



上海大学未来技术学院 | 上海大学人工智能研究院
SCHOOL OF FUTURE TECHNOLOGY, SHANGHAI UNIVERSITY | INSTITUTE OF ARTIFICIAL INTELLIGENCE, SHANGHAI UNIVERSITY

人工智能导论

——第7课：大模型与智能体

叶林奇

未来技术学院（人工智能研究院）

2024秋季学期



提纲

一、大模型

二、智能体



上海大学
SHANGHAI UNIVERSITY



大模型

自1956年达特茅斯会议至今，人工智能经历三个主要发展阶段

呈现从手工总结到自动学习，从离散符号到连续参数，从**专用智能到通用智能**的趋势

符号智能

时间：1950-1990

知识来源：领域专家

获取方法：手工总结

知识形式：以**知识库**等**符号系统**
存储**专家知识**

专用智能

时间：1990-2018

知识来源：特定任务带标注数据

获取方法：机器学习（自动）

知识形式：以**任务专用小模型**
存储**任务知识**

通用智能

时间：2018至今

知识来源：通用域无标注数据

获取方法：预训练学习（自动）

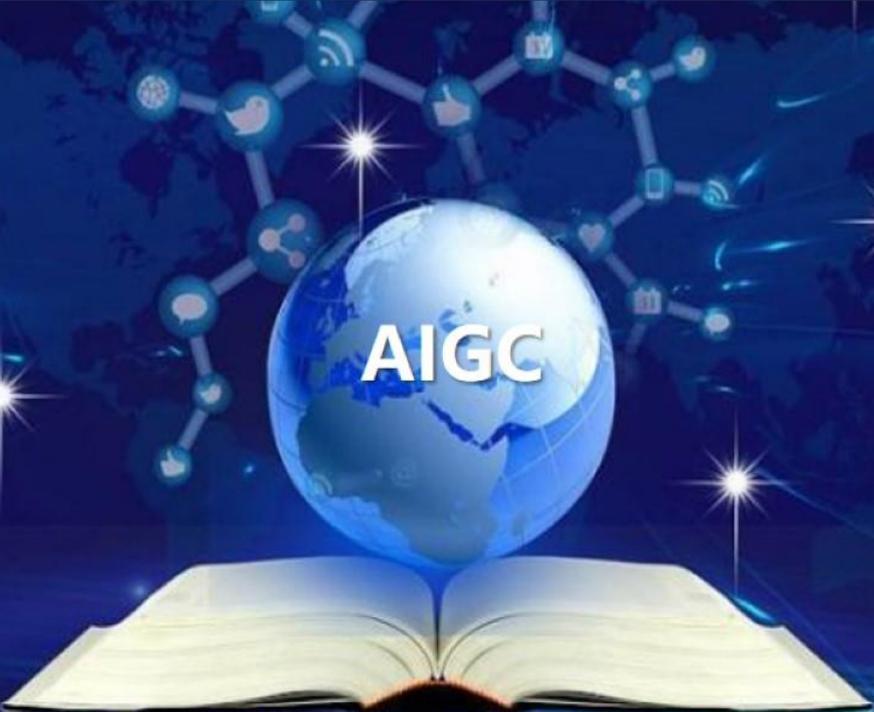
知识形式：以**通用大模型**存储
通用知识



究竟啥是 大语言模型

AI大模型科普课 ②

国内的百模大战



AIGC

科大讯飞（星火大模型）、百度（文心一言）、商汤科技（商量SenseChat）、智谱AI（GLM智谱清言）、华为（盘古大模型）、腾讯（混元大模型）、百川智能（百川大模型）、抖音（云雀大模型）、中科院（紫东太初）、MiniMax（ABAB大模型）、上海人工智能实验室（书生大模型）、360智脑、.....



大模型

GPT的全称，是Generative Pre-Trained Transformer（生成式预训练Transformer模型）是一种基于互联网的、可用数据来训练的、文本生成的深度学习模型。

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

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

Łukasz Kaiser*
Google Brain
lukaszkaizer@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

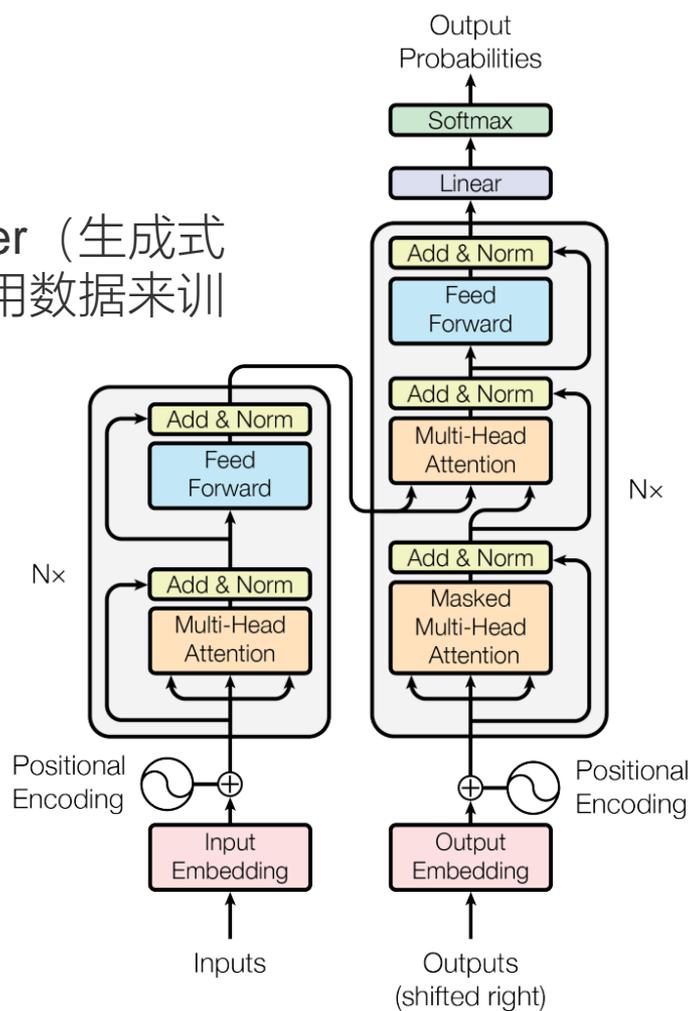
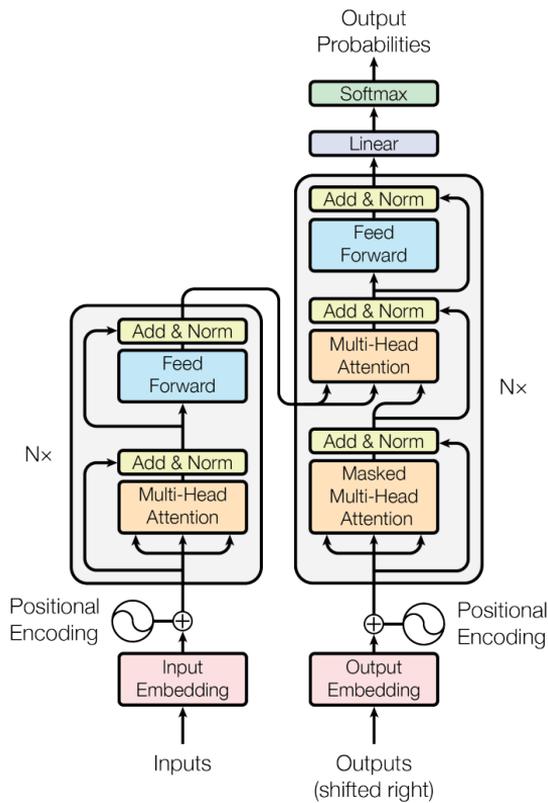


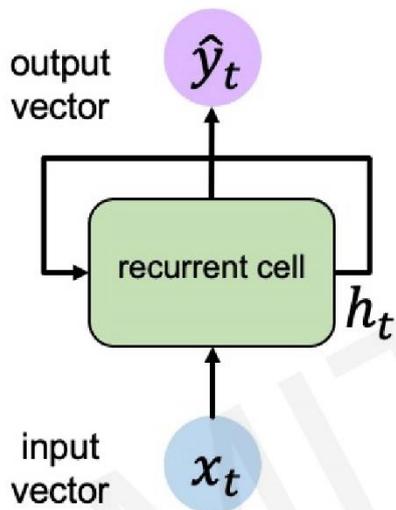
Figure 1: The Transformer - model architecture.



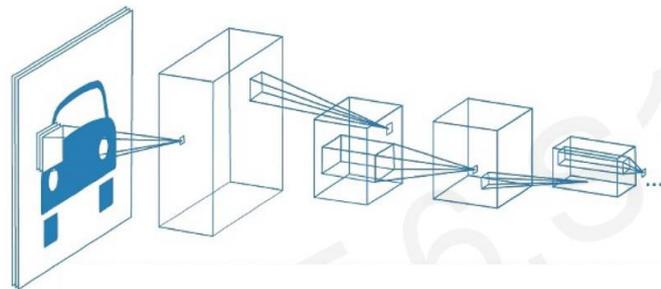
大模型



Transformer



RNN



CNN

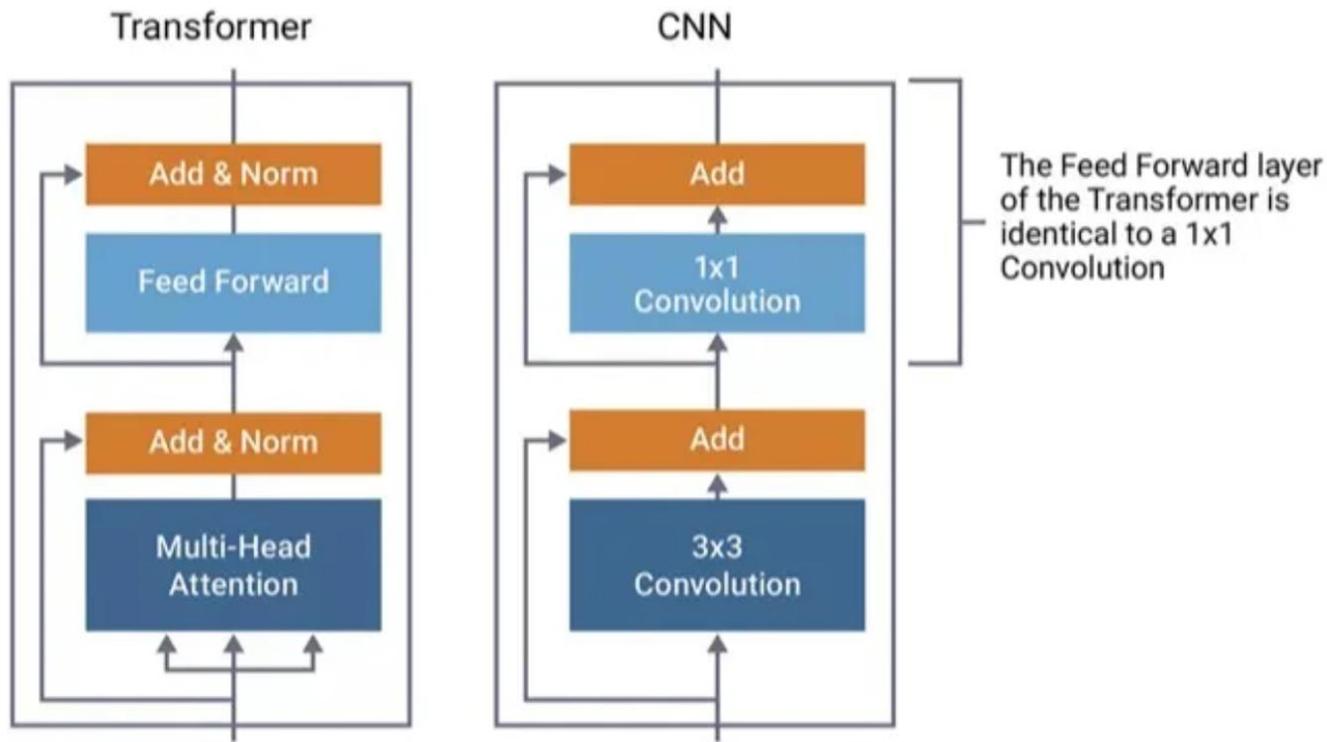
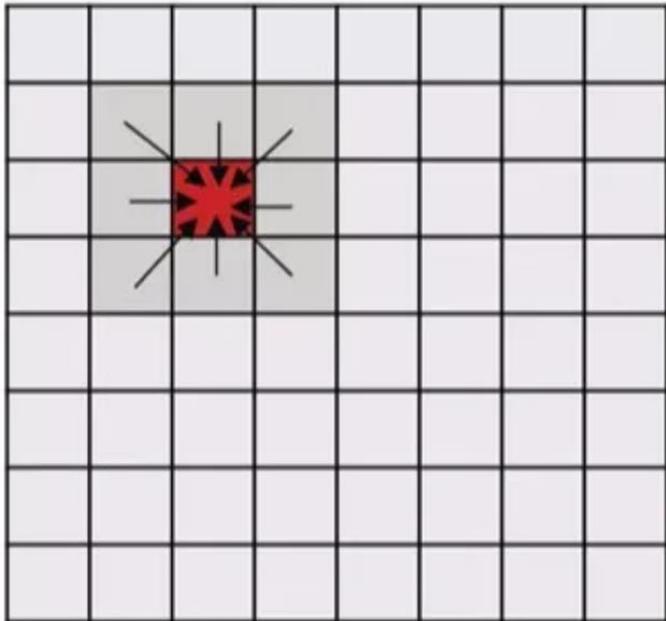


Fig. 2: Comparing Transformer and CNN structures.



大模型

Convolution



Attention

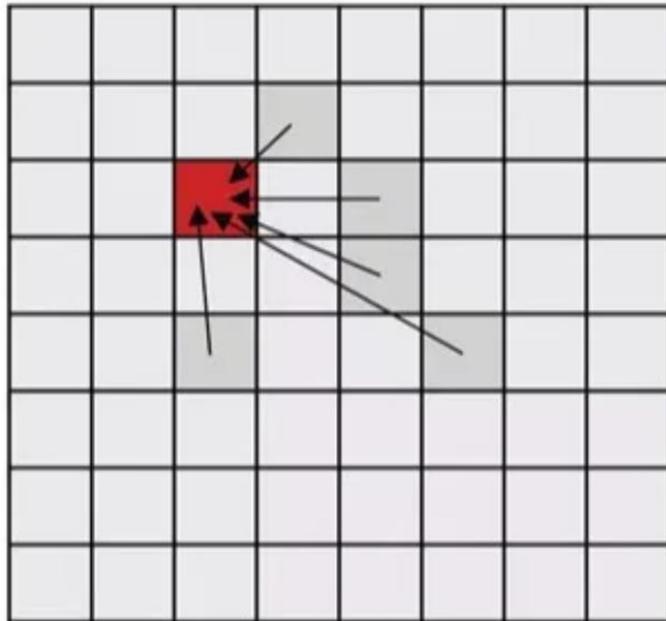


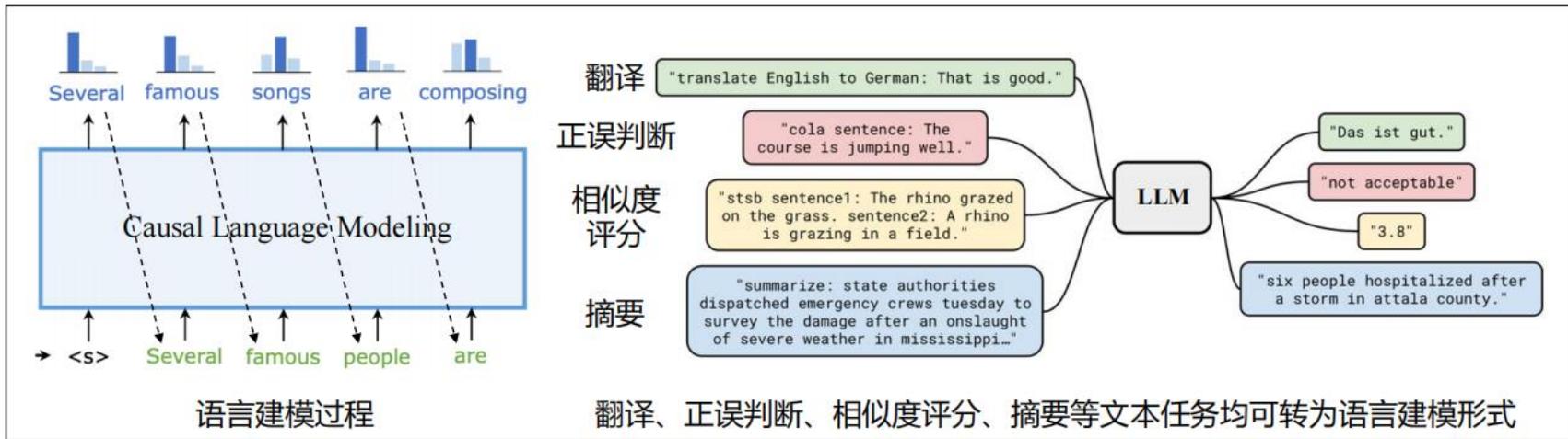
Fig. 3: Illustrating the difference between how a CNN's convolution and a transformer's attention networks mix in features of other tokens/pixels.



