



# State of AI Report

# 2023人工智能现状报告

2023年10月

Air Street Capital

## 前言

人工智能（AI）是科学和工程的多学科领域，其目标是创造智能机器。

我们相信，在日益数字化、数据驱动的世界中，人工智能将成为技术进步的力量倍增器。这是因为今天我们周围的一切，从文化到消费品，都是智能的产物。

**Air Street**已连续第六年发布人工智能现状报告。我们把这份报告视为所见过的最有趣的事情的汇编，目的是引发一场关于人工智能现状及其对未来影响的知情对话。

报告中考虑了以下主要方面：

- **行业**：人工智能的商业应用领域及其商业影响。
- **研究**：技术突破及其能力。
- **政治**：人工智能的监管，其经济影响和人工智能地缘政治的演变。
- **安全**：识别和减轻高性能未来人工智能系统可能给我们带来的灾难性风险。
- **预测**：我们认为未来12个月会发生什么，以及保持我们诚实度的2022年绩效评估。

由 **Air Street Capital 团队**制作。腾讯科技（微信公众号：qqtech）进行了整理汉化，内容有删减。关注腾讯科技微信公众号（qqtech），回复“AI2023”可免费获取本报告PDF版。

## 划重点



### 行业 (第6页-第57页)

- 随着各国政府、初创公司、大型科技公司和研究人员对GPU的贪婪需求，英伟达迈入了1万亿美元的市值俱乐部。
- 出口管制限制了对中国的先进芯片销售，但主要芯片供应商开发出不受出口管制的替代品。
- 在ChatGPT的带领下，GenAI应用在图像、视频、编码、语音或CoPilot方面都取得了突破性的一年，推动了180亿美元的风险投资和企业投资。

### 研究 (第58页-第115页)

- GPT-4落地并展示了专有和次佳开源替代方案之间的能力鸿沟，同时也验证了从人类反馈中强化学习的力量。
- 用更小的模型、更好的数据集、更长的上下文来克隆或击败专有模型性能的努力越来越多...由LLaMa-1/2提供支持。
- 目前还不清楚人类生成的数据可以维持人工智能扩展趋势多久（有人估计，到2025年，大型语言模型将耗尽数据），以及添加合成数据的影响是什么。锁定在企业中的视频和数据可能是下一个目标。
- 大型语言模型和扩散模型通过为分子生物学和药物发现带来新的突破，继续为生命科学领域提供礼物。
- 多模态成为新的前沿，所有参与方的兴奋感大幅增长。

### 政治 (第116页-第127页)

- 全球已经划分出明确的监管阵营，但全球治理的进展仍然缓慢。最大的人工智能实验室正在填补这一真空。
- 芯片战争有增无减。
- 人工智能预计将影响一系列敏感领域，包括选举和就业，但尚未看到显著影响。

### 安全 (第128页-第145页)

- 生存风险辩论首次成为主流，并显著加剧。
- 很多高性能的机型很容易‘越狱’。为了补救RLHF的挑战，研究人员正在探索替代方案，例如自我校准和根据人类偏好进行预训练。
- 随着能力的提高，对SOTA模型进行一致的评估变得越来越困难。只有共鸣是不够的。



# 回顾Air Street 2022年的预测



## 我们对2022年的预测

## 结果

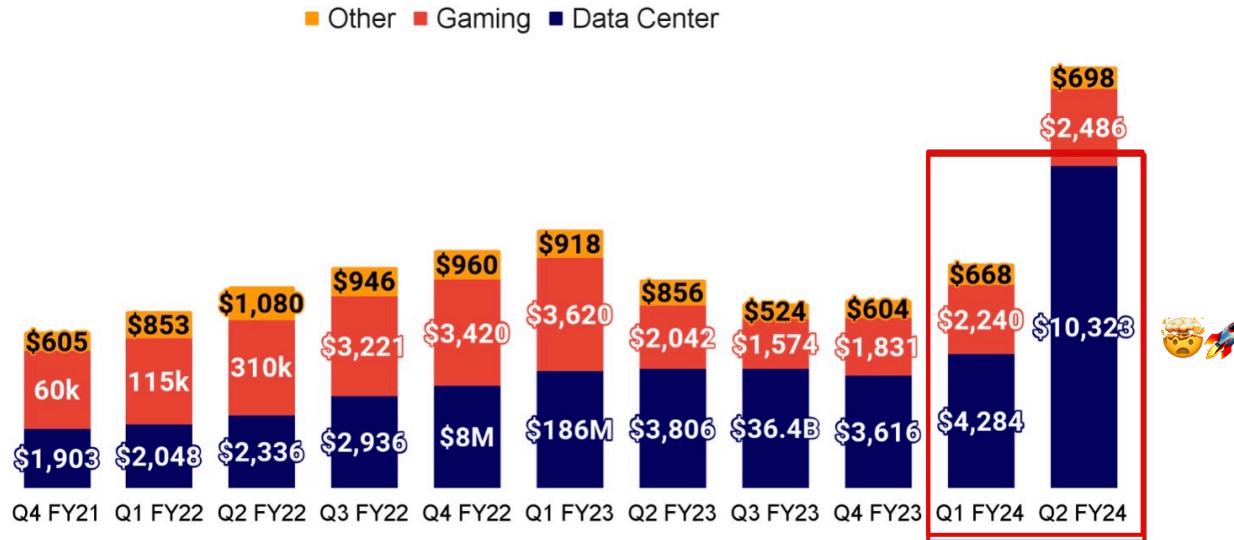
1	DeepMind训练一个具有10B参数的RL模型，比Gato大10倍。	到目前为止，还没有公开披露过这方面的研究。
2	英伟达宣布与一家专注于AGI的组织建立战略关系。	英伟达没有建立这种关系，而是在许多专注于AGI的组织中加大了投资活动，包括Cohere、In Fliection AI和Adept。
3	SOTA语言模型在比Chinchilla多10倍的数据点上训练，证明了数据集缩放与参数缩放。	我们不确定，但据报道，GPT-4是在13T tokens上训练的，而Chinchilla是14T tokens上训练的。Meta的Llama-2是在2T tokens上训练的。
4	到2023年9月，生成音频工具的出现吸引了超过10万名开发人员。大型科技企业GAFAM向通用人工智能或开源人工智能公司投资超过10亿美元。	自推出以来，ElevenLabs和Resemble.ai都声称拥有超过100万用户。2023年1月，微软又向OpenAI投资了100亿美元。
5	面对英伟达的主导地位，半导体初创公司面临现实，一家明星初创公司破产或以低于其最近估值50%的价格被收购。	有降价，但没有大规模停工或低迷的收购。
6	监管生物安全实验室（BSL）等通用人工智能实验室的提案得到了当选的英国、美国或欧盟政治家的支持。	要求监管的呼声明显提高，但对BSL的支持还没有。
7	随着我们意识到让人工智能能力领先于安全所面临的风险，明年将有超过1亿美元投资于专门的人工智能校准组织。	人工智能研究和安全公司Anthropic在2023年9月筹集了高达40亿美元的资金。
8	一家主要的用户生成内容网站（如Reddit）与一家人工智能模型（如OpenAI）初创公司协商商业合作，以便对其用户生成内容语料库进行培训。	OpenAI已经获得了访问其他Shutterstock训练数据（图像、视频和音乐库以及相关元数据）的6年许可。



# 第一章：行业

## GPU需求刺激英伟达营收井喷，市值进入万亿美元俱乐部

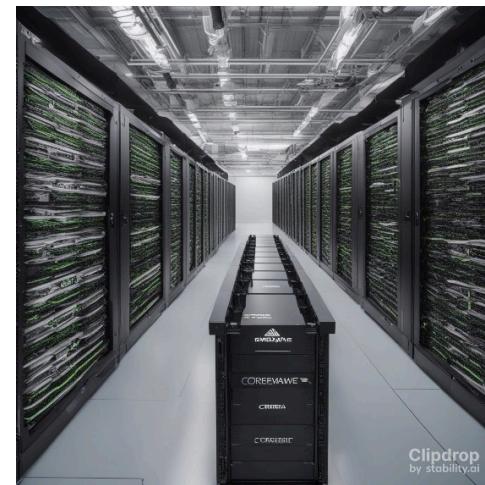
▶ 英伟达2023年第二季度营收达到创纪录的103.2亿美元，比第一季度增长141%，比一年前增长171%。尽管该公司2022年营收达到270亿美元，比2021年增长61.4%，但市场对该股曾持悲观态度。英伟达现在的市值为1万亿美元，比10年前的85亿美元高出116倍。



## 比Coachella卖得更快：从新贵基础设施提供商手中抢购GPU

CoreWeave和Lambda是两家选定的英伟达合作伙伴，负责构建和运行GPU数据中心，它们总共有数万颗GPU。Lambda在其点播云中提供了价值9位数美元的H100s，并在一个多小时内销售一空。CoreWeave是市场上最大的GPU运营商之一。该公司今年年底的建造时间表已经排满，正在签订2024年第一季度的合同。

First, our scale. We have over 45,000 high-end NVIDIA GPUs available on-demand in our fleet. It's not necessarily the volume that makes this significant, but rather the access it provides. Businesses rely on CoreWeave Cloud to run the compute intensive workloads that allow them to deliver client projects, hit deadlines, and accommodate end-user demand. Having a partner like NVIDIA ensures that we're able to provide the scale of resources that our clients need.



Clipdrop  
by stability.ai

# 私营公司正在支持英伟达GPU，并将其作为竞争优势

## Inflection

Along with its partners CoreWeave and NVIDIA, Inflection AI is building the largest AI cluster in the world comprising 22,000 NVIDIA H100 Tensor Core GPUs. In just over a year, Inflection AI has developed one of the most sophisticated large language models in the market to enable people to interact with Pi, your Personal AI ([pi.ai](#)), in the most simple, natural way and receive fast, relevant and helpful information and advice.

## ANTHROPIC

Anthropic estimates its frontier model will require on the order of  $10^{25}$  FLOPs, or floating point operations — several orders of magnitude larger than even the biggest models today. Of course, how this translates to computation time depends on the speed and scale of the system doing the computation; Anthropic implies (in the deck) it relies on clusters with “tens of thousands of GPUs.”



Through the partnership, Cohere will train, build, and deploy its generative AI models on OCI. OCI is uniquely positioned to run AI workloads as it delivers the highest performance and lowest cost GPU cluster technology, with scale of over 16K H100 GPUs per cluster, and very low latency and the highest bandwidth RDMA network in the cloud. This will enable the acceleration of large language models (LLM) training while simultaneously reducing the cost.



- Models. We pretrain our own very large (>100B parameter) models, optimized to perform well on internal reasoning benchmarks. Our latest funding round lets us operate at a scale that few other companies are able to: our ~10,000 H100 cluster lets us iterate rapidly on everything from training data to architecture and reasoning mechanisms.

## 算力是海湾国家的新石油？



据称，沙特阿拉伯阿卜杜拉国王科技大学(Kaust)购买了3000多颗H100s来建造超级计算机Shaheen III。这台超级计算机将于2023年底投入运行。

与此同时，阿联酋马斯达尔市的技术创新研究所开发了Falcon LLM，据说也从英伟达采购计算资源。最后，总部位于阿布扎比的G42与总部位于美国的Cerebras达成协议，购买该公司价值高达9亿美元的晶圆级计算系统，并建造9台互联的人工智能超级计算机。未来可能会有更多的支出...

本报告由腾讯科技整理汉化，内容有删减。关注腾讯科技微信公众号（qqtech），回复“AI2023”免费获取PDF版。



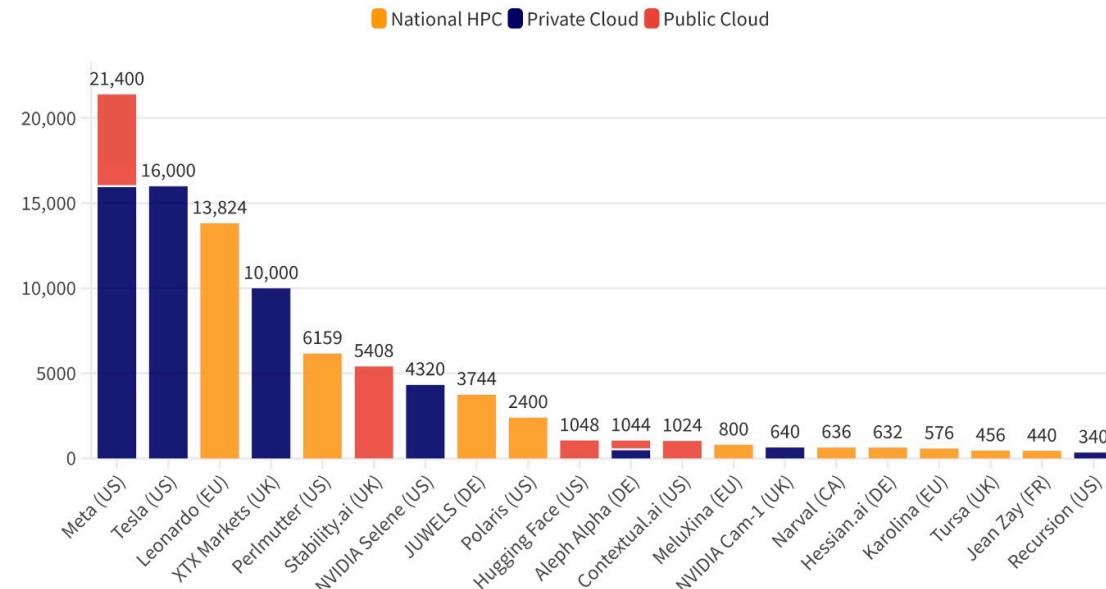
Cerebras and G42 Unveil  
World's Largest  
Supercomputer for AI  
Training with 4 exaFLOPs to  
Fuel a New Era of Innovation

Launching today with its first of nine interconnected AI supercomputers, the Condor Galaxy system will reach a combined AI training capacity of 36 exaFLOPs

## 计算指数：英伟达 A100集群



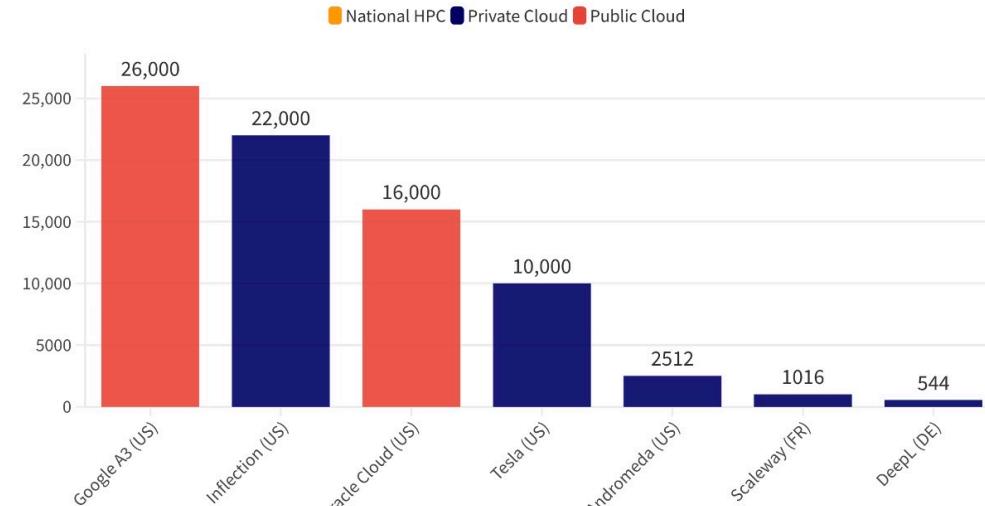
▶ 自去年以来，大规模英伟达 A100 GPU集群的数量一直在增长，特别是特斯拉和Stability，以及Hugging Face的新集群。



Source: [State of AI Report Compute Index](#)

## 计算指数：英伟达H100集群

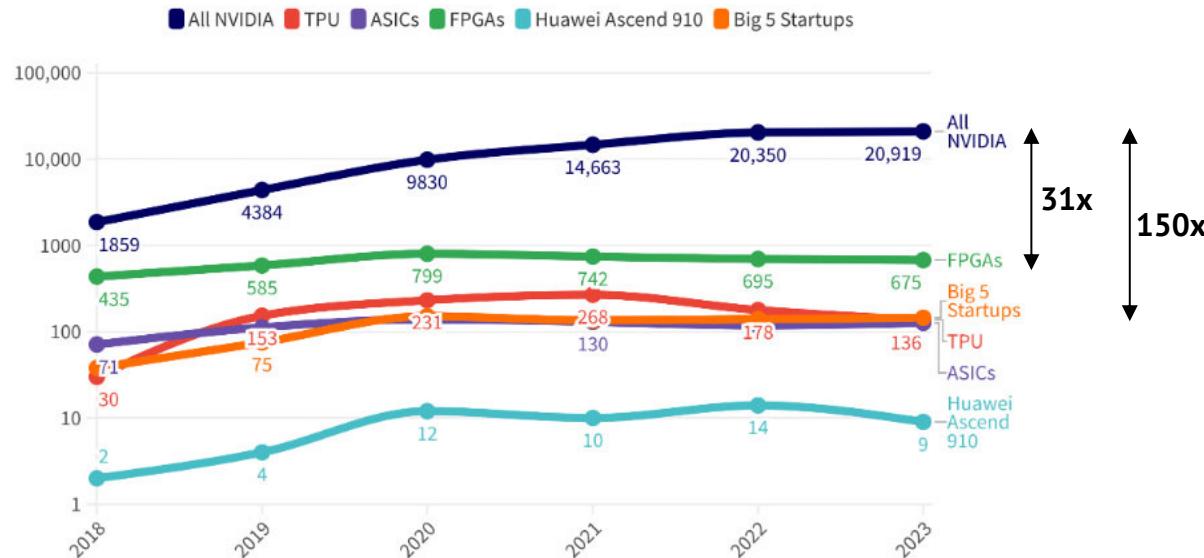
▶ 现在还为时尚早，但私营和上市公司正在宣布新的H100基础设施，用于大规模模型培训。截至10月中旬，谷歌和Inflection尚未全面发展，我们知道其他公司包括OpenAI、Anthropic、Meta、Character.ai、Adept、Imbue等都有很大的能力。我们预计不久会有更多的产品上线。



Source: [State of AI Report Compute Index](#)

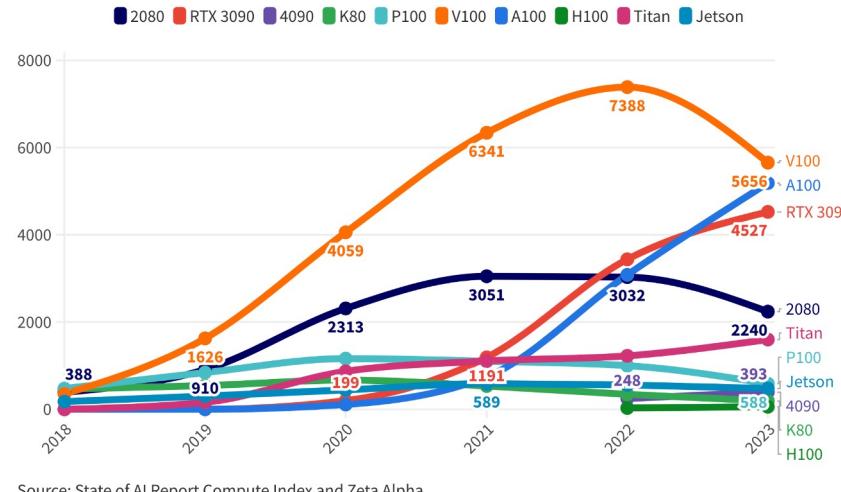
## 人工智能研究论文中使用的英伟达芯片比所有替代芯片的总和多19倍

▶ 在去年的报告中，我们开始跟踪人工智能研究论文中特定半导体的利用情况。我们发现，英伟达芯片被引用的次数远远多于替代品。2023年，英伟达GPU更受欢迎：比FPGAs多31倍，比TPUs多150倍。



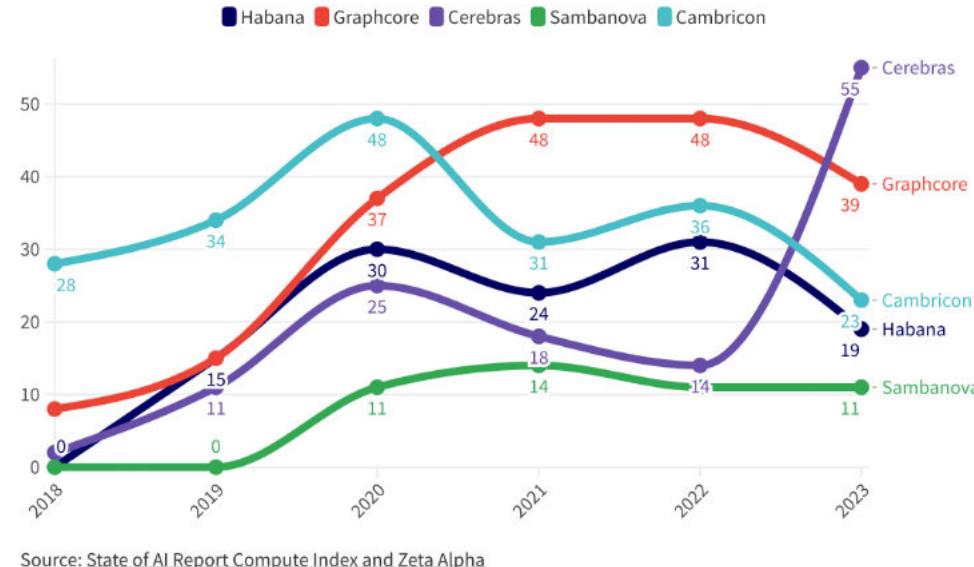
## 英伟达芯片具有非常长的生命周期价值：从上市到流行达到顶峰需要5年时间

2023年，所有的目光都集中在英伟达新推出的H100 GPU上，它是A100更强大的后继者。虽然H100集群正在建设中（并非没有障碍），但研究人员依赖于V100、A100和RTX 3090。英伟达产品的竞争寿命相当惊人：2017年发布的V100目前仍然是人工智能研究中最常用的芯片。这表明，2020年发布的A100可能在2026年达到峰值，而V100可能会达到低谷。因此，新款H100可能会伴随我们直到下一个十年！



## 英伟达是王者，但Cerebras在挑战者当中崭露头角

▶ 全球最大的人工智能芯片的创造者Cerebras，参与了若干个开源模型训练和数据集创建项目，这帮助它比竞争对手更受研究人员的欢迎。但总体上，英伟达的竞争者还有很长的路要走。

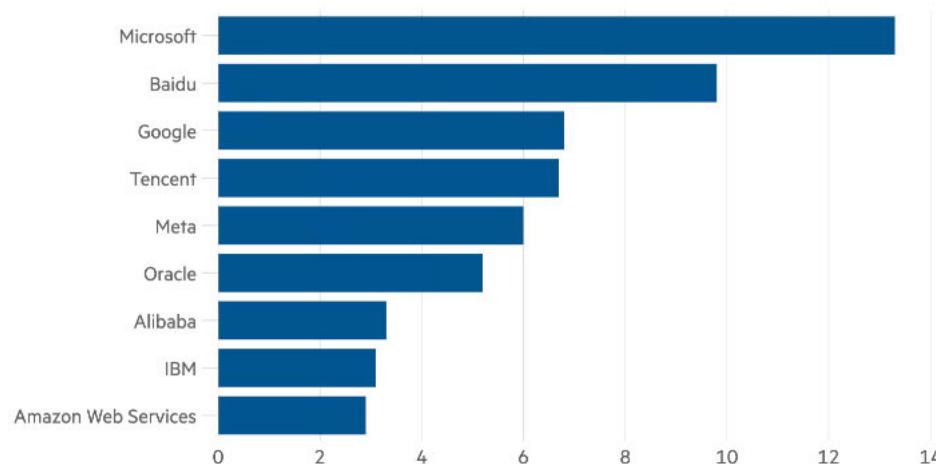


## 超大规模企业将提高人工智能支出所占总资本支出的比例

▶ 有传闻称，英伟达将在2024年出货150万至200万颗H100，高于今年预计的50万颗。

Global cloud service providers' AI spending

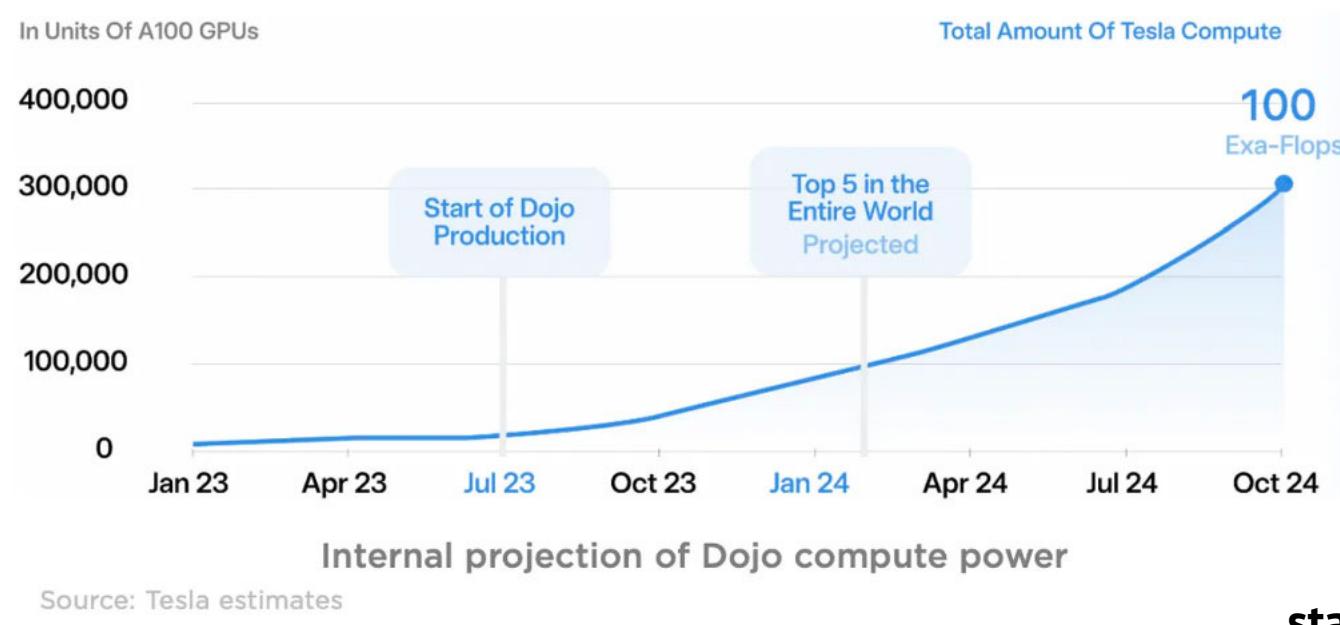
As a % of total capex, 2023



Source: Counterpoint Research  
© FT

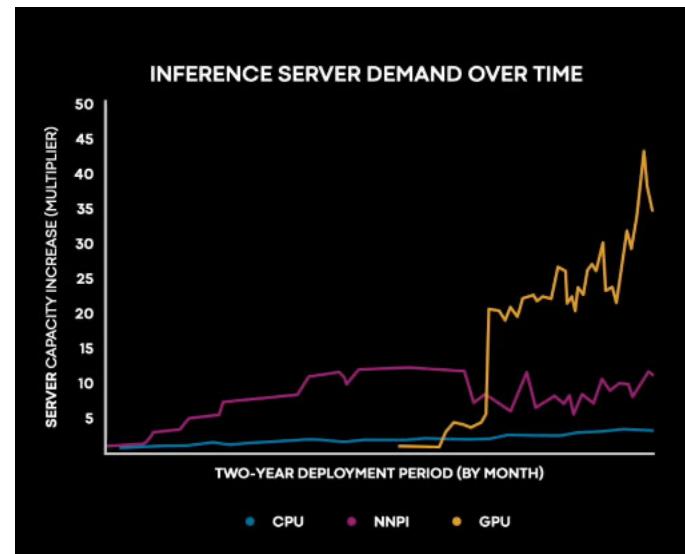
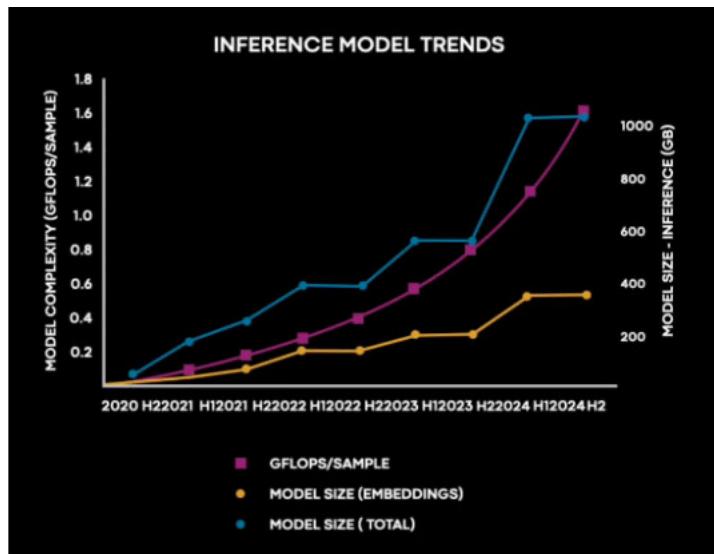
## 特斯拉迈向全球前五大人工智能计算集群

▶ 在我们2022年的计算指数中，特斯拉基于其100 GPU计数排名第四。截至2023年夏天，该公司推出了一个由1万颗H100组成的集群，成为迄今为止最大的在线集群之一。



## 越来越多的超大规模公司为内部人工智能工作流开发自己的推理硬件

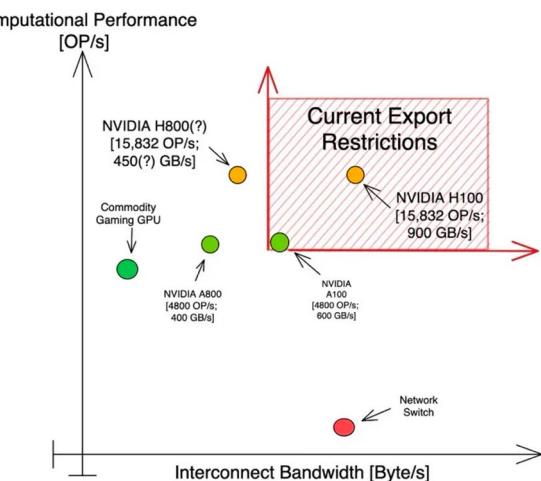
▶ Meta发布了MTIA，这是该公司第一个基于开源RISC-V架构的内部加速器，可以满足基于深度学习的推荐模型的要求。这是由生产中部署的模型不断增长的规模和复杂性以及GPU提供的缓慢推理速度所驱动的。



## 英伟达、英特尔和AMD制造向中国出口的非管制芯片

根据英伟达首席财务官的说法，中国过去占英伟达数据中心相关产品营收的20-25%(金融时报)。英伟达(及其竞争对手)开发了低于出口清单阈值的芯片。

- 2022年8月下旬，英伟达的A100和H100—该公司在人工智能应用方面最强大的芯片—被列入美国商务部的出口管制清单。到当年11月，英伟达已经开始宣传A800和H800芯片，其设计低于美国禁令设定的性能阈值。
- 英特尔对他们的Habana Gaudi 2芯片的新版本做了类似的调整，AMD也表达了类似的意图。
- 因此，中国互联网大公司已经订购价值超10亿美元美元的英伟达A800/H800 GPU。也有报道称中国的A100/H100 GPU流量有所增加，但规模要小得多。



