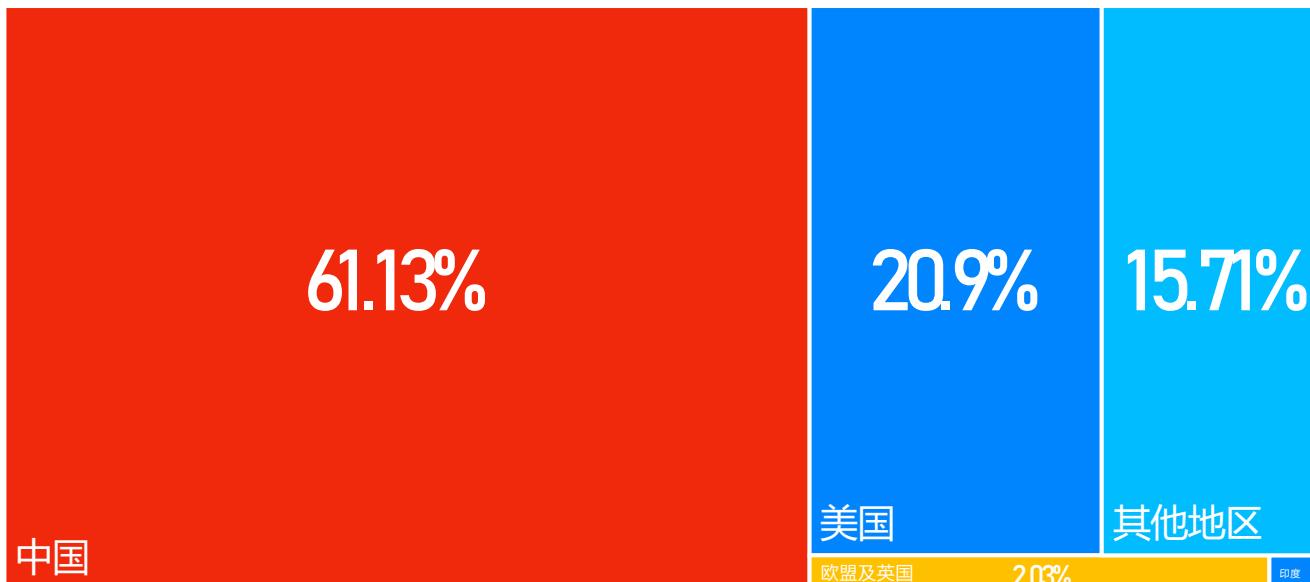
2010年-2022年全球人工智能出版物数量快速增长



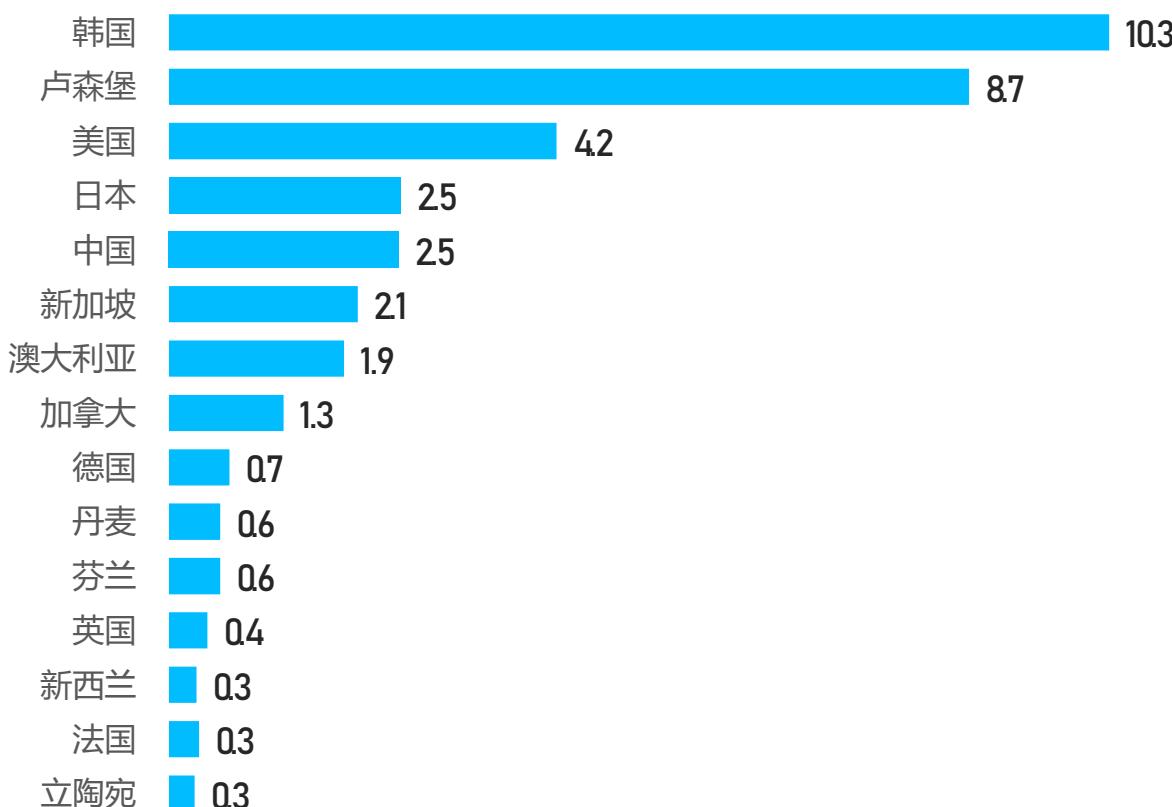
中国在全球AI专利数量处于领先地位，2022年中国AI专利数量占全球的61.13%

2022年全球AI专利国别分布分析

■ 中国 ■ 美国 ■ 欧盟及英国 ■ 印度 ■ 其他地区



2022年各国每10万人获得的AI专利数量 (件)

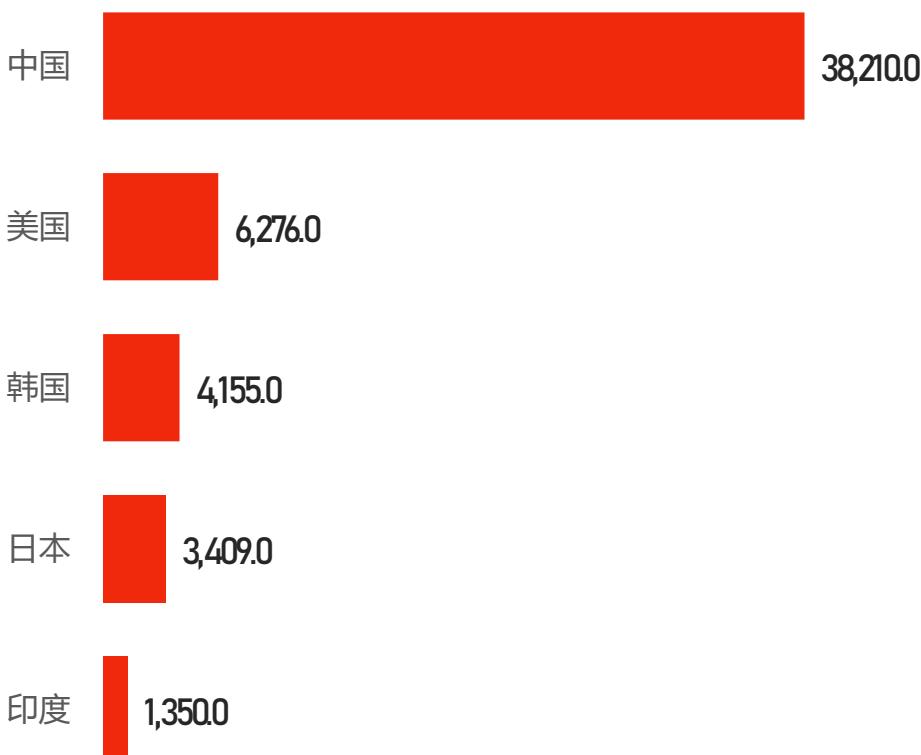


中国申请的生成式人工智能专利数量同样处于全球领先地位

世界知识产权组织(WIPO)的最新报告显示，中国发明家申请的生成式人工智能(GenAI)专利数量最多，远远超过排名前五的美国、韩国、日本和印度。《WIPO生成式人工智能专利态势报告》记录了截至2023年的十年间54,000项GenAI专利，其中仅在去年一年就有超过25%的发明诞生。

GenAI允许用户创建包括文本、图像、音乐和计算机代码在内的内容，为一系列工业和消费产品提供支持，包括ChatGPT、GoogleGemini或百度的ERNIE等聊天机器人。2014年至2023年期间，中国将有超过38,000项GenAI发明，是排名第二的美国的六倍。印度是GenAI发明的第五大来源地，在前五大国家中，其年均增长率最高，达到56%。报告显示，GenAI已经遍及生命科学、制造业、交通运输、安全和电信等行业。

2014年-2023年各国生成式人工智能专利数量 (件)



邓达仁
WIPO总干事

- “GenAI已经成为一项改变游戏规则的技术，有可能改变我们的工作、生活和娱乐方式。通过分析专利趋势和数据，WIPO希望让每个人都更好地了解这项快速发展的技术的发展方向和发展方向。这可以帮助政策制定者塑造GenAI的发展，造福我们共同利益，并确保我们继续将人类置于创新和创意生态系统的中心。我们相信，这份报告将使创新者、研究人员和其他人能够驾驭快速发展的生成式人工智能格局及其对世界的影响，”

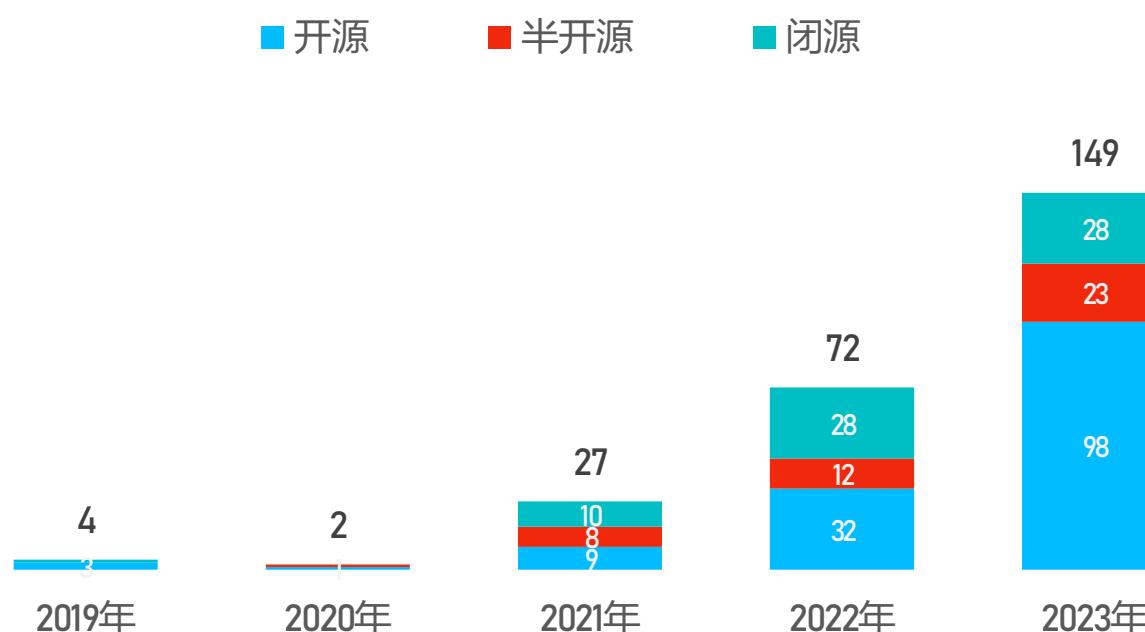
中国生成式AI参与主体对专利申请尤为重视，全球前十大生成式AI专利拥有平台，中国占据六席

2014年-2023年获得GenAI专利数量主要公司/机构排名

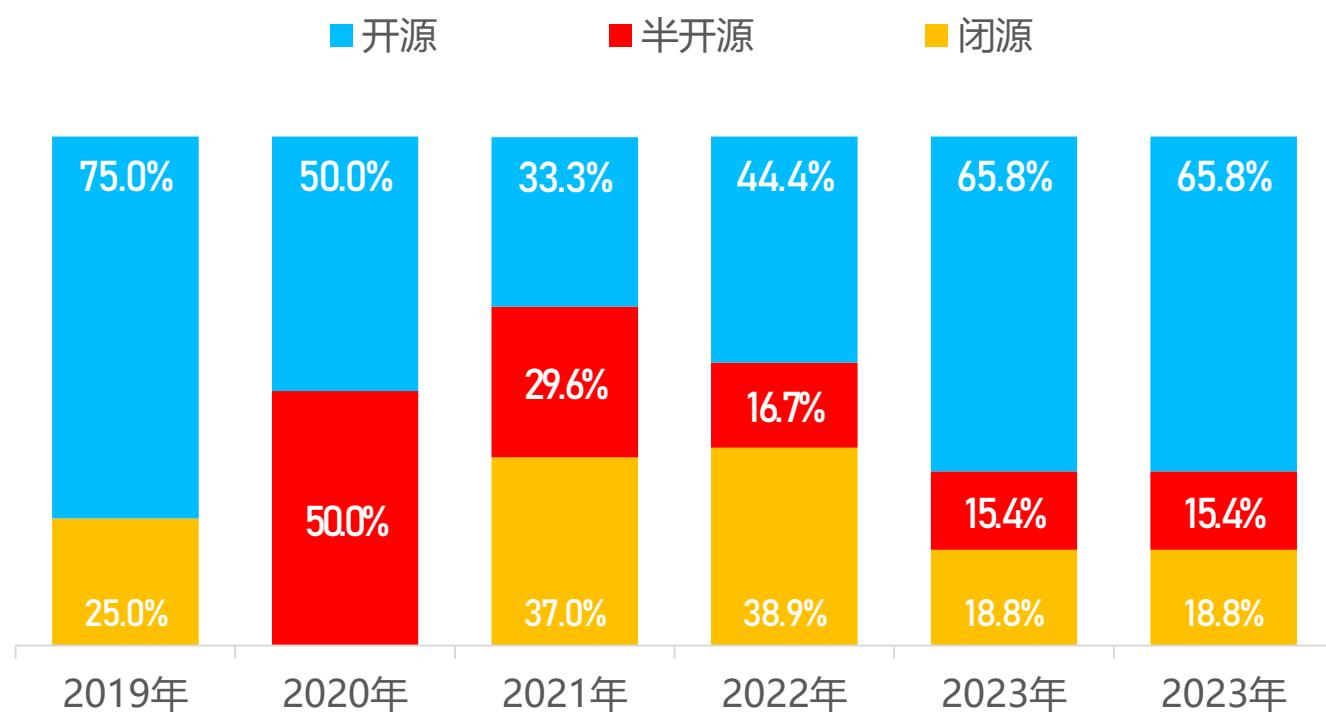


2021年开始，全球基础模型数量大幅飙升，开源模型占据主导地位，半开源及闭源模型也大量涌现

2019年-2023年全球不同可访问类型基础模型数量

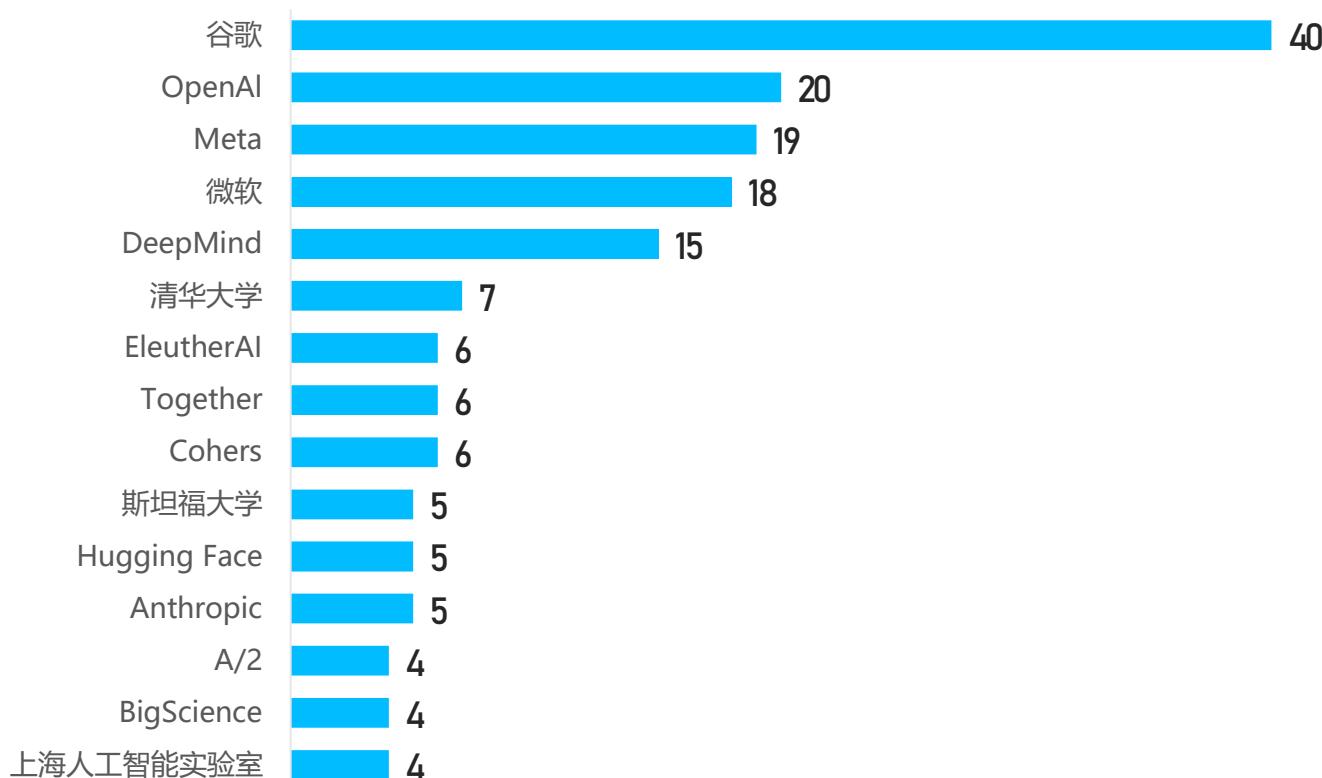


2019年-2023年全球不同可访问类型基础模型数量占比

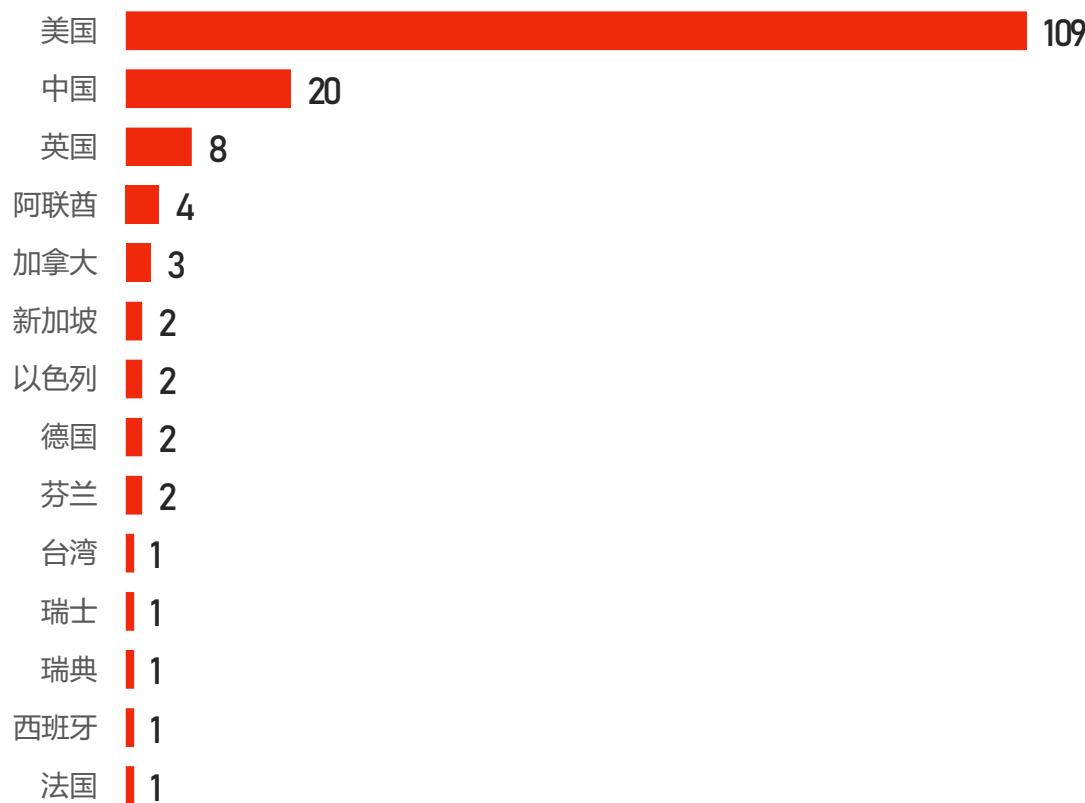


美国企业在人工智能领域一骑绝尘，几乎贡献了全部高价值基础模型

2019年-2023年各组织基础人工智能模型发布数量 (个)

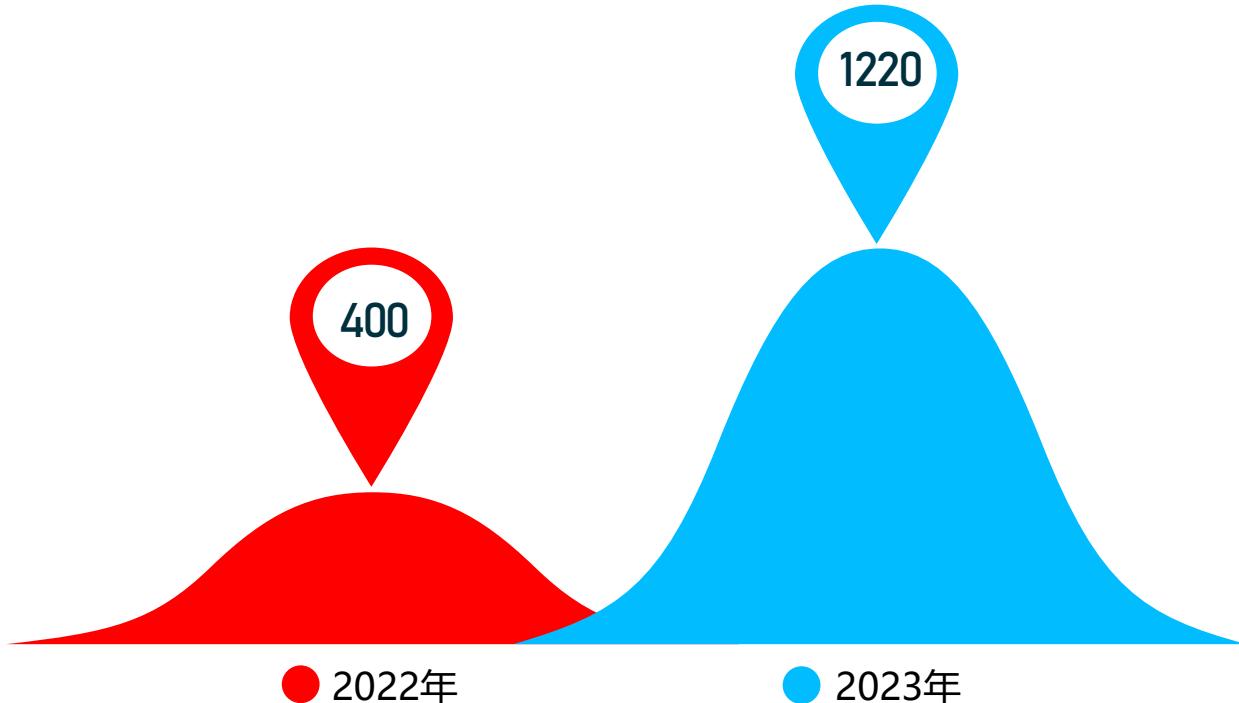


2023年各组织基础人工智能模型发布数量 (个)



