从ChatGPT到DeepSeek：人工智能大模型技术现状与展望

刘群

华为AI语音语义首席科学家

脱敏版

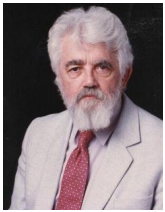
2025.04.15

HUAWEI

Content

Content

什么是人工智能 — 来自学术界的解释



John McCarthy

Stanford University

Q. What is artificial intelligence? 什么是人工智能？

A. It is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable.
它是制造智能机器，特别是智能计算机程序的科学和工程。它与使用计算机理解人类智能的类似任务有关，但AI不必局限于生物学上可观察的方法。

Q. Yes, but what is intelligence? 是的，但什么是智能？

A. Intelligence is the computational part of the ability to achieve goals in the world. Varying kinds and degrees of intelligence occur in people, many animals and some machines.
智能是实现目标能力的计算部分。不同种类和程度的智能发生在人、许多动物和一些机器中。

Q. Isn't there a solid definition of intelligence that doesn't depend on relating it to human intelligence? 智能难道没有一个可靠的定义，不依赖于将其与人类智能联系起来吗？

A. Not yet. The problem is that we cannot yet characterize in general what kinds of computational procedures we want to call intelligent. We understand some of the mechanisms of intelligence and not others.
还没有。问题是，我们还不能概括地描述我们想要称之为智能的计算过程的类型。我们了解智能的某些机制，而不是其他机制。

More in: http://www-formal.stanford.edu/jmc/whatisai/node1.html

什么是人工智能 — 来自企业界的解释

究竟人工智能是什么？回答这一问题并不像看起来那么简单。事实上，就连统一的“人工智能”定义也尚未出现。这是因为，从本质来看，我们所谈论的人工智能并不真的特指某项技术。

从实际层面出发，人工智能涵盖了一系列不同的技术，通过有效的组合，机器便能够以类似人类的智能水平展开行动。

弱人工智能（WEAK AI）

是指具有“模拟”思维的系统，也就是说，虽然看上去能够明智行动，但其实对于正在从事的工作，却并不拥有任何意识。例如，聊天机器人似乎可以保持自然的对话，但它其实不知道自己是谁，或为什么与对方交谈。

强人工智能（STRONG AI）

是指具备“真实”思维的系统——即运用有意识、主观性的头脑，像人类一样思考，进而展开睿智的行动。譬如当两个人谈话时，他们很可能确切地知晓对方是谁、自己在做什么、以及为何如此。

窄人工智能（NARROW AI）

是指仅针对单个或特定数量任务的人工智能。例如，1997年击败国际象棋世界冠军加里·卡斯帕罗夫（Gary Kasparov）的计算机“深蓝”，其功能仅限于下棋。它无法在简单的井字格游戏里获胜——甚至不知晓基本规则。

通用人工智能（GENERAL AI）

这类人工智能可用于在各种环境中执行广泛的任务。因此，它更接近于人类智慧。谷歌DeepMind使用强化学习技术开发了一款人工智能，使之学会参与诸多需要不同技能的竞赛。该人工智能系统在29款经典的雅达利（Atari）电子游戏中，仅使用屏幕上的像素作为数据输入，便取得了与人类相当的成绩[13]。

超级智能

“超级智能（superintelligence）”如果存在的话，通常是指超越人类智慧的通用人工智能和强人工智能。

我们并未像许多人那样，不断尝试去明确地描述人工智能，而是倾向于将此类技术视为一套能力框架。毫无疑问，这是了解人工智能、知晓其背后广泛技术的最佳方式。我们的框架以**人工智能支持机器实现的主要功能**为核心，其中包括：

感知。人工智能使机器可以通过获取并处理图像、声音、语言、文字和其他数据，察觉周围的世界。

理解。人工智能使机器可以通过识别模式来理解所收集到的信息。这类似于人类的信息诠释过程：解读信息的呈现方式及其背景——尽管这种方式未必能推导出真正的“含义”。

行动。人工智能使机器可以基于上述理解，在实体或数字世界中采取行动。

学习。人工智能使机器可以从成功或失败的行动中汲取经验教训，不断优化自身性能。

引自埃森哲高管指南《人工智能应用之道》

通用人工智能(AGI) / 强人工智能(StrongAI) / 深度人工智能(DeepAI)

通用人工智能也被称为强人工智能或深度人工智能。这是一个具有通用智能的机器的概念，它模仿人类智能，具有思考、理解、学习和应用其智能来解决任何问题的能力，就像人类在任何给定情况下所做的那样。

强人工智能使用思维理论AI框架不是复制或模拟，而是训练机器理解人类，以区分需求、情绪、信念和思维过程。

但是，这一切并不容易！AI研究人员和科学家需要找到一种方法，让机器有意识，编程出一整套认知能力。

翻译自 https://www.mygreatlearning.com/blog/artificial-general-intelligence/

就像帮助我们更深入地观察太空的哈勃望远镜一样，这些工具已经在扩展人类的知识，并产生积极的全球影响。

我们的长期目标是解决智能问题，开发更通用和更有能力的问题解决系统，称为通用人工智能（AGI）。

在安全和伦理的指导下，这项发明可以帮助社会找到世界上一些最紧迫和最基本的科学挑战的答案。

翻译自 DeepMind: https://deepmind.com/about

达特茅斯会议

1955年8月，四位学者起草了一份被称为“达特茅斯建议”的文件。它考察了这一时期研究领域的一些主要主题，包括神经网络、可计算性理论、创造力和自然语言处理和识别。该文件提议在接下来的夏天讨论这些话题。

文档以以下语句开头：

> “我们建议1956年夏天在新罕布什尔州汉诺威的达特茅斯学院进行为期2个月、10个人的人工智能研究。这项研究是基于这样一个猜想：学习的每个方面或智能的任何其他特征原则上都可以如此精确地描述，从而可以让一台机器来模拟它。我们试图找到如何让机器使用语言，形成抽象和概念，解决现在留给人类的各种问题，并改进自己。我们认为，如果一群经过精心挑选的科学家一起工作一个夏天，就可以在这些问题中的一个或多个问题上取得重大进展。”

次年夏天，这群学者在“达特茅斯学院人工智能夏季研究项目”（Dartmouth College for the Dartmouth Summer Research Project on Artificial Intelligence）中会面。
因此，1956年标志着一个新的研究领域的正式开始，数学家**约翰·麦卡锡（John McCarthy）**，达特茅斯学院的教授，也是这个事件的主要推动者，提议将其称为**人工智能**。

达特茅斯会议

August 1956. From left to right: Oliver Selfridge, Nathaniel Rochester, Ray Solomonoff, Marvin Minsky, Trenchard More, John McCarthy, Claude Shannon.

图灵测试

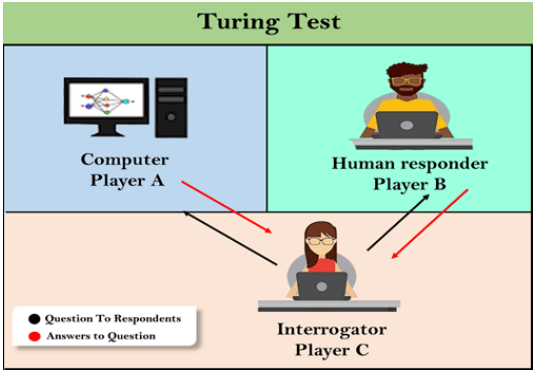


image from: https://1investing.in/worlds-first-ai-robot-citizen-sophia/

- 1950年艾伦·图灵的论文《计算机械与智能》描述了现在所谓的“图灵测试”。
- 图灵预测，在大约50年后，“一个普通的审问者在5分钟的审问后做出正确辨认的机会不会超过70%”。

人工智能发展历程

1956以前 孕育期：电子计算机、机器翻译、图灵测试、计算机下棋、早期神经网络

1956 达特茅斯会议，正式宣告人工智能诞生

1956-1973 乐观思潮：搜索式推理，聊天机器人

1973-1980 所有的AI程序都只是玩具：计算能力限制、计算复杂性、常识与推理

1980-1987 实用系统出现：专家系统、知识工程、五代机、神经网络重生

1987-1993 再次陷入低潮：未达预期，投入减少

1993-2006 机器学习兴起：摩尔定律导致计算成本降低，互联网带来更多数据，应用范围扩大

2006-2017 深度学习在图像识别、语音识别、机器翻译取得突破，AlphaGo战胜李世石

2018 预训练语言模型（以BERT为代表），单一模型通过微调解决各种语言理解问题

2020 大规模生成语言模型（以GPT-3为代表），单一模型无需微调，通过提示词解决各类语言理解和生成问题

2022 大语言模型（以ChatGPT/GPT-3.5代表），理解各种复杂指令和隐含的意图，跟人进行流畅对话，善解人意，遵从人类价值观

2024 推理大模型（以OpenAI o1/DeepSeek R1代表），能够通过慢思考解决复杂的数学推理和代码生成问题

人工智能发展历程

1956以前 孕育期：电子计算机、机器翻译、图灵测试、计算机下棋、早期神经网络

1956 达特茅斯会议，正式宣告人工智能诞生

1956-1973 乐观思潮：搜索式推理，聊天机器人

1973-1980 所有的AI程序都只是玩具：计算能力限制、计算复杂性、常识与推理

1980-1987 实用系统出现：专家系统、知识工程、五代机、神经网络重生

1987-1993 再次陷入低潮：未达预期，投入减少

1993-2006 机器学习兴起：摩尔定律导致计算成本降低，互联网带来更多数据，应用范围扩大

2006-2017 深度学习在图像识别、语音识别、机器翻译取得突破，AlphaGo战胜李世石

2018 预训练语言模型（以BERT为代表），单一模型通过微调解决各种语言理解问题

2020 大规模生成语言模型（以GPT-3为代表），单一模型无需微调，通过提示词解决各类语言理解和生成问题

2022 大语言模型（以ChatGPT/GPT-3.5代表），理解各种复杂指令和隐含的意图，跟人进行流畅对话，善解人意，遵从人类价值观

2024 推理大模型（以OpenAI o1/DeepSeek R1代表），能够通过慢思考解决复杂的数学推理和代码生成问题

人工智能发展历程

1956以前 孕育期：电子计算机、机器翻译、图灵测试、计算机下棋、早期神经网络

1956 达特茅斯会议，正式宣告人工智能诞生

1956-1973 乐观思潮：搜索式推理，聊天机器人

1973-1980 所有的AI程序都只是玩具：计算能力限制、计算复杂性、常识与推理

1980-1987 实用系统出现：专家系统、知识工程、五代机、神经网络重生

1987-1993 再次陷入低潮：未达预期，投入减少

1993-2006 机器学习兴起：摩尔定律导致计算成本降低，互联网带来更多数据，应用范围扩大

2006-2017 深度学习在图像识别、语音识别、机器翻译取得突破，AlphaGo战胜李世石

2018 预训练语言模型（以BERT为代表），单一模型通过微调解决各种语言理解问题

2020 大规模生成语言模型（以GPT-3为代表），单一模型无需微调，通过提示词解决各类语言理解和生成问题

2022 大语言模型（以ChatGPT/GPT-3.5代表），理解各种复杂指令和隐含的意图，跟人进行流畅对话，善解人意，遵从人类价值观

2024 推理大模型（以OpenAI o1/DeepSeek R1代表），能够通过慢思考解决复杂的数学推理和代码生成问题

人工智能发展历程

1956以前 孕育期：电子计算机、机器翻译、图灵测试、计算机下棋、早期神经网络

1956 达特茅斯会议，正式宣告人工智能诞生

1956-1973 乐观思潮：搜索式推理，聊天机器人

1973-1980 所有的AI程序都只是玩具：计算能力限制、计算复杂性、常识与推理

1980-1987 实用系统出现：专家系统、知识工程、五代机、神经网络重生

1987-1993 再次陷入低潮：未达预期，投入减少

1993-2006 机器学习兴起：摩尔定律导致计算成本降低，互联网带来更多数据，应用范围扩大

2006-2017 深度学习在图像识别、语音识别、机器翻译取得突破，AlphaGo战胜李世石

2018 预训练语言模型（以BERT为代表），单一模型通过微调解决各种语言理解问题

2020 大规模生成语言模型（以GPT-3为代表），单一模型无需微调，通过提示词解决各类语言理解和生成问题

2022 大语言模型（以ChatGPT/GPT-3.5代表），理解各种复杂指令和隐含的意图，跟人进行流畅对话，善解人意，遵从人类价值观

2024 推理大模型（以OpenAI o1/DeepSeek R1代表），能够通过慢思考解决复杂的数学推理和代码生成问题

人工智能发展历程

1956以前 孕育期：电子计算机、机器翻译、图灵测试、计算机下棋、早期神经网络

1956 达特茅斯会议，正式宣告人工智能诞生

1956-1973 乐观思潮：搜索式推理，聊天机器人

1973-1980 所有的AI程序都只是玩具：计算能力限制、计算复杂性、常识与推理

1980-1987 实用系统出现：专家系统、知识工程、五代机、神经网络重生

1987-1993 再次陷入低潮：未达预期，投入减少

1993-2006 机器学习兴起：摩尔定律导致计算成本降低，互联网带来更多数据，应用范围扩大

2006-2017 深度学习在图像识别、语音识别、机器翻译取得突破，AlphaGo战胜李世石

2018 预训练语言模型（以BERT为代表），单一模型通过微调解决各种语言理解问题

2020 大规模生成语言模型（以GPT-3为代表），单一模型无需微调，通过提示词解决各类语言理解和生成问题

2022 大语言模型（以ChatGPT/GPT-3.5代表），理解各种复杂指令和隐含的意图，跟人进行流畅对话，善解人意，遵从人类价值观

2024 推理大模型（以OpenAI o1/DeepSeek R1代表），能够通过慢思考解决复杂的数学推理和代码生成问题

人工智能发展历程

1956以前 孕育期：电子计算机、机器翻译、图灵测试、计算机下棋、早期神经网络

1956 达特茅斯会议，正式宣告人工智能诞生

1956-1973 乐观思潮：搜索式推理，聊天机器人

1973-1980 所有的AI程序都只是玩具：计算能力限制、计算复杂性、常识与推理

1980-1987 实用系统出现：专家系统、知识工程、五代机、神经网络重生

1987-1993 再次陷入低潮：未达预期，投入减少

1993-2006 机器学习兴起：摩尔定律导致计算成本降低，互联网带来更多数据，应用范围扩大

2006-2017 深度学习在图像识别、语音识别、机器翻译取得突破，AlphaGo战胜李世石

2018 预训练语言模型（以BERT为代表），单一模型通过微调解决各种语言理解问题

2020 大规模生成语言模型（以GPT-3为代表），单一模型无需微调，通过提示词解决各类语言理解和生成问题

2022 大语言模型（以ChatGPT/GPT-3.5代表），理解各种复杂指令和隐含的意图，跟人进行流畅对话，善解人意，遵从人类价值观

2024 推理大模型（以OpenAI o1/DeepSeek R1代表），能够通过慢思考解决复杂的数学推理和代码生成问题

人工智能发展历程

1956以前 孕育期：电子计算机、机器翻译、图灵测试、计算机下棋、早期神经网络

1956 达特茅斯会议，正式宣告人工智能诞生

1956-1973 乐观思潮：搜索式推理，聊天机器人

1973-1980 所有的AI程序都只是玩具：计算能力限制、计算复杂性、常识与推理

1980-1987 实用系统出现：专家系统、知识工程、五代机、神经网络重生

1987-1993 再次陷入低潮：未达预期，投入减少

1993-2006 机器学习兴起：摩尔定律导致计算成本降低，互联网带来更多数据，应用范围扩大

2006-2017 深度学习在图像识别、语音识别、机器翻译取得突破，AlphaGo战胜李世石

2018 预训练语言模型（以BERT为代表），单一模型通过微调解决各种语言理解问题

2020 大规模生成语言模型（以GPT-3为代表），单一模型无需微调，通过提示词解决各类语言理解和生成问题

2022 大语言模型（以ChatGPT/GPT-3.5代表），理解各种复杂指令和隐含的意图，跟人进行流畅对话，善解人意，遵从人类价值观

2024 推理大模型（以OpenAI o1/DeepSeek R1代表），能够通过慢思考解决复杂的数学推理和代码生成问题

人工智能发展历程

1956以前 孕育期：电子计算机、机器翻译、图灵测试、计算机下棋、早期神经网络

1956 达特茅斯会议，正式宣告人工智能诞生

1956-1973 乐观思潮：搜索式推理，聊天机器人

1973-1980 所有的AI程序都只是玩具：计算能力限制、计算复杂性、常识与推理

1980-1987 实用系统出现：专家系统、知识工程、五代机、神经网络重生

1987-1993 再次陷入低潮：未达预期，投入减少

1993-2006 机器学习兴起：摩尔定律导致计算成本降低，互联网带来更多数据，应用范围扩大

2006-2017 深度学习在图像识别、语音识别、机器翻译取得突破，AlphaGo战胜李世石

2018 预训练语言模型（以BERT为代表），单一模型通过微调解决各种语言理解问题

2020 大规模生成语言模型（以GPT-3为代表），单一模型无需微调，通过提示词解决各类语言理解和生成问题

2022 大语言模型（以ChatGPT/GPT-3.5代表），理解各种复杂指令和隐含的意图，跟人进行流畅对话，善解人意，遵从人类价值观

2024 推理大模型（以OpenAI o1/DeepSeek R1代表），能够通过慢思考解决复杂的数学推理和代码生成问题

人工智能发展历程

1956以前 孕育期：电子计算机、机器翻译、图灵测试、计算机下棋、早期神经网络

1956 达特茅斯会议，正式宣告人工智能诞生

1956-1973 乐观思潮：搜索式推理，聊天机器人

1973-1980 所有的AI程序都只是玩具：计算能力限制、计算复杂性、常识与推理

1980-1987 实用系统出现：专家系统、知识工程、五代机、神经网络重生

1987-1993 再次陷入低潮：未达预期，投入减少

1993-2006 机器学习兴起：摩尔定律导致计算成本降低，互联网带来更多数据，应用范围扩大

2006-2017 深度学习在图像识别、语音识别、机器翻译取得突破，AlphaGo战胜李世石

2018 预训练语言模型（以BERT为代表），单一模型通过微调解决各种语言理解问题

2020 大规模生成语言模型（以GPT-3为代表），单一模型无需微调，通过提示词解决各类语言理解和生成问题

2022 大语言模型（以ChatGPT/GPT-3.5代表），理解各种复杂指令和隐含的意图，跟人进行流畅对话，善解人意，遵从人类价值观

2024 推理大模型（以OpenAI o1/DeepSeek R1代表），能够通过慢思考解决复杂的数学推理和代码生成问题

人工智能发展历程

1956以前	孕育期：电子计算机、机器翻译、图灵测试、计算机下棋、早期神经网络
1956	达特茅斯会议，正式宣告人工智能诞生
1956-1973	乐观思潮：搜索式推理，聊天机器人
1973-1980	所有的AI程序都只是玩具：计算能力限制、计算复杂性、常识与推理
1980-1987	实用系统出现：专家系统、知识工程、五代机、神经网络重生
1987-1993	再次陷入低潮：未达预期，投入减少
1993-2006	机器学习兴起：摩尔定律导致计算成本降低，互联网带来更多数据，应用范围扩大
2006-2017	深度学习在图像识别、语音识别、机器翻译取得突破，AlphaGo战胜李世石
2018	预训练语言模型（以BERT为代表），单一模型通过微调解决各种语言理解问题
2020	大规模生成语言模型（以GPT-3为代表），单一模型无需微调，通过提示词解决各类语言理解和生成问题
2022	大语言模型（以ChatGPT/GPT-3.5代表），理解各种复杂指令和隐含的意图，跟人进行流畅对话，善解人意，遵从人类价值观
2024	推理大模型（以OpenAI o1/DeepSeek R1代表），能够通过慢思考解决复杂的数学推理和代码生成问题