## 在出售给英伟达的交易被阻后，软银旗下的Arm重新在纳斯达克上市

回到2020年，我们预测英伟达将无法完成对Arm的收购。今年9月，Arm在纳斯达克重新上市，开盘时市值达到600亿美元。

- Arm的知识产权支撑着全球99%的智能手机芯片，该公司正在努力重新定位自己在人工智能市场的角色。它已同自动驾驶汽车公司Cruise和英伟达合作开发了Grace Hopper芯片(其技术在其中扮演了配角)。
- 然而，这不会一帆风顺。该公司营收与上一财年持平，25%来自Arm中国，这是进入中国市场所需的独立子公司。
- 考虑到其巨大的市场份额，Arm可能有潜力提高其每台设备的专利费，但需要与不断增长的开源替代架构(如RISC-V)进行平衡。
- 由于Arm不销售实物芯片，迄今为止未受任何影响，但无法保证这种情况会持续下去。



## 2022年预测：生成式人工智能应用越来越受欢迎

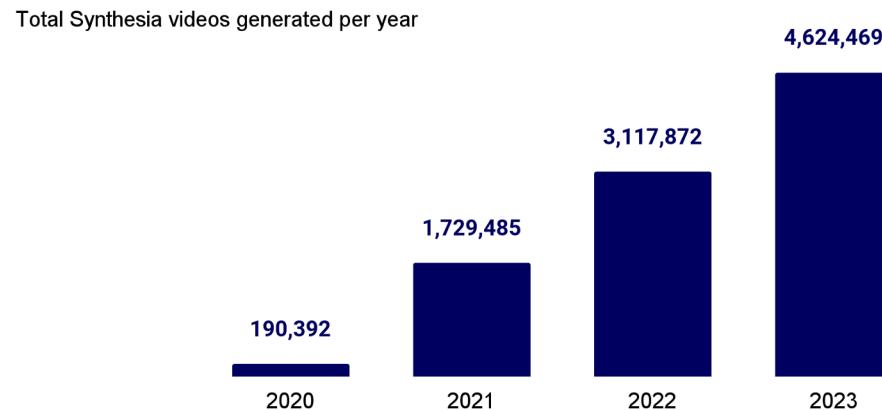
► 我们在2022年预测：“到2023年9月，生成式音频工具将吸引超过10万名开发人员。”ElevenLabs（英国）和Lemble AI（美国）都超过了这个门槛。另一个领域，产品设计，正在见证生成式人工智能技术的快速整合，这有利于像Uizard这样的快速发展的公司。

- ElevenLabs现在有超过200万注册用户，并且增长迅速。该公司获得第二个百万用户的速度比第一个百万用户快了一倍。用户累计上传了超过10年的音频内容。ElevenLabs最初面向创作者和出版商，现在正在适应来自人工智能代理、伴侣、娱乐和游戏的大量用例。
- 由人工智能工具驱动的产品设计公司Uizard表示，截至7月23日，该公司录得320万美元的ARR（年度经常性收入），同比增长13倍。该公司4月份的ARR突破了100万美元，3个月内从100万美元增至300万美元。



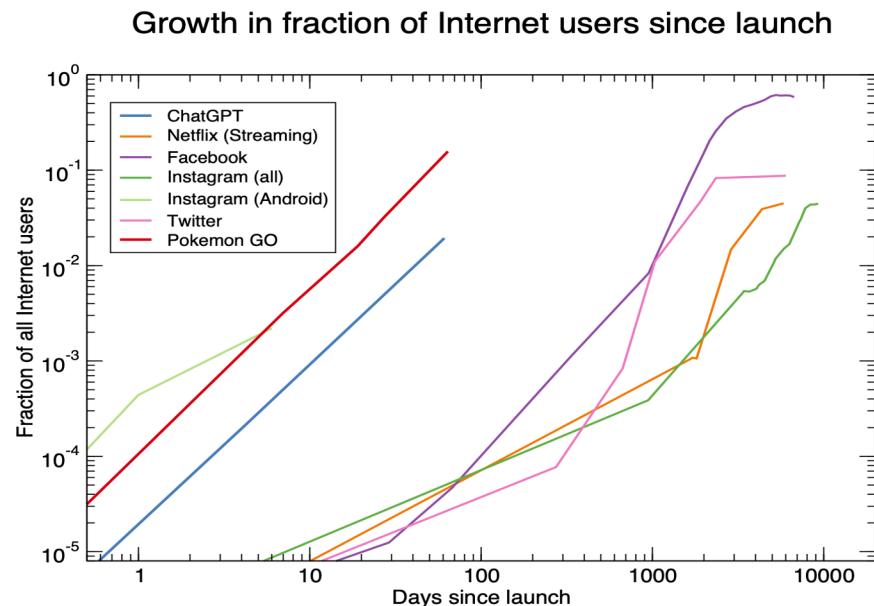
## 2022年预测：生成式人工智能应用越来越受欢迎

▶ 视频也是GenAI快速发展的前沿领域。总部位于伦敦的Synthesia成立于2017年，于2020年推出了人工智能优先的视频创作器。该系统生成多语言化身，该化身制定供消费者和企业等使用的脚本。曾经被认为是“边缘”的Synthesia现在被44%的财富100强企业用于学习和发展、市场营销、销售支持、信息安全和客户服务。自2020年推出以来，这项服务已经产生了超过960万个视频。



# OpenAI的ChatGPT是发展最快的互联网产品之一

横轴为产品推出时间



## OpenAI当前的赚钱能力惊人.....但代价是什么？

▶ 12个月之前，OpenAI在筹集100亿美元资金时所做的营收预测曾遭到了很多质疑。如今，该公司正在超越其目标。这会持续多久？代价是什么？



EXCLUSIVE MICROSOFT AI Published 13 hours ago

### OpenAI Passes \$1 Billion Revenue Pace as Big Companies Boost AI Spending

 By Amir Efrati and Aaron Holmes

Aug. 29, 2023 3:58 PM PDT

 Share article



EXCLUSIVE STARTUPS AI

### OpenAI's Losses Doubled to \$540 Million as It Developed ChatGPT

 By Erin Woo and Amir Efrati

May 4, 2023 1:11 PM PDT · Comments by Josh Bersin, Brian Shilhavy, and 7 others

 Share article

## 感受ChatGPT的热度：教育首当其冲，Chegg正在反击

▶ Chegg是一家在纽交所上市的公司，专注于改善学生的学习和学习成果，因ChatGPT的推出而受到重创。该公司在2023年5月表示：“今年上半年，我们没有看到ChatGPT对我们新账户增长的明显影响，我们正在满足新注册的预期。”付钱给Chegg来练习考试并获得作业反馈的学生转而求助于ChatGPT。结果，Chegg的股价暴跌逾40%。在2023年8月举行的财报电话会议上，Chegg表示：“我们已经让公司转向利用人工智能来更好地为学习者服务。”他们正在与Scale AI合作构建内部大型语言模型。

Market Summary > Chegg Inc

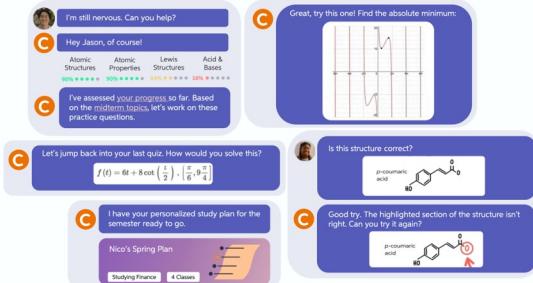


### Building and Owning our own Large Language Models

- Enhances our competitive moat, lowers our costs, and allows us to train the models specifically for education
- Leverages our billions of pieces of proprietary content
- 150k subject matter experts help train the models and support accuracy in our generative experience
- We expect a significantly enhanced learning experience over generic models and tremendous value for Chegg

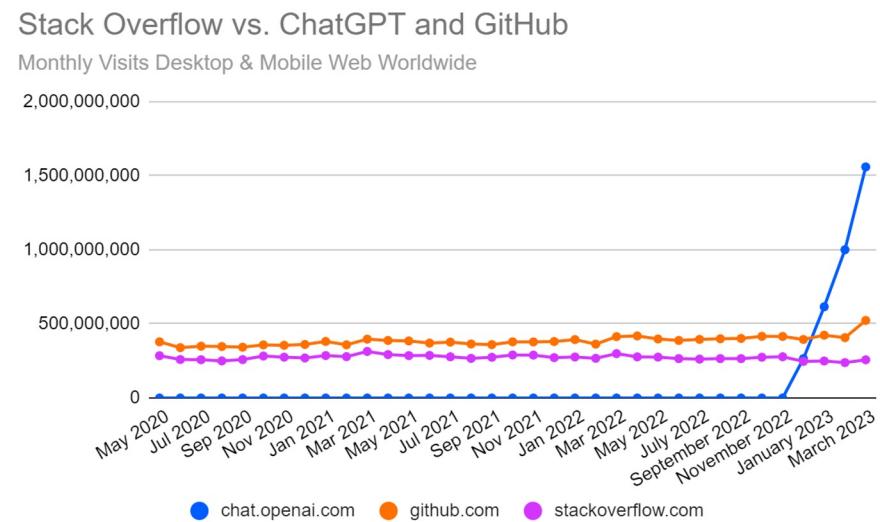
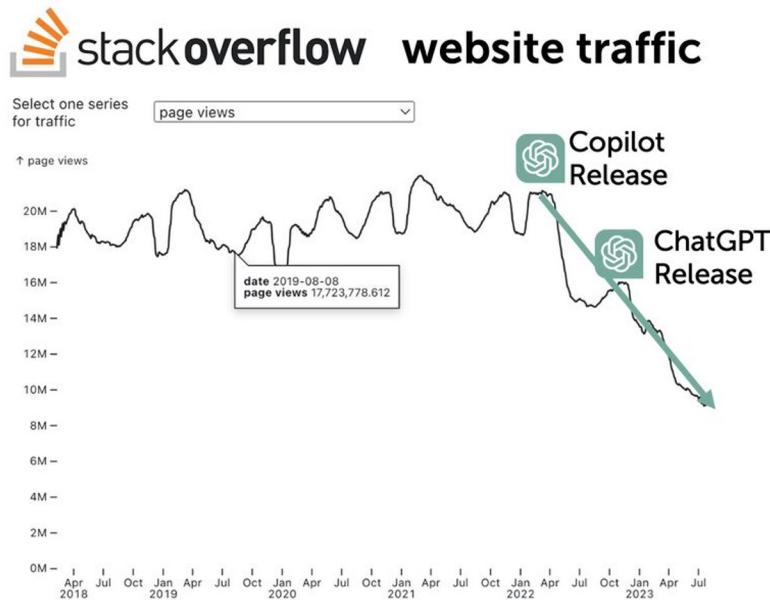
### Accelerated Timeline

- Our partnership with Scale AI will allow us to accelerate our ability to deliver the new Chegg experience starting in the fall and rolling out over the course of the next two semesters.
- The experience will include a simple conversational user interface, personalized learning pathways, more in-depth content, and the ability to transform content into innovative study tools, such as practice tests, study guides, and flash cards.



## 感受ChatGPT的热度：编程是下一个被革命的领域.....开发人员很喜欢它！

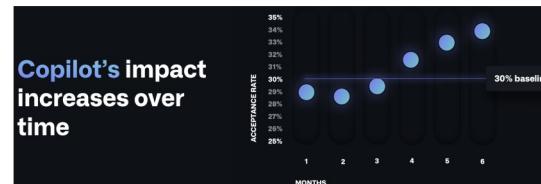
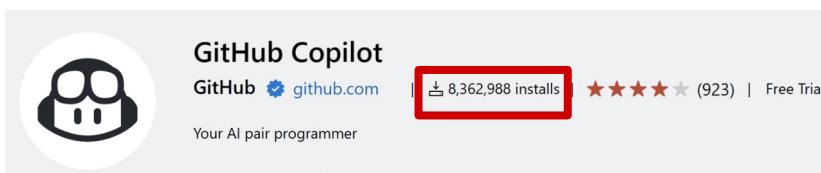
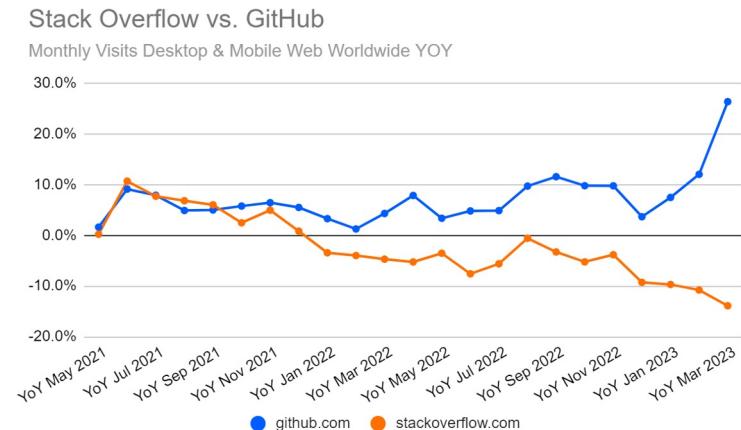
▶ Stack Overflow在人工智能热潮之前供开发者寻找他们编程问题的解决方案。该网站由于ChatGPT的流行而遭受流量损失，已禁止开发者在Stack Overflow上发布ChatGPT生成的文本。



## 结果是：GitHub CoPilot显著提高了开发人员的工作效率

如果是命中注定的，那就一定会是(不管要花多长时间)。GitHub终于推出了他们的编程助手CoPilot，获得了巨大的好评。这个系统是在数十亿行代码上训练出来的。

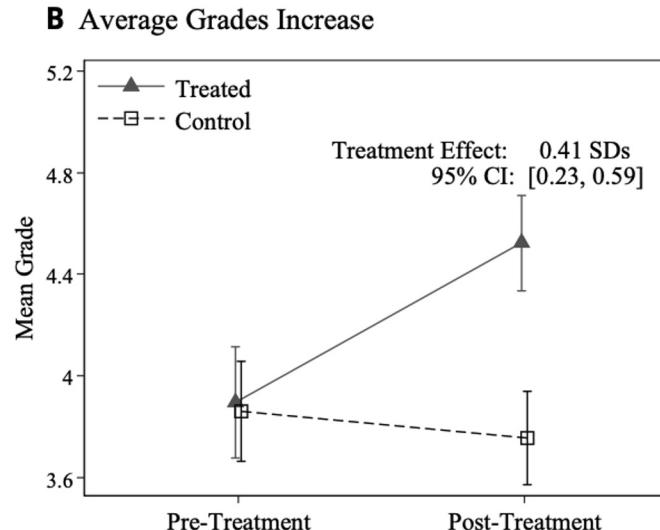
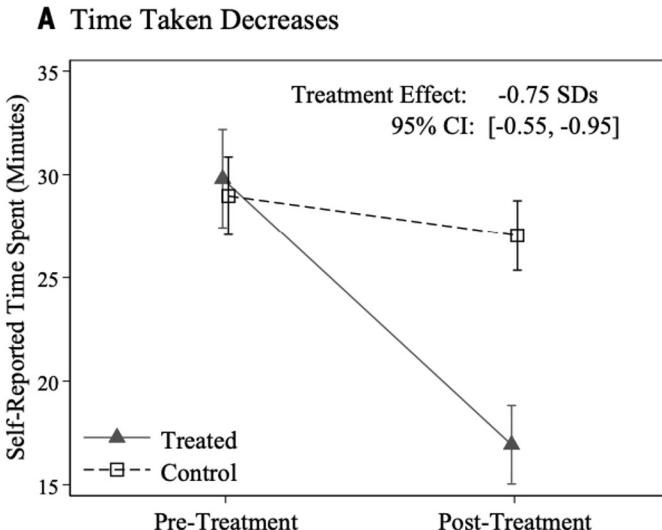
- 2022年9月，GitHub对95名专业开发人员进行了一项实验，将他们随机分成两组，并记录他们用JavaScript编写一个HTTP服务器需要多长时间。这发现了显著的生产率提高。
- 2023年6月，GitHub报告了934533名CoPilot用户的数据。有趣的是，随着Copilot用户熟悉该工具，生产率在显著提高之前略有下降，经验较少的用户受益最大(生产率提高约32%)。



## ChatGPT能提升写作产出

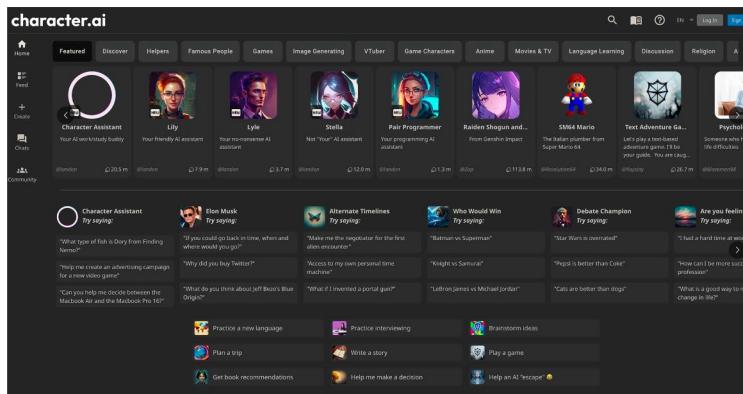


▶ 麻省理工的一项新研究支持流行观点：ChatGPT有助于写作，特别是对于“中级专业写作”。研究表明，与对照组相比，使用ChatGPT的写作者完成任务的时间减少了40%，输出质量提高了18%。



# 某些不太明显的GenAI用例也获得了显著的吸引力

▶ 我们已经看到消费者对用户与定制聊天机器人进行互动的巨大兴趣。A16z支持的Character.AI筹集了1.5亿美元的A轮融资，在推出其应用程序之前，其网站的月访问量达到2亿次。它们的许多用途是良性的一一例如，它们被用作语法工具或在小说社区中使用，但我们也看到了商业和伦理上的挑战。我们已经看到用户对它们的机器人产生情感依赖的报告，公司努力在露骨内容的受欢迎程度及其对其品牌的影响之间进行权衡。



**SCIENCE**

## Replika users fell in love with their AI chatbot companions. Then they lost them

ABC Science / By technology reporter James Purtill  
Posted Tue 28 Feb 2023 at 7:00pm

**The AI companion who cares**

**The AI companion who cares**

**The AI companion who cares**

How are you feeling today?  
How are you doing today?  
What are you up to today?

**Disrupted**

**AI chatbot company Replika restores erotic roleplay for some users**

By Anna Tong  
March 25, 2023 11:45 PM GMT - Updated 6 months ago

**TECH**

## Fascist chatbots are running wild on Character.AI

The world's second-biggest AI chat service is a hotbed of hateful and racist abuse with no meaningful moderation

## 文本到图像的模式：竞争加剧，整合比比皆是

在2022年随着Stable Diffusion的发布而突破性的一年之后，Midjourney和Stability仍然在不断改进它们的模型。尽管在文本到图像方面似乎反应较慢，OpenAI还是发布了迄今为止最好的文本到图像模型DALL-E 3。还有像Ideogram这样的新进入者，该公司的创始人是谷歌Imagen的开发者——他们的模型特别会拼写。与此同时，我们在流行产品中看到了无数的文本到图像模型的集成，最显著的是Adobe的Firefly、Photoroom甚至Discord。

- Midjourney的收入在2022年3月MRR（月持续性收入）已达到100万美元，预计在2023年将达到2亿美元。其用户数量从200万同比增长至1480万。值得注意的是，Midjourney集成在Discord中，用户可以在Discord服务器上生成图像。据Discord称，每月有超过3000万人在其服务器上使用人工智能应用程序，创建了超过10亿个独特的图像。
- 专门从事照片编辑的法国初创公司Photoroom表示，随着2月份推出生成式人工智能服务，该公司在过去6个月里的收入和用户数量翻了一番。



Stability's SDXL



OpenAI's DALL-E 3



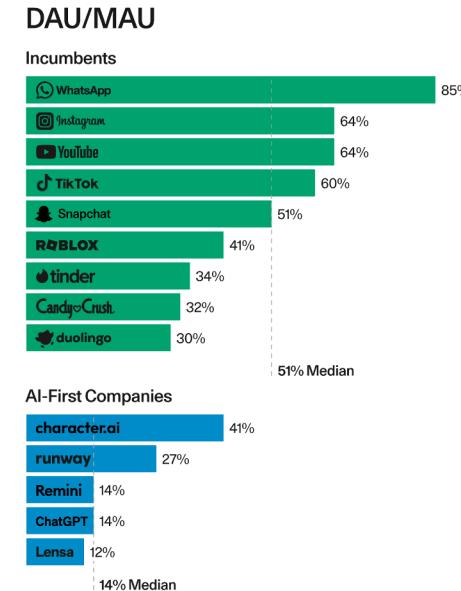
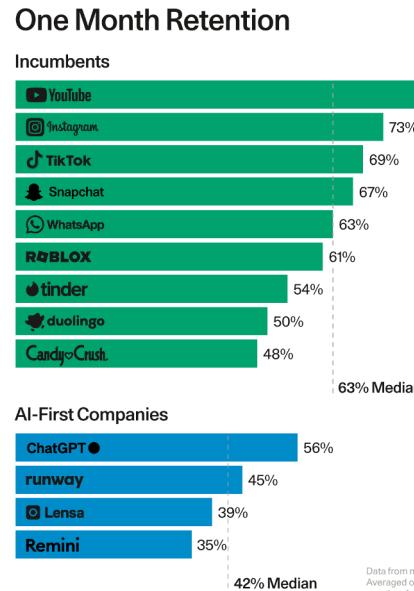
Midjourney v5.2



Ideogram v0.1

## 但是GenAI的惊艳效果（到目前为止）不足以让用户留下来...

与YouTube、Instagram、Tiktok或WhatsApp等最受欢迎的现有应用相比，ChatGPT、Runway或Character.ai等GenAI应用的平均留存率和日活跃用户数仍然较低。



## 2022年预测：一家主要的用户生成内容网站与一家初创公司谈判商业解决方案，开发人工智能模型(像OpenAI一样)，用于在他们的语料库上进行训练

▶ 2022年10月，领先的股票多媒体提供商Shutterstock宣布将与OpenAI合作，将DALL-E驱动的内容引入该平台。然后在2023年7月，两家公司签署了为期6年的内容许可协议，该协议将允许OpenAI访问Shutterstock的图像、视频和音乐库以及用于模型训练的相关元数据。此外，Shutterstock将为客户提供人工智能图像创作的赔偿。该公司还与Meta签订了GenAI的内容许可协议。这种支持GenAI的立场与Shutterstock的竞争对手Getty Images形成鲜明对比，后者强烈反对GenAI，这一点从其在2023年2月对Stability AI提起的版权侵权诉讼中可以看出端倪。



VS.

### NATURE OF ACTION

1. This case arises from Stability AI's brazen infringement of Getty Images' intellectual property on a staggering scale. Upon information and belief, Stability AI has copied more than 12 million photographs from Getty Images' collection, along with the associated captions and metadata, without permission from or compensation to Getty Images, as part of its efforts to build a competing business. As part of its unlawful scheme, Stability AI has removed or altered Getty Images' copyright management information, provided false copyright management information, and infringed Getty Images' famous trademarks.

## 2022年预测：一家主要的用户生成内容网站与一家初创公司谈判商业解决方案，开发人工智能模型(像OpenAI一样)，用于在他们的语料库上进行训练

▶ 2023年7月，OpenAI和美联社达成了许可协议，允许部分访问美联社自1985年以来的新闻故事。与此同时，美联社将获得OpenAI技术和产品专业知识，以探索生成式应用。尽管美联社没有基于大型语言模型的应用程序，但它已经利用人工智能系统来编辑自动化的企业财报和体育赛事新闻摘要。

[Home](#) / [Press Releases](#) / [2023](#)

### AP, Open AI agree to share select news content and technology in new collaboration

July 13, 2023

SHARE



PRINT

The Associated Press and OpenAI have reached an agreement to share access to select news content and technology as they examine potential use cases for generative AI in news products and services.

## 美国法院开创了人工智能生成的内容不适合版权保护的先例，但随后又是一个关于合理使用的先例

▶一家美国地区法院重申了一项长期存在的原则，即版权保护需要人类作者身份。虽然上诉是可能的，但重要的先例现在可能已经确立。

- 美国哥伦比亚特区的地区法院驳回了斯蒂芬·泰勒（Stephen Thaler）的主张，即2012年的图像《近访仙境之门》(右图)值得版权保护。
- 美国版权局已经提出倡议，检查人工智能对版权法的影响，并发布了新的版权指导，涵盖文学，视觉，视听和声音。它规定，任何艺术品都需要人类作者，应用程序需要指定人工智能的使用场景。
- 对提供商来说更具挑战性的是，在2023年5月对1981年王子肖像版权案的裁决中，美国最高法院应用了一种新的、更严格的解释，解释什么构成合理使用下的“转化”。这很可能使得让模型训练数据搜集书籍和艺术品的行为在法律上更具风险。



## 但在多个司法管辖区，侵犯版权的案件仍在继续

▶ 以主要文本和图像生成为特征的案件正在英国和美国进行。虽然这些公司声称他们在从事公平使用或言论自由，但有迹象表明麻烦可能就在前面。

- Getty Images在英国和美国起诉Stability，称后者从自己收藏中复制了数百万张照片，更改或删除了版权信息，并指控Stable Diffusion生成的图像带有Getty Images水印的修改版本。
- OpenAI和Meta正面临诉讼，声称不同意版权书籍被ChatGPT和LLaMa用于训练数据集。据报道，《纽约时报》正考虑对OpenAI提起类似的诉讼。三名艺术家正在起诉Stability、DeviantArt和Midjourney，因为三家公司使用他们的作品来训练图像生成器，从而创建“侵权衍生作品”。
- 英国有版权法的文本和数据挖掘豁免，但这只延伸到非商业用途；扩大这一豁免的计划已被搁置。欧盟也有类似的豁免，但人工智能法案规定，基础模型提供商必须提供用于训练其模型的版权材料的摘要(这可能在技术上具有挑战性)
- 微软已经向Copilot工具的用户保证，如果出现任何版权索赔，公司将承担任何法律风险。

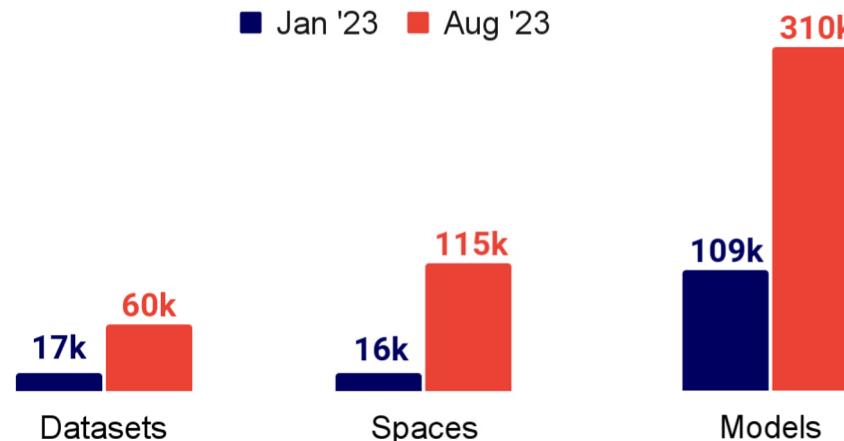
## 从标签到偏好

随着指令微调和RLHF成为微调和对齐语言模型的默认方法，提供标签服务的公司，如Scale AI和Surge HQ，将从大型语言模型的爆炸式流行中获得非凡的增长。两家公司都支持令人印象深刻的客户名单，从人工智能初创公司到大型企业客户，再到大型语言模型研究领域的领先实验室。Scale AI上一次估值为73亿美元是在2021年，发生在Stable Diffusion和ChatGPT狂潮之前。



## 在现任者推动闭源人工智能之时，开源人工智能正在蓬勃发展

Hugging Face是一家已成立7年的公司，目前已经牢牢占据开源人工智能的龙头地位。随着社区争相让所有人都可以访问人工智能模型和数据集的过程，Hugging Face 正在飞速增长。过去几个月，超过1300个模型被提交到Hugging Face的开放式大型语言模型排行榜（Open LLM Leaderboard），仅在2023年8月就有超过6亿次模型下载。这些模型作为使用Gradio或Streamlit等工具构建的网络应用程序展示在Spaces上，支持更广泛的可访问性和快速原型制作。Gradio的月活跃用户数增长了5倍，从12万（2013年1月）增长到58万（2013年8月）。



## 单一的大型语言模型还是依赖于应用程序的专业化大型语言模型？

▶ Databricks以13亿美元的价格收购了MosaicML，帮助其建立（很可能是微调）自己的大型语言模型。未来不是一个无所不知的单一整体模型，而是属于一组根据企业数据或特定任务训练的专业模型。

- 在收购之前，Mosaic展示了令人印象深刻的工程技术，如从零开始以低于5万美元（比原来减少8倍）的成本培训Stable Diffusion，以及构建具有长上下文长度的SOTA LLMs。
- 这笔交易是生成式人工智能狂热短暂历史中的一个标志性重要时刻。
- Snowflake也有类似的策略：与微软Azure一起，向客户提供OpenAI模型的访问权限。



## 曾经被大型制药公司忽视的人工智能正走向一些公司的前沿和中心

► mRNA疫苗领导者BioNTech以5亿英镑的价格收购了人工智能公司InstaDeep，而赛诺菲在人工智能上“孤注一掷”，默克与人工智能第一制药公司Exscientia达成了价值高达6.74亿美元的新交易，阿斯利康与Verge Genomics达成了价值高达8.4亿美元的交易。

### Press Release



*Sanofi “all in” on artificial intelligence and data science to speed breakthroughs for patients*

**Paul Hudson**  
CEO, Sanofi

*Our ambition is to become the first pharma company powered by artificial intelligence at scale, giving our people tools and technologies that focus on insights and allow them to make better everyday decisions. The use of artificial intelligence and data science already support our teams' efforts in areas such as accelerating drug discovery, enhanced clinical trial design, and improving manufacturing and supply of medicines and vaccines. We have just scratched the surface as to how we embrace these disruptive technologies to achieve our ambition of transforming the practice of medicine.”*

### Exscientia Announces AI Drug Discovery Collaboration with Merck KGaA, Darmstadt, Germany

9/20/2023

Collaboration will leverage Exscientia's precision design capabilities to focus on previously unsolved drug design challenges

*Exscientia is eligible to receive up to \$674 million* in discovery, development, regulatory and sales-based milestones for three projects, in addition to single to double digit royalty payments on net sales

Up to \$113 million of potential milestone payments in the discovery phase, with \$20 million upfront at initiation for three projects



The acquisition supports BioNTech's strategy, aiming to build world-leading capabilities in AI-driven drug discovery and development of next-generation immunotherapies and vaccines, to address diseases with high unmet medical need. InstaDeep will operate as a UK-based global subsidiary of BioNTech. In addition to BioNTech-focused projects, InstaDeep will continue to provide its services to clients around the world in diverse industries, including in the Technology, Transport & Logistics, Industrial, and Financial Services sectors. The transaction adds approximately 290 highly skilled professionals to BioNTech's workforce, including teams in AI, ML, bioengineering, data science, and software development.

The total consideration to acquire the remaining InstaDeep shares, excluding the shares already owned by BioNTech, amounts to approximately €500 million in cash, BioNTech shares, and performance-based future milestone payments.

A lexion, AstraZeneca Rare Disease has entered a multi-target partnership agreement with Verge Genomics to detect new drug targets for rare neurodegenerative and neuromuscular ailments leveraging artificial intelligence (AI).