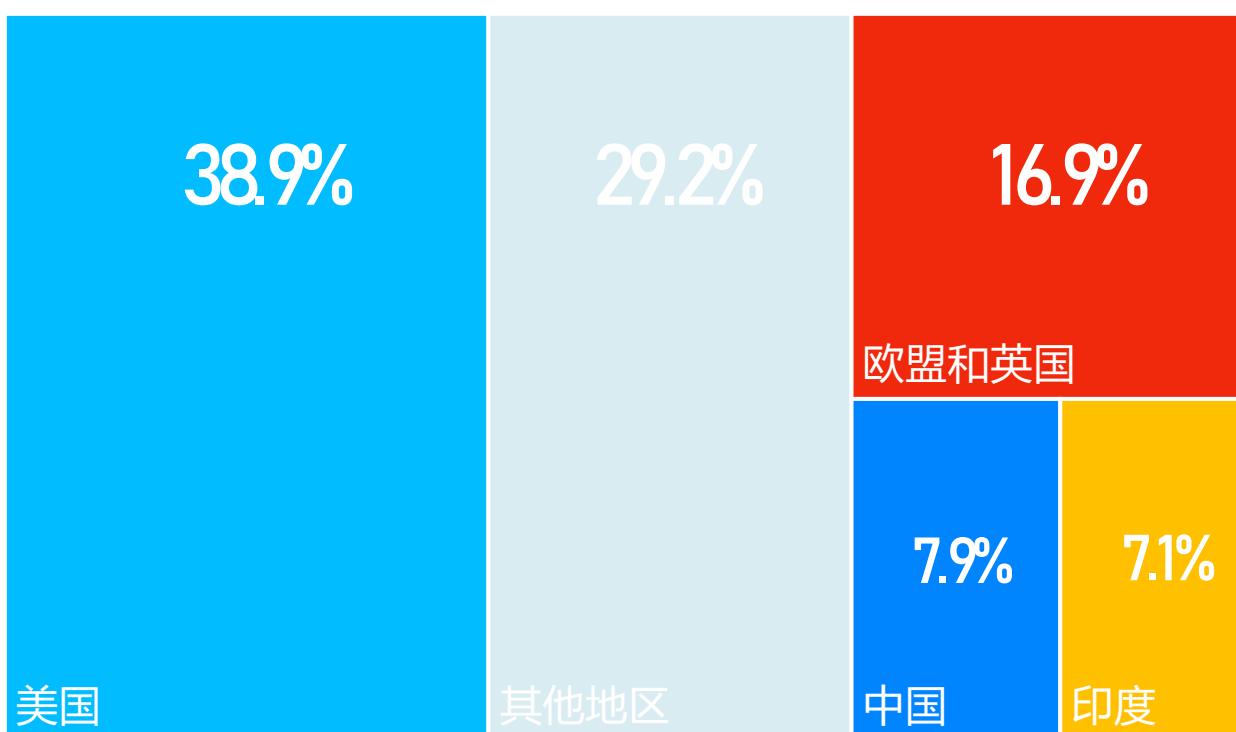
人工智能开源项目数量大幅飙升，2023年GitHub人工智能项目星标达1220万，同比增长205%

2011年-2023年GitHub人工智能项目星标数量（万）



2023年全球各主要地区GitHub人工智能项目星标数量分布

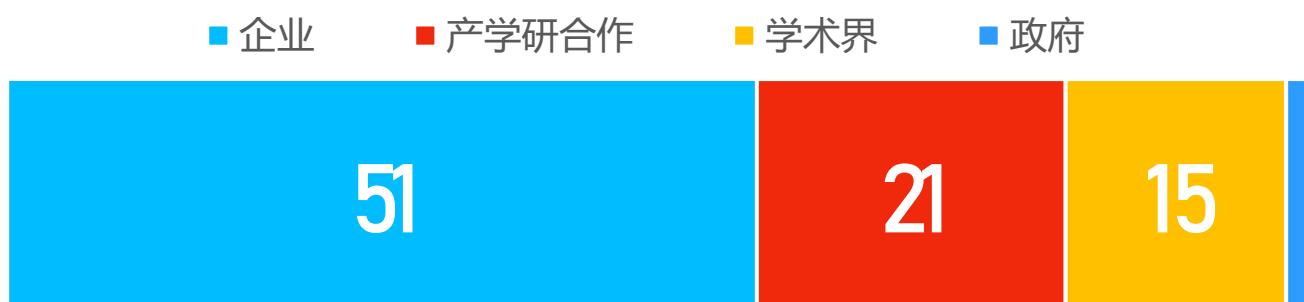
■ 美国 ■ 欧盟和英国 ■ 中国 ■ 印度 ■ 其他地区



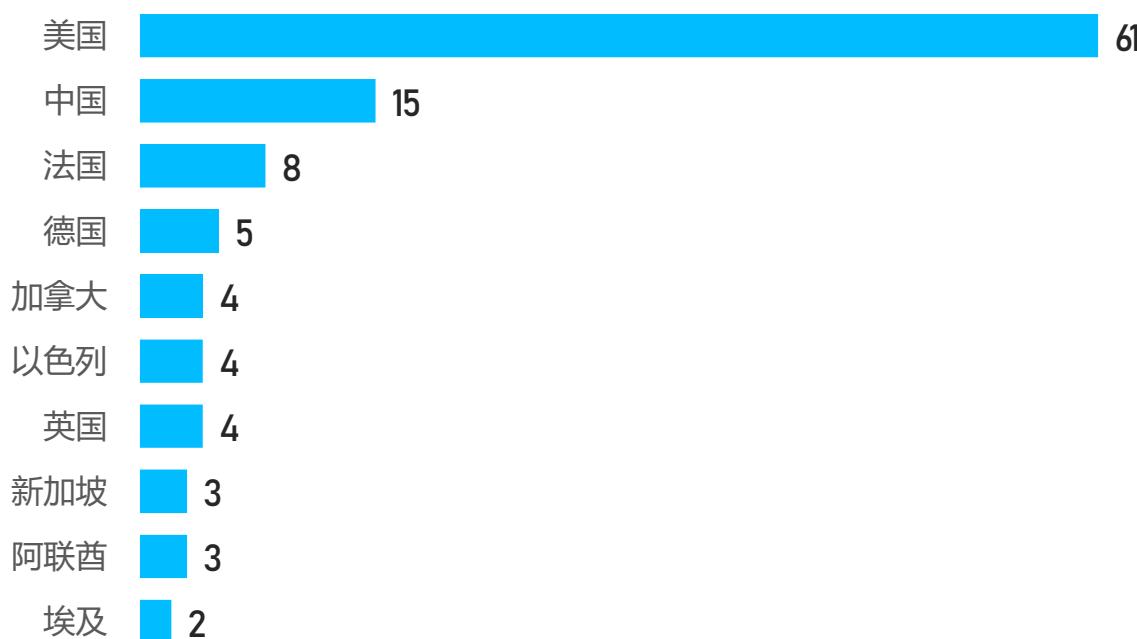
企业在全球AI研究领域处于绝对领导地位，科研机构及高校成果产出相对低效

直到2014年，学术界在人工智能模型的发布方面一直处于领先地位。2014年之后企业就占据了主导地位。2023年，全球AI领域产生了51个值得注意的人工智能模型，而学术界只有15个。值得注意的是，2023年，有21个值得注意的模型来自产学研合作。现在创建尖端的人工智能大模型需要大量的数据、算力和资金投入，而这些是学术界所不具备的。这种由企业主导领先人工智能模型研发的趋势和去年相比基本保持不变，并且还将持续下去。

2023年全球值得关注人工智能大模型发布主体分布

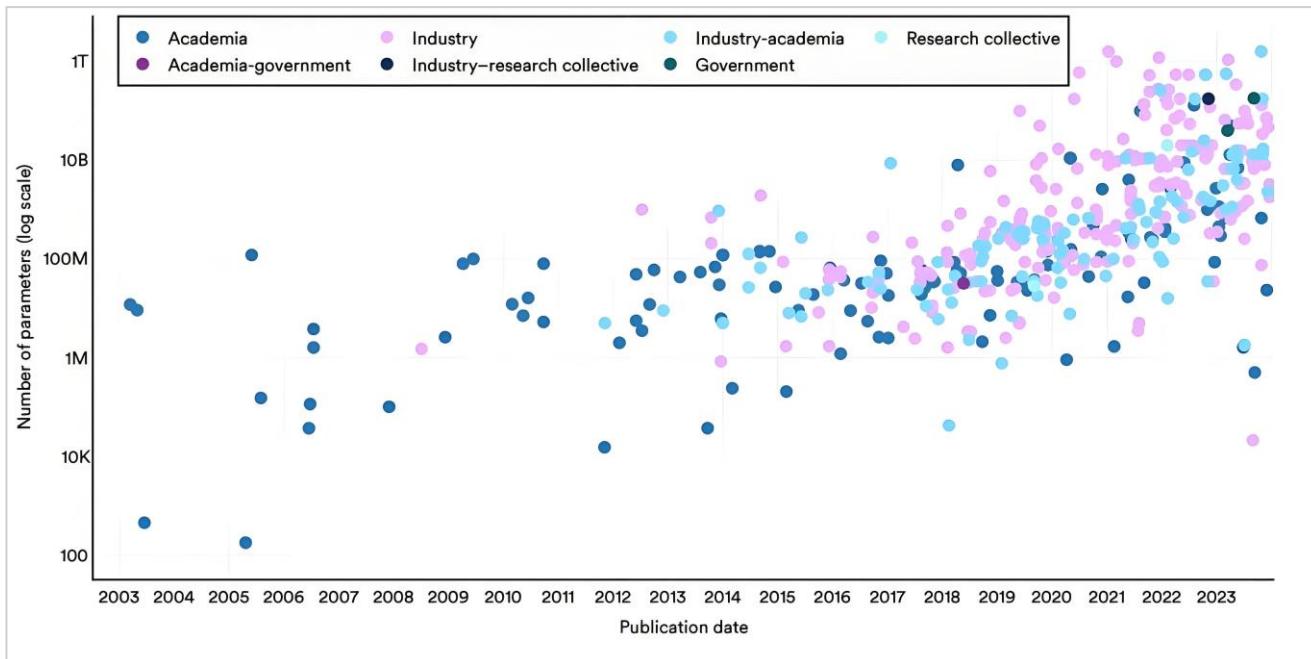


2023年值得关注人工智能大模型发布国家分布

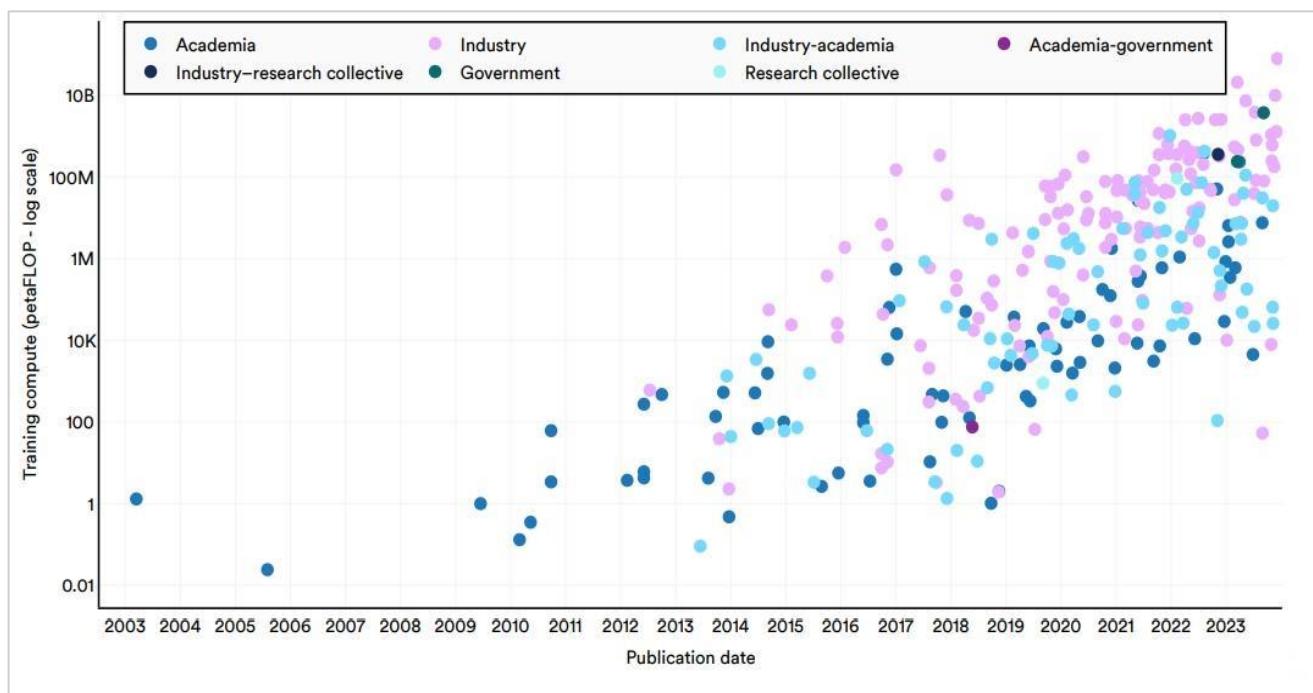


大模型参数及训练算力需求呈现指数级增长，大模型的训练门槛在急速增高

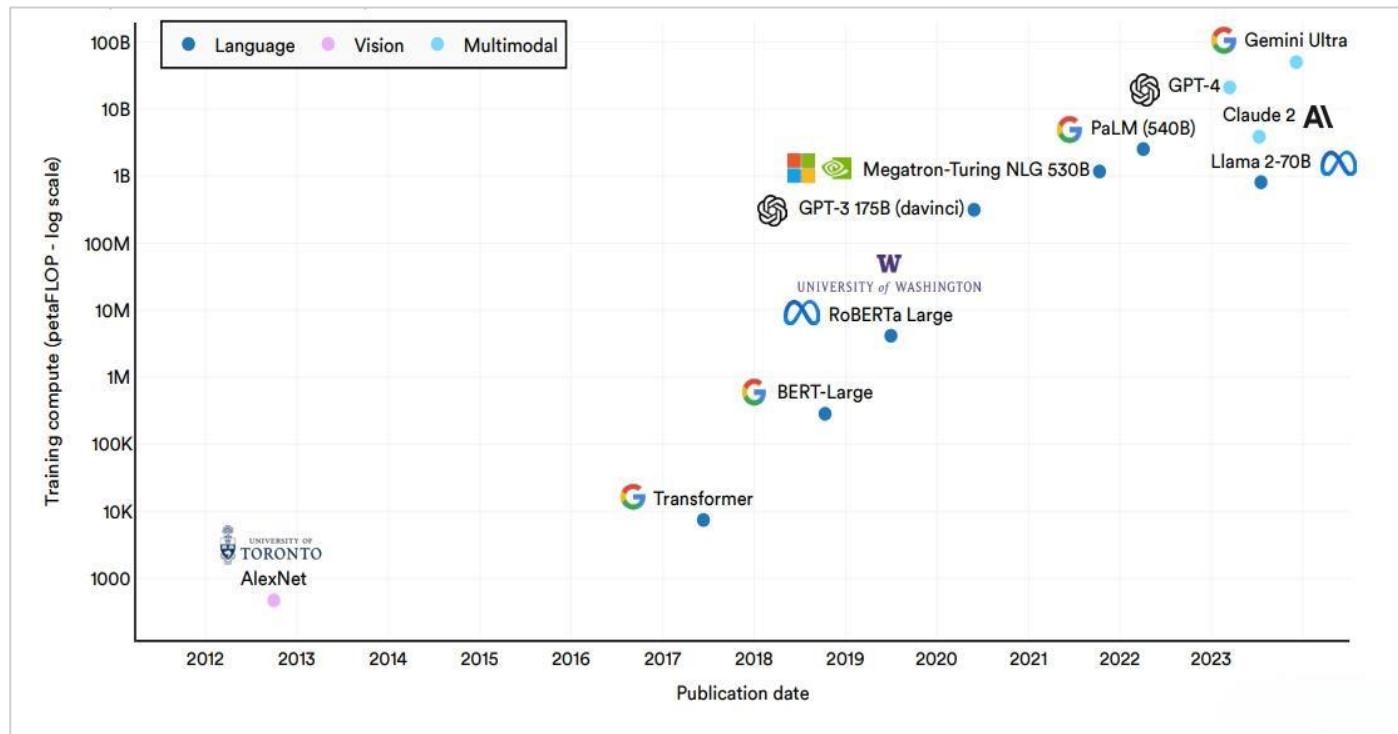
2003年-2023年主要人工智能大模型参数数量



2003年-2023年主要人工智能大模型训练算力需求



2012年-2023年各领域著名人工智能大模型训练计算量



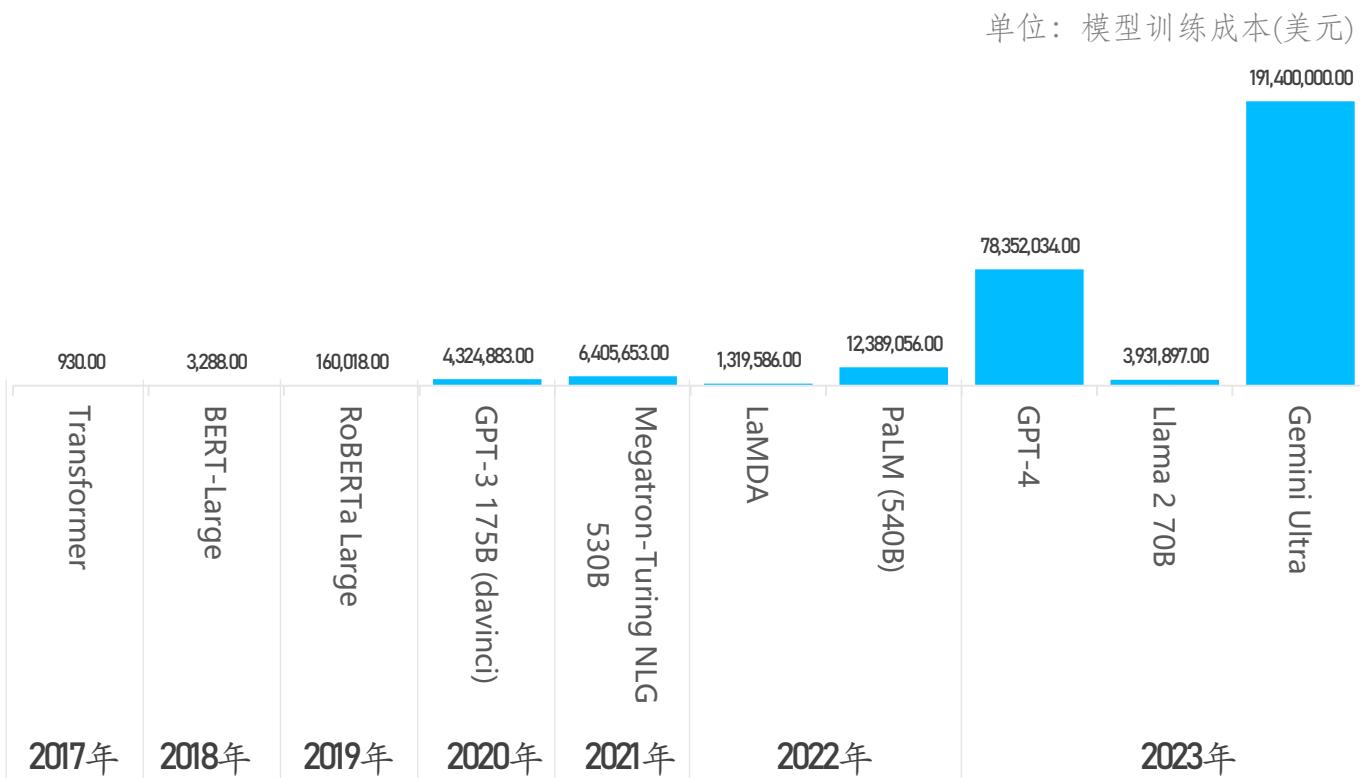
自2012年以来著名机器学习模型的训练算力中。例如，AlexNet是推广使用GPU改进AI模型的现有标准做法的论文之一，它需要估计470petaFLOP进行训练。