大模型

使用**预测下一个字符 (Token)** 的**语言建模任务 (Language Modeling)**，从无标注数据中学习通用知识

各类自然语言处理任务均可**转化为序列生成**任务进行处理





大模型

ChatGPT标志着人工智能进入新时代，加速引领新一轮科技革命

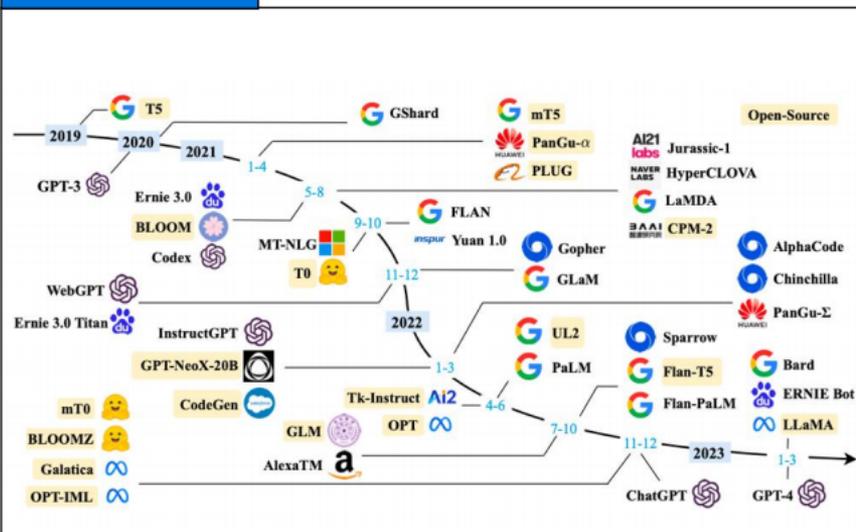
主要特点

基于大数据和人类反馈进行学习，与用户通过对话进行交互，能够处理多任务，深层推理与常识运用能力大幅度提升

里程碑事件

- 2022/11/30: OpenAI 发布文本对话模型 ChatGPT
- 2023/01/31: ChatGPT 月活用户突破 1亿
- 2023/02/07: 微软产品全线整合 ChatGPT
- 2023/03/15: OpenAI 发布图文对话模型 GPT-4
- 2023/05/15: OpenAI 发布人工智能助手 AutoGPT
- 2023/10/05: OpenAI 发布多模态模型 GPT-4V

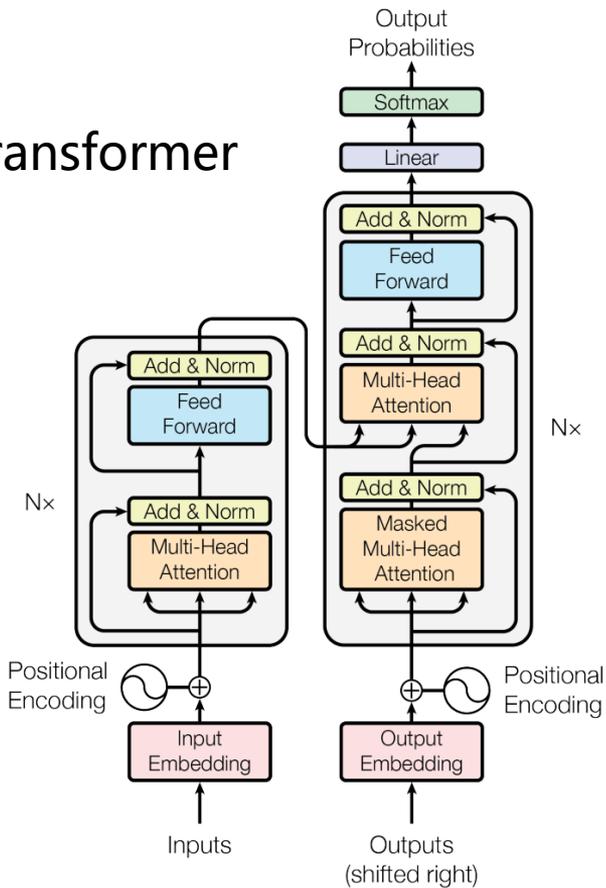
大模型发展历程



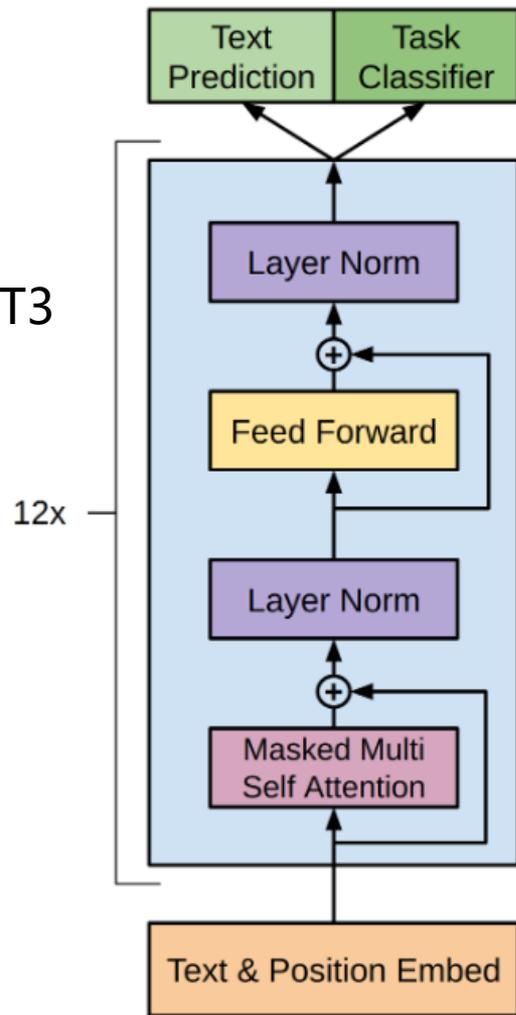


大模型

Transformer

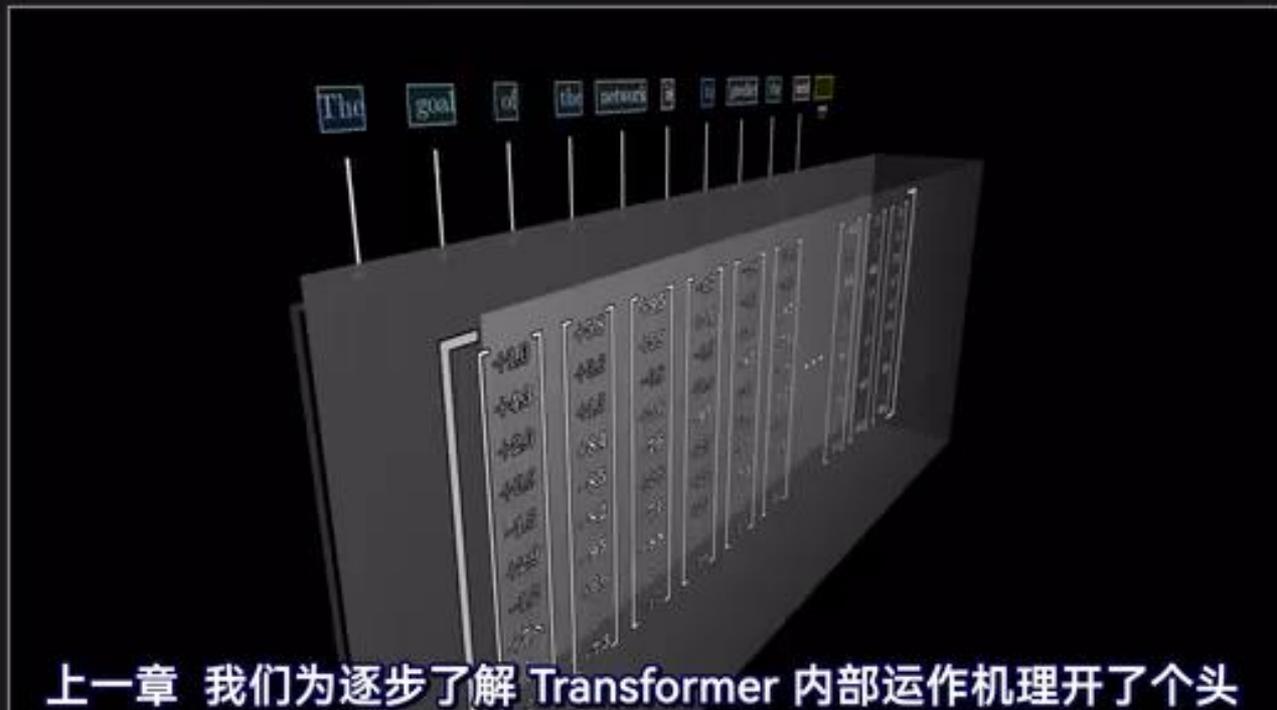


GPT3





Previously: Transformers



上一章 我们为逐步了解 Transformer 内部运作机理开了个头

In the last chapter, you and I started to step through the internal workings of a transformer.



大模型

GPT 的输入和输出是什么?

<s>	not	all	heroes	wear
0	1	2	3	4

Input Sequence

GPT

capes	90%
pants	5%
socks	2%
⋮	⋮

Output guess

输入是 N 个单词（也称为 Token，可译为“词元”）的序列。输出是对最有可能放在输入序列末尾单词的猜测。



大模型

所有令人印象深刻的 GPT 对话、故事和示例都是通过这种简单的输入输出方案完成的：给它一个输入序列——得到接下来的一个词。

- **Not all heroes wear -> capes**
- **Not all heroes wear capes -> but**
- **Not all heroes wear capes but -> all**
- **Not all heroes wear capes but all -> villains**
- **Not all heroes wear capes but all villains -> do**

得到下一个单词后，将其添加到输入序列中，再得到下一个单词。

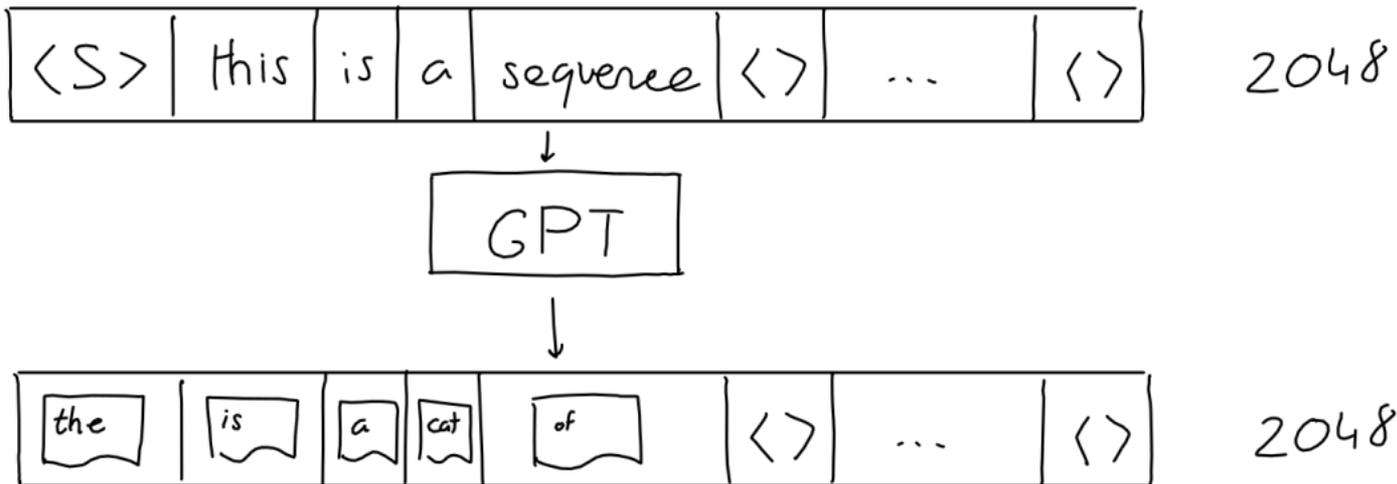
像这样一直重复，最终生成你需要的长文本。



大模型

确切来说还有两点需要纠正：

1. 输入序列实际上固定为 2048 个 Token（对于 GPT-3 来说）。仍然可以传递短序列作为输入：只需用“空”值填充其他额外的位置。
2. GPT 输出不仅是单个预测，而是一个多预测值（每个可能单词的概率）构成的序列（长度为 2048），每组预测值对应输入序列中的每个单词的“下一个”位置。但是在生成文本时，我们通常只查看序列中最后一个单词的预测。





大模型

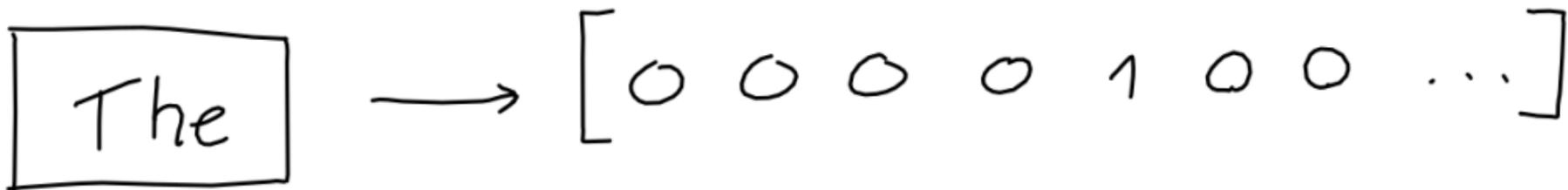
编码 (Encoding)

GPT 实际上并不能理解单词含义。作为一种机器学习算法，它是对数字向量进行运算的。那么如何将单词转换为向量呢？

第一步是将所有单词整理为一个词汇表，这使我们能够为每个单词赋予一个值。

(GPT 的词汇表包含 50257 个单词)

最后，我们可以将每个单词转换为 50257 长度的独热编码 (one-hot) 向量，其中仅索引 i 处的维 (单词的值) 为 1，其他维均为 0。

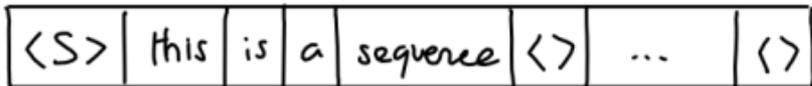


对序列中的每个单词都执行此操作

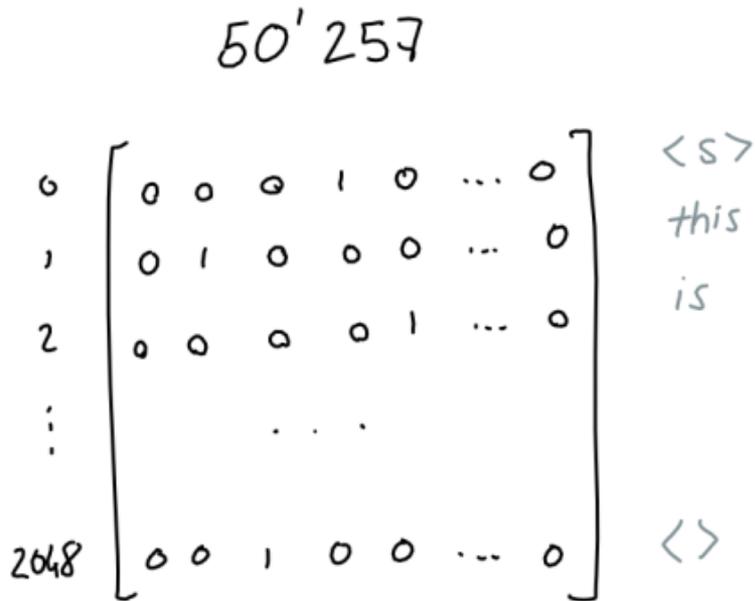


大模型

结果是一个 2048 x 50257 的 1/0 矩阵。



VOC



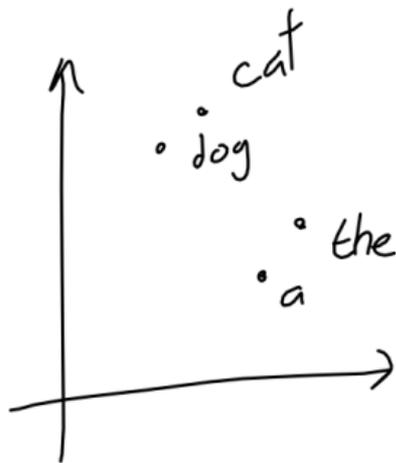
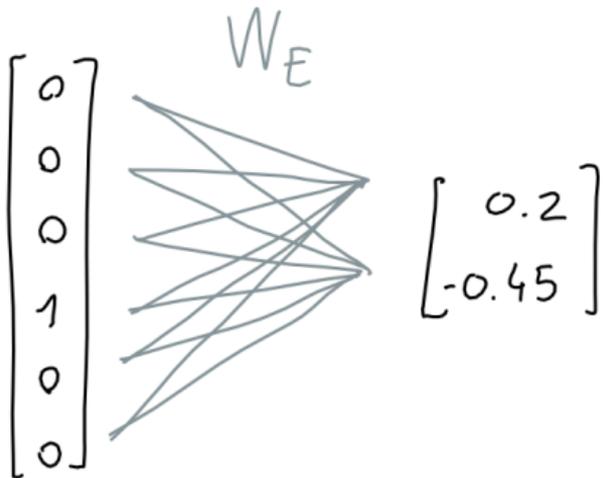


大模型

嵌入 (Embedding)

对于一个向量来说，50257 相当大了，并且其中大部分都用 0 填充，这样会浪费很多空间。为了解决这个问题，可以学习一个嵌入 (Embedding) 函数：一个神经网络，以 50257 长度的 $1/0$ 向量为输入，输出一个长度为 n 的数字向量。尝试将单词含义的信息存储（或投影）到较小的空间中。

