

夯实基础，了解LLM大语言模型-章节介绍

LLM大语言模型快速认识

- ◆ 介绍什么是LLM大语言模型，大语言模型在企业中的价值与市场需求，LLM在企业中落地的一些简单案例。
- ◆ 掌握ChatGPT的注册与使用，引出大模型的局限性。
- ◆ 了解大模型对软件构建有什么影响，未来的编程范式，探讨程序员在 AI 2.0 时代的新机遇。

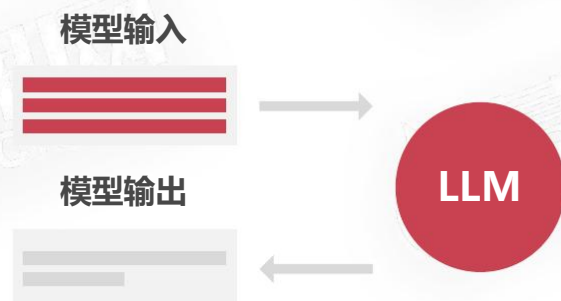
LLM应用开发入门及交互方式

- ◆ 探索AI研究的两个分支——基础研究与**应用研究**，快速扫盲LLM应用开发的常见专有名词，以及什么是LLM应用开发。
- ◆ 学习在AI新时代下，LLM/AI Agent应用的交互发生的变化，以及自然交互模式的一些思考。

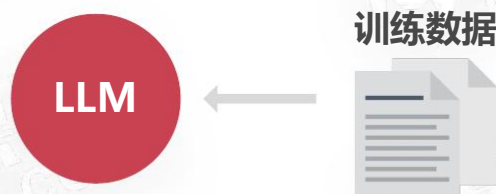
什么是大语言模型（LLM）

LLM大语言模型定义

- ◆ LLM全称Large Language Model，即大语言模型，是一种用大量数据训练的深度学习模型，给模型一些输入，它可以预测并返回相应的输出。



LLM大语言模型的训练数据大



GPT-3公开训练数据

Dataset	Quantity (tokens)	Weight in Training mix	Epochs elapsed when Training for 300B tokens
Common Crawl(filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	5 billion	3%	3.4

<https://arxiv.org/abs/2005.14165>

Common Crawl

网络爬虫公开数据集（海量互联网内容）

WebText2

Reddit 论坛的网页文本

Books1, Books2

互联网书籍语料库

Wikipedia

整个英文维基百科知识库



书籍



新闻



论文



维基百科



社交媒体

...