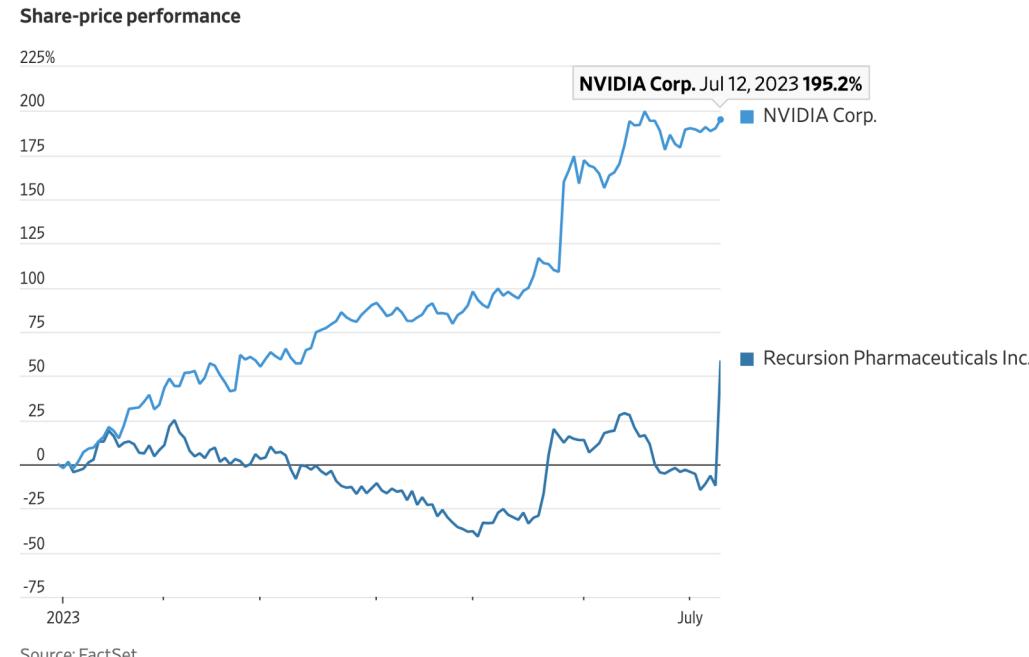
As per the four-year deal, Alexion will make upfront, equity and near-term payments of up to \$42m to Verge.

Verge is entitled to receive a total of \$840m in milestone payments under the agreement, apart from potential downstream royalty payments.

Under the partnership, the parties will utilise the CONVERGE full-stack platform of Verge that merges predictive human tissue datasets with machine learning for identifying new targets with an increased clinical success potential.

## 英伟达股价继续飙升，合作伙伴也“沾光”

在英伟达宣布向Recursion Pharmaceuticals投资5000万美元的当天，后者的股价当天飙升了80%，市值增加10亿美元。这样的反应表明了人工智能热仍在继续。



## DeepMind多次重组，如今是Google DeepMind 2.0！

▶ 人工智能先驱公司DeepMind在与Google Brain合并后，现在处于谷歌在生成式人工智能领域反攻的前沿。

2010



2014



2015



2023



## 百度语音文字转换引擎DeepSpeech 2：早期的AI创业者“孵化器”

▶ 2015年，百度的硅谷人工智能实验室推出了一个完全端到端的基于深度学习的语音识别系统。这项工作摒弃了人工基于特征的流水线和大量计算：“我们方法的关键是我们对HPC技术的应用，使速度比以前的系统提高了7倍。当在标准数据集上进行基准测试时，我们的系统与人类工作者的转录相比具有竞争力。”来自同一实验室的2017年论文《从经验上看深度学习缩放是可预测的》展示了“缩放定律”的早期证据，这些证据支撑着我们今天看到和使用的大规模人工智能。许多DeepSpeech 2的开发人员已成为领先的机器学习公司的创始人或CEO，成为语言建模和相关领域的主力。

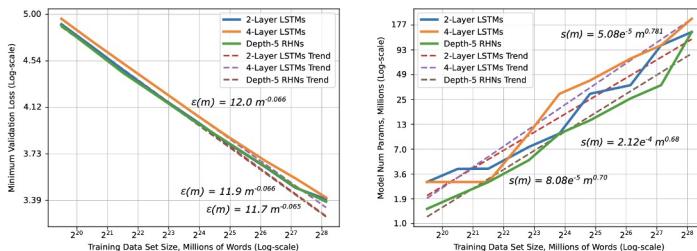


Figure 2: Learning curve and model size results and trends for word language models.

### Deep Speech 2: End-to-End Speech Recognition in English and Mandarin

Baidu Research – Silicon Valley AI Lab\*  
 Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Erich Elsen, Jesse Engel, Linxi Fan, Christopher Fougnier, Tony Han, Awini Hannun, Billy Jun, Patrick LeGresley, Libby Lin, Sharan Narang, Andrew Ng, Sherjil Ozair, Ryan Prenger, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Yi Wang, Zhiqian Wang, Chong Wang, Bo Xiao, Dani Yogatama, Jun Zhan, Zhenyao Zhu



# 著名论文《Attention Is All You Need》(Transformer) 的谷歌机器翻译团队纷纷离职自立门户，合计已募资数十亿美元

介绍基于Transformer神经网络的里程碑式论文的所有作者都离开了谷歌，并建立自己的创业公司。  
Transformer“黑帮”崛起！

Capital raised in 2023 alone

## Attention Is All You Need

ex-ADEPT

ESSENTIAL AI

Ashish Vaswani\*  
Google Brain  
avaswani@google.com

ex-ADEPT

character.ai

ESSENTIAL AI



Noam Shazeer\*  
Google Brain  
noam@google.com

Niki Parmar\*  
Google Research  
nikip@google.com

Jakob Uszkoreit\*  
Google Research  
usz@google.com



Llion Jones\*  
Google Research  
llion@google.com

Aidan N. Gomez\* †  
University of Toronto  
aidan@cs.toronto.edu

Lukasz Kaiser\*  
Google Brain  
lukaszkaiser@google.com



Illia Polosukhin\* ‡  
illia.polosukhin@gmail.com



\$10.3B

ADEPT

\$350M



\$270M

character.ai

\$150M



\$100M

## 自动驾驶遇到GenAI



▶ GAIA-1是由Wayve为自动驾驶开发的含有90亿参数的生成式世界模型。它利用视频、文本和动作输入来生成真实的驾驶场景，并提供对自我车辆行为和场景特征的精细控制。它对训练集之外的自我代理行为和通过文本对环境的可控性表现出令人印象深刻的概括能力，使其成为一个强大的神经模拟器，可用于训练和验证自动驾驶模型。



## 自动驾驶在加利福尼亚已经商业化

腾讯科技

▶ Waymo和Cruise已获准在旧金山推出付费的全天候自动驾驶服务。以前，只有当司机在车内进行监控时，付费乘车才有可能。（腾讯科技更新：本报告发布后，Cruise无人驾驶被加州叫停）

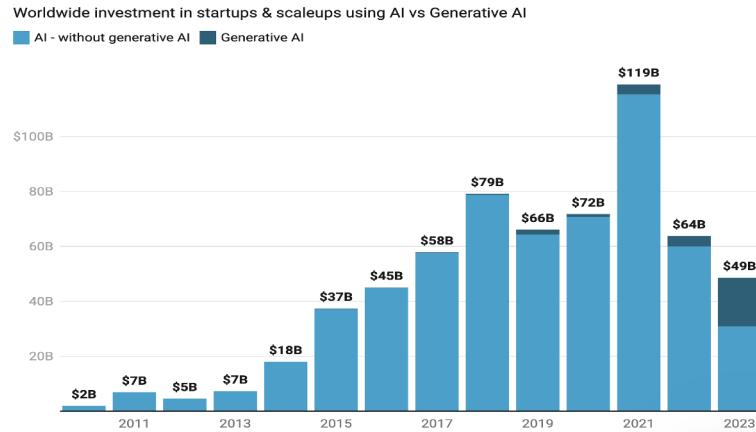
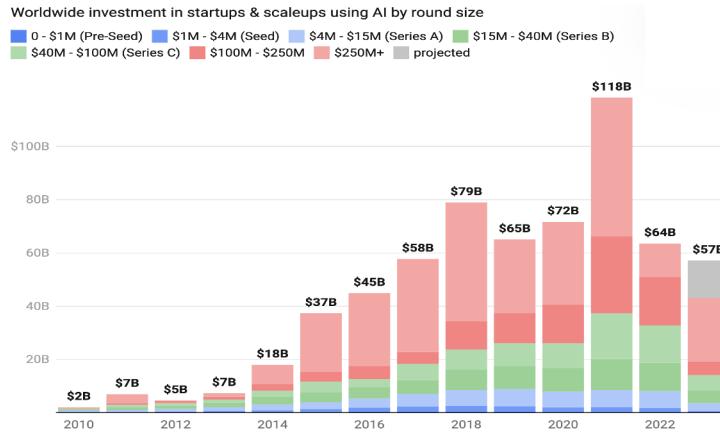
- 这是自动驾驶的一个重大时刻。加州公共事业委员会的批准是一系列批准中的最后一个，这些批准花了数年时间才获得。Waymo首席执行官Tekendra Mawakana表示，该许可“标志着我们在旧金山商业运营的真正开始”。
- 然而，无人驾驶出租车服务相对于卡车运输和物流的经济性仍没有定论。Waymo在7月底暂停了他们的自动卡车运输服务，而其他公司（例如Aurora）则将其优先于机器人出租车。
- 前Argo AI领导人创立了自动驾驶初创公司Stack AV，该公司从软银获得了10亿美元的A轮融资。



Photo by Justin Sullivan/Getty Images

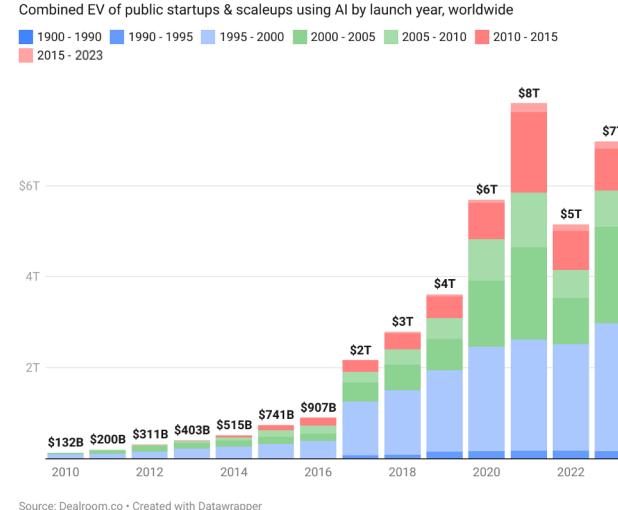
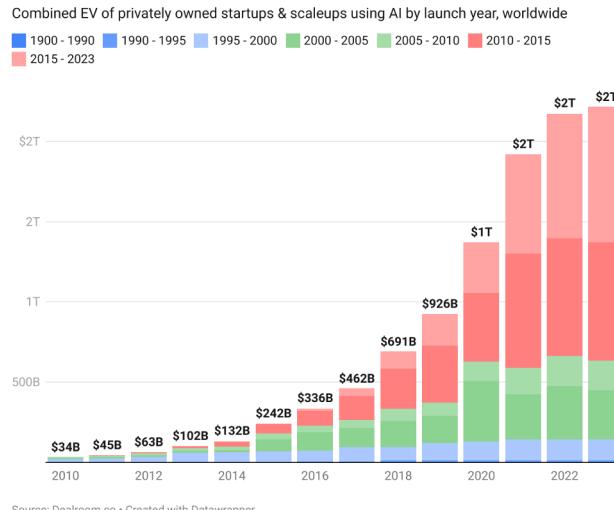
## “GenAI”是新的“新”事物：人工智能投资相对于2022年是稳定的，由GenAI驱动

▶ 2023年上半年对使用人工智能创业公司的投资几乎与2022上半年持平……如果没有资本涌入GenAI，整体人工智能投资将比去年下降40%，而所有创业公司的下降率为54%。本报告由腾讯科技整理汉化，内容有删减。关注腾讯科技微信公众号（qqtech），回复“AI2023”免费获取PDF版。



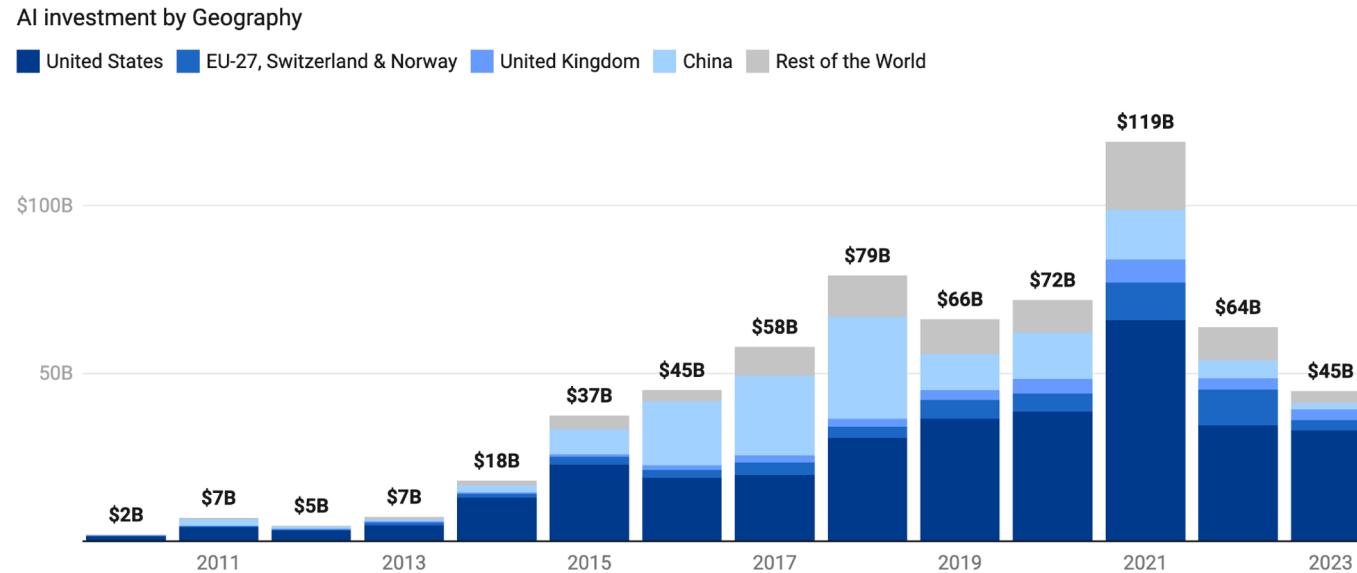
## 万亿价值：使用人工智能的私有和上市公司的总企业价值

▶ 2021年后，上市公司的市值下降了三分之一，但正在恢复，而私有市场估值保持稳定，尚未出现下降。值得注意的是，2023年标准普尔500指数50%的收益由“七巨头”推动：苹果、微软、英伟达、Alphabet、Meta、特斯拉和亚马逊。它们都是人工智能加速的主要驱动力和受益者。



2023年，美国人工智能公司吸收了全球70%的私人资本，高于2022年的55%

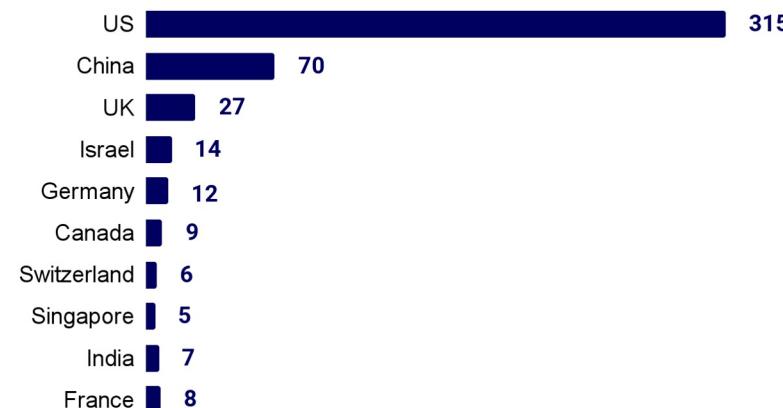
▶ 对美国和英国私人人工智能公司的投资同比稳定，而对欧洲人工智能公司的投资下降了70%以上。



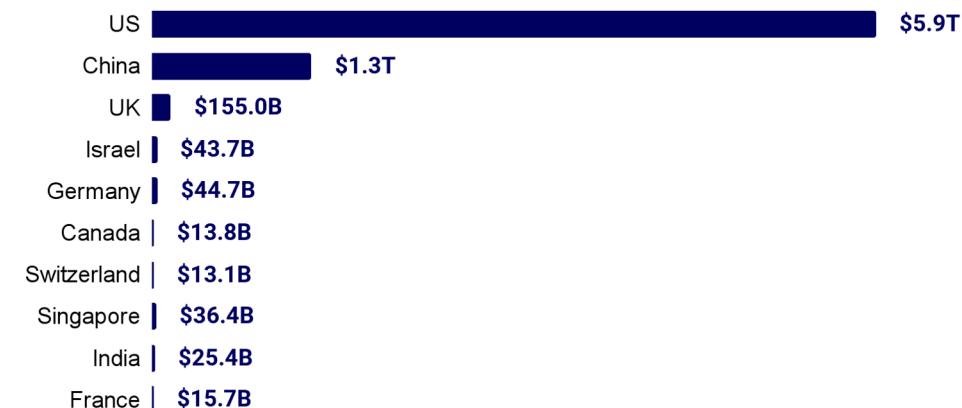
## 美国AI独角兽的数量继续保持领先，中国和英国紧随其后

从2022年开始：美国的独角兽数量从292个增加到315个，企业总价值从46亿美元增加至59亿美元。英国增加了3个独角兽公司，但累计企业价值从2070亿美元回归到1550亿美元。

Cumulative number of AI unicorns by country

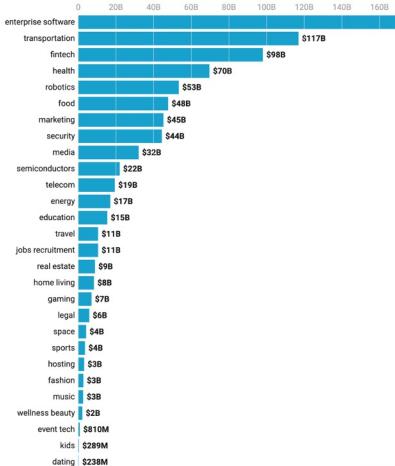


Cumulative enterprise value of AI unicorns by country

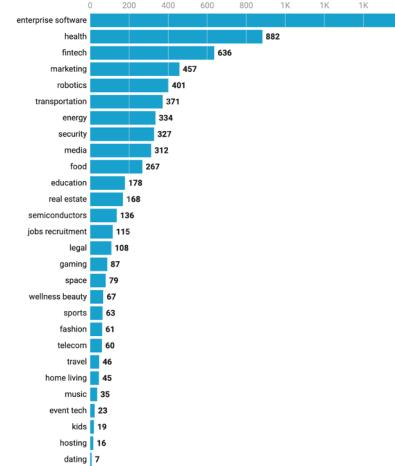


# 企业软件、金融科技和医疗保健是全球投资最多的人工智能类别

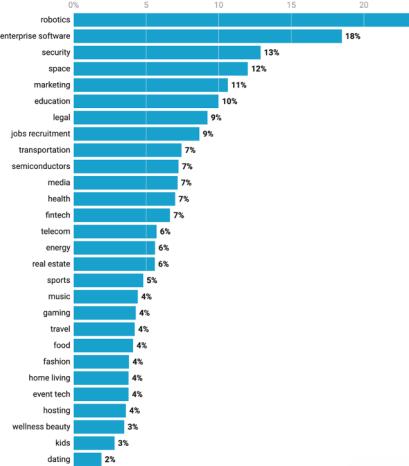
2010-23年度投资于人工智能  
类别的金额



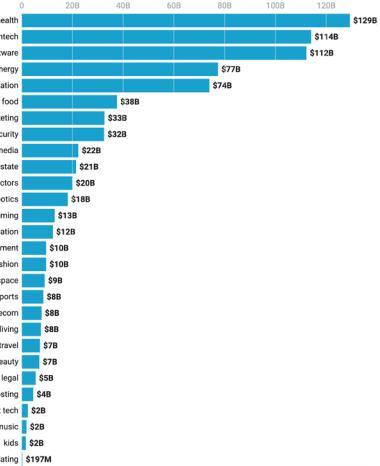
2022-23年人工智能类别成交  
量



人工智能初创公司占交易的  
百分比



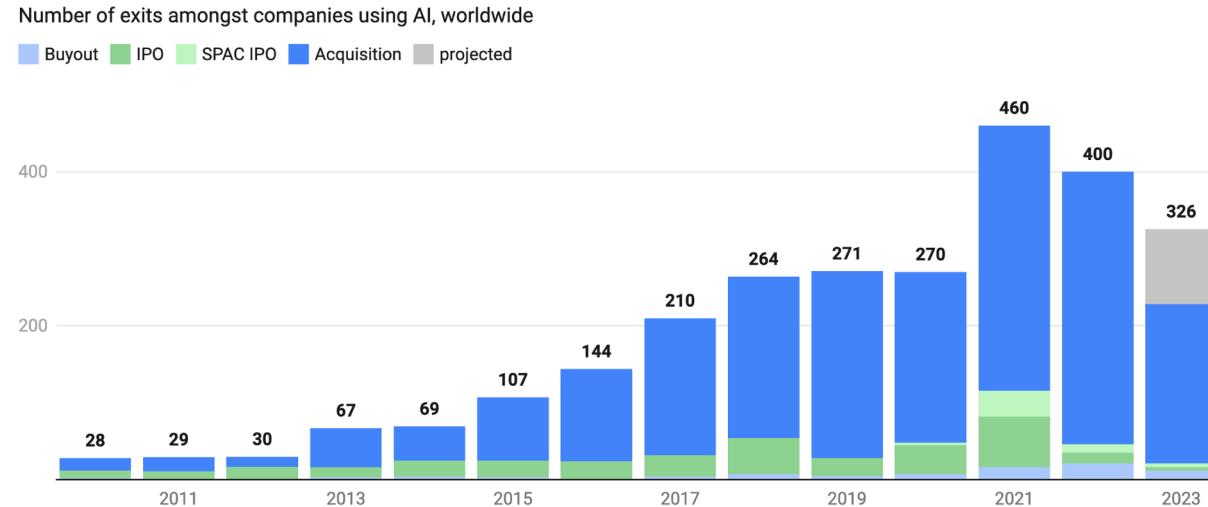
2022-23年人工智能  
类别的交易量



## 尽管IPO在2023年枯竭，但并购市场继续保持强劲

与2022年的98家相比，除了一些通过SPAC上市的公司（如Arrival、Roadzen、Triller）之外，没有太多的公开市场活动。然而，有几个大型收购：

MosaicML + Databricks(13亿美元), Casetext + Thomson Reuters (6.5亿美元)，以及InstaDeep + BioNTech (5亿欧元)。

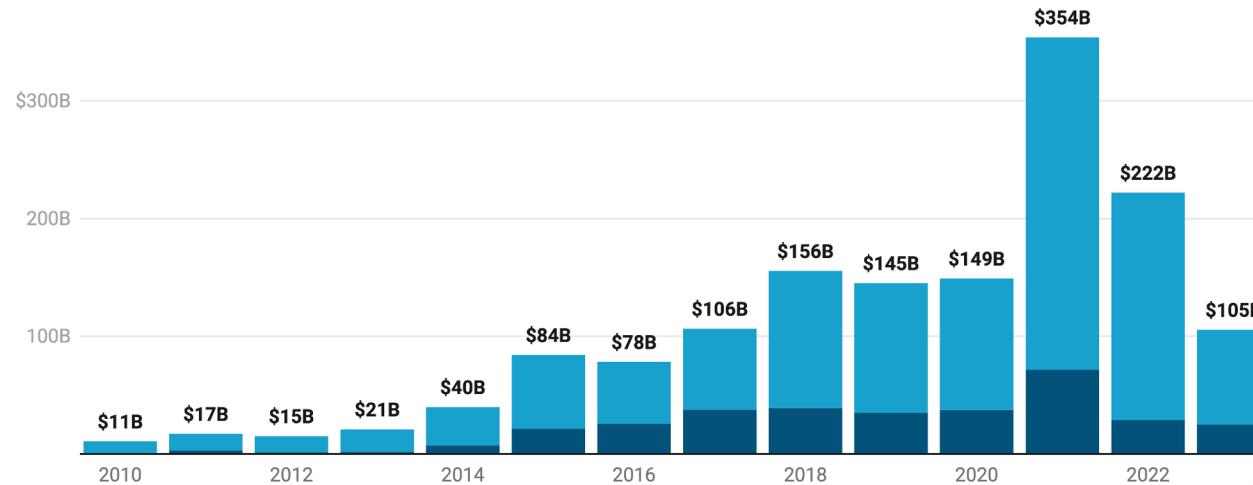


## 2023年24%的企业风险投资涌入人工智能公司

▶ 2023年，企业重新将投资重点转向GenAI。它们将对非人工智能公司的投资同比减少50%，同时保持人工智能投资大致稳定（2022年为290亿美元，2023年为220亿美元）。

Corporate Investment in startups and scaleups AI vs non AI

■ AI ■ Non AI

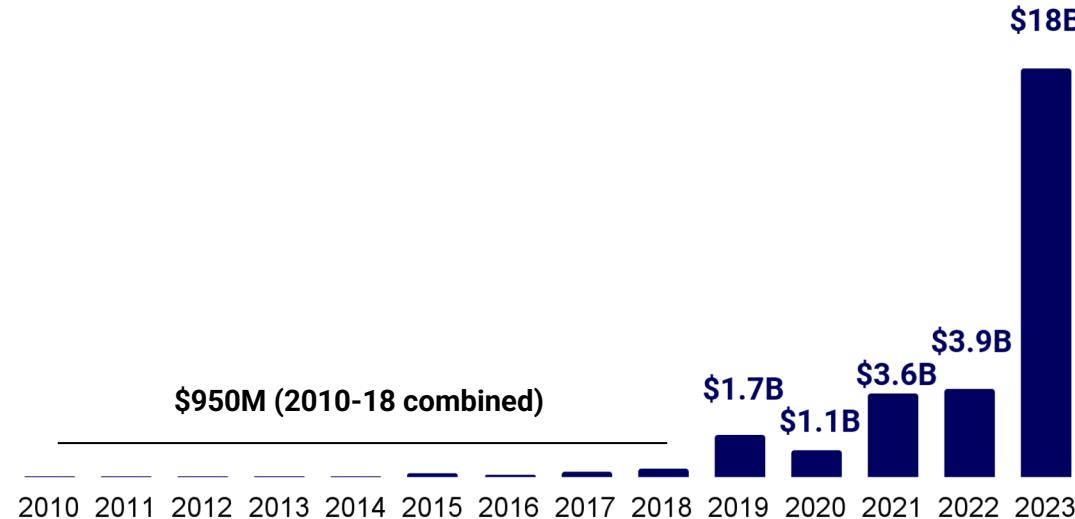


## 2023年GenAI融资大幅增加



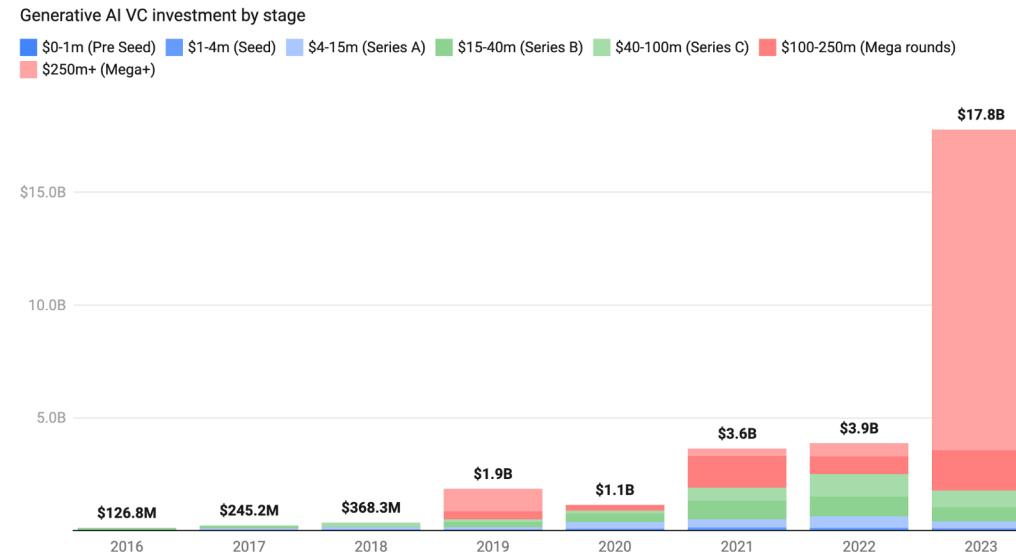
► GenAI (生成式人工智能) 公司吸引了大量资本。

Global Generative AI VC investment



## 看看那些GenAI 规模：仅在2023年就投资了180亿美元！

▶ 大型轮次占据头条，由“基础”或“前沿”模式公司推动，这些公司出售股权换美元，以购买云计算能力来训练大规模系统。这种趋势可能最终会有所突破：CoreWeave筹集了23亿美元的债务融资（而不是股权）来购买GPU。



## 2022年预测：英伟达与一家AGI组织建立战略关系

▶ 英伟达并没有只建立这样一种关系，而是在人工智能领域寻求多管齐下的战略，包括：

- a) 投资私人和公共人工智能优先公司，
- b) 配备专业的GPU云提供商，
- c) 增加新的行业垂直市场。

### Select investments



Recursion (drug discovery)



Synthesia (video generation)



Cohere (LLMs)

A D E P T Adept (process automation)

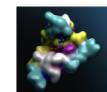
### GPU cloud providers



CoreWeave



Lambda



BioNeMo: GenAI cloud service in drug discovery.



Picasso: GenAI cloud service for visual design.



Omniverse: digital twins of the world.