最初的Transformer于2017年发布，

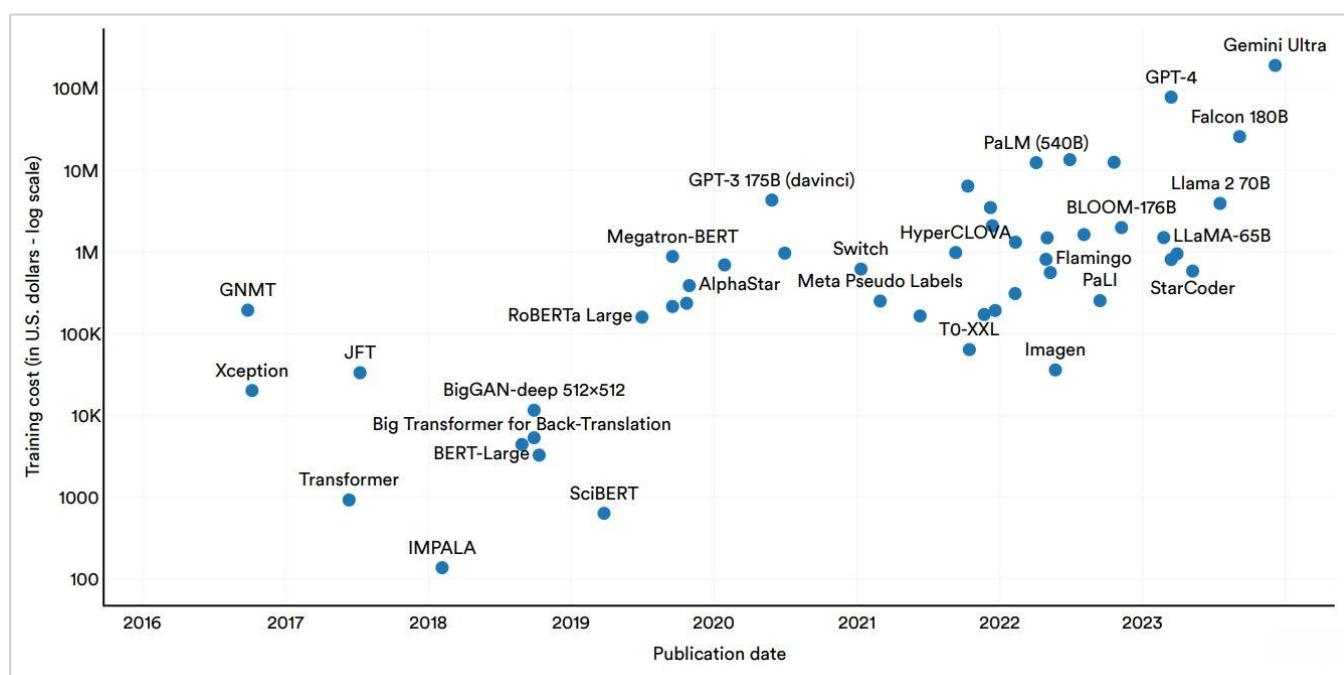


模型训练成本指数级飙升，进入人工智能领域的门槛越来越高

2017年-2023年部分模型训练成本预估



2016-23年部分AI模型的训练成本估算



风投资本大举押注人工智能领域，美国是全球AI创投的绝对中心，创业项目质量及获得投资均大幅领先

2024年海外市场AI相关企业重点融资案例盘点

| 时间 | 企业名称 | 细分赛道 | 融资阶段 | 融资金额 | 投资方 | 国家 |
|----------|---------------|---------------------|--------|---------|--|----|
| 2024年12月 | Luma AI | 3D模型研发商 | C轮 | 9000万美元 | 亚马逊Amazon AMD Andreessen Horowitz-a16z Amplify Partners 经纬创投 Factorial Funds LDV Capital 韩华集团 | 美国 |
| 2024年12月 | Nebius | 人工智能基础设施服务商 | 战略投资 | 7亿美元 | Accel Partners 英伟达NVIDIA Orbis Investment | 荷兰 |
| 2024年11月 | /dev/agents | AI Agent通用操作系统研发商 | 种子轮 | 5600万美元 | CapitalG谷歌资本 Index Ventures | 美国 |
| 2024年11月 | Argil | 人工智能预编辑平台运营商 | 天使轮 | 490万欧元 | Y Combinator EQT Ventures | 法国 |
| 2024年11月 | Play.AI | 语音人工智能平台提供商 | 天使轮 | 2100万美元 | Pioneer Fund Soma Capital Y Combinator Race Capital 500 Global Kindred Ventures TRAC | 美国 |
| 2024年11月 | Talus Network | 基于MoveVM的链上人工智能网络平台 | Pre-A轮 | 600万美元 | [领投] Polychain Capital | 美国 |
| 2024年11月 | OpenAI | 人工智能研究公司 | 战略投资 | 15亿美元 | 软银愿景基金 SoftBank 软银 | 美国 |
| 2024年11月 | Anthropic | 人工智能大模型 Claude开发商 | 战略投资 | 40亿美元 | 亚马逊Amazon | 美国 |
| 2024年11月 | xAI | 马斯克发起设立的人工智能公司 | 战略投资 | 50亿美元 | 卡塔尔投资局QIA Valor Equity Partners Andreessen Horowitz-a16z Sequoia Capital 红杉 | 美国 |
| 2024年11月 | Rox | AI Agent解决方案提供商 | A轮 | 未透露 | [领投] General Catalyst Partners | 美国 |
| 2024年11月 | Rox | AI Agent解决方案提供商 | 种子轮 | 未透露 | [领投] Sequoia Capital 红杉 Google Ventures(GV) | 美国 |

2024年海外市场AI相关企业重点融资案例盘点

| 时间 | 企业名称 | 细分赛道 | 融资阶段 | 融资金额 | 投资方 | 国家 |
|----------|-----------------------|-----------------|------|---------|--|----|
| 2024年11月 | Gendo | 人工智能驱动建筑平台 | 种子轮 | 430万英镑 | [领投] LEA Partners [领投] PT1 Concept Ventures Koro Capital | 英国 |
| 2024年11月 | SERENITY | 人工智能软件解决方案提供商 | A轮 | 550万美元 | Base10 Partners Allegion Ventures | 美国 |
| 2024年11月 | workflow | AI创意工作流平台 | 种子轮 | 300万美元 | [领投] Venrex 8VC Sequoia Capital 红杉 Index Ventures Octopus Ventures | 英国 |
| 2024年11月 | Bluespine | AI理赔成本降低解决方案提供商 | 种子轮 | 720万美元 | [领投] Team8 | 美国 |
| 2024年11月 | Ennoventure | 医药智能包装系统供应商 | A轮 | 890万美元 | Tanglin Venture Partners Fenice Investment Group | 美国 |
| 2024年11月 | Writer | 企业级AI写作协助工具 | C轮 | 2亿美元 | ICONIQ Growth Premji Invest Radical Ventures Salesforce Adobe IBM Workday Ventures Insight Partners 花旗Citi Ventures B Capital Group 先锋领航 | 美国 |
| 2024年11月 | Zap Surgical Systems | 美国放射外科手术设备研发商 | 战略投资 | 数千万美元 | 百洋医药 | 美国 |
| 2024年11月 | Perplexity AI | 智能对话式搜索引擎提供商 | C轮 | 5亿美元 | [领投] Institutional Venture Partners | 美国 |
| 2024年11月 | Physical Intelligence | 通用机器人系统的人工智能 | A轮 | 4亿美元 | Lux Capital Thrive Capital Sequoia Capital 红杉 Khosla Ventures OpenAI Jeff Bezos Bond Capital Redpoint Ventures 红点全球基金 | 美国 |
| 2024年11月 | Spot AI | 智能摄像头系统开发商 | C轮 | 3100万美元 | Marcy Venture Partners Bessemer Venture Partners Redpoint Ventures 红点全球基金 Cheyenne Ventures GSBackers 行健资本 Scale Ventures 高通Qualcomm Ventures | 美国 |
| 2024年11月 | Jugemu.ai | 人工智能DePIN项目 | 种子轮 | 100万美元 | 未透露 | 美国 |

2024年海外市场AI相关企业重点融资案例盘点

| 时间 | 企业名称 | 细分赛道 | 融资阶段 | 融资金额 | 投资方 | 国家 |
|----------|------------------|-------------------|------|---------|--|----|
| 2024年11月 | Interface.ai | 金融领域AI Agent平台提供商 | 种子轮 | 3000万美元 | [领投] Avataar Venture Partners | 美国 |
| 2024年11月 | iGent | 人工智能软件开发商 | 种子轮 | 630万英镑 | HV Capital Dhyan VC Twin Path Ventures XTX Ventures 10X Founders | 英国 |
| 2024年10月 | Coframe | AI网站前端优化工具研发商 | 种子轮 | 930万美元 | Hack VC Soma Capital Khosla Ventures HFO 三星电子 TI Platform Management | 美国 |
| 2024年10月 | Smashing | AI网络内容社区 | 种子轮 | 340万美元 | Blockchange Ventures True Ventures Power of N Offline Ventures | 美国 |
| 2024年10月 | Urbint | 事故预防人工智能平台 | 战略投资 | 3500万美元 | [领投] S2G Ventures | 美国 |
| 2024年10月 | Sierra | 人工智能客服服务商 | 战略投资 | 1.75亿美元 | [领投] Greenoaks Capital Management | 美国 |
| 2024年10月 | Paretos | 智能决策服务提供商 | A轮 | 850万欧元 | [领投] Acton Capital Etsy UVC Partners HomeToGo Cyberport LEA Partners Interface Capital | 德国 |
| 2024年10月 | Cascade AI | 员工援助AI解决方案提供商 | 种子轮 | 375万美元 | [领投] Gradient Ventures Success Venture Partners Myriad Venture Partners | 美国 |
| 2024年10月 | Waymo | 汽车自动驾驶技术研发商 | 战略投资 | 56亿美元 | Alphabet Tiger Global 老虎海外 淡马锡 Temasek | 美国 |
| 2024年10月 | Reality Defender | 人工智能生成媒体检测平台提供商 | A+轮 | 3300万美元 | IBM DCVC Bio 埃森哲 Accenture Illuminate Financial Booz Allen Hamilton The Jeffries Family Office | 美国 |
| 2024年10月 | CrewAI | 自动化AI代理服务商 | A轮 | 1800万美元 | Insight Partners BOLDstart Ventures | 美国 |
| 2024年10月 | Decagon | 生成式人工智能服务提供商 | B轮 | 6500万美元 | [领投] 贝恩资本 BainCapital Bond Capital Accel Partners ACME Ventures | 美国 |
| 2024年10月 | SynthBee | 新型计算智能技术研发商 | 种子轮 | 2000万美元 | [领投] Crosspoint Capital Partners | 美国 |
| 2024年10月 | Lightmatter | 芯片研发商 | D轮 | 4亿美元 | CapitalG 谷歌资本 Fidelity Management and Research Company T. Rowe Price | 美国 |

2024年海外市场AI相关企业重点融资案例盘点

| 时间 | 企业名称 | 细分赛道 | 融资阶段 | 融资金额 | 投资方 | 国家 |
|----------|-----------------|------------------|------|---------|--|----|
| 2024年10月 | Beyond Presence | 计算机视觉技术研发商 | 种子轮 | 310万美元 | [领投] HV Capital Meta DeepMind 10x Founders Zalando Alba VC | 美国 |
| 2024年10月 | Gladia | AI转录和音频智能硬件开发商 | A轮 | 1600万美元 | [领投] XAnge Private Equity Gaingels Soma Capital Athletico Ventures Illuminate Financial Mana Ventures | 法国 |
| 2024年10月 | Numeric | AI会计自动化服务商 | A轮 | 2800万美元 | Menlo Ventures Institutional Venture Partners Socii Capital | 美国 |
| 2024年10月 | Suki | 医用语音笔记助手研发商 | D轮 | 7000万美元 | [领投] Hedsophia Breyer Capital Flare Capital Partners March Capital Venrock Healthcare Capital Partners InHealth Ventures | 美国 |
| 2024年10月 | Open Gradient | 去中心化开源AI平台 | 种子轮 | 850万美元 | Coinbase Ventures Foresight Ventures SV Angel Near Illia Polosukhin Pragma | 美国 |
| 2024年10月 | Distributional | 人工智能安全服务商 | A轮 | 1900万美元 | [领投] Two Sigma Ventures | 美国 |
| 2024年10月 | EvenUp | 智能法务服务商 | D轮 | 1.35亿美元 | Capital Group Bessemer Venture Partners SignalFire Premji Invest Lightspeed Venture Partners 光速 全球 贝恩资本 BainCapital | 美国 |
| 2024年10月 | Everday | AI技能管理服务商 | 种子轮 | 未透露 | Builders Studio | 荷兰 |
| 2024年10月 | Poolside | AI软件及代码开发解决方案提供商 | B轮 | 5亿美元 | [领投] 贝恩资本 BainCapital DST Global 英伟达 NVIDIA LG集团 eBay 汇丰创投 | 美国 |
| 2024年10月 | Voyage AI | RAG人工智能应用 | A轮 | 2000万美元 | [领投] CRV Snowflake Databricks Wing Venture Partners Pear Ventures Tectonic Capital | 美国 |

2024年海外市场AI相关企业重点融资案例盘点

| 时间 | 企业名称 | 细分赛道 | 融资阶段 | 融资金额 | 投资方 | 国家 |
|----------|------------|-----------------|------|---------|---|----|
| 2024年10月 | OpenAI | 人工智能研究公司 | 战略投资 | 66亿美元 | [领投] Thrive Capital 微软 Microsoft 英伟达 NVIDIA 软银愿景基金 SoftBank 软银 Khosla Ventures Altimeter Capital Fidelity Management and Research Company Tiger Global 老虎海外 MGX | 美国 |
| 2024年9月 | Prepared | 紧急通信AI平台提供商 | B轮 | 2700万美元 | Andreessen Horowitz-a16z NewView Capital M13 First Round Capital | 美国 |
| 2024年9月 | Atomicwork | 人工智能员工支持平台提供商 | 天使轮 | 300万美元 | Z47-Matrix Partners India Storm Ventures Blume Ventures | 美国 |
| 2024年9月 | Pyka | 美国农业电动无人机研发生产商 | B轮 | 4000万美元 | Y Combinator Prelude Ventures Obvious Ventures Metaplanet Holdings Piva Capital | 美国 |
| 2024年9月 | Supermaven | AI编程服务提供商 | A轮 | 1200万美元 | [领投] Bessemer Venture Partners | 美国 |
| 2024年9月 | Harmonic | AI视频处理播放技术研发商 | A轮 | 7500万美元 | [领投] Sequoia Capital 红杉 Index Ventures GreatPoint Ventures DST Global Jasper Lau's Era Funds Jared Leto Nikesh Arora | 美国 |
| 2024年9月 | Scribenote | 利用人工智能为兽医生成医疗记录 | 种子轮 | 820万美元 | [领投] Andreessen Horowitz-a16z iNovia Capital Velocity Fund | 美国 |
| 2024年9月 | Mercor | 全自动AI人才评估平台提供商 | A轮 | 3000万美元 | [领投] Bill Gurley [领投] Victor Lazarte General Catalyst Partners Chris Re Larry Summers Adam D'Angelo Jack Dorsey Peter Thiel | 美国 |
| 2024年9月 | fal | 内容生成平台 | A轮 | 1400万美元 | [领投] Kindred Ventures | 美国 |
| 2024年9月 | Brightband | 利用人工智能进行天气预报 | A轮 | 1000万美元 | [领投] Prelude Ventures Bain Capital Ventures Future Back Ventures | 美国 |

2024年海外市场AI相关企业重点融资案例盘点

| 时间 | 企业名称 | 细分赛道 | 融资阶段 | 融资金额 | 投资方 | 国家 |
|---------|-------------------|---------------|------|---------|--|-----|
| 2024年9月 | OffDeal | AI并购服务提供商 | 种子轮 | 470万美元 | [领投] adical Ventures | 美国 |
| 2024年9月 | XP Health | 美国AI视觉福利平台 | B轮 | 3320万美元 | Valor Capital Group American Family Ventures Canvas Ventures Manchester Story | 美国 |
| 2024年9月 | World Labs | 空间智能技术服务商 | B轮 | 2.3亿美元 | [领投] Andreessen Horowitz-a16z [领投] NEA恩颐投资-New Enterprise Associates [领投] Radical Ventures AMD 英特尔投资Intel Capital 英伟达NVIDIA | 加拿大 |
| 2024年9月 | Glean | 工作助手和知识管理平台 | E轮 | 2.6亿美元 | [领投] DST Global [领投] Altimeter Capital Craft Ventures Sapphire Ventures 软银愿景基金 | 美国 |
| 2024年9月 | SmartCAT | 一个自动化翻译平台 | C轮 | 4300万美元 | [领投] Left Lane Capital | 美国 |
| 2024年9月 | SeekfilAI | 人工智能文档搜索工具开发商 | A轮 | 510万美元 | [领投] Kih Fund 汉十投资 百联资本 九禾资本 贵州国建集团 | 美国 |
| 2024年9月 | WOMBO Dream | AI画作制作软件 | 种子轮 | 900万美元 | 英伟达NVIDIA SBI投资(思佰益) CoreWeave Round13 Digital Assets Fund Web3.com Ventures | 加拿大 |
| 2024年9月 | Palm Technologies | AI金融科技服务商 | 种子轮 | 未透露 | Target Global SpeedInvest Liquid 2 Ventures GREENS Upfin | 英国 |
| 2024年9月 | Champion AI | AI驱动客户宣传平台 | 种子轮 | 330万美元 | High Alpha Flyover Capital Stage 2 Capital Bread and Butter Ventures | 美国 |
| 2024年9月 | SSI | 安全智能服务商 | 战略投资 | 10亿美元 | Andreessen Horowitz-a16z Sequoia Capital红杉 DST Global SV Angel AI Grant NFDG Ventures | 美国 |
| 2024年9月 | Dotyon.AI | 人工智能服务提供商 | 天使轮 | 320万美元 | 百联资本 贵州国建集团 | 美国 |
| 2024年9月 | Magic AI | 人工智能编码助手 | 战略投资 | 3.2亿美元 | Atlassian 埃里克·施密特 Capital G | 美国 |

2024年海外市场AI相关企业重点融资案例盘点

| 时间 | 企业名称 | 细分赛道 | 融资阶段 | 融资金额 | 投资方 | 国家 |
|---------|------------|------------------|--------|---------|--|-----|
| 2024年9月 | fal | 内容生成平台 | 种子轮 | 900万美元 | [领投] Andreessen Horowitz-a16z | 美国 |
| 2024年8月 | Codeium | 人工智能编码工具包提供商 | C轮 | 1.5亿美元 | Greonoaks Capital Management Kleiner Perkins Caufield & Byers(KPCB全球) General Catalyst Partners | 美国 |
| 2024年8月 | Pryzm | 人工智能工具开发商 | 种子轮 | 未透露 | Amplify.LA XYZ Venture Capital First In | 美国 |
| 2024年8月 | Opus Clip | 生成式AI视频编辑工具 | A轮 | 3000万美元 | Samsung NEXT DCM海外 Millennium New Horizons GTMFund | 美国 |
| 2024年8月 | Bland AI | 对话式人工智能平台提供商 | A轮 | 1600万美元 | Y Combinator Scale Ventures | 美国 |
| 2024年8月 | Viggle | AI动画生成商 | A轮 | 1900万美元 | [领投] Andreessen Horowitz-a16z Two Small Fish | 加拿大 |
| 2024年8月 | Opkey | 人工智能持续测试自动化平台提供商 | B轮 | 4700万美元 | PeakSpan Capital | 美国 |
| 2024年8月 | Pangeam | 工业人工智能平台提供商 | Pre-A轮 | 280万美元 | Neotribe Ventures | 美国 |
| 2024年8月 | ModelOp | 人工智能治理软件开发商 | B轮 | 1000万美元 | 贝雅资本 | 美国 |
| 2024年8月 | Cosine | 类人自主人工智能软件开发商 | 种子轮 | 250万美元 | [领投] Soma Capital [领投] Uphonest Capital 威诚资本 Lakestar Focal | 美国 |
| 2024年8月 | CodeRabbit | AI代码审查平台 | A轮 | 1600万美元 | CRV Flex Capital Engineering Capital | 美国 |
| 2024年8月 | EvenUp | 智能法务服务商 | C轮 | 3500万美元 | Lightspeed Venture Partners 光速全球 | 美国 |
| 2024年8月 | Ragie AI | 人工智能技术提供商, | 种子轮 | 550万美元 | Craft Ventures Valor Chapter One Ventures Saga VC | 美国 |
| 2024年8月 | Nepoe | 人工智能技术开发商 | 天使轮 | 900万美元 | [领投] Bitkraft Ventures OKX Ventures Mask Network Animoca Brands Big Brain Holdings Galaxy Interactive Hanow Fund Martin P Tin | 美国 |

2024年海外市场AI相关企业重点融资案例盘点

| 时间 | 企业名称 | 细分赛道 | 融资阶段 | 融资金额 | 投资方 | 国家 |
|---------|--------------------|---------------|--------|---------|---|-----|
| 2024年8月 | OW | 气味数字化研发商 | Pre-A轮 | 220万英镑 | [领投] Parkwalk Advisors Innovate UK | 英国 |
| 2024年8月 | Mechanical Orchard | 人工智能增强工具研发商 | B轮 | 5000万美元 | [领投] Google Ventures(GV) | 美国 |
| 2024年8月 | Datch | 工业语音AI技术开发商 | A+轮 | 1500万美元 | Blue Bear Capital Blackhorn Ventures SIG海纳亚洲 Third Prime | 美国 |
| 2024年8月 | Groq | AI芯片研发商 | 战略投资 | 6.4亿美元 | [领投] BlackRock 贝莱德 Samsung Catalyst Fund Cisco Investments | 美国 |
| 2024年8月 | Contextual AI | AI大模型方案解决提供商 | A轮 | 8000万美元 | Greycroft Partners Bain Capital Ventures 英伟达NVIDIA Lightspeed Venture Partners | 美国 |
| 2024年8月 | Protect AI | 网络安全服务商 | B轮 | 6000万美元 | [领投] Evolution Equity Partners 01 Advisors Salesforce Samsung Electronics Co. | 美国 |
| 2024年7月 | Harvey | AI法律助手服务商 | C轮 | 1亿美元 | [领投] Google Ventures(GV) OpenAI Kleiner Perkins Caufield & Byers(KPCB全球) Sequoia Capital红杉 | 美国 |
| 2024年7月 | Uplimit | 人工智能学习平台 | A轮 | 1100万美元 | [领投] Salesforce Workday Ventures Translink Capital Cowboy Ventures Greylock Partners | 美国 |
| 2024年7月 | TES AI | AI算力协议提供商 | A轮 | 500万美元 | [领投] X FORCE Foundation CMCC Global Group Newman Capital OIG Investment | 美国 |
| 2024年7月 | Cohere | 自然语言处理平台 | D轮 | 5亿美元 | [领投] PSP Investments Cisco思科 AMD EDC 英伟达NVIDIA Salesforce Fujitsu富士通 | 加拿大 |
| 2024年7月 | World Labs | 空间智能技术服务 | A轮 | 1亿美元 | [领投] NEA恩颐投资-New Enterprise Associates Andreessen Horowitz-a16z Radical Ventures | 加拿大 |
| 2024年7月 | MP | 药物制造图像处理平台运营商 | 种子轮 | 未透露 | Silverton Partners Alumni Ventures Group | 美国 |
| 2024年7月 | Anysphere | 人工智能工具和助手开发商 | 战略投资 | 未透露 | Andreessen Horowitz-a16z | 美国 |
| 2024年7月 | Wanderboat AI | AI驱动的智能旅行规划平台 | 种子轮 | 数百万美元 | Sequoia Capital红杉 | 美国 |