3000亿文本单位

大语言模型的参数大

模型输入



模型输出

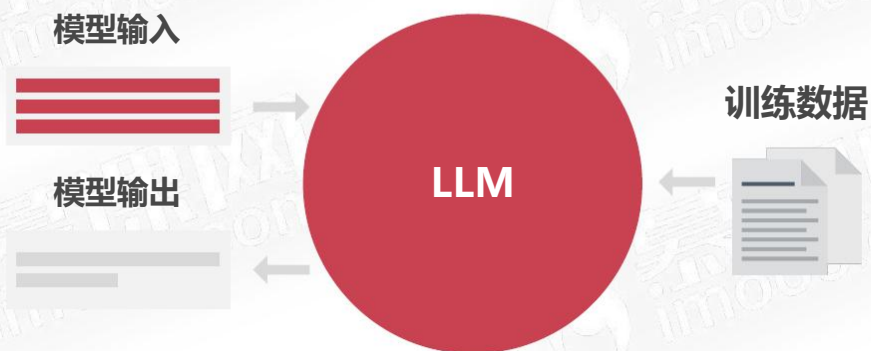


LLM

训练数据



大语言模型的参数大



OpenAI历代公开大模型参数



GPT-1
1.17亿

OpenAI历代公开大模型参数

GPT-1
1.17亿

GPT-2
15亿

OpenAI历代公开大模型参数

GPT-1
1.17亿

GPT-2
15亿

GPT-3
1750亿

LLM大语言完成通用的任务



大语言模型参数量基础



- ◆ 参数量大一般指的是7b~100b+, 即70亿~1000亿+参数

70亿(4字节)28G

- ◆ 参数其实就是一个浮点数 (2字节或4字节), 比如3.1415, 所以一个7b的模型, 本质上就是一坨70亿以上的数字而已。

大语言模型的基础单位Token

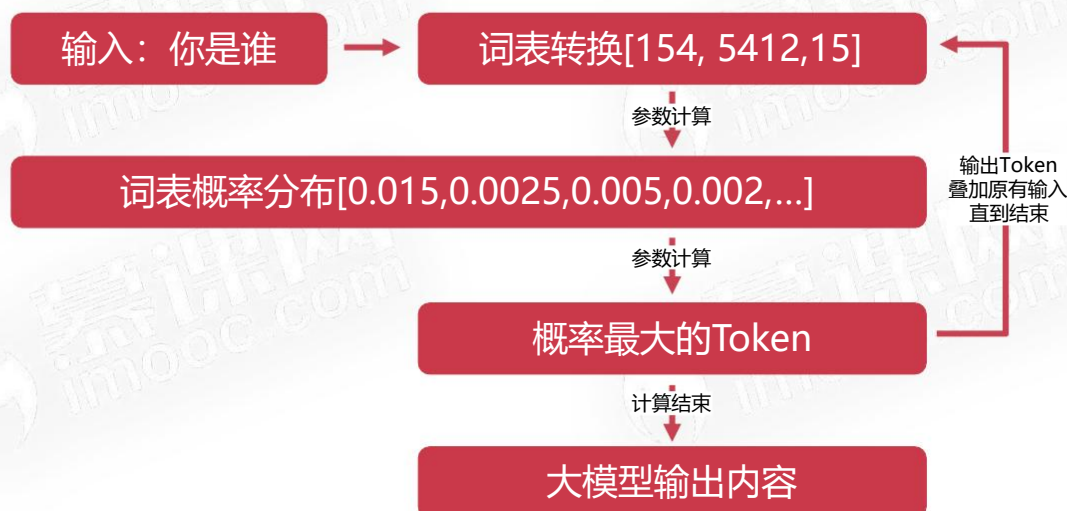
- ◆ Token其实就是文本的片段，是大模型计算长度的单位，对于汉字，可以是字、词、甚至是半个字或者三分之一个字。
- ◆ 对于一个仅支持英语的模型，它的词表可以只有a-z 26个字母，加上逗号句号空格等标点符号，Token 数可以非常少。
- ◆ 汉字字词更多，语义更复杂多样，所以包含的 Token 数会更多，很多语言模型都支持多语言，包含各种符号、单词、单词片段等，所以往往会有几十万个 Token 甚至更多。

大语言模型词表介绍与示例

- ◆ 词表，就是这个模型的所有 Token 映射，每个 Token 有其对应的 id，一般从 0 开始，如下：

```
1 "vocab": {  
2     # 开头是一些特殊符号  
3     "<unk>": 0,  
4     "<|startoftext|>": 1,  
5     ...  
6     # 这是字节token，如果出现不在词表中的特殊符号会回退到字节表示  
7     "<0x00>": 305,  
8     "<0x01>": 306,  
9     ...  
10    # 下面是正常的英文token，有_的表示是单词的开头，没有的是单词中间  
11    "ct": 611,  
12    "_re": 612,  
13    ...  
14    # 有中文token出现  
15    "安徽省": 28560,  
16    "子和": 28561,  
17    ...  
18 }
```