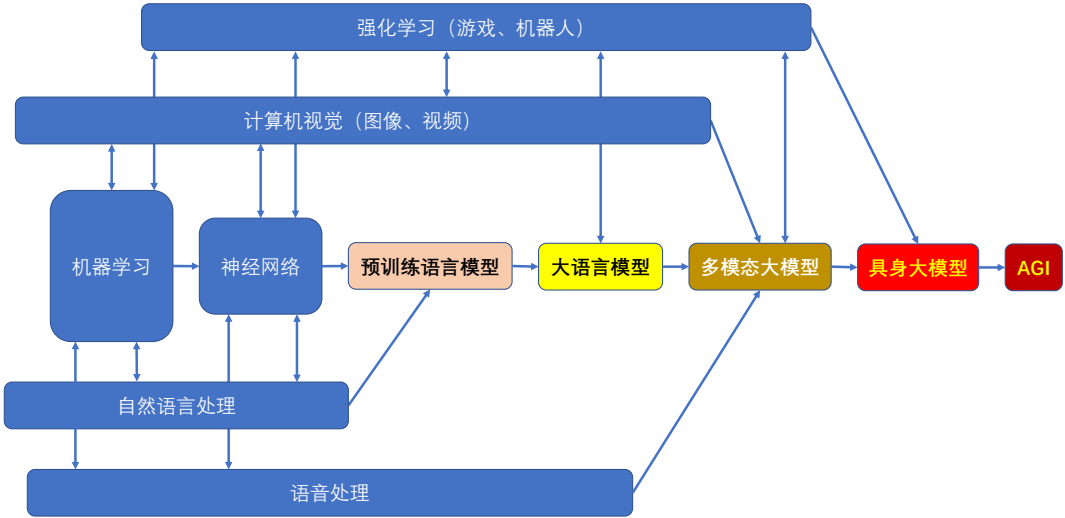
人工智能发展历程

1956以前	孕育期：电子计算机、机器翻译、图灵测试、计算机下棋、早期神经网络
1956	达特茅斯会议，正式宣告人工智能诞生
1956-1973	乐观思潮：搜索式推理，聊天机器人
1973-1980	所有的AI程序都只是玩具：计算能力限制、计算复杂性、常识与推理
1980-1987	实用系统出现：专家系统、知识工程、五代机、神经网络重生
1987-1993	再次陷入低潮：未达预期，投入减少
1993-2006	机器学习兴起：摩尔定律导致计算成本降低，互联网带来更多数据，应用范围扩大
2006-2017	深度学习在图像识别、语音识别、机器翻译取得突破，AlphaGo战胜李世石
2018	预训练语言模型（以BERT为代表），单一模型通过微调解决各种语言理解问题
2020	大规模生成语言模型（以GPT-3为代表），单一模型无需微调，通过提示词解决各类语言理解和生成问题
2022	大语言模型（以ChatGPT/GPT-3.5代表），理解各种复杂指令和隐含的意图，跟人进行流畅对话，善解人意，遵从人类价值观
2024	推理大模型（以OpenAI o1/DeepSeek R1代表），能够通过慢思考解决复杂的数学推理和代码生成问题

人工智能发展历程

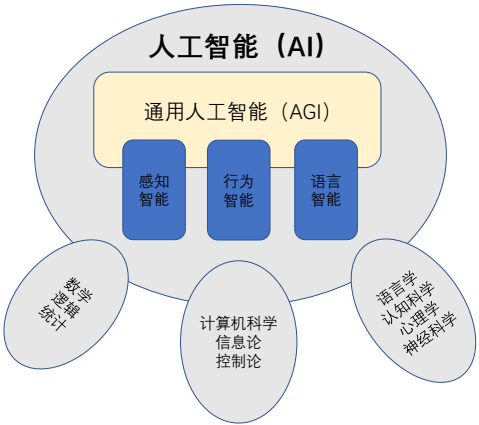
1956以前	孕育期：电子计算机、机器翻译、图灵测试、计算机下棋、早期神经网络
1956	达特茅斯会议，正式宣告人工智能诞生
1956-1973	乐观思潮：搜索式推理，聊天机器人
1973-1980	所有的AI程序都只是玩具：计算能力限制、计算复杂性、常识与推理
1980-1987	实用系统出现：专家系统、知识工程、五代机、神经网络重生
1987-1993	再次陷入低潮：未达预期，投入减少
1993-2006	机器学习兴起：摩尔定律导致计算成本降低，互联网带来更多数据，应用范围扩大
2006-2017	深度学习在图像识别、语音识别、机器翻译取得突破，AlphaGo战胜李世石
2018	预训练语言模型（以BERT为代表），单一模型通过微调解决各种语言理解问题
2020	大规模生成语言模型（以GPT-3为代表），单一模型无需微调，通过提示词解决各类语言理解和生成问题
2022	大语言模型（以ChatGPT/GPT-3.5代表），理解各种复杂指令和隐含的意图，跟人进行流畅对话，善解人意，遵从人类价值观
2024	推理大模型（以OpenAI o1/DeepSeek R1代表），能够通过慢思考解决复杂的数学推理和代码生成问题

人工智能发展路径



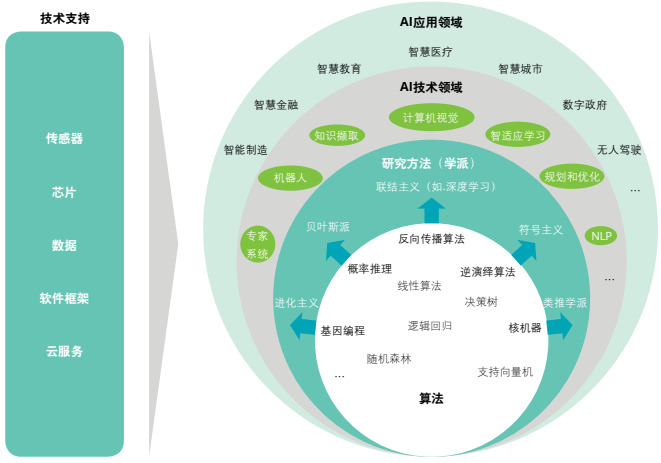
人工智能与其他学科的关系

- 人工智能主要包括感知智能、语言智能、行为智能三类；
- 通用人工智能是人工智能的一种特殊情况，也是人工智能的终极理想；
- 通用人工智能试图采用单一的方法解决所有智能问题；
- 人工智能是一门典型的交叉学科，与很多基础科学、工程科学甚至人文科学都有密切的关系。



人工智能技术生态

图表1-2：人工智能各层级图示



资料来源：德勤研究

Content

大模型名词解释

- AI大模型通常指参数规模达到10亿以上的模型。
- AI大模型根据参数规模大致分为几个档次：十亿级、百亿级、千亿级、万亿级。
- AI大模型通常是指大语言模型（Large Language Models，LLM），因为最早出现的AI大模型就是语言模型，而且后来出现的大模型通常也是以语言模型为核心构建的。
- 多模态大模型指同时支持图像、音频、视频等多模态理解和生成的大模型，但这些模型通常还是以语言模型为核心，因此又叫做多模态大语言模型（Multimodal Large Language Models），简称MLLM。
- 推理模型（Reasoning Models）又叫慢思考模型（Slow Thinking Models），它可以先进行思考再输出结果，而不是直接输出结果，适合解决一些较复杂的问题。
- AI大模型有时也被叫做基础模型（Foundation Models），这些基本上都是同义词。
- 大语言模型具备用一个模型解决各种不同问题的能力，被认为是向通用人工智能（Artificial General Intelligence，AGI）跨出了一大步。

大模型名词解释

- AI大模型通常指参数规模达到10亿以上的模型。
- AI大模型根据参数规模大致分为几个档次：十亿级、百亿级、千亿级、万亿级。
- AI大模型通常是指大语言模型（Large Language Models，LLM），因为最早出现的AI大模型就是语言模型，而且后来出现的大模型通常也是以语言模型为核心构建的。
- 多模态大模型指同时支持图像、音频、视频等多模态理解和生成的大模型，但这些模型通常还是以语言模型为核心，因此又叫做多模态大语言模型（Multimodal Large Language Models），简称MLLM。
- 推理模型（Reasoning Models）又叫慢思考模型（Slow Thinking Models），它可以先进行思考再输出结果，而不是直接输出结果，适合解决一些较复杂的问题。
- AI大模型有时也被叫做基础模型（Foundation Models），这些基本上都是同义词。
- 大语言模型具备用一个模型解决各种不同问题的能力，被认为是向通用人工智能（Artificial General Intelligence，AGI）跨出了一大步。

大模型名词解释

- AI大模型通常指参数规模达到10亿以上的模型。
- AI大模型根据参数规模大致分为几个档次：十亿级、百亿级、千亿级、万亿级。
- AI大模型通常是指大语言模型（Large Language Models，LLM），因为最早出现的AI大模型就是语言模型，而且后来出现的大模型通常也是以语言模型为核心构建的。
- 多模态大模型指同时支持图像、音频、视频等多模态理解和生成的大模型，但这些模型通常还是以语言模型为核心，因此又叫做多模态大语言模型（Multimodal Large Language Models），简称MLLM。
- 推理模型（Reasoning Models）又叫慢思考模型（Slow Thinking Models），它可以先进行思考再输出结果，而不是直接输出结果，适合解决一些较复杂的问题。
- AI大模型有时也被叫做基础模型（Foundation Models），这些基本上都是同义词。
- 大语言模型具备用一个模型解决各种不同问题的能力，被认为是向通用人工智能（Artificial General Intelligence，AGI）跨出了一大步。

大模型名词解释

- AI大模型通常指参数规模达到10亿以上的模型。
- AI大模型根据参数规模大致分为几个档次：十亿级、百亿级、千亿级、万亿级。
- AI大模型通常是指大语言模型（Large Language Models，LLM），因为最早出现的AI大模型就是语言模型，而且后来出现的大模型通常也是以语言模型为核心构建的。
- 多模态大模型指同时支持图像、音频、视频等多模态理解和生成的大模型，但这些模型通常还是以语言模型为核心，因此又叫做多模态大语言模型（Multimodal Large Language Models），简称MLLM。
- 推理模型（Reasoning Models）又叫慢思考模型（Slow Thinking Models），它可以先进行思考再输出结果，而不是直接输出结果，适合解决一些较复杂的问题。
- AI大模型有时也被叫做基础模型（Foundation Models），这些基本上都是同义词。
- 大语言模型具备用一个模型解决各种不同问题的能力，被认为是向通用人工智能（Artificial General Intelligence，AGI）跨出了一大步。

大模型名词解释

- AI大模型通常指参数规模达到10亿以上的模型。
- AI大模型根据参数规模大致分为几个档次：十亿级、百亿级、千亿级、万亿级。
- AI大模型通常是指大语言模型（Large Language Models，LLM），因为最早出现的AI大模型就是语言模型，而且后来出现的大模型通常也是以语言模型为核心构建的。
- 多模态大模型指同时支持图像、音频、视频等多模态理解和生成的大模型，但这些模型通常还是以语言模型为核心，因此又叫做多模态大语言模型（Multimodal Large Language Models），简称MLLM。
- 推理模型（Reasoning Models）又叫慢思考模型（Slow Thinking Models），它可以先进行思考再输出结果，而不是直接输出结果，适合解决一些较复杂的问题。
- AI大模型有时也被叫做基础模型（Foundation Models），这些基本上都是同义词。
- 大语言模型具备用一个模型解决各种不同问题的能力，被认为是向通用人工智能（Artificial General Intelligence，AGI）跨出了一大步。

大模型名词解释

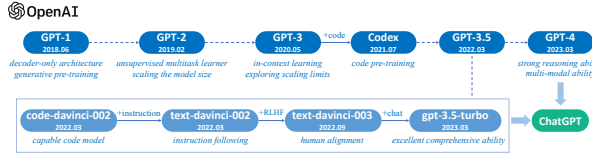
- AI大模型通常指参数规模达到10亿以上的模型。
- AI大模型根据参数规模大致分为几个档次：十亿级、百亿级、千亿级、万亿级。
- AI大模型通常是指大语言模型（Large Language Models，LLM），因为最早出现的AI大模型就是语言模型，而且后来出现的大模型通常也是以语言模型为核心构建的。
- 多模态大模型指同时支持图像、音频、视频等多模态理解和生成的大模型，但这些模型通常还是以语言模型为核心，因此又叫做多模态大语言模型（Multimodal Large Language Models），简称MLLM。
- 推理模型（Reasoning Models）又叫慢思考模型（Slow Thinking Models），它可以先进行思考再输出结果，而不是直接输出结果，适合解决一些较复杂的问题。
- AI大模型有时也被叫做基础模型（Foundation Models），这些基本上都是同义词。
- 大语言模型具备用一个模型解决各种不同问题的能力，被认为是向通用人工智能（Artificial General Intelligence，AGI）跨出了一大步。

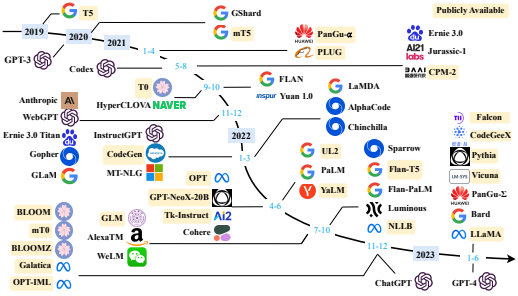
大模型名词解释

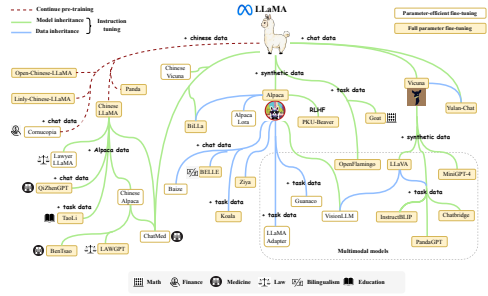
- AI大模型通常指参数规模达到10亿以上的模型。
- AI大模型根据参数规模大致分为几个档次：十亿级、百亿级、千亿级、万亿级。
- AI大模型通常是指大语言模型（Large Language Models，LLM），因为最早出现的AI大模型就是语言模型，而且后来出现的大模型通常也是以语言模型为核心构建的。
- 多模态大模型指同时支持图像、音频、视频等多模态理解和生成的大模型，但这些模型通常还是以语言模型为核心，因此又叫做多模态大语言模型（Multimodal Large Language Models），简称MLLM。
- 推理模型（Reasoning Models）又叫慢思考模型（Slow Thinking Models），它可以先进行思考再输出结果，而不是直接输出结果，适合解决一些较复杂的问题。
- AI大模型有时也被叫做基础模型（Foundation Models），这些基本上都是同义词。
- 大语言模型具备用一个模型解决各种不同问题的能力，被认为是向通用人工智能（Artificial General Intelligence，AGI）跨出了一大步。