例如，如果嵌入维数为 2，就类似于将每个单词存储在二维空间中的特定坐标处。

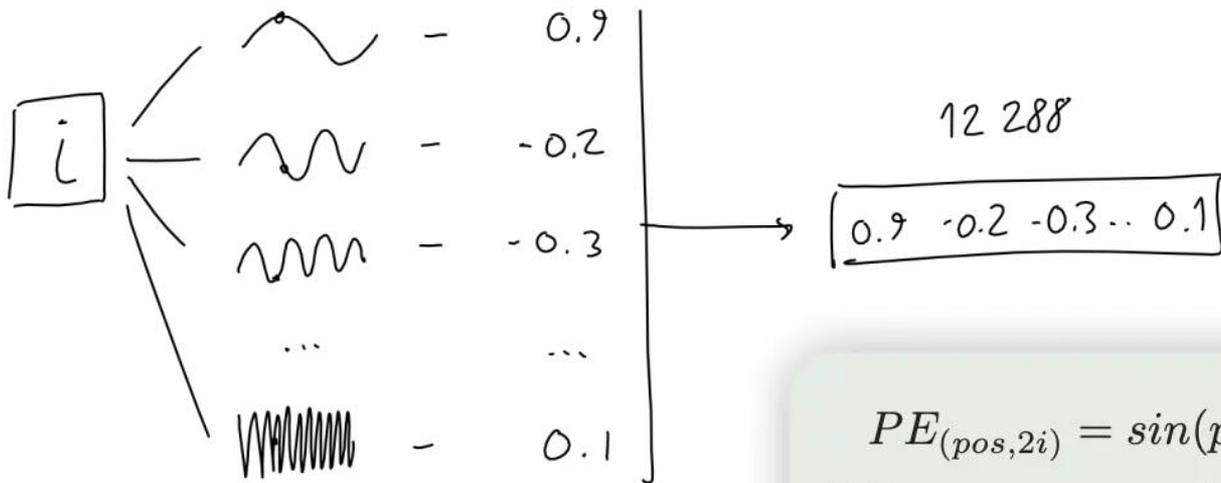




大模型

位置编码

为了对当前 Token 在序列中的位置进行编码，作者将 Token 的位置（标量 i ，取值范围 $[0-2047]$ ）作为参数传递给 12288 个频率不同的正弦函数。



$$PE_{(pos, 2i)} = \sin(pos/10000^{2i/d_{model}})$$

$$PE_{(pos, 2i+1)} = \cos(pos/10000^{2i/d_{model}})$$



大模型

这种做法为什么会有效？作者的解释是，生成很多相对位置编码，这对模型很有用。用其他可能的理论来分析这一选择：考虑到信号经常表示为周期性样本之和的方式（参见傅立叶变换），或者语言自然地呈现不同长度循环的可能性（例如诗歌）。

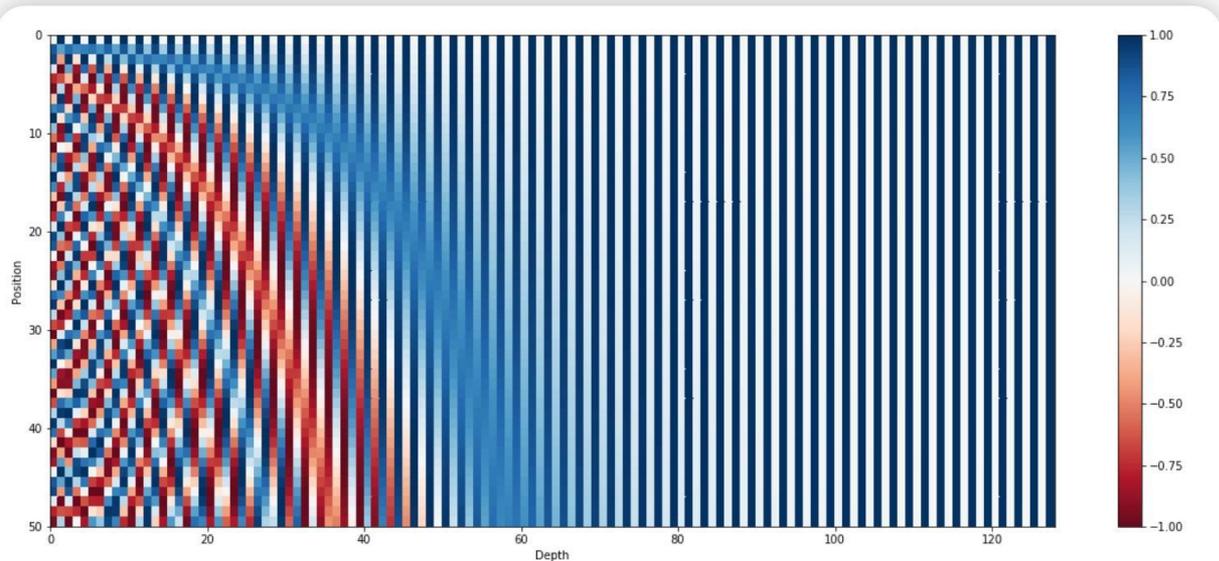
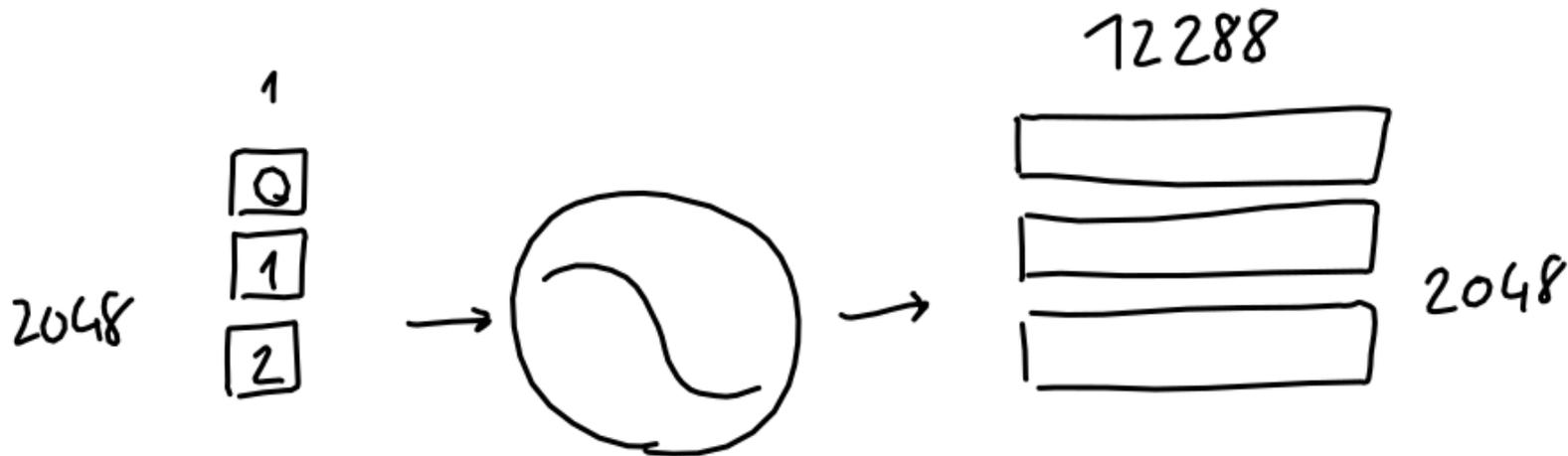


Figure 2 - The 128-dimensional positional encoding for a sentence with the maximum length of 50. Each row represents the embedding vector \vec{p}_t



大模型

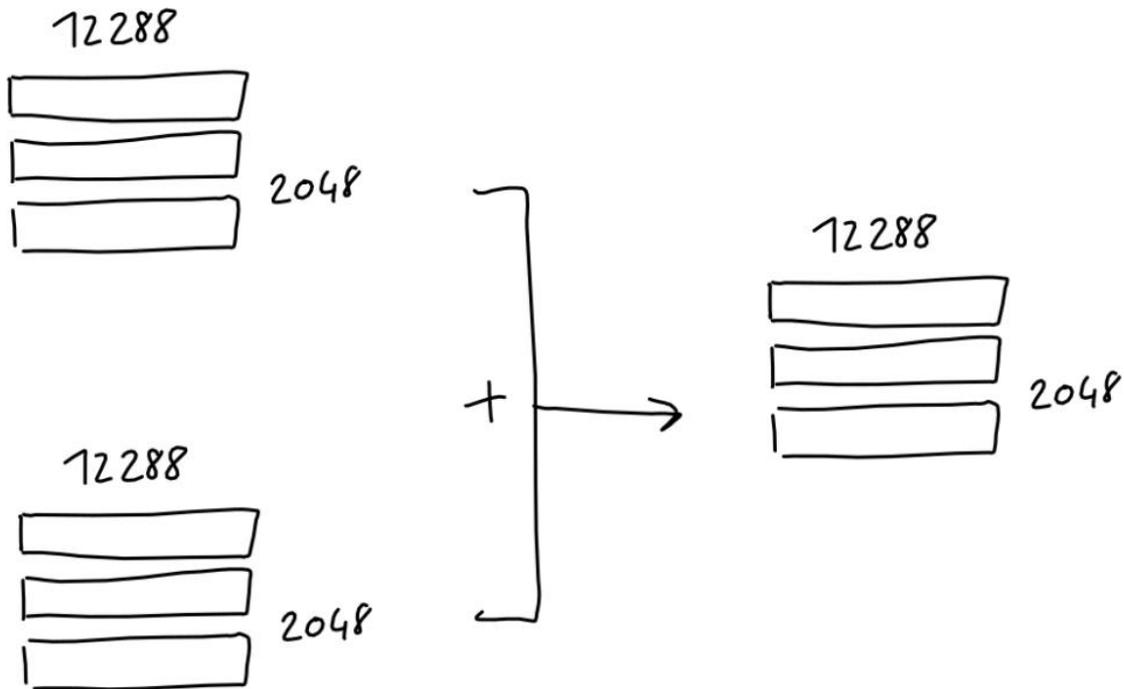
这一过程的结果是，每个 Token 对应一个 12288 维的数字向量。和嵌入操作一样，我们将这些向量组合成具有 2048 行的单一矩阵，其中每一行是序列中每个 Token 的 12288 列位置编码。





大模型

最后，与序列嵌入矩阵相同形状的序列位置编码矩阵可以直接添加到嵌入矩阵中。



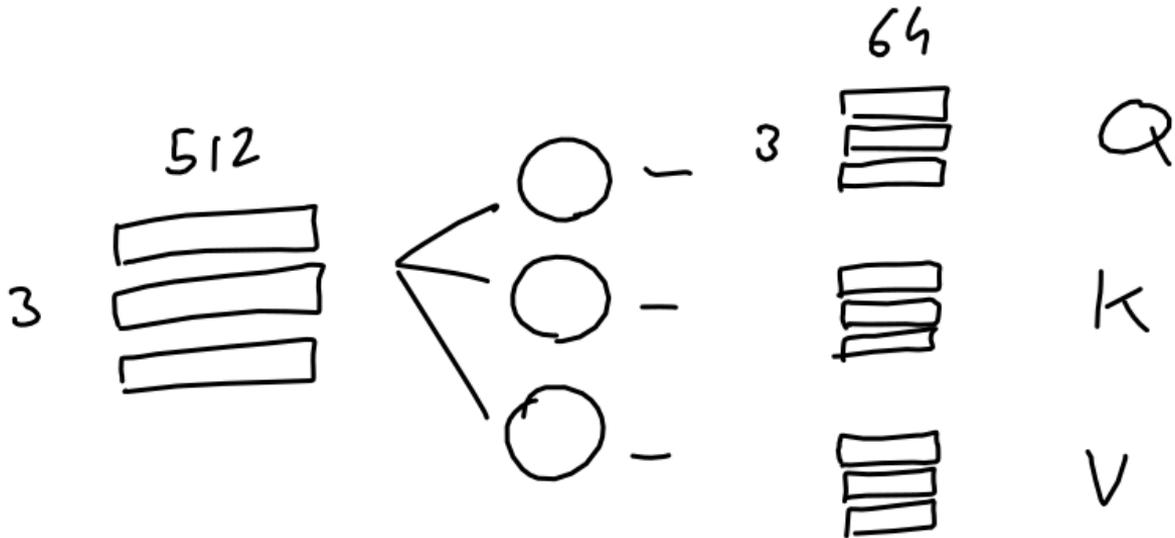


大模型

注意力 (简化版)

简单来说，注意力 (Attention) 机制的目的是：对于序列中的每个输出，预测需要关注的输入 Token 是哪些，以及有多少。这里，想象一个由 3 个 Token 组成的序列，每个 Token 都由 512 个值的嵌入表示。

该模型学习 3 个线性投影，所有这些投影都应用于序列嵌入。也可以说是学习了 3 个权重矩阵，这些矩阵将我们的序列嵌入转换为 3 个单独的 3×64 矩阵，每个矩阵分别用于不同的任务。





大模型

第三个矩阵（“值 Value”）与这个重要性矩阵相乘，从而为每个 Token 生成所有其他 Token 值的混合（按各个 Token 的重要性加权）。

$$\text{Softmax}(QK^T) \begin{bmatrix} \text{---} \\ \text{---} \\ \text{---} \end{bmatrix}_3 = 3 \begin{bmatrix} \text{---} \\ \text{---} \\ \text{---} \end{bmatrix}_{512}$$

(Note: The diagram uses colored lines to represent rows in the matrices. The result matrix has a '3' multiplier, indicating a 3x512 matrix.)

$$\begin{matrix} v \\ \text{---} \\ \text{---} \\ \text{---} \end{matrix} \times \begin{matrix} \text{---} \\ \text{---} \\ \text{---} \end{matrix}$$

(Note: The diagram uses colored boxes to represent rows in the matrices. The result matrix has a 'v' label above it.)

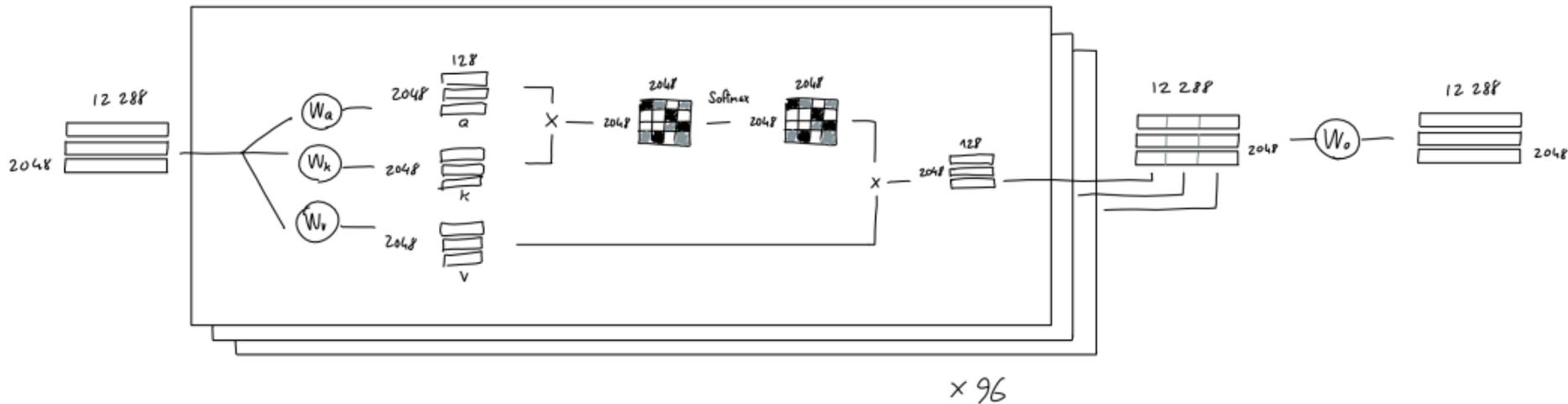


大模型

多头注意力 (Multi-Head Attention)

GPT 模型中使用了多头注意力。这仅仅意味着，上述过程被重复了很多次（GPT-3 中为 96 次），每次过程都有不同的学习查询、键、值投影权重。

每个注意力头的结果（单个 2048×128 矩阵）被拼接在一起，生成一个 2048×12288 矩阵，然后将其乘以一个线性投影（不会改变矩阵形状），以达到良好的度量。



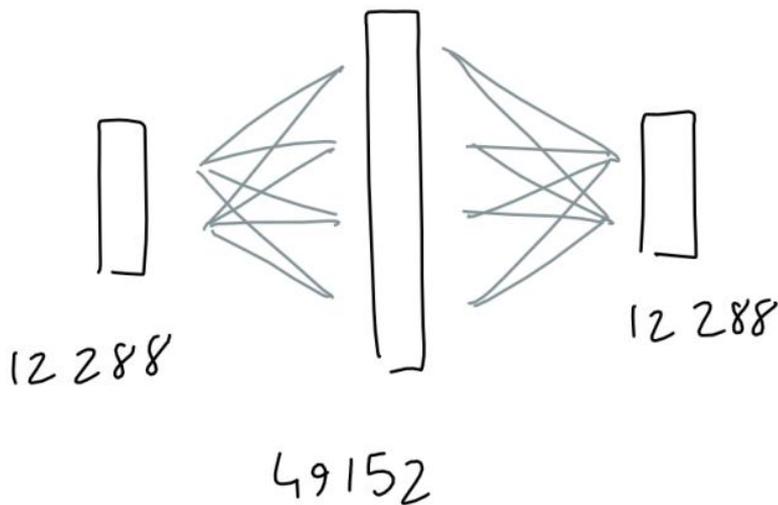
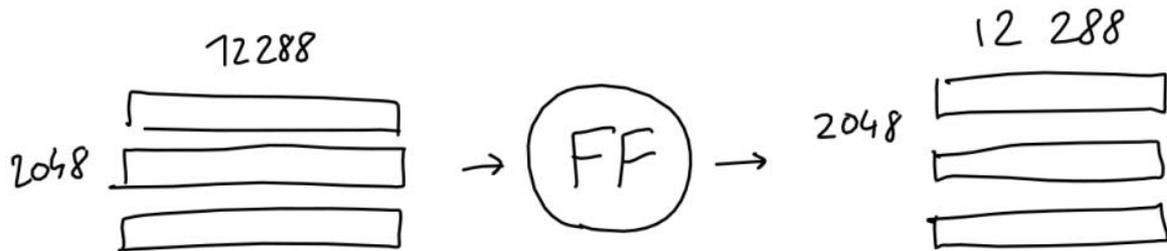