大语言模型的工作流程

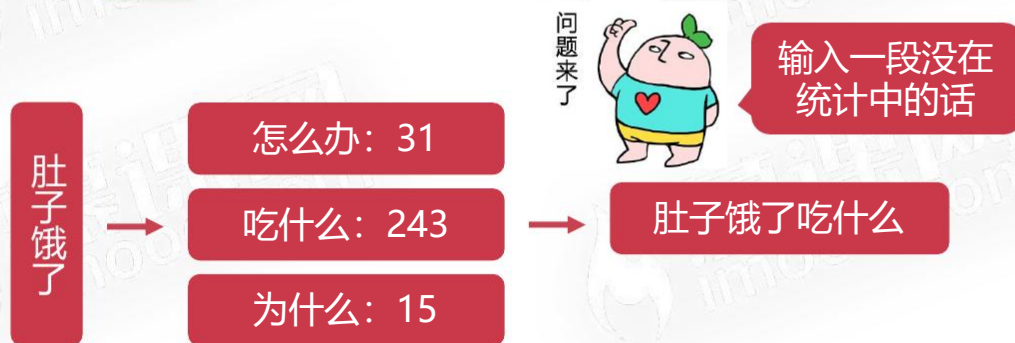


大语言模型预测Token的机制

- ◆ 如何快速预测下一个 token 是什么呢？一种最简单的办法就是基于统计，通过大量数据的统计，找到下一个 token。
- ◆ 采集大量文本进行扫描计算，并记录所有片段的输入以及下一个文本出现的次数，得到一张巨大的分布表。
- ◆ 将输入的文本对照分布表查询，找到所有 token 的出现次数或概率，找到出现次数最大的 token 即为预测结果。

大语言模型预测Token的机制

- ◆ 如何快速预测下一个 token 是什么呢？一种最简单的办法就是基于统计，通过大量数据的统计，找到下一个 token。



大语言模型预测Token的机制

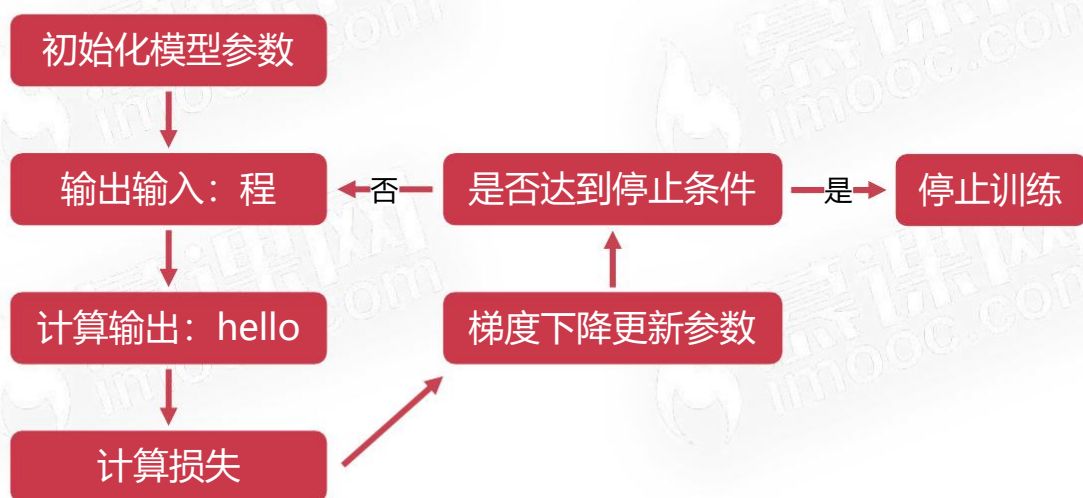
- ◆ 如何快速预测下一个 token 是什么呢？一种最简单的办法就是基于统计，通过大量数据的统计，找到下一个 token。



模型训练-从随机到智能

- ◆ 训练指的是将大量文本输入给模型，进而得到模型参数，目前LLM训练一般用到了大量文本，一般在 2T token 以上。

模型训练流程图（程序员的梦工厂）



课程总结

- ◆ 学习什么是 LLM、Token、词表、上下文长度，以及为什么有些模型不支持中文，或中文支持较差等。
- ◆ 初步了解了一个极简的 LLM 工作流程——不断预测下一个词是什么，直到结束。
- ◆ 从非机器/机器学习角度了解了 LLM 的预测流程、训练流程等。