从GPT-3到ChatGPT和GPT-4：业界大模型概览（至2023年3月）

- 2023年是大模型爆发的一年
- 2023年又被称为AGI元年
- 大模型已经深刻影响了AI
- 大模型还将深刻影响我们的社会







Zhao, et al. “A Survey of Large Language Models.” arXiv2303.18223

GPT-4以后发布的主要大模型（至2024年底）

模型	公司	发布时间	模型规格	特点
GPT-4	OpenAI	2023-03-14	1.8T MoE	语言能力大幅提高
Claude	Anthropic	2023-03-14		安全性强
文心3.0	百度	2023-03-18		中文强
Claude 2	Anthropic	2023-07-11		能够处理图像和PDF/Word等格式文件
混元大模型	腾讯	2023-09-07	>100B	图像和视频生成
文心大模型Ernie4.0	百度	2023-10-17		知识增强、检索增强和对话增强
Grok-1	xAI	2023-11-03	314B MoE A25%	开源
Gemini	Google	2023-12-06	Nano, Pro, Ultra	原生多模态
Sora宣布	OpenAI	2024-02-15		视频生成
Gemma	Google	2024-02-22	2B, 7B	开源
Claude 3	Anthropic	2024-03-01	Haiku, Sonnet, Opus	能力强大，支持操作计算机应用
Kimi	月之暗面	2024-03-17		支持2M长上下文
Llama 3	Meta	2024-04-19	8B, 70B	最强开源，支持多模态输入（图像、音频、视频）
千问Qwen2.5	阿里云	2024-05-09	5B - 72B	开源，多语言
GPT-4o	OpenAI	2024-05-13	200B MoE?	多模态理解，语音交互能力强大
豆包大模型	字节跳动	2024-05-15		性能好
GLM-4	智谱AI	2024-06-05		长序列处理、视频通话
Claude 3.5	Anthropic	2024-06-21	Sonnet	代码能力强大
Gemma 2	Google	2024-06-27	2B, 9B, 27B	开源
Gemini 1.5 Pro	Google	2024-06-27		长文档
Llama 3.1	Meta	2024-07-23	8B, 70B, 405B	128k长序列、多语言、工具使用
Grok-2	xAI	2024-08-13	大约600B	开源，长序列、代码生成
o1 mini	OpenAI	2024-09-12	100B MoE?	推理模型，Test-time scaling, 代码数学能力强，响应快
o1-preview	OpenAI	2024-09-12	300B MoE?	推理模型，Test-time scaling，代码数学推理能力强，
Llama 3.2	Meta	2024-09-26	1B, 3B, 11B, 90B	能力更强，多模态，边缘应用，眼镜，VR
Kimi 探索版	月之暗面	2024-10-11		推理能力强
千问QwQ	阿里巴巴	2024-11-01	32B	推理模型
o1	OpenAI	2024-12-05	300B MoE?	推理模型，Test-time scaling，代码数学推理能力强，
Sora	OpenAI	2024-12-09		视频生成
Gemini 2.0	Google	2024-12-11	Flash, Pro	支持2M token的上下文长度；视频处理能力突出
豆包通用模型Pro	字节跳动	2024-12-18		语言对话，视频通话
o3 (announce)	OpenAI	2024-12-20		推理能力再次提高，ARC-AGI突破
DeepSeek v3	深度求索	2024-12-27	671B MoE	高效预训练，节省10多倍训练算力

- 模型发布密集紧凑，竞争激烈
- 模型规模增大、序列长度加长、推理成本减低趋势明显
- 开源模型能力呈追赶之势
- 多模态大模型发展迅速
- 推理模型开辟了一个新赛道
- 大模型应用处于井喷之势，但还没有出现杀手级别应用

大模型发展的一些关键节点

2019.02 OpenAI GPT-2 XL：首个大语言模型（15亿参数）

2020.05 OpenAI GPT-3：首个千亿模型（1750亿参数），实现单一模型无需微调，通过提示词解决各类语言理解和生成问题

2022.11 OpenAI ChatGPT：理解各种复杂指令和隐含的意图，跟人进行流畅对话，善解人意，遵从人类价值观

2023.03 OpenAI GPT-4：数学、逻辑推理、代码生成能力大幅提高，支持文本和图像的多模态理解，支持128k长上下文

2024.02 OpenAI Sora发布：自然语言提示直接生成长达1分钟前后一致的高清视频，逼真模拟物体运动和自然现象，支持场景视角切换

2024.05 OpenAI GPT-4o：文本、图像、视频多模态理解，自然的语音交互

2024.09 OpenAI ChatGPT o1-preview/o1-mini：能够通过慢思考解决复杂的数学推理和代码生成问题

2025.01 DeepSeek V3/R1：训练推理成本大幅降低，完全开源的慢思考模型，解密了训练慢思考模型的强化学习算法

大模型发展的一些关键节点

2019.02 OpenAI GPT-2 XL：首个大语言模型（15亿参数）

2020.05 OpenAI GPT-3：首个千亿模型（1750亿参数），实现单一模型无需微调，通过提示词解决各类语言理解和生成问题

2022.11 OpenAI ChatGPT：理解各种复杂指令和隐含的意图，跟人进行流畅对话，善解人意，遵从人类价值观

2023.03 OpenAI GPT-4：数学、逻辑推理、代码生成能力大幅提高，支持文本和图像的多模态理解，支持128k长上下文

2024.02 OpenAI Sora发布：自然语言提示直接生成长达1分钟前后一致的高清视频，逼真模拟物体运动和自然现象，支持场景视角切换

2024.05 OpenAI GPT-4o：文本、图像、视频多模态理解，自然的语音交互

2024.09 OpenAI ChatGPT o1-preview/o1-mini：能够通过慢思考解决复杂的数学推理和代码生成问题

2025.01 DeepSeek V3/R1：训练推理成本大幅降低，完全开源的慢思考模型，解密了训练慢思考模型的强化学习算法

大模型发展的一些关键节点

2019.02 OpenAI GPT-2 XL：首个大语言模型（15亿参数）

2020.05 OpenAI GPT-3：首个千亿模型（1750亿参数），实现单一模型无需微调，通过提示词解决各类语言理解和生成问题

2022.11 OpenAI ChatGPT：理解各种复杂指令和隐含的意图，跟人进行流畅对话，善解人意，遵从人类价值观

2023.03 OpenAI GPT-4：数学、逻辑推理、代码生成能力大幅提高，支持文本和图像的多模态理解，支持128k长上下文

2024.02 OpenAI Sora发布：自然语言提示直接生成长达1分钟前后一致的高清视频，逼真模拟物体运动和自然现象，支持场景视角切换

2024.05 OpenAI GPT-4o：文本、图像、视频多模态理解，自然的语音交互

2024.09 OpenAI ChatGPT o1-preview/o1-mini：能够通过慢思考解决复杂的数学推理和代码生成问题

2025.01 DeepSeek V3/R1：训练推理成本大幅降低，完全开源的慢思考模型，解密了训练慢思考模型的强化学习算法

大模型发展的一些关键节点

2019.02 OpenAI GPT-2 XL：首个大语言模型（15亿参数）

2020.05 OpenAI GPT-3：首个千亿模型（1750亿参数），实现单一模型无需微调，通过提示词解决各类语言理解和生成问题

2022.11 OpenAI ChatGPT：理解各种复杂指令和隐含的意图，跟人进行流畅对话，善解人意，遵从人类价值观

2023.03 OpenAI GPT-4：数学、逻辑推理、代码生成能力大幅提高，支持文本和图像的多模态理解，支持128k长上下文

2024.02 OpenAI Sora发布：自然语言提示直接生成长达1分钟前后一致的高清视频，逼真模拟物体运动和自然现象，支持场景视角切换

2024.05 OpenAI GPT-4o：文本、图像、视频多模态理解，自然的语音交互

2024.09 OpenAI ChatGPT o1-preview/o1-mini：能够通过慢思考解决复杂的数学推理和代码生成问题

2025.01 DeepSeek V3/R1：训练推理成本大幅降低，完全开源的慢思考模型，解密了训练慢思考模型的强化学习算法

大模型发展的一些关键节点

2019.02 OpenAI GPT-2 XL：首个大语言模型（15亿参数）

2020.05 OpenAI GPT-3：首个千亿模型（1750亿参数），实现单一模型无需微调，通过提示词解决各类语言理解和生成问题

2022.11 OpenAI ChatGPT：理解各种复杂指令和隐含的意图，跟人进行流畅对话，善解人意，遵从人类价值观

2023.03 OpenAI GPT-4：数学、逻辑推理、代码生成能力大幅提高，支持文本和图像的多模态理解，支持128k长上下文

2024.02 OpenAI Sora发布：自然语言提示直接生成长达1分钟前后一致的高清视频，逼真模拟物体运动和自然现象，支持场景视角切换

2024.05 OpenAI GPT-4o：文本、图像、视频多模态理解，自然的语音交互

2024.09 OpenAI ChatGPT o1-preview/o1-mini：能够通过慢思考解决复杂的数学推理和代码生成问题

2025.01 DeepSeek V3/R1：训练推理成本大幅降低，完全开源的慢思考模型，解密了训练慢思考模型的强化学习算法

大模型发展的一些关键节点

2019.02 OpenAI GPT-2 XL：首个大语言模型（15亿参数）

2020.05 OpenAI GPT-3：首个千亿模型（1750亿参数），实现单一模型无需微调，通过提示词解决各类语言理解和生成问题

2022.11 OpenAI ChatGPT：理解各种复杂指令和隐含的意图，跟人进行流畅对话，善解人意，遵从人类价值观

2023.03 OpenAI GPT-4：数学、逻辑推理、代码生成能力大幅提高，支持文本和图像的多模态理解，支持128k长上下文

2024.02 OpenAI Sora发布：自然语言提示直接生成长达1分钟前后一致的高清视频，逼真模拟物体运动和自然现象，支持场景视角切换

2024.05 OpenAI GPT-4o：文本、图像、视频多模态理解，自然的语音交互

2024.09 OpenAI ChatGPT o1-preview/o1-mini：能够通过慢思考解决复杂的数学推理和代码生成问题

2025.01 DeepSeek V3/R1：训练推理成本大幅降低，完全开源的慢思考模型，解密了训练慢思考模型的强化学习算法

大模型发展的一些关键节点

2019.02 OpenAI GPT-2 XL：首个大语言模型（15亿参数）

2020.05 OpenAI GPT-3：首个千亿模型（1750亿参数），实现单一模型无需微调，通过提示词解决各类语言理解和生成问题

2022.11 OpenAI ChatGPT：理解各种复杂指令和隐含的意图，跟人进行流畅对话，善解人意，遵从人类价值观

2023.03 OpenAI GPT-4：数学、逻辑推理、代码生成能力大幅提高，支持文本和图像的多模态理解，支持128k长上下文

2024.02 OpenAI Sora发布：自然语言提示直接生成长达1分钟前后一致的高清视频，逼真模拟物体运动和自然现象，支持场景视角切换

2024.05 OpenAI GPT-4o：文本、图像、视频多模态理解，自然的语音交互

2024.09 OpenAI ChatGPT o1-preview/o1-mini：能够通过慢思考解决复杂的数学推理和代码生成问题

2025.01 DeepSeek V3/R1：训练推理成本大幅降低，完全开源的慢思考模型，解密了训练慢思考模型的强化学习算法

大模型发展的一些关键节点

2019.02 OpenAI GPT-2 XL：首个大语言模型（15亿参数）

2020.05 OpenAI GPT-3：首个千亿模型（1750亿参数），实现单一模型无需微调，通过提示词解决各类语言理解和生成问题

2022.11 OpenAI ChatGPT：理解各种复杂指令和隐含的意图，跟人进行流畅对话，善解人意，遵从人类价值观

2023.03 OpenAI GPT-4：数学、逻辑推理、代码生成能力大幅提高，支持文本和图像的多模态理解，支持128k长上下文

2024.02 OpenAI Sora发布：自然语言提示直接生成长达1分钟前后一致的高清视频，逼真模拟物体运动和自然现象，支持场景视角切换

2024.05 OpenAI GPT-4o：文本、图像、视频多模态理解，自然的语音交互

2024.09 OpenAI ChatGPT o1-preview/o1-mini：能够通过慢思考解决复杂的数学推理和代码生成问题

2025.01 DeepSeek V3/R1：训练推理成本大幅降低，完全开源的慢思考模型，解密了训练慢思考模型的强化学习算法

ChatGPT发布引发轰动效应

- 用户数：5天100万，2个月达到1亿
- 所有人都开始讨论ChatGPT，成为现象级产品
- Google内部拉响红色警报
- Google紧急仅仅发布Bard，但因发布现场出现错误导致股票蒸发8%
- 微软追加投资OpenAI一百亿美元
- 微软迅速推出加载了ChatGPT的New Bing，并计划将ChatGPT接入Office套件
- 国内外大厂迅速跟进

DeepSeek R1发布是ChatGPT之后AI领域影响最大的事件

- DeepSeek上线 20 天日活超 2000 万是 ChatGPT 的 40%。
- 2025年1月20日DeepSeek-R1 发布几天后，在1月的最后一周迎来了爆发，DeepSeek在1月份累计获得1.25亿用户（含网站(Web)、应用(App)累加不去重）。其中 80% 以上用户来自最后一周，即 DeepSeek 7天完成了1亿用户的增长，在没有任何广告投放的情况下。
- 梁文锋：外部看到的是幻方 2015 年后的部分，但其实我们做了 16 年。（来自于丽丽采访梁文锋，揭秘 DeepSeek：一个更极致的中国技术理想主义故事，原文：https://36kr.com/p/2872793466982535）

Source: https://news.qq.com/rain/a/20250208A05G7J00

大模型带来的AI领域格局的变化

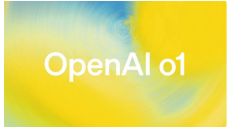
以算法为中心

以模型为中心

OpenAI o1推理模型取得突破

- OpenAI o1 是 2024 年 9 月 12 日 OpenAI 正式对外发布的一款新模型，是该公司下一代 “推理” 模型中的第一个。
- o1具有强大的推理能力：
  - 在竞争性编程问题 （Codeforces） 中排名第 89 个百分位，
  - 在美国数学奥林匹克竞赛 （AIME） 资格赛中跻身美国前 500 名学生之列，
  - 在物理、生物和化学问题的基准 （GPQA） 上超过了人类博士水平的准确性。
- 此次发布同时包括了两个版本：o1-preview 和 o1-mini。o1-mini是一个更小，更便宜的版本，在编码方面特别有效。
- 正式的o1版本预期将在1-2个月内发布。
- 相比GPT-4o等模型，o1速度相对较慢，成本相对更高。
- 应用场景：适用于解决复杂的数学、科学、编程等多领域的复杂问题，为需要深度推理和复杂任务处理的场景提供了新的解决方案。

- 模型命名从1开始一个新的序列，而且不再以GPT为前缀
- 传说中的草莓模型(strawberry)，或猎户座(orion)
- o代表OpenAI，不代表猎户座



Introducing OpenAI o1-preview

We've developed a new series of AI models designed to spend more time thinking before they respond. They can reason through complex tasks and solve harder problems than previous models in science, coding, and math.

OpenAI o1-mini

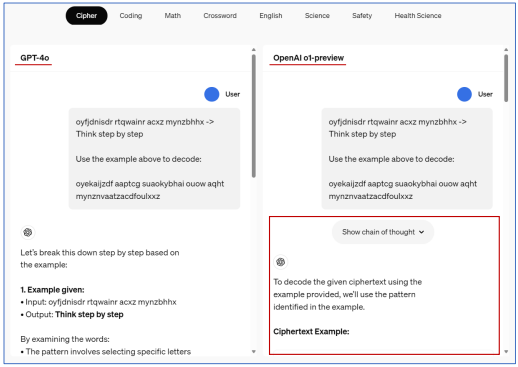
OpenAI o1-mini excels at STEM, especially math and coding—nearly matching the performance of OpenAI o1 on evaluation benchmarks such as AIME and Codeforces. We expect o1-mini will be a faster, cost-effective model for applications that require reasoning without broad world knowledge.