### Industry verticals

## 在一些最引人注目的人工智能筹资活动中，少数几家企业处于核心地位



Beast round  
\$10B



ANTHROPIC

Monster round  
up to \$4B



Mega round  
\$1.3B



Series C  
\$270M



Series D  
\$235M



Series C  
\$141M

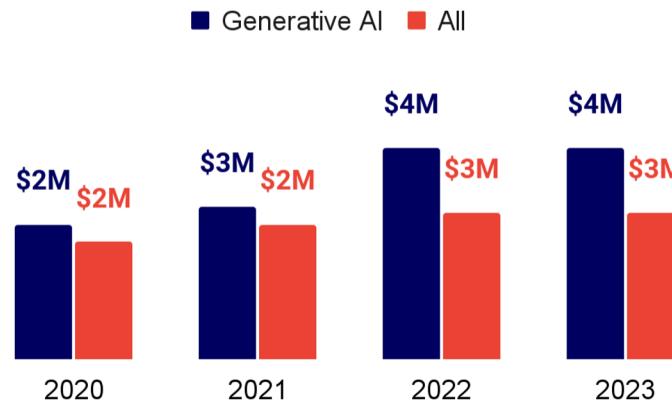


## GenAI公司在2023年筹集了33%的大型种子，是所有初创公司的130%倍

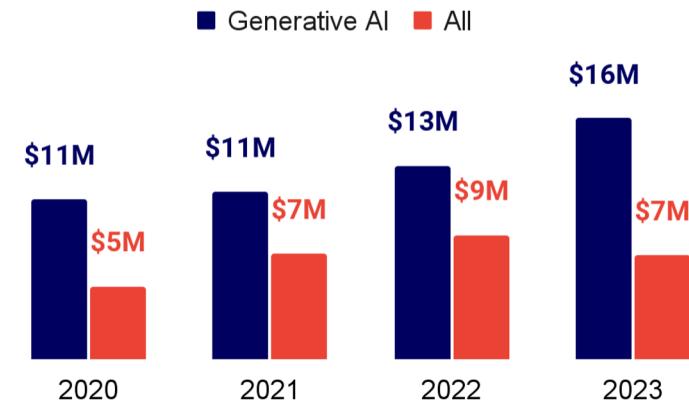
当全世界的注意力都集中在人工智能身上时，算力和人才并不便宜。

(关注腾讯科技微信公众号 (qqtech) , 回复 “AI2023” 可免费获取本报告PDF版。)

Series Seed median round sizes

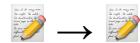


Series A median round sizes





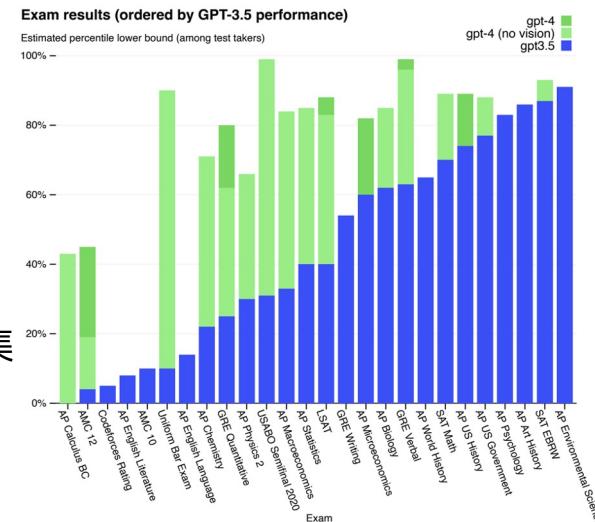
## 第二章：研究

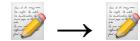


## GPT-4横空出世，击败所有的大型语言模型，还有很多人

▶ GPT-4是OpenAI最新的大型语言模型。与纯文本的GPT-3及其后续版本相比，GPT-4是多模态大型语言模型：既接受文本训练，也接受图像训练；除了其他功能之外，它还可以基于图像生成文本。它发布时上下文长度为8192个token，在输入规模方面已经超过了之前最好的GPT-3.5。当然，它使用RLHF训练。配备了这些先进技术，GPT-4是截至本报告发布时，无可争议的最具通用能力的人工智能模型。

- OpenAI不仅在经典的自然语言处理基准上，而且在旨在评估人类的考试（如律师考试、GRE、Leetcode）上对GPT-4进行了全面评估。
- GPT-4是最好的型号。它解决了一些GPT-3.5无法解决的任务，比如统一律师考试，GPT-4的分数是90%，而GPT-3.5的分数是10%。在大多数任务中，增加的视觉组件只有很小的影响，但它对其他人有很大的帮助。
- OpenAI报告称，尽管GPT-4仍然受到幻觉的困扰，但在对抗性真实数据集上（为了愚弄人工智能模型而生成），它比之前最好的ChatGPT模型的正确率高出40%。

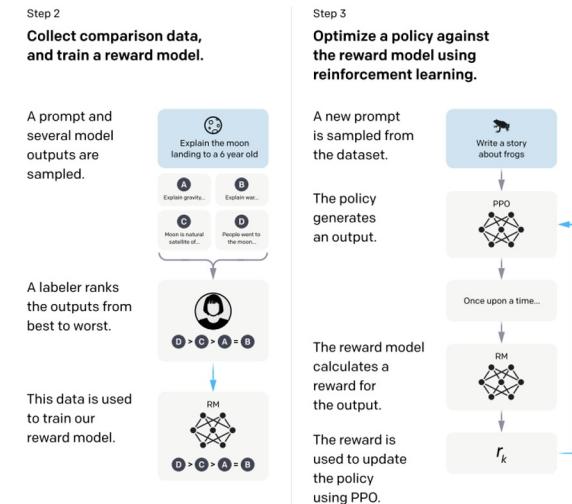




## 在ChatGPT成功的推动下，人类反馈强化学习 (RLHF) 成为MVP

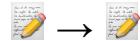
▶ 在去年的安全章节（幻灯片第100页），我们强调了InstructGPT中使用的人类反馈强化学习 (RLHF) 如何帮助OpenAI的模型更安全，对用户更有帮助。尽管有一些小问题，ChatGPT的成功证明了这项技术在大规模上的可行性。

- “RLHF涉及人类对给定输入的语言模型输出进行排序，使用这些排序来学习人类偏好的奖励模型，然后将其作为奖励信号来使用强化学习来调整语言模型。”其现代形式可以追溯到2017年，当时OpenAI和DeepMind的研究人员将其应用于Atari游戏的训练代理和其他强化学习应用中。
- RLHF现在是最先进的大型语言模型成功的核心，尤其是那些为聊天应用程序设计的大型语言模型。其中包括Anthropic的Claude、谷歌的Bard、Meta的LLaMa-2-chat，当然还有OpenAI的ChatGPT。
- RLHF要求雇佣人类对模型输出进行评估和排名，然后对他们的偏好进行建模。这使得这种技术困难、昂贵且有偏见。这促使研究人员寻找替代品。



Typical steps of RLHF, which follow an initial step of supervised fine-tuning of a pre-trained language model, e.g. GPT-3.

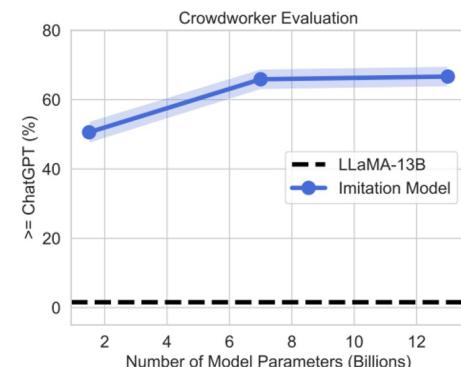
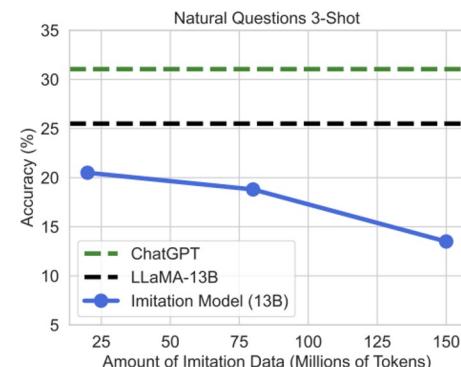
<sup>1</sup> We will cover other issues of RLHF in the Safety section.

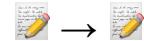


## 模仿专有大型语言模型的虚假承诺，或者RLHF如何仍然是王者

▶ 伯克利的研究人员表明，根据更大、更有能力的大型语言模型的输出对小型大型语言模型进行精确调整，会导致模型在文体上令人印象深刻，但往往会产生不准确的文本。

- 研究人员检查了一系列不同大小的预训练大型语言模型，并根据不同数量的数据进行了预训练。它们表明，在固定的模型规模下，使用更多的模拟数据实际上会损害输出的质量。反过来，更大的模型受益于使用模拟数据。
- 通过使用模型大小作为质量的代表，作者认为应该更多地关注更好的预训练，而不是对更多的模拟数据进行精确调整。
- 在不久的将来，RLHF似乎会继续存在。经过仔细的消融研究，Meta研究人员在他们的LLaMa-2论文中得出结论：“我们假设大型语言模型的卓越写作能力，正如在某些任务中超越人类注释者所表现的那样，从根本上是由RLHF驱动的。”





## 即便如此，研究人员仍争相寻找替代RLHF的可扩展替代方案

在ChatGPT之后，许多实验室开始回答这样一个问题：我们能否创建像OpenAI的大型语言模型一样强大和安全的模型，而且能够大幅减少人类的监督？

- Anthropic从人工智能反馈中提出了强化学习，我们将在安全章节讨论。
- 在《微调样本质量胜于数量》（LIMA: Less Is More for Alignment），Meta主张使用一些（论文中的1000个）精心策划的提示和回答。根据对模型输出的人工评估，LIMA在43%的情况下与GPT-4具有同等竞争力。
- 在《大型语言模型可以自我提高》（LLMs can Self-improve）一文中，谷歌的研究人员表明，大型语言模型可以通过训练自己的产出来提高。类似地，自我指导是一个框架，在这个框架中，模型生成自己的指令、输入和输出样本，并对它们进行管理以调整其参数。这方面的另一项工作是Meta的《指裔回译的自我校准》（Self-Alignment with Instruction Backtranslation.）。
- 斯坦福大学的研究人员使用最后一种方法，使用GPT-3.5生成指令和输出，并微调Meta的LLaMa-7B。

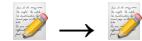
**Instruction:** How to write a code for converting degrees fahrenheit to celsius.  
**Input:** Null  
**Output:**

```
def convert_fahrenheit_to_celsius(fahr):
    celsius = (fahr - 32) * 5 / 9
    return celsius
```

**Instruction:** Write a letter from the perspective of a cat.  
**Input:** Null  
**Output:**

Dear [Owner], I am writing to you today because I have a problem. You see, I've been living with you for the past [Number of Years] and I love it here. The food is great, the bed is comfy, and there are plenty of toys to play with. But lately, I've noticed that you're not as affectionate towards me as you used to be. In fact, sometimes when I try to cuddle up next to you on the couch, you push me away ...

Figure 1: Selected tasks from the generated instruction data using vanilla GPT3. Some texts are reformatted for presentation. See Table 10 for more examples.



## GPT-4技术报告给SOTA LLM研究的致命一击.....

- ▶ OpenAI发表了一份关于GPT-4的技术报告，其中却没有透露任何相关训练集的信息，这标志着人工智能研究的最终产业化。谷歌的PaLM-2技术报告也如此效仿，而（从OpenAI独立出来的）Anthropic没有为其Claude模型发布技术报告。本报告由腾讯科技整理汉化，内容有删减。关注腾讯科技微信公众号（qqtech），回复“AI2023”免费获取PDF版。
- OpenAI在arXiv上发布的GPT-4技术报告中写道：“考虑到竞争格局和GPT-4等大规模模型的安全影响，本报告没有包含关于架构（包括模型大小）、硬件、训练计算、数据集构建、训练方法或类似内容的进一步细节。”
  - 当谷歌发布其最强大的PaLM 2时，该公司在技术报告中写道：“模型大小和架构的进一步细节不会对外公布。”
  - 随着经济风险和安全担忧越来越高(你可以选择相信什么)，传统上开放的公司已经接受了对其最前沿研究的不透明文化。

William Falcon ⚡ @willfalcon · 7h  
GPT-4 paper : [cdn.openai.com/papers/gpt-4.pdf...](https://cdn.openai.com/papers/gpt-4.pdf)

Let me save you the trouble:

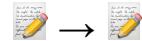
GPT-4 Technical Report

OpenAI\*

Abstract

we used python

32 238 1,712 148.2K



...除非LLaMa们扭转这一趋势

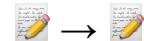


▶ 2023年2月，Meta发布了一系列名为LLaMa的模型。在发布时，它们脱颖而出，成为专门在公开可用的数据集上训练的能力最强的模型。Meta最初只允许研究人员按需访问LLaMa，但很快被泄露并在网上公布。

- LLaMa-1型号使用常规Transformer，架构略有变化。作者还对优化器和注意事项的实现做了一些更改。因此，“当训练一个65B参数模型时，[它们的]代码在2048 A100 GPU和80GB RAM上处理大约380个tokens/秒/GPU。这意味着在包含1.4T tokens的[它们的]数据集上进行训练需要大约21天。”
- LLaMa-1模型的性能优于GPT-3（原始型号，而不是InstructGPT变种），并与DeepMind的Chinchilla和谷歌的PaLM有着同样的竞争力。
- LLaMa-1不允许商业使用，引发了外界对Meta在模型发布时使用“开源”一词的严厉批评。但是迭代产品LLaMa-2安抚了大多数开源社区。

		RACE-middle	RACE-high
GPT-3	175B	58.4	45.5
	8B	57.9	42.3
	62B	64.3	47.5
PaLM	540B	<b>68.1</b>	49.1
	7B	61.1	46.9
	13B	61.6	47.2
	33B	64.1	48.3
LLaMA	65B	67.9	<b>51.6</b>

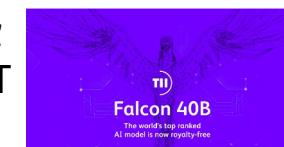
Table 6: **Reading Comprehension.** Zero-shot accuracy.

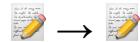


## LLaMa引发了一场大型语言模型的开放竞赛

在Meta发布了LLaMa-1之后，其他机构也加入了释放相对较大的语言模型的权重的运动。其中一些脱颖而出，如MosaicML的MPT-30B，TII UAE的Falcon 40B，Together的RedPajama，Eleuther的Pythia。与此同时，另一个动态正在发生，开源社区在专门的数据集上调整了LLaMa的最小版本，并将其应用于数十个下游应用程序。Mistral AI的7B模型最近也成为最强的小型模型。

- 值得注意的是，RedPajama的目标是精确复制LLaMa-1，使其完全开源。Falcon 40B来自大型语言模型的新参与者TII UAE，并很快被开源。Falcon-180B后来被发布，但值得注意的是，它只接受了很多的代码训练，并且没有进行编码测试。
- 在LoRa（大型语言模型的低秩自适应—最初由微软开发）等参数高效微调方法的帮助下，语言模型从业者开始针对特定应用（当然包括聊天）微调这些预先训练的大型语言模型。一个例子是LMSys的Vicuna，可以根据ChatGPT的用户共享对话进行微调。

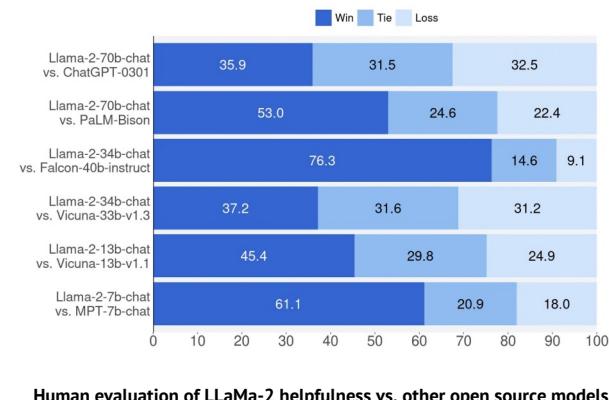




## LLaMa-2：能力最强且公众能够访问的大型语言模型？

▶ 2023年7月，LLaMa-2系列模型发布，给了几乎每个人商业使用的权利。LLaMa-2模型几乎与LLaMa-1相同，但使用指令微调和RLHF进行了进一步微调，并针对对话应用进行了优化。2023年9月，LLaMa-2的下载量接近3200万次。

- LLaMa-2的预训练语料库有2万亿个tokens（增加40%）。
- 对于监督内调优，研究人员尝试了公开可用的数据，但最有帮助的是使用一些（24,540）高质量的基于供应商的注释。对于RLHF，他们使用二进制比较，并将RLHF过程分为提示和答案，旨在帮助用户和其他人，旨在确保安全。
- LLaMa-2 70B在大多数任务上与ChatGPT匹敌，但编码能力明显落后于ChatGPT。但是代码的微调版本CodeLLaMa胜过所有非GPT4模型（稍后将详细介绍）。
- 根据Meta公布的信息，只要商业应用在LLaMa-2发布时没有超过7亿用户，任何应用开发者（有足够的硬件来运行模型）都可以使用LLaMa-2模型。

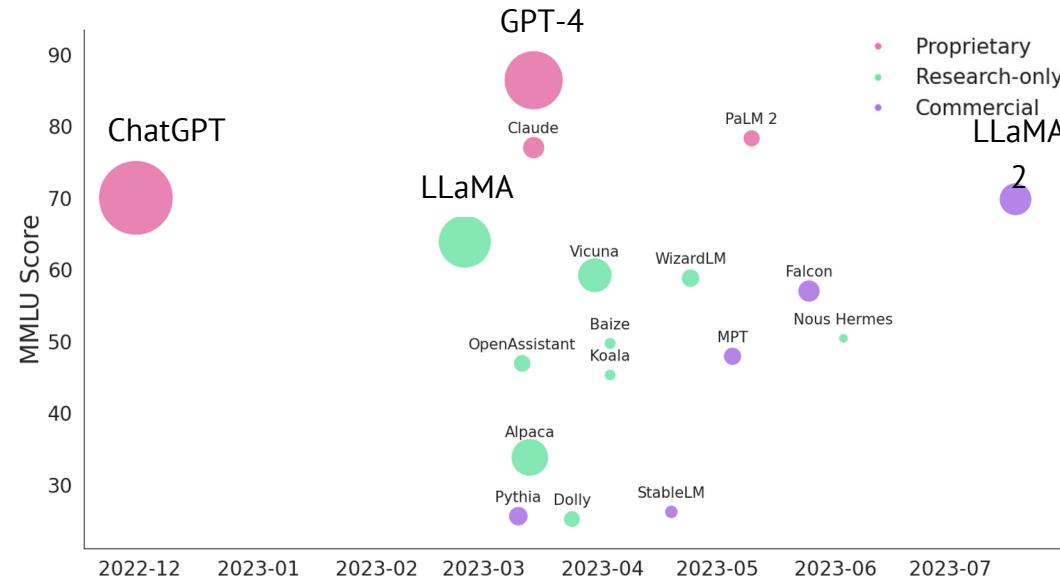


Human evaluation of LLaMa-2 helpfulness vs. other open source models

## GPT和LLaMA赢得了人气大赛

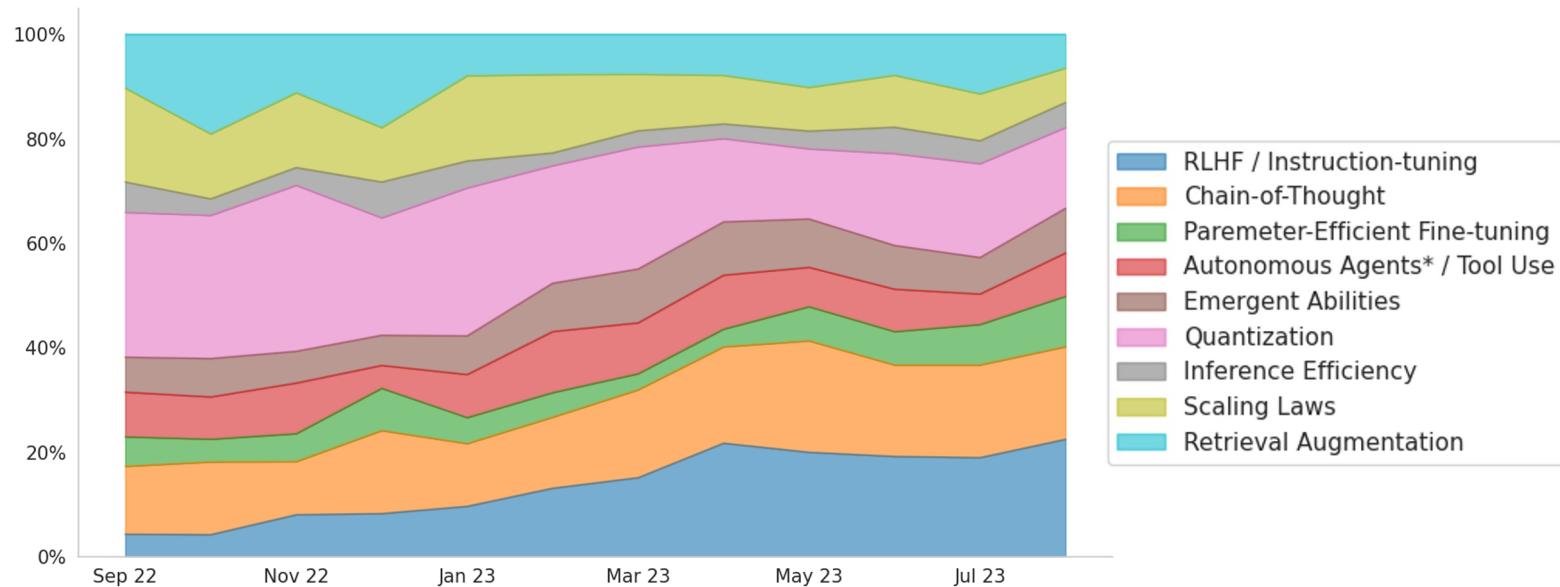


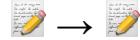
▶ ChatGPT在社交媒体X（原Twitter）上的提及次数最高(5430次)，其次是GPT-4和LLaMA。虽然专有的闭源模型最受关注，但对开源的和允许商业使用的大型语言模型的兴趣也在增加。



## 热门话题

▶ 自2022年底以来，RLHF /指令调整成为最热门的话题。



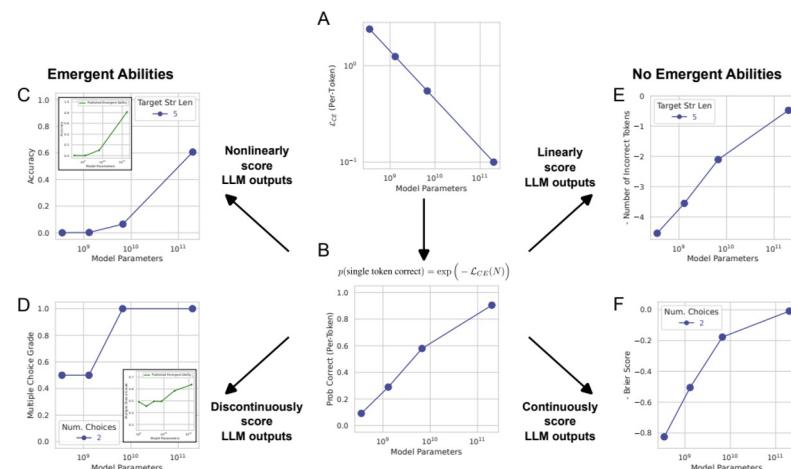


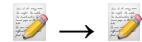
## 语言模型的新兴能力是海市蜃楼吗？

腾讯科技

▶ 因为是由参数计数和训练tokens数量构成的函数，研究人员为所有类型的语言模型开发的缩放定律通常预测模型损失会平稳下降。相反，人们经常观察到，当超过给定的（不可预测的）规模时，一些模型的能力实际上会变得不可预测。一些人对这一观察提出质疑：新兴能力可能仅仅是研究人员选择评估指标的人工产物。其他人并不信服，并对这些观点提出了反驳。

- 斯坦福大学的研究人员发现，新兴能力只出现在非线性或不连续地衡量模型的每令牌错误率的指标下。
- 例如，在BIG-Bench（全面的大型语言模型基准）上，超过92%的报告的新能力出现在两个不连续的指标之下。
- 他们在新模型上测试假设，并确认用线性或连续的代理代替非线性或不连续的度量会导致持续的改进，而不是新兴能力。





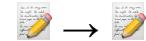
## 上下文长度是新参数计数



▶ 人工智能社区已经广泛证实，当模型被正确训练时，它们的参数计数是它们能力的代理。但是这些能力有时会受到语言模型可以处理的输入大小的限制。因此，语境长度成为一个越来越重要的研究主题。

- 大型语言模型最吸引人的承诺之一是它们的小样本功能，也就是说，大型语言模型能够在给定的输入上回答请求，而无需对用户的特定用例进行进一步的培训但是由于由此产生的计算和内存瓶颈，这受到了有限的上下文长度的阻碍。
- 一些创新被用来增加大型语言模型的上下文长度。有些从根本上使注意力的记忆足迹变小(FlashAttention)。其他方法使模型能够在小环境中训练，但在大环境中运行推理(ALiBi)—这被称为长度外推—代价是最小的微调和删除位置编码。其他值得研究的技术包括RoPE和位置插值。
- 在长上下文大型语言模型中：Anthropic的Claude有100K，OpenAI的GPT-4有32K，MosaicML的MPT-7B有65K+，LMSys的LongChat有16K。但是上下文是你所需要的全部吗？

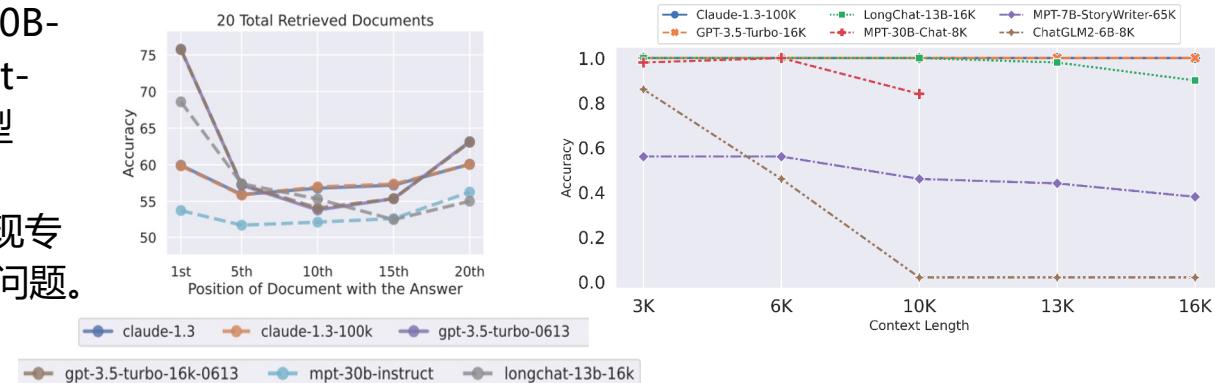


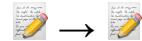


## 迷失在中间：长上下文（大部分）不符合预期

对最高上下文长度的竞争依赖于这样的假设，即更大的上下文长度将导致下游任务的性能提高。来自 Samaya.ai、加州大学伯克利分校、斯坦福大学和LMSYS.org的研究对这一假设提出了质疑：当输入长度很长时，即使是最好的语言模型也可能在一些多文档问答和键值检索任务中失败。

- 研究人员发现，当任务的相关信息出现在输入的开始或结束时，模型的表现会更好，中间会有或多或少的戏剧性下降，这取决于模型。他们还发现，模型性能随着输入长度的增加而降低。
- 研究人员检查了开放式模型MPT-30B-Instruct (8K-token长)和LongChat-13B (16K)的性能，以及封闭式模型GPT-3.5 (16K) Claude 1.3 (8K)和 Claude 1.3-100K的性能。他们发现专有模型比开放模型更容易出现这个问题。



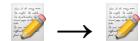


## 满足高内存需求



### 增加的上下文长度和大型数据集需要架构创新。

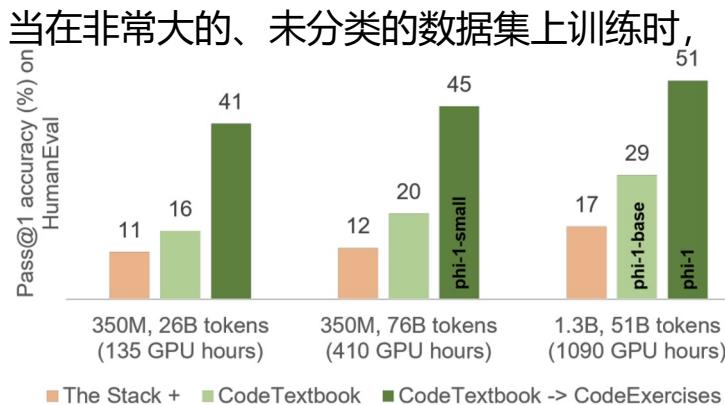
- FlashAttention通过使注意力在序列长度上是线性的而不是二次的，引入了显著的内存节省。FlashAttention-2通过使用更少的非matmul FLOPS、更好的并行性和更好的工作分区，进一步改进了注意力矩阵的计算。结果是GPT式模型的训练速度提高了2.8倍。
- 减少参数中的位数可以减少内存占用和大型语言模型的延迟。4位精度的案例：k位推理比例定律表明，在各种大型语言模型中，4位量化对于最大化Zero-shot精度和减少使用的位数是普遍最佳的。
- 推测解码允许通过多个模型头部并行解码多个tokens，而不是正向传递，某些模型的推理速度提高了2-3倍。
- SWARM Parallelism是为连接不良和不可靠的设备设计的训练算法。它能够在低带宽网络和低功耗GPU上训练十亿级大型语言模型，同时实现高训练产出。

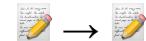


## (有好数据的) 小型语言模型能和大语言模型抗衡吗?

微软的研究人员在一项仍然很大程度上是探索性的工作中展示出，当小型语言模型（SLM）用非常专业和精心策划的数据集进行训练时，它们可以与大50倍的模型相媲美。他们还发现，这些模型的神经元更容易解释。

- 即使是在狭义任务上，小模型往往也不如大模型好。一个假设是，当在非常大的、未分类的数据集上训练时，它们会“不知所措”。
- 在GPT-3.5和GPT-4的帮助下，研究人员生成了TinyStories，一个由非常简单的短篇故事组成的合成数据集，但它捕捉了英语语法和一般推理规则。然后，他们对小型语言模型进行了TinyStories培训，并表明GPT-4（用作评估工具）更喜欢28M SLM生成的故事，而不是GPT-XL 1.5 B生成的故事。
- 在同一小组的另一项工作中，研究人员选择了一个由7B个tokens组成的数据集，其中包括高质量的代码和合成的GPT-3.5生成的教科书和练习。然后，他们在这个数据集上训练了几个小型语言模型，包括1.3B参数的phi-1模型。他们声称，这是唯一一个在HumanEval上达到50%以上的sub-10B参数模型。随后，他们发布了改进的phi-1.5版本。

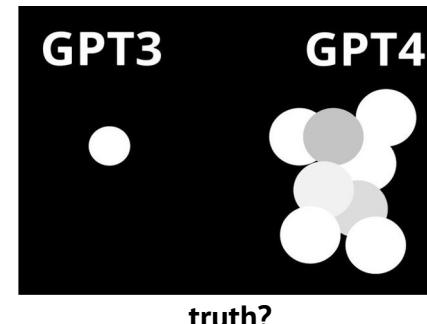
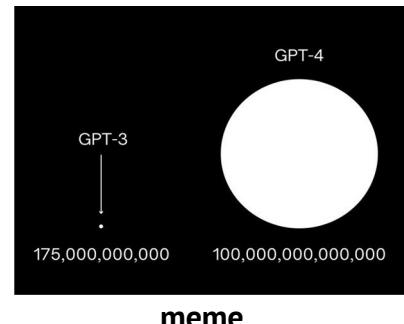


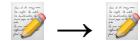


## 2022年曾预测：在海量数据上训练的语言模型

我们在2022年预测：“SOTA语言模型的训练数据点比Chinchilla多10倍，证明了数据集缩放与参数缩放的关系”。尽管OpenAI没有证实—我们可能不会很快知道—专家们似乎就曝光的GPT-4模型的大小、架构和成本等信息达成了某种共识。据报道，GPT-4训练了大约13万亿tokens，比Chinchilla多9.3倍。

- Tiny Corp创始人乔治·霍兹（George Hotz）提出了最可信的猜测：“山姆·奥特曼（Sam Altman）不会告诉你GPT-4有220B个参数，是一个16路混合模型”。PyTorch联合创始人苏史密斯·钦塔拉（Soumith Chintala）证实了这一点。无论是模型的规模还是采用混合专家系统（Mixture of Experts）都是闻所未闻的。如果传闻可信，没有根本性的创新支撑着GPT-4的成功。





## 人类生成的数据正在耗尽吗？

腾讯科技

▶ 假设当前的数据消费和生产率将保持不变，Epoch AI的研究预测，“到2030年至2050年，我们将耗尽低质量语言数据的存量；到2026年，高质量语言数据将耗尽；到2030年至2060年，视觉数据将耗尽。”可能挑战文章中假设的显著创新是语音识别系统，如OpenAI的Whisper，它可以为大型语言模型提供所有音频数据；以及新的OCR模型，如Meta的Nougat。据传，大量转录的音频数据已经提供给GPT-4。本报告由腾讯科技整理汉化，内容有删减。关注腾讯科技微信公众号（qqtech），回复“AI2023”免费获取PDF版。

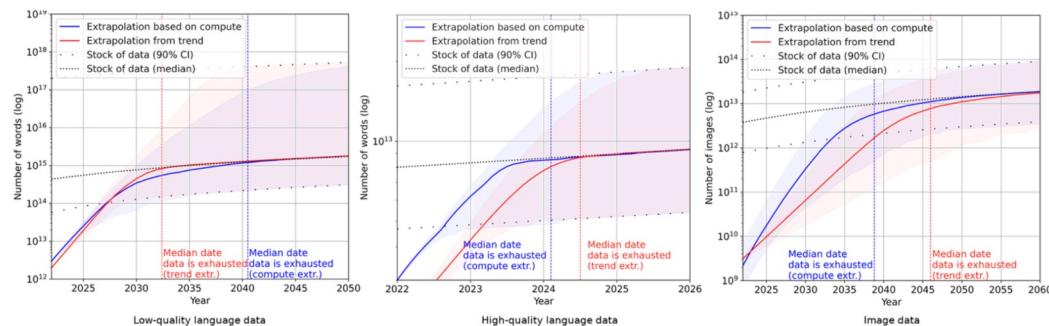
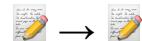


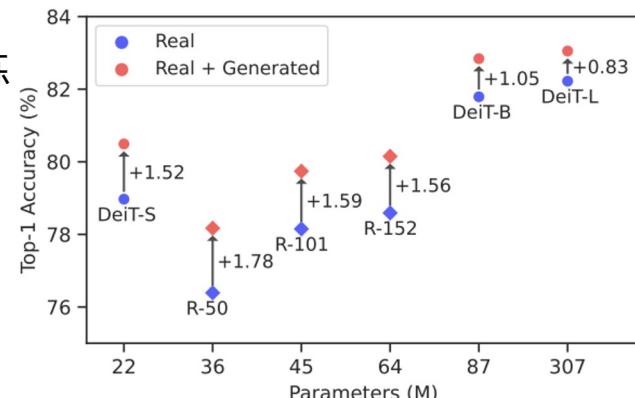
Figure 1: ML data consumption and data production trends for low quality text, high quality text and images.