LLM在企业中的价值与市场需求

所有产品都值得用AI重做一遍

- ◆ 2023年 GPT-4 发布以后，张勇在阿里云峰会上宣布阿里所有产品都将接入大模型全面升级，并强调**所有产品都值得用AI 重做一遍**。

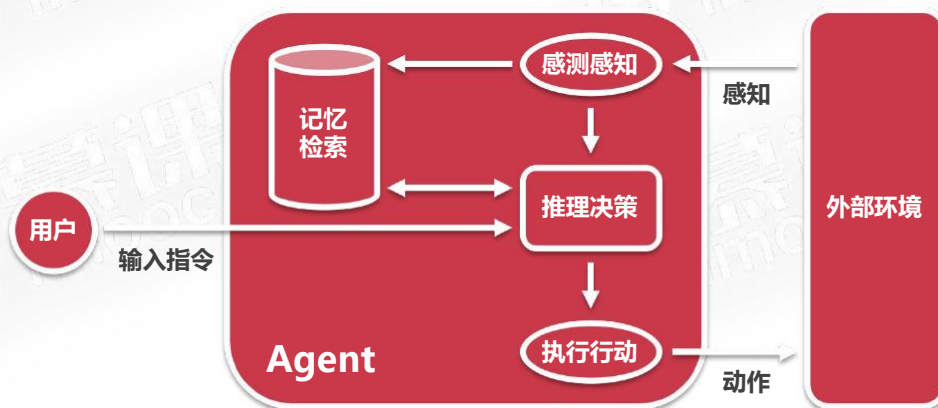
- ◆ 通用人工智能（AGI）将是 AI 的终极形态，几乎已成为业界共识。类比之，构建**智能体 Agent**则是 AI 工程应用当下的“终极形态”。

- ◆ 大模型出现后，AI Agents 衍生出一种新的架构模式，将最重要的**规划/决策部分或全部交由LLM**完成。



走进人工智能——Agent/LLM

- ◆ 智能体指代具有**自主性和智能的程序或系统**，能够通过感知、规划、决策并执行相关任务。



Agent/LLM在企业应用场景落地



案例1: 基于Agent/LLM的智能客服

- ◆ 企业的传统客服存在效率低、培养成本大、工作压力大、提供服务参差不齐等问题，在线客服是企业 and 客户沟通的桥梁，可以为企业带来更多的商机。
- ◆ 利用 LLM+企业知识库搭建 AI 智能客服，对比传统客服可以实现智能沟通、无间断工作、智能分析、自动分类、跨语言沟通等。

案例2: 基于Agent/LLM的数据分析

- ◆ 几乎所有企业每天都会产生大量的数据，除了新数据，还有海量的历史沉淀数据，涉及：日志、历史代码、订单、会员信息、行业数据等。
- ◆ 这些数据量庞大、结构差异明显、利用率低，导致企业很难精确获知和建立各项数据之间的关联。
- ◆ 利用Agent/LLM参与数据的自动分类、信息提取、数据分析等，建立起各项数据之间的关联。

案例3: 人工智能即服务-AlaaS

- ◆ AI as a Service(AlaaS, AI 即服务)是一项非常新颖的变革，将 AI 视为应用后端服务，让公司发布一款产品变得及其简单，无需投入昂贵的硬件、专业人才或耗时的开发流程。
- ◆ 翻译服务 = 限定的预设 Prompt + LLM 规范化输出
- ◆ 自动化运维 = 日志采集 + LLM 推理决策 + 工具包

行业需求爆发——奇点已来

- ◆ 人工智能浪潮已来，不加入就淘汰的说法虽然很夸张，但是 AI Agent/LLM 会像过去的 PPT 和 Excel 一样，成为几乎所有职业必学的技能。



- 1. 代替搜索引擎
- 2. 半工作自动化or全自动化
- 3. 通过GPT更快速学习知识
- 4. ...