Learning to Reason with LLMs

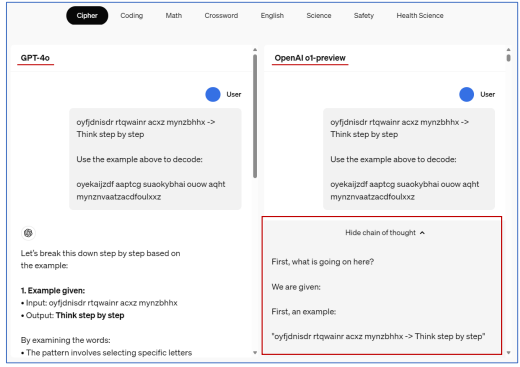
OpenAI o1 ranks in the 89th percentile on competitive programming questions (Codeforces), places among the top 500 students in the US in a qualifier for the USA Math Olympiad (AIME), and exceeds human PhD-level accuracy on a benchmark of physics, biology, and chemistry problems (GPQA). While the work needed to make this new model as easy to use as current models is still ongoing, we are releasing an early version of this model, OpenAI o1-preview, for immediate use in ChatGPT and to trusted API users.

o1推理示例

- oyfjdnisdr rtqwainr acxz mynzbhhx -> Think step by step
- 使用上面这个例子破解下面文字：
- oyekaijzdf aaptcg suaokybhai ouow aqht mynznvaatzacdfoulxxz
- （答案：THERE ARE THREE R'S IN STRAWBERRY）



隐藏CoT，只有约80步推理



显示CoT，约720步推理

OpenAI连续12天发布会

日期	主题	主要内容
Day 1	o1	1. 完整版 o1 模型。据称该模型运行速度更快，在解决复杂实际问题时的重大错误率降低了34%。 2. o1 模型新增了图像分析功能，且添加网页浏览和文件上传功能，API 版本将支持视觉任务、函数调用和结构化输出，便于系统集成。 3. 推出月费 200 美元的 ChatGPT Pro 订阅，提供"无限"使用 o1、GPT-4o 和 Advanced Voice 功能的权限。
Day 2	RFT	1. 强化微调 (RFT) 技术，能让开发者针对特定任务调整"o-系列"模型的方法，只需要几十个样本，就可以通过强化学习让模型在反复实践中大幅提升推理能力。 2. 2025年初向公众开放 RFT。
Day 3	SORA	OpenAI 将文本转视频模型 Sora 打造成独立产品，不过模型效果并不算好，但产品方面设计还行。
Day 4	Canvas	提供了一个专门的界面，用于处理超出常规聊天格式的长篇写作和编程项目，现已与 GPT-4o 模型完全整合。
Day 5	Apple	Apple Intelligence与ChatGPT集成的重新发布。
Day 6	Video Input	ChatGPT 的语音功能增添了两项新内容：面向 ChatGPT Plus 和 Pro 用户的支持屏幕共享的"视频通话"功能，以及应景的圣诞老人语音。
Day 7	Project	设置Projects来将相关对话和文件分类整理，并且支持将同一个项目中的其他内容作为context进行问答。仅限付费用户。被人戏称为新建文件夹功能。
Day 8	Search	ChatGPT 的搜索功能向所有免费用户开放，还包括新的地图界面，并支持在语音对话中进行搜索。
Day 9	DevDay	1. o1 模型，新增了函数调用、开发者消息和视觉处理功能。 2. 大幅降低了 GPT-4o 的音频处理价格，降幅达 60%，并推出了更经济的 GPT-4o mini 版本，价格仅为原音频服务的十分之一。 3. 简化了实时应用的 WebRTC 接入流程。 4. 推出了让开发者能够自定义模型的偏好微调功能（官方文档说就是DPO）。
Day 10	接入公众电话与IM	通过免费电话号码 (1-800-CHATGPT) 和 WhatsApp 提供 ChatGPT 服务，To C的模型厂都应该抄。
Day 11	PC端适配应用	1. 自动化处理桌面任务，如实时分析第三方软件中的内容、实时完善代码，让工作更高效。Mac版已上线，Wins版也即将上线。 2. 增加了对Apple Notes、Notion、Quip和Warp等更多笔记和编程应用的支持，让GPT成为学习的好帮手。 3. 引入“与应用对话”选项，用高级语音模式与应用程序互动，实时调试、思考文档内容，让编程和写作更流畅。
Day 12	o3预告	预告了两个新的模拟推理模型 o3 和 o3-mini，o3-mini 可能会在 1 月底推出。（传说中的Orion猎户座）

Credit: TIAN Yong @ Huawei

模型　开发工具　产品

OpenAI连续12天发布会

- 更大的模型：
  - GPT-4的参数为1.8万亿，虽然传言GPT-5训练遇到瓶颈，但各大AI厂商都在加紧投资构建更大规模集群（几十万卡集群）
  - AGI：一种观点认为，达到人类智能水平的AGI的模型参数规模约为100万亿，这是OpenAI、DeepMind追逐的愿景，现在看并不遥远
- 多模态模型（4o/Gemini/Sora）：
  - 多模态理解和生成：GPT-4o、Sora、可灵都展现了令人惊艳的多模态能力
- 推理模型（o1/o3）：
  - 长序列：长序列在支持复杂推理和完成复杂任务方面重要性日显，成为很多公司追求的热点，目前最好模型可支持百万token长度，精度不损失
- 大模型应用：
  - 智能体（Agent）和多智能体（Multi-agent）
  - 小的大模型（sLLM）：可以部署在手机等端侧设备
  - 高效训练和推理部署：增量训练、投机推理、量化、FastAttention……每百万token推理成本降低到个位数美元/人民币

DeepSeek-V3在模型架构和训练方法上的创新

- 2024年12月27日，DeepSeek V3 是深度求索（DeepSeek）推出的一款高性能大语言模型，其技术特点包括：
  - DeepSeek V3 采用 6850 亿参数（685B）的 MoE 架构，每个 Token 仅激活 37 亿参数，在计算效率和性能之间取得平衡。
  - 多头潜在注意力（MLA）：优化长序列处理，降低内存占用，支持128K上下文窗口，适用于长文档分析、代码库理解等任务。
  - 多Token预测（MTP）：增强文本连贯性，提升长文本生成能力。
  - DeepSeekMoE负载均衡：采用无辅助损失负载均衡策略, 优化专家路由, 提高计算效率。
- DeepSeek V3大大降低了训练成本：仅使用2048个H800 GPU，在57天内完成训练，全部训练成本总计仅为557.6万美元。这远低于通常训练大型语言模型所需的数亿美元。
- 2025年2月16日，DeepSeek发布了最新的研究成果——原生稀疏注意力（Native Sparse Attention, NSA），硬件友好而且高效，并可以实现训推一体化。
- 2025年2月24-28日，DeepSeek连续五天开放五个重要的底层技术项目代码库，以全透明的方式与全球开发者社区分享其在通用人工智能基础设施方面的最新成果。此举被视为 DeepSeek 推动 AI 开源生态和技术变革的里程碑式事件，引发了社区的强烈关注和反响。
- 2025年3月24日，最近发布的新版本DeepSeek-V3-0324，优化了数学、代码、报告撰写能力，超越 GPT-4.5，并且再次大幅降低了训练成本。

DeepSeek开源周

- 2月24日：发布 Flash MLA，这是首个开源的代码库，是针对英伟达 Hopper GPU 优化的高效 MLA 解码内核，针对可变长度序列作了优化，能依据序列长度动态调配计算资源。
- 2月25日：开源 DeepEP，这是首个用于 MoE（混合专家模型）训练和推理的开源 EP 通信库。它还原生支持 FP8 低精度运算调度，降低计算资源消耗，并且在节点内和节点间都支持 NVLink 和 RDMA，拥有用于训练和推理预填充的高吞吐量内核以及用于推理解码的低延迟内核。
- 2月26日：开源 DeepGEMM，它支持普通和混合专家（MoE）分组的 GEMM，是矩阵乘法加速库，为 V3/R1 的训练和推理提供支持。该库采用 CUDA 编写，在安装过程中无需编译，通过使用轻量级的即时编译（JIT）模块在运行时编译所有内核，在 Hopper GPU 上最高可达到 1350 + FP8 TFLOPS（每秒万亿次浮点运算）的计算性能。
- 2月27日：一次性开源了 DualPipe、EPLB（专家并行负载均衡器）以及训练和推理框架的性能分析数据。其中，DualPipe 是一种双向管道并行算法，EPLB 用于增强训练效率。此外，DeepSeek 还在 Github 上详细讲解了 DeepSeek-V3 和 R1 模型背后的并行计算优化技术。
- 2月28日：开源 3FS，即 Fire-Flyer 文件系统，它是所有 Deepseek 数据访问的助推器。此外，还开源了基于 3FS 的数据处理框架 Smallpond，可进一步优化 3FS 的数据管理能力。

DeepSeek R1破解了慢思考推理模型强化学习训练方法

- 2024年11月20日，DeepSeek-R1-Lite预览版正式上线网页端。2025年1月20日，DeepSeek正式发布 DeepSeek-R1模型，并同步开源模型权重。
- 使用强化学习训练，推理过程包含大量反思和验证，思维链长度可达数万字。在后训练阶段大规模使用强化学习技术，在仅有极少标注数据的情况下，极大提升了模型推理能力。
- 在数学、代码、自然语言推理等任务上，性能比肩OpenAI o1正式版。在GSM 8K等评估数学和逻辑技能的基准测试中，能生成详细思维链推理，解决复杂问题能力强。
- DeepSeek R1开源：采用宽松的MIT许可协议，支持免费商用、任意修改和衍生开发，为全球开发者提供了广泛的使用和创新空间，打破了OpenAI尝试通过闭源建立的技术壁垒。
- DeepSeek R1破解了OpenAI在推理模型技术上布下的迷局，公开了大量技术细节，澄清了RL训练中的一些关键问题，使得推理模型的RL训练不再神秘。
- DeepSeek R1的发布产生了轰动效应，DeepSeek应用登顶苹果中国地区和美国地区应用商店免费App下载排行榜，在美区下载榜上超越了ChatGPT，在全球140个市场的应用商店下载榜上也强势夺冠。引发行业关注：在海外开发者社区中引发了轰动，成为美国顶尖大学研究人员的首选模型，也让华尔街和投资者感到震撼。同时也引发了资本市场的剧烈波动。

OpenAI近期在Agent和多模态方面发布的新成果

- 2025年1月23日发布 Operator：是 OpenAI 首款 AI 智能体产品。
- 2025年2月2日发布 Deep Research：是为 ChatGPT 打造的“深度研究” 智能体，主要为金融、科学、政策和工程等领域从事高强度知识工作的人员设计。
- 2025年3月11日发布 Agent-first 开发套件：
  - Responses API：堪称“智能体操作系统”，整合了原对话 API 的流畅交互与助手 API 的工具调用能力。
  - 内置工具：包含 Web 搜索、文件搜索和计算机操作三大工具。
  - 开源 Agents SDK：用于简化多代理工作流的编排。开发者可通过可视化的工作流编排定义不同角色的 Agent，设置任务交接规则，并实时监控执行路径。
- 2025年3月22日发布 3 款全新语音模型：包括 gpt-4o-transcribe、gpt-4o-mini-transcribe 和 gpt-4o-mini-tts。
- 2025年3月22日宣布GPT-4o支持文生图，效果惊艳。
- 此外，3月27日，OpenAI 对其 Agent SDK 进行重大更新，正式支持 Model Context Protocol（MCP）服务。

Manus的发布推动了Agent的发展

- 2025年3月6日，中国AI创业公司Monica发布全球首款通用AI Agent产品Manus，具有“自主执行”与“通用性” 两大核心优势，定位于性能强大的通用型助手，不仅能生成想法，还能独立思考并采取行动，从规划到执行全流程自主完成复杂任务，如撰写报告、制作表格、操作电脑、编写代码等，可交付完整成果，展现出前所未有的通用性和执行能力。
- Manus采用多模型协同架构：能够独立完成从任务拆解到成果交付的全流程，覆盖研究、金融分析、教育等51个具体场景。例如，可直接解压简历文件、分析候选人信息并生成评估报告，展现出接近人类工作流的执行效率。
- Manus自发布后迅速引爆全网，官网注册流量激增导致服务器崩溃，二手平台邀请码价格一度飙升至10万元，Discord社区涌入超10万开发者，用户自发构建2000余条任务模板库。
- MetaGPT团队4人仅用3小时即复刻了开源 AI Agent 产品Open Manus，CAMEL-AI团队也实现“0 天复刻”，并将系统中涉及的每个部件单独开源，这一方面推动了技术普惠化进程，但也说明Manus的技术护城河有限。
- Manus在信息整合能力、任务执行稳定性、速度和用户体验等方面还存在诸多问题，其表现高度依赖具体场景，离成熟产品还有一段距离。

MCP协议成为Agent工具调用事实标准

- 诞生背景：随着AI技术的发展，企业和组织在将AI系统与外部数据源和工具集成时，面临着每个数据源和工具都有独特的API和接口规范的问题，开发者需编写定制化代码，这耗时费力且降低了系统的可扩展性和灵活性。在此背景下，MCP应运而生，旨在标准化AI模型访问和利用外部上下文的过程，简化AI与外部数据源的集成。
- Anthropic 公司于 2024年11月推出的开源框架Model Context Protocol（MCP）。包括规范和多种语言的软件开发工具包（SDK），还建立了收集基于MCP 的不同服务器实现的存储库。
- 早期应用与发展：发布后，MCP迅速获得关注并在多个领域得到应用。
- 持续改进与拓展：MCP在不断发展，通过标准化的打包、简化的安装、用于增强安全性的沙箱以及集中式服务器注册表，使MCP服务器更易于访问。同时，社区也在扩大对音频和视频等新模式的支持，并探索正式的标准化。
- 2025年3月，OpenAI对其 Agent SDK 进行重大更新，正式支持MCP，使得开发者可以通过统一接口标准，为智能体无限接入各种第三方工具，大幅提升复杂自动化应用的开发效率。
- 2025年4月4日，谷歌官宣支持MCP。4月9日，在Google Cloud Next 25大会上，在开源了首个标准智能体交互协议——Agent2Agent Protocol（简称A2A）。这些协议将对于整个AI生态的构建起到至关重要的作用。

大模型持续演进带来算力需求快速增长

大模型

千亿稠密/万亿稀疏（24年） → 十万亿稀疏（25年）

语言模型 → 多模态模型

直接生成 → 慢思考推理

应用探索 → 应用爆发

算力平台

计算

更多模型并行技术的应用

- **算力密度&精度**: 单柜算力密度持续上升，多种计算精度支持（FP16/BF16/FP32）
- **超大规模单体集群**:
  - ✓ 千亿稠密/万亿稀疏需要万卡集群
  - ✓ 十万亿稀疏， TP&EP规模扩大，需要更大规模超节点集群，提升MOE模型训练效率

网络

更大规模/更复杂的组网

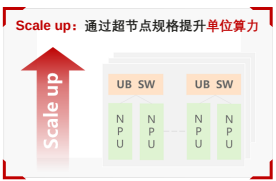
- **超大组网规模**：2层1:1无收敛的AI网络架构，最支持128K集群规模组网，千卡到万卡平滑演进
- **网络级负载均衡**：独家NSLB算法，算网协同调度，优化AI训练参数面HASH冲突

存储

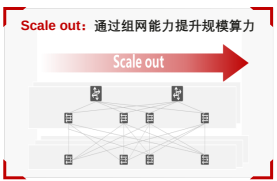
更大数据量存储和读写
(模型参数、CheckPoint …)

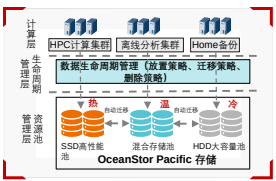
- **极致读写性能**：高IOPS和并行大带宽，大小I/O自适应，混合负载高效，提升训　练效率
- **智能分级存储**：全闪+混闪的智能分级存储，PB级存取+TB级带宽，综合效率更高

Scale up：通过超节点规格提升单位算力



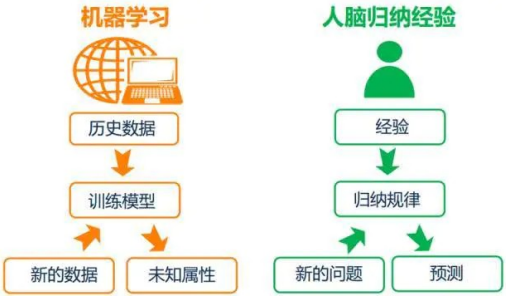
Scale out：通过组网能力提升规模算力





Content

机器学习



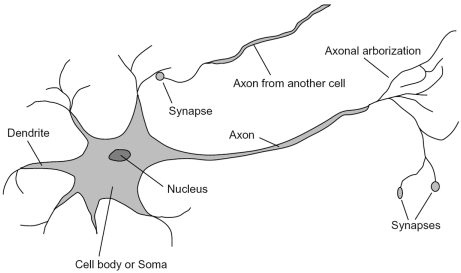
https://www.jianshu.com/p/effdd03547d4

Table: A simple retail dataset

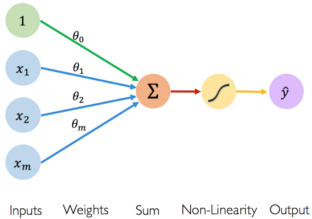
ID	BBY	ALC	ORG	GRP
1	no	no	no	couple
2	yes	no	yes	family
3	yes	yes	no	family
4	no	no	yes	couple

John Kelleher and Brian Mac Namee and Aoife D' Arcy
Fundamentals of Machine Learning for Predictive Data Analytics

神经网络



生物神经元网络



MIT: Alexander Amini, 2018 introtodeeplearning.com

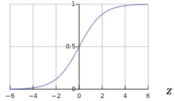
人工神经元网络

Activation Functions

$$\hat{y} = g(\theta_0 + \boldsymbol{X}^T\boldsymbol{\theta})$$

- Example: sigmoid function

$$g(z) = \sigma(z) = \frac{1}{1+e^{-z}}$$



深度学习

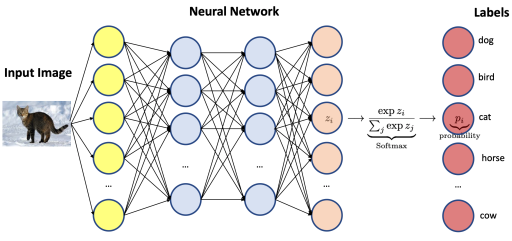
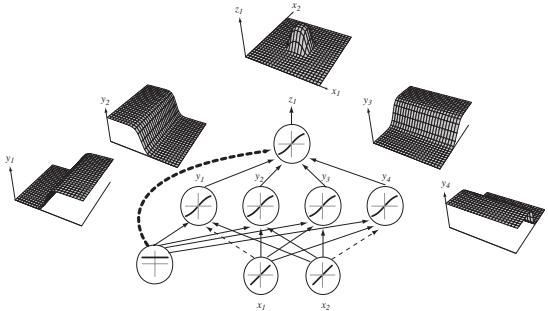
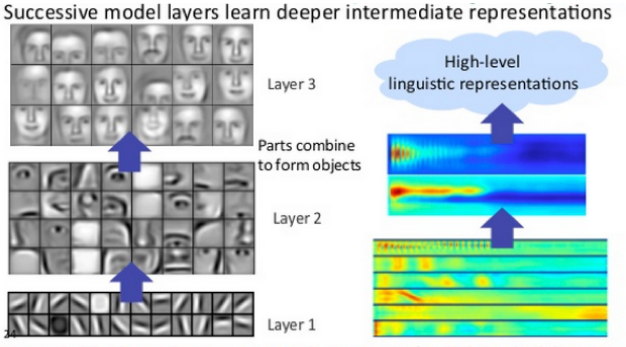