大模型

前馈 (Feed Forward)

前馈模块是我们熟知的多层感知器 (Multi-Layer Perceptron, MLP), 具有一个隐含层。获取输入, 乘以学习的权重, 添加学习的偏差, 重复该过程, 获得结果。

此处, 输入和输出形状都相同 (2048×12288), 但是隐藏层的大小为 4×12288 。

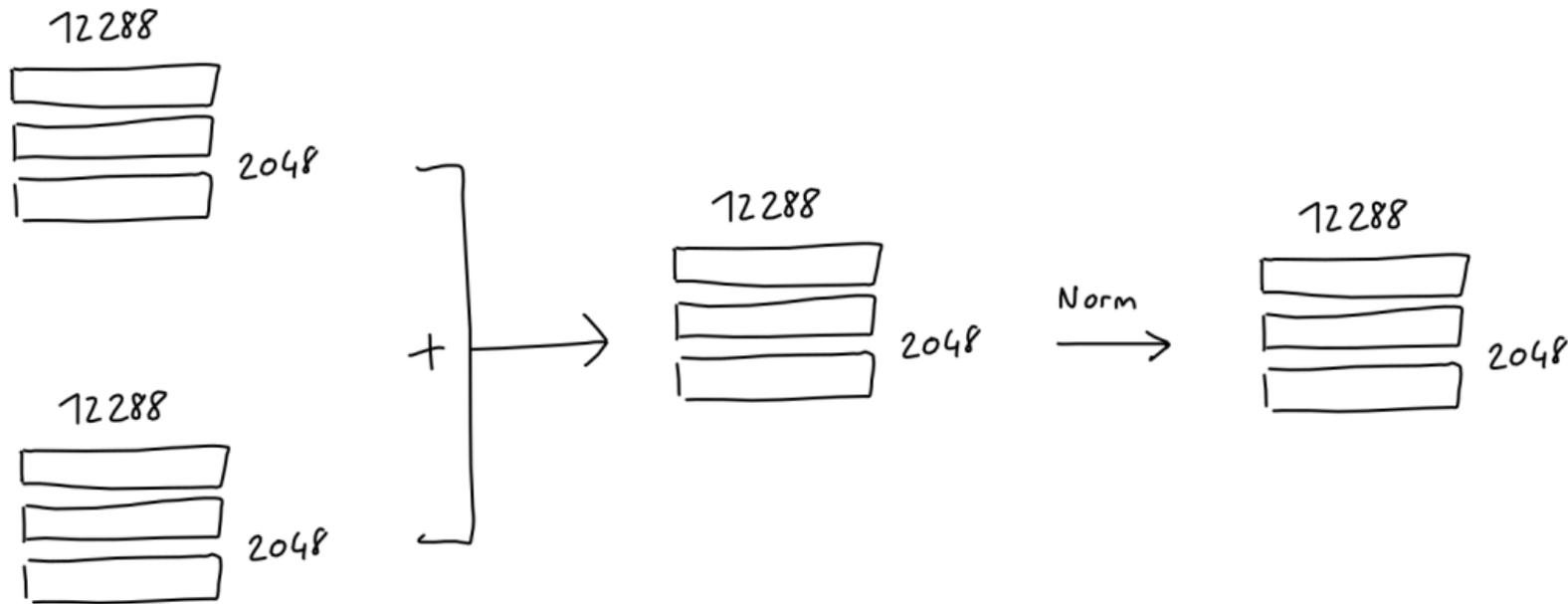




大模型

相加和归一化

在“多头注意力”和“前馈”模块之后，将模块的输入添加到输出中，然后对结果进行归一化。这在深度学习模型中很常见（自从 ResNet 之后）。



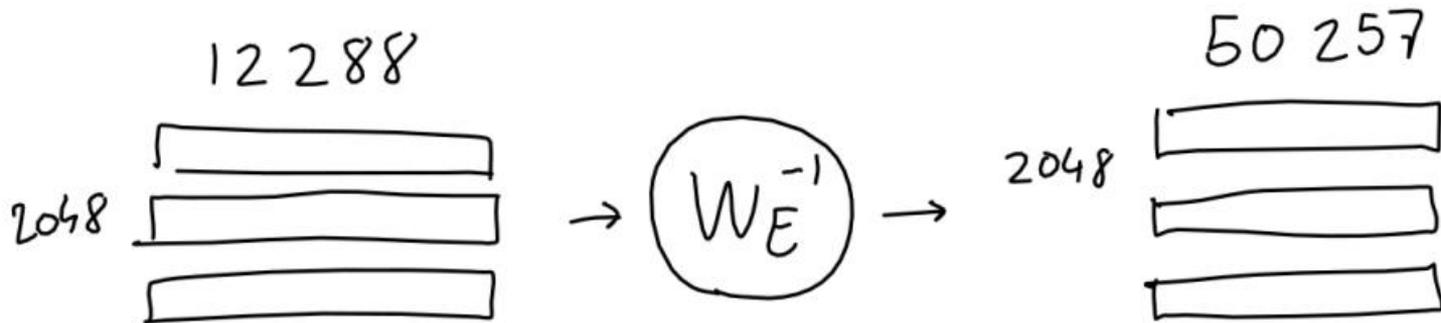


大模型

解码 (Decoding)

马上就要完成了！通过所有 96 层 GPT-3 的注意力/神经网络机制后，输入已处理为一个 2048×12288 矩阵。对于序列中 2048 个输出位置，该矩阵都应该对应包含一个 12288 维向量，其中包含了可能的单词信息。那么，要如何将这些信息提取出来呢？

回想“嵌入”部分，我们学习了一种映射，该映射将给定单词（的独热编码）转换为一个 12288 维向量嵌入。实际上，我们可以反转此映射，将输出的 12288 维向量嵌入转换回 50257 维单词编码。这一思路就是，既然已经花费了大量精力学习从单词到数字的良好映射，不妨重新利用！

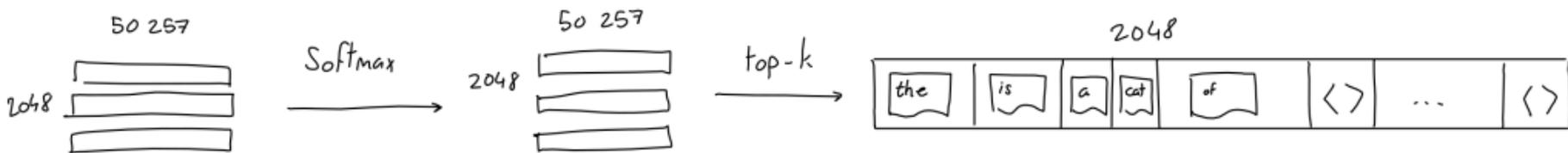




大模型

当然，这样操作不会得到与最开始相同的 1/0 值，但这是一件好事：在快速 softmax 之后，我们可以将结果值视为每个单词的概率。

此外，GPT 论文还提到了参数 top-k，该参数将输出中要采样的可能单词数量限制为 k 个最可能的值。例如，当 top-k 参数为 1 时，选择的的就是最有可能的单词。





大模型

新石器公园  CNBC

ChatGPT的工作原理是什么?



大模型

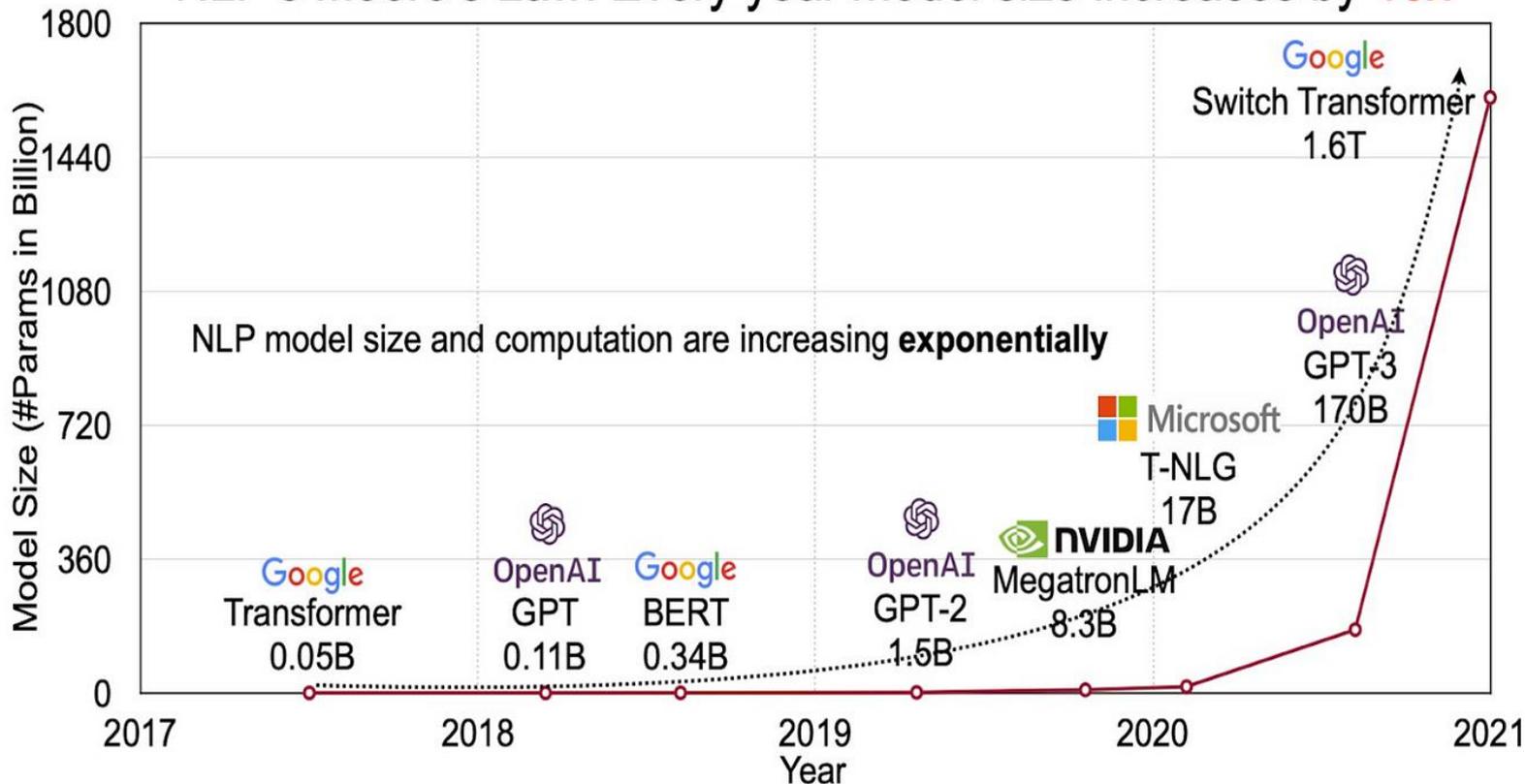
模型	年份	关键创新	训练集	参数数量	上下文窗口
GPT-1	2018	Transformer 解码器, 自监督训练 + 微调	4.5 GB	117M	512
GPT-2	2019	修改的归一化层,	40 GB	1.5B	1,024
GPT-3	2020	稀疏注意力层, 支持零样本	57 TB	175B	2,048 令牌
GPT-4	2023	多模态输入 (文本 + 图像)	?	1.76T	32,000 令牌



大模型

模型参数量指数增长

NLP's Moore's Law: Every year model size increases by 10x





大模型

智能涌现



Sam Altman

@sama

a new version of moore's law that could start soon:

the amount of intelligence in the universe doubles every 18 months

12:24 AM · Feb 27, 2023 · **3.8M** Views

1,923 Retweets **702** Quotes **14.8K** Likes



大模型训练成本

Optimal LLM Training Cost				
Model	Size (# Parameters)	Tokens	GPU	Optimal Training Compute Cost
MosaicML GPT-30B	30 Billion	610 Billion	A100	\$ 325,855
Google LaMDA	137 Billion	168 Billion	A100	\$ 368,846
Yandex YaLM	100 Billion	300 Billion	A100	\$ 480,769
Tsinghua University Zhipu.AI GLM	130 Billion	400 Billion	A100	\$ 833,333
Open AI GPT-3	175 Billion	300 Billion	A100	\$ 841,346
AI21 Jurassic	178 Billion	300 Billion	A100	\$ 855,769
Bloom	176 Billion	366 Billion	A100	\$ 1,033,756
DeepMind Gopher	280 Billion	300 Billion	A100	\$ 1,346,154
DeepMind Chinchilla	70 Billion	1,400 Billion	A100	\$ 1,745,014
MosaicML GPT-70B	70 Billion	1,400 Billion	A100	\$ 1,745,014
Nvidia Microsoft MT-NLG	530 Billion	270 Billion	A100	\$ 2,293,269
Google PaLM	540 Billion	780 Billion	A100	\$ 6,750,000



大模型

大模型训练流程图

四大步骤

预训练

监督式微调

奖励建模

强化学习

数据集

互联网公开数据集
2万亿token

问答对
1万-10万

人工反馈评价
10万-100万

人工提示词
1万-10万

算法

语言模型
预测下一个token

语言模型
预测下一个token

二元分类器
输出奖励

强化学习
最大化奖励

模型

基础模型

SFT模型

奖励模型

强化学习模型

所需资源

上千块GPU
训练数月

几十块GPU
训练几天

几十块GPU
训练几天

几十块GPU
训练几天

提纲

一、大模型

二、智能体



上海大学
SHANGHAI UNIVERSITY

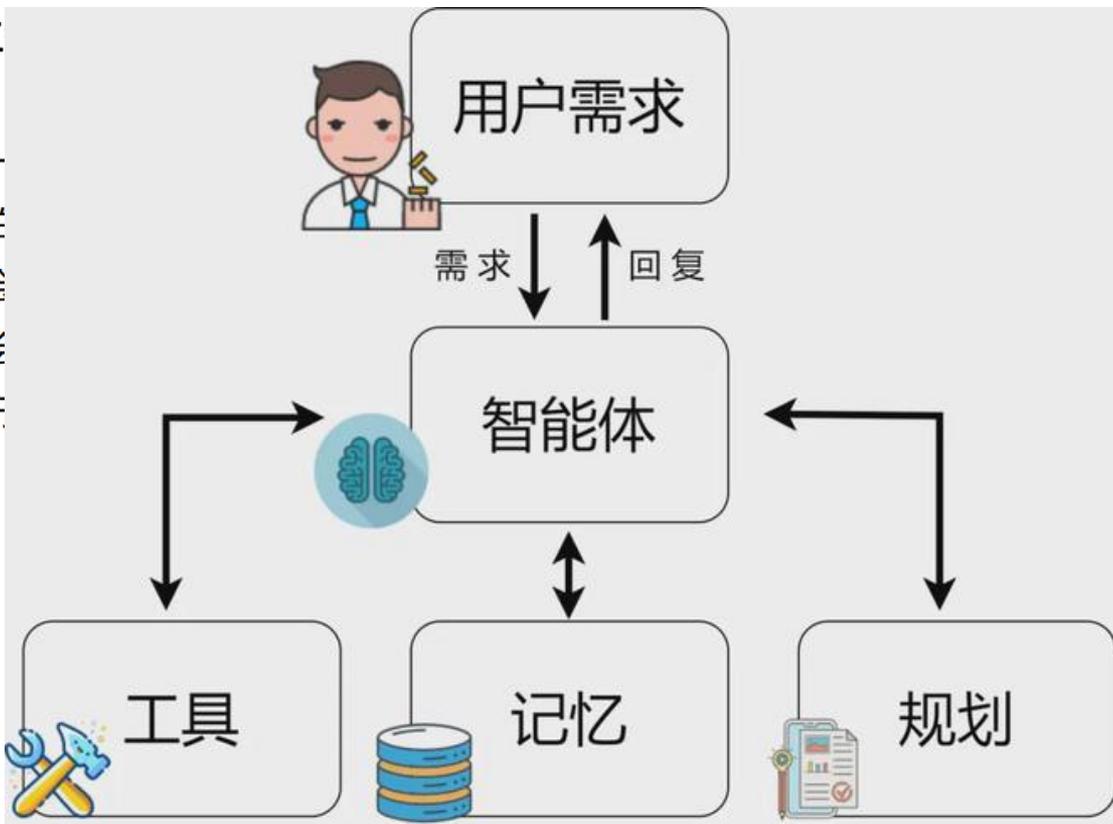




智能体

什么是 AI Agent?