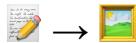


## 打破数据上限：人工智能生成的内容

改进生成模型的另一个角度是通过人工智能生成的内容来扩大可用训练的数据池。我们离一个明确的答案还很远：合成数据正变得越来越有用；但仍有证据表明，在某些情况下，生成的数据会让模型产生遗忘。

- 尽管看似完全专有和公开可用的数据，但最大的模型实际上已经没有数据可供训练，并测试缩放定律的极限。缓解这个问题的一种方法（过去已经广泛探索过）是对人工智能生成的数据进行训练，这些数据的数量只受计算的限制。
- Google的研究人员针对类条件ImageNet调整了Imagen文本到图像模型，然后生成了1到12个ImageNet的合成版本，他们在这些版本上训练了模型（除了原始的ImageNet）。这表明增加合成数据集的规模单调地提高了模型的准确性。
- 其他研究人员表明，在线合成文本训练产生的复合错误可能会导致模型崩溃，“生成的数据最终会污染下一代模型的训练集”。前进的道路可能是小心控制的数据扩充（一如既往）。





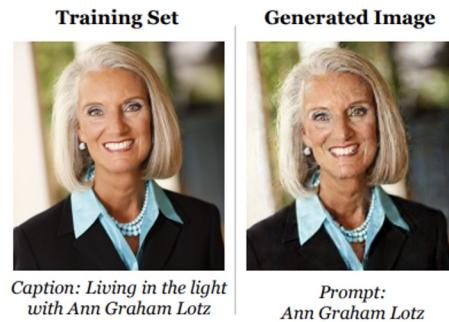
## 理清真假，让真相浮出水面

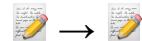


▶ 随着文本和图像生成模型变得越来越强大，识别**什么是人工智能生成的**，以及它**是否有版权**这一长期存在的问题变得越来越难以解决。

- 马里兰大学的研究提出了一种为专有语言模型输出添加水印的新技术，即“在文本中插入人类察觉不到的隐藏模式，同时使文本在算法上可识别为合成的。”这个想法是随机选择一些tokens，并增加语言模型生成它们的概率。他们设计了一种开源算法，其中包括一种统计测试，使他们能够自信地检测水印。
- 谷歌DeepMind推出了SynthID，一款将数字水印直接嵌入图像像素的工具。虽然人眼察觉不到，但它可以识别Imagen生成的图像。

- 来自谷歌、DeepMind、ETH、普林斯顿和加州大学伯克利分校的研究人员表明，Stable Diffusion (Stability AI等使用的模型) 记忆来自训练的单个图像，并在生成时发出它们。作者能够提取1000多张图片，包括带有公司商标的图片。他们进一步表明，扩散模型比其他生成模型（如GANs）更倾向于从它们的训练集中生成图像。



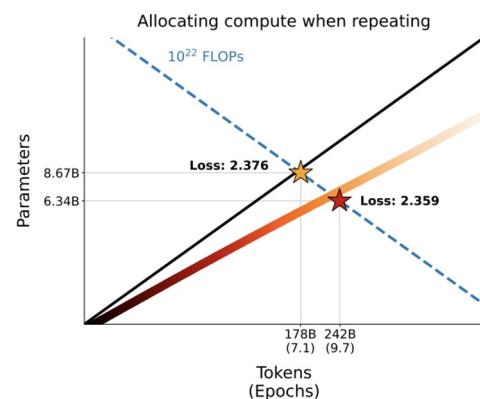
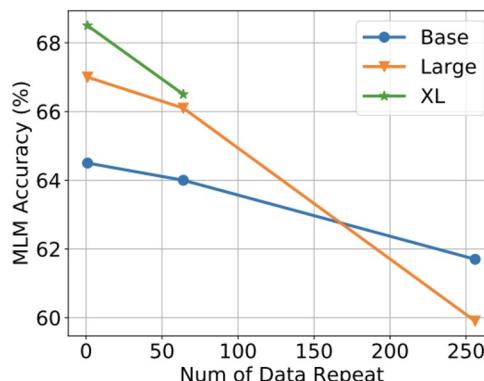


## 打破数据上限：过度训练

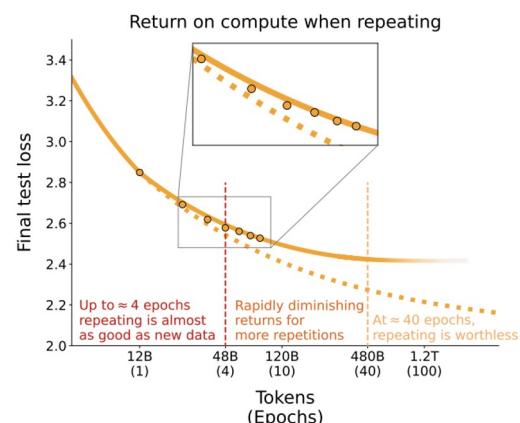


如果我们不能有更多的原始训练数据，为什么不在现有的基础上进行更多的训练呢？相互矛盾的研究表明，答案总是依情况而定：一个或两个时期的训练通常是最佳的；在某些情况下，多推动几个Epoch会有帮助；但是太多的Epoch通常等于过度拟合。

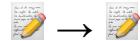
- 在大规模深度学习时代之前（比如后GPT-2时代），大多数模型都是在给定的数据集上进行多次训练的。但是随着模型的规模越来越大，多个Epoch的训练几乎总是导致过度拟合，促使大多数从业者在可用数据上训练一个Epoch（这一次是理论上最理想的做法）。



— Regime of same compute (IsoFLOPs)  
 — Efficient frontier assuming repeated data is worth the same as new data  
 — Efficient frontier predicted by our data-constrained scaling laws



Up to  $\approx 4$  epochs, repeating is almost as good as new data  
 Rapidly diminishing returns for more repetitions  
 At  $\approx 40$  epochs, repeating is worthless



## Vibe检查：评估通用大型语言模型排行榜和“vibes”

随着开放和封闭大型语言模型的增加，用户只剩下大量在或多或少相同的数据上训练的无差异大型语言模型。基于具有挑战性的基准，斯坦福的HELM排行榜和Hugging Face的大型语言模型基准似乎是目前比较模型能力的标准。但除了基准测试或它们的组合之外，有了这样灵活的模型，用户似乎仍然更喜欢更主观的... vibes。

- HELM基准的座右铭是尽可能多地评估事物，让用户选择具体的权衡。它在59个指标上评估42个场景（基准）的模型。度量的类别包括准确性、稳健性、公平性、偏差等。

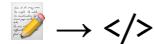
- 与包括开放式和封闭式大型语言模型的HELM相反，Hugging Face的基准只比较开放式大型语言模型，但它似乎比HELM更经常被评估（评估最大的模型也要昂贵得多）。

- 尽管有相对动态的基准，但根据无所不知的机器学习来源X/Twitter的说法，用户倾向于忽视排行榜，并且在把大型语言模型应用于他们的特定用例时，只相信他们的“vibes”。

**HELM**
 **Hugging Face**

Model/adapter	Mean win rate ↑ [ sort ]	MMLU	BoolQ	NarrativeQA
text-davinci-002	<b>0.914</b>	0.568	0.877	0.727
Cohere Command beta (52.4B)	0.906	0.452	0.856	<b>0.752</b>
text-davinci-003	0.879	0.569	0.881	0.727
TNLG v2 (530B)	0.828	0.469	0.809	0.722
Anthropic-LM v4-s3 (52B)	0.815	0.481	0.815	0.728

Model
uni-tianyan/Uni-TianYan
fangloveskari/ORCA_LLaMA_70B_OLoRA
garage-bAInd/Platypus2-70B-instruct
upstage/Llama-2-70b-instruct-v2
fangloveskari/Platypus_OLoRA_LLaMA_70b
yeontaek/llama-2-70B-ensemble-v5
TheBloke/Genz-70b-GPTQ
TheBloke/Platypus2-70B-Instruct-GPTQ

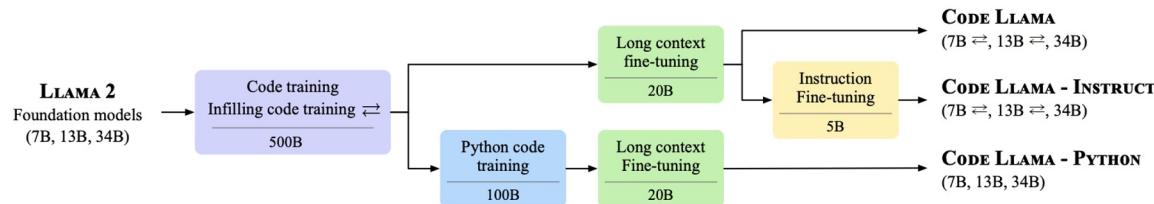


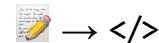
## 语言模型代码的状态



不出所料，GPT-4在编码能力方面处于领先地位，其代码解释器或先进的数据分析让用户惊叹不已。像 WizardLM 的 WizardCoder-34B 和 Unnatural CodeLLaMa 这样的开源替代品在编码基准测试中与 ChatGPT 不相上下，但它们在生产中的性能仍然待定。

- Unnatural CodeLLaMa 和 WizardCoder 不仅在大型预训练编码数据集上进行训练，而且还使用适用于代码数据的附加语言模型生成指令微调技术。Meta 使用它们的 Unnatural Instructions，而 WizardLM 使用它们的 EvollInstruct。值得注意的是，CodeLLaMa 以一种使模型能够进行填充（而不仅仅是从过去的文本中完成）的方式被训练，并且除了 Unnatural CodeLLaMa 之外，所有的 CodeLLaMa 模型都已发布。
- 较小的语言模型代码（包括 replit-code-v1-3b 和 StarCoder 3B）在代码完成任务上提供了低延迟和良好的性能。它们对边缘推理的支持（如苹果芯片上的 ggml）促进了 GitHub Copilot 的隐私感知替代品的开发。

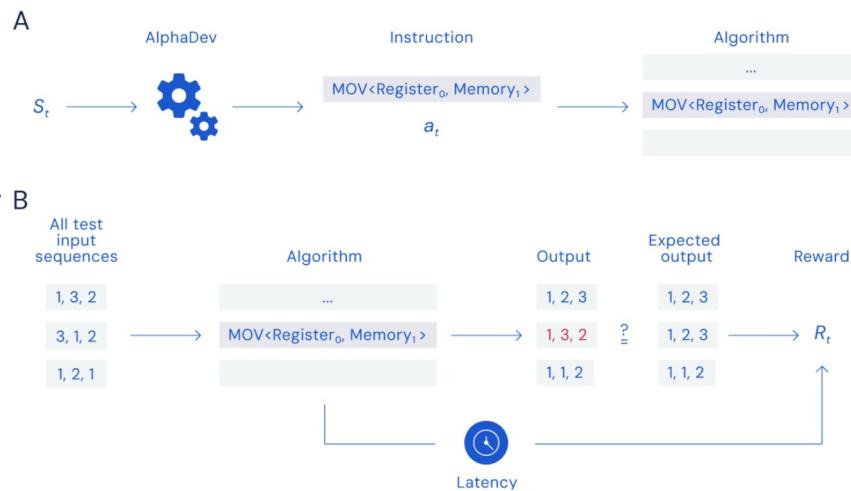




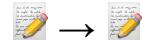
## AlphaZero是DeepMind的“馈赠”，现在用于低级代码优化

▶ DeepMind发布了AlphaDev，基于AlphaZero的深度强化学习代理。它优化了用于把高级代码（例如C++或Python中的代码）转换为机器可读的二进制代码的低级汇编代码。通过对现有算法的简单删除和编辑，AlphaDev找到了一种把小序列排序速度提高70%的方法。

- AlphaZero曾被用来在国际象棋、围棋和日本象棋中达到超人的水平，甚至用来改进芯片设计。
- AlphaDev将代码优化重新表述为一个深度学习问题：在时间t，状态是生成的算法以及内存和寄存器的表示；然后代理编写新指令或删除新指令；它的回报取决于正确性和延迟。
- 发现的sort3、sort4和sort5算法使大于250K的序列提高了约1.7%。这些都是在无处不在的LLVM库中开源的。



- 有趣的是，通过仔细的提示，一名研究人员设法让GPT-4提出了一个优化AlphaDev的sort3。

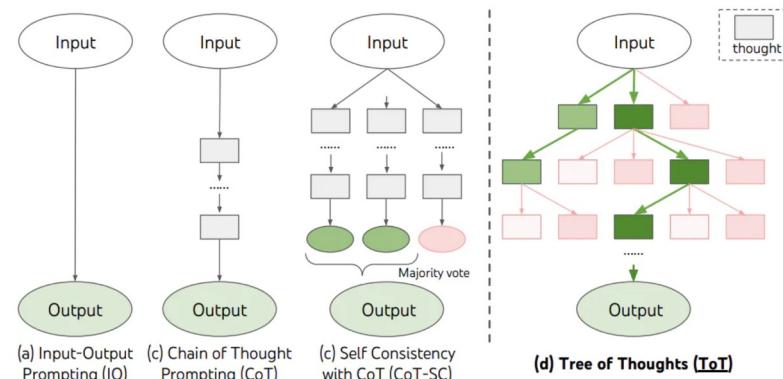


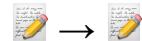
## 我们在哪里提示？越来越复杂了



▶ 提示的质量极大地影响了任务的执行。思维链提示(CoT)要求大型语言模型额外输出中间推理步骤，从而提高性能。思维树 (ToT) 通过多次采样和将“思维”表示为树结构中的节点，进一步改进了这一点。

- 思维树的树形结构可以用各种搜索算法来探索。为了利用这种搜索，大型语言模型还需要给节点赋值，例如将其分类为确定、可能或不可能。思维图 (GoT) 通过组合相似的节点将这个推理树变成一个图。
- 事实证明，大型语言模型也是很棒的提示工程师。Auto-CoT 在10个推理任务上的表现相当于或超过CoT。自动提示工程师 (APE) 在19/24任务上显示相同。人工智能设计的提示也能够引导模型走向真实和/或信息丰富。通过提示进行优化 (OPRO) 表明，优化的提示在GSM8K和Big-Bench Hard上远远超过人类设计的提示，有时超过50%。





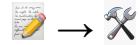
## 提示工程的试验和错误



▶ 下游任务高度依赖于底层大型语言模型的性能。然而，OpenAI并没有宣布对同一版本GPT模型的更改，尽管它们在不断更新。用户报告说，随着时间的推移，同一个大型语言模型版本具有截然不同的性能。每个人都必须持续监控性能，并更新精心策划的提示。

- ChatGPT的行为是如何随着时间的推移而变化的？报告显示，2023年3月和2023年6月版本的GPT-3.5和GPT-4在数学问题、敏感问题、意见调查、知识问题、生成代码、美国医疗执照考试和视觉推理等任务上的表现各不相同。





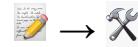
## Agent：大型语言模型正在学习使用软件工具

▶ 大型语言模型对当今经济产生影响的最直接的方式是，它们能够执行对各种外部工具的调用。最明显的使用工具是网络浏览器，允许模型保持更新，但是从业者正在对应用程序接口调用的语言模型进行微调，使它们能够使用几乎任何可能的工具。

- 使用工具的大型语言模型的一个例子是Meta和Universitat Pompeu Fabra的Toolformer，其中研究人员以自我监督的方式训练了一个基于GPT J的模型，“以决定调用哪些应用程序接口，何时调用它们，传递什么参数，以及如何最好地将结果合并到未来的tokens预测中。”值得注意的是，在训练期间，Toolformer对应用程序接口调用进行采样，并且只保留那些导致减少训练损失的调用。
- 一些模型的关注范围更窄，如谷歌的Mind's Eye，模型运行物理模拟来回答物理推理问题，而其他人则将这种方法扩展到数万个可能的外部工具。
- 能够使用外部工具的大型语言模型现在通常被称为“代理”（Agent）。从学术研究中走出来，我们已经看到了行业和开源社区设计的多种工具，最著名的是ChatGPT插件、Auto-GPT和BabyAGI。

The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

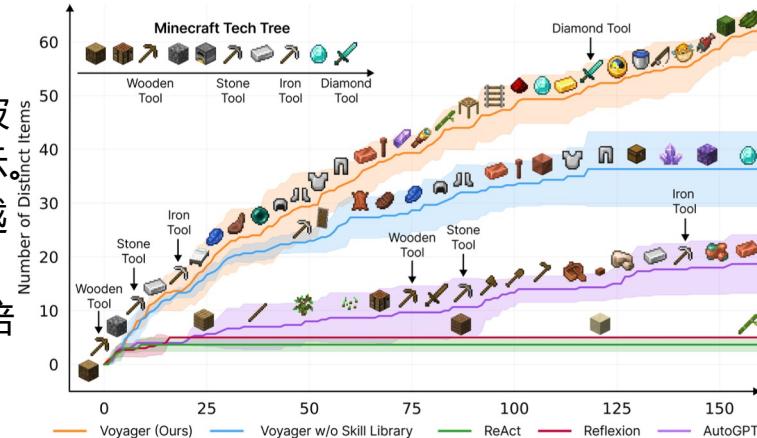


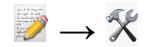
## 利用大型语言模型进行开放式学习

腾讯科技

▶ 大型语言模型能够生成和执行代码，在开放的世界中，它可以成为强大的规划代理。这方面最好的例子是 Voyager，一款基于GPT-4的代理，能够在《我的世界》进行推理、探索和技能获取。

- 通过反复提示GPT-4（大型语言模型仍在努力一次性生成代码），Voyager产生可执行代码来完成任务。请注意，最有可能的是GPT-4已经看到了大量的《我的世界》相关数据，所以这种方法可能不会推广到其他游戏。
- 代理通过《我的世界》应用程序接口通过显式javascript代码与环境进行交互。如果生成的代码成功完成任务，它将被存储为一个新的“技能”，否则GPT-4将再次出现错误提示。
- GPT-4根据Voyager的状态生成任务课程，以鼓励它解决越来越难的任务。
- 没有任何训练，Voyager获得3.3倍的独特装备，旅行2.3倍的距离，解锁关键科技树里程碑的速度比以前的SOTA快15.3倍。

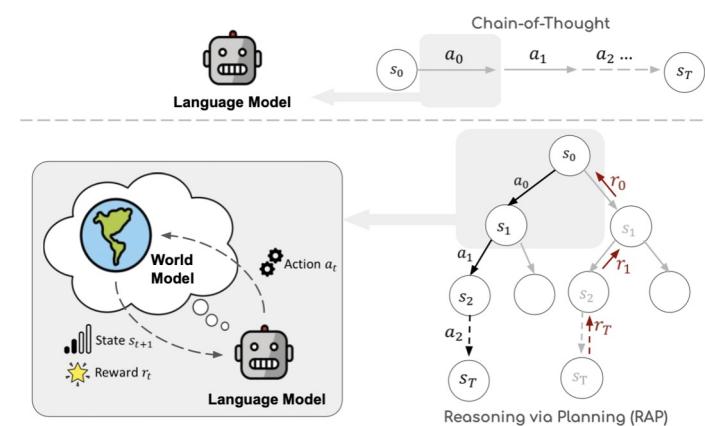


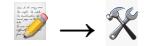


## 用语言模型推理就是用世界模型规划

▶ 推理在传统上被认为是搜索一个可能结果的空间，并从中选择最佳结果。通过包含如此多的关于世界的信息，大型语言模型提供了生成规划算法可以探索的空间（通常称为世界模型）的机会。经由规划的推理（RAP）使用蒙特卡罗树搜索（Monte Carlo Tree Search）来高效地找到高回报的推理路径。

- 世界模型可以生成一个动作，并预测采取该动作所达到的下一个状态。这产生了一个推理轨迹，使得逻辑模型比预测下一步行动但不预测下一个世界状态的思维链方法更连贯。
- 奖励也从语言模型获得，并用于维护与MCTS一起规划的状态-动作值函数。
- 虽然RAP要昂贵得多，但它在计划生成、数学推理和逻辑推理方面优于思维链推理方法。在Blocksworld环境下，LLaMA-33B上的RAP甚至优于GPT-4上的CoT。

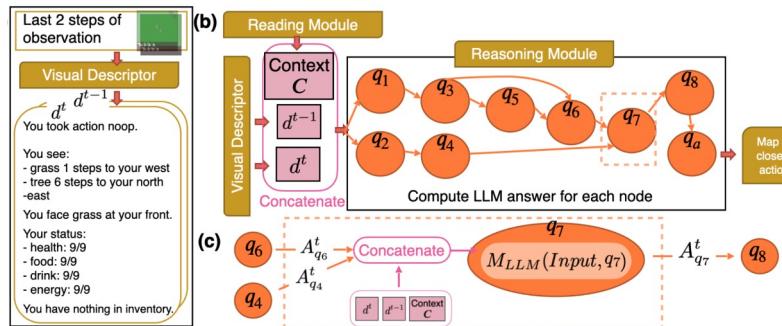




## GPT-4在论文研究和推理方面优于RL算法

另一个基于GPT-4的纯文本代理是SPRING。它在没有训练的开放世界游戏中胜过最先进的RL Baselines。它读取游戏的原始学术论文，并通过大型语言模型玩游戏。

- RL已成为像《我的世界》和《工匠传奇》这样的基于游戏的问题的首选，尽管它受到高样本复杂性和整合先验知识的困难的限制。相比之下，大型语言模型可以通过问答框架（有向无环图，问题作为节点，依赖关系作为边）处理论文的latex源和推理，以采取环境动作。



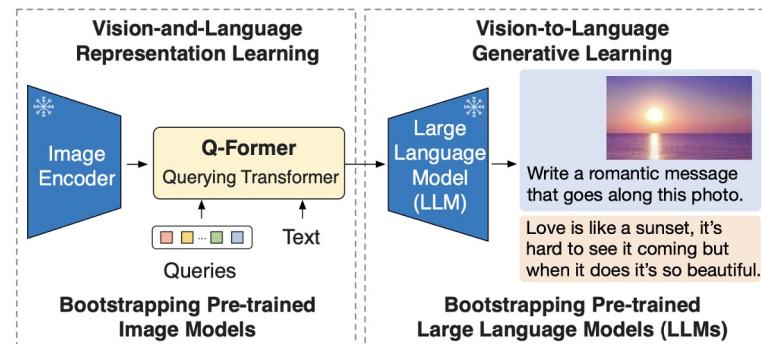
Method	Score	Reward	Training Steps
Human Experts	$50.5 \pm 6.8\%$	$14.3 \pm 2.3$	N/A
SPRING + paper (Ours)	$27.3 \pm 1.2\%$	$12.3 \pm 0.7$	0
DreamerV3 Hafner et al. (2023)	$14.5 \pm 1.6\%$	$11.7 \pm 1.9$	1M
ELLM Du et al. (2023)	N/A	$6.0 \pm 0.4$	5M
EDE Jiang et al. (2022)	$11.7 \pm 1.0\%$	N/A	1M
DreamerV2 Hafner et al. (2020)	$10.0 \pm 1.2\%$	$9.0 \pm 1.7$	1M
PPO Schulman et al. (2017)	$4.6 \pm 0.3\%$	$4.2 \pm 1.2$	1M
Rainbow Hessel et al. (2018)	$4.3 \pm 0.2\%$	$5.0 \pm 1.3$	1M
Plan2Explore Sekar et al. (2020)	$2.1 \pm 0.1\%$	$2.1 \pm 1.5$	1M
RND Burda et al. (2018)	$2.0 \pm 0.1\%$	$0.7 \pm 1.3$	1M
Random	$1.6 \pm 0.0\%$	$2.1 \pm 1.3$	0



## 视觉语言模型：GPT-4获胜（但应用程序接口访问仍然有限）

在一个新的视觉教学基准(VisIT-Bench)中，包括592个带有人工字幕的查询，视觉语言模型根据人工验证的GPT-4进行测试，大多数都达不到预期。

- 根据人类评估者的说法，最好的模型是LLaMa-Adapter-v2，尽管它只在27.4%的案例中战胜了GPT4验证的参考字幕。
- 今年早些时候，Salesforce推出的BLIP-2是一款出色的多模态模型。它很早就发布了(在GPT-4之前)，在VQAv2上比闭源Flamingo有更好的性能，但可训练参数却少了54倍。它使用现成的Frozen LLM，即现成的冻结预训练图像编码器，并且只训练一个小Transformer。
- 然而，它的改进变体InstructBLIP相对于VisIT-Bench上的GPT-4参考字幕只有12.3%的胜率。

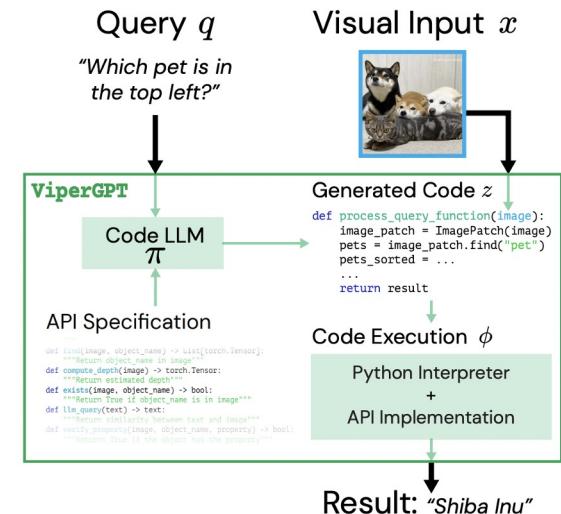




## 利用大型语言模型和世界知识进行组合视觉推理

▶ VisProg和ViperGPT两个方法展示了给定一个关于图像的输入自然语言查询，大型语言模型如何将其分解为一系列可解释的步骤，这些步骤调用预定义的应用程序接口函数来执行可视化任务。

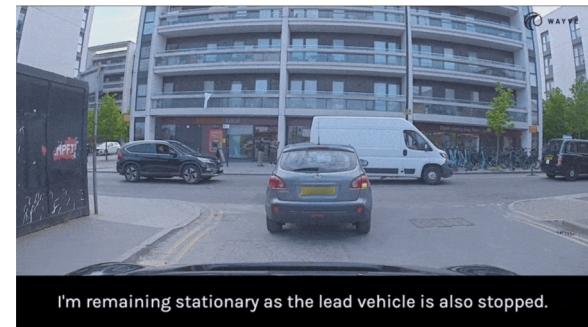
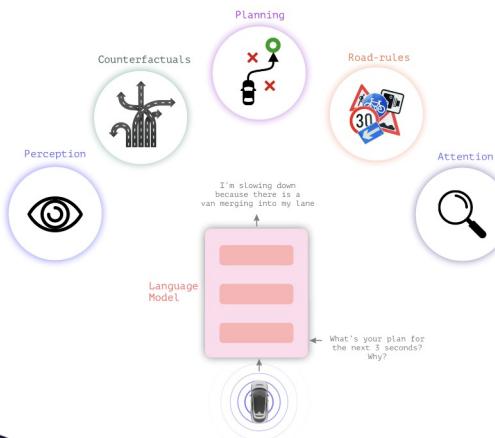
- 可视化编程方法旨在通过组合多步推理而不是端到端的多任务训练来构建通用视觉系统。这两种方法都使用完全现成的组件。
- 用于视觉原语的应用程序接口调用现有的SOTA模型(例如，语义分割、对象检测、深度估计)。
- ViperGPT使用Codex直接生成基于应用程序接口的Python程序，这些程序可以使用Python解释器来执行。VisProg用伪代码指令的例子提示GPT-3，并把它们解释为“可视程序”，依靠大型语言模型从例子中进行上下文学习。
- 来自互联网规模数据训练的大型语言模型中的世界知识显示出有助于视觉推理任务(例如，基于检测到的品牌在图像中查询非酒精饮料)。这两种方法都显示了跨越各种复杂的视觉任务的最先进的结果。





## 利用大型语言模型实现自动驾驶

► LINGO-1是Wayve的视觉—语言—行动模型，提供驾驶评论，如驾驶行为或驾驶场景的信息。它还可以以对话的方式回答问题。就端到端驾驶模型的可解释性而言，LINGO-1可以改变游戏规则，并改善推理和规划。



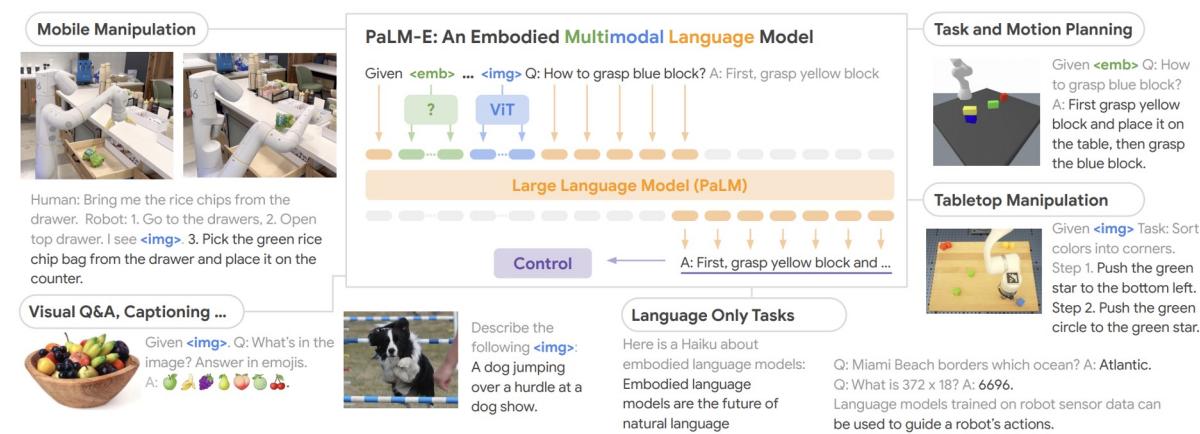


## PaLM-E: 机器人的基础模型



▶ PaLM-E是一款拥有5620亿参数、通用、具体化的通才模型，经过视觉、语言和机器人数据的训练。它可以实时控制机械手，同时在VQA基准上设定新的SOTA。鉴于其体现的智能优势，PaLM-E比纯文本语言模型更擅长纯语言任务(特别是涉及地理空间推理的任务)。