提升职场竞争力/大模型工程师待遇

- ◆ 现阶段，AI 不会淘汰任何岗位，但会大大减少某些岗位的需求，并创造出大量的**高薪/前景好的新岗位。**

项目名称	领域	业务	投资轮次	投资金额
WorkMagic	智能营销	多Agent营销Sass平台	战略融资	1800万
幂律智能	智能法务	AI驱动的智能法律服务产品提供商	Pre-B轮	8000万
海纳AI	智能招聘	AI招聘产品开发商	A轮	数千万

学习LLM知识对职业发展的影响

- ◆ 技术领先：掌握更高效的Prompt编写技巧，让LLM更清楚你的需求，更高效完成工作上的任务、疑难杂症等。
- ◆ 薪资水平：由于LLM是一个相对新兴的领域，人才短缺，具备相关技能的人才往往能够获得更具竞争力的薪酬。
- ◆ 扩充职业前景：随着人工智能的发展和普及，LLM应用的前景非常广阔，不一定所有公司都需要训练LLM，但绝大部分公司都会基于LLM开发应用。

课程总结

- ◆ 学习了AI/Agent/LLM之间的关联，并分享了大模型在企业中的价值与一些常见的案例。
- ◆ 简单聊了当前行业的发展状况、LLM对职场竞争力的提升以及大模型工程师的待遇，学习LLM对职业发展的影响等。

ChatGPT聊天机器人的使用与局限性

ChatGPT介绍与账号注册/使用

- ◆ ChatGPT是OpenAI基于GPT架构开发的聊天机器人，可以通过邮箱/Google/Apple账户进行注册登录，推荐使用

Gmail邮箱注册。

- ◆ GPT-3.5目前可以有限的使用，



是GPT-4需要付费才
辅助代码编写。

- ◆ GPT-4的权益：使  404 您访问的页面飞走了 高级数据分析等。

大模型的使用与局限性

- ◆ ChatGPT不仅可以生成流畅自然的文本，还能理解上下文，根据用户指令生成相关内容，无论是**创作诗歌、编写代码还是撰写报告**，ChatGPT都能提供帮助。

- ◆ ChatGPT**很强**，拥有**镭射激光眼、透视视力、无穷的力量等...**，但**也有缺陷**，无法**免疫氪石**的伤害...



大模型的局限性与案例演示

- ◆ 无法获取实时更新的信息，提问：**2024年无锡马拉松的冠军是谁？**
- ◆ ChatGPT底层的GPT模型虽然更新的频率飞快，每次更新都会新增训练的内容，但是**训练高昂的费用与时间，无法实时将信息投喂给模型进行训练。**

大模型的局限性与案例演示

- ◆ 无法基于企业内部信息进行回答，提问：广东合虎餐饮有限公司有什么产品？
- ◆ GPT模型的训练资料都来源于互联网公开资料集，公司的内部资料并没有在GPT的语料库中，所以不会回答关于公司内部信息的问题，或者随意回答。

大模型的局限性与案例演示

- ◆ 默认无法与外部环境进行交互，提问：帮我看下慕课网上关于LangChain的课程有哪些？
- ◆ 默认情况下大模型没有接口与外部进行交互，例如通过爬虫采集特定网站的信息，只有配置后才可以实现。

大模型的局限性与案例演示

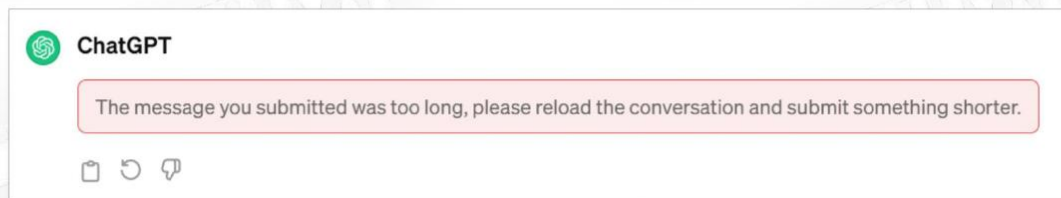
- ◆ 数学能力有限，处理复杂数学问题时，结果往往是错的，提问：方程有没有解 $x^3+15x^2-25x+0.5x^{0.5}=105.4$ ，请直接告诉我答案是什么？
- ◆ 大模型通过大量参数预测下一个词出现的概率，对于数学问题，大模型采用预测并不一定能计算得准确，特别是复杂的任务。