2024年海外市场AI相关企业重点融资案例盘点

| 时间 | 企业名称 | 细分赛道 | 融资阶段 | 融资金额 | 投资方 | 国家 |
|---------|-------------------|-------------------|------|---------|---|----|
| 2024年7月 | Fireworks AI | AI实验和生产平台开发商 | B轮 | 5200万美元 | Sequoia Capital 红杉 英伟达 NVIDIA AMD Benchmark Capital MongoDB Databricks | 美国 |
| 2024年7月 | Captions | AI创意工作室 | C轮 | 6000万美元 | [领投] Index Ventures HubSpot Adobe Andreessen Horowitz-a16z Sequoia Capital 红杉 | 美国 |
| 2024年7月 | Accend | 人工智能解决方案提供商 | 种子轮 | 320万美元 | [领投] Adverb Ventures 645 Ventures Y Combinator Stripe | 美国 |
| 2024年7月 | Thrive AI Health | 线上“AI健康教练” | 种子轮 | 数百万美元 | OpenAI Thrive Global Alice L. Walton 基金会 | 美国 |
| 2024年7月 | SeekfilAI | 人工智能文档搜索工具开发商 | 天使轮 | 500万美元 | [领投] Hanow Fund ParaFi Capital Moonrock Capital Alpha Intelligence Capital | 美国 |
| 2024年7月 | Redactive | 人工智能增强应用程序开发商 | 种子轮 | 1150万美元 | Felicis Ventures Blackbird Ventures Zapier Atlassian Ventures | 美国 |
| 2024年7月 | Sentient | 开源AI平台提供商 | 种子轮 | 8500万美元 | [领投] Founders Fund [领投] Framework Ventures [领投] Pantera Capital HashKey Capital | 美国 |
| 2024年7月 | Gendo | 人工智能驱动建筑平台 | 种子轮 | 85万英镑 | Ascension Ventures Baobab Network Concept Ventures | 英国 |
| 2024年7月 | Axelera AI | 人工智能半导体制造商 | B轮 | 6800万美元 | Samsung Catalyst Fund Verve Ventures | 荷兰 |
| 2024年7月 | Dust | AI技术服务商 | 战略投资 | 1600万美元 | [领投] Sequoia Capital 红杉 | 法国 |
| 2024年7月 | LeyLine | AI赋能人机混合智能生态系统开发商 | 种子轮 | 数百万美元 | [领投] 春华创投 Taihill Venture | 美国 |
| 2024年6月 | Dotyon.AI | 人工智能服务提供商 | 种子轮 | 590万美元 | ParaFi Capital Moonrock Capital ARCA | 美国 |
| 2024年6月 | EvolutionaryScale | 生物学前沿人工智能模型开发商 | 种子轮 | 1.42亿美元 | [领投] Lux Capital [领投] Daniel Gross [领投] Nat Friedman AWS NVIDIA 英伟达 | 美国 |
| 2024年6月 | Groq | AI芯片研发商 | 战略投资 | 未透露 | [领投] BlackRock 贝莱德/黑岩 | 美国 |

2024年海外市场AI相关企业重点融资案例盘点

| 时间 | 企业名称 | 细分赛道 | 融资阶段 | 融资金额 | 投资方 | 国家 |
|---------|---------------------|-------------------|------|---------|---|-----|
| 2024年6月 | Stability AI | 元宇宙及数字媒体工具开发商 | A轮 | 1亿美元 | Greycroft Partners Coatue Management | 英国 |
| 2024年6月 | Iambic Therapeutics | 人工智能医疗平台开发商 | B+轮 | 5000万美元 | [领投] Exor Ventures [领投] Mubadala Capital 卡塔尔投资局 | 美国 |
| 2024年6月 | Groq | AI芯片研发商 | 战略投资 | 未透露 | Opus Capital Seven Rivers Capital | 美国 |
| 2024年6月 | Decagon | 生成式人工智能服务提供商 | A轮 | 3500万美元 | Airtable Klaviyo Lattice Okta Ventures | 美国 |
| 2024年6月 | CuspAI | 人工智能材料搜索引擎开发商 | 种子轮 | 3000万美元 | [领投] Hoxton Ventures Basis Set Ventures Lightspeed Venture Partners 光速全球 LocalGlobe Northzone Ventures FJ Labs Zero Prime Ventures | 美国 |
| 2024年6月 | Genspark | 人工智能搜索服务提供商 | 种子轮 | 6000万美元 | [领投] 蓝驰创投 真格基金 | 美国 |
| 2024年6月 | Private AI | AI语言识别工具开发商 | 战略投资 | 未透露 | DAO Maker Meezan Ventures Snova Capital Samara Asset Group | 加拿大 |
| 2024年6月 | Veeton | 生成式AI视觉创作服务商 | 种子轮 | 200万欧元 | [领投] Founders Future [领投] Armilar | 法国 |
| 2024年6月 | Apparate | Proteus创新AI视频生成模型 | 天使轮 | 数百万美元 | 真格基金 | 美国 |
| 2024年6月 | Mistral AI | 人工智能大模型服务商 | B轮 | 6亿欧元 | [领投] General Catalyst Partners BNP Paribas 英伟达NVIDIA Cisco思科 Samsung Ventures三星 Salesforce ServiceNow IBM DST Global | 法国 |
| 2024年6月 | Anterior | 医疗人工智能服务商 | A轮 | 2000万美元 | [领投] NEA恩颐投资 | 美国 |
| 2024年6月 | Contextual AI | AI大模型方案解决提供商 | 种子轮 | 2000万美元 | [领投] 贝恩资本BainCapital Lightspeed Venture Partners Greycroft Partners SVA | 美国 |
| 2024年6月 | Thoughtly | AI语音代理服务提供商 | 种子轮 | 300万美元 | Expansion Venture Capital Greycroft Partners Afore Capital | 美国 |
| 2024年6月 | AirMDR | AI自主管理检测和响应服务提供商 | 种子轮 | 500万美元 | Storm Ventures Foundation Capital | 美国 |

2024年海外市场AI相关企业重点融资案例盘点

| 时间 | 企业名称 | 细分赛道 | 融资阶段 | 融资金额 | 投资方 | 国家 |
|---------|---------------|----------------|------|---------|---|----|
| 2024年6月 | Pika | 视频生成AI公司 | B轮 | 8000万美元 | [领投] Spark Capital Greycroft Partners | 美国 |
| 2024年6月 | Tobiko | 扩展数据基础设施 | A轮 | 1730万美元 | [领投] Theory Ventures [领投] Unusual Ventures | 美国 |
| 2024年6月 | Viaduct | 人工智能技术开发服务商 | B轮 | 1000万美元 | Innovation Endeavors Stellantis | 美国 |
| 2024年6月 | Perplexity AI | 智能对话式搜索引擎提供商 | B+轮 | 2.5亿美元 | [领投] Bessemer Venture Partners Associates Databricks | 美国 |
| 2024年5月 | Maven AGI | 生成式人工智能平台 | A轮 | 2000万美元 | Lux Capital M13 The E14 Fund | 美国 |
| 2024年5月 | Sagetap | 人工智能技术解决方案提供商 | 种子轮 | 680万美元 | Emergent Ventures Uncorrelated Ventures NFX | 美国 |
| 2024年5月 | ThinkLabs | 人工智能开发和部署服务商 | 种子轮 | 500万美元 | Powerhouse Ventures Blackhorn Ventures | 美国 |
| 2024年5月 | xAI | 马斯克发起设立的人工智能公司 | B轮 | 60亿美元 | Valor Equity Partners VY Capital Andreessen Horowitz-a16z Sequoia Capital红杉 | 美国 |
| 2024年5月 | DeepL | 人工智能翻译软件和写作助手 | C轮 | 3亿美元 | [领投] Index Ventures ICONIQ Growth Bessemer Venture Partners Teachers' Venture Growth | 德国 |
| 2024年5月 | Scale AI | AI数据平台 | F轮 | 10亿美元 | [领投] Accel Partners Meta 亚马逊Amazon AMD 英特尔投资Intel Capital Y Combinator Founders Fund Coatue Management Thrive capital Spark Capital 英伟达NVIDIA 老虎Tiger Global 海外Greenoaks Capital Management | 美国 |
| 2024年5月 | H | 人工智能服务提供商 | 种子轮 | 2.2亿美元 | FirstMark Capital Eurazeo Elaia Partners Creandum | 法国 |
| 2024年5月 | Artisan AI | 人工智能解决方案提供商 | 种子轮 | 730万美元 | [领投] Oliver Jung | 美国 |
| 2024年5月 | Voxel51 | 人工智能视觉服务商 | B轮 | 3000万美元 | [领投] Shasta Ventures [领投] Drive Capital [领投] Bessemer Venture Partners | 美国 |

2024年海外市场AI相关企业重点融资案例盘点

| 时间 | 企业名称 | 细分赛道 | 融资阶段 | 融资金额 | 投资方 | 国家 |
|---------|----------------|----------------|------|---------|---|-----|
| 2024年5月 | CoreWeave | GPU云服务算力厂商 | 战略投资 | 75亿美元 | [领投] 黑石集团 Blackstone [领投] Magnetar Capital 凯雷投资 BlackRock 贝莱德/黑岩 | 美国 |
| 2024年5月 | Poly AI | 对话式人工智能语音助手 | C轮 | 5000万美元 | Georgian Partners 英伟达NVIDIA Passion Capital | 英国 |
| 2024年5月 | Gamma App | AI自动生成PPT | A轮 | 1200万美元 | [领投] Accel Partners Fellows Fund | 美国 |
| 2024年5月 | weka.io | AI深度学习和技术问题解决商 | E轮 | 1.4亿美元 | 高通Qualcomm Atreides Management 英伟达NVIDIA Generation Investment Management MoreTech Ventures Lumir Ventures Key1 Capital | 美国 |
| 2024年5月 | SmarterDx | 临床人工智能服务商 | B轮 | 5000万美元 | FLOODGATE Fund Flare Capital Partners | 美国 |
| 2024年5月 | 0xGen | AI整合DeFi平台 | 战略投资 | 未透露 | X Labs Tensor Investment Corporation Phoenix Chain | 美国 |
| 2024年5月 | Subtle Medical | AI医疗成像解决方案提供商 | B+轮 | 近千万美元 | Fusion Fund BlueRun Ventures海外文周投资 嘉加资本 | 美国 |
| 2024年5月 | Datology AI | AI大模型训练服务商 | A轮 | 4600万美元 | [领投] Felicis Ventures Radical Ventures 亚马逊Amazon Amplify Partners 微软Microsoft | 美国 |
| 2024年5月 | Panax | 人工智能现金流管理平台提供商 | 战略投资 | 1500万美元 | TLV Partners Team8 | 美国 |
| 2024年5月 | Numeric | AI会计自动化服务商 | 天使轮 | 1000万美元 | Founders Fund Menlo Ventures 8VC Long Journey Ventures Friends & Family Capital | 美国 |
| 2024年5月 | Tekst | 人工智能平台 | 战略投资 | 70万欧元 | Entourage | 美国 |
| 2024年5月 | World Labs | 空间智能技术服务商 | 种子轮 | 千万级美元 | Andreessen Horowitz-a16z Radical Ventures | 加拿大 |
| 2024年5月 | Lamini | 构建企业级AI应用程序 | A轮 | 2500万美元 | [领投] Amplify Partners First Round Capital AMD | 美国 |
| 2024年5月 | CoreWeave | GPU云服务算力厂商 | C轮 | 11亿美元 | Coatue Management Altimeter Capital Fidelity Management and Research | 美国 |

2024年海外市场AI相关企业重点融资案例盘点

| 时间 | 企业名称 | 细分赛道 | 融资阶段 | 融资金额 | 投资方 | 国家 |
|---------|--------------|-----------------|------|---------|--|----|
| 2024年4月 | FlexAI | 法国人工智能计算服务商 | 种子轮 | 3000万美元 | [领投] Alpha Intelligence Capital [领投] Elaia Partners [领投] Heartcore Capital Bpifrance | 法国 |
| 2024年4月 | Augment Code | 人工智能编码辅助公司 | B轮 | 2.27亿美元 | Index Ventures Innovation Endeavors Lightspeed Venture Partners 光速全球 Meritech Capital Partners Sutter Hill Ventures | 美国 |
| 2024年4月 | Cognition | AI程序员技术应用开发商 | A轮 | 1.75亿美元 | [领投] Founders Fund Conviction Partners | 美国 |
| 2024年4月 | Parloa | 对话式人工智能平台开发商 | B轮 | 6170万欧元 | [领投] Altimeter Capital Mosaic Ventures Newion Investments | 德国 |
| 2024年4月 | Airchat | AI社交媒体应用 | 战略投资 | 未透露 | Sam Altman | 美国 |
| 2024年4月 | AI Squared | 低代码 AI 集成平台 | A轮 | 1380万美元 | NEA 恩颐投资-New Enterprise Associates Ansa Capital | 美国 |
| 2024年4月 | FYLD | 人工智能现场工作执行平台提供商 | 战略投资 | 1200万英镑 | [领投] 安大略省教师退休基金 Ontario Teachers | 英国 |
| 2024年4月 | sapien.io | AI数据标签提供商 | 种子轮 | 500万美元 | Animoca Brands Yield Guild Games Ravikant Capital Primitive Ventures | 美国 |
| 2024年4月 | OpenAI | 人工智能研究公司 | 战略投资 | 未透露 | ARK Invest | 美国 |
| 2024年4月 | Symbolica | AI基础模型开发商 | A轮 | 3100万美元 | [领投] Khosla Ventures Abstract Ventures General Catalyst Partners Buckley Ventures | 美国 |
| 2024年4月 | GoodGist | 人工智能技术开发商 | 战略投资 | 100万美元 | 微软 Microsoft 亚马逊 Amazon DX Partners Cedar Ridge Ventures FortyTwo.VC | 美国 |
| 2024年4月 | Captions | AI创意工作室 | A轮 | 1100万美元 | Andreessen Horowitz-a16z Ludlow Ventures Sequoia Capital 红杉 20VC Adjacent Chapter One Ventures | 美国 |
| 2024年4月 | Celestial AI | 人工智能加速器产品提供商 | C轮 | 1.75亿美元 | [领投] 美国创新技术基金 | 美国 |
| 2024年4月 | Brandtech | 生成式AI平台提供商 | C轮 | 1.15亿美元 | Mousse Partners Bansk Group Nendo Labs Fimalac | 美国 |

2024年海外市场AI相关企业重点融资案例盘点

| 时间 | 企业名称 | 细分赛道 | 融资阶段 | 融资金额 | 投资方 | 国家 |
|---------|-----------------------|-----------------------|------|---------|---|-----|
| 2024年4月 | SOAI | AI社交平台 | 天使轮 | 500万美元 | Archer Capital XForce Capital | 美国 |
| 2024年4月 | Rug.AI | 人工智能安全平台 | 种子轮 | 110万美元 | [领投] No Limit Holdings Mask Network Andrej Radonjic MacnBTC | 美国 |
| 2024年3月 | Anthropic | 人工智能大模型 Claude 开发商 | 战略投资 | 27.5亿美元 | 亚马逊Amazon | 美国 |
| 2024年3月 | Fireworks AI | AI实验和生产平台 开发商 | A轮 | 2500万美元 | Benchmark Capital Sequoia Capital 红杉 Databricks | 美国 |
| 2024年3月 | Borderless AI | AI人力资源管理平台 | 战略投资 | 2700万美元 | [领投] SGE [领投] Aglaé Ventures | 加拿大 |
| 2024年3月 | Quilt | 服务于销售的AI助手 开发商 | 种子轮 | 250万美元 | Sequoia Capital 红杉 | 美国 |
| 2024年3月 | Sprih | 人工智能技术平台 提供商 | 种子轮 | 300万美元 | [领投] 利奥资本 | 美国 |
| 2024年3月 | Lumino | 人工智能基础设施 提供商 | 种子轮 | 280万美元 | TRGC Fenbushi Capital Protocol Labs OrangeDAO | 美国 |
| 2024年3月 | Hippocratic AI | 生成式AI服务商 | A轮 | 5300万美元 | Premji Invest General Catalyst Partners SV Angel Memorial Hermann Hospital a16z Bio+Health Cincinnati Children's | 美国 |
| 2024年3月 | Loop.ai | 人工智能驱动餐 厅后台平台 | 种子轮 | 600万美元 | Afore Capital Base10 Partners | 美国 |
| 2024年3月 | Axion Ray | 人工智能可观测 指挥中心提供商 | A轮 | 1750万美元 | Amplio BVP RTX Ventures Inspired Capital Partners | 美国 |
| 2024年3月 | Physical Intelligence | 通用机器人系统 的人工智能 | 战略投资 | 7000万美元 | Khosla Ventures Lux Capital Thrive Capital Greencoaks Capital Management Sequoia Capital 红杉 OpenAI | 美国 |
| 2024年3月 | Adaptive | 人工智能服务商 | 种子轮 | 2000万美元 | Index Ventures Iconiq Capital Motier Ventures Databricks Ventures Iris Ventures Financiere Saint James HuggingFund by Factori | 法国 |
| 2024年3月 | Cognition | AI程序员技术研 发应用商 | 种子轮 | 2100万美元 | [领投] Founders Fund Elad Gil Conviction Partners | 美国 |

2024年海外市场AI相关企业重点融资案例盘点

| 时间 | 企业名称 | 细分赛道 | 融资阶段 | 融资金额 | 投资方 | 国家 |
|---------|----------------|---------------------|------|---------|---|-----|
| 2024年3月 | Bluwhale | AI Web3服务提供商 | 种子轮 | 700万美元 | [领投] SBI Baselayer Capital Spyre Capital Ghaf Capital Partners | 美国 |
| 2024年3月 | Kaedim | 人工智能2D图像转换3D模型应用研发商 | A轮 | 1500万美元 | [领投] Andreessen Horowitz-a16z Pioneer Fund | 英国 |
| 2024年3月 | Cogna | 人工智能驱动SaaS平台提供商 | 战略投资 | 376万英镑 | [领投] Hoxton Ventures Notion Capital Octopus First Check Fund | 英国 |
| 2024年3月 | Jabali | 生成型AI游戏引擎开发商 | 种子轮 | 500万美元 | [领投] Bitkraft Ventures Sony Innovation Fund 亚马逊Amazon 谷歌Google OpenAI | 美国 |
| 2024年3月 | Theia Insights | 人工智能解决方案提供商 | 战略投资 | 650万美元 | [领投] Unusual Ventures 富达投资Fidelity Investments Clocktower Ventures | 英国 |
| 2024年3月 | Ema | AI办公自动化解决方案提供商 | 战略投资 | 2500万美元 | [领投] Prosus Ventures-Naspers [领投] Section 32 [领投] Accel Partners Firebolt Ventures AME Cloud Ventures | 美国 |
| 2024年3月 | Taalas | 人工智能和芯片创新服务商 | 战略投资 | 5000万美元 | [领投] Quiet Capital [领投] Pierre Lamond | 加拿大 |
| 2024年3月 | Baseten | 人工智能基础设施提供商 | B轮 | 4000万美元 | [领投] Institutional Venture Partners [领投] Spark Capital | 美国 |
| 2024年3月 | Wordware | 使用自然语言创建AI代理工具 | 种子轮 | 数百万美元 | Y Combinator | 美国 |
| 2024年3月 | SynthLabs | 人工智能技术提供商 | 种子轮 | 未透露 | M12 First Spark Ventures | 美国 |
| 2024年2月 | Ideogram | 生成式人工智能服务商 | A轮 | 8000万美元 | 未透露 | 加拿大 |
| 2024年2月 | Collow | 智能室内设计及家居电商平台 | A轮 | 1000万美元 | [领投] GoldenHome | 美国 |
| 2024年2月 | KIT-AR | 工业AR解决方案提供商 | 战略投资 | 330万欧元 | [领投] 3XP Global [领投] Blue Crow Capital Sintef Venture V Armilar Venture Partners Criteria Venture Tech | 英国 |
| 2024年2月 | Glean | 工作助手和知识管理平台 | D轮 | 2亿美元 | General Catalyst Partners Workday Ventures 花旗Citi Ventures 纬度创投 ICONIQ Growth Coatue Management Lightspeed Venture Partners KPCB凯鹏华盈中国 | 美国 |

2024年海外市场AI相关企业重点融资案例盘点

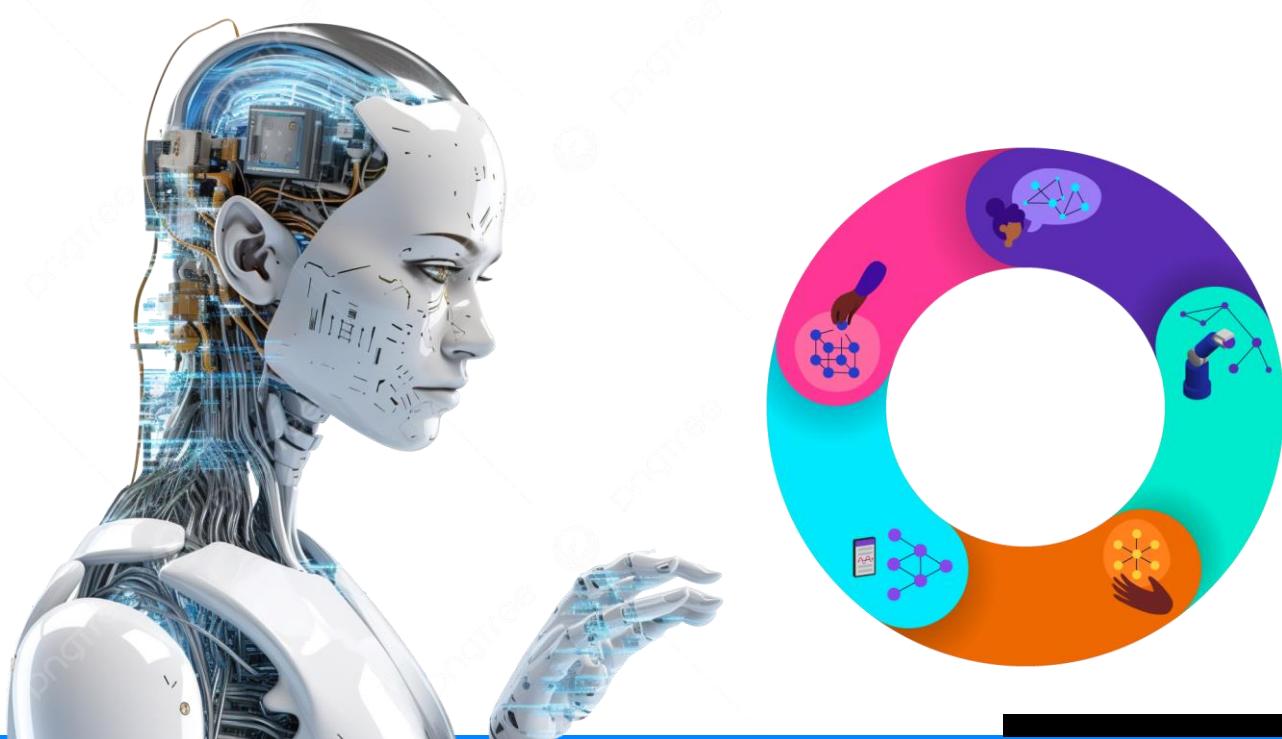
| 时间 | 企业名称 | 细分赛道 | 融资阶段 | 融资金额 | 投资方 | 国家 |
|---------|--------------------|---------------------|--------|----------|---|----|
| 2024年2月 | FlowGPT | 在线聊天平台服务商 | Pre-A轮 | 1000万美元 | Goodwater Capital | 美国 |
| 2024年2月 | NLX | 对话式人工智能服务提供商 | A轮 | 1200万美元 | Cercano Management Thayer Ventures HL Ventures IAG Capital Partners | 美国 |
| 2024年2月 | Genmo | 人工智能创意内容生成平台 | 战略投资 | 3000万美元 | NEA恩颐投资-New Enterprise Associates | 美国 |
| 2024年2月 | Talus Network | 基于MoveVM的链上人工智能网络平台 | 种子轮 | 300万美元 | [领投] Polychain Capital IBM 英伟达NVIDIA TRGC | 美国 |
| 2024年2月 | Mistral AI | 人工智能大模型服务商 | 战略投资 | 1630万美元 | 微软Microsoft | 法国 |
| 2024年2月 | Datology AI | AI大模型训练服务商 | 种子轮 | 1170万美元 | [领投] Amplify Partners Radical Ventures Quiet Capital Douwe Kiela Geoffrey Hinton | 美国 |
| 2024年2月 | Magic AI | 人工智能编码助手 | A轮 | 1.45亿美元 | Nat Friedman Daniel Gross AI Grant CapitalG谷歌资本 Elad Gil | 美国 |
| 2024年2月 | Bioptimus | 通用人工智能基础模型开发商 | 种子轮 | 3500万美元 | [领投] Sofinnova Partners FORLIFE [领投] Bpifrance Large Venture NJF Capital Hummingbird Ventures Cathay Innovation | 法国 |
| 2024年2月 | Altavo | 人工智能语音康复服务商 | A轮 | 540万美元 | High-Tech Gründerfonds Occident Novalis Biotech Beteiligungsmanagement Thüringen TGFS Technologiegründerfonds Sachsen Saxonia Systems Holding TUDAG TU Dresden AG | 德国 |
| 2024年2月 | Anthropic | 人工智能大模型Claude开发商 | 战略投资 | 数亿美元 | 高通Qualcomm INTUIT | 美国 |
| 2024年2月 | Praktika | AI语言学习应用研发商 | Pre-A轮 | 未透露 | Yellow Rock Blue Wire Capital Creator Ventures | 美国 |
| 2024年2月 | Mechanical Orchard | 人工智能增强工具研发商 | A轮 | 2400万美元 | [领投] Emergence Capital Partners Bloomberg Beta Industry Ventures | 美国 |
| 2024年2月 | Brix | 海外AI跨境招聘平台 | 天使轮 | 2000万人民币 | [领投] 峰瑞资本 Plug and Play | 美国 |
| 2024年2月 | CARPL | 放射科技服务商 | 种子轮 | 600万美元 | [领投] Stellaris Venture Partners | 美国 |