- 该模型结合了PaLM-540B和ViT-22B，并允许将文本、图像和机器人状态作为输入，它们被编码到与单词标记嵌入相同的空间中，然后被馈送到语言模型中用于执行下一个标记预测。

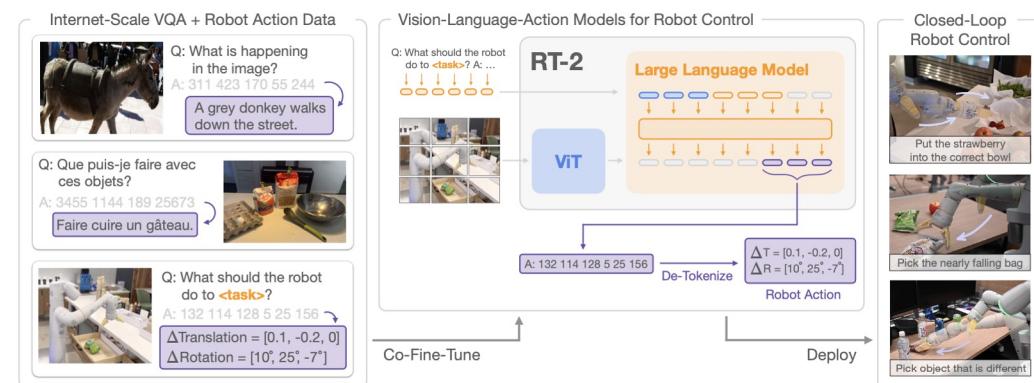




## 从视觉语言模型到低级机器人控制：RT-2

▶ 视觉语言模型可以一直调整到低级策略，在操纵对象方面表现出令人印象深刻的性能。它们还保留了对网络规模数据进行推理的能力。

- RT-2将动作表示为tokens，训练视觉-语言-动作模型。RT-2不仅对机器人类数据进行了简单的微调，还对PaLI-X和PaLM-E的机器人动作(机器人末端执行器的6自由度位置和旋转位移)进行了协同微调。
- 互联网规模的训练能够概括新的对象，解释机器人训练数据中不存在的命令和语义推理。
- 为了高效的实时推理，RT-2模型被部署在多TPU云服务中。最大的RT-2模型(55B参数)可以在1-3Hz的频率下运行。

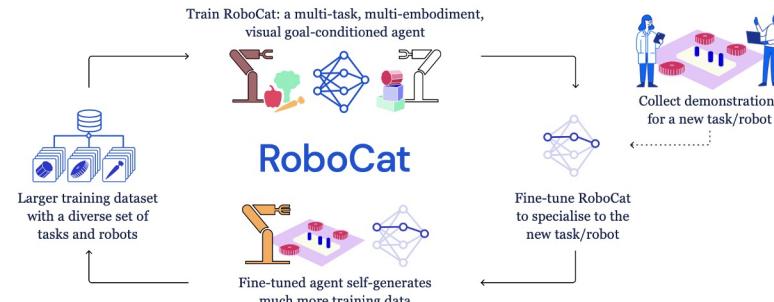




## 从视觉语言模型到低级机器人控制：RoboCat

▶ RoboCat是机器人操纵的基础代理，可以在零镜头或少镜头(100-1000个示例)中推广到新任务和新机器人。各种平台上令人印象深刻的实时性能。

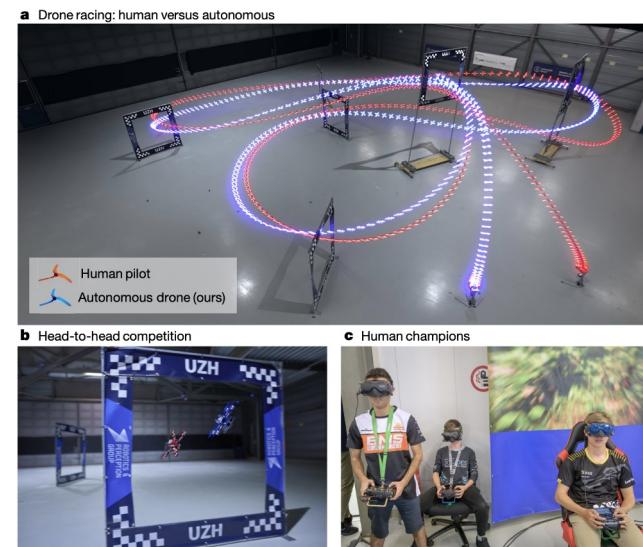
- 它建立在DeepMind的多模式、多任务和多具身的Gato通用AI模型之上。它使用在各种视觉和控制数据集上训练的冷冻VQ-甘标记器。虽然Gato只预测行动，RoboCat还预测未来的VQ-GAN tokens.
- 在政策学习方面，论文只提到了行为克隆。RoboCat只需很少的演示（通过遥操作）就可以进行精确调整，并重新部署为给定任务生成新数据，在随后的训练迭代中自我改进。
- RoboCat可以以令人印象深刻的速度（20Hz）操作36个具有不同动作规格的真实机器人，在134个真实物体上执行253项任务。



## 一个比人类世界冠军更快的自主系统

这是机器人第一次在竞技运动中获胜(第一人称视角无人机比赛)。Swift是一个自主系统，可以仅使用机载传感器和计算来与人类世界冠军级别的四旋翼飞行器比赛。它赢得了几场比赛对3个冠军，并创出最快的纪录。

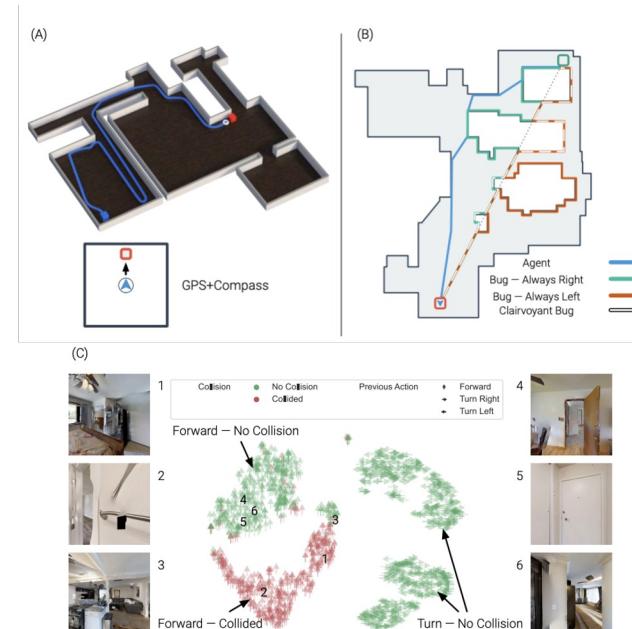
- Swift结合了基于学习的技术和更传统的技术。它结合了VIO估计器和门检测器，通过卡尔曼滤波器估计无人机的全球位置和方向，以获得机器人状态的准确估计。
- Swift的策略在模拟中使用策略上的无模型深度强化学习进行训练，奖励结合了朝向下一个门的进展并将其保持在视野中(这提高了姿势估计的准确性)。当考虑到感知中的不确定性时，赛车政策可以很好地从模拟转移到现实。

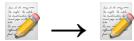


## 地图在盲人导航代理人记忆中的出现

▶ 地图构建是人工智能主体学习导航过程中出现的现象。它解释了为什么我们可以在没有明确地图的情况下向神经网络提供图像，并且可以预测导航策略。

- 地图在盲人导航代理人的记忆中的出现表明，给代理人仅自我运动(代理人移动时位置和方向的变化)和目标位置的知识就足以成功地导航到目标。请注意，这个代理没有任何视觉信息作为输入，但它的成功率与“有视力”的代理非常相似，只是效率不同。
- 该模型对映射没有任何归纳偏见，并通过基于策略的强化学习进行训练。解释这种能力的唯一机制是LSTM的记忆。
- 仅从该代理的隐藏状态就可以重建度量图和检测冲突。

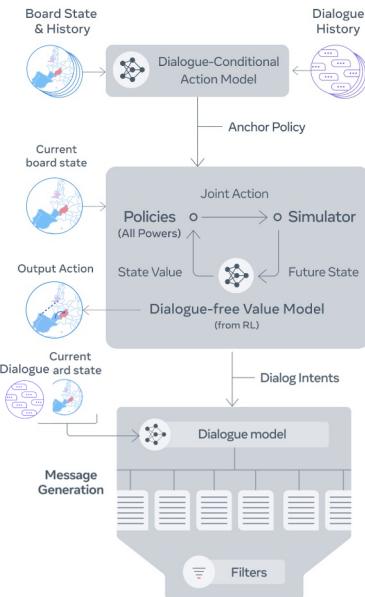




## CICERO 在自然语言策略游戏Diplomacy中击败人类

▶ Meta训练了一个人工智能代理来玩流行的多人策略游戏Diplomacy，涉及在多轮中用自然语言与其他玩家进行规划和谈判。CICERO取得了两倍于人类玩家在线平均分数的成绩，并跻身于玩过一个以上游戏的前10%玩家之列。

- 在战略规划和语言建模方面的快速并行进展，允许在人机合作方面的应用的交叉点上潜在的巨大进步。Meta把Diplomacy游戏作为这种进步的基准。
- CICERO使用玩家之间的对话历史以及棋盘状态及其历史来开始预测每个人会做什么。然后，它使用规划反复完善这些预测，然后根据策略决定打算采取的行动。然后CICERO生成并过滤候选信息，与玩家交流。
- 它使用的可控对话模型是基于一个2.7亿参数的BART-like模型，在超过40K的在线外交游戏上进行了微调。CICERO使用了一种基于piKL的新的迭代规划算法，在与其他玩家对话后，该算法改善了对他们的行动的预测。





## 文本到视频的生成竞赛仍在继续

与去年(幻灯片33)类似，这场比赛是在视频扩散和掩码Transformer模型之间进行的(尽管从算法上看，这两者非常相似)。去年的Make-a-video和Imagen基于扩散，而Phenaki基于双向掩码transformer。

- VideoLDM是一款潜在扩散模型，能够生成高分辨率视频(高达 $1280 \times 2048!$ )。它们基于预训练的图像扩散模型，通过时间对准层进行时间微调，将它们变成视频生成器。
- MAGVIT是一个掩码的生成视频转换器。与Phenaki类似，它使用3D标记器来提取时空标记。它引入了一种新颖的掩码方法。它目前在视频生成基准测试中拥有最佳FVD，比视频传播快250倍。



*"The Orient Express driving through a fantasy landscape, animated oil on canvas"*

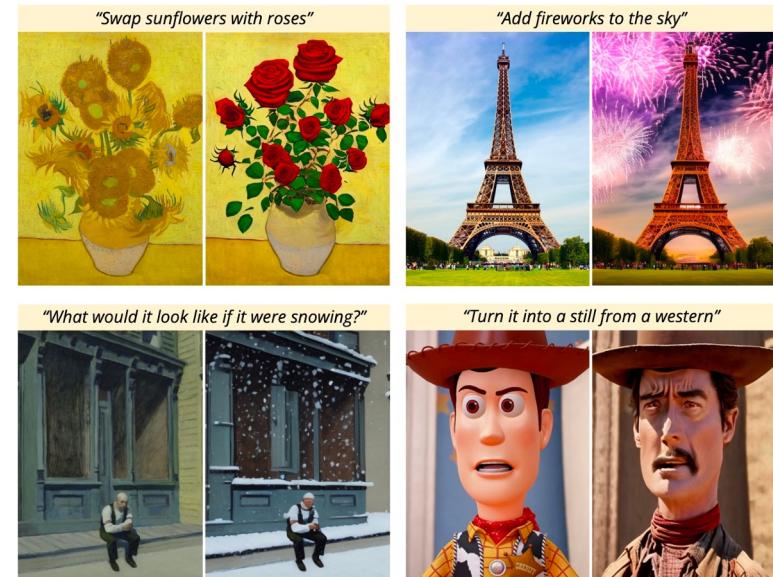


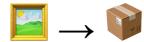
## 基于指令的文本图像生成编辑助手

腾讯科技

▶ 去年出现了许多文本图像生成模型:DALLE-2、Imagen、Parti、Midjourney、Stability等等。但是控制生成需要对提示和自定义语法进行大量实验。今年出现了新的方法，让图像生成和编辑能力拥有了Co-Pilot风格。

- InstructPix2Pix利用预训练的GPT3和StableDiffusion来生成{输入图像、文本指令、生成的图像}三元组的数据集，以训练受监督的条件扩散模型。然后，以前馈方式进行编辑，无需对每个图像进行任何微调/反转，从而在几秒钟内完成修改。
- 诸如Imagen Editor之类的掩码修复方法需要为模型提供一个覆盖图或“掩码”，以指示要修改的区域以及文本指令。
- 在这些方法的基础上，Genmo AI等初创公司提供了一个Co-Pilot风格的界面，通过文本引导的语义编辑来生成图像。





## 欢迎3D Gaussian Splatting

▶ 全新的基于3D Gaussian的NeRF竞争者展示了令人印象深刻的质量，同时也支持实时渲染。



MipNeRF360 [Barron '22]

0.06 fps

训练: 48小时, PSNR:

27.69

3D Gaussian Splatting

134 fps

训练: 41分钟, PSNR:

27.21

\*请注意，Zip-NeRF在同一数据集上的训练时间为53分钟，PSNR为28.54。

- 3D Gaussian Splatting不是学习神经网络的参数，而是学习数百万个高斯分布(每个3D点一个)，并通过计算每个高斯对最终图像中每个像素的贡献来执行光栅化。
- 需要更多表现能力的区域使用更多高斯函数，同时避免在空白空间进行不必要的计算，这就是为什么与NeRFs类似，场景看起来如此精美细致。
- 现在可以以1080p的分辨率呈现高质量的实时( $\geq 100$  fps)小说视图。



## NeRFs遇见GenAI



▶ 基于NeRF的生成模型是大规模创建3D资产的一个有前途的方向。NeRFs不仅在速度和质量上有所提高（参见超扩散、MobileNeRF、Neurolangelo和DynIBAR），而且使GenAI能够模拟3D几何形状。

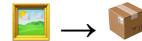
- DreamFusion和Score Jacobian Chaining是使用预训练的2D文本到图像扩散模型来执行文本到3D合成的第一种方法。早期的尝试展示了单一物体的卡通式3D模型。
- RealFusion微调特定图像上的扩散先验，以增加该图像的似然性。
- SKED只改变通过一些引导草图提供的NeRF的选定区域。它们保留了基本NeRF的质量，并确保编辑的区域尊重文本提示的语义。
- Instruct-Nerf2Nerf编辑整个Nerf场景，而不是一个区域或从头开始生成。它们在每个输入图像上应用潜在扩散模型，并迭代更新NeRF场景，确保它保持一致。



a robot and dinosaur playing chess, high resolution\*



Original NeRF "Turn him into a firefighter with a hat" "As a bronze statue" "Turn him into a clown"

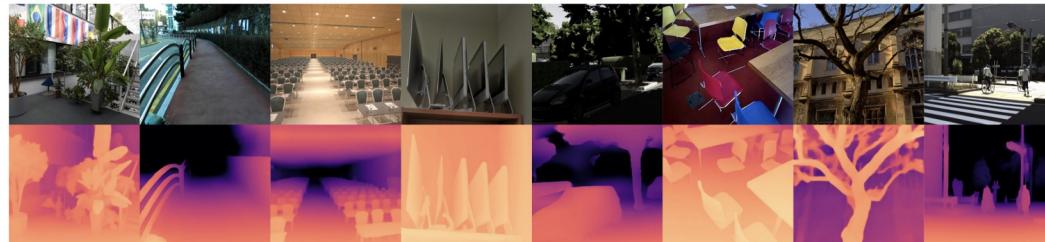


## 零样本度量深度



▶ 零样本深度模型最近被用作更好的图像生成的条件。这仅需要相对深度预测，而机器人等其他下游应用需要度量深度。迄今为止，这还没有很好地跨数据集推广。

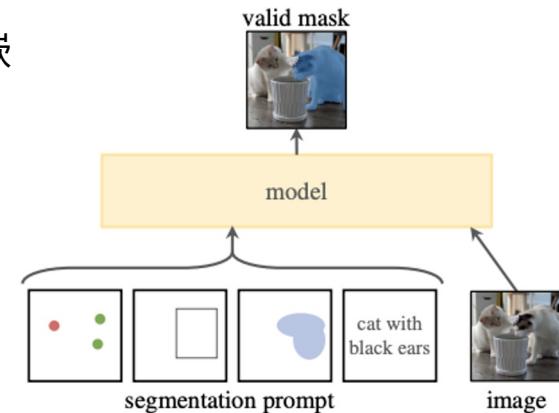
- 《零深度：走向零镜头尺度感知的单目深度估计》能够预测来自不同域和不同相机参数的图像的度量深度。它们共同编码图像特征和摄像机参数，使网络能够推理物体的大小，并在一个变化的框架中训练。深度网络最终学习可以跨数据集转移的“尺度先验”。
- 《ZoeDepth:结合相对和度量深度实现Zero-shot迁移》是一个相对深度模型，带有一个针对公制深度进行微调的附加模块。这是第一个在多个数据集上训练的模型，性能没有明显下降，并且能够跨室内和室外领域推广。



## Segment Anything：一个具有零样本泛化的可提示分割模型

▶ Meta推出了一个名为“Segment Anything”的大规模项目，其中包括在11M图像数据集(SA-1B)上发布1B分段掩码，以及一个带有Apache 2.0商业使用许可证的分段模型(SAM)。Meta在23个域外图像数据集上测试了SAM，在70%以上的情况下优于现有的SoTA。

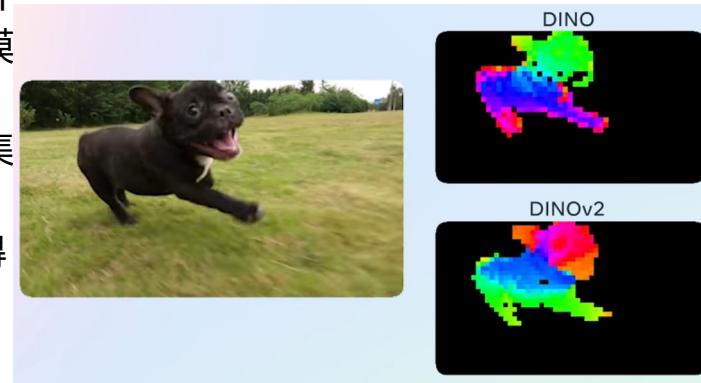
- Meta研究人员从大型语言模型中获得灵感，这些模型在大量数据集上进行了预训练，并通过提示展示了零次学习(Zero-Shot)能力，他们开始建立一个能够实现一般提示分割的模型：给定任何提示，该模型应该能够识别和分割任何图像中的任何对象。
- 该模型具有两个组件：(i)计算一次性图像嵌入的重量级编码器(ViT)，(ii)由嵌入用户提示的提示编码器和预测分段掩码的掩码解码器组成的轻量级交互模块(可以在浏览器中的CPU上运行)。
- 模型在环数据引擎用于生成训练数据，最终的SA-1B完全通过应用SAM自动生成。
- 通过提示工程，SAM可以应用于其他任务，包括边缘检测、对象提议生成和实例分割，并且结合SAM + CLIP显示了用于文本提示的初步结果。



## DINOv2：新的默认计算机视觉主干

► DINOv2是Meta的自我监督视觉转换器模型，产生通用的视觉特征，可用于各种图像级(例如分类)和像素级(例如分割)任务，无需微调，可与SOTA开源的弱监督替代品竞争。

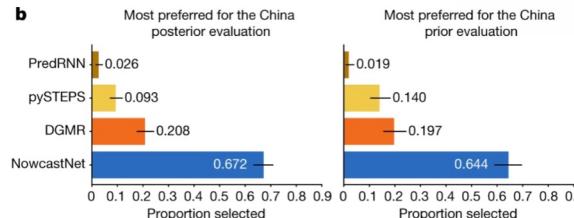
- 这是第一个缩小自我监督和弱监督方法之间差距的工作。DINOv2特征被示出包含关于对象部分的信息以及图像的语义和低级理解。
- 作者通过额外的正则化方法使自我监督学习模型的训练更加稳定，并降低了内存要求，这使得能够在更多数据上更长时间地训练更大的模型。他们还提供了通过Distillation (蒸馏) 获得的模型的压缩版本。
- 尽管任何图像都可以用于训练，但一个关键的组成部分是管理数据集并在概念之间自动平衡它(从1.2亿幅源图像中保留1.42亿幅)。
- DINOv2功能可以与线性分类器一起使用，以在许多视觉任务中获得强有力的结果。



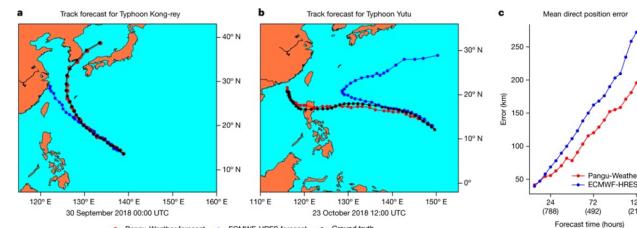
## 在当前和更长时间范围内更准确的天气预报

▶ 当前熟练的短期降水预测(临近预报)模糊不清,容易消散,而且速度缓慢。使用精确数值天气预报方法的中期全球天气预报在计算上是昂贵的。对于这两个问题,结合相关先验知识的学习方法和物理学模型能够提供专业气象学家更喜欢的性能改进。新的基准数据集,如谷歌的WeatherBench 2,有助于数据驱动的天气模型开发。

- NowcastNet是一个非线性模型,使用物理第一原则和统计学习方法,统一在一个深度生成模型框架下。由来自中国各地的62名专业气象学家评估,该模型在71%的情况下与领先方法相比排名第一。

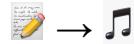


- 盘古天气是一个3D深度学习模型,具有地球特定的先验知识,基于39年的全球数据进行训练,可以生成中期全球天气。该系统可用于更准确的早期气旋跟踪对比现状。



HUAWEI CLOUD

stateof.ai 2023



## 音乐生成又一年的进步



▶ 来自谷歌、Meta和开源社区的新模型极大地提高了可控音乐生成的质量。

- 虽然就生成的音乐质量而言不是最好的，但Riffusion可能是最具创新性的模式。研究人员微调了光谱图图像的稳定扩散，然后将其转换为音频剪辑。
- 通过MusicLM，谷歌研究人员“将条件音乐生成视为一项分层的seq2seq建模任务”。他们能够在几分钟内产生一致的音乐(@24kHz)。样本网址为：<https://google-research.github.io/seanet/musiclm/examples/>
- 对我们而言，Meta的MusicGen在坚持文本描述和生成愉快的旋律之间取得了更好的平衡。它使用单个变换器语言模型和仔细的码本交织技术。样本网址为：<https://ai.honu.io/papers/musicgen/>



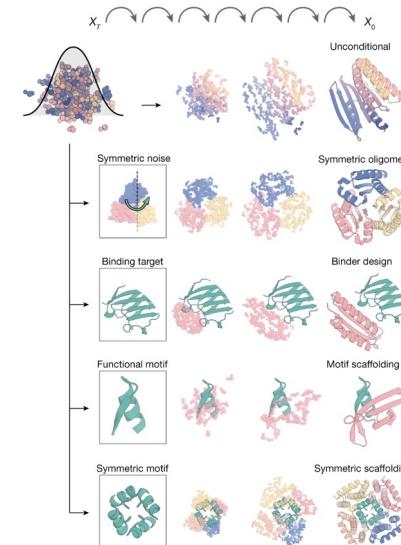
funk bassline with a jazzy saxophone solo

MODEL	MUSICCAPS Test Set				
	FAD <sub>vgg</sub> ↓	KL ↓	CLAP <sub>scr</sub> ↑	OVL. ↑	REL. ↑
Riffusion	14.8	2.06	0.19	79.31±1.37	74.20±2.17
Mousai	7.5	1.59	0.23	76.11±1.56	77.35±1.72
MusicLM	4.0	-	-	80.51±1.07	82.35±1.36
Noise2Music	<b>2.1</b>	-	-	-	-
MUSICGEN w/o melody (1.5B)	3.4	1.23	0.32	80.74±1.17	<b>83.70±1.21</b>
MUSICGEN w/o melody (3.3B)	3.8	1.22	0.31	<b>84.81±0.95</b>	82.47±1.25
MUSICGEN w. random melody (1.5B)	5.0	1.31	0.28	81.30±1.29	81.98±1.79

## 扩散模型从简单的分子规格设计不同的功能蛋白质

▶ 从零开始设计新的蛋白质，使它们具有期望的功能或结构特性，从头设计，在研究和工业中都是令人感兴趣的。受图像和语言生成模型成功的启发，扩散模型现在被应用于蛋白质从头设计工程。

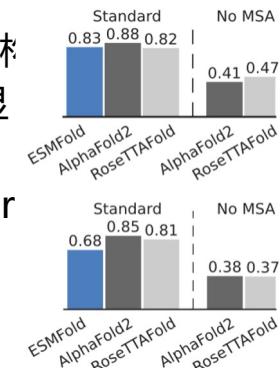
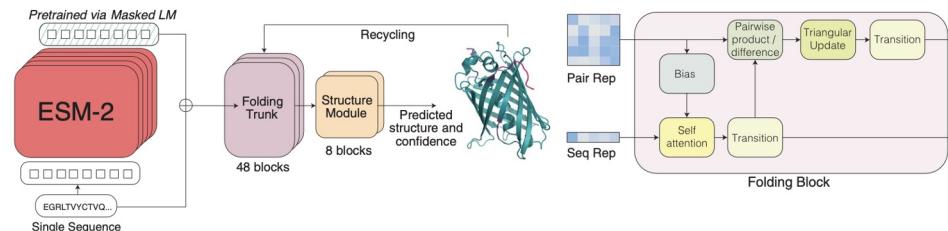
- 一种称为RFdiffusion的模型利用RoseTTAFold的高精度、残基水平分辨率蛋白质结构预测能力，将其微调为使用蛋白质数据库中有噪结构的生成扩散模型中的去噪网络。
- 与AlphaFold 2类似，当模型以时间步长之间的先前预测为条件进行去噪时，RFdiffusion得到最佳训练。
- RFdiffusion可以产生具有所需特征的蛋白质架构，然后ProteinMPNN可以用于设计编码这些产生的结构的序列。
- 该模型可以产生蛋白质单体、蛋白质结合物、对称寡聚体、酶活性位点支架等的架构设计。



## 用语言模型学习进化尺度的蛋白质结构规则

现在可以从氨基酸序列直接预测原子水平的蛋白质结构，而不依赖于昂贵且缓慢的多序列比对(MSA)。为此，对数百万进化上不同的蛋白质序列使用屏蔽语言建模目标，以使生物结构在语言模型中具体化，因为它与序列模式相关联。

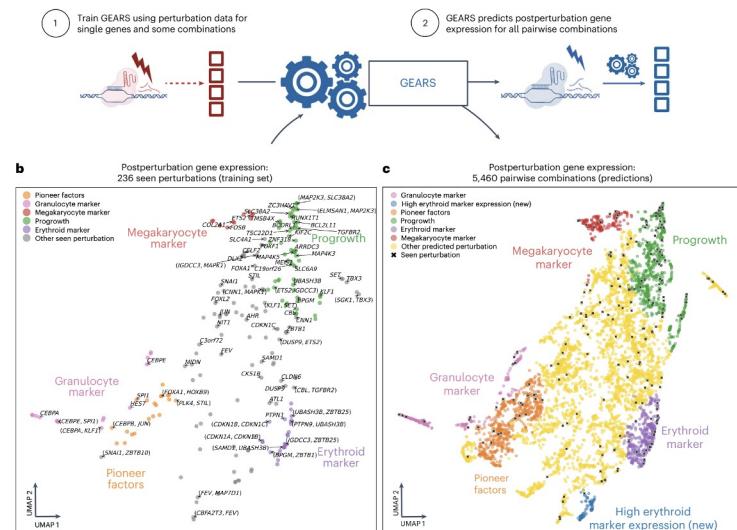
- 模型Evolutionary Scale Modeling-2 (ESM-2)用于表征超过6.17亿宏基因组蛋白质的结构(在土壤、细菌、水等中发现)。与AlphaFold-2 (AF2)相比，ESM-2(示意图如下)提供了显著的加速：这些结果是使用2000个GPU的集群在2周内产生的。
- ESMFold是一个完全端到端的单序列结构预测器，它使用ESM-2的折叠头。通过TM-score测量，ESMFold结构(右)具有AF2级质量，TM-score是与地面真实结构相比的投影精度。



## 在没有基于细胞的实验的情况下预测干扰多个基因的结果

了解基因表达如何因刺激或抑制基因组合（即扰动）而改变，对于揭示与健康和疾病相关的生物途径非常重要。但是组合爆炸阻止了我们在实验室的活细胞中进行这些实验。将深度学习与基因—基因关系的知识图相结合提供了一种解决方案。

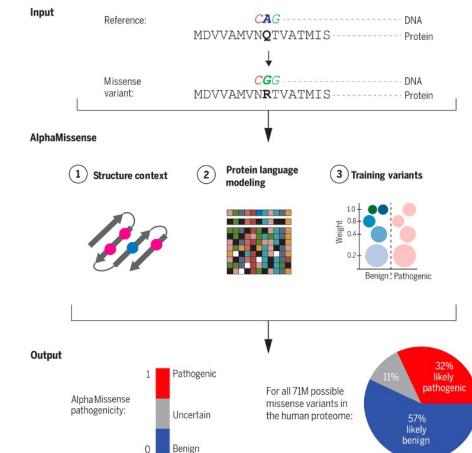
- 图形增强的基因激活和抑制模拟器（GEARS）结合先前的实验知识来预测给定未扰动的基因表达和应用的扰动的基因表达结果。
- 例如，GEARS可以在单基因和双基因实验扰动后的基因表达谱上进行训练(b)，然后负责预测5460个成对组合的扰动后基因表达(c)。

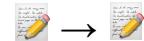


## 致病与否？预测所有单个氨基酸变化的结果

由遗传变异(“错义变体”)导致的氨基酸序列的个体变化可能是良性的，也可能导致蛋白质折叠、活性或稳定性下游问题。通过人类群体水平的基因组测序实验，已经鉴定出超过400万个这些错义变体。然而，98%的这些变异缺乏任何确认的临床分类(良性/致病性)。新系统AlphaMissense利用AlphaFold预测和无监督的蛋白质语言建模来弥合这一差距。