AI Agent (人工
的人工智能, AI
Agent 帮忙下单-
类去指定每一步的
Minsky 认为社会
Agent 应具有社会
研究热点。但苦



不同于传统
，告诉 AI
；无需人
书中提出，
他还认为
迅速成为
丰。



智能体

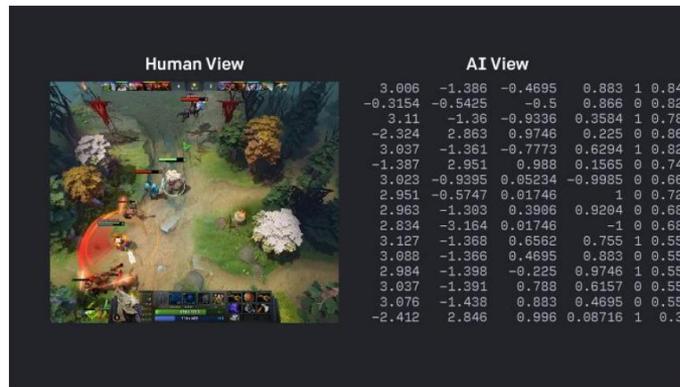
大语言模型和 AI Agent 的区别在于 AI Agent 可以独立思考并做出行动，和 RPA 的区别在于它能够处理未知环境信息。ChatGPT 诞生后，AI 从真正意义上具备了和人类进行多轮对话的能力，并且能针对相应问题给出具体回答与建议。随后各个领域的“Copilot”推出，如 Microsoft 365 Copilot、GitHub Copilot、Adobe Firefly 等，让 AI 成为了办公、代码、设计等场景的“智能副驾驶”。AI Agent 和大模型的区别在于，大模型与人类之间的交互是基于 prompt 实现的，用户 prompt 是否清晰明确会影响大模型回答的效果，例如 ChatGPT 和这些 Copilot 都需要明确任务才能得到有用的回答。而 AI Agent 的工作仅需给定一个目标，它就能够针对目标独立思考并做出行动，它会根据给定任务详细拆解出每一步的计划步骤，依靠来自外界的反馈和自主思考，自己给自己创建 prompt，来实现目标。如果说 Copilot 是“副驾驶”，那么 Agent 则可以算得上一个初级的“主驾驶”。和传统的 RPA 相比，RPA 只能在给定的情况条件下，根据程序内预设好的流程来进行工作的处理，在出现大量未知信息、难以预测的环境中时，RPA 是无法进行工作的，AI Agent 则可以通过和环境进行交互，感知信息并做出对应的思考和行动。



智能体

Agent 的最终发展目标：通用人工智能 AGI

AI Agent 并不是一个新兴的概念，早在多年前就已在人工智能领域有了研究。例如 2014 年由 DeepMind 推出的引发全球热议的围棋机器人 AlphaGo，也可以看做是 AI Agent 的一种。与之类似的还有 2017 年 OpenAI 推出的用于玩《Dota2》的 OpenAI Five，2019 年 DeepMind 公布用于玩《星际争霸 2》的 AlphaStar 等，这些 AI 都能根据对实时接收到的信息的分析来安排和规划下一步的操作，均满足 AI Agent 的基本定义。当时的业界潮流是通过强化学习的方法来对 AI Agent 进行训练，主要应用场景是在游戏这类具有对抗性、有明显输赢双方的场景中。但如果想要在真实世界中实现通用性，基于当时的技术水平还难以实现。





智能体

大语言模型的浪潮推动了 AI Agent 相关研究快速发展。 AI Agent 需要做到能够像人类一样进行交互，大语言模型强大的能力为 AI Agent 的突破带来了契机。大模型庞大的训练数据集中包含了大量人类行为数据，为模拟类人的交互打下了坚实基础；另一方面，随着模型规模不断增大，大模型涌现出了上下文学习能力、推理能力、思维链等类似人类思考方式的多种能力。将大模型作为 AI Agent 的核心大脑，就可以实现以往难以实现的将复杂问题拆解成可实现的子任务、类人的自然语言交互等能力。大模型的快速发展大幅推动了 AI Agent 的发展。

通往 AGI 的道路仍需探索，AI Agent 是当前的主要路线。在大模型浪潮席卷全球之时，很多人认为大模型距离真正的通用人工智能 AGI 已经非常接近，很多厂商都投入了基础大模型的研究。但经过了一段时间后，大家对大模型真实的能力边界有了清晰的认知，发现大模型仍存在大量的问题如幻觉、上下文容量限制等，导致其无法直接通向 AGI，于是 AI Agent 成为了新的研究方向。通过让大模型借助一个或多个 Agent 的能力，构建成为具备自主思考决策和执行能力的智能体，来继续实现通往 AGI 的道路。OpenAI 联合创始人 Andrej Karpathy 在一次开发者活动中讲到，OpenAI 内部对 AI Agents 非常感兴趣，AI Agent 将是未来 AI 的前沿方向。扎克伯格也在 Meta 的一季度财报电话会上提到，Meta 将会把 AI Agents 介绍给数十亿用户。

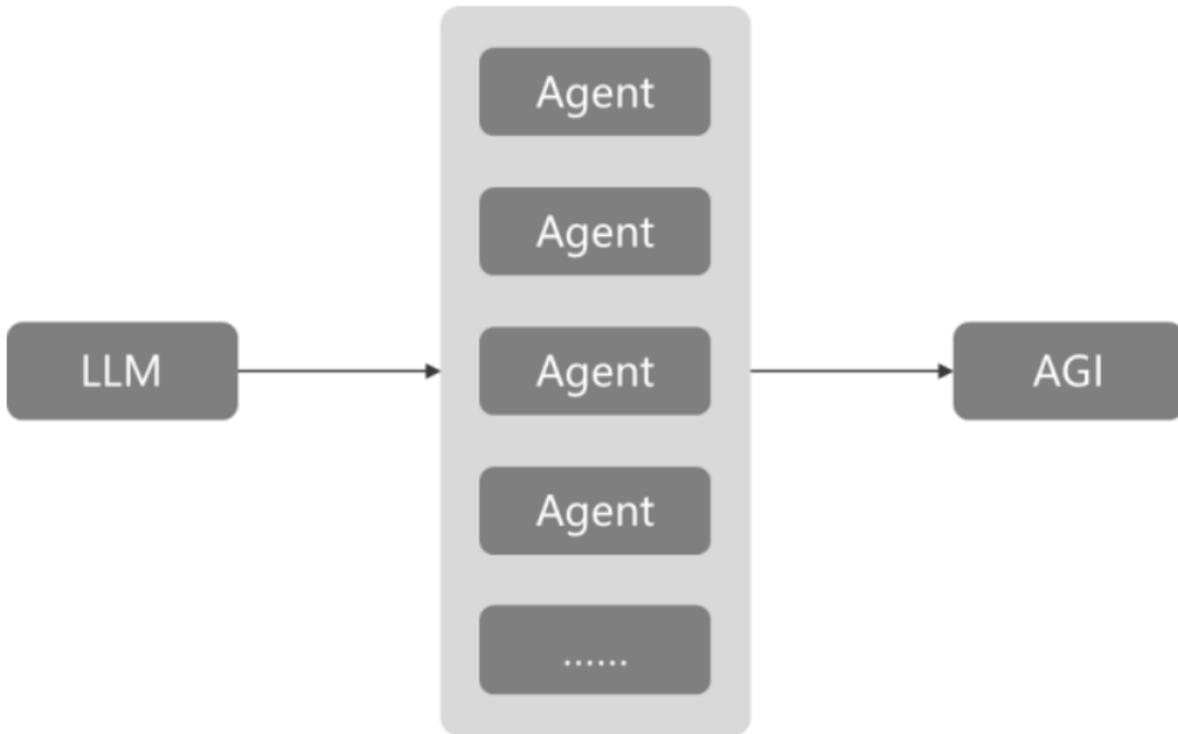


5分钟大模型 思维链

00:00:00



研究 AI Agent 的最终目标是通向 AGI





智能体

将 AI 和人类协作的程度类比自动驾驶的不同阶段

AI 等级 (类比自动驾驶)	名称	特点	示例
L1	Tool	人类完成所有工作，没有任何显性的 AI 辅助	目前绝大多数软件产品
L2	Chatbot	人类完成绝大部分工作。人类向 AI 询问意见，了解信息，AI 提供信息和建议但不直接处理工作	初代 ChatGPT 等 Chatbot
L3	Copilot	人类和 AI 进行协作，工作量相当。AI 根据人类 prompt 完成工作初稿，人类进行目标设定、修改调整，最后确认	GitHub Copilot、Midjourney、Jasper 等
L4	Agent	AI 完成绝大部分工作，人类负责设定目标、提供资源和监督结果。AI 完成任务拆分，工具选择，进度控制，实现目标后自主结束工作	AutoGPT 等
L5	Species	完全无需人类监督，AI 自主拆解目标、寻找资源、选择并使用工具、完成全部工作，人类只需给出目标	机器人？

AI Agent 可以类比为人类的协作程度可以和自动驾驶，人类可以向 AI 别的自动驾驶，人类而 Agent 则进一步升但就如同 L4 级别的 Agents 的真正应用还

报告，AI 与人 L2 级别自动类比为 L3 级进行修改调整；工具调用等。能以实现，AI

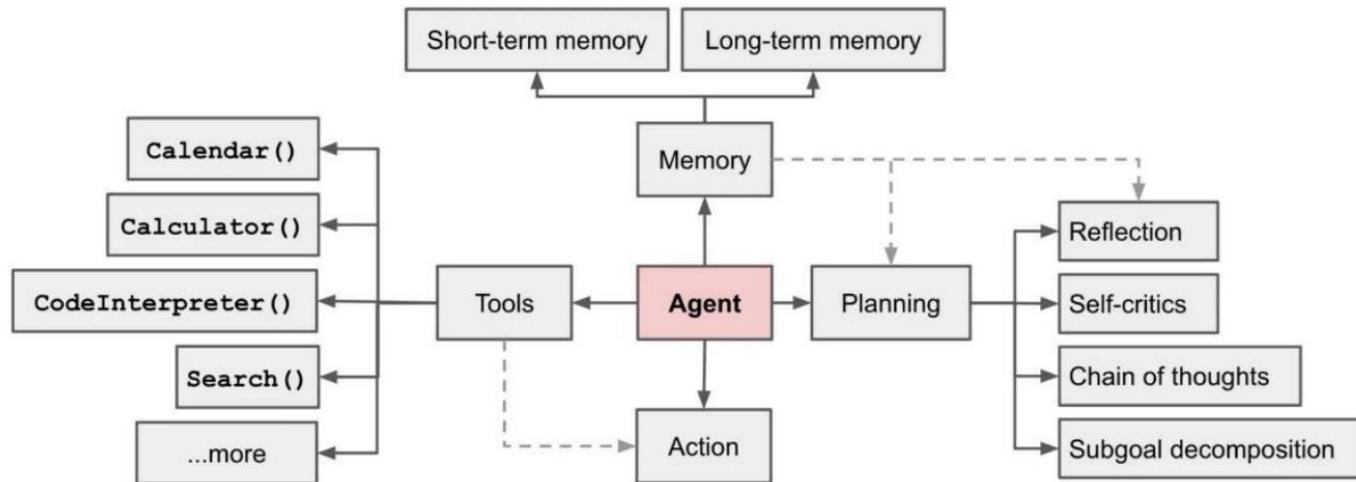


智能体

AI Agent 拆解：大模型、规划、记忆与工具

一个基于大模型的 AI Agent 系统可以拆分为大模型、规划、记忆与工具使用四个组件部分。6 月，OpenAI 的应用研究主管 Lilian Weng 撰写了一篇博客，认为 AI Agent 可能会成为新时代的开端。她提出了 Agent = LLM + 规划技能 + 记忆 + 工具使用的基础架构，其中 LLM 扮演了 Agent 的“大脑”，在这个系统中提供推理、规划等能力。

图 8：由 LLM 驱动的自主智能体系统的架构

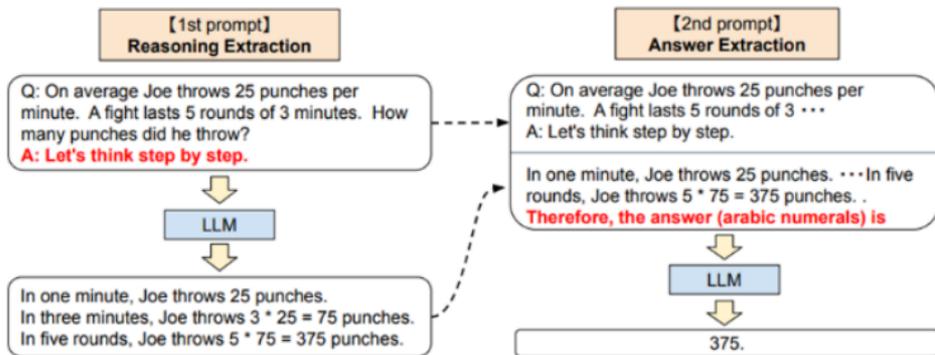




通过调整 prompt 可以提升大模型推理效果

大模型+推理

LLM 具备逻辑推理能力，LLM 本身在解决逻辑推理问题上，LLM 的推理能力因在于 prompt 的优化，大幅提升 LLM 的推理准确率。在向 LLM 提供合适的 prompt 时，



Zero-shot CoT: 例子

	Arithmetic					
	SingleEq	AddSub	MultiArith	GSM8K	AQUA	SVAMP
zero-shot	74.6/78.7	72.2/77.0	17.7/22.7	10.4/12.5	22.4/22.4	58.8/58.7
zero-shot-cot	78.0/78.7	69.6/74.7	78.7/79.3	40.7/40.5	33.5/31.9	62.1/63.7
	Common Sense		Other Reasoning Tasks		Symbolic Reasoning	
	Common SenseQA	Strategy QA	Date Understand	Shuffled Objects	Last Letter (4 words)	Coin Flip (4 times)
zero-shot	68.8/72.6	12.7/54.3	49.3/33.6	31.3/29.7	0.2/-	12.8/53.8
zero-shot-cot	64.6/64.0	54.8/52.3	67.5/61.8	52.4/52.9	57.6/-	91.4/87.8

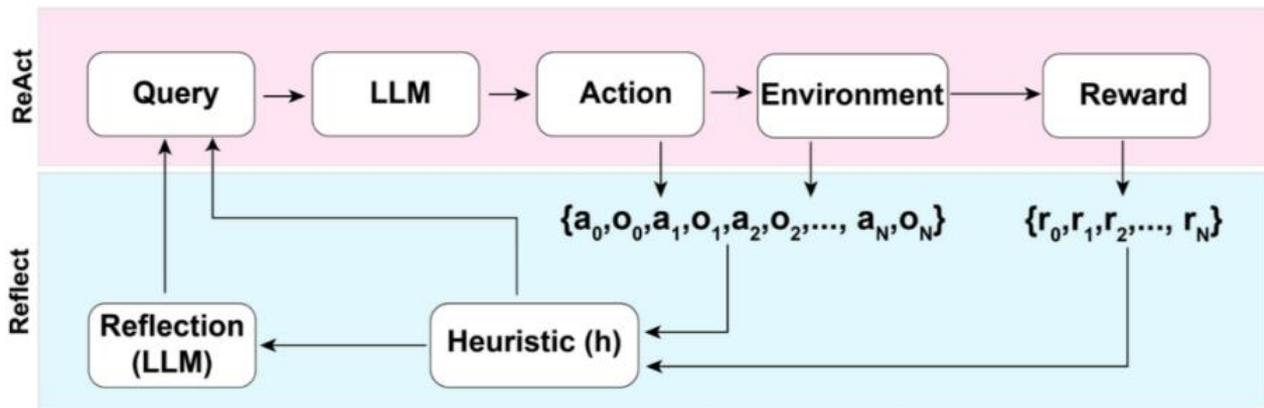
Zero-shot CoT: 效果明显提升



智能体

对于需要更多步骤的复杂任务，Agent 能够调用 LLM 通过思维链能力实现任务分解与规划。在 AI Agent 的架构中，任务分解规划的过程是基于大模型的能力来实现的。大模型具备思维链（Chain of Thoughts, CoT）能力，通过提示模型“逐步思考”，利用更多的计算时间来将困难任务分解为更小，更简单的步骤，降低每个子任务的规模。

通过反思与自省框架，Agents 可以不断提升任务规划能力。AI Agent 可以对过去的行为进行自我批评和反思，从错误中学习，并为未来的步骤进行完善，从而提高最终结果的质量。自省框架使 Agents 能够修正以往的决策、纠正之前的失误，从而不断优化其性能。在实际任务执行中，尝试和错误是常态，反思和自省两个框架在这个过程中起到了核心作用。





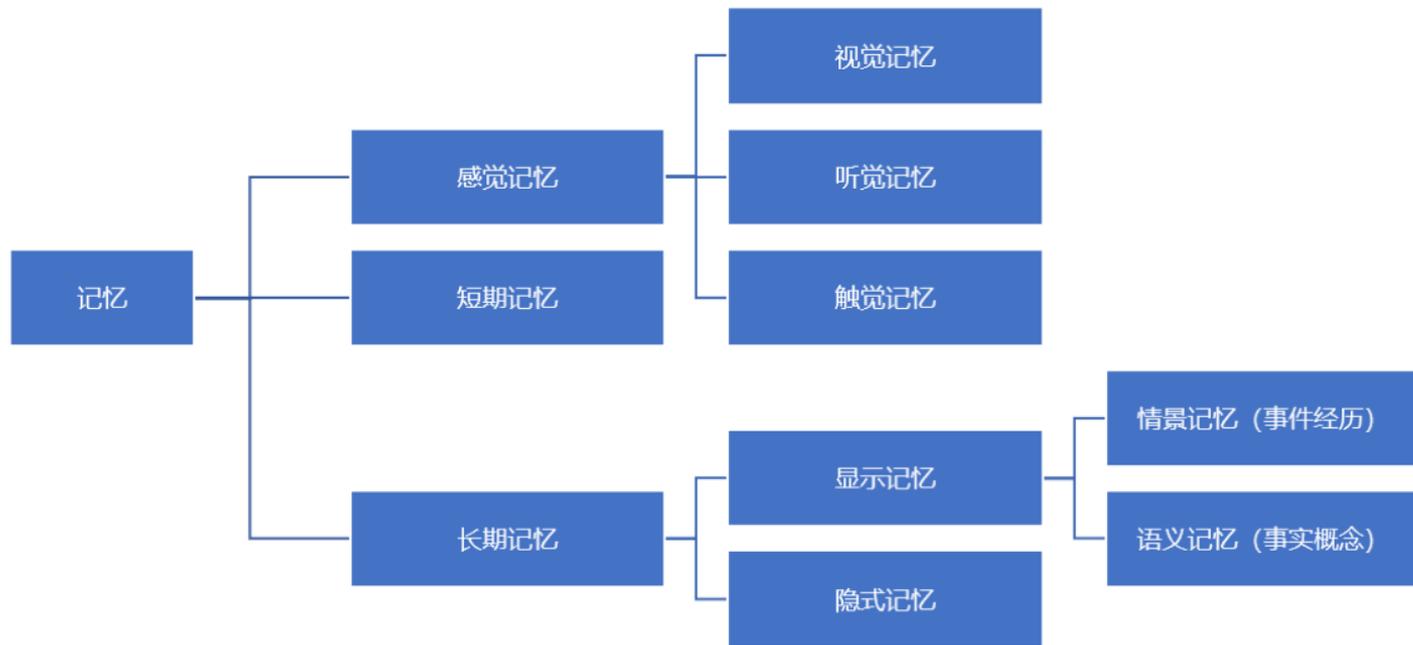
记忆：用有限的上下文长度实现更多的记忆

对 AI 智能体系统的输入会成为系统的记忆，与人类的记忆模式可实现一一映射。记忆可以定义为用于获取、存储、保留以及随后检索信息的过程。人脑中有多种记忆类型，如感觉记忆、短期记忆和长期记忆。而对于 AI Agent 系统而言，用户在其交互过程中产生的内容都可以认为是 Agent 的记忆，和人类记忆的模式能够产生对应关系。感觉记忆就是作为学习嵌入表示的原始输入，包括文本、图像或其他模态；短期记忆就是上下文，受到有限的上下文窗口长度的限制；长期记忆则可以认为是 Agent 在工作时需要查询的外部向量数据库，可通过快速检索进行访问。目前 Agent 主要是利用外部的长期记忆，来完成很多的复杂任务，比如阅读 PDF、联网搜索实时新闻等。任务与结果会储存在记忆模块中，当信息被调用时，储存在记忆中的信息会回到与用户的对话中，由此创造出更加紧密的上下文环境。