2024年海外市场AI相关企业重点融资案例盘点

| 时间 | 企业名称 | 细分赛道 | 融资阶段 | 融资金额 | 投资方 | 国家 |
|---------|--------------------|----------------------|------|---------|--|----|
| 2024年2月 | HireAra | 招聘技术开发商 | 种子轮 | 45万英镑 | Candidate.ID One Up Sales SourceWhale Mercury | 英国 |
| 2024年2月 | SEMRON | 人工智能芯片研发商 | 种子轮 | 793万美元 | Join Capital SquareOne | 德国 |
| 2024年1月 | Codeium | 人工智能编码工具包提供商 | B轮 | 6500万美元 | [领投] Kleiner Perkins KPCB | 美国 |
| 2024年1月 | Sema4.ai | 开源人工智能技术应用服务商 | 战略投资 | 3050万美元 | Benchmark Capital Mayfield Fund Canvas Ventures | 美国 |
| 2024年1月 | Kore.ai | 对话式AI平台提供商 | D轮 | 1.5亿美元 | 英伟达NVIDIA | 美国 |
| 2024年1月 | OnPoint Healthcare | 人工智能技术服务提供商 | 战略投资 | 未透露 | [领投] Peloton Equity Fort Maitland Capital | 美国 |
| 2024年1月 | Sierra | 人工智能客服服务商 | A轮 | 8500万美元 | [领投] Sequoia Capital红杉 Benchmark Capital | 美国 |
| 2024年1月 | Cultivo | 人工智能驱动科技平台 | A轮 | 1400万美元 | Conviction Partners [领投] MassMutual Ventures [领投] Octopus Energy Generation Peña Verde Salkantay Ventures | 美国 |
| 2024年1月 | ViralMoment | 人工智能社交视频洞察和分析解决方案提供商 | 种子轮 | 250万美元 | [领投] Supernode Global Techstars Ventures 卡内基梅隆大学 Duo Partners Crush Ventures | 美国 |
| 2024年1月 | Empeq | 人工智能技术服务商 | 战略投资 | 未透露 | Leonid Capital Partners | 美国 |
| 2024年1月 | Eleven Labs | AI语音合成软件研发商 | B轮 | 8000万美元 | [领投] Andreessen Horowitz-a16z [领投] Nat Friedman Sequoia Capital红杉 SV Angel Credo Ventures Smash Capital BroadLight Capital | 英国 |
| 2024年1月 | Mercor | 全自动AI人才评估平台提供商 | 战略投资 | 360万美元 | [领投] General Catalyst Partners Soma Capital 2 Twelve Angels Link Ventures | 美国 |
| 2024年1月 | Poe AI | 生成式AI聊天助手聚合平台 | A轮 | 7500万美元 | Andreessen Horowitz-a16z | 美国 |
| 2024年1月 | Luma AI | 3D模型研发商 | B轮 | 4300万美元 | [领投] Andreessen Horowitz-a16z 英伟达NVIDIA Matrix Partners经纬海外 Amplify Partners | 美国 |

2024年海外市场AI相关企业重点融资案例盘点

| 时间 | 企业名称 | 细分赛道 | 融资阶段 | 融资金额 | 投资方 | 国家 |
|---------|----------------------------|------------------------|------|---------|--|-----|
| 2024年1月 | Geminus.AI | 物理信息技术 研发商 | A轮 | 1300万美元 | SLB | 美国 |
| 2024年1月 | Agrawal's AI initiative | 人工智能技术 研发和应用服 务商 | 种子轮 | 3000万美元 | [领投] Khosla Ventures Index Ventures First Round Capital | 美国 |
| 2024年1月 | Durable | 人工智能网站 建设应用研发 商 | 战略投资 | 1400万美元 | Spark Capital Torch Capital South Park Ventures 南方公 园 Soma Capital Infinity Ventures Crypto Dash Fund Altman Capital | 加拿大 |
| 2024年1月 | Perplexity AI | 智能对话式搜 索引擎提供商 | B轮 | 7360万美元 | NEA 恩颐投资- New Enterprise Associates Institutional Venture Partners 英伟达NVIDIA Jeff bezos | 美国 |
| 2024年1月 | Tobiko | 扩展数据基础 设施 | 种子轮 | 450万美元 | Fivetran Census 20Sales MotherDuck | 美国 |



中国AI创投主要聚焦应用层，利用开源模型拓展应用场景，初创企业数量多，但具备全球影响力项目较少

2024年中国市场AI相关企业重点融资案例盘点

| 时间 | 企业名称 | 细分赛道 | 融资阶段 | 融资金额 | 投资方 | 城市 |
|------------|------|--------------------|--------|--------|--|----|
| 2024-12-11 | 面壁智能 | 人工智能大模型加速与应用落地赋能公司 | B轮 | 数亿人民币 | [财务顾问] 万甲资本 [领投] 鼎晖投资 [领投] 中关村科学城 [领投] 赛富基金SAIF Partners [领投] 龙芯创投 北京人工智能产业基金 清科创投 | 北京 |
| 2024-11-15 | 宾果智能 | AI教育机器人研发商 | B轮 | 近亿人民币 | 九颂山河基金 三合资本 | 北京 |
| 2024-10-17 | 平方和 | 工业4.0解决方案提供商 | B轮 | 数亿人民币 | [领投] 同创伟业 元禾璞华 雅惠投资 云晖资本 | 北京 |
| 2024-10-15 | 玻色量子 | 量子计算服务商 | A轮 | 数亿人民币 | [领投] 啓赋资本 阿米巴资本 元和资本 盈富泰克 | 北京 |
| 2024-10-11 | 基流科技 | 高性能网络服务提供商 | Pre-A轮 | 近亿人民币 | 光速光合 星连资本 中关村科学城 | 北京 |
| 2024-9-27 | 氢源智能 | 智能装备与技术创新服务商 | B轮 | 近亿人民币 | 上海海际朴诚 吴兴交投集团 | 北京 |
| 2024-9-26 | 易航智能 | 智能驾驶科技公司 | C轮 | 数亿人民币 | 北汽产投 财通资本 浙江金控 德清产投 | 北京 |
| 2024-9-26 | 潞晨科技 | 高性能计算解决方案提供商 | A+轮 | 数亿人民币 | [财务顾问] 义柏资本 北京人工智能产业基金 领沨资本 石溪资本 Capstone Partners | 北京 |
| 2024-9-13 | 分子之心 | AI蛋白质设计平台 | A轮 | 数亿人民币 | [领投] 谢诺投资 [领投] 深创投 商汤国香资本 久奕投资 | 北京 |
| 2024-9-11 | 风平智能 | 数字人AIGC平台 | A轮 | 近亿人民币 | [领投] 璀璨资本 [领投] 华鲲资本 | 北京 |
| 2024-9-5 | 智谱AI | 中文认知大模型平台 | D轮 | 数十亿人民币 | [领投] 中关村科学城 红杉中国 高瓴投资 君联资本 | 北京 |

2024 年中国市场 AI 相关企业重点融资案例盘点

| 时间 | 企业名称 | 细分赛道 | 融资阶段 | 融资金额 | 投资方 | 城市 |
|-----------|-------------|------------------|------|---------|--|----|
| 2024-9-2 | 无问芯穹 | 提供 AGI 算力解决方案 | A 轮 | 5 亿人民币 | [财务顾问] 光源资本 [领投] 君联资本 [领投] 启明创投 [领投] 洪泰基金 联想创投 小米集团 达晨财智 国开科创 临港科创 德同资本 徐汇科投 申万宏源 顺为资本 | 北京 |
| 2024-8-14 | 水木分子 | 生物医药基础大模型研发商 | 天使轮 | 近亿人民币 | [领投] 华山资本 WestSummit Capital 道彤投资 科大讯飞 | 北京 |
| 2024-8-12 | 卓世科技 | MaaS 赋能企业数字转型 | B+ 轮 | 数亿人民币 | [领投] 同创伟业 启迪之星 青岛国资平台 青岛海发 | 北京 |
| 2024-8-7 | 宽凳科技 | 高精地图综合解决方案服务商 | B+ 轮 | 近亿人民币 | 融泰资本 浙江德清政府产业基金 | 北京 |
| 2024-8-7 | 零一万物 | 人工智能 AIGC 大模型服务商 | A 轮 | 数亿美元 | 未透露 | 北京 |
| 2024-7-29 | LiblibAI 哩布 | AI 绘画原创模型网站 | A 轮 | 数亿人民币 | [财务顾问] 远识资本 [领投] 明势资本 源码资本 高榕资本 金沙江创投 | 北京 |
| 2024-7-25 | 百川智能 | AI 人工智能技术公司 | A+ 轮 | 50 亿人民币 | [财务顾问] 光源资本 顺禧基金 临港科创 深创投 阿里巴巴 腾讯投资 小米集团 中金资本 好未来 慕华科创 三七互娱 中贝通信 信雅达 亚洲投资基金 | 北京 |
| 2024-7-17 | 耀速科技 | 人工智能药物研发商 | 天使轮 | 1 亿人民币 | [财务顾问] 浩悦资本 [领投] 鼎泰集团 正轩投资 天图投资 君联资本 雅亿资本 | 北京 |

2024年中国市场AI相关企业重点融资案例盘点

| 时间 | 企业名称 | 细分赛道 | 融资阶段 | 融资金额 | 投资方 | 城市 |
|-----------|----------|--------------------|--------|---------|---|----|
| 2024-7-4 | 硅基流动 | AI大模型推理和部署系统 | 天使轮 | 近亿人民币 | [财务顾问] 华兴资本 智谱AI 奇虎360 哈勃投资(华为旗下) | 北京 |
| 2024-6-28 | 51WORLD | 中国领先的VR+AI科技公司 | F轮 | 2亿人民币 | [财务顾问] 金元证券 南宁焕新资本 | 北京 |
| 2024-6-6 | 深思考 | 专注于类脑人工智能与深度学习核心科技 | B轮 | 数亿人民币 | 鼎信泰和 | 北京 |
| 2024-6-5 | 赋乐科技 | 大数据与信息安全服务提供商 | 战略投资 | 近亿人民币 | 微智数科 中国电科 华迪创投管理 | 北京 |
| 2024-6-5 | 生数科技 | 多模态生成式大模型与应用产品开发商 | A+轮 | 数亿人民币 | [财务顾问] 华兴资本 [领投] 北京人工智能产业基金 [领投] 百度 中关村科学城 启明创投 卓源亚洲 锦秋基金 哈勃投资(华为) | 北京 |
| 2024-5-31 | 智谱AI | 中文认知大模型平台 | C轮 | 4亿美元 | Prosperity7 Ventures | 北京 |
| 2024-5-21 | 月之暗面Kimi | AI初创大模型公司 | B轮 | 3亿美元 | 腾讯投资 高榕资本 源码资本 | 北京 |
| 2024-5-9 | 超星未来 | 边缘侧人工智能芯片提供商 | Pre-B轮 | 数亿人民币 | 中安资本 龙鼎投资 讯飞创投 梁溪科创 天智投资 陕汽智能汽车基金 | 北京 |
| 2024-5-7 | 主线科技 | L4自动驾驶卡车服务提供商 | B+轮 | 数亿人民币 | 顺义科创 民航投资基金 常州钟楼金控 | 北京 |
| 2024-4-26 | 沐言智语 | AIGC产品开发服务商 | Pre-A轮 | 1.2亿人民币 | [领投] 高瓴创投 明势资本 Monolith Management 砺思资本 | 北京 |
| 2024-4-24 | 爱诗科技 | AIGC视觉多模态算法开发商 | A+轮 | 1亿人民币 | [财务顾问] 光源资本 [领投] 蚂蚁集团 | 北京 |
| 2024-4-16 | 诺谛智能 | 认知与决策人工智能企业 | Pre-A轮 | 近亿人民币 | [领投] 武岳峰资本 清智资本 联想集团 三叶虫创投 | 北京 |
| 2024-4-12 | 瑞莱智慧 | AI行业应用服务提供商 | 战略投资 | 数亿人民币 | [财务顾问] 光源资本 顺禧基金 北京人工智能产业基金 | 北京 |
| 2024-4-11 | 深势科技 | 药物模拟研发平台 | C+轮 | 数亿人民币 | 顺禧基金 北京人工智能产业基金 中关村科学城 | 北京 |

2024年中国市场AI相关企业重点融资案例盘点

| 时间 | 企业名称 | 细分赛道 | 融资阶段 | 融资金额 | 投资方 | 城市 |
|-----------|-----------|--------------------|--------|---------|--|----|
| 2024-4-11 | 面壁智能 | 人工智能大模型加速与应用落地赋能公司 | A轮 | 数亿人民币 | [领投] 春华创投 [领投] 哈勃投资(华为旗下) 顺禧基金 知乎 | 北京 |
| 2024-5-21 | 月之暗面 Kimi | AI初创大模型公司 | B轮 | 3亿美元 | 腾讯投资 高榕资本 源码资本 五源资本 云九资本 | 北京 |
| 2024-4-16 | 诺谛智能 | 认知与决策人工智能企业 | Pre-A轮 | 近亿人民币 | [领投] 武岳峰资本 清智资本 联想集团 | 北京 |
| 2024-4-12 | 瑞莱智慧 | AI行业应用服务提供商 | 战略投资 | 数亿人民币 | [财务顾问] 光源资本 顺禧基金 北京人工智能产业基金 | 北京 |
| 2024-4-11 | 深势科技 | 药物模拟研发平台 | C+轮 | 数亿人民币 | 顺禧基金 北京人工智能产业基金 中关村科学城 | 北京 |
| 2024-3-27 | 新石器无人车 | 载物型无人车研发、制造及服务提供商 | C轮 | 6亿人民币 | 中金汇融 前海母基金 中金资本 壳牌 Shell Ventures | 北京 |
| 2024-3-14 | 智谱AI | 中文认知大模型平台 | 战略投资 | 数亿人民币 | 北京人工智能产业基金 | 北京 |
| 2024-3-12 | 生数科技 | 多模态生成式大模型与应用产品开发商 | A轮 | 数亿人民币 | [领投] 启明创投 达泰资本 Z基金-智谱AI 百度风投 | 北京 |
| 2024-3-11 | 爱诗科技 | AIGC视觉多模态算法开发商 | A轮 | 1亿人民币 | [财务顾问] 光源资本 达晨财智 | 北京 |
| 2024-3-11 | 中科世通亨奇 | 数据智能解决方案服务商 | B轮 | 2亿人民币 | 乐礼资本 达晨财智 | 北京 |
| 2024-3-7 | 雷科智途 | 商用车自动驾驶研发商 | A轮 | 1亿人民币 | [领投] 中关村发展集团 [领投] 中关村创投 | 北京 |
| 2024-2-19 | 月之暗面 Kimi | AI初创大模型公司 | A轮 | 10亿美元 | [领投] 阿里巴巴 [领投] 蚂蚁集团 红杉中国 小红书 美团 Monolith Management 砾思资本 九安医疗 蓝驰创投 襄禾资本 宿华 | 北京 |
| 2024-1-22 | 山景智能 | 企业业务超自动化产品和解决方案提供商 | A+轮 | 近亿人民币 | [领投] 中科海创 | 北京 |
| 2024-1-22 | 卓翼智能 | 无人机系统解决方案提供商 | B轮 | 2.5亿人民币 | 中关村科学城 中航融富 陕西光子强链 | 北京 |

2024年中国市场AI相关企业重点融资案例盘点

| 时间 | 企业名称 | 细分赛道 | 融资阶段 | 融资金额 | 投资方 | 城市 |
|------------|-----------|----------------------|--------|---------|--|----|
| 2024-1-22 | DPM | 高端医疗设备和创新医疗技术提供商 | C轮 | 2亿人民币 | [领投] 道禾志医 | 北京 |
| 2024-1-5 | 的卢深视 | 专注于计算机视觉和人工智能的高新技术企业 | B轮 | 1.5亿人民币 | 国科新能 创东方投资 合肥创新投资 | 北京 |
| 2024-1-4 | 千挂科技 | 无人驾驶货运卡车技术研发商 | Pre-A轮 | 数亿人民币 | 凯辉基金 亦庄国投 IDG资本 浦发硅谷银行 | 北京 |
| 2024-1-3 | 斯年智驾 | 港口全栈式无人集卡运输解决方案 | B轮 | 数亿人民币 | [领投] 力合科创 [领投] 浙江金融控股 | 北京 |
| 2024-12-5 | 穹彻智能 | 具身智能公司 | Pre-A轮 | 数亿人民币 | [领投] 红杉中国 Prosperity7 Ventures 小苗朗程 Plug and Play | 上海 |
| 2024-10-25 | 奔曜科技 | 生命科学领域自动化解决方案提供商 | A+轮 | 数亿人民币 | [领投] 璞霖资本 启明创投 博远资本 | 上海 |
| 2024-9-19 | VAST哇嘶嗒科技 | 通用3D大模型研发商 | Pre-A轮 | 数亿人民币 | [领投] 达晨财智 [领投] 春华创投 英诺天使基金 水木清华校友种子基金 | 上海 |
| 2024-9-12 | 快仓 | 智能仓储机器人系统解决方案提供商 | D轮 | 1亿美元 | 金杜鹃资本 远雄集团 无锡梁溪科创 潍坊鸢飞产业基金 | 上海 |
| 2024-9-6 | 穹彻智能 | 具身智能公司 | Pre-A轮 | 数亿人民币 | [领投] Prosperity7 Ventures [领投] 广发信德 创新工场 奇绩创坛 Plug and Play 魔量资本MFund 泽羽资本 | 上海 |
| 2024-8-9 | 秘塔科技 | 智能语义数据服务提供商 | A轮 | 1亿人民币 | [领投] 蚂蚁集团 光速光合 [财务顾问] 指数资本 [领投] 鼎晖投资 新尚资本 | 上海 |
| 2024-8-2 | 思朗科技 | 微处理器技术IP授权及芯片服务提供商 | 战略投资 | 数亿人民币 | 贵州茅台 金石投资 鸣渠资本 国鑫创投 交控金石基金 | 上海 |
| 2024-7-25 | 鲸鱼机器人 | 人工智能及编程教育服务商 | B轮 | 1亿人民币 | 五星体育 上海久事集团 陆家嘴集团 | 上海 |
| 2024-7-19 | 感图科技 | 计算机视觉技术及产品开发商 | C+轮 | 数亿人民币 | 博华资本 | 上海 |

2024年中国市场AI相关企业重点融资案例盘点

| 时间 | 企业名称 | 细分赛道 | 融资阶段 | 融资金额 | 投资方 | 城市 |
|------------|--------------|------------------|------|-------|--|----|
| 2024-7-3 | 医日健 | 医药行业数智化解决方案提供商 | 战略投资 | 1亿人民币 | 香港汇金股份有限公司 | 上海 |
| 2024-3-4 | Minimax稀宇科技 | 通用人工智能科技公司 | 战略投资 | 6亿美元 | [领投] 阿里巴巴 红杉中国 高瓴创投 经纬创投 | 上海 |
| 2024-7-8 | 边界智控 | 人工智能 | A轮 | 近亿人民币 | [财务顾问] 投中资本 基石资本 南山战新投 北航投资 普华资本 | 深圳 |
| 2024-4-10 | 墨芯人工智能 | AI芯片设计商 | B轮 | 数亿人民币 | [财务顾问] 告捷资本 [领投] 蚂蚁集团 盛景网联(盛景嘉成) | 深圳 |
| 2024-4-2 | 墨影科技 | 机器人及智能制造解决方案提供商 | A+轮 | 近亿人民币 | 天奇创投 | 深圳 |
| 2024-2-23 | 今日人才 | 人才供应服务提供商 | C轮 | 近亿人民币 | [领投] 盛景网联(盛景嘉成) [领投] 凯思博 今日资本 | 深圳 |
| 2024-1-4 | 光鉴科技 | 光学技术及3D视觉解决方案服务商 | B轮 | 2亿人民币 | 中金资本 一村淞灵 重庆科兴 | 深圳 |
| 2024-10-21 | 网思科技 | IT系统解决方案提供商 | A+轮 | 1亿人民币 | 范式基金 | 广州 |
| 2024-8-1 | SynSense时识科技 | 类脑计算及类脑芯片设计与研发商 | B轮 | 数亿人民币 | [财务顾问] 势能资本 [领投] 宁波通商基金 Samsung Ventures三星 | 南京 |
| 2024-7-15 | 后摩智能 | 原创新型智能计算芯片研发商 | 战略投资 | 数亿人民币 | 中移创新产业基金(中国移动) | 南京 |
| 2024-2-20 | 新格视讯 | 智能视频解决方案提供商 | B轮 | 近亿人民币 | [领投] 同创伟业 | 南京 |
| 2024-10-12 | 亿铸科技 | AI大算力芯片公司 | A轮 | 数亿人民币 | [领投] Prosperity7 Ventures 行至资本 盛视科技 co-founders | 徐州 |
| 2024-8-9 | 太初元碁 | 智能算力系统研发生产商 | A+轮 | 数亿人民币 | [财务顾问] 告捷资本 金蚂投资 霜叶创投 | 无锡 |
| 2024-5-31 | 斯帝尔 | 柔性打磨机器人研发商 | A轮 | 近亿人民币 | 浙商创投 航天科工资产 | 无锡 |