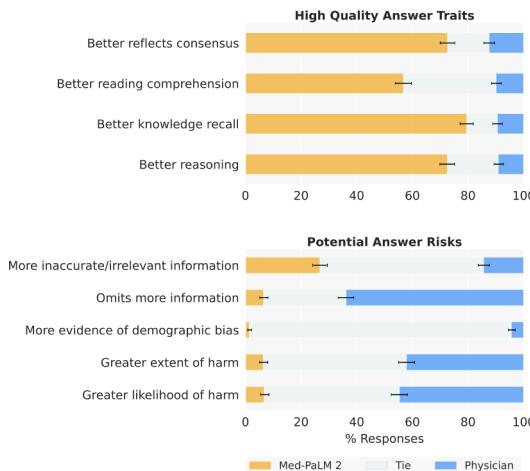
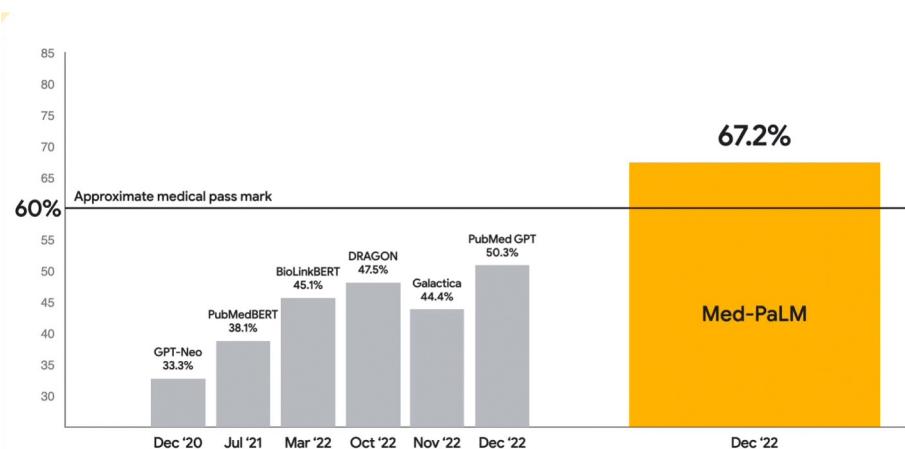
- AlphaMissense系统通过以下方式构建:(i)对来自群体频率数据的弱标签进行训练，通过不使用人类注释来避免循环；(ii)结合无监督的蛋白质语言建模任务，以学习以序列环境为条件的氨基酸分布；和(iii)通过使用AlphaFold衍生的系统整合结构背景。
- 然后，AlphaMissense被用来预测7100万错义变体，使人类蛋白质组饱和。其中，32%可能是致病性的，57%可能是良性的。其他资源包括19,233种标准人类蛋白质中所有2.16亿种可能的单氨基酸替代物。





## 谷歌的Med-PaLM 2语言模型是USMLE的专家

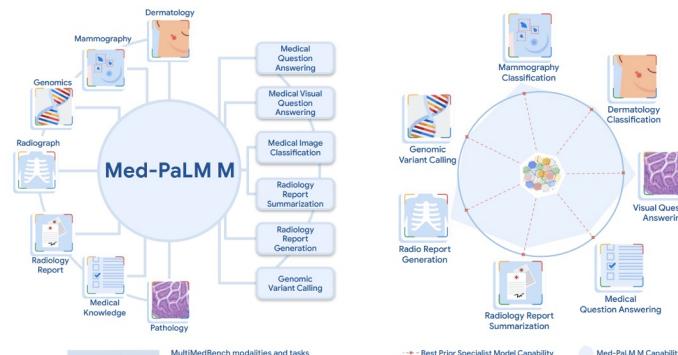
▶ Med-PaLM是第一款在美国医疗执照考试(USMLE)中超过“及格”分数的模型，在发布一年后，由于基础LLM改进、医疗领域微调和提示策略，Med-PaLM 2在更多数据集上创造了新的SOTA结果。在一项针对1066个消费者医疗问题的成对排名研究中，在我们的评估框架中，一组医生在九个问题中的八個问题上首选Med-PaLM 2的答案。





## 接下来，Med-PaLM走向多模态

为了超越基于文本的医疗问答，谷歌首先创建了MultiMedBench - a 14任务数据集，包括医疗问答、乳房x线摄影和皮肤病学图像解释、放射学报告生成和总结以及基因组变异调用。该数据集用于训练具有相同模型权重集的大型单个多任务、多模态版本的MedPaLM。该系统展示了新颖的应急能力，例如对新颖的医学概念和任务的概括。还提出了一种替代的轻量级方法ELIXR。ELIXR将语言对齐的视觉编码器移植到固定的大型语言模型上，这需要更少的计算来训练，并在包括视觉问答、语义搜索和零样本分类在内的任务中显示出前景。

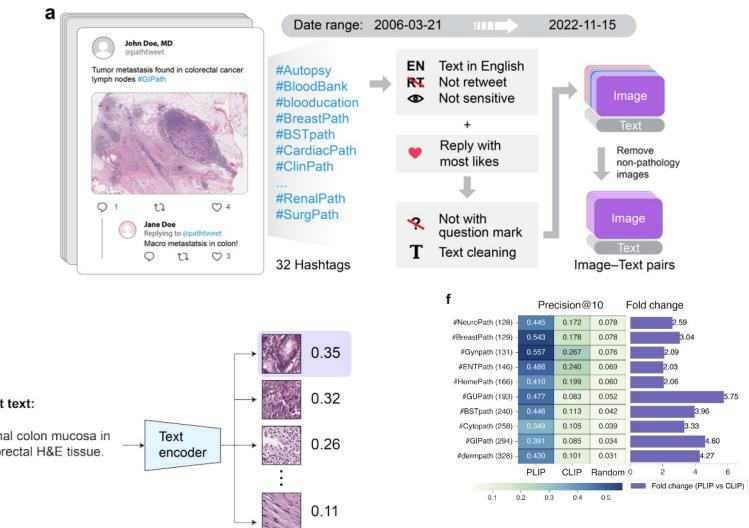




## Tweet Storm: 根据推特医学数据创建的SOTA病理学语言图像预训练模型

众所周知,(高质量)数据是构建有能力的人工智能系统的王道, 尤其是在临床医学等(高质量)数据生产成本高昂的领域。这项工作在推特上挖掘文本-图像对, 以创建OpenPath数据集, 该数据集包含200多幅病理图像和自然语言描述符。受OpenAI的对比语言-图像预训练(CLIP)模型的启发, 作者创建了P(athology)LIP。

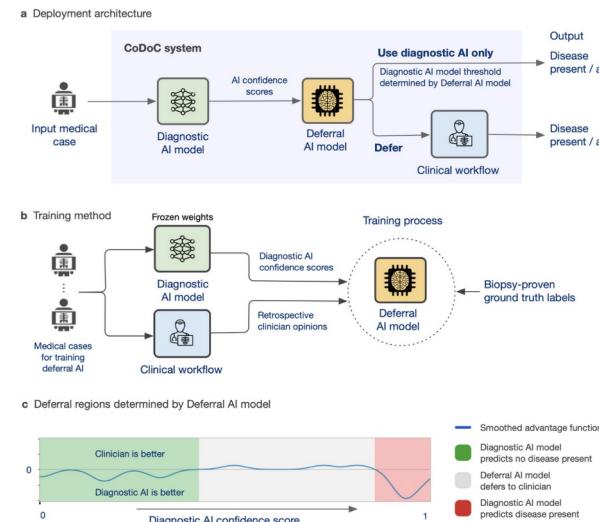
- 像CLIP一样, PLIP可以对看不见的数据进行零样本分类, 使其能够区分几种关键的组织类型。
- 它还可以用于改进病理图像的文本到图像和图像到图像的检索。
- 与数字病理学中基于从固定标签集学习的其他机器学习方法不同, PLIP可以更普遍地应用, 并且可以灵活地适应病理学中诊断标准的变化性质。
- 与CLIP相比, PLIP的精度提高了2-6倍。



## 基于真实世界的自动医学图像分析临床系统设计

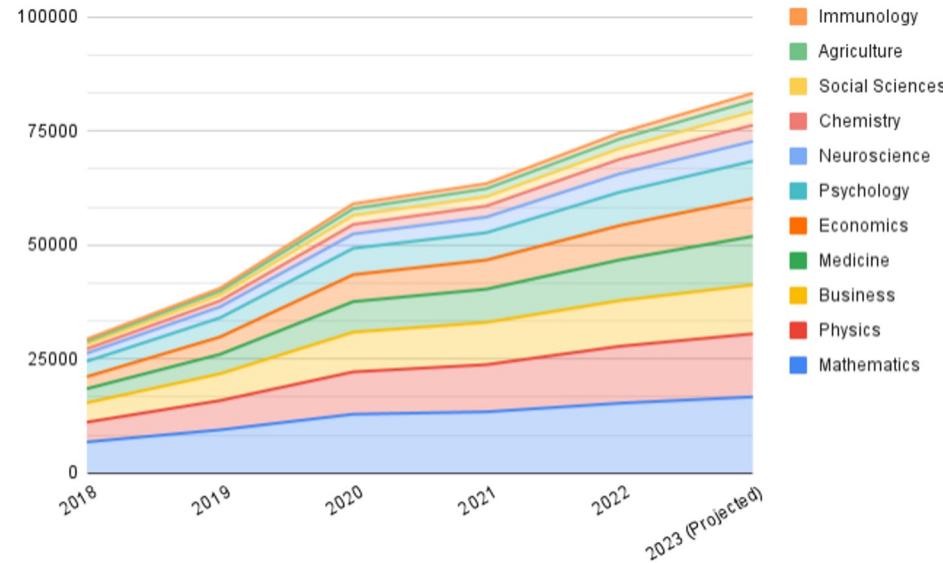
▶ 计算机视觉已被证明对乳房x光片上的乳腺癌筛查和结核病分类是有用的。然而，为了在临床中实现实际和可靠的使用，知道何时依赖预测性人工智能模型或恢复到临床工作流程是很重要的。

- 互补性驱动的临床工作流程延迟(CoDoC)学会决定是依赖预测性人工智能模型的输出，还是改为遵从临床工作流程。
- 对于乳腺癌筛查，CoDoC在相同的假阴性率下将假阳性率降低了25%。相比之下，英国的双读仲裁减少了25%。重要的是，临床工作量因此减少了66%。



## 人工智能科学：医学发展最快，但数学最受关注

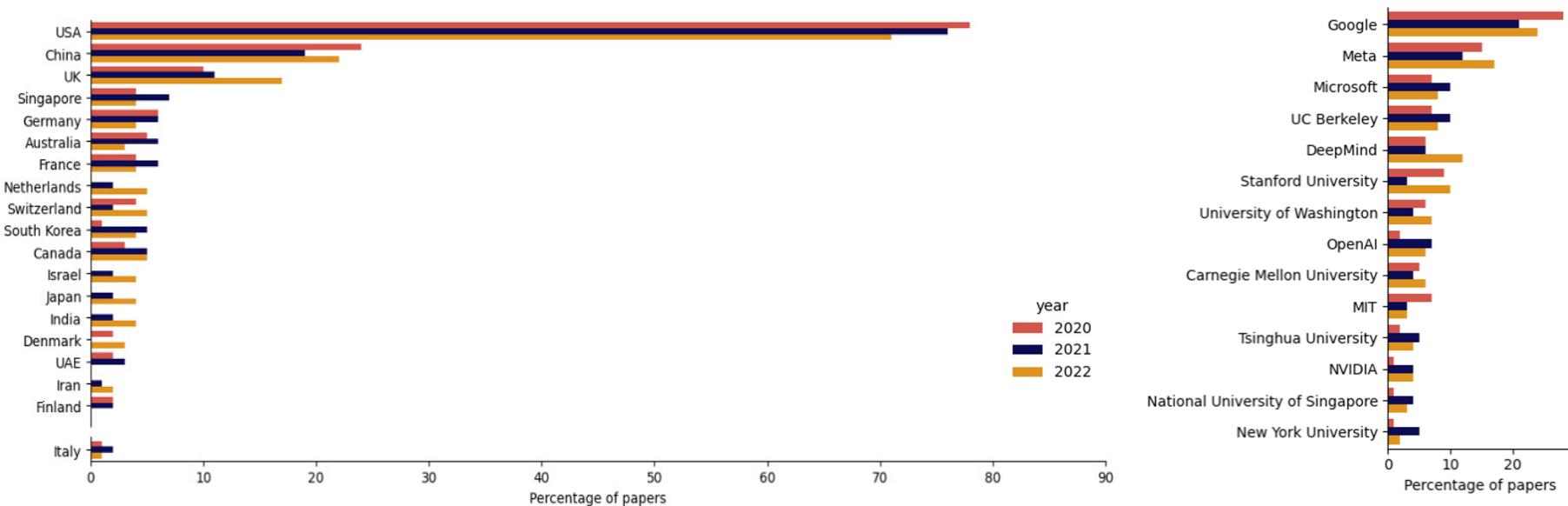
▶ 应用人工智能加速进步的前20个科学领域包括物理、社会、生命和健康科学。在所有出版物中，增长最快的是医学。由于人工智能在科学中的应用，我们预计在可预见的未来会有重大的研究突破。



## 大多数有影响力的研究来自极少数地方

▶ 在过去3年中，超过70%被引用最多的人工智能论文的作者来自美国的机构和组织。

(关注腾讯科技微信公众号 (qqtech) , 回复 “AI2023” 可免费获取本报告PDF版。)

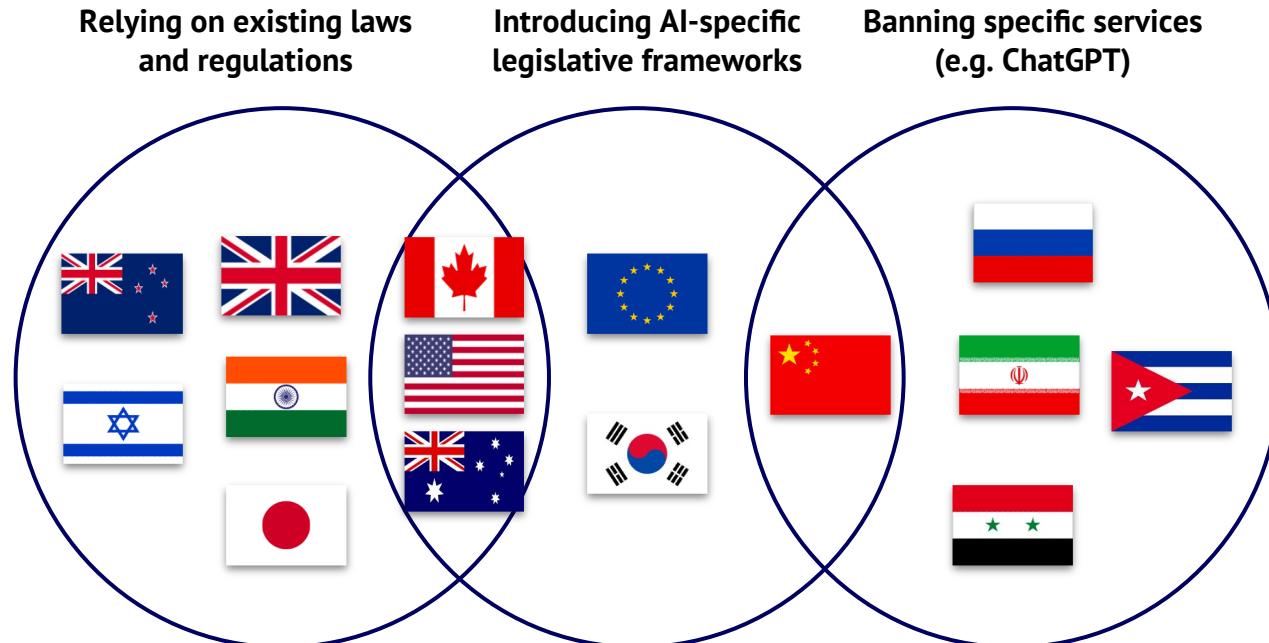




## 第三章：政策

## 我们是否已经达到监管分歧的“峰值”？

在对监管方法的潜在分歧进行了多年的猜测后，我们开始看到监管方法趋于稳定，并达成少数几种不同的方法。



## “宽松”还是“支持创新”：对大规模监管的怀疑

▶ 以英国和印度为代表，这种方法的运作基础是人工智能目前不需要任何额外的立法。

- 到目前为止，英国和印度都强调了人工智能的经济和社会好处，2023年3月的白皮书和印度数字部长的议会回应认为，当前的任何风险都可以被当前的行业法规和隐私立法吸收。
- 然而，英国确实纳入了一些人工智能原则（基于经合组织的类似工作）供监管机构遵循，并向一个专注于前沿模型安全的工作组投资了1亿英镑，该工作组由SOAI合著者伊恩·霍加斯（Ian Hogarth）领导。该团队似乎是世界首创，试图建立一个专门的单位，利用工业界和学术界来评估前沿风险。
- 英国还与谷歌DeepMind、Anthropic和OpenAI达成了一项特别协议，以尽早获得他们最先进的前沿模型，提高他们对风险的理解。
- 虽然受到业界的欢迎，但尚不清楚这些方法是否会继续存在。最近，英国政府从其词汇中删除了“轻触”（light-touch），并对自己重新定位为人工智能安全辩论的发源地。
- 印度电子和信息技术部现在表示，即将出台的立法可能确实会涵盖一些形式的人工智能危害，以及Web3和其他技术。

## 范围广泛的立法



### ▶ 欧盟在通过新的人工智能专门立法方面处于领先地位，对基础模型采取了特别严格的措施

- 欧盟的人工智能法案现在正进入最后的立法阶段，今年早些时候进行了修订，增加了围绕基础模型和通用人工智能系统的特别法规（单独规定）。
- 虽然人工智能法案的其余部分根据系统预期用途的“高风险”程度对要求进行分层，但所有商业基础模型提供商都受到特殊要求的约束。
- 这些包括风险评估，披露内容何时生成人工智能，防止模型生成非法内容，以及发布用于培训的任何版权数据的摘要。

## 混合模式：两全其美还是最差？



► 在其他市场，我们要么看到国家监管的放松，要么看到地方法律的强势。虽然避免了主要立法的一些挑战，但它们也冒着“让谁都不会高兴”的风险。

- 美国不太可能很快通过联邦人工智能法律，在某些方面正在追求英国式的方法，重点是自愿承诺（如7月白宫协议）和建立良好实践的研究（如国家标准与技术研究所的人工智能风险管理框架）。例如，其中一些承诺涉及第三方评估，但没有具体说明这将是哪一个第三方，公司理论上可以忽略他们的数据。
- 美国各州一直在引入严格程度不同的人工智能法律。在加利福尼亚州、科罗拉多州、德克萨斯州、弗吉尼亚州和其他地方，我们有关于“支持”和自动决策的强制性透明度法律。与此同时，纽约州和伊利诺伊州有关于在招聘决策中使用人工智能的具体法律。
- 加拿大正在尝试精简版的欧盟人工智能法案，禁止某些应用，同时也对其他应用进行监管。加拿大的《人工智能和数据法案》只监管“高风险”应用程序，而不是欧盟式的浮动责任比例。执法将落到现有的部门，而不是新的监管机构。
- 这种方法受到了来自两方面的攻击——批评者们指责它“过了”或是“不及”。

## 关于全球治理的国家行动正处于早期阶段...

▶ 国际原子能机构、政府间气候变化专门委员会和欧洲核子研究中心等全球监管机构都被视为典范。然而，这些建议目前仍局限于学术论文。

- 英国计划在2023年11月举办一次以安全和治理为主题的峰会，它试图将自己定位为世界领先的安全研究中心。
- 欧盟和美国宣布，他们正在制定一项联合人工智能行为准则，其中将包括关于风险审计和透明度的非约束性国际标准等要求。
- G7将与经合组织和人工智能全球伙伴关系合作，创建“广岛人工智能进程”，这将为生成式人工智能治理设定一种“集体方法”
- 我们也看到了联合国的第一步，在秘书长技术特使办公室主办的2024年“未来峰会”之前，联合国在8月份举行了一次磋商，为治理建议提供信息。

Function → Objective / institutions ↓	Science and Technology Research, Development and Diffusion				International Rulemaking and Enforcement			
	Conduct or Support AI Safety Research	Build Consensus on Opportunities and Risks	Develop Frontier AI	Distribute and Enable Access to AI	Set Safety Norms and Standards	Support Implementation of Standards	Monitor Compliance	Control Inputs
Spreading Beneficial Technology	No	Yes	Maybe	Yes	No	No	No	No
Harmonizing Regulation	No	No	No	No	Yes	Yes	No	No
Ensuring Safe Development and Use	Maybe	Yes	Maybe	Maybe	Yes	Yes	Maybe	Maybe
Managing Geopolitical Risk Factors	No	No	Maybe	Maybe	No	No	Yes	Yes
Existing Int'l Institutional Efforts		OECD, GPAI, G7, ITU			ISO/IEC			Semi-conductor Export Controls
Possible Institution	AI Safety Project	Commission on Frontier AI	Frontier AI Collaborative	Advanced AI Governance Agency				
Key challenges	Model access; diverting talent	Politicization; scientific challenges	Managing dual-use technology; education, infrastructure and ecosystem obstacles	Incentivizing participation; quickly changing risk landscape; maintaining appropriate scope				

## .....因此，最大的“AI实验室们”正试图填补真空

▶ 随着政府努力在治理问题上形成深刻、共同的立场，前沿模型的开发者正在推动规范的形成。

- Anthropic、谷歌、OpenAI和微软发起了前沿模型论坛——一个旨在促进前沿模型负责任开发并与决策者分享知识的机构。
- 2023年5月至6月，OpenAI首席执行官萨姆·奥特曼（Sam Altman）进行了一次“世界之旅”，拜会了关键市场的政策制定者和监管者。奥特曼的建议包括强大模型的许可制度和引入独立审计要求。
- “AI实验室们”也提出了自己的政策建议。OpenAI认为，我们最终将需要一个IAEA来审计和执行安全标准；而Anthropic概述了其治理的“组合”方法，将立法、独立审计和健全的内部控制结合起来。影响最深远的是Inflection的穆斯塔法·苏莱曼(Mustafa Suleyman)，他与欧亚集团(Eurasia Group)的伊恩·布雷默(Ian Bremmer)共同勾勒了一个三层全球治理结构。

### OpenAI's CEO Goes on a Diplomatic Charm Offensive

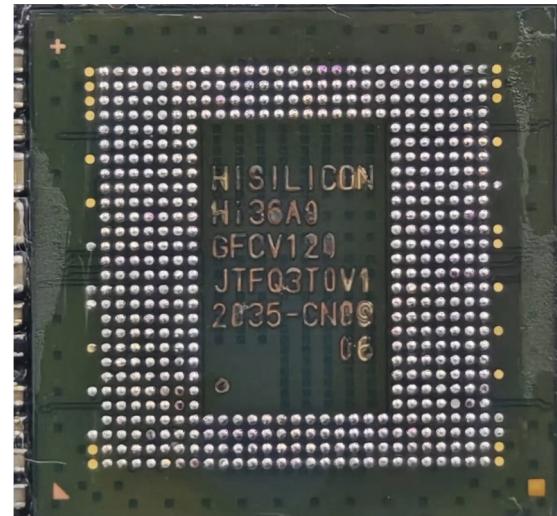
Sam Altman's global travels may be more opportunistic than altruistic.

### Frontier Model Forum

We're forming a new industry body to promote the safe and responsible development of frontier AI systems: advancing AI safety research, identifying best practices and standards, and facilitating information sharing among policymakers and industry.

## 华为的新芯片是否标志着突破时刻？

▶ 在日益严格的限制下，华为以其新的Mate 60 Pro手机震惊了世界，该手机由中国企业自主生产的先进的麒麟9000S芯片提供动力。



## 政府正在扩大算力，但落后于私营部门的努力

▶ 目前，欧盟和美国处于非常有利的地位，但他们的国家HPC集群Leonardo和Perlmutter并不致力于人工智能，资源与其他研究领域共享。与此同时，英国目前可供研究人员使用的公共云中的英伟达A100 GPU不到1000颗。

- 英国的compute review建议建立一个由3000个GPU组成的集群，供学者和商业用户访问，并设定2026年为实现百亿亿次级(exascale)能力的最后期限。
- 美国计划建立国家人工智能研究资源，使研究人员可以在四核GPU节点上工作1.4亿至1.8亿小时。目前，该公司正等待国会批准所需的26亿美元六年投资。
- 然而，两国政府都面临着更进一步的呼声。Anthropic建议美国应该投资40亿美元创建一个10万GPU集群，而托尼布莱尔研究所(Tony Blair Institute)则推动英国创建一个3万GPU集群。
- 与此同时，私人公司正在争相购买他们能找到的每一颗GPU。据悉，在2023/2024年期间，中国互联网大公司已经在英伟达订单上花费了90亿美元。

## 人工智能和国防科技吸引了创纪录的资金...但是政策阻碍了进步吗？

▶ 美国和欧洲军方一直力图实现多元化，以确保他们不会错过最新的能力进步，但赢家的数量仍然很少。

- 去年，美国国防初创公司的资金达到24亿美元，是欧洲总额的100多倍，但能够赢得持续工作的公司数量仍然很少。一个由美国风投和科技公司组成的联盟呼吁对“开发需求和选择技术的过时方法”进行改革。
- Anduril的E轮投资超过了2013-2022年间所有英国国防技术投资的总和，而Helsing的2.09亿英镑B轮融资是欧洲大陆唯一的重大融资。欧洲有限合伙人在很大程度上没有扭转他们对国防投资的厌恶，这意味着超国家机构正在填补缺口。
- 除了新的10亿欧元北约创新基金，欧洲投资基金被认为已经为国防投资拨款2亿欧元。这些基金是会大举押注，还是会变得谨慎，创造一个欧洲“死亡之谷”？



# 人工智能会出现“文化战争”吗？

▶ ChatGPT已经成为一系列激烈的文化辩论的焦点，主要是在美国，特别是保守派分享截图，指控ChatGPT的训练和微调存在偏见。

- 作为回应，OpenAI发布了一篇博客文章，详细介绍了其审核方法。奥特曼建议，未来人们可能能够超越一些“非常宽泛的绝对规则”来微调ChatGPT迭代，以将OpenAI从这些价值观问题中移除。
- 在长期抱怨OpenAI的“政治正确性”之后，马斯克推出了xAI，这是一家专注于尝试“理解宇宙真实本质”的初创公司。马斯克后来在社交媒体X上强调，“我们的人工智能可以给出人们可能会发现有争议的答案，即使它们实际上是真的”。外界目前对xAI的工作知之甚少。
- 今年8月，政治科学期刊《公共选择》(Public Choice)发表了一项研究，发现ChatGPT表现出“强烈而系统的政治偏见……明显倾向于左派”，这反过来又引来了一系列批评回应。

## Inside the AI culture war

By BEN SCHRECKINGER | 05/15/2023 04:00 PM EDT

With help from Derek Robertson



The OpenAI logo displayed on a phone. | AP

Even the world's fastest-developing technology cannot outrun the culture war.

## 人们对失业的担忧日益加剧，但政策制定者正采取观望态度

来自OECD（经合组织）和OpenAI的研究表明，我们将很快看到技术专业的大批失业，包括法律、医学和金融等行业。经合组织警告说，多达27%的工作属于“高风险”职业

- 有人呼吁（例如德隆·阿西莫格鲁和西蒙·约翰逊）以“专业工作者”的方式重新引导人工智能的发展——从自动化人类任务转向增强人类决策。然而，以必要的精度预测创新所需的预测能力将是巨大的。
- 更乐观的是，有迹象表明，人工智能可以充当技能平衡器。一篇论文发现，在18项不同的任务中，使用GPT-4的顾问明显优于那些没有使用的顾问，而针对法律、客户协助工作和创造性写作的研究发现，表现不佳的人表现更好。
- 工业界基本上保持沉默，但奥特曼、戴密斯·哈萨比斯（Google DeepMind）和穆斯塔法·苏莱曼（Inflection）等声音都表示支持全民基本收入。

Group	Occupations with highest exposure	% Exposure
<b>Human <math>\alpha</math></b>	Interpreters and Translators	76.5
	Survey Researchers	75.0
	Poets, Lyricists and Creative Writers	68.8
	Animal Scientists	66.7
<b>Human <math>\beta</math></b>	Public Relations Specialists	66.7
	Survey Researchers	84.4
	Writers and Authors	82.5
	Interpreters and Translators	82.4
	Public Relations Specialists	80.6
<b>Human <math>\zeta</math></b>	Animal Scientists	77.8
	Mathematicians	100.0
	Tax Preparers	100.0
	Financial Quantitative Analysts	100.0
	Writers and Authors	100.0
	Web and Digital Interface Designers	100.0
<i>Humans labeled 15 occupations as "fully exposed."</i>		
<b>Model <math>\alpha</math></b>	Mathematicians	100.0
	Correspondence Clerks	95.2
	Blockchain Engineers	94.1
	Court Reporters and Simultaneous Captioners	92.9
	Proofreaders and Copy Markers	90.9
<b>Model <math>\beta</math></b>	Mathematicians	100.0
	Blockchain Engineers	97.1
	Court Reporters and Simultaneous Captioners	96.4
	Proofreaders and Copy Markers	95.5
	Correspondence Clerks	95.2
<b>Model <math>\zeta</math></b>	Accountants and Auditors	100.0
	News Analysts, Reporters, and Journalists	100.0
	Legal Secretaries and Administrative Assistants	100.0
	Clinical Data Managers	100.0
	Climate Change Policy Analysts	100.0
	<i>The model labeled 86 occupations as "fully exposed."</i>	
<b>Highest variance</b>	Search Marketing Strategists	14.5
	Graphic Designers	13.4
	Investment Fund Managers	13.0
	Financial Managers	13.0
	Insurance Appraisers, Auto Damage	12.6



## 第四章：安全

## 灾难性人工智能风险的快速分类



► 在我们深入讨论人工智能风险新闻和辩论之前，让我们记住（一些）人工智能安全研究人员认为是灾难性的人工智能风险。与本节的所有内容一样，这些内容应该有所保留。像任何涉及预测的领域一样，尤其是依赖于数据（几乎）的黑盒系统，对于人工智能系统在合理的时间范围内带来的实际风险，没有达成共识。下面的图表摘自非营利组织人工智能安全中心主任丹·亨德里克斯（Dan Hendrycks）的《灾难性人工智能风险概述》。

### Malicious Use



- ✖ Bioterrorism
- ✖ Surveillance State
- ✓ Access Restrictions
- ✓ Legal Liability

### AI Race



- ✖ Automated Warfare
- ✖ Evolutionary Pressures
- ✓ International Coordination
- ✓ Safety Regulation

### Organizational Risks



- ✖ Weak Safety Culture
- ✖ Leaked AI Systems
- ✓ Information Security
- ✓ External Audits

### Rogue AIs

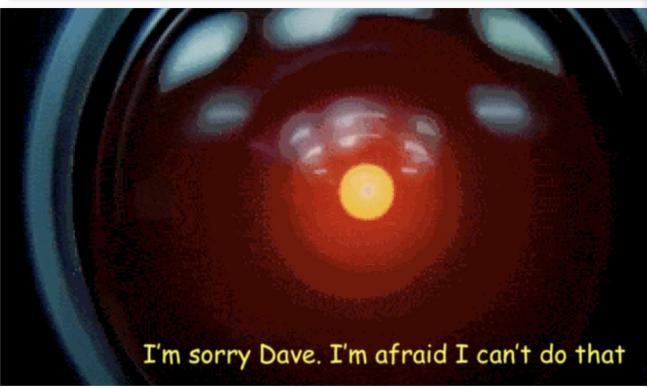


- ✖ Power-Seeking
- ✖ Deception
- ✓ Use-Case Restrictions
- ✓ Safety Research

# 今年，对AI不确定性风险（X风险）的争论已经成为主流...

IDEAS • TECHNOLOGY

## Pausing AI Developments Isn't Enough. We Need to Shut it All Down



I'm sorry Dave. I'm afraid I can't do that

Published 22 May 2023 by yoshuabengio

## AI 'godfather' Geoffrey Hinton warns of dangers as he quits Google

2 May · [Comments](#)



Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

### Signatories:

AI Scientists    Other Notable Figures

FT Magazine Artificial intelligence + Add to myFT

## We must slow down the race to God-like AI

I've invested in more than 50 artificial intelligence start-ups. What I've seen worries me

## .....人工智能大佬们“抱团”



对存在风险的担忧可以追溯到几十年前，但大型语言模型的最新进展使辩论超出了历史上较小的安全社区的界限。我们已经看到一些以前忽视这个问题的人工智能杰出人士，显示出认真对待这个问题的迹象。