大模型的局限性与案例演示

- ◆ 自信地将虚假或者错误的信息呈现给用户（幻觉），如果不核实，很难发现错误，提问：请使用LangChain+Flask开发一个简单的机器人聊天接口。
- ◆ 大模型语料库如果没有涵盖某些信息，某些场合下，大模型并不会拒绝回答，而是自信地开始胡说八道。

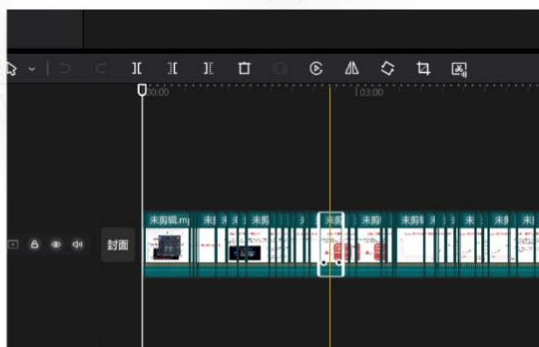
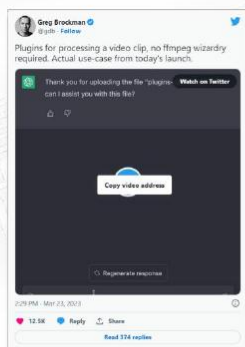
大模型的局限性与案例演示

- ◆ 处理长上下文，复杂上下文时不连贯，例如向ChatGPT提问超过模型能处理的上下文，亦或者对话次数过多时，会出现无法响应的服务器错误。



大模型的局限性与案例演示

- ◆ 人机对话很强，但不**一定**是一种**很**适合的交互方式，例如：通过GPT+插件的方式完成剪辑一个视频的前5s，整体体验不如在剪映中切一刀直接输出流畅。



课程总结

- ◆ 分享了ChatGPT如何注册，GPT-3.5与付费的GPT-4的使用差异，涵盖插件使用、升级指南等。
- ◆ 使用案例演示了大模型的9个局限性，大模型虽然很强，但是仍然存在不少缺陷，在当前可以[通过一些技巧来规避](#)。

大语言模型如何影响软件的构建

LLM在软件开发过程中的单点提效

- ◆ 智能代码提示
- ◆ 重复代码检查
- ◆ SQL语句的智能生成
- ◆ 跨端代码快速转换
- ◆ 静态代码检查与自动修复
- ◆ 代码注释生成
- ◆ 单元/接口测试代码生成
- ◆ 更精准的技术问答
- ◆ 代码评审与代码重构
- ◆ 失败用例自动分析与归因

GitHub Copilot与CodeGeeX

- ◆ GitHub Copilot和CodeGeeX分别是微软和智谱推出的AI编程助手，目前可以在VSCode中使用，支持对代码进行生成与补全、生成注释、代码翻译、代码问答等功能。

未来的架构可能是什么样的

- ◆ 在过去几个月里，有大量的KOL都在说：**所有的应用都值得用 AI 重写一遍**，而落地到现有的DevOps工具里，假设都需要重写，那么未来的架构可能是什么样的？

未来的架构可能是什么样的



AI时代的自动化编程5个等级

- ◆ C1：基于当前代码自动补全。
- ◆ C2：编写代码时AI可以预测下一行代码。
- ◆ C3：基于自然语言生成代码与编程语言的翻译。
- ◆ C4：高度自动化编程包括自然语言生成代码及注释、自动化测试、编程语言互译、代码补全与生成、调试及检查。
- ◆ C5：完全自动编程，AI可以看成是一个任意的软件，甚至无需代码，基于AI本身即可提供对应的服务。

短期看AI时代对程序员的影响

- ◆ AI可以帮助程序员完成代码的**自动化测试、代码审查**等重复、容易出错的工作，但是自动化测试、代码审查只是程序员的一小部分工作。
- ◆ 在创造性问题上，例如在项目的**需求分析、设计和策划**过程中，需要程序员进行创造性思考和判断，AI技术还没有完全达到这一点。
- ◆ 对于目前的AI进展来说，并不会导致程序员的完全失业，而是在一定程度上提高了程序员工作的效率和准确性。