智能体

人类记忆的分类





人类记忆与 AI Agent 记忆的映射

记忆类型	映射	例子
感觉记忆	学习原始输入的嵌入表示，包括文本、图像或其他形式，短暂保留感觉印象。	看一张图片，然后在图片消失后能够在脑海中回想起它的视觉印象。
短期记忆	上下文学习（比如直接写入 prompt 中的信息），处理复杂任务的临时存储空间，受有限的上下文长度限制。	在进行心算时记住几个数字，但短期记忆是有限的，只能暂时保持几个项目。
长期记忆	在查询时 Agent 可以关注的外部向量存储，具有快速检索和基本无限的存储容量。	学会骑自行车后，多年后再次骑起来时仍能掌握这项技能，这要归功于长期记忆的持久存储。



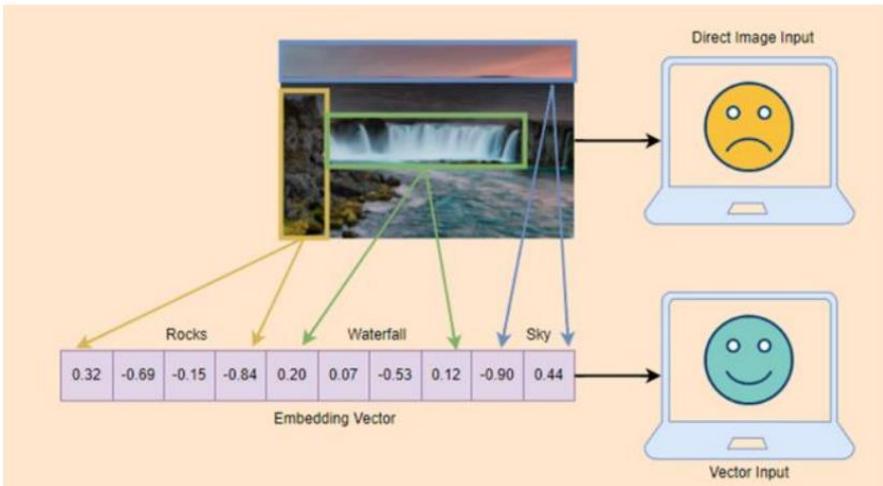
智能体

向量数据库通过将数据转化为向量存储，解决大模型海量知识的存储、检索、匹配问题。向量是 AI 理解世界的通用数据形式，大模型需要大量的数据进行训练，以获取丰富的语义和上下文信息，导致了数据量的指数级增长。向量数据库利用人工智能中的 Embedding 方法，将图像、音视频等非结构化数据抽象、转换为多维向量，由此可以结构化地在向量数据库中进行管理，从而实现快速、高效的数据存储和检索过程，赋予了 Agent “长期记忆”。同时，将高维空间中的多模态数据映射到低维空间的向量，也能大幅降低存储和计算的成本，向量数据库的存储成本比存到神经网络的成本要低 2 到 4 个数量级。

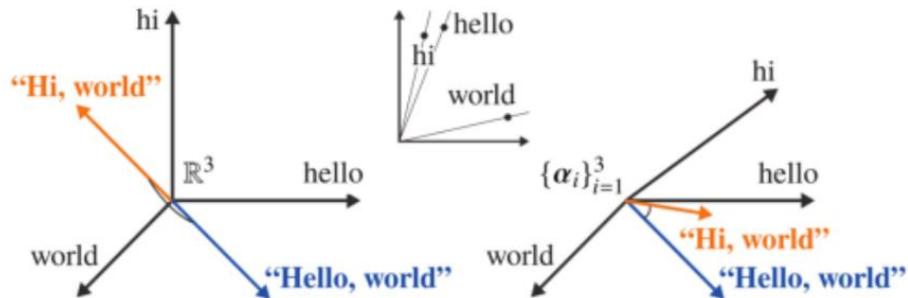
Embedding 技术和向量相似度计算是向量数据库的核心。 Embedding 技术是一种将图像、音视频等非结构化数据转化为计算机能够识别的语言的方法，例如常见的地图就是对于现实地理的 Embedding，现实的地理地形的信息其实远远超过三维，但是地图通过颜色和等高线等来最大化表现现实的地理信息。在通过 Embedding 技术将非结构化数据例如文本数据转化为向量后，就可以通过数学方法来计算两个向量之间的相似度，即可实现对文本的比较。向量数据库强大的检索功能就是基于向量相似度计算而达成的，通过相似性检索特性，针对相似的问题找出近似匹配的结果，是一种模糊匹配的检索，没有标准的准确答案，进而更高效地支撑更广泛的应用场景。



非结构化数据的向量化表征



不同文本在向量空间中的相似度计算





智能体

工具：懂得使用工具才会更像人类

AI Agent 与大模型的一大区别在于能够使用外部工具拓展模型能力。懂得使用工具是人类最显著和最独特的地方，同样地，我们也可以为大模型配备外部工具来让模型完成原本无法完成的工作。ChatGPT 的一大缺点在于，其训练数据只截止到了 2021 年底，对于更新一些的知识内容它无法直接做出回答。虽然后续 OpenAI 为 ChatGPT 更新了插件功能，能够调用浏览器插件来访问最新的信息，但是需要用户来针对问题指定是否需要使用插件，无法做到完全自然的回答。AI Agent 则具备了自主调用工具的能力，在获取到每一步子任务的工作后，Agent 都会判断是否需要通过调用外部工具来完成该子任务，并在完成后获取该外部工具返回的信息提供给 LLM，进行下一步子任务的工作。OpenAI 也在 6 月为 GPT-4 和 GPT-3.5 更新了函数调用的功能，开发者现在可以向这两个大模型描述函数，并让模型智能地选择输出包含调用这些函数的参数的 JSON 对象。这是一种更可靠地将 GPT 的功能与外部工具和 API 相连的新方法，允许开发者更可靠地从模型中获得结构化的数据，为 AI 开发者提供了方便。



函数调用示例



波士顿 现在的天气怎么样?

步骤 1 · OpenAI API

使用函数和用户的输入调用模型



步骤 2 · 第三方接口

使用模型响应调用 API



步骤 3 · OpenAI API

将响应发送回模型进行汇总



波士顿的天气目前晴朗，气温为22摄氏度。



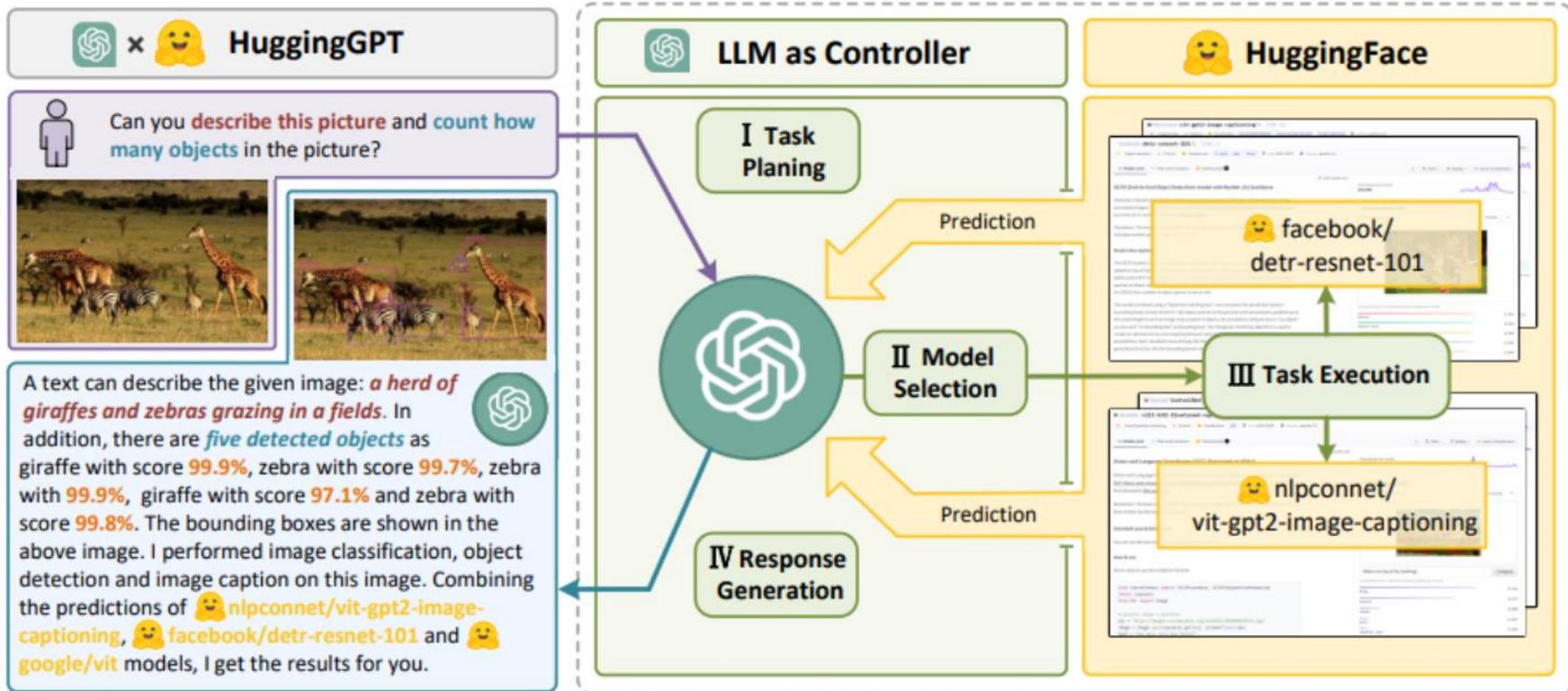
以 HuggingGPT 为例，HuggingGPT 将模型社区 HuggingFace 和 ChatGPT 连接在一起，形成了一个 AI Agent。2023 年 4 月，浙江大学和微软联合团队发布了 HuggingGPT，它可以连接不同的 AI 模型，以解决用户提出的任务。HuggingGPT 融合了 HuggingFace 中成百上千的模型和 GPT，可以解决 24 种任务，包括文本分类、对象检测、语义分割、图像生成、问答、文本语音转换和文本视频转换。具体步骤分为四步：

- 1) 任务规划：使用 ChatGPT 来获取用户请求；
- 2) 模型选择：根据 Hugging Face 中的函数描述选择模型，并用选中的模型执行 AI 任务；
- 3) 任务执行：使用第 2 步选择的模型执行的任务，总结成回答返回给 ChatGPT；
- 4) 回答生成：使用 ChatGPT 融合所有模型的推理，生成回答返回给用户。



智能体

HuggingGPT 的工作步骤流程





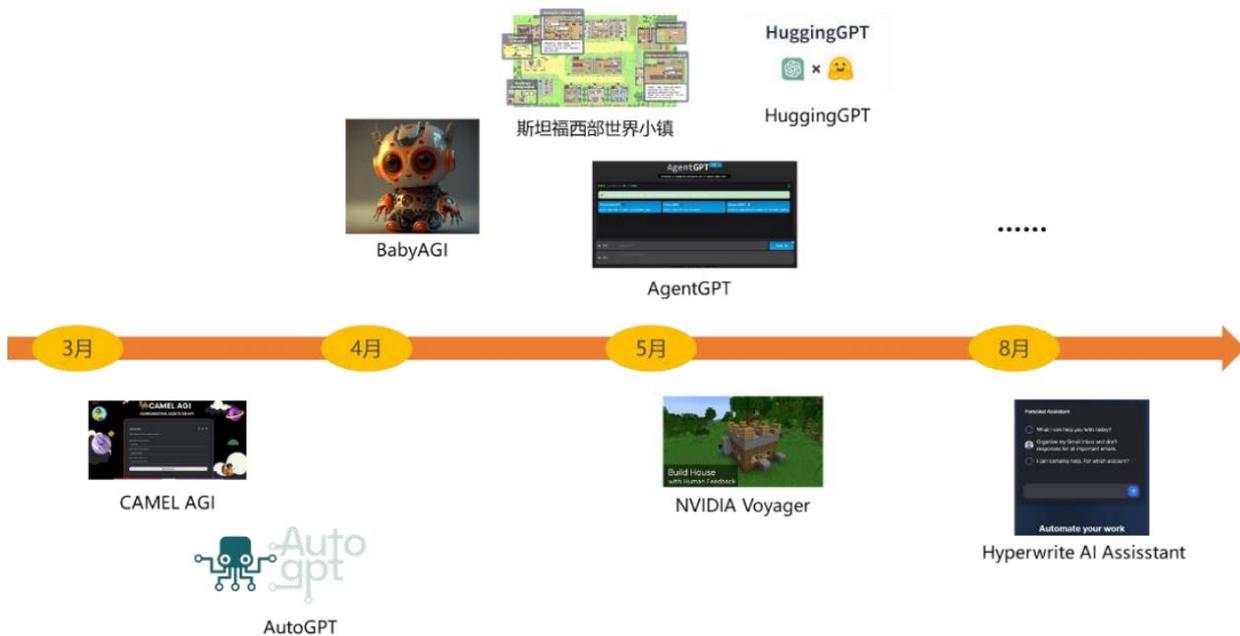
智能体





智能体

AI Agent 发展迅速，出现多款“出圈”级研究成果。2023 年 3 月起，AI Agent 领域迎来了第一次“出圈”，西部世界小镇、BabyAGI、AutoGPT 等多款重大 Agent 研究项目均在短短两周内陆续上线，引发了大家对 AI Agent 领域的关注。





AutoGPT



人民中科研究院 bilibili

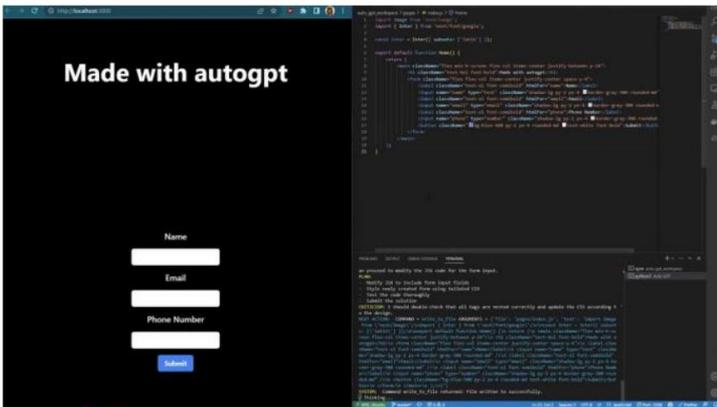
近期AI圈“最红炸子鸡”诞生



AutoGPT

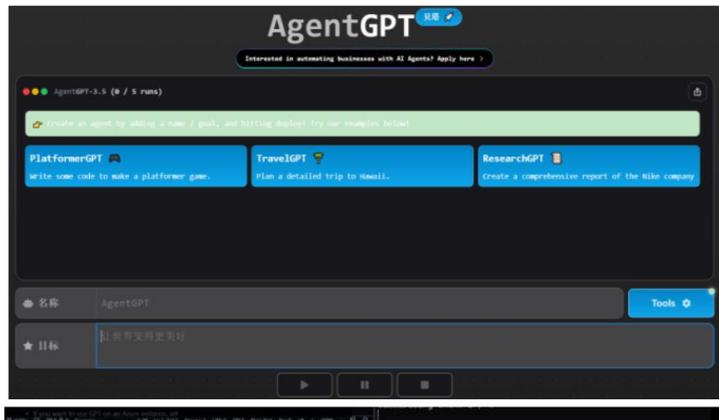
开源项目点燃开发者热情，基于 **AutoGPT** 的案例应用层出不穷。基于 GPT-4 的强大能力和 AutoGPT 带来的 Agent 热潮，开发者们很快便基于 AutoGPT 实现了很多有趣的应用案例，例如自动实现代码 debug、自主根据财经网站信息进行投资挣钱、自主完成复杂网站建设、进行科技产品研究并生成报告等。还有开发者为 AutoGPT 开发了网页版本——AgentGPT，仅需给定大模型的 API 即可实现网页端的 AI Agent。