Figure source: https://gab41.lab41.org/entropic-ghosts-35670292bc87

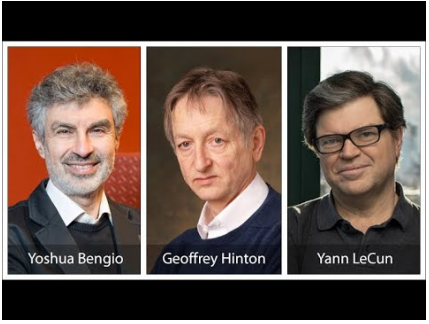


(from Pascal Vincent's slides)

深度学习

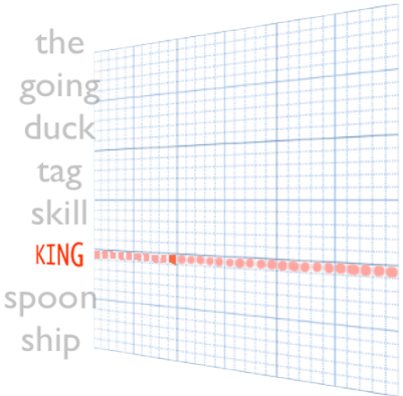


from https://wiki.pathmind.com/neural-network

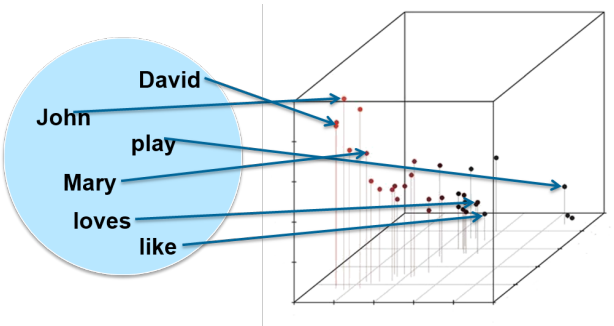


Turing Awards 2018

深度学习

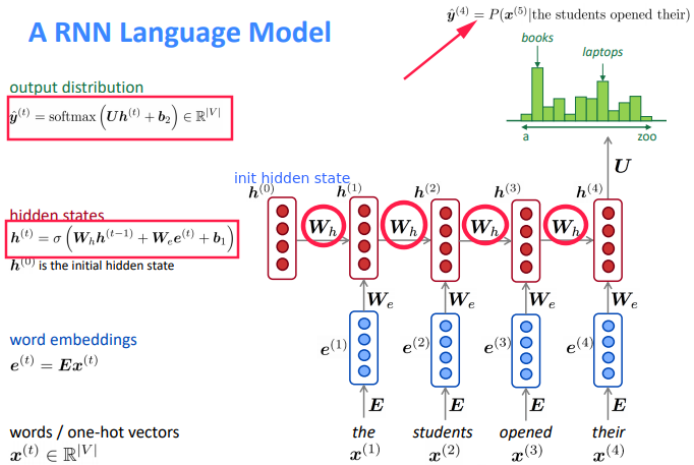


词语嵌入表示



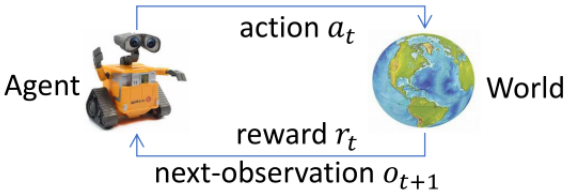
从符号空间到向量空间的映射

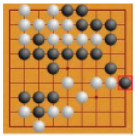
深度学习



Christopher Manning, CS224n, Lecture 6 – Language Models and RNNs

强化学习





ANDREW BARTO
AND
RICHARD SUTTON

2024 ACM A.M. TURING AWARD
RECIPIENTS

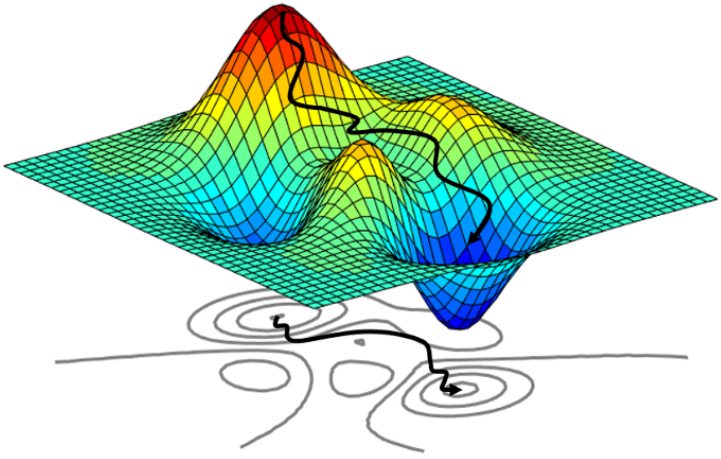
Goal of RL

At each step t, given history so far s_t, take action a_t to maximize long-term reward ("return"):

$$R_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots$$

"Reinforcement Learning: An Introduction", 2nd ed., Sutton & Barto
Jianfeng Gao, Michel Galley, Neural Approaches to Conversational AI (slides), ICML 2019

最优化方法



https://towardsdatascience.com/an-introduction-to-surrogate-optimization-intuition-illustration-case-study-and-the-code-5d9364aed51b

模型训练，就是在整个模型空间内搜索，寻找一个模型，使得模型在训练数据上的损失函数达到最小值。
而最优化算法的目标，就是设计一种算法，使得整个搜索的过程越快越好。

- 模糊逻辑
- 遗传算法
- 蚁群优化算法
- 粒子群优化算法
- 免疫算法
- 分布估计算法
- Memetic算法
- 模拟退火算法
- 禁忌搜索算法

语言模型定义

- 语言也可以定义为由该语言的词表中的单词组成的所有合法句子的集合。
- 一个统计语言模型就是由一个给定的词表中的单词组成的所有可能的句子上的一个概率分布：

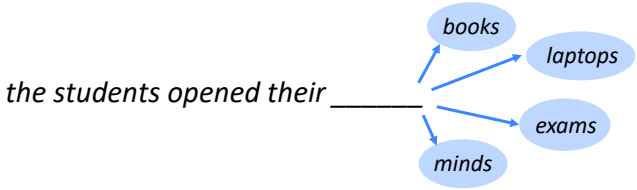
$$\sum_{s \in V^+} P_{LM}(s) = 1$$

- 或者:

$$\sum_{\substack{s=w_1 w_2 \ldots w_n \\ w_i \in V, n>0}} P_{LM}(s) = 1$$

语言模型定义

- 语言建模的目的是完成预测下一个词的任务：



- 给定一个单词序列$x^{(1)}, x^{(2)}, ..., x^{(t)}$，计算下一个词$x^{(t+1)}$的概率分布：

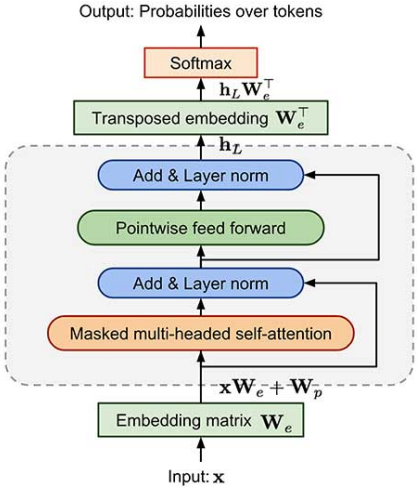
$$P(\boldsymbol{x}^{(t+1)} | \boldsymbol{x}^{(t)}, \ldots, \boldsymbol{x}^{(1)})$$

其中$x^{(t+1)}$可以是词表$V = \{w_1, w_2, ..., w_{|V|}\}$中的任意一个单词。

- 一个能够完成上述任务的系统就可以称为一个语言模型。

Christopher Manning, Natural Language Processing with Deep Learning, Standford U. CS224n

Transformer模型

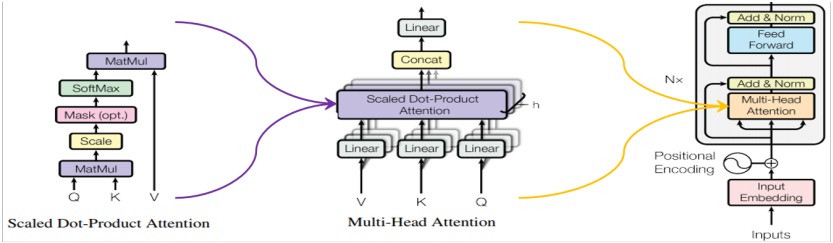


Transformer Block
Repeat x L=12

$$\mathbf{h}_\ell = \text{transformer\_block}(\mathbf{h}_{\ell-1})$$
$$\ell = 1, \ldots, L$$

Liliang Wen, Generalized Language Models: Ulmfit & OpenAI GPT (blog)

自注意力机制（self-attention）



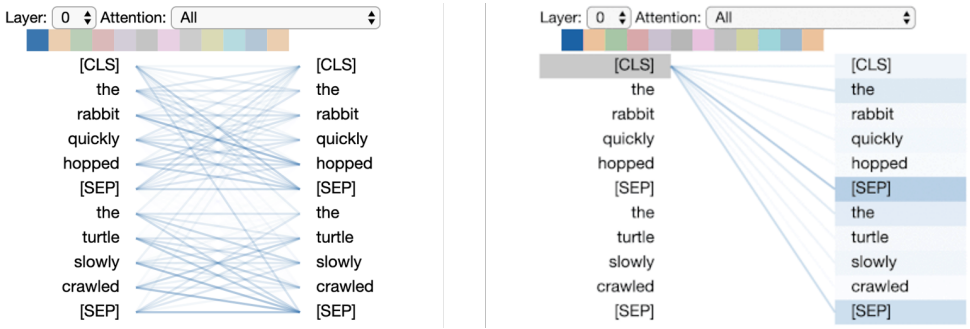
$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,...,\text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

(Vaswani et al., 2017)

自注意力机制（self-attention）

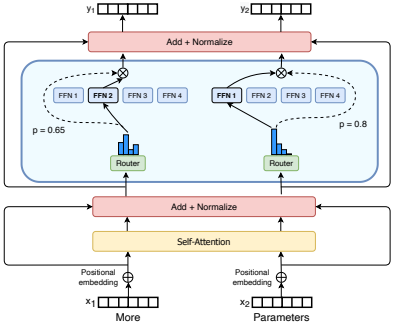
- 每个token是通过所有词动态加权得到
- 动态权重会随着输入的改变而变化



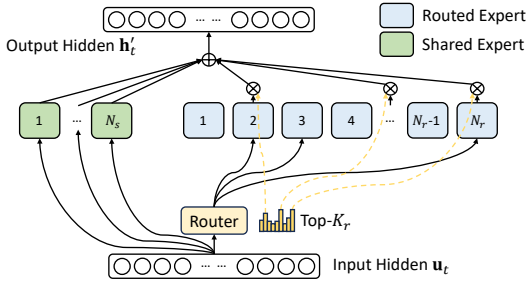
(BertViz tool, Vig et al., 2019)

大语言模型的模型架构：稠密 vs 稀疏

- 目前大语言模型的主体架构都收敛到Transformer Decoder及其变种
- 稀疏架构的Transformer（如MoE）被一些模型采用，以节省算力
- 由于算力约束，目前超大语言模型越来越多采用稀疏架构。如GPT-4架构被披露是采用了MLP111B*16Expert+Attention*55B形式。DeepSeek V3采用独创的DeepSeekMoE稀疏架构，每个MoE层包含1个共享专家和256个路由专家，每个Token选择8个路由专家。

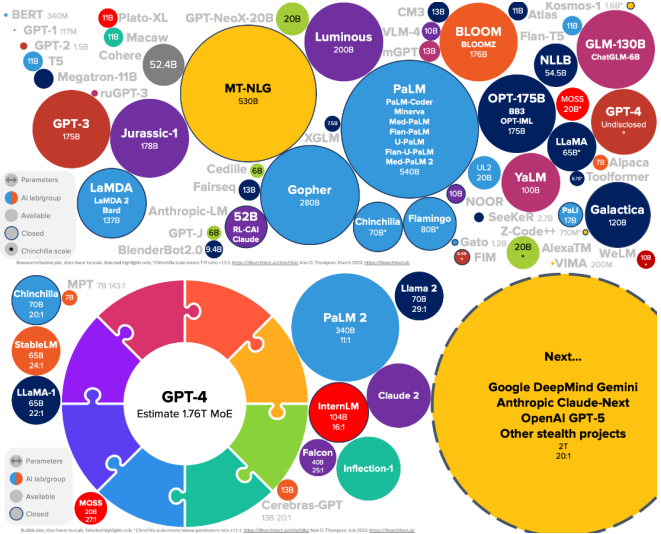


Fedus, et al. “Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity” 2021. arxiv2101.03961v1.

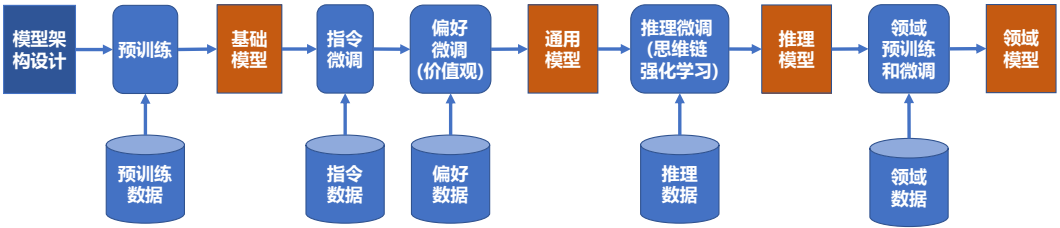


DeepSeek, DeepSeek-V3 Technical Report. 2024.12.27.

大语言模型的参数规模

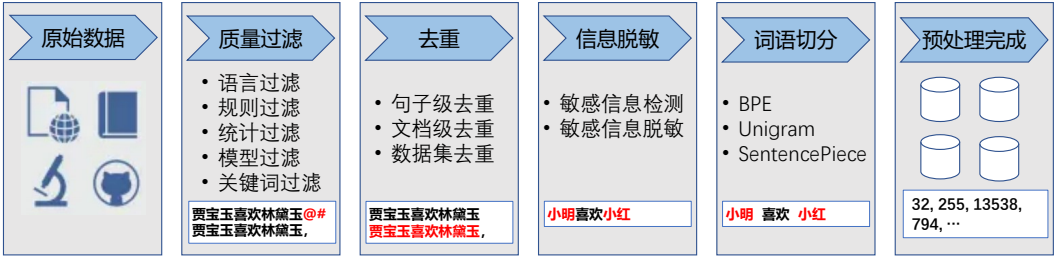


大语言模型训练流程

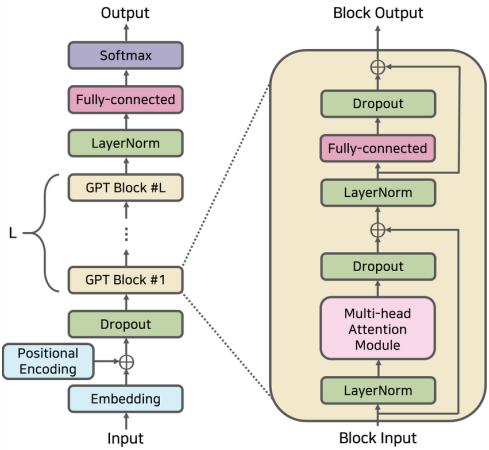


- 预训练 – 监督学习 – 预测下一个词 – 死记硬背
- 指令微调 – 监督学习 – 自己做练习题 – 学会遵从指令
- 偏好微调 – 人类反馈的强化学习（RLHF）– 老师指导做练习题 – 学会人类偏好（价值观等）
- 推理微调 – 思维链强化学习（CoT RL）– 闯社会 – 学会推理
- 领域训练和微调 – 使用领域数据进行继续预训练和微调（含前面训练过程的任意组合）

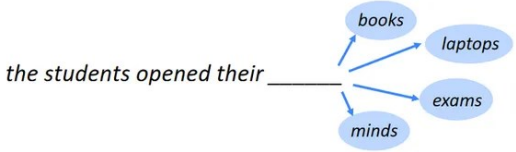
大语言模型数据预处理



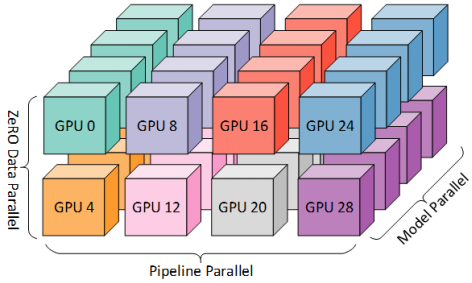
大语言模型预训练



https://www.mdpi.com/2227-7390/11/10/2320



https://www.analyticsvidhya.com/blog/2023/07/next-word-prediction-with-bidirectional-lstm/



https://huggingface.co/docs/transformers/v4.15.0/en/parallelism

大语言模型的预训练数据

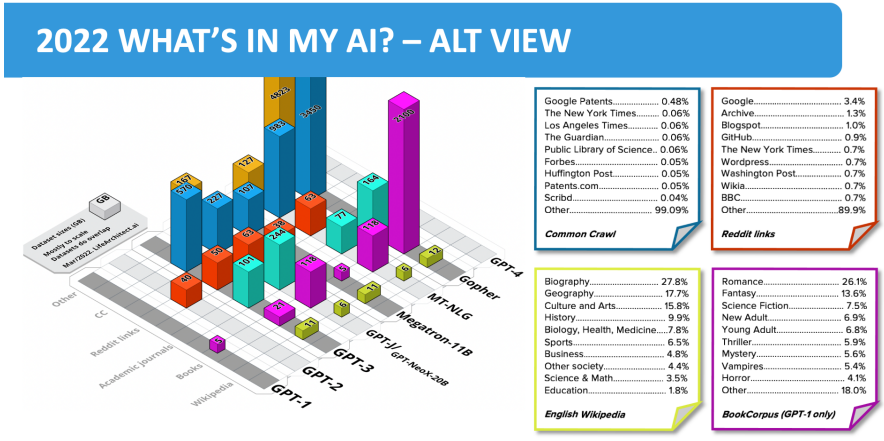
Dataset	Tokens (billion)	Assumptions	Tokens per byte (Tokens / bytes)	Ratio	Size (GB)
Web data	**410B**	–	*0.71*	*1:1.9*	**570**
WebText2	**19B**	*25% > WebText*	*0.38*	*1:2.6*	*50*
Books1	**12B**	*Gutenberg*	*0.57*	*1:1.75*	*21*
Books2	**55B**	*Bibliotik*	*0.54*	*1:1.84*	*101*
Wikipedia	**3B**	*See RoBERTa*	*0.26*	*1:3.8*	*11.4*
Total	**499B**			*753.4GB*	

Table. GPT-3 Datasets. Disclosed in **bold**. Determined in *italics*.