长期看AI时代对程序员的影响

- ◆ 不久的将来AI有望取代一些低水平（比如仅会CURD）的程序员，但是以目前的进展来说，还有一段路需要走。
- ◆ 长远的未来，人人都具备编程能力，但是程序员不太应该被单独拿出来讨论，实际上对大模型来说，编程仍然是最具挑战的任务之一。

LLM应用开发专有名词解释

LLM-大型语言模型

- ◆ LLM是基于深度学习技术构建的人工智能模型，由具有数以亿计参数的人工神经网络组成，通过自监督学习或半监督学习在大量无标签文本上进行训练。
- ◆ LLM于2018年左右出现，并在各种任务上表现出色，LLM改变了自然语言处理研究的重点，使其不再是以训练特定任务的专门监督模型为范式。
- ◆ ChatGPT、文心一言、Kimi等都是基于LLM开发的聊天机器人应用。

AIGC-AI生成内容

- ◆ AIGC (AI-Generated Content) 通过对已有数据进行学习和模式识别，**以适当的泛化能力生成相关内容的技术。**
- ◆ AIGC生成的内容很多，涵盖文字、图像、视频、音频、游戏、虚拟人等等。
- ◆ AIGC是AI大模型，而ChatGPT则是AIGC在聊天对话场景的一个具体应用。

AGI-人工通用智能

- ◆ AGI (Artificial General Intelligence) 全称**人工通用智能**，是指能够理解、学习和应用广泛的知识和技能的人工智能系统。
- ◆ 以人类角度来看，AGI就是一种能够“思考”和“理解”各种问题的智能生命体，就像人类一样。

Agent-智能代理

- ◆ Agent (智能代理) 一个能够**自主感知环境并采取行动的计算实体**，其目标是最大化某种预定义的效用或实现特定的目标。
- ◆ AGI可以看成是一种非常高级的Agent，具备广泛适应性和自我学习能力，Agent也是现阶段AGI的最佳实现方式。

Prompt-提示词

- ◆ Prompt是指给定的一段文本或问题，用于引导和启发人工智能模型生成相关的回答或内容。
- ◆ Prompt是目前人类与LLM大语言模型交互的核心方式。

GPT-生成型预训练变换模型

- ◆ GPT (Generative Pre-trained Transformer) 是一种基于深度学习的**大型语言模型**。
- ◆ GPT模型最初由OpenAI开发，旨在通过训练模型预测下一个单词或字符来学习自然语言的统计规律。

Token-文本基础单元

- ◆ Token是指在自然语言处理和文本处理任务中，将文本分解成较小单元的基本单位。这些单元可以是单词、字符、子词或其他语言单位，具体取决于任务和处理方式。
- ◆ 大语言模型中的上下文长度计算一般都是基于Token，而不是字符，例如GPT-4的16K上下文意味着传递的消息不能超过16K个Token。

LoRA-插件式微调

- ◆ LoRA (Low-Rank Adaptation of LLM) 即插件式微调，用于对大语言模型进行个性化的特定任务的定制。
- ◆ LoRA通过将模型的权重矩阵分解成低秩的相似矩阵，降低了参数空间的复杂性，从而减少微调的计算成本和模型存储要求。

矢量/向量数据库

- ◆ 矢量数据库是一种用于存储**矢量/向量数据**的数据库。
- ◆ 矢量数据库可以存储和管理大量的矢量数据，例如图像、视频、音频、文本等，同时提供高效检索功能。