2024年中国市场AI相关企业重点融资案例盘点

| 时间 | 企业名称 | 细分赛道 | 融资阶段 | 融资金额 | 投资方 | 城市 |
|------------|--------------|--------------------|---------|---------|---|----|
| 2024-4-11 | 卡尔曼 | 自动驾驶智能装备企业 | C轮 | 数亿人民币 | [领投] 百联挚高 创新工场 深投控 浚源资本 中小企业专精特新基金 瑞世财富 | 无锡 |
| 2024-11-11 | 九识智能 | L4级自动驾驶产品研发企业 | B轮 | 1亿美元 | [财务顾问] 光源资本 [领投] 蓝湖资本 | 苏州 |
| 2024-6-6 | 星逻智能 | 无人机数据采集分析服务商 | B轮 | 1亿人民币 | [领投] 粤科金融 [领投] 临创投资 远瞻资本 常春藤资本 宁波诺登创投 苏州工业园区科创基金 遨问创投 | 苏州 |
| 2024-3-18 | 光本位 | 光通信器件及光计算芯片研发制造商 | 天使轮 | 近亿人民币 | [财务顾问] 慕石资本 [领投] 中瀛投资 慕石资本 小苗朗程 峰瑞资本 创业接力 | 苏州 |
| 2024-2-27 | 九识智能 | L4级自动驾驶产品研发企业 | A轮 | 1亿美元 | [财务顾问] 光源资本 [领投] 美团 百度风投 索道投资 蓝湖资本 | 苏州 |
| 2024-1-29 | 辅易航 | 智能停车公司 | B轮 | 近亿人民币 | [领投] 苏高新金控/苏高新创投 [领投] 元禾重元 | 苏州 |
| 2024-1-10 | 思必驰 | 对话式人工智能平台 | Pre-IPO | 2亿人民币 | 未透露 | 苏州 |
| 2024-12-5 | 行芯科技 | 专注集成电路芯片的设计软件与IP开发 | D轮 | 数亿人民币 | 国家集成电路产业投资基金 青呈投资 | 杭州 |
| 2024-12-2 | 百应科技 | AI解决方案提供商 | C轮 | 1亿人民币 | [领投] 武汉创新投 [领投] 湖北产融资本 | 杭州 |
| 2024-9-13 | 中昊芯英 | AI芯片研发商 | B+轮 | 2.5亿人民币 | 艾布鲁 | 杭州 |
| 2024-9-11 | Mindflow曼孚科技 | AI基础架构与数据智能平台服务商 | B+轮 | 数亿人民币 | 前海信诺 闲庭基金 | 杭州 |
| 2024-8-1 | 中昊芯英 | AI芯片研发商 | B轮 | 近亿人民币 | 浙江正方资产 | 杭州 |
| 2024-6-22 | 剂泰医药 | 人工智能 | C轮 | 1亿美元 | [领投] 中金资本 中国太平 | 杭州 |
| 2024-3-13 | 百应科技 | 人工智能 | 战略投资 | 2亿人民币 | 百度 湖北高投 | 杭州 |

2024 年中国市场 AI 相关企业重点融资案例盘点

| 时间 | 企业名称 | 细分赛道 | 融资阶段 | 融资金额 | 投资方 | 城市 |
|-----------|--------------|---------------------|---------|-------|---|----|
| 2024-2-29 | 明度智云 | 人工智能 | C轮 | 数亿人民币 | [财务顾问] 指数资本 [领投] 深创投 招商健康 | 杭州 |
| 2024-2-28 | 联汇科技 | 人工智能 | 战略投资 | 数亿人民币 | [领投] 中移创新产业基金(中国移动) 前海母基金 | 杭州 |
| 2024-2-18 | 实在智能 | 人工智能 | C轮 | 2亿人民币 | [领投] 金泰富资本 [领投] 安吉智慧谷 卓源亚洲 两山国控集团 | 杭州 |
| 2024-2-1 | 杰毅生物 | 人工智能 | C轮 | 数亿人民币 | 余杭国投 余杭转型升级产业基金 昆泰投资 | 杭州 |
| 2024-5-31 | 渊亭科技 DataExa | 一站式认知智能平台与服务厂商 | B+轮 | 4亿人民币 | 奇安投资 创东方投资 达晨财智 厦金创新 思明科创 福建省电子信息 天健周行 国家战新产业基金 重庆制造业转型升级基金 财信中金 | 厦门 |
| 2024-8-9 | 考拉悠然 | AI 科技创新公司 | B 轮 | 1亿人民币 | 策源资本 深圳诺辰实业投资 | 成都 |
| 2024-4-9 | 星凡星启 | 一站式行业 AIGC 技术服务提供商 | Pre-A 轮 | 近亿人民币 | [领投] 盛景网联(盛景嘉成) 高捷资本 开普云 | 成都 |
| 2024-2-21 | 万像电子 | 计算机图像人机交互协议及零终端专用芯片 | A+ 轮 | 近亿人民币 | 西高投 西交大一八九六资本 中财融商 空天院 常青资本 瑞圭 | 西安 |
| 2024-1-17 | 欧卡智舶 | 水面无人驾驶技术及水面服务机器人研发商 | B 轮 | 数亿人民币 | 戈壁大湾区 Gobi GBA 天善资本 西安财金 | 西安 |

03

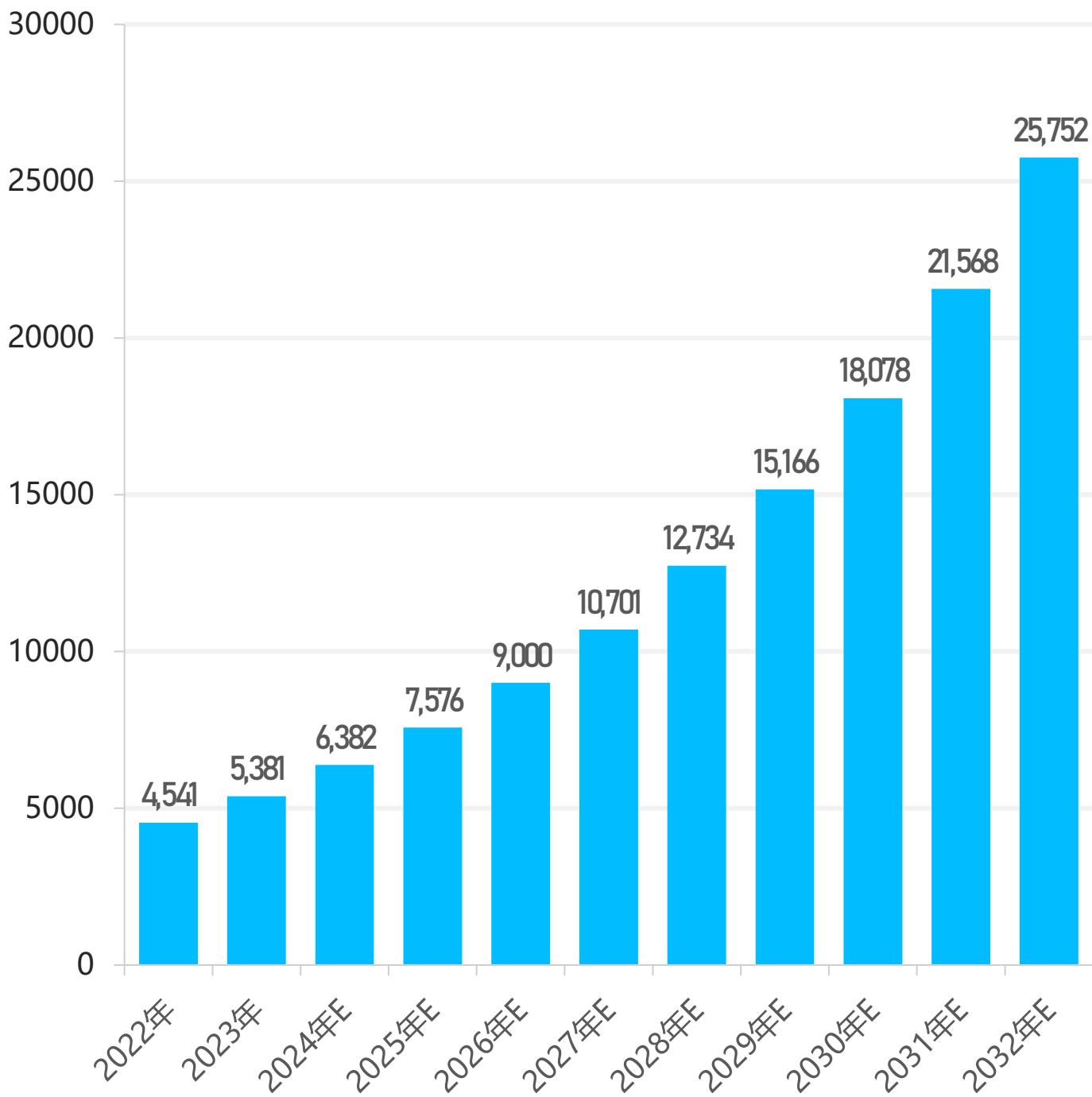
走向应用

2023年全球人工智能市场规模高达惊人的5381亿美元，预计2032年全球人工智能市场将超2.5万亿美元

全球人工智能市场规模

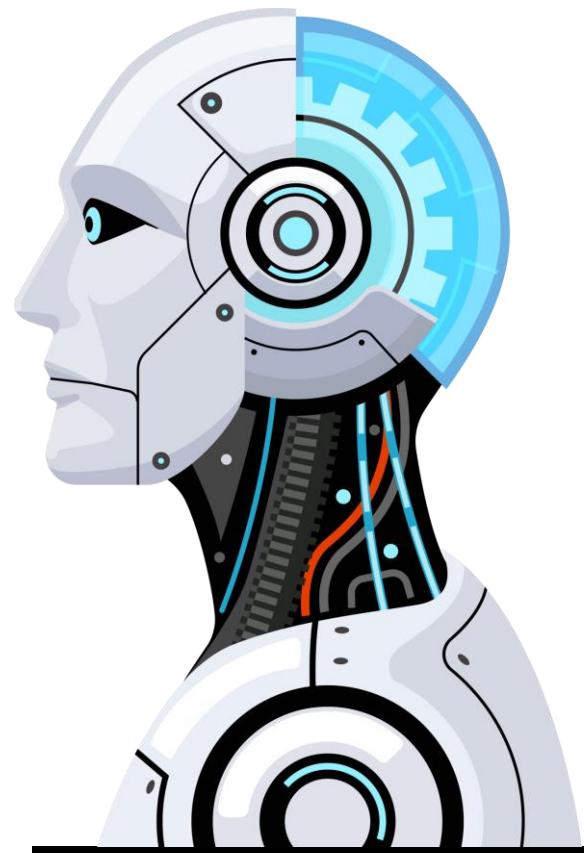
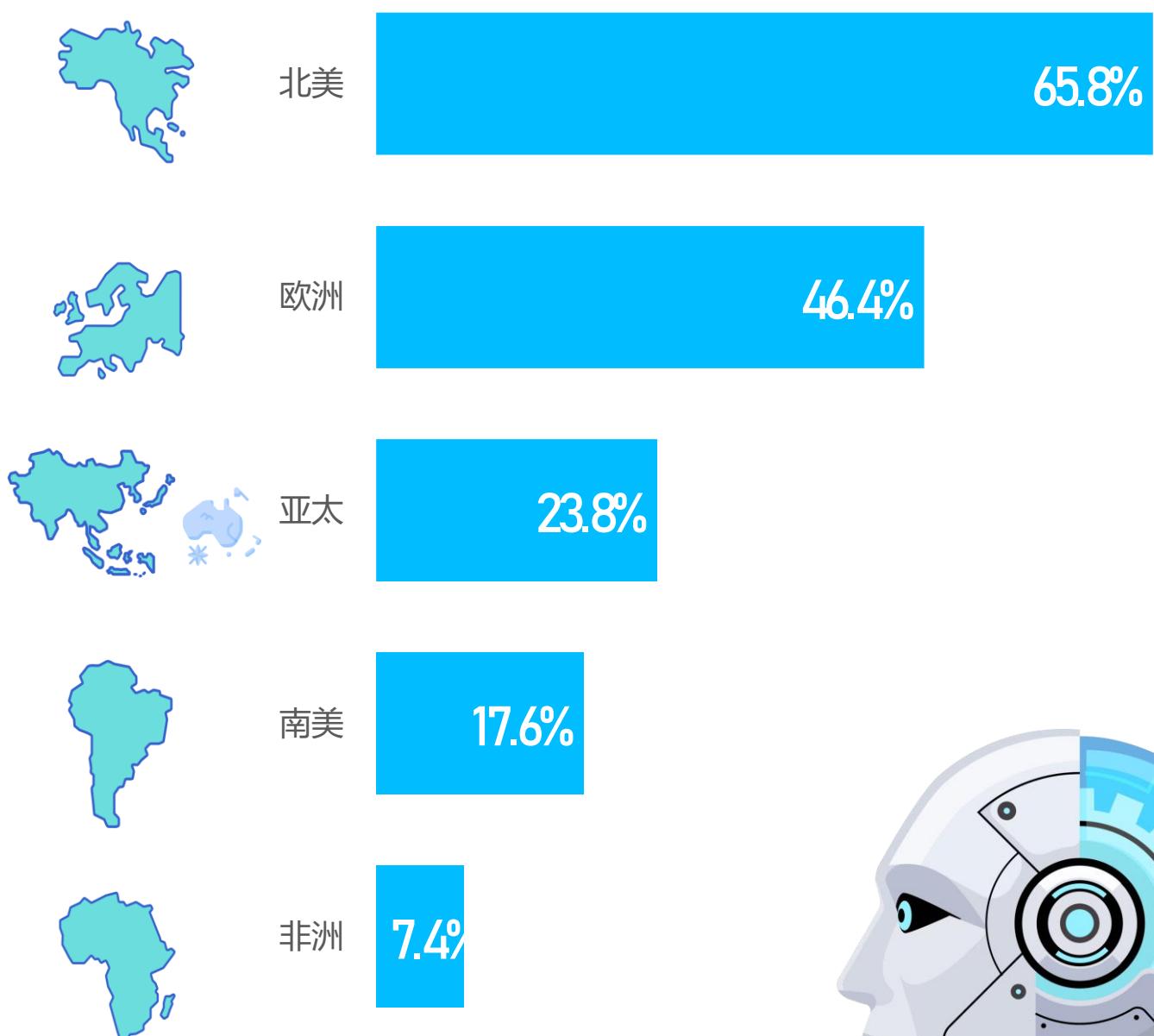
2022年-2032年全球人工智能市场规模及预估

单位：亿美元

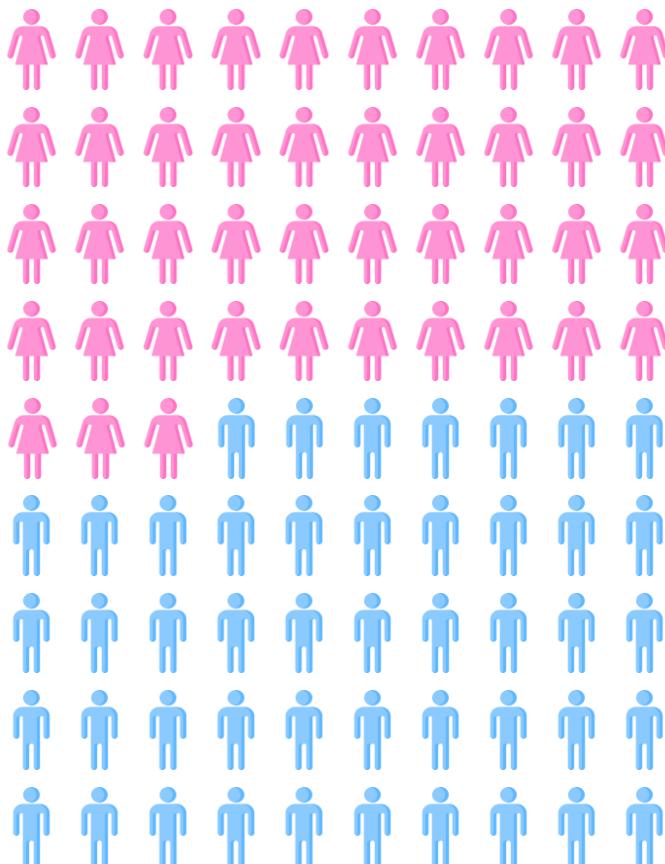


人工智能工具在以前所未有的速度普及，2024年上半年，近七成北美洲用户每周使用人工智能工具

2024年上半年全球各地区每周至少使用一次人工智能工具的用户比例



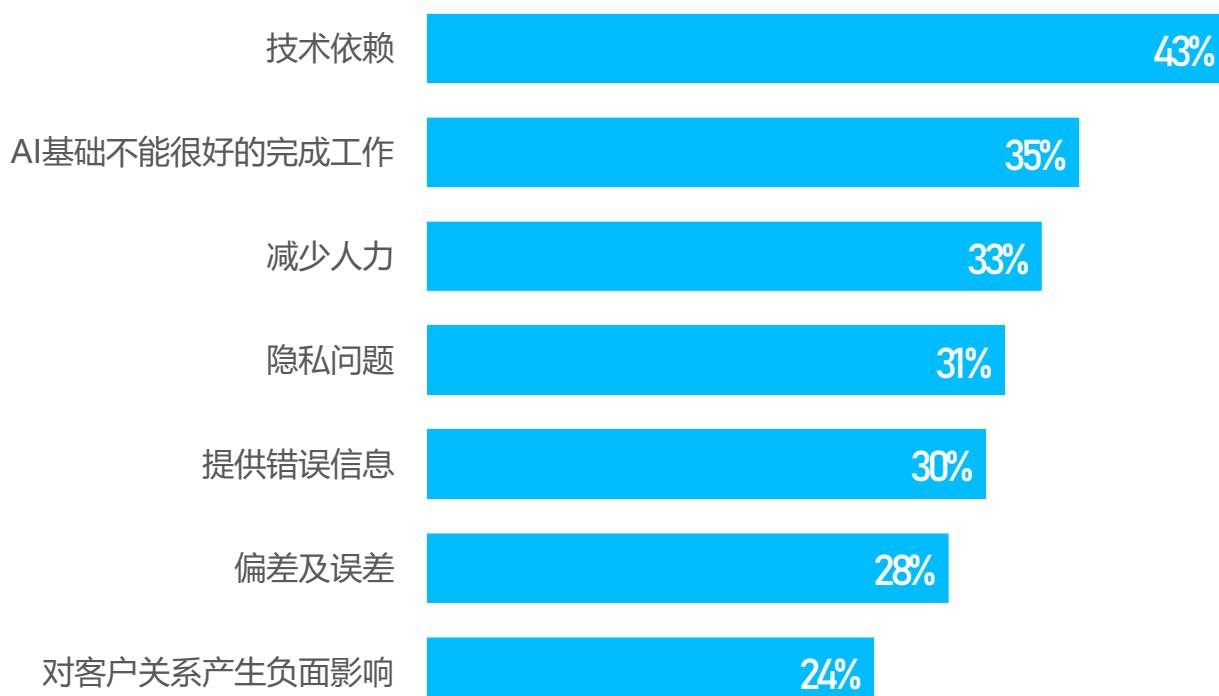
企业主对是否应用人工智能仍存在疑虑，男性用户对新技术的接受程度显著高于女性



Female
44.1%

Male
55.9%

2024年上半年中小企业家对使用人工智能的担忧



超八成用户认为AI可以为其工作带来正面影响，愈七成用户认为AI可以带来新的工作机会

2024年10月全球用户对AI在就业市场应用的观点分析



愿意在工作中使用AI工具的用户比例



使用AI工具的用户中，认为AI工具为自己工作带来正面影响的比例

62.9%

81.4%



认为AI可以大幅优化工作职能的用户比例



认为AI会重塑现有工作内容的用户比例

68.1%

45.4%



认为至少有20%工作岗位会消失的用户比例



认为AI会带来新工作机会的用户比例

59.5%

72.3%



人工智能应用已在企业中快速普及



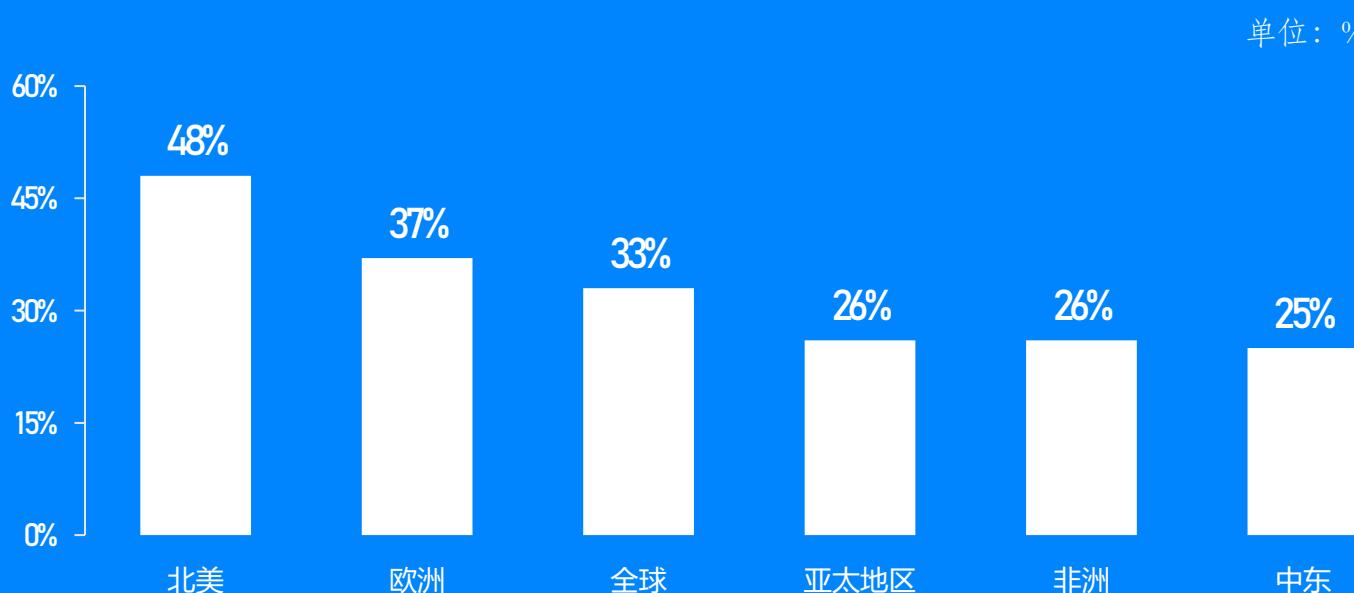
关键发现：

人工智能最常见的应用已经从2023年长远战略次要组成部分，转变为2024年广泛应用并推动企业关键价值增长。

提高产品或服务质量并通过提高IT效率节省成本是开发AI应用程序的主要驱动力。

人工智能正成为许多组织战略的一个基本方面，越来越被视为广泛实施和至关重要的因素。表示人工智能是其组织“更广泛战略的一个次要组成部分”的受访者比例比去年的调查减少了一半，而认为人工智能“广泛实施，推动关键价值”的受访者比例从28%上升到33%，成为最常见的答案。对于北美受访者来说，这一比例甚至更高，为48%，而亚太地区（26%）和欧洲、中东和非洲（25%）则为26%。