Alan D. Thompson, GPT-3.5 + ChatGPT: An illustrated overview, https://lifearchitect.ai/chatgpt/

大语言模型的预训练数据

数据来源：各个大语言模型的对比



大语言模型的预训练数据

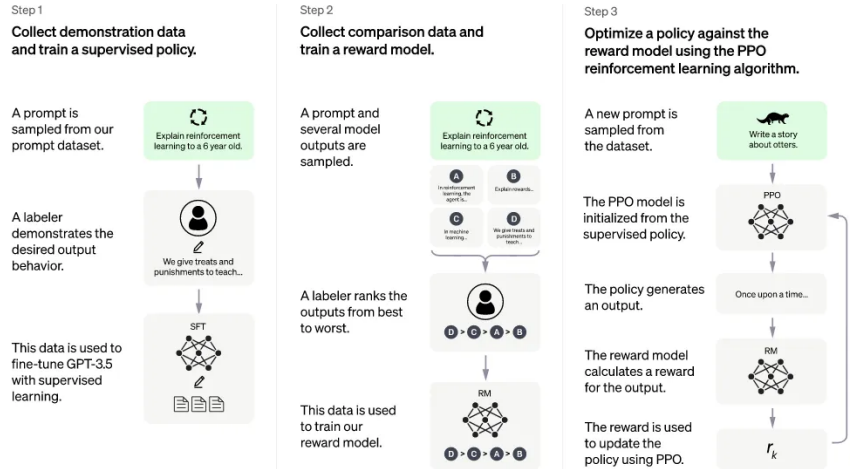
看一下大语言模型训练的token数量：

- GPT-3（2020.5）是500B（5000亿）tokens，目前最新数据未知；
- Google的PaLM（2022.4）是780B tokens；
- DeepMind的Chinchilla是1400B=1.4T tokens；
- 开源的Llama的训练数据是1.5T tokens，Llama2的训练数据是2T tokens。
- GPT-4（2023）的训练数据13T文本（包括代码）tokens+2T图像tokens。

大语言模型的数据工程：

- LLM训练需要海量的数据；
- 数据质量对最后的模型有巨大影响；
- 如果收集过滤海量高质量数据，是非常关键的。

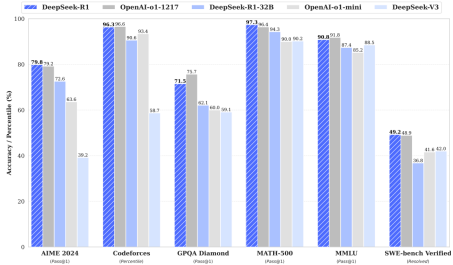
大语言模型后训练（指令微调和人类偏好训练）



Credit to: Ouyang et al., Training language models to follow instructions with human feedback, arXiv:2203.02155

大语言模型的思维链强化学习训练：DeepSeek-R1核心创新点

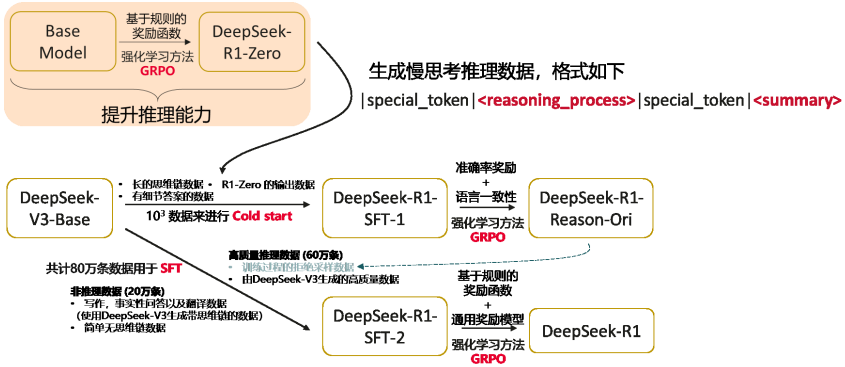
- 无需SFT的复杂推理能力涌现
  - 模型自主演化出了诸如**自我反思**、**自我进化**等类人推理行为，并显著提高了模型的复杂高阶推理能力。
- 多阶段训练框架：平衡推理与语言能力
- 结构化慢思考模板：提升结果可解释性
  - 将推理过程与最终答案进行分离（例如采用 <think>...</think> <answer>...</answer> 的格式）
- 基于规则的奖励函数设计：提升推理的准确性
  - 结果正确：对于每一个训练使用的问题，提前准备正确答案，而不是另一个reward model来给模型的输出打分
  - 格式正确：要求模型必须以固定格式输出思维链，即必须在“<think></think>”中间



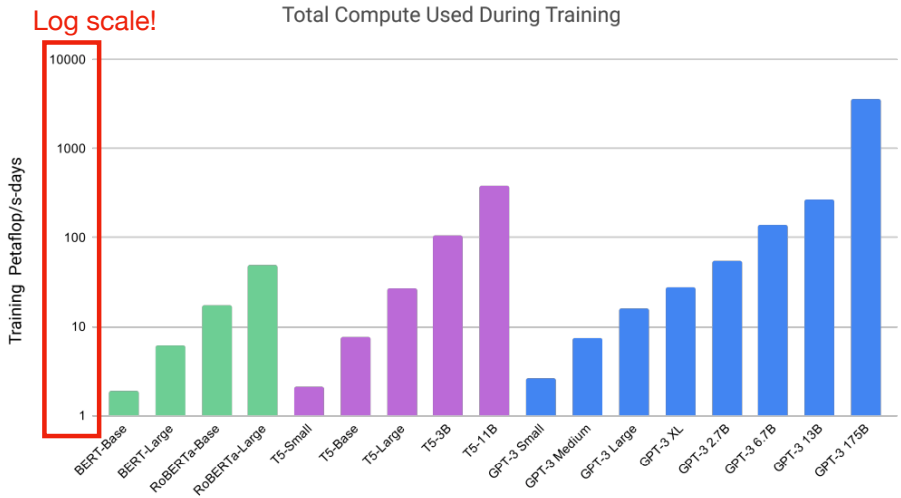
Deepseek R1在推理任务和通用任务显著超越OpenAI-o1

大语言模型的思维链强化学习训练：DeepSeek-R1训练流程

- **R1-Zero训练阶段**：以预训练模型为基础，暂时搁置对通用能力的追求，借助强化学习直接激活模型的慢思考推理能力，大幅拔高推理能力。
- **R1训练阶段**：利用具备强大推理能力的 R1-Zero 反向生成数据，采用多轮RL+SFT迭代训练，激活模型慢思考推理能力和通用能力。

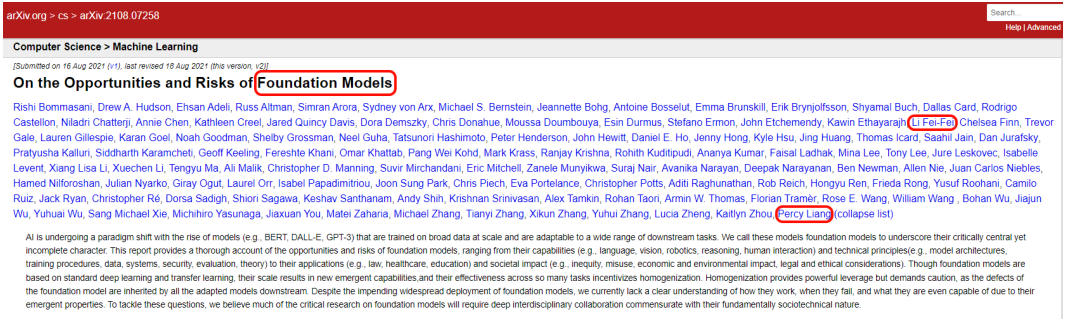


大语言模型的算力消耗



Mohit Iyyer, slides for CS685 Fall 2020, University of Massachusetts Amherst

涌现（Emergence）和同一化（homogenization）



arXiv.org > cs > arXiv:2108.07258

Search...

Help | Advanced

Computer Science > Machine Learning

[Submitted on 16 Aug 2021 (v1), last revised 18 Aug 2021 (this version, v2)]

On the Opportunities and Risks of Foundation Models

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Kohd, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang , Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, Percy Liang (collapse list)

AI is undergoing a paradigm shift with the rise of models (e.g., BERT, DALL-E, GPT-3) that are trained on broad data at scale and are adaptable to a wide range of downstream tasks. We call these models foundation models to underscore their critically central yet incomplete character. This report provides a thorough account of the opportunities and risks of foundation models, ranging from their capabilities (e.g., language, vision, robotics, reasoning, human interaction) and technical principles(e.g., model architectures, training procedures, data, systems, security, evaluation, theory) to their applications (e.g., law, healthcare, education) and societal impact (e.g., inequity, misuse, economic and environmental impact, legal and ethical considerations). Though foundation models are based on standard deep learning and transfer learning, their scale results in new emergent capabilities,and their effectiveness across so many tasks incentivizes homogenization. Homogenization provides powerful leverage but demands caution, as the defects of the foundation model are inherited by all the adapted models downstream. Despite the impending widespread deployment of foundation models, we currently lack a clear understanding of how they work, when they fail, and what they are even capable of due to their emergent properties. To tackle these questions, we believe much of the critical research on foundation models will require deep interdisciplinary collaboration commensurate with their fundamentally sociotechnical nature.

Bommasani et al., On the Opportunities and Risks of Foundation Models, arXiv:2108.07258 [cs.LG]

涌现（Emergence）和同一化（homogenization）

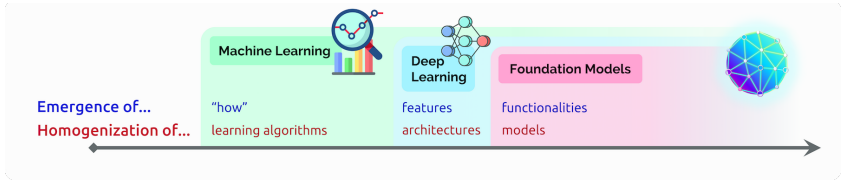


Fig. 1. The story of AI has been one of increasing *emergence* and *homogenization*. With the introduction of machine learning, *how* a task is performed emerges (is inferred automatically) from examples; with deep learning, the high-level features used for prediction emerge; and with foundation models, even advanced functionalities such as in-context learning emerge. At the same time, machine learning homogenizes learning algorithms (e.g., logistic regression), deep learning homogenizes model architectures (e.g., Convolutional Neural Networks), and foundation models homogenizes the model itself (e.g., GPT-3).

	机器学习	深度学习	基础模型
涌现	解决问题的方法	特征	模型能力
同一化	学习算法	模型结构	模型本身

Bommasani et al., On the Opportunities and Risks of Foundation Models, arXiv:2108.07258 [cs.LG]

少样本和零样本学习 (上下文学习 in-context learning)

Zero-shot

The model predicts the answer given only a natural language discription of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

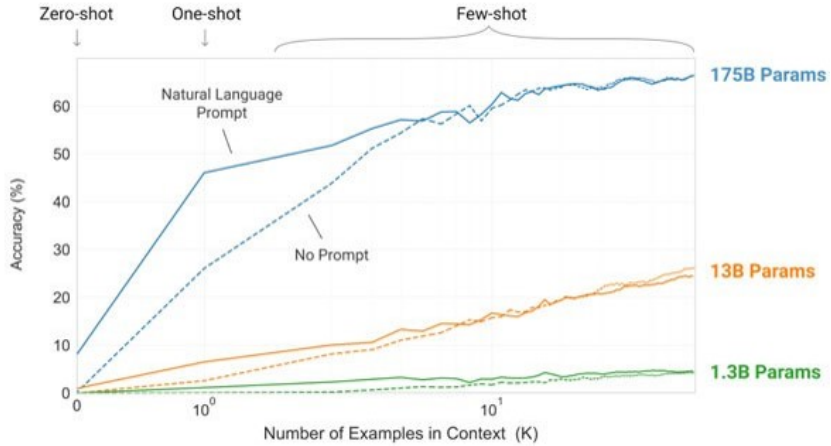
```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée
4  plush girafe => girafe peluche
5  cheese =>                           ← prompt
```

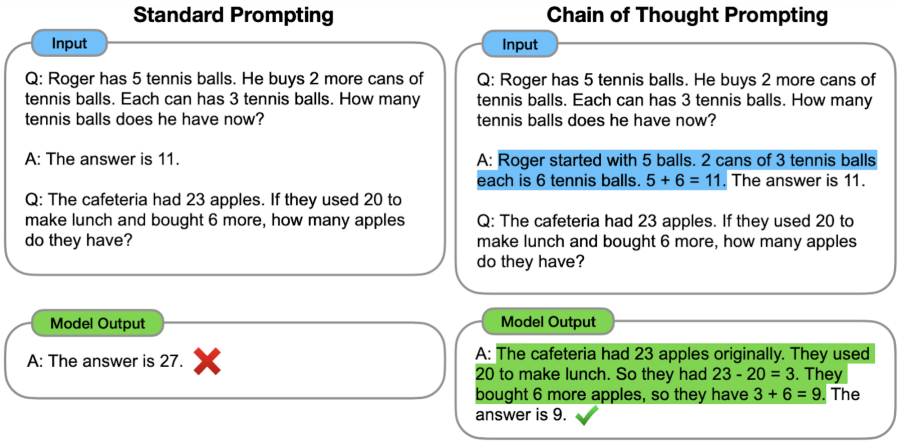
少样本和零样本学习 (上下文学习 in-context learning)



Brown et al., Language Models are Few-Shot Learners,

arXiv:2005.14165, 2021

思维链 Chain-of-thought



Preprint: https://arxiv.org/pdf/2201.11903.pdf

零样本思维链 Zeor-shot CoT: Let's think step-by-step

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The answer is 8. ✗

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are 16 / 2 = 8 golf balls. Half of the golf balls are blue. So there are 8 / 2 = 4 blue golf balls. The answer is 4. ✓

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: The answer (arabic numerals) is

(Output) 8 ✗

(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: ***Let's think step by step.***

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

Preprint: http://arxiv.org/abs/2205.11916

大语言模型的能力涌现（Emergent Ablities）

上下文学习（零样本/少样本学习）

Zero-shot

The model predicts the answer given only a natural language discription of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

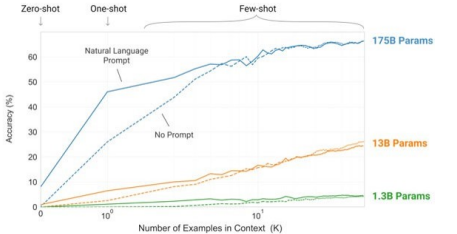
```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

Few-shot

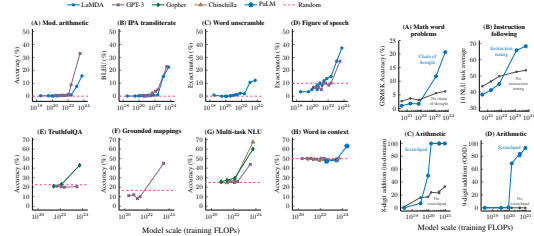
In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ←
4  plush girafe => girafe peluche      ←
5  cheese =>                           ← prompt
```

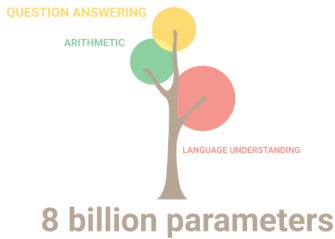
上下文学习的能力涌现



其他能力涌现



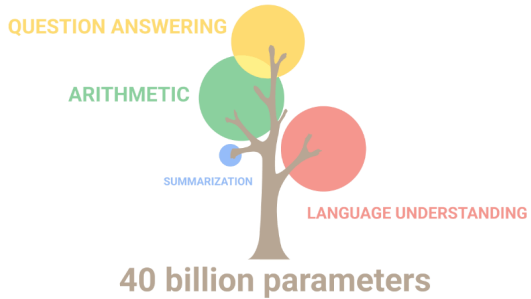
能力涌现随着模型规模增加一直在持续



As the scale of the model increases, the performance improves across tasks while also unlocking new capabilities.