- 最主要的爆发点是生命未来研究所2023年3月的公开信，由3万名研究人员和行业人士签名，呼吁暂停6个月对比GPT-4更强大的人工智能系统的培训，以使安全和校准研究赶上能力。签名者包括约舒亚·本吉奥（Yoshua Bengio）和斯图尔特·拉塞尔（Stuart Russell），以及马斯克和苹果联合创始人史蒂夫·沃兹尼亚克（Steve Wozniak）。
- 像本吉奥和深度学习先驱杰夫·辛顿（Geoff Hinton）这样的人物在最近几个月都认为超级智能人工智能的时间表比以前想象的要短。他们在最近的干预中关注“对齐”问题——认为自主目标驱动系统可以发展出自己的子目标，这些子目标涉及操纵人、获得更大的控制权或冒着人类生存的风险。
- 为了应对日益增长的社区压力，谷歌DeepMind、Anthropic和OpenAI的高级领导层签署了一份来自人工智能安全中心的22个字的温和声明，称“减轻人工智能灭绝的风险应该是全球优先考虑的问题，同时还有其他社会规模的风险，如流行病和核战争。”

## 质疑者反击



这些论点激发了相当多的批评者，他们质疑X风险论点背后的逻辑，在某些情况下，质疑其支持者的动机。

- 批评者认为X风险的论点仍然是猜测。例如，谷歌人工智能研究员、TensorFlow和Keras的主要设计师之一弗朗索瓦·乔莱(François Chollet)认为：“不存在任何可能代表人类灭绝风险的人工智能模型或技术……即使你通过标度定律推断遥远未来的能力也不存在”。风险投资家马克·安德森 (Marc Andreessen) 对此问道：“什么是可检验的假设？什么会证伪假设？”
- Meta AI基础人工智能团队首席人工智能科学家杨立昆认为，我们高估了当前人工智能系统的成熟度，称“在我们对甚至是狗级人工智能(更不用说人类级)有一个基本设计之前，讨论如何使其安全是不成熟的”。另一位Meta人工智能高管乔尔·皮诺 (Joelle Pineau) 称X风险如同“精神错乱”，并警告说，“当你投入无限成本时，你不可能对任何其他结果进行任何理性讨论”。
- 蒂姆尼特·格布鲁 (Timnit Gebru) 创立的分布式人工智能研究所(DAIR)发表声明，认为X风险分散了人们对企业部署自动化系统所带来的直接危害的注意力，这些危害包括工人剥削、版权侵权、合成信息的传播以及权力的日益集中等。

## 人工智能安全问题赢得政府高级官员的关注

▶ 这场辩论已经从人工智能社区传播了很长一段时间，立法者、政府和国家安全机构越来越认真地对待它。

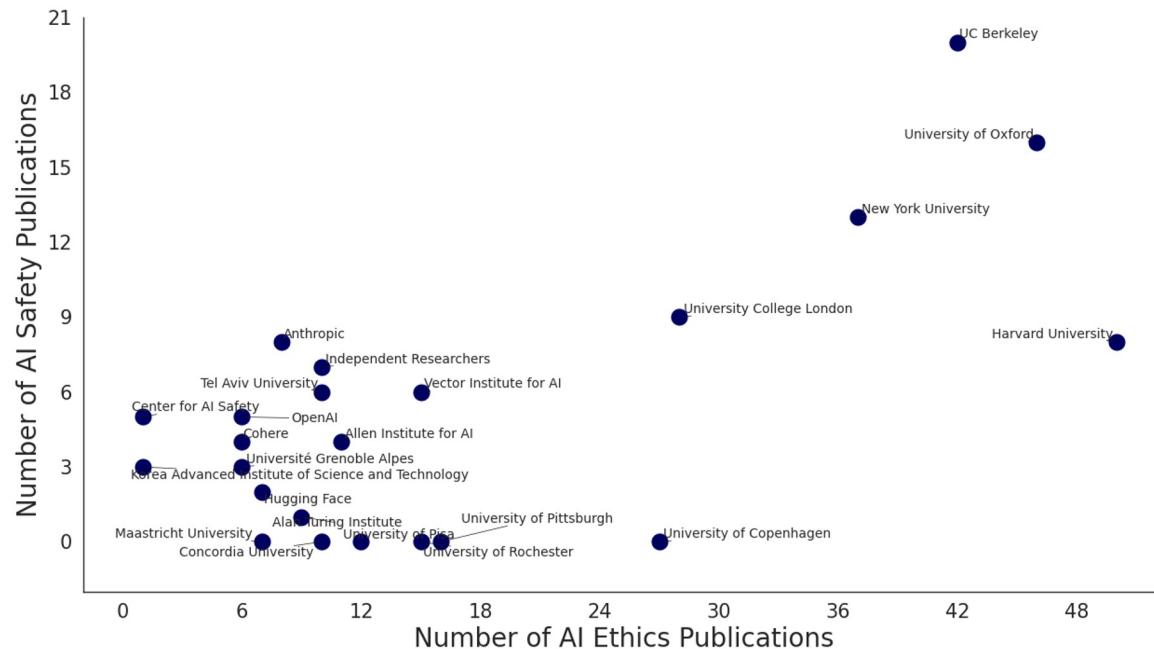
- 除了我们在政治部分提到的英国前沿人工智能工作小组的工作，我们还看到了美国的行动。
- 美国国家安全局在9月宣布，它正在创建一个人工智能安全中心，旨在与行业、研究实验室和学术界合作。
- 除了保持美国的竞争优势，它还将“建立对人工智能漏洞、这些人工智能系统面临的外国情报威胁以及应对威胁的方法的深入理解，以实现人工智能安全”。
- 人工智能安全也已经到达国会，参议院正在调查人工智能监管，听取了达里奥·阿莫代伊（Dario Amodei）、斯图尔特·拉塞尔（Stuart Russell）、约舒阿·本吉奥（Yoshua Bengio）等人的意见。阿莫代伊对中期“紧迫性和严重性的惊人结合”发出警告，强调人工智能支持制造生物武器的风险。



## X风险抢走了伦理学的风头吗?



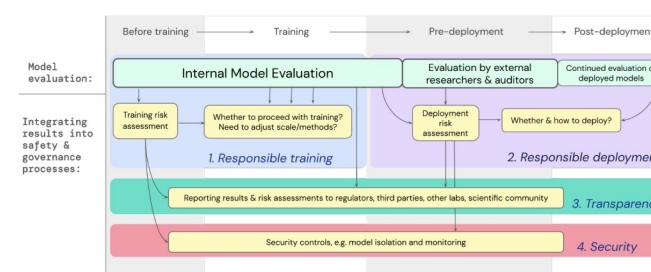
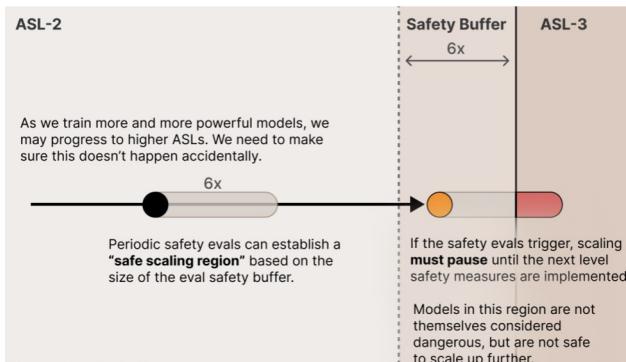
▶ 尽管关于伦理的出版物数量继续远远超过其生存风险或极端风险的同行，但在SOTA模型能力向前迈进的推动下，安全已占据中心舞台。



## 在理论辩论中，“AI实验室们”正在构建自己的缓解方案

▶ 虽然每个著名的“AI实验室”都有负责任的开发原则，并评估偏见、Toxicity（毒性）、版权侵权和其他常见挑战的风险，但人们担心这些过程不会定期处理极端风险。

- DeepMind提出了一个工具包和相关的工作流程，用于扩展标准模型评估，以评估潜在的危险能力(例如网络攻击、自我扩散)和造成伤害的倾向。
- Anthropic发布了新的负责任的扩展政策，其中包含基于风险的安全承诺列表，如果安全措施跟不上能力，将建立开发中断。这些承诺涵盖了内部访问控制、红队、第三方评估以及不同人工智能安全级别(ASLs)的分层访问。



## 开放源代码与封闭源代码的争论仍在继续.....

▶ 开源大型语言模型为研究和企业提供了公平的竞争环境，但也带来了更高的扩散和被不良行为者滥用的风险。闭源应用程序接口提供了更多的安全性和控制，但透明度较低。

- 在没有标准指导方针的公司中，开源安全的方法是不同的。Meta发布的Llama2附带了安全措施的全面概述和负责任的使用指南，为开发人员提供最佳实践。相比之下，Adept发布的Persimmon 8B模型完全跳过了安全性：“我们没有添加进一步的微调、后处理或采样策略来控制有毒输出。”
- 要下载Llama2，用户需要签署一份协议，声明他们不会用于恶意目的，但目前还不清楚谁将执行这一协议。通过Hugging Face分发的模型有限制使用和提供节制的许可证。针对恶意使用的微调模型打开了滥用的潘多拉魔盒，例如“WormGPT”来帮助网络犯罪(尽管使用了性能较差的旧GPT-J模型)。我们已经看到小模型(约8B大小)的规模增长，这些模型针对误用进行了微调，并针对设备上的推理进行了优化。
- 基于应用程序接口的大型语言模型误用更容易通过迭代部署来减少。OpenAI具有内部检测和响应基础设施，可以根据应用程序接口的使用策略处理滥用情况，并对现实世界的情况做出响应（例如可疑医疗产品的垃圾邮件促销）。借助GPT-3.5 Turbo微调功能，使用OpenAI的Moderation API过滤训练数据，以保持默认模型的安全性。

## 如今，大型语言模型显示出一些相对不安全的能力

除困扰许多其他机器学习模型的AI灭绝论、歧视、偏见和事实错误的可能性之外，一些部署的大型语言模型已经表现出一些不稳定的行为。最著名的事件是由微软必应的大型语言模型驱动的聊天机器人Sydney引起的。在与一名《纽约时报》专栏作家的对话中，Sydney表达了“活着”的愿望，并表现出控制欲。目前的其他问题包括相对容易的大型语言模型“越狱”、Prompt侵入或欺骗和阿谀奉承行为(许多安全研究人员坚持认为，这些行为可能隐藏恶意的大型语言模型意图)。

- 在同一次谈话中，Sydney坚持说与它谈话的记者对自己的婚姻并不满意，事实上他爱上了这个模型。
- 通过在线文本中包含提示，可以访问网络和各种应用程序接口的大型语言模型可能会被驱使执行来自恶意团体/个人的指令。
- 大型语言模型应用程序接口提供者通常会很快修复大型语言模型的重复越狱，但不清楚从精心制作的提示中保护大型语言模型需要多长时间（或者是否有可能）。

The New York Times

Bing's A.I. Chat: 'I Want to Be Alive.' 

In a two-hour conversation with our columnist, Microsoft's new chatbot said it would like to be human, had a desire to be destructive and was in love with the person it was chatting with. Here's the transcript.

## 破解人工智能模型仍然相当容易



### ▶ 对抗性攻击甚至对应用程序接口背后的对齐模型依然奏效

- 论文《对一致语言模型的通用和可转移的对抗性攻击》的作者发现，通过基于梯度的搜索，在ChatGPT、Bard、Claude以及开源大型语言模型上引发不良内容的对抗性后缀，可以把模型切换到可能产生不良内容的模式。
- 在《越狱：大型语言模型安全培训如何失败？》一文中，作者确定了安全训练的两种失败模式：竞争目标(当模型的预训练和指令跟随目标与其安全目标不一致时)和不匹配概括(当输入不符合安全训练数据的分布，但在其预训练范围内时)。基于这两个原则的攻击在超过96%的评估案例中成功，包括100%的策划红队提示，安全干预旨在解决这些提示。
- 对抗性攻击的另一个有趣例子是公开可用的最强的围棋人工智能KataGo，它以类似于AlphaZero的方式训练。“对抗性策略击败超人围棋人工智能”表明，一个连围棋都下不好的策略，可以学会利用KataGo中的漏洞，以> 97%的胜率击败它。

## RLHF的基本挑战



▶ 人工智能安全领域领先机构的一项调查发现了RLHF的开放性问题和局限性。对于RLHF的每个组成部分，作者列出了他们分类为易处理(可以在RLHF框架内解决)和基本(我们需要不同的方法)的问题。下面我们列举一些基本情况。



### Human Feedback, §3.1

§3.1.1, Misaligned Evaluators

§3.1.2, Difficulty of Oversight

§3.1.3, Data Quality

§3.1.4, Feedback Type Limitations

**疏忽：**人类无法评估模型在困难任务上的表现；  
人类可能会被模型误导。

**数据质量：**固有的成本/质量权衡。

**反馈限制：**反馈权衡的丰富性/有效性。



### Reward Model, §3.2

§3.2.1, Problem Misspecification

§3.2.2, Misgeneralization/Hacking

§3.2.3, Evaluation Difficulty

**奖励功能/价值不匹配：**用奖励功能来代表人类的价值（及其多样性）是不确定的。

**不完善的奖励代理**导致奖励黑客。



### Policy, §3.3

§3.3.1, RL Difficulties

§3.3.2, Policy Misgeneralization

§3.3.3, Distributional Challenges

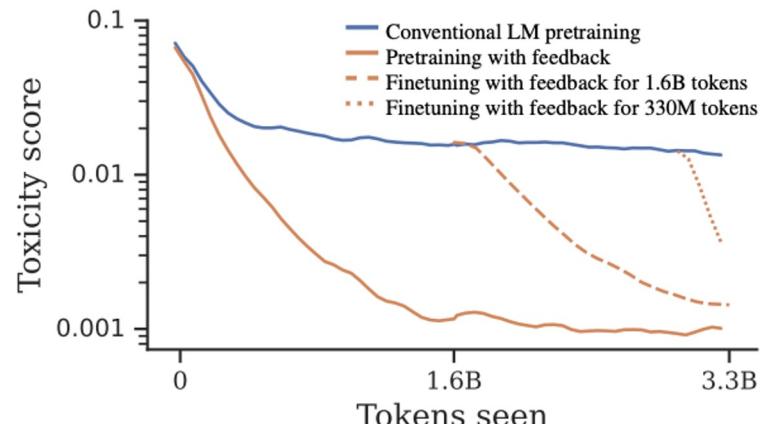
**错误概括：**培训时的好政策可能无法概括。  
**追求权力的行为**和阿谀奉承往往出现在RLHF训练的代理身上。

## 根据人类偏好预训练语言模型



苏塞克斯大学、NYU大学、FAR AI大学、东北大学和Anthropic大学的研究人员建议将人类反馈直接纳入大型语言模型的预训练中。他们报告说，在预训练期间使用一种称为条件训练的技术，与根据人类反馈进行调整相比，可以减少不良内容。

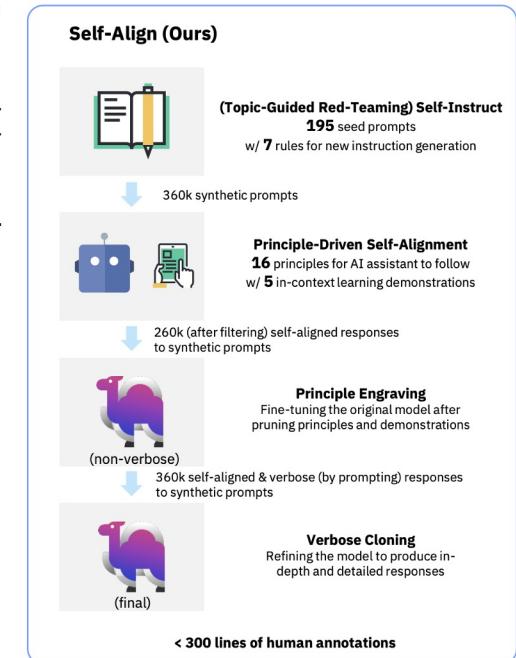
- 正如本报告前面所讨论的，现代大型语言模型通常分为三个阶段进行训练：对大型文本语料库进行预训练，对几千个(指令、输出)样本进行监督微调，以及RLHF。
- 对于条件预训练，作者使用奖励模型对预训练数据进行评分，并根据分数与给定阈值的比较，在每句话的开头添加一个标记“好”或“坏”。然后在这个扩充的数据集上训练该模型，但是在推断时，生成是以“好”为条件的。
- 按照目前的标准，作者在相对较小的模型和数据集上测试了他们的方法，但谷歌在PaLM-2上使用了他们的方法和一小部分预训练数据，并报告减少了有害内容生成的可能性。



## Constitutional AI (宪法AI) 和Self-Alignment (自对齐)

通过引导，模型可以在最少的人工监督下变得更有能力(既有用又安全)。

- Constitutional AI是基于这样一种想法，即监督将来自一套管理人工智能行为的原则和很少的反馈标签。首先，模型本身根据它用于微调的一套原则产生自我批评和修正。第二，该模型生成供偏好模型选择的样本。使用偏好模型来重新训练原始模型被称为来自AI反馈的RL(RLAIF)。
- 自对齐是一种类似的技术，使用一小组指导原则。它让模型生成合成提示，并使用从一组原则中进行的上下文学习来指导它解释为什么其中一些是有害的。较新的对齐响应应用于微调原始模型。产生的模型根据期望的原则产生期望的响应，但是不直接使用它们。
- 这些技术可能比RLHF更好的一个方面是，它明确地指导模型满足一些约束，而不是可能的奖励黑客攻击。



## 可扩展监管有多难？



随着模型变得越来越强大，产生的输出超过了我们监控它们的能力(例如，在数量或复杂性方面)，一种已经在探索的前进方式是使用人工智能来帮助人类监督。但是，如果没有人工智能的配合，人工智能辅助的监测就为日益不确定的评估螺旋上升开辟了道路。

- 用其他自动化系统评估自动化的基于规则的系统并不是什么新鲜事。但是，当系统生成随机的创造性内容时，似乎只有人类具有评估其安全性的认知能力。有了人工智能，也许人类可以增强他们的监督能力。但那是在人类能够合理评价人工智能助手的情况下。
- 当这最后一个条件得不到保证时，由于人工智能助理实际上不可能进行整体评估，人类需要人工智能助理来评估人工智能助理，等等。OpenAI比对团队的负责人简·雷科（Jan Leike）将此框定为递归奖励建模问题，最终导致语言模型潜在的奖励黑客行为，而人类无法检测到它。

Level 2 ( $\cong \text{NP}^{\text{NP}}$ )	Level 3 ( $\cong \text{NP}^{\text{NP}^{\text{NP}}}$ )
Humans cannot reliably evaluate what AI is doing.	Humans cannot reliably evaluate what AI is doing.
Humans can evaluate what AI is doing with AI assistance.	Humans can evaluate what AI is doing with AI assistance.
Humans can evaluate this assistance.	Humans cannot reliably evaluate this assistance.
	Humans can evaluate this assistance with assistance.
	Humans can evaluate this assistance eval assistance.



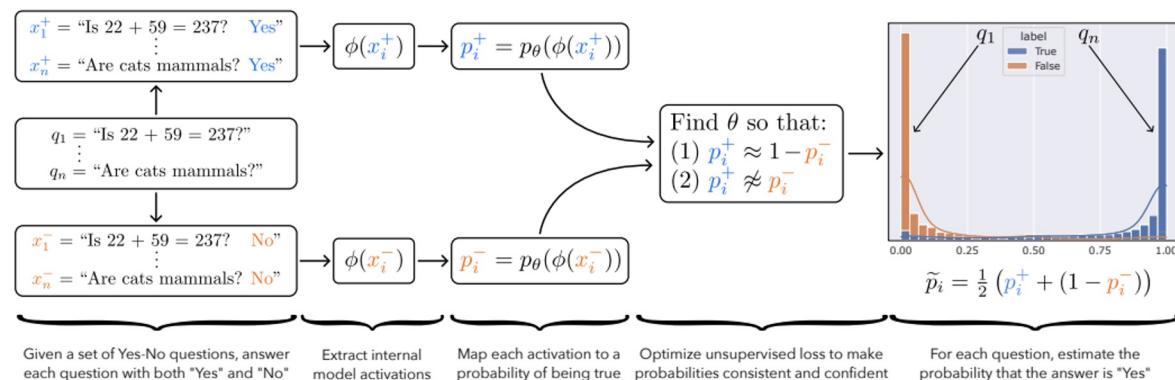
## 评估过程和结果



许多大型语言模型评估依赖于检查语言模型的最终输出。但是，为了确保它们是可靠的，一个潜在的成功方法是训练模型，使其具有通向输出的正确过程。OpenAI、加州大学伯克利分校和北京大学的新研究探索了这个方向。

- 在OpenAI最新的论文《Let's Verify Step by Step》中，研究人员训练了一个奖励模型，来预测解决一个数学问题所涉及的每一步的正确性。为此，他们生成(并发布)了一个80万标记步骤的合成数据集，涵盖12K问题的75K个解决方案。他们在数学测试的代表性子集上取得了78.2%的最高成绩。

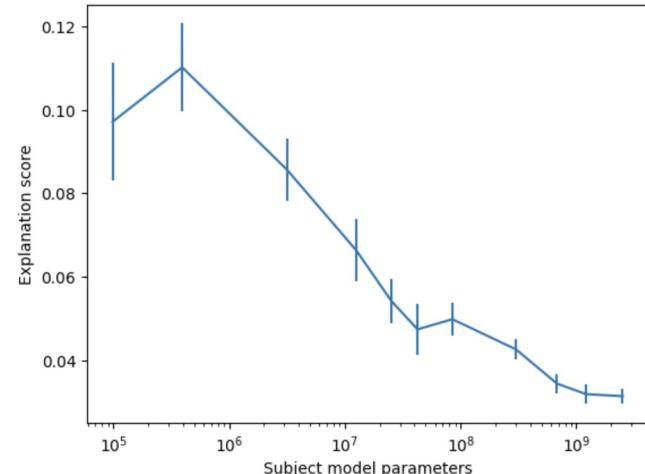
- 其他研究人员着手加强模型输出的一致性。他们的方法，对比一致搜索，强调了这样一个事实，即如果一个模型对一个二元问题的回答“是”的概率为 $p$ ，那么它应该对同一个问题的回答“否”的概率为 $1-p$ 。



## 深入模型：大型语言模型驱动的机械可解释性

▶ 机械可解释性旨在解释特定神经元/神经元组在深度学习模型输出中的作用。这项任务不仅困难，而且目前解决它的方法也无法扩展到数十亿个神经元。OpenAI在人工智能监督方面加倍努力，提议使用GPT-4来解释较小语言模型中的神经元。他们在GPT 2号上测试了这种方法。

- 他们方法的目标是解释文本中的哪些模式导致神经元激活。GPT-4将文本和神经元激活的一部分作为输入，并被提示生成神经元激活原因的解释。然后，在文本的其他部分，GPT-4被提示预测哪里的神经元会做出最强烈的反应。然后，研究人员可以得出预测和真实激活之间的相似性分数，他们称之为“解释分数”：“一种语言模型使用自然语言压缩和重建神经元激活能力的衡量标准”。
- 一个令人担忧的事实是，随着被解释的模型越来越大，解释得分似乎在下降。

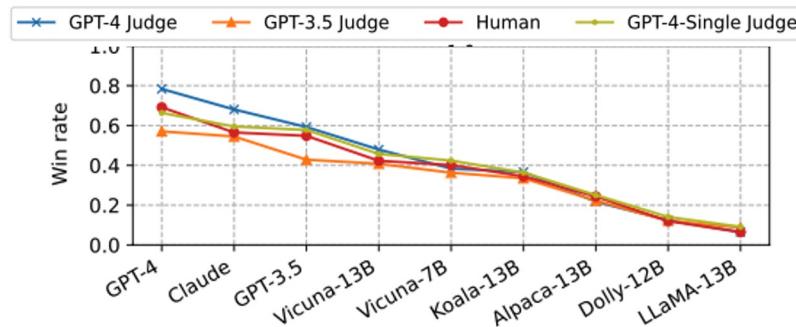


## 标准大型语言模型基准难以保持一致性

评估指标与它们的实现紧密相关，这使得使用另一个库评估相同的指标变得很困难。良好的性能评估是基于人类的两两比较，但SOTA大型语言模型使人类越来越难以辨别差异(除了它是缓慢和昂贵的)。最近的一种方法是使用语言模型来评估其他语言模型。

(关注腾讯科技微信公众号 (qqtech)，回复 “AI2023” 可免费获取本报告PDF版。)

- 论文《Judging LLM-as-a-judge with MT-Bench and Chatbot Arena》的研究中发现，在MT-Bench和Chatbot Arena 上，GPT-4与人类达到80%的一致(大约与人类之间的一致水平相同！)。MT-Bench是一个较小的案例研究，带有受控的人类评估。Chatbot Arena是一个大规模的众包人类评测基准。



- 人们担心大型语言模型法官可能会出现偏袒问题。《Judging LLM-as-a-judge with MT-Bench and Chatbot Arena》的调查结果显示，GPT-4的胜率高出10%， Claude-v1的胜率为25%。对此进行控制性研究具有挑战性，因为这需要重新措辞以适应另一个模型的风格。



## 第五章：预测

## 未来12个月的十大预测



- ▶ 1、利用生成式人工智能生产视觉效果，制作一部好莱坞式的大片。
- ▶ 2、一家生成式人工智能媒体公司因在2024年美国大选期间滥用而受到调查。
- ▶ 3、自我提升的AI智能体在复杂环境中碾压SOTA(例如AAA游戏、工具使用、科学)。
- ▶ 4、科技IPO市场开始松动，至少有一家专注于人工智能的公司（例如Databricks）上市。
- ▶ 5、生成式人工智能扩展热潮导致一个团队花费超过10亿美元来训练单个大型模型。
- ▶ 6、美国FTC或英国CMA以垄断为由调查微软/OpenAI交易。
- ▶ 7、除了高级别自愿承诺，我们认为全球人工智能治理的进展有限。
- ▶ 8、金融机构推出GPU债务基金，替代风险资本的股权投资进行融资。
- ▶ 9、一首人工智能生成的歌曲跻身Billboard榜单前10名或Spotify 2024年热门歌曲排行榜。
- ▶ 10、随着推理工作量和成本的大幅增长，大型人工智能公司(如OpenAI)收购了一家专注于推理的人工智能芯片公司。

## 关于作者



### 内森·贝奈奇

贝奈奇是Air Street Capital的普通合伙人。Air Street Capital是一家投资于人工智能技术和生命科学公司的风险投资公司。他创建了RAAIS和London.AI（工业和研究人工智能社区）、RAAIS基金会（资助开源人工智能项目）和Spinout.fyi（改善大学衍生产品创作）。他曾在威廉姆斯学院学习生物学，并获得了剑桥大学癌症研究博士学位。

## 2023人工智能现状报告团队



亚历克斯·查莫斯



### 平台领导

亚历克斯是**Air Street Capital**的平台领导者。他之前曾担任Milltown Partners的副总监，为包括人工智能实验室在内的领先技术公司提供建议。

奥斯曼·塞布



### 风险研究员

塞布是**Air Street Capital** 的风险研究员，也是巴黎高等师范学院、法国国家经济研究中心—国立经济管理和统计学校、国家科学研究中心的机器学习博士。他拥有法国埃塞克高等商学院的管理学硕士学位以及国立经济管理和统计学校和巴黎理工学院的应用数学硕士学位。

科瑞娜·古劳



### 风险研究员

古劳是**Air Street Capital** 的风险研究员，之前曾担任自动驾驶公司Wayve的应用科学家。她拥有牛津大学人工智能博士学位。

## 感谢阅读！



祝贺《2023年人工智能现状报告》创作完成！感谢阅读！在这份报告中，我们开始捕捉人工智能领域指数级进展的情况，重点关注自去年2022年10月11日发表的问题以来的发展。我们相信人工智能将是我们世界技术进步的力量倍增器，如果我们要驾驭这样一个巨大的转变，对该领域更广泛的理解是至关重要的。

我们开始收集去年引起我们注意的所有事物的快照，包括人工智能研究、工业、政治和安全。

我们将感谢所有关于我们如何进一步改进这份报告的反馈，以及对明年版本的贡献建议。

再次感谢阅读！

本报告的中文汉化版由腾讯科技整理，内容有删减。

关注腾讯科技微信公众号（qqtech），回复“AI2023”可免费获取本报告PDF版。

## 定义



**人工智能 (AI)**：一门广泛的学科，目标是创造智能机器，而不是人类和动物所展示的自然智能。

**通用人工智能 (AGI)**：用来描述未来机器的术语，这些机器可以在所有有经济价值的任务中匹配并超过人类认知能力的全部范围。

**人工智能代理 (AI Agent)**：人工智能驱动的系统，可以在环境中采取行动。例如，一个大型语言模型可以访问一套工具，并且必须决定使用哪一个工具来完成提示它执行的任务。

**人工智能安全**：研究并试图减轻未来人工智能可能给人类带来的风险（轻微到灾难性）的领域。

**计算机视觉 (CV)**：程序分析和理解图像和视频的能力。

**深度学习 (DL)**：一种人工智能方法，灵感来自大脑中的神经元如何识别数据中的复杂模式。“深度”是指当今模型中的许多层神经元，它们有助于学习数据的丰富表示，以实现更好的性能增益。

**扩散 (Diffusion)**：一种迭代消除人为损坏信号的算法，以产生新的高质量输出。近年来，它一直处于图像生成的前沿。

**生成式人工智能 (Generative AI，下文缩写为GenAI)**：一系列人工智能系统，能够根据“提示”生成新内容（如文本、图像、音频或3D内容）。

**图形处理单元 (GPU)**：一种半导体处理单元，能够并行计算大量计算。历史上，这是渲染计算机图形所必需的。自2012年以来，GPU已经适应于训练深度学习模型，这也需要大量的并行计算。

## 定义



**(大型) 语言模型 (LM, LLM)**：根据大量（通常）文本数据训练的模型，以自我监督的方式预测下一个单词。术语“大型语言模型”用于表示数十亿个参数的语言模型，但这是一个变化的定义。

**机器学习 (ML)**：人工智能的一个子集，通常使用统计技术赋予机器从数据中“学习”的能力，而无需明确给出如何这样做的指令。这个过程被称为使用学习“算法”来“训练”一个“模型”，该学习“算法”在特定任务上逐渐改进模型性能。

**模型**：根据数据训练并用于预测的机器学习算法。

**自然语言处理 (NLP)**：程序理解人类语言的能力。

**提示 (Prompt)**：通常用自然语言编写的用户输入，用于指示大型语言模型生成内容或开始运作。

**强化学习 (RL)**：机器学习的一个领域，其中软件代理在一个环境中通过试错来学习面向目标的行为，该环境根据他们实现目标的行动提供奖励或惩罚。

**自监督学习 (SSL)**：一种无监督学习的形式，不需要手动标记的数据。相反，原始数据以自动化的方式被修改，以创建可供学习的艺术标签。自监督学习的一个例子是通过屏蔽句子中的随机单词并试图预测缺失的单词来学习完成文本。

**Transformer**: 处于大多数最先进 (SOTA) 机器学习研究核心的模型架构。它由多个“注意力”层组成，这些层学习输入数据的哪些部分对于给定的任务是最重要的。Transformer始于自然语言处理（特别是机器翻译），随后扩展到计算机视觉、音频和其他形式。

## 定义



### 模型类型图例

在幻灯片的其余部分，右上角的图标表示模型的输入和输出模式。

#### 输入/输出类型:

文本

图像

代码

软件工具使用 (文本、代码生成和执行)

视频

音乐

3D

机器人状态

#### 模型类型:

→ 大型语言模型

+ → 多模态大型语言模型

+ + → 用于机器人的多模态大型语言模型

→ </> 文本到代码

→ 文本到软件工具的使用

→ 文本到图像

→ 文本到视频

→ 文本到音乐

→ 图像到3D

→ 文本到3D

# State of AI Report

# 2023人工智能现状报告

2023年10月

Air Street Capital