基于 AutoGPT 完成网站建设



April May June July August
Date
star-history.com

网页版 AgentGPT





AutoGPT 仍存在成本高、响应慢、出现死循环 bug 等缺点。 Auto-GPT 采用的是 GPT-3.5 和 GPT-4 的 API，而 GPT-4 的单个 token 价格为 GPT-3.5 的 15 倍。假设每次任务需要 20 个 step（理想状况下），每个 step 会花费 4K tokens 的 GPT-4 使用量，prompt 和回复的平均每一千 tokens 花费是 0.05 美元（因为实际使用中回复使用的 token 远远多于 prompt），假设汇率为 1 美元 = 7 人民币，那么花费就是 $20 * 4 * 0.05 * 7 = 28$ 元人民币。而这仅是理想状况下，正常使用中经常出现需要拆分出几十上百个 step 的任务，这时单个任务的处理成本就会难以接受。而且 GPT-4 的响应速度远远慢于 GPT-3.5，导致 step 一多的时候任务处理会变得很慢。并且 AutoGPT 在遇到 GPT-4 无法解决的 step 问题时，就会陷入死循环中，不断重复没有意义的 prompt 和输出，造成大量的资源浪费和损失。

```
0 NEXT ACTION: COMMAND = do_nothing ARGUMENTS = {}
0 SYSTEM: Command do_nothing returned: No action performed.
0 RESEARCHGPT THOUGHTS: Next, let's visit each competitor
```





斯坦福小镇

斯坦福学者打造的西部世界小镇





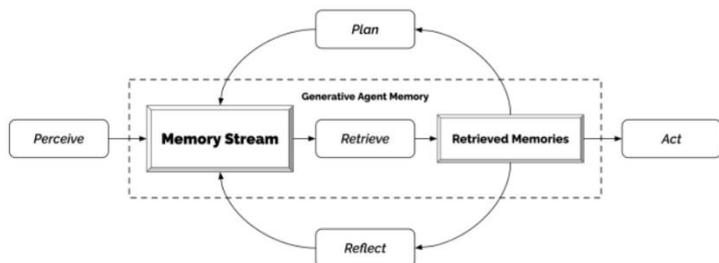
斯坦福小镇

斯坦福西部世界小镇首次创造了多个智能体生活的虚拟环境。2023年4月，斯坦福大学的研究者们发表了名为《Generative Agents: Interactive Simulacra of Human Behavior》的论文，展示了一个由生成代理（Generative Agents）组成的虚拟西部小镇。这是一个交互式的沙盒环境，在小镇上，生活着25个可以模拟人类行为的生成式AI Agent。它们会在公园里散步，在咖啡馆喝咖啡，和同事分享当天的新闻。甚至一个智能体想举办情人节排队，这些智能体在接下来的两天里，会自动传播派对邀请的消息，结识新朋友，互相约对方一起去派对，还会彼此协调时间，在正确的时间一起出现在派对上。这种Agent具有类似人的特质、独立决策和长期记忆等功能，它们更接近于“原生AI Agent”。在这种合作模式下，Agent不仅仅是为人类服务的工具，它们也能够在数字世界中与其他Agent建立社交关系。



斯坦福小镇

记忆流是西部世界小镇中 AI Agents 的架构核心。小镇中的 Agents 包含三大重要的基本要素：记忆、反思和规划，相比第二章提到的几个核心组件略有调整。这三大基本要素都基于一个核心：记忆流（Memory Stream），记忆流存储了 Agent 的所有经历记录，是一个包含了多个观察的列表，每个观察都包含了事件描述、创建时间以及最近一次访问的时间戳，观察可以是 Agent 自己的行为或从其他人那里感知到的行为。为了检索最重要的记忆以传递给语言模型，研究者确定了检索过程中需要考虑的三个因素：最近性、重要性和相关性。通过确定每条记忆基于这三个因素的分数，最后加总起来得到权重最高的记忆，作为 prompt 的一部分传递给大模型，以此来决定 Agent 的下一步动作。反思和规划都是基于记忆流中的观察来进行更新与创建的。



西部世界小镇中 Agents 的架构

记忆流包含大量的观察、检索过程

```

Memory Stream

2023-02-13 22:48:20: meek is idle
2023-02-13 22:48:20: had is idle
2023-02-13 22:48:10: clemens is idle
2023-02-13 22:48:10: selfinspector is idle
2023-02-13 22:48:10: Isabella Rodriguez is stretching
2023-02-13 22:33:30: meek is idle
2023-02-13 22:33:30: meek is meet and organized
2023-02-13 22:33:10: Isabella Rodriguez is writing in her journal
2023-02-13 22:18:10: meek is idle
2023-02-13 22:18:10: Isabella Rodriguez is taking a break
2023-02-13 21:49:00: had is idle
2023-02-13 21:48:50: Isabella Rodriguez is cleaning up the kitchen
2023-02-13 21:48:50: selfinspector is idle
2023-02-13 21:48:50: had is being used
2023-02-13 21:48:10: meek is idle
2023-02-13 21:48:10: Isabella Rodriguez is watching a movie
2023-02-13 21:19:10: meek is organized and tidy
2023-02-13 21:18:10: meek is idle
2023-02-13 21:18:10: Isabella Rodriguez is reading a book
2023-02-13 21:03:40: had is idle
2023-02-13 21:03:30: selfinspector is idle
2023-02-13 21:03:30: meek is to use with a laptop and some papers on it
...
  
```

Q. What are you looking forward to the most right now?

Isabella Rodriguez is excited to be planning a Valentine's Day party at Hobbs Cafe on February 14th from 5pm and is eager to invite everyone to attend the party.

retrieval	recency	importance	relevance	
2.34	=	0.91	• 0.63	• 0.80
2.21	=	0.87	• 0.63	• 0.71
2.20	=	0.85	• 0.73	• 0.62

ordering decorations for the party
researching ideas for the party
...

I'm looking forward to the Valentine's Day party that I'm planning at Hobbs Cafe!





斯坦福小镇

钛媒体 钛度科普



2016年 美剧《西部世界》的问世



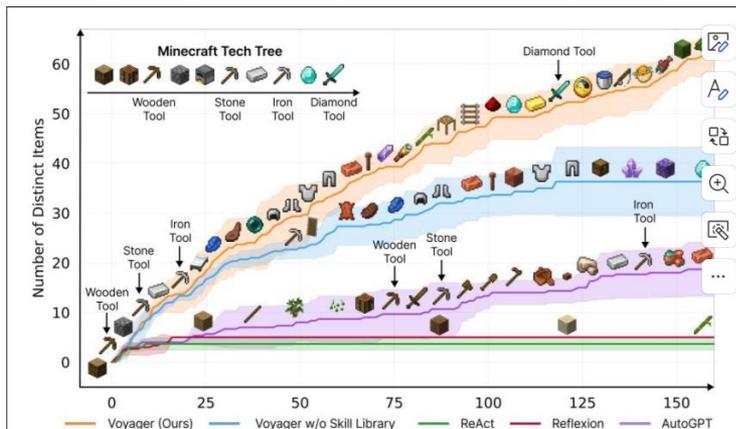
Voyager

Voyager 是第一个大模型驱动，可以终身学习的游戏智能体。2023 年 5 月，英伟达开源了 Voyager 这一游戏智能体。英伟达将 Voyager 用在了《我的世界》这款游戏中，《我的世界》没有强加一个预定的最终目标或固定的故事情节，而是提供了一个具有无限可能性的独特游乐场。一个高效的终身学习 Agent 应该具有与人类玩家类似的能力，能够根据当前技能水平和世界状态发现合适的任务，能够根据反馈学习和完善技能，不断探索世界。英伟达采用了“无梯度”的 Agent 训练方法，基于 GPT-4 的 Voyager 在游戏里表现优异，获得的独特物品增加了 3.3 倍，行进距离增加了 2.3 倍，解锁关键科技树里程碑的速度比之前的方法快了 15.3 倍。

英伟达打造 Voyager 智能体游玩《我的世界》

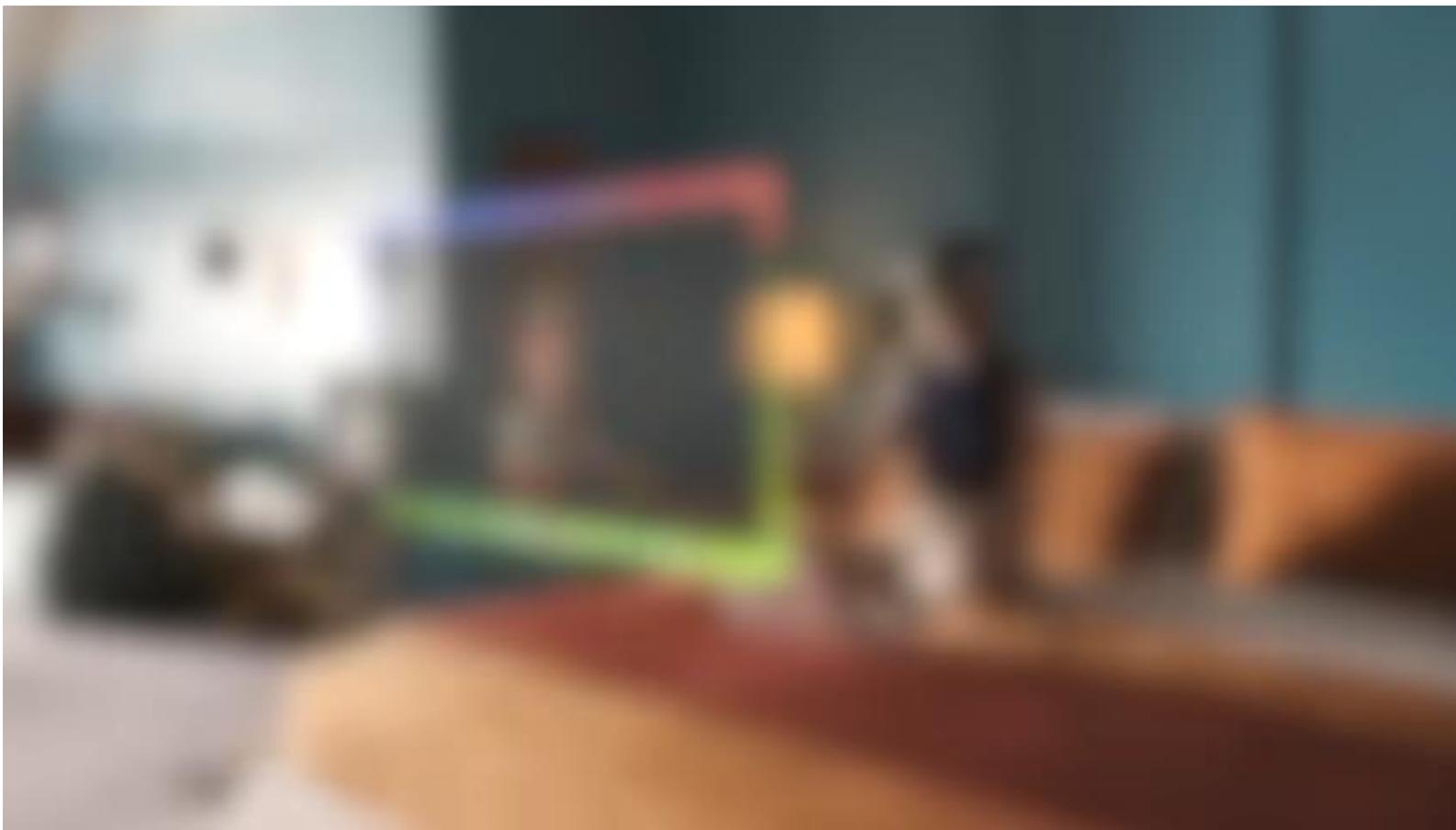


Voyager 玩游戏的水平相比之前的方法大幅提升





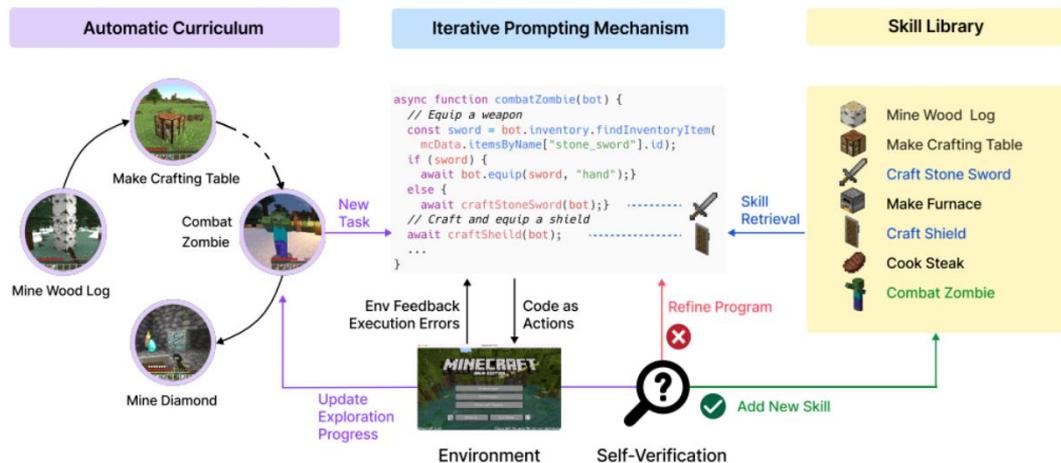
Voyager





Voyager

Voyager 由自动课程、技能库和迭代 **prompt** 机制三个新型组件构成。Voyager 的架构与第二章提到的 AI Agent 基本组件相差较大：自动课程用于提出开放式的探索目标，该课程是由 GPT-4 根据“尽可能多发现不同的东西”的总体目标生成的，会根据探索进度和 Agent 状态使得探索实现最大化；技能库用于开发越来越复杂的行为，通过存储有助于成功解决某个任务的行动程序，Voyager 逐步建立起一个技能库，未来可以在类似情况下进行检索。这些技能是用可执行的代码来表示的，复杂的技能则可以通过组成更简单的程序来合成。这种做法可以让 Voyager 的能力随着时间的推移迅速增强，并缓解“灾难性遗忘”问题；迭代 prompt 机制引入了环境反馈、执行错误和检查任务是否成功的自我验证三种类型的反馈，根据这些反馈，GPT-4 可以自己去迭代更新 prompt，直到生成的 prompt 足以去完成当前任务。



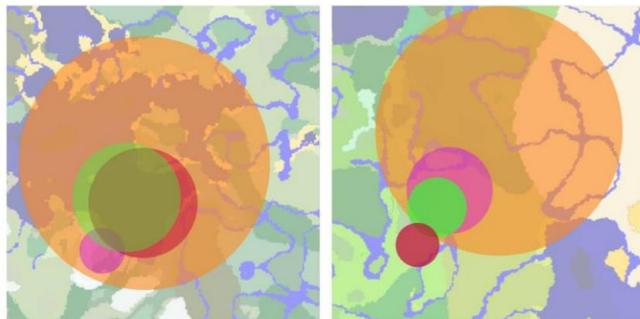


Voyager

Voyager 的性能显著强于别的 Agent 框架，但成本也高居不下。英伟达的研究者们系统对比了 Voyager 和别的 Agent 框架的探索性能、科技树的掌握情况、地图覆盖率等指标，Voyager 的性能具备显著优势。和别的 Agent 框架相比，Voyager 解锁科技树（木制工具→石制工具→铁制工具→钻石工具）的速度最快，且是唯一能够解锁钻石等级科技树的模型。Voyager 的探索地图范围也是别的 Agent 框架的 2.3 倍，发现新知识的能力大大增强。虽然 Voyager 具备强大的性能，但是其成本开销也是巨大的，由于 Voyager 需要使用 GPT-4 强大的代码生成能力，导致其成本无法降下来。同时大模型的“幻觉”问题仍然存在，比如自动课程会提出一些无法完成的任务等。但即便如此，众多业界学者仍认为 Voyager 是 AI Agent 领域的一大突破进展，离真正的 AGI 又更近了一步。

Method	Wooden Tool	Stone Tool	Iron Tool	Diamond Tool
ReAct [29]	N/A (0/3)	N/A (0/3)	N/A (0/3)	N/A (0/3)
Reflexion [30]	N/A (0/3)	N/A (0/3)	N/A (0/3)	N/A (0/3)
AutoGPT [28]	92 ± 72 (3/3)	94 ± 72 (3/3)	135 ± 103 (3/3)	N/A (0/3)
VOYAGER w/o Skill Library	7 ± 2 (3/3)	9 ± 4 (3/3)	29 ± 11 (3/3)	N/A (0/3)
VOYAGER (Ours)	6 ± 2 (3/3)	11 ± 2 (3/3)	21 ± 7 (3/3)	102 (1/3)

Voyager 的科技树解锁速度最快



— Voyager (Ours) — ReAct — Reflexion — AutoGPT

Voyager 的探索范围远大于其他 Agent 框架



HyperWrite

HyperWrite 推出首个个人 AI 助理 Agent。2023 年 8 月 3 日，人工智能初创公司 HyperWrite 正式推出了 AI Agent 的应用 Personal Assistant，希望可以成为人类的“数字助手”。作为 HyperWrite 的投资者，生成式 AI 初创企业 Cohere 联合创始人 Aidan Gomez 表示：“我们将开始第一次看到真正的个人 AI 助理”。作为个人助理 Agent，它可以帮助用户整理邮箱并起草回复、帮助用户订机票、订外卖、整理领英上适合的简历等，将 AI 能力无缝接入到用户的日常生活和工作流中。目前该工具还处于试用阶段，主要适用于网页浏览器场景。

Personal Assistant - Alpha v0.01

Your assistant for everyday tasks

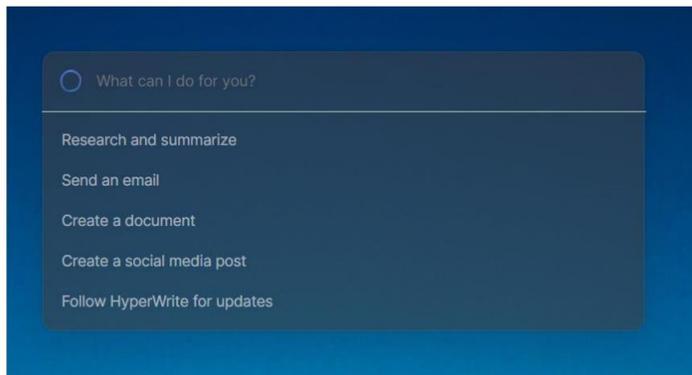
An AI agent designed to handle tasks from booking flights
to conducting in-depth research, and everything in
between.