Blog: Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance

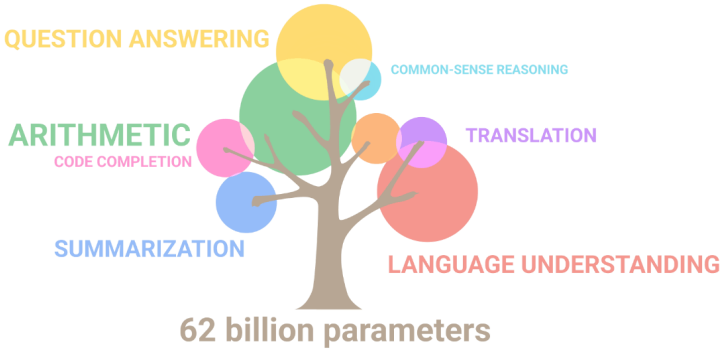
能力涌现随着模型规模增加一直在持续



As the scale of the model increases, the performance improves across tasks while also unlocking new capabilities.

Blog: Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance

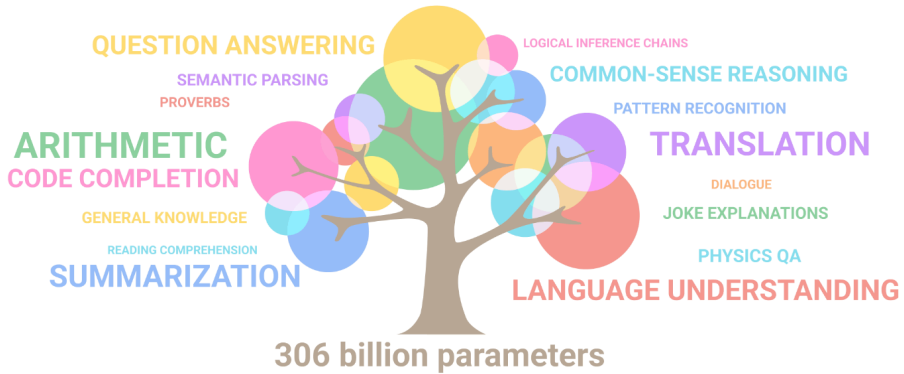
能力涌现随着模型规模增加一直在持续



As the scale of the model increases, the performance improves across tasks while also unlocking new capabilities.

Blog: Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance

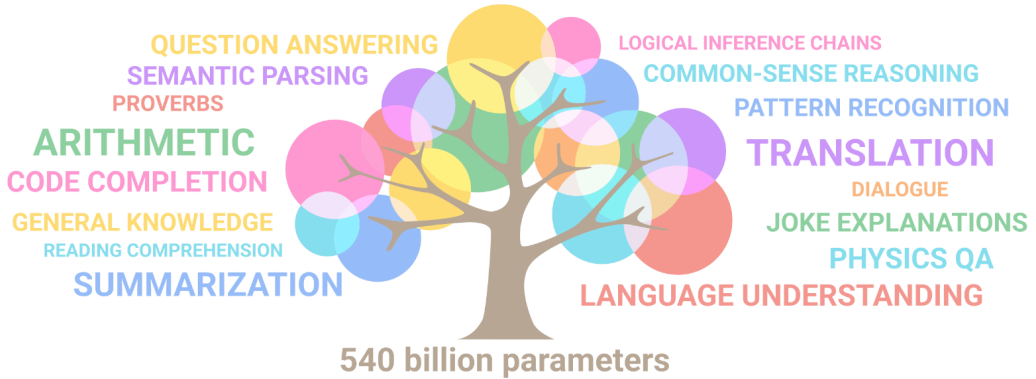
能力涌现随着模型规模增加一直在持续



As the scale of the model increases, the performance improves across tasks while also unlocking new capabilities.

Blog: Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance

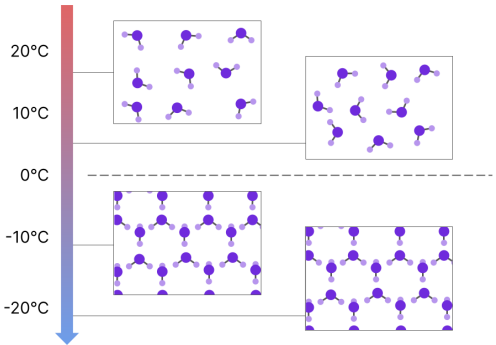
能力涌现随着模型规模增加一直在持续

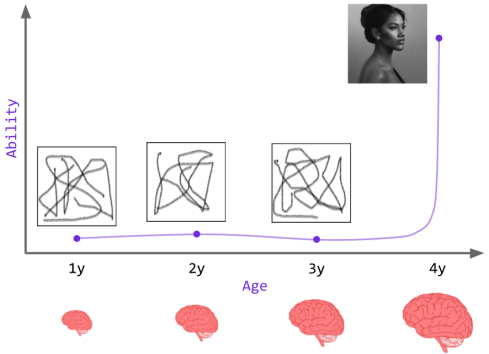


As the scale of the model increases, the performance improves across tasks while also unlocking new capabilities.

Blog: Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance

能力涌现 Ability Emergence的解释



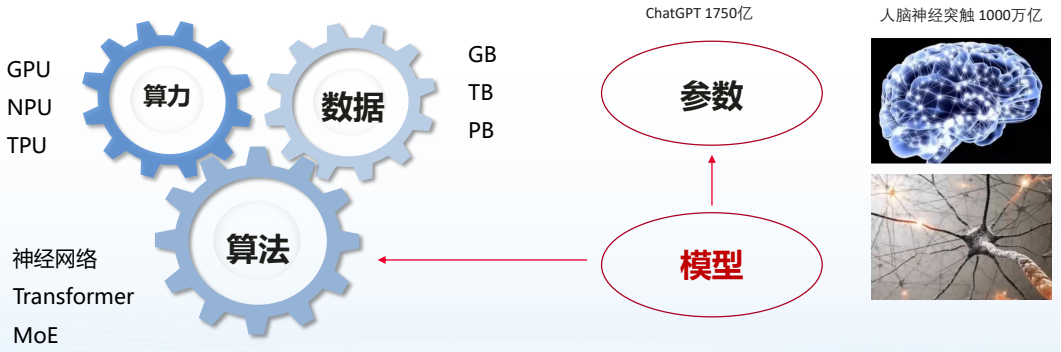


Blog: Emergent Abilities of Large Language Models, url

大模型训练的基本要素：算力、算法、数据、模型、参数

算力、算法、数据三大要素构筑人工智能基础

人工智能进入大模型、大算力、大数据时代



大模型预训练增长定律（Pre-training Scaling Law）

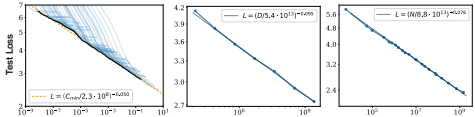


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

Kaplan et al. “Scaling Laws for Neural Language Models.” 2000.

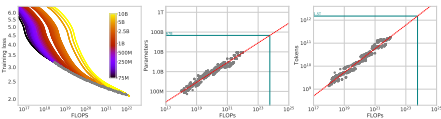
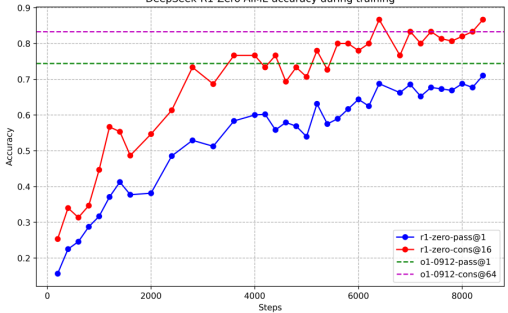


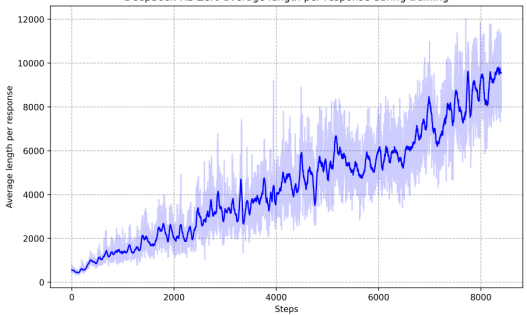
Figure 2 | **Training curve envelope.** On the **left** we show all of our different runs. We launched a range of model sizes going from 70M to 10B, each for four different cosine cycle lengths. From these curves, we extracted the envelope of minimal loss per FLOP, and we used these points to estimate the optimal model size (**center**) for a given compute budget and the optimal number of training tokens (**right**). In green, we show projections of optimal model size and training token count based on the number of FLOPs used to train *Gopher* (5.76×10^{23}).

Hoffmann et al. “Training Compute-Optimal Large Language Models.” 2022.

- 在大语言模型的预训练中，观察预训练效果最重要的指标是模型的训练loss，只要训练的loss在不断降低，我们就可以有信心这个模型在不断地学到新的知识
- 研究人员发现，模型的训练loss跟模型的参数规模、模型的训练数据大小、模型训练所使用的算力几乎都呈现某种线性关系（对数坐标下），因此，可以根据模型的参数规模、训练数据大学和训练所使用的算力来预测模型的loss，这就是所谓的LLM预训练Scaling Law
- 另有研究人员发现，在给定的有限算力（模型规模乘以训练量）下，存在一个最优的模型规模，并不是模型规模越大越好。太大的模型可能因为训练不充分反而达不到最优的效果（Chinchilla Law）。
- LLM预训练Scaling Law随着训练的算力、能源成本越来越高、数据逐渐枯竭，GPT-5遥遥无期，Grok3性能也远没有做到经验。这一切似乎显示出预训练增长定律带来的大模型性能不断增长趋势已经放缓。但这方面的投资还没有终止，如xAI和星际之门项目。

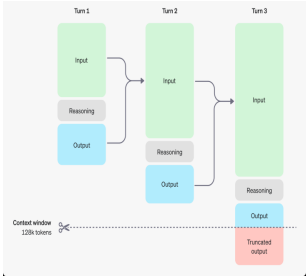
大模型强化学习后训练增长定律 RL Post Training Scaling Law

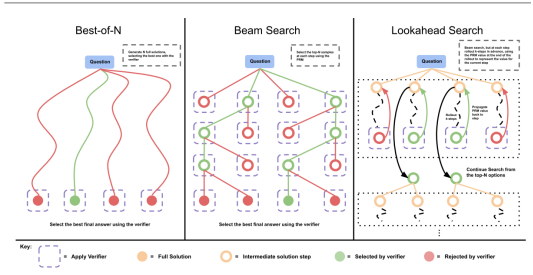


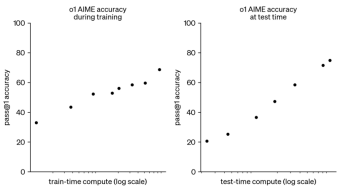


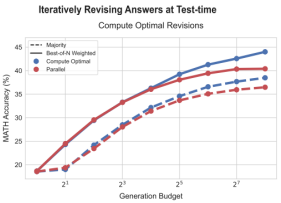
DeepSeek R1 Technique Report

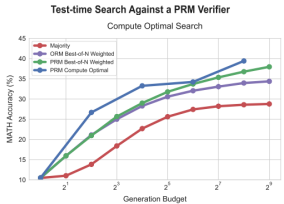
大模型推理增长定律 Test Time Scaling Law







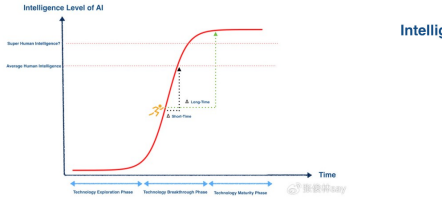


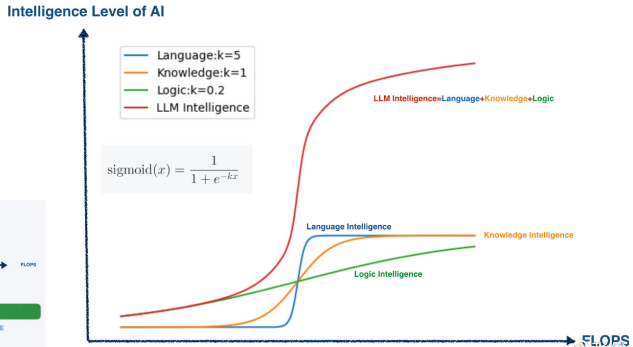


OpenAI o1

Snell et al. Scaling LLM Test-Time Compute……, arXiv:2408.03314v1

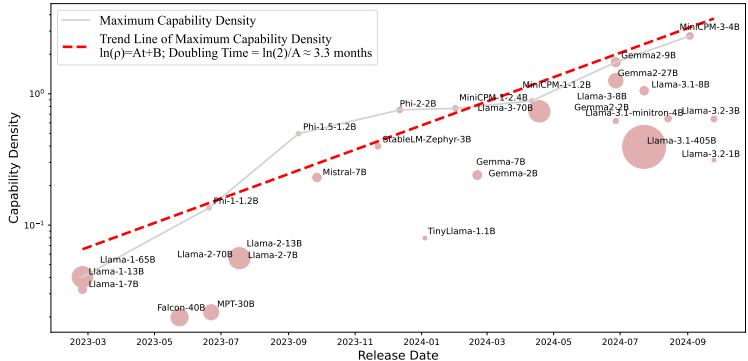
多重增长定律的叠加效应





张俊林，S型智能增长曲线：从Deepseek R1看Scaling Law的未来，https://weibo.com/1064649941/PdBGubw

大模型能力密度定律



- LLM能力密度定律：模型能力密度随时间呈指数级增长，2023年以来能力密度约每3.3个月（约100天）翻一倍。
- “能力密度”定义为了“有效参数量”与实际参数量的比值。其“有效参数量”被定义为实现与目标模型一样的效果（5个最主要的banchmark评分），参考模型需要的最少参数量。
- 根据拟合曲线，到了2025年底，只要8B参数就能实现和GPT-4一样的效果。

大模型技术概览

模型架构与预训练

- 线性模型和混合架构
- 分布式并行训练
- 稀疏架构（MoE、MoD）
- 模型蒸馏、剪枝、扩增
- 低精度、混合精度训练
- 长序列训练
- 扩散模型

模型Infra和推断部署

- PD分离部署
- 参数量化、KVCatch量化
- 长序列压缩推理
- 投机推理
- 负载均衡

数据工程

- 数据挖掘
- 数据配比
- 数据合成
- 长文本数据
- 思维链推理数据
- 多模态推理数据

Agent与具身智能

- 检索 DeepSearch, DeepResearch
- 代码生成/代码注释/代码修改
- 工具调用/代码解释器
- 数字助手/虚拟人/GUI Agent
- 多智能体
- 自动驾驶
- 机器人

多模态

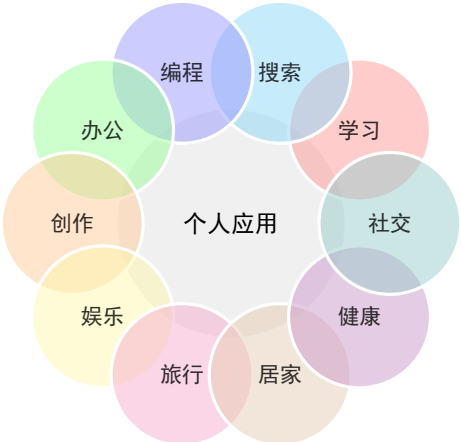
- 语音理解和生成
- 图像理解和生成
- 视频理解和生成
- 全模态理解和生成
- 全模态理解生成一体化
- 视觉语言行为模型VLA

后训练与强化学习

- 监督学习SFT
- 强化学习算法DPO/PPO/GRPO
- 强化学习框架 BoN, Beam Search, MTS
- 奖励设计

Content

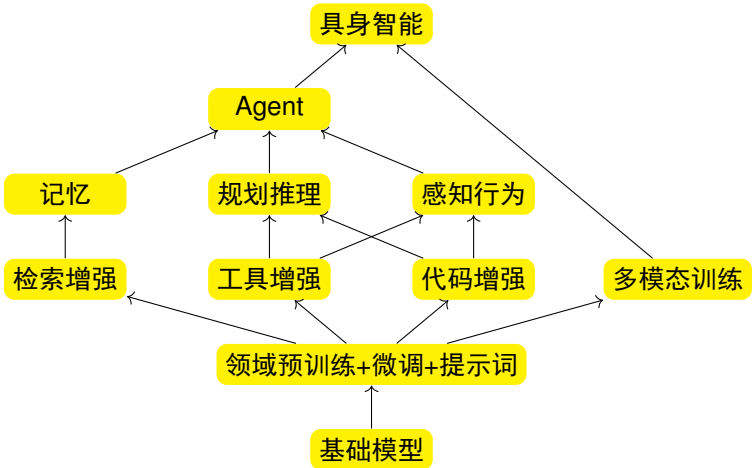
大模型应用场景



大模型应用类型

对话	闲聊	客服	情感抚慰	知识问答	角色扮演
文本生成	文学创作	应用文写作	广告创意生成		
翻译	文本翻译	OCR翻译	字幕翻译	语音翻译	视频翻译(对口型)
文摘	文本摘要	对话摘要	会议摘要		
检索问答	信息查询	文档问答	领域问答		
软件开发	代码理解	代码生成	代码补全	注释生成	
多模态	语音识别	语音合成	图像理解	图像生成	视频理解	视频生成	多模聊天
任务完成	购物	游戏	订餐订票	APP操作	
科学探索	数学求解	定理证明	生物制药	天气预报	新材料发现