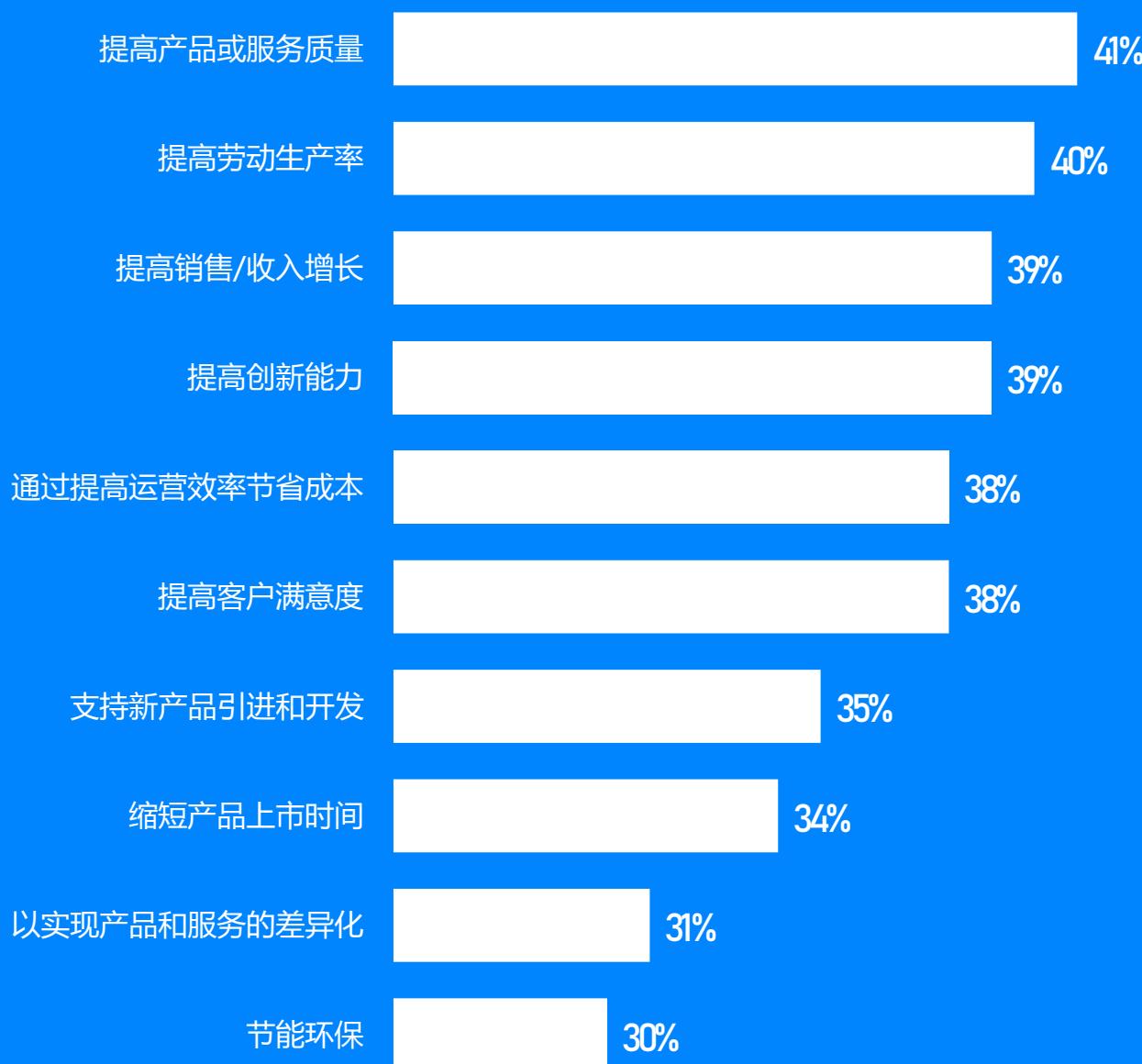
2024年10月全球主要地区认为AI被广泛应用且推动关键价值的用户比例



人工智能的影响力不仅限于实施的广度，还包括该技术的战略影响。从历史上看，人工智能的价值主张与降低成本密切相关。例如，之前人工智能在机器人流程自动化方面的进步与裁员或降低外包成本等目标密切相关。这并不是说人工智能带来的降低成本的机会被挤占了，事实上，通过提高 IT 效率来节省成本是人工智能的第二大目标，而是成本驱动因素正与更具

战略性的目标相结合。例如，在我们的调查中，超过三分之一 (39%) 的受访者将收入增长视为其人工智能计划的主要驱动力。公司不仅试图利用人工智能实现比去年更多的目标，而且他们还看到了与收入驱动因素更清晰的一致性。与去年相比，他们更加意识到人工智能可用于获得产品差异化并缩短上市时间的机会。

2024年人工智能应用开发驱动因素



• 许多人工智能项目无法扩展；遗留的数据架构是罪魁祸首

人工智能日益增长的战略重要性正在推动各企业大幅增加相关计划。广泛的实验和教育是企业应尽的义务，如果不鼓励，企业将失去其应有的职责。然而，缺乏明确价值识别途径的项目正因数据挑战而受到阻碍，从而扼杀这一机会。人工智能项目可能会在有限的部署困境中停滞不前，从而浪费公司的金钱、时间和资源，而无法达到预期的使用水平。数据孤岛、数据质量差以及数据和模型管道效率低下等问题正阻碍这些计划的实施。



关键发现：

在普通组织中，51% 的 AI 项目正在生产但尚未大规模交付。

将 AI 项目引入生产环境时，数据质量是最大的阻碍。

35% 的组织认为，存储和数据管理是 AI 计划最常见的基础设施抑制因素；然而，那些已经广泛实施 AI 的组织对这些挑战的感受不那么强烈。

随着企业投资将 AI 应用于不断增长的目标，企业的项目流程中出现了一个问题。虽然越来越多的计划被集中到 AI 项目团队，但仍有大量计划仅得到部分部署。平均而言，受访企业被归类为有限部署的生产项目多于扩大能力的项目。在追求新计划时，许多组织可能无法最大限度地发挥其现有投资的价值。问题的关键似乎是数据质量和可用性，许多企业组织的遗留数据架构导致这一

流程停滞。

当企业组织将项目从试点阶段推进到生产阶段时，数据质量是最常被提及的挑战。数据质量问题（42% 的组织认为是其三大障碍之一）甚至比技能短缺（32%）和预算限制（31%）更为严重。媒体和娱乐（59%）、高等教育（53%）以及航空航天和国防（48%）等行业的组织尤其强烈地感受到数据质量挑战。

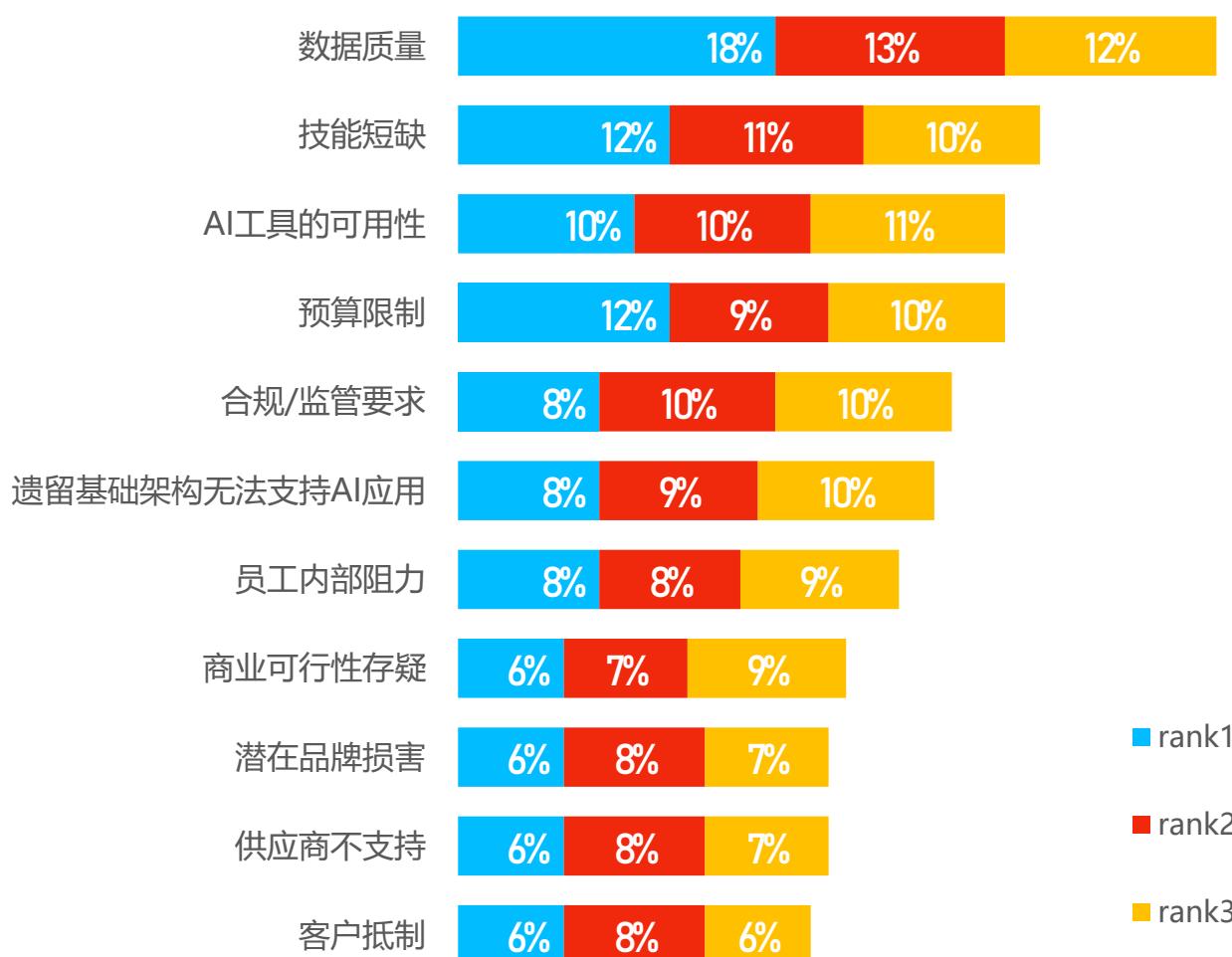
数据质量挑战并非缺乏构建高性能模型的数据，而是数据没有以项目团队可以充分利用的方式进行设置。当被问及将项目转移到生产环境的主要数据挑战时，受访者表示，高质量数据的可用性是比识别相关数据更显著的障碍。34% 的组织认为高质量数据的可用性是三大数据挑战之一，仅次于数据隐私问题 (35%)，显然许多组织在有效数据管理方面的设置不佳。

遗留数据技术似乎是造成这些数据管理缺陷的主要原因。数据管理和存储最常被视为阻碍人工智能应用开发的基础设施组件。超过三分之一 (35%) 的受访者认

为它们比安全 (23%)、计算 (26%) 和网络资源 (15%) 更严重。

值得注意的是，最有效地扩展 AI 计划的组织较少受到这些数据管理和存储组件的限制。在报告其组织内广泛实施 AI 的受访者中，只有 28% 的人认为存储和数据管理挑战是他们最大的阻碍；相反，他们感受到来自网络或计算资源的最大压力。相比之下，42% 的受访者认为 AI 仅限于其组织内的少数用例或项目。大规模实施 AI 的组织似乎专注于投资升级用于存储或管理数据的系统和技术。

组织将AI应用从试点转移到生产环境的三大障碍分析



• 生成式人工智能已迅速超越其他人工智能应用

各组织纷纷投资生成式人工智能，其兴趣超过了对长期存在的人工智能形式的兴趣。随着这波投资热潮的尘埃落定，一小部分生成式人工智能先驱者应运而生。这些组织拥有更广泛的整合能力，并从该技术在新产品开发、增强创新和更快上市时间方面获得了显著的竞争优势。随着生成式人工智能先驱者开始建立与他人之间的显著差距，这些竞争优势可能会不断增长，而这些差距是由他们的投资和基础设施优势决定的。



关键发现：

88% 的组织正在积极研究生成式人工智能。

24% 的受访者已将生成式人工智能视为其整个组织的一项综合能力。

大多数生成式人工智能开拓者都认为，生成式人工智能计划对提高创新率（79%）、支持新产品推出（76%）和缩短产品上市时间（76%）等竞争差异化领域具有“高”或“非常高”的影响。

随着企业投资将 AI 应用于不断增长的目标，企业的项目流程中出现了一个问题。虽然越来越多的计划被集中到 AI 项目团队，但仍有大量计划仅得到部分部署。平均而言，受访企业被归类为有限部署的生产项目多于扩大能力的项目。在追求新计划时，许多组织可能无法最大限度地发挥其现有投

生成式人工智能的采用正在迅速推进。一组开拓者（占 24%）已经将生成式人工智能投资转化为规模化生产能力。相比之下，11% 的公司尚未投资生成式人工智能，29% 的公司仍在试验该技术，37% 的公司已将生成式人工智能投入生产但尚未实现规模化。对于一项直到 2022 年 11 月推出 ChatGPT 后才引起公众关注的技术而言，这是一个了不起的采用水平。

已集成并广泛部署生成式 AI 的组织将获得广泛的好处。重要的是，这些好处通常体现在提供竞争优势的领域。超过四分之三 (79%) 的先驱者认为生成式 AI 对其创新率具有“高”或“非常高”的影响，76% 对其新产品上市时间具有影响，76% 对其支持新产品推出具有影响，74% 对其产品或服务质量的改进具有影响，67% 对其产品和/或服务差异化具有影响。这些水平超过了“AI 成熟度”较低的组织，这表明生成

式 AI 的相对采用可能会决定行业的赢家和输家。未能迅速实施有意义的生成式 AI 项目的组织最终可能会输给那些能够迅速实施的组织。

生成式AI的先驱者在其支持基础设施和策略方面更加成熟。他们使用更广泛的场所进行AI模型训练和推理。但更根本的是，在AI基础设施规划方面，他们考虑的因素要多得多。与没有进行同等程度投资的组织相比，他们更有可能在规划基础设施时考虑安全性、AI加速器访问、数据隐私、可扩展性、客户支持以及对AI工具和框架的访问。与那些尝试生成式AI的人相比，这些先驱者不太可能考虑的唯一因素是前期成本，他们认为前期成本不如长期运营支出重要。通过在基础设施决策之初考虑这些因素，这些组织可以确保这些问题不会在项目进展过程中出现。



• GPU 可用性继续受到限制，影响基础设施决策

AI 加速器在优化 AI 性能方面发挥着重要作用。这些专用硬件设备（最突出的例子是 GPU）旨在加速模型训练和推理；对于 AI 工作负载，它们比 CPU 更快、更高效。组织在访问 GPU 时可能会面临挑战，而这种稀缺性提升了它们在基础设施规划中的地位，并鼓励采用专业的 AI 云计算平台。



关键发现：

在安全性之后，44% 的组织认为加速器可用性是基础设施决策的主要因素。

超大规模公共云是通往 GPU 的一条途径，但许多人也转向了专业的 AI 云。GPU 云正成为训练（近三分之一，即 32% 的组织采用）和推理（31%）的关键场所。

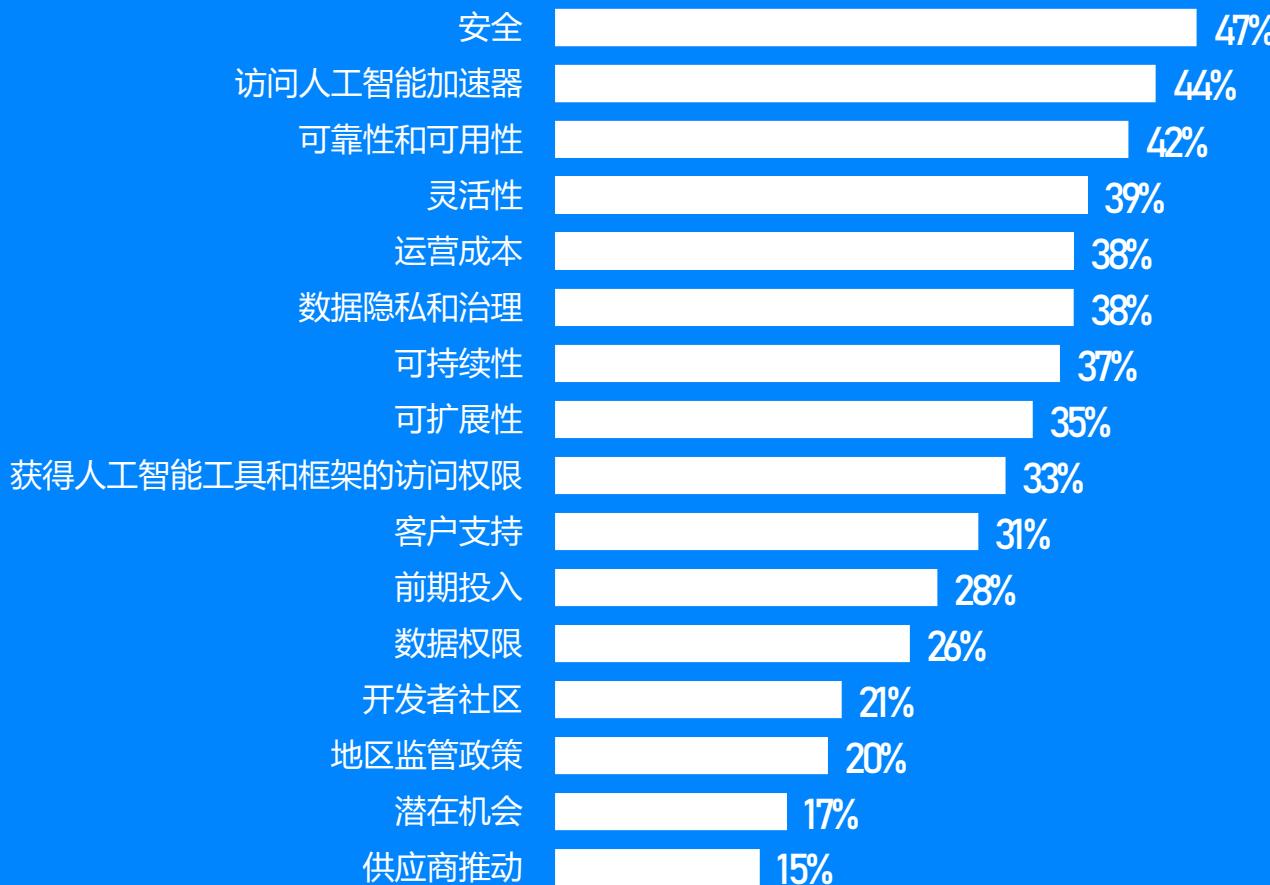
在某些地区，特别是亚太地区，缺乏人工智能加速器已经限制了组织将模型投入生产。

基础设施决策的主要因素与安全性、AI 加速器访问以及可靠性和可用性有关。如图 8 所示，AI 加速器的访问排名很高，甚至超过了运营成本和灵活性等长期关注的领域。电信公司（53%）、高等教育（53%）和制造组织（51%）特别重视这种访问。

超大规模公共云为寻求 GPU 的组织提供了一条重要途径，但它们并不是唯一的选择。虽然超大规模云计算现有

企业代表了最受欢迎的 AI 训练和推理场所，分别有 46% 和 40% 的组织表示，但专业 AI 云已成为辅助甚至替代场所。GPU 云的普及度激增，反映了对 GPU 的高需求。近三分之一（32%）的已投资 AI 的组织正在使用 GPU 云执行训练工作负载，31% 的组织正在执行推理。这些专业云产品在信息技术和服务公司中特别受欢迎，51% 的组织将 GPU 云列为训练场所。

2024年影响组织AI基础设施决策的主要因素分析



随着 AI 开发和部署领域的不断壮大，GPU 云有望进一步增长。各组织预计，未来 12 个月内，训练和推理场所的使用量将会增加，在这种增长环境下，GPU 云的推理和训练采用率预计将增长至 34%。高等教育机构似乎是一个增长特别快的客户群体。我们的数据显示，可扩展性是组织认为 GPU 云将要扮演的主要角色；组织能够轻松且经济高效地管理波动的 AI 工作负载，这是采用

GPU 云的明显驱动因素。

一些国家（包括一些亚太主要经济体）的组织强烈感受到 GPU 可用性挑战；印度、台湾、新西兰和澳大利亚更有可能将 GPU 可用性列为将模型投入生产的三大挑战之一。瑞典（39%）和阿联酋（35%）在这方面也表现突出。





结论：

2024 年全球人工智能趋势报告展现了与 2023 年报告截然不同的人工智能应用前景。人工智能正在得到更广泛的应用，更加注重提供产品和服务质量改进和收入增长。生成式人工智能的成熟是这一转变的关键驱动力。然而挑战依然存在。许多组织正在努力将投资转移到他们可以大规模提供的能力上，他们承认业务运营的可持续性面临压力。

构建强大的数据架构，助力AI成功



- 组织必须建立清晰的途径，将 AI 项目扩展到生产中，确保高效的数据管理和存储。在开展大量试点项目之前，投资建立强大的数据基础至关重要。这将有助于实现无缝 AI 价值交付。

明智的投资是生成式AI成功的关键



- 受益于生成式AI的组织已重新分配预算，以专注于这些计划。成功取决于复杂的决策和强大的基础设施。为了效仿这一点，组织应确保全面的采购实践并最大限度地提高GPU的可用性和利用率，包括研究专门的GPU和AI云服务。