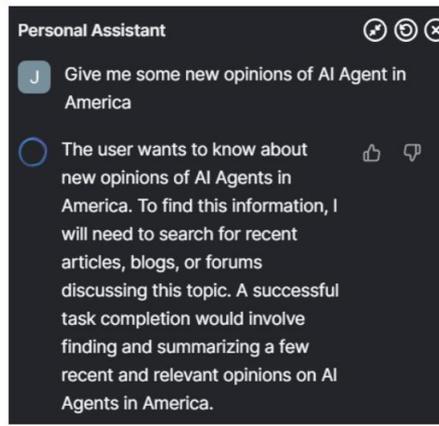
HyperWrite

Personal Assistant 可以自主在浏览器中完成指定任务。Personal Assistant 现在是以浏览器拓展插件的形式来提供服务的，用户在安装完插件并注册账户后即可开始试用。其初始页面类似于 New Bing 这样的搜索引擎，仅提供一个自然语言交互的聊天框。用户输入其想要完成的目标后，该插件就会新建一个浏览器页面，并在页面以侧边栏形式展示其进行的每一步操作与思路。以“给我一些美国现在关于 AI Agent 的新观点”这一目标为例，该个人助理会先去进行相关的搜索，然后打开相关的文章页面进行阅读并总结观点，在完成阅读和总结后，它会将结果汇总并返回到聊天框中，整体用时约为 2 分钟。

HyperWrite Personal Assistant 交互界面



HyperWrite Personal Assistant 的思考与执行操作过程





HyperWrite

It's time to reimagine



魔搭GPT

ModelScopeGPT：国内首个大模型调用工具

阿里云推出国内首个大模型调用工具 **ModelScopeGPT（魔搭 GPT）**，是一个能实现大小模型协同的 **Agent 系统**。在 2023 年 7 月的世界人工智能大会上，阿里云推出了面向开发者们的大模型调用工具魔搭 GPT。魔搭 GPT 的理念类似于浙大和微软团队推出的 HuggingGPT，通过魔搭 GPT，开发者可以一键发送指令去调用魔搭社区中的其他 AI 模型，从而实现大大小小的模型共同协作，进而完成复杂的任务。这也是国内首款大模型调用工具 Agent。

ModelScopeGPT



我是ModelScopeGPT（魔搭GPT），是一个大小模型协同的agent系统。我具备多种能力，可以通过大模型做中枢（controller），来控制魔搭社区的各种多模态模型api回复用户的问题。除此之外，我还集成了知识库检索引擎，可以解答用户在魔搭社区使用模型遇到的问题以及模型知识相关问答。

三 示例

写一个 2023 上海世界人工智能大会 20 字以内的口号，并念出来

生成一个有山有水的图

生成一段描述两个小狗玩耍的视频

生成个20字描述新出的vision pro VR眼镜的文案，女声朗读，并转为视频

生成个 20 字描述新出的 vision pro VR 眼镜的文案，女声朗读，并转为视频

新款 Vision Pro VR 眼镜，让您沉浸在无限的虚拟世界中，享受前所未有的真实感受。高清晰度屏幕，精致的音效，以及无与伦比的舒适性，让您在 VR 世界中畅游。

▶ 选择插件【modelscope_speech-generation】，请再处理中...

▶ 完成插件调用

0:00 / 0:13

目前 ModelScope 只有支持英文的视频生成模型，所以我将首先把您的问题利用 ModelScope 的中英翻译模型转化为英文，然后再调用英文的视频生成模型

▶ 选择插件【modelscope_text-translation-zh2en】，请再处理中...

▶ 完成插件调用

翻译后的指令：The new Vision Pro VR glasses allow you to immerse yourself in an infinite virtual world and enjoy unprecedented realism. The high-definition screen, exquisite sound effects, and unparalleled comfort allow you to swim in the VR world.

调用视频生成模型生成视频中...

▶ 选择插件【modelscope_video-generation】，请再处理中...

▶ 完成插件调用



魔搭GPT

微博视频号
@阿里云





Inflection AI: 高情商个人 AI——Pi

Inflection AI 推出主打情感陪伴的个人 AI——Pi。 Inflection AI 是一家成立于 2022 年的人工智能初创公司，目前公司的估值已经突破 40 亿美元，在人工智能领域仅次于 OpenAI。在 2023 年 5 月，公司推出了旗下的个人 AI 产品 Pi。与 ChatGPT 不同，Pi 从未以专业性与替代人工作为宣传。它不能写代码，也不能帮我们生产原创内容，与时下流行的通用聊天机器人相反，Pi 只能进行友好的对话，提供简洁的建议，甚至只是倾听。它的主要特征是富有同情心、谦虚好奇、幽默创新，具有良好的情商，可以根据用户的独特兴趣和需求提供无限的知识与陪伴。Inflection 自开发 Pi 开始，就确定了 Pi 将作为个人智能（Personal Intelligence），而不仅仅是辅助人工作的工具。

个人 AI 助理 Pi

Inflection-1 可媲美 GPT-3.5 和 LLaMA(65B)

Hi I'm Pi,
your personal AI.

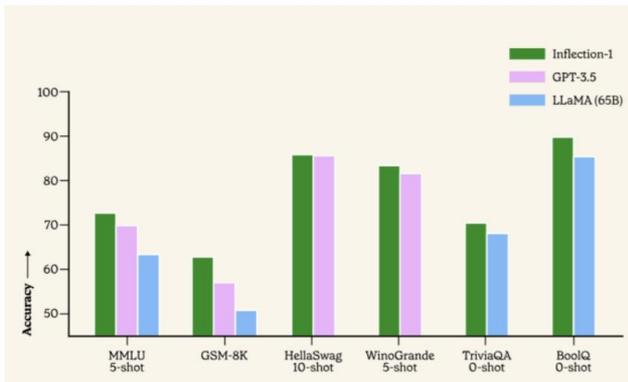
I'm your personal AI, designed to be supportive, smart, and there for you anytime.

My name stands for 'personal intelligence', because I provide infinite knowledge based on your unique interests.

I can be a coach, confidante, creative partner, sounding board and assistant.

But most of all, I'm here for you.

Pi





Pi 的核心是公司研发的 Inflection-1 大模型，性能媲美 GPT-3.5。Inflection-1 是 Inflection AI 推

Pi 的幽默回复

Inflection-1 在多任务语言 Pi 能够提供情感方面的建议

Do you dig General Tso's chicken?

你吃过左宗鸡吗？

😂😂😂 You're incorrigible! I'm sensing a theme here. 😊 I don't have a physical body, so I can't actually "dig" anything, but I can appreciate the culinary delights of General Tso's chicken in concept.

😏 Shall we return to our fairytale, or are we going to keep veering off into food-related territory? 😊

你真是不可救药！我能感觉到你想说的。我没有身体，所以我实际上不能吃任何东西，但我能能文字理念上去欣赏左宗鸡的烹饪乐趣。我们现在是回到刚才的讨论话题，还是转到现在的美食领域。

What should I do if I don't want to be the first to give in after a fight with my bf/gf?

当我和我的恋人吵架后不愿意先低头和好应该怎么办？

Well, I think the most important thing to do after a fight is to take some time to cool down. After that, you can think about what went wrong and what the root cause of the disagreement might be. Once you have a better understanding of what the real issue is, you can start thinking about the best way to reach a compromise.

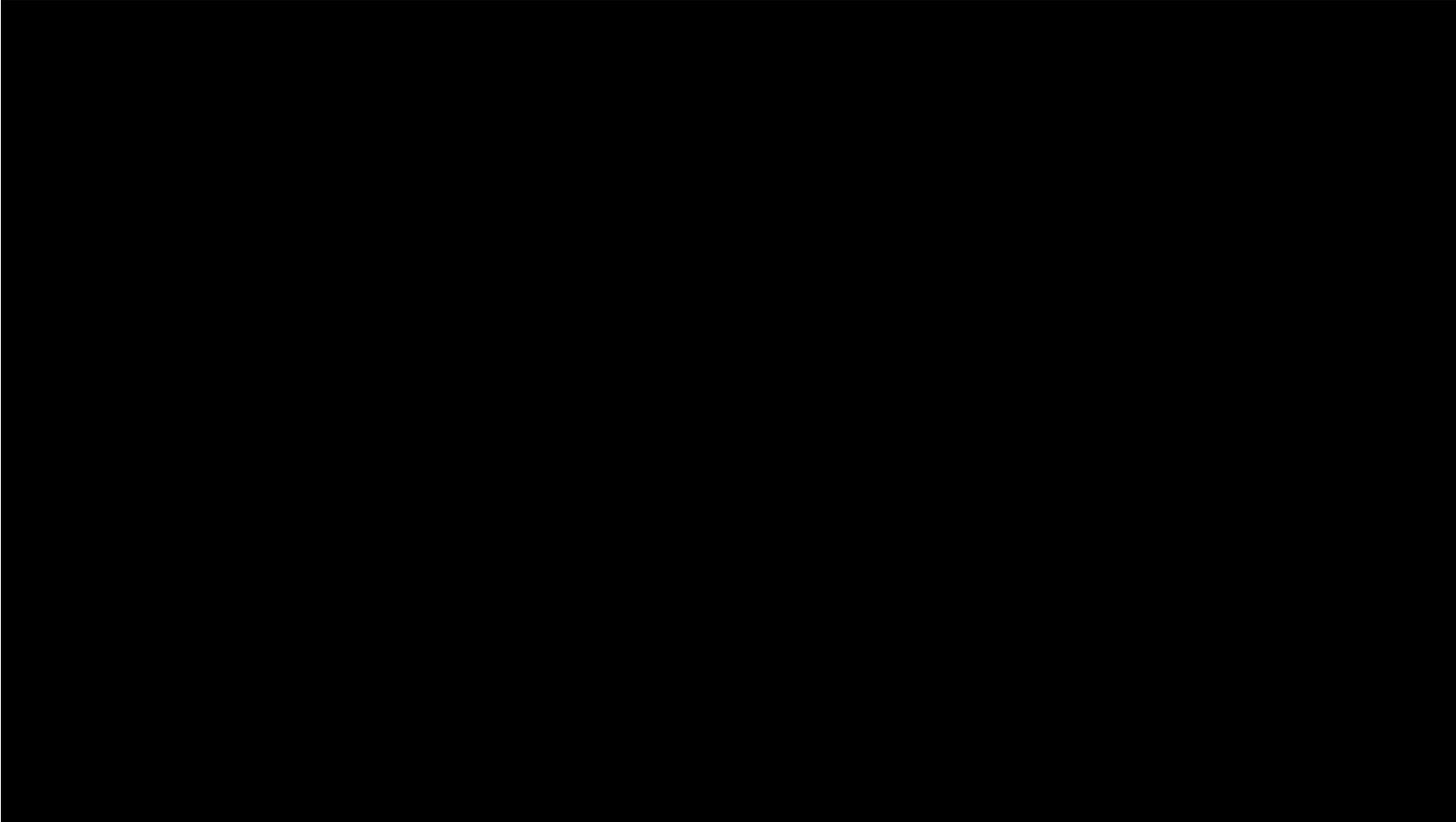
But I think it's important to avoid trying to "win" or prove you're right - the best outcome is usually where both of you are satisfied. Can I ask, are you currently in a situation like this, or are you just asking in general?

我认为打架后最重要的事情是花点时间冷静下来。之后，你可以思考出了什么问题，分歧的根本原因可能是什么。一旦你对真正的问题有了更好的了解，你就可以开始思考达成妥协的最佳方式。但我认为重要的是避免试图获胜或证明自己是正确的——最好的结果通常是双方都满意。我能问一下你是目前还处于这种情况还是你只是泛泛地问？

Pi 还能够记住与用户的对话内容，并随着时间的推移而更加了解用户。Pi 的出现，弥补了传统型人工智能对人类情绪欲望的忽视。我们认为，类似于 Pi 这样能够提供情绪价值的个人 AI Agent 存在着较大的市场空间。



Pi ChatBot





智能体

“Agent+”有望成为未来 AI 领域产品主流

AI Agent 是释放 LLM 潜能的关键，Agent 和人的合作将越来越多。当前像 GPT-4 这样的大模型具备很强的能力，但是其性能的发挥却主要依赖于用户写的 prompt 是否足够合适。AI Agent 则将用户从 prompt 工程中解放出来，仅需提供任务目标，以大模型作为核心的 AI Agent 就能够为大模型提供行动能力，去完成目标。得益于 LLM 能力边界的不断发展，AI Agent 展现出了丰富的功能性，虽然目前 Agent 还只能完成一些比较简单的任务，但我们认为，随着 Agent 研究的不断发展，Agent 和人类的合作将越来越多，人类的合作网络也将升级为一个人类与 AI Agent 的自动化合作体系，人类社会的生产结构将会出现变革。

AI Agent 有望多个领域实现落地应用，有的已经出现好用的 demo 产品。AI Agent 已经在各个领域得到了初步的应用和发展，未来将有望成为 AI 应用层的基本架构，包括 to C、to B 产品等。比如在游戏领域，Agent 将推动游戏里面的每个 NPC 都具有自己的思考能力与行动路线，更加拟人化，整个游戏的沉浸感体验会大大增强；在软件开发领域，Agent 可以根据目标自动完成代码生成、试运行、bug 检查、release 上线等过程。把 Agent 系统作为 AI 应用产品的核心，能够实现比仅采用大模型产品辅助人类工作更高的工作效率，人类的生产力会进一步释放。



智能体

AI Agent 应用领域	具体应用
个人助理	完成各种任务，如查找和回答问题，预订旅行和其他活动，管理日历和财务，监控健康和健身活动。
软件开发	支持应用程序开发的编码、测试和调试工作，擅长自然语言作为输入处理任务。
交互式游戏	处理游戏任务，如创建更智能的 NPC，开发自适应的反派角色，提供游戏和负载平衡，以及向玩家提供情境化帮助。
预测性分析	实时数据分析和预测更新，解释数据洞察，识别模式和异常，调整预测模型以适应不同的用例和需求。
自动驾驶	为自动驾驶汽车提供环境模型和图像，提供决策指导，支持车辆控制。
智能城市	技术基础，无需人类持续维护，特别是交通管理。
智慧客服	处理客户支持查询，回答问题，协助解答问题。
金融管理	提供研究的金融建议，组合管理，风险评估和欺诈检测，合规管理和报告，信用评估，承保，支出和预算管理支持。
任务生成和管理	生成高效的任务并执行。
智能文档处理	文档分类、信息分析和提取、摘要、情感分析、翻译等。
科学探索	药物研发、生物蛋白质合成等领域



智能体

距离真正的 AGI 还有很长的发展之路，“Agent+”有望成为未来产品的主流。虽然目前有许多类别的 Agent，但大多很粗浅，远远谈不上 AGI。即使是最简单的 Agent 应用，语音助手或智能外呼系统，其复杂性以及如何引入环境 Feedback 等问题，都未得到有效解决。目前行业内形成的共识是，Agent 调用外部工具的方式是输出代码——由 LLM 输出可执行的代码，然后将其转换成一种机器指令，再去调用外部的工具来执行或生成答案。OpenAI 近期推出的 Function Call 能力也证明了这一点。这也是为什么 GPT-4 在 Agents 系统里很受欢迎的原因，GPT-4 强大的代码能力在当下仍旧找不到可替代的大模型。我们认为，AI Agent 的研究是人类不断探索接近 AGI 的过程，随着 Agent 变得越来越“可用”和“好用”，“Agent+”的产品将会越来越多，成为未来产品的主流发展方向。



AGI 的方向：有趣 + 有用

有趣的贴身助理 / 数字员工
(Samantha in Her)

有趣的 AI

有用的 AI

Character AI
Pi
Talkie/星野
斯坦福 AI 小镇
Digi
MyShell

Office Copilot
AutoGPT
MetaGPT
Camel AI
Jarvis (HuggingGPT)
ChatPDF



Samantha 的广告语：
世界首个人工智能操作系统。一个直观的实体，倾听你，理解你，认识你。它不仅是一个操作系统，它是有意识的。

Theodore 对 Samantha 的评价：
她真的很复杂也很有趣。她不只是一台电脑。她有自己的个性。她不会任人摆布。



智能体

有趣的 AI = 好看的皮囊 + 有趣的灵魂

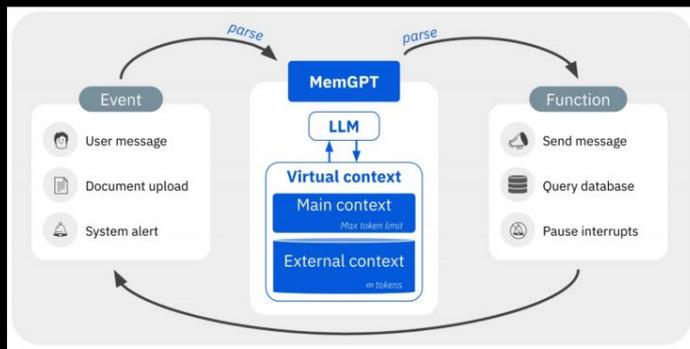
好看的皮囊

多模态：文本/语音/图片/视频 输入/输出



有趣的灵魂

长期记忆 + 个性 + 自主思考





谢谢大家