大模型训练和应用路径

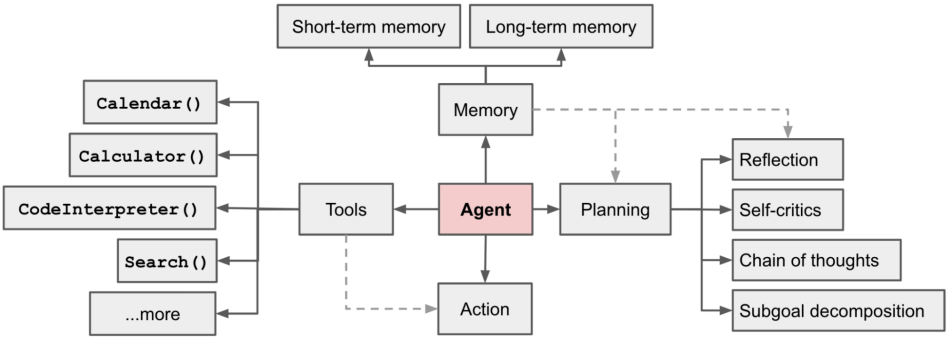


大模型与知识图谱结合：GraphRAG

- 什么是GraphRAG？
  - GraphRAG是一种基于知识图谱的检索增强技术。通过构建图模型的知识表达，将实体和关系之间的联系用图的形式展示出来，然后利用大语言模型（LLM）进行检索增强。
- GraphRAG 的工作原理：
  - 提取实体：从用户输入的查询中提取关键实体。
  - 构建子图：根据提取的实体构建相关的子图，形成上下文。
  - 生成答案：将构建好的子图输入大语言模型，生成答案。
- GraphRAG引起了较多的重视，取得了一定的成功。

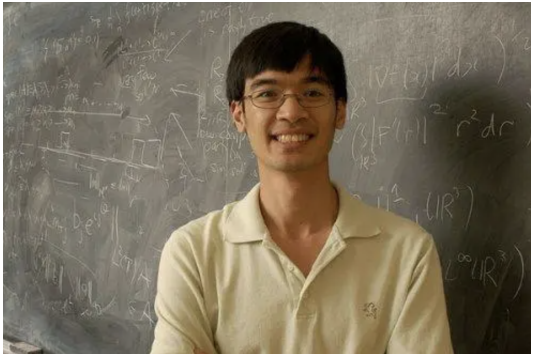
资料来源：CSDN Blog: GraphRAG：知识图谱+大模型, 作者：Python_金钱豹

大语言模型智能体（Agent）架构



Credit to: Lilian Weng, LLM Powered Autonomous Agents

大模型辅助数学研究：陶哲轩把ChatGPT加入日常工作流



Terence Tao

@tao

1 天

Traditional computer software tools resemble the standard mathematical concept of a function $f: X \to Y$: given an input x in the domain X, it reliably returns a single output $f(x)$ in the range Y that depends on x in a deterministic fashion, but is undefined or gives nonsense if fed an input outside of the domain. For instance, the LaTeX compiler in my editor will take my LaTeX code, and - provided that it is correctly formatted, all relevant packages and updates have been installed, etc. - return a perfect PDF version of that LaTeX every time, with no unpredictable variation. On the other hand, if one tries to compile some LaTeX with a misplaced brace or other formatting problem, then the output can range from compilation errors to a horribly mangled PDF, but such results are often obvious to detect (though not always to fix).

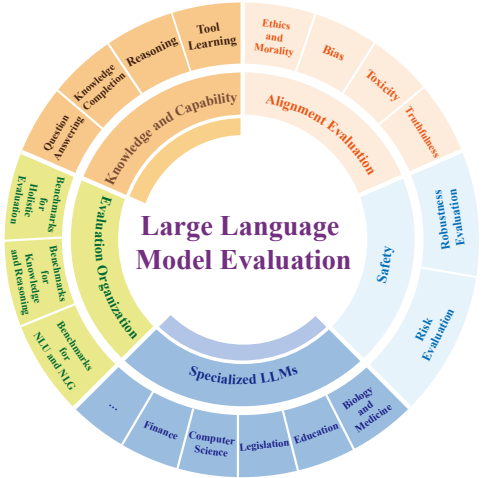
#AI tools, on the other hand, resemble a probability kernel $\mu: X \to \Pr(Y)$ instead of a classical function: an input x now gives a random output sampled from a probability distribution μ_x that is somewhat concentrated around the perfect result $f(x)$, but with some stochastic deviation and inaccuracy. In many cases the inaccuracy is subtle; the random output superficially resembles $f(x)$ until inspected more closely. On the other hand, such tools can handle noisy or badly formatted inputs x much more gracefully than a traditional software tool.

Because of this, it seems to me that the way AI tools would be incorporated into one's workflow would be quite different from what one is accustomed to with traditional tools. An AI LaTeX to PDF compiler, for instance, would be useful, but not in a "click once and forget" fashion; it would have to be used more interactively.

Content

大语言模型评价任务的多样性

与传统NLP任务的评价不同，大语言模型是一个通用模型，并不存在单一的评价标准。



Guo et al., Evaluating Large Language Models: A Comprehensive Survey, arXiv:2310.19736v3, 2023

大语言模型评价的挑战

- 大语言模型的评价面临的挑战主要体现在：
  - 任务的多样性
  - 评价的主观性
  - 评测数据的代表性
  - 评测数据的时效性
  - 评测数据泄露和污染
  - 评测结果的可解释性

AI大模型面临的问题

- 幻象问题：大模型存在幻象，很难用在关键任务（如医疗）中
- AI安全问题：伦理（歧视、冒犯、、价值观）安全、滥用问题
- 可持续发展问题：能源消耗现有承受能力，数据枯竭
- 大模型是否是通向通用人工智能（AGI）的正确路径？
  - 否：大模型只是部分的智能，远远不是智能的全部，也不可能只通过大模型达到通用人工智能
  - 是：大模型可以通向通用人工智能（AGI），甚至达到超人智能（ASI），这一天很快将到来
- 如果AI的智力水平超过人类会导致什么后果？
  - 后果很严重，AI可能操纵人类、统治人类甚至毁灭人类
  - 不会有严重后果，AI只是工具，智力水平再高也不会统治人类

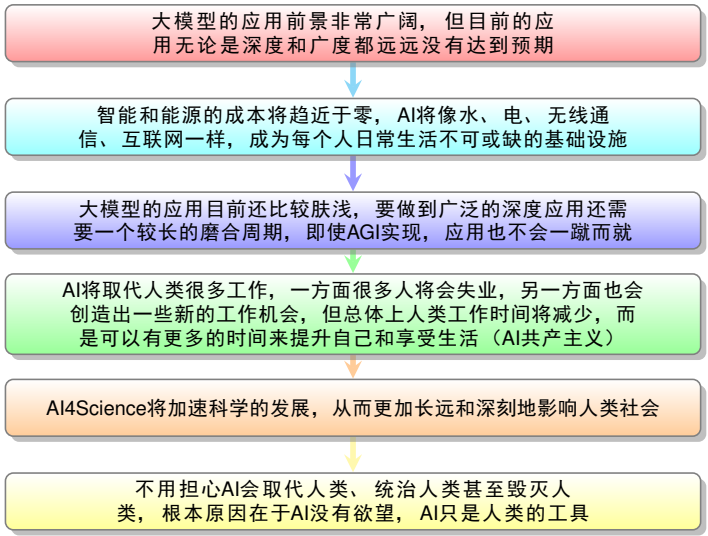
幻象问题

- 现有神经网络本身无法避免幻象问题：
  - 神经网络的知识都存储在参数之中，参数本身是无法区分事实和幻象的。
  - 神经网络生成的文本或图像也无法区分事实和幻象。
- 更大的模型（如GPT-4）可以更好地对数据建模，有助于减少幻象。
- 引入外部知识（如RAG），可以很好地减少幻象。
- 人类也有幻象（儿童、病人、梦境、文学创作），幻象不一定都是坏事。
- 一种可能的消除幻象的方案，是在模型内部引入事实性判断模块。
- 在特定应用领域，只要能把幻象减少到足够低，就可以满足需求，并不一定需要彻底消除幻象。

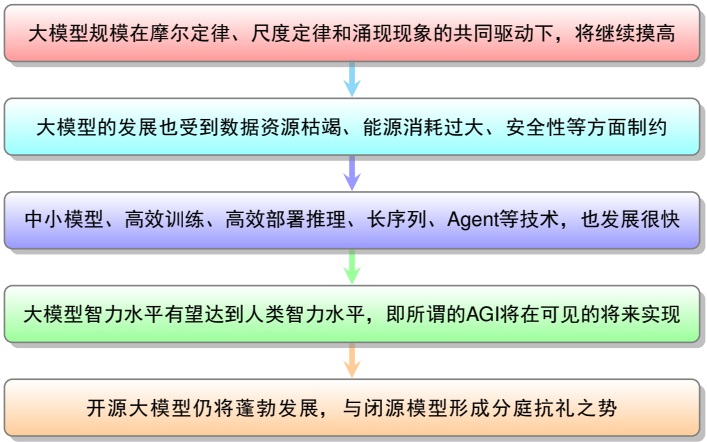
超人智能和毁灭人类的风险

- AI能力将在越来越多的专业领域内超过大部分普通人、甚至超过一般的专家，成为超人智能（Superhuman Intelligence）。
- AI在日常生活中超过人类，还有很长的路要走。
- AI毁灭人类的可能性：
  - AI没有毁灭人类的意图（有能力并不等于有意图）。
  - AI可能无意中毁灭人类：回形针思想实验。
  - 回形针思想实验的问题（个人意见）：
    - 资源的授权；
    - 从失败中恢复的意志。

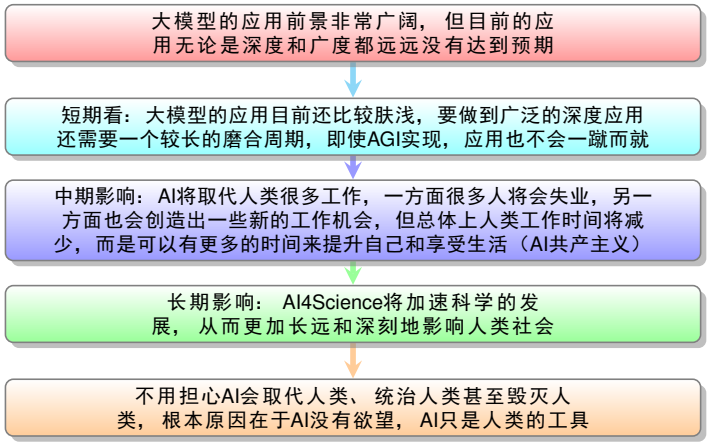
大语言模型应用趋势及对人类社会的影响



大语言模型技术发展趋势



大语言模型应用发展趋势



OpenAI定义的AI发展的五个级别

OpenAI Imagines Our AI Future

Stages of Artificial Intelligence

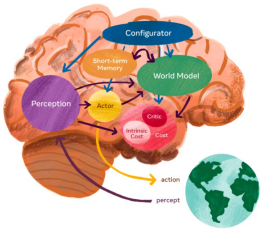
Level 1	Chatbots, AI with conversational language	**L1：聊天机器人，具有对话能力的AI。**
Level 2	Reasoners, human-level problem solving	**L2：推理者，像人类一样能够解决问题的AI。**
Level 3	Agents, systems that can take actions	**L3：智能体，不仅能思考，还可以采取行动的AI系统。**
Level 4	Innovators, AI that can aid in invention	**L4：创新者，能够协助发明创造的AI。**
Level 5	Organizations, AI that can do the work of an organization	**L5：组织者，可以完成组织工作的AI。**

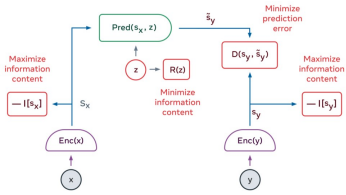
Source: Bloomberg reporting

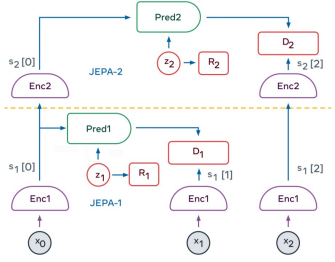
李飞飞团队提出空间智能

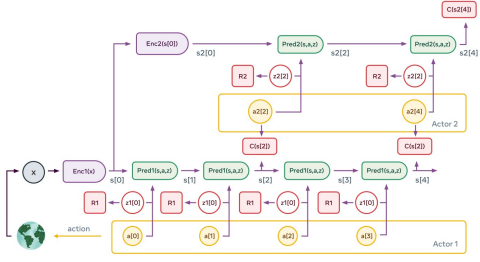
- 李飞飞提出的空间智能是指机器在 3D 空间和时间中感知、推理和行动的能力。
- 李飞飞认为空间智能与语言智能一样重要，甚至在某些方面更古老、更基础。从进化角度看，智能的进化使动物尤其是人类能够在三维的真实物理环境中移动、互动并创造文明，而空间智能是实现这些的关键，是构建未来 AI 生态的重要部分。
- 传统基于语言的 AI 底层表示是一维的，而空间智能是通过 3D 模型来推动机器理解物理世界的本质，不仅仅是对图像或视频的 2D 处理，而是让机器真正理解并应对三维空间。
- 在实现空间智能时，需要维持物体的永久性和遵守物理法则，例如让生成的 3D 场景中的物体根据重力或其他物理规则正确地与环境交互，这对技术实现有较高要求。
- 空间智能具有广泛的应用前景，可应用于游戏、教育、虚拟摄影等领域，能创建充满活力和交互性的 3D 世界，降低 3D 内容制作成本，激发更多沉浸式体验。在增强现实（AR）和虚拟现实（VR）中也有重要作用，有望成为 AR/VR 的“操作系统”，帮助人类增强能力，如佩戴 AR 眼镜的人可借助它修理汽车或完成复杂操作，还可能减少人们对手机、平板等屏幕的依赖。

Yann LeCun的自主智能构想和联合嵌入预测架构（JEPA）









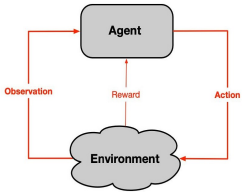
阿尔伯塔计划（DeepMind阿尔伯塔实验室，Richard Sutton）

愿景（Research Vision）

智能作为随时间进行的信号处理（Intelligence as signal processing over time）。智能体的智能行为不仅仅是对即时刺激的反应，还包括感知、分析和在时间背景下对信号作出反应的能力。智能被视为一种连续而动态的过程，它随着时间的推移获取、处理和利用信息，以便做出明智的决策并采取适当的行动。
特点一：是基于模型的连续学习。
特点二：是持续学习和元学习。
特点三：是考虑算力的因素。
特点四：是包括与其他智能代理进行交互的情况。

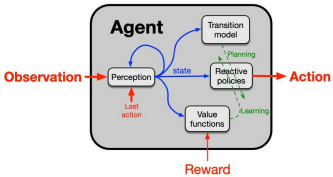
路线图

1. 表示 I: 给定特征的持续监督学习。（Representation I: Continual supervised learning with given features.）
2. 表示 II: 监督特征发现。（Representation II: Supervised feature finding.）
3. 预测 I: 持续广义值函数（GVF）预测学习。（Prediction I: Continual Generalized Value Function (GVF) prediction learning.）
4. 控制 I: 持续的演员-评论者控制。（Control I: Continual actor-critic control.）
5. 预测 II: 平均奖励GVF学习。（Prediction II: Average-reward GVF learning.）
6. 控制 II: 持续的控制问题。（Control II: Continuing control problems.）
7. 规划 I: 带平均奖励的规划。（Planning I: Planning with average reward.）
8. AI原型 I: 具有持续函数逼近的单步模型基于强化学习。（Prototype-AI I: One-step model-based RL with continual function approximation. ）
9. 规划 II: 搜索控制和探索。（Planning II: Search control and exploration.）
10. AI原型 II: STOMP进展。（Prototype-AI II: The STOMP progression.）
11. AI原型 III: Oak。（Prototype-AI III: Oak.）
12. IA原型: 智能增强。（Prototype-IA: Intelligence amplification.）

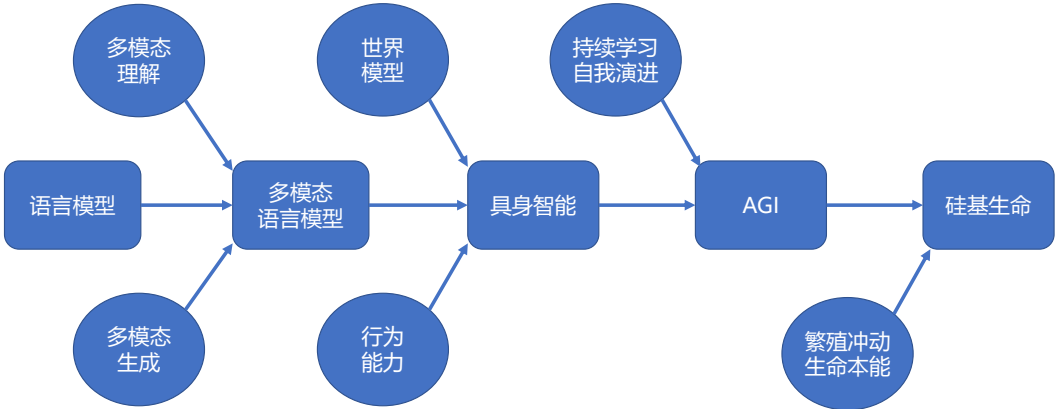


基本智能体有四个主要部件：

1. 感知（Perception）
2. 反应性策略（Reactive policies）
3. 值函数（Value functions）
4. 转移模型（Transition model）



通向AGI之路



Content

总结

人工智能(AI)简介和发展历程

AI大模型现状及近期热点

AI大模型技术简介

AI大模型应用

AI大模型面临的问题、发展趋势和对策

Thank you!

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。

Bring digital to every person, home and organization for a fully connected, intelligent world.