数据蒸馏

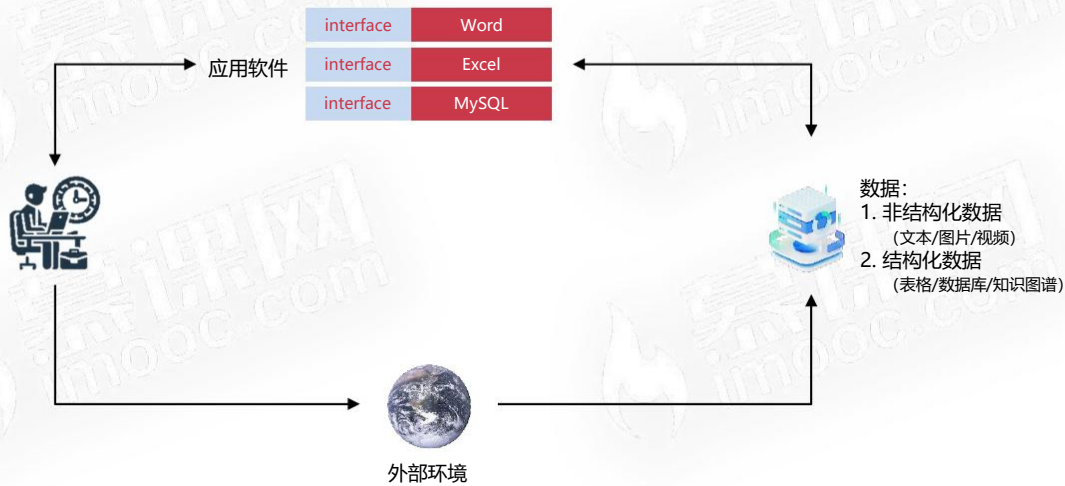
- ◆ 数据蒸馏指将给定的原始大数据集浓缩并生成一个小型数据集，使得在小数据集上训练出来的模型与原数据集上训练的模型相似。
- ◆ 数据蒸馏在深度学习领域被广泛应用，可以帮助将复杂的模型转换成更轻量级的模型，提高模型的鲁棒性和泛化能力。

课程总结

- ◆ 学习了一些常见的LLM应用开发的专有名词，涵盖：LLM、AIGC、AGI、Agent、Prompt、GPT、Token、LoRA、向量数据库、数据蒸馏等。

LLM&AI Agent应用的交互模式

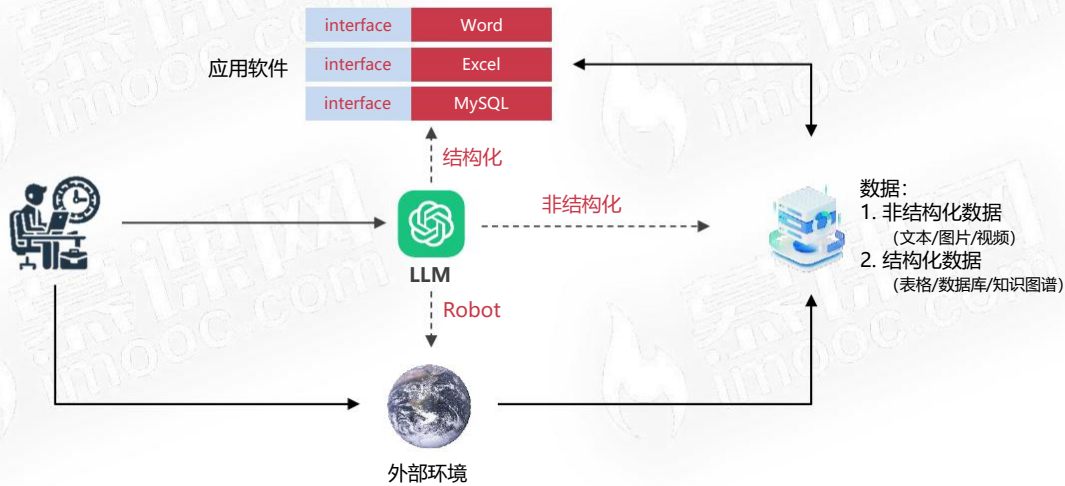
传统人机交互范式



传统人机交互范式的缺点

- ◆ 交互接口**复杂繁杂且多样化**，要使用多一款软件，要专门花时间去学习，每个软件之间接口差异巨大。
- ◆ 人在环境中进行一些行为，会产生各类的数据，涵盖结构化和非结构化数据，而人不能和数据直接发生关系，需要一个中介，这个中介就是软件。
- ◆ 处理不同的数据需要不同的软件，而有些软件非专业人士很难上手，比如说PhotoShop。