探索生成式AI驱动的IT效率



- 生成式AI可以自动执行常规模型开发任务并改善IT决策，从而推动更精简的交付。这种自我强化的方法可以为更可持续的 AI 路线图奠定基础。

扩大可持续发展实践



- 更换基础设施供应商或修改 AI 项目范围可能会对总体排放量产生重大影响。可持续发展措施的价值在结合使用时会成倍增加，因此组织应授权项目团队采用各种方法。

制定全面的AI战略



- 生成式AI提供了巨大的机遇，但组织应该制定全面的 AI 战略。狭隘的 AI 方法（即未能研究多种技术的混合方法）忽略了将不同类型模型结合在一起的机会，并关闭了许多影响深远的用例。

04

趨勢及展望

The Trend 全球人工智能发展趋势

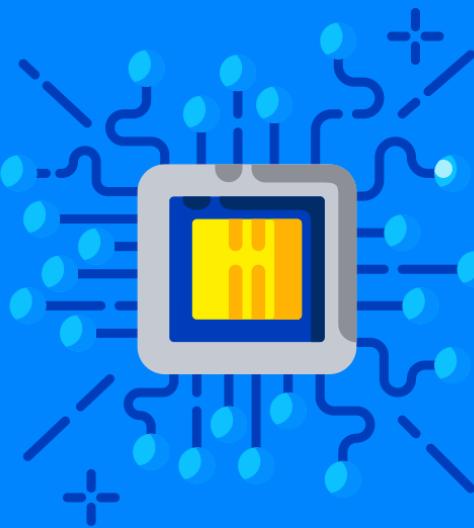
2022年是生成式人工智能(AI)爆发式进入公众视野的一年，2023年是它开始在商业世界扎根的一年。因此，2024年将成为人工智能未来的关键一年，因为研究人员和企业都在寻求确定如何将这一技术的革命性飞跃最实际地融入到我们的日常生活中。

生成式人工智能的发展与计算机的发展如出一辙，只不过速度要快得多。少数公司生产的大型集中式计算机逐渐被企业和研究机构使用的小型高效机器取代。在随后的几十年里，技术逐渐进步，爱好者们可以随意摆弄家用计算机。随着时间的推移，功能强大的个人计算机和直观的无代码界面变得无处不在。

生成式人工智能已经进入了“业余爱好者”阶段——与计算机一样，进一步的进步旨在以更小的成本实现更高的性能。2023年，具有开放许可的基础模型呈爆炸式增长，首先是Meta推

出的LlaMa系列大型语言模型(LLM)，随后是StableLM、Falcon、Mistral和Llama2等。Deep Floyd和Stable Diffusion已与领先的专有模型实现了相对平价和易用。借助开源社区开发的微调技术和数据集，许多开放模型现在可以在大多数基准测试中胜过除最强大的闭源模型之外的所有模型，尽管参数数量要少得多。

随着进步步伐的加快，最先进模型不断扩展的功能将获得最多的媒体关注。但最具影响力的发展可能是那些专注于治理、中间件、训练技术和数据管道的发展，这些发展使生成式人工智能对企业和最终用户来说都更加值得信赖、可持续和易于访问。以下是未来一年值得关注的一些重要当前人工智能趋势。



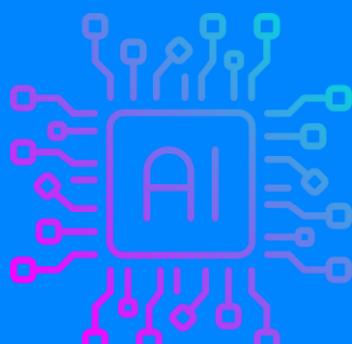
• 现实检验：更现实的期望

当生成式人工智能首次引起大众关注时，典型的商业领袖的知识主要来自营销材料和令人振奋的新闻报道。实际经验（如果有的话）仅限于摆弄ChatGPT和DALL-E。现在尘埃落定，商界现在对人工智能解决方案有了更深入的了解。有机构将用技术成熟度曲线将生成式人工智能定位在“预期膨胀的顶峰”，即将滑入“幻灭的低谷”——换句话说，即将进入一个（相对）令人失望的过渡期——而德勤2024年第一季度的“企业生成式人工智能状况”报告指出，许多领导者“预计短期内将产生重大的变革影响”。而现实情况可能介于两者之间：生成式人工智能提供了独特的机会和解决方案，但它并不会满足所有人的需求。

现实世界的结果与炒作相比如何，在一定程度上取决于观点。像ChatGPT这样的独立工具通常占据大众想象的焦点，但顺利集成到现有服务中往往会产生更大的持久力。在当前的炒作周期之

前，生成机器学习工具（如谷歌在2018年推出的“智能撰写”功能）并没有被视为范式转变，尽管它们是当今文本生成服务的先驱。同样，许多影响深远的生成式AI工具正在作为企业环境的集成元素实施，它们增强和补充现有工具，而不是彻底改变或取代它们：例如，Microsoft Office中的“Copilot”功能、Adobe Photoshop中的“生成填充”功能或生产力和协作应用程序中的虚拟代理。

生成式人工智能在日常工作流程中首先获得发展势头的地方，将对人工智能工具的未来产生比任何特定人工智能功能的假设优势更大的影响。根据IBM最近对1,000多名企业级公司员工进行的一项调查，推动人工智能采用的三大因素是人工智能工具的进步使其更易于访问、降低成本和自动化关键流程的需求，以及嵌入到标准现成业务应用程序中的人工智能数量不断增加。

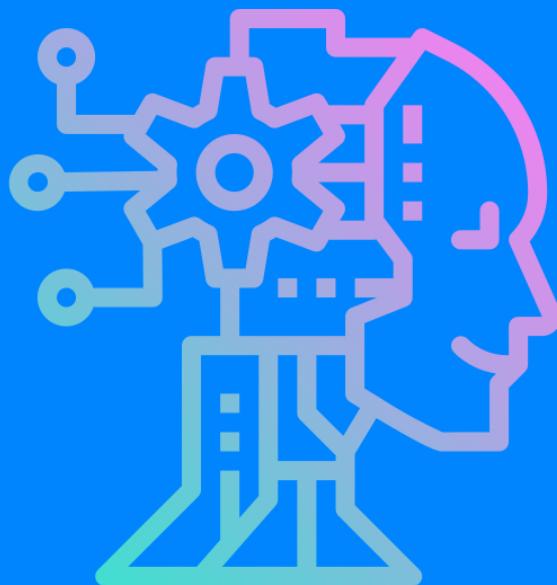


• 多模态人工智能将获得快速发展

最先进的生成式人工智能的野心正在不断增长。下一波进步不仅将侧重于提高特定领域的性能，还将侧重于能够将多种类型的数据作为输入的多模态模型。虽然跨不同数据模态运行的模型并不是一个严格意义上的新现象——文本到图像模型（如CLIP）和语音到文本模型（如Wave2Vec）已经存在多年——但它们通常只在一个方向上运行，并且经过训练以完成特定任务。

新一代跨学科模型包括OpenAI的GPT-4V或Google的Gemini等专有模型，以及LLaVa、Adept或Qwen-VL等开源模型，它们可以在自然语言处理(NLP)和计算机视觉任务之间自由切换。新模型还将视频纳入其中：2024年1月底，Google宣布推出Lumiere，这是一种文本转视频的传播模型，它还可以执行图像转视频的任务或使用图像作为风格参考。

多模态人工智能最直接的好处是更加直观、功能多样的人工智能应用和虚拟助手。例如，用户可以询问图像并收到自然语言答案，或者大声询问修复某些东西的说明，并获得视觉辅助和分步文本说明。从更高层次来看，多模态人工智能允许模型处理更多样化的数据输入，从而丰富和扩展可用于训练和推理的信息。尤其是视频，为整体学习提供了巨大的潜力。



• 小型语言模型和开源的持续进步

在特定领域的模型中，我们可能已经达到了参数数量增加带来的收益递减点。OpenAI 首席执行官 Sam Altman (GPT-4 模型有大约 1.76 万亿个参数) 在 2023 年 4 月麻省理工学院的“想象力在行动”活动上提出了同样的观点：“我们认为我们正处于一个巨型模型时代的终结，我们会以其他方式让它们变得更好，”他预测道。

“我认为人们过于关注参数数量。”

大规模模型开启了人工智能黄金时代，但它们并非没有缺点。只有最大的公司才有资金和服务器算力来训练和维护具有数千亿个参数的耗能模型。根据华盛顿大学的一项估计，训练一个 GPT-3 大小的模型需要消耗 1,000 多户家庭一年的用电量；标准的 ChatGPT 查询日消耗的电量相当于 33,000 个美国家庭一天的用电量。

与此同时，较小的模型所需的资源要少得多。Deep mind 于 2022 年 3 月发表了一篇颇具影响力的论文，该论文表明，在更多数据上训练较小的模型比在较少数据上训练较大的模型能产生更好的性能。因此，LLM 中正在进行的大部分创新都集中在从较少的参数中产生更大的输出。正如 30 亿到 700 亿参数范围内的模型的最新进展所证明的那样，特别是 2023 年基于 LLaMa、LLaMa2 和 Mistral 基础模型构

建的模型，模型可以在不牺牲太多性能的情况下缩小规模。

开源模型的力量将不断增强。2023 年 12 月，Mistral 发布了“Mixtral”，这是一个集成了 8 个神经网络的混合专家(MoE)模型，每个神经网络都有 70 亿个参数。Mistral声称，Mixtral 不仅在大多数基准测试中以 6 倍的推理速度超越了 Llama2 的 70B 参数变体，而且在大多数标准基准测试中甚至匹敌或超越了 OpenAI 规模更大的 GPT-3.5。此后不久，Meta 发布的 Llama3 模型，它依然开源。较小模型的进步具有三个重要优点：

它们有助于实现人工智能的民主化：可以在更易获得的硬件上以更低成本运行的小型模型使更多的业余爱好者和机构能够研究、训练和改进现有模型。

它们可以在较小的设备上本地运行：这允许在边缘计算和物联网(IoT)等场景中实现更复杂的 AI。此外，在本地运行模型（例如在用户的智能手机上）有助于避免与敏感的个人或专有数据交互而产生的许多隐私和网络安全问题。

• GPU 短缺和高算力成本还将持续

随着硬件可用性下降导致云计算成本增加，小型模型的趋势将受到必要性和创业活力的双重驱动。大公司（以及更多大公司）都在尝试将 AI 功能引入内部，GPU 也出现了一定程度的抢购。这不仅会给 GPU 产量带来巨大压力，还会迫使创新者想出更便宜、更易于制造和使用的硬件解决方案。正如 2023 年末那样，云提供商目前承担了大部分计算负担：相对较少的 AI 采用者维护自己的基础设施，而硬件短缺只会增加设置内部服务器的障碍和成本。从长远来看，这可能会给云成本带来上行压力，因为提供商将更新和优化自己的基础设施以有效满足生成式 AI 的需求。对于企业来说，驾驭这种不确定的环境需要灵活性，包括模型（在必要时依靠更小、更高效的模型，或在实际情况下依靠更大、更高性能的模型）和部署环境。

• 模型优化变得越来越容易

开源社区最近的成果很好地满足了最大化更紧凑模型性能的趋势。许多关键进步不仅受到（并将继续受到）新基础模型的推动，还受到用于训练、调整、微调或调整预训练模型的新技术和资源（如开源数据集）的推动。2023 年流行值得注意的与模型无关的技术包括：

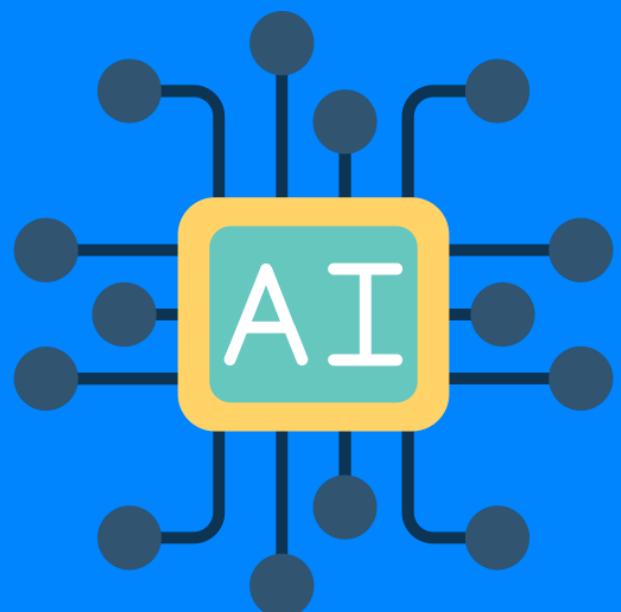
低秩自适应 (LoRA): LoRA 不是直接微调数十亿个模型参数，而是需要冻结预先训练的模型权重并在每个 Transformer 块中注入可训练层（将模型权重变化矩阵表示为 2 个较小的（低秩）矩阵）。这大大减少了需要更新的参数数量，进而大大加快了微调速度并减少了存储模型更新所需的内存。

• 定制的本地化模型

2024 年的企业可以通过定制模型开发来追求差异化，而不是围绕“大人工智能”重新包装的服务构建包装器。有了正确的数据和开发框架，现有的开源人工智能模型和工具可以适应几乎任何现实场景，从客户支持用途到供应链管理再到复杂的文档分析。

开源模型让组织有机会快速开发强大的自定义 AI 模型（使用专有数据进行训练并根据特定需求进行微调），而无需昂贵的基础设施投资。这在法律、医疗保健或金融等领域尤其重要，因为基础模型在预训练中可能没有学习过高度专业化的词汇和概念。

法律、金融和医疗保健也是可以从足够小的模型中受益的典型行业，这些模型可以在适中的硬件上本地运行。将 AI 训练、推理和检索增强生成 (RAG) 保持在本地可避免专有数据或敏感个人信息被用于训练闭源模型或以其他方式通过第三方之手的风险。使用 RAG 访问相关信息而不是将所有知识直接存储在 LLM 本身中有助于减小模型大小，进一步提高速度并降低成本。随着 2024 年模型竞争环境继续趋于公平，竞争优势将越来越多地由能够实现行业最佳微调的专有数据管道所驱动。



• 更强大的虚拟代理

凭借更加复杂、高效的工具以及一年的市场反馈，企业已准备好扩展虚拟代理的用例，而不仅仅是简单的客户体验聊天机器人。

随着人工智能系统加速并整合新的信息流和格式，它们不仅扩大了通信和指令遵循的可能性，还扩大了任务自动化的可能性。2023年是能够与人工智能聊天的一年。多家公司推出了一些产品，但交互方式一直是你输入一些内容，然后它回复一些内容，到2024年，我们看到代理能够为您完成工作。预订、计划旅行、连接到其他服务。

尤其是多模态人工智能，大大增加了与虚拟代理无缝交互的机会。例如，用户不必简单地向机器人询问食谱，而是可以将摄像头对准打开的冰箱，并请求使用现有食材制作的食谱。Be My Eyes 是一款移动应用程序，可将盲人和视力低下人士与志愿者联系起来，帮助他们完成快速任务，该应用程序正在试用人工智能工具，帮助用户通过多模态人工智能直接与周围环境互动，而无需等待人类志愿者。

• 监管、版权和人工智能道德问题

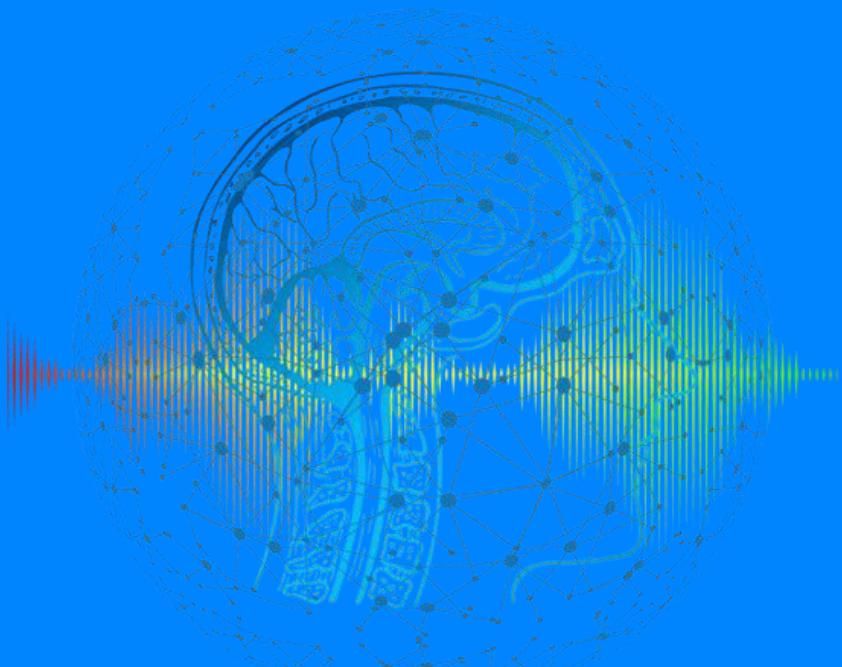
多模态能力的提升和准入门槛的降低也为滥用打开了新的大门：对于不良行为者来说，深度伪造、隐私问题、延续偏见甚至逃避 CAPTCHA 保障措施可能会变得越来越容易。2024年1月，一波名人深度伪造风潮席卷社交媒体；2023年5月的研究表明，与2022年同期相比，网上发布的语音深度伪造数量增加了8倍。

监管环境的模糊性可能会在短期至中期内减缓采用，或至少减缓更积极的实施。对新兴技术或做法进行任何重大、不可逆转的投资都存在固有风险，这些投资可能需要在未来几年内根据新立法或不断变化的政治逆风进行重大调整，甚至成为非法行为。

2023年12月，欧盟(EU)就《人工智能法案》达成临时协议。除其他措施外，该法案禁止不加区别地抓取图像以创建面部识别数据库、可能存在歧视性偏见的生物特征分类系统、“社交评分”系统以及将人工智能用于社会或经济操纵。该法案还试图定义一类“高风险”人工智能系统，这些系统有可能威胁安全、基本权利或法治，将受到额外监督。同样，该法案还为所谓的“通用人工智能(GPAI)”系统(基础模型)设定了透明度要求，包括技术文档和系统对抗测试。

但是，尽管Mistral等一些关键参与者位于欧盟，但大多数突破性的人工智能发展都发生在美国，而私营部门的人工智能实质性立法将需要国会采取行动——这在选举年可能不太可能实现。拜登政府发布了一项全面的行政命令，详细列出了联邦机构使用人工智能技术的150项要求；该政府获得了知名人工智能开发商的自愿承诺，以遵守某些信任和安全准则。值得注意的是，加利福尼亚州和科罗拉多州都在积极寻求有关人工智能个人数据隐私权的立法。

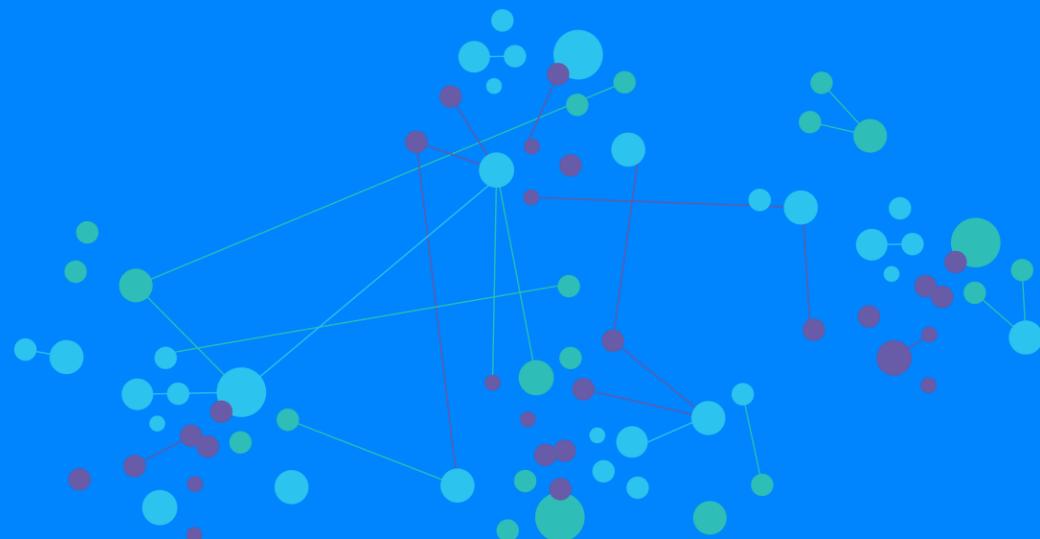
中国已更积极地对人工智能进行正式限制，禁止社交媒体上推荐算法的价格歧视，并要求对人工智能生成的内容进行清晰的标注。未来对生成人工智能的监管旨在要求用于培训大语言模型的训练数据和随后由模型生成的内容必须“真实准确”，专家们已采取这些措施来审查大预言模型的产出。



• 影子人工智能

对于企业而言，生成式人工智能工具的普及和易用性加剧了法律、监管、经济或声誉方面的潜在影响。组织不仅必须制定谨慎、连贯且清晰的生成式人工智能企业政策，还必须警惕影子人工智能：员工在工作场所“非正式”地使用人工智能。影子人工智能也被称为“影子 IT”或“BYOAI”，当急躁的员工寻求快速解决方案（或只是想以比谨慎的公司政策允许的速度更快的速度探索新技术）时，就会在工作场所实施生成式人工智能，而无需经过IT部门的批准或监督。许多面向消费者的服务（有些是免费的）甚至允许非技术人员即兴使用生成式人工智能工具。安永的一项研究表明，90%的受访者表示他们在工作中使用人工智能。这种进取精神在真空中可能很棒——但热切的员工可能缺乏与安全、隐私或合规相关的信息或观点。这可能会使企业面临巨大风险。例如，员工可能会在不知情的情况下将商业机密提供给面向公众的AI模型，该模型会不断根据用户输入进行训练，或者使用受版权保护的材料来训练专有的内容生成模型，从而使公司面临法律诉讼。和许多正在进行的发展一样，这凸显了生成式人工智能的危险性几乎随着其能力的提高而线性上升。能力越大，责任越大。

随着我们进入人工智能的关键一年，理解和适应新趋势对于最大限度地发挥潜力、最大限度地降低风险和负责任地扩大生成性人工智能的采用至关重要。





数据说明

移动端数据：通过SDK的形式获取用户移动端APP使用数据。包括但不限于频次、时长、浏览路径、订单、移动支付等维度数据的收集，上报、存储及统计分析。

PC端数据：针对特定类型平台进行不同维度及口径的数据抓取、数据结构化处理、存储及统计分析。

宏观数据：来源渠道主要包括Wind、choice、彭博、各国相关统计机构、国际组织、第三方数据机构等。

统计周期：报告最新数据截止日期为2024年5月31日。

研究对象：本报告着重研究全球AI发展历史、现状及发展趋势。

免责声明：本报告基于独立、客观、实事求是的分析研究，但不对任何机构及个人，构成投资及其他决策建议，不分享相关收益，也不承担相关责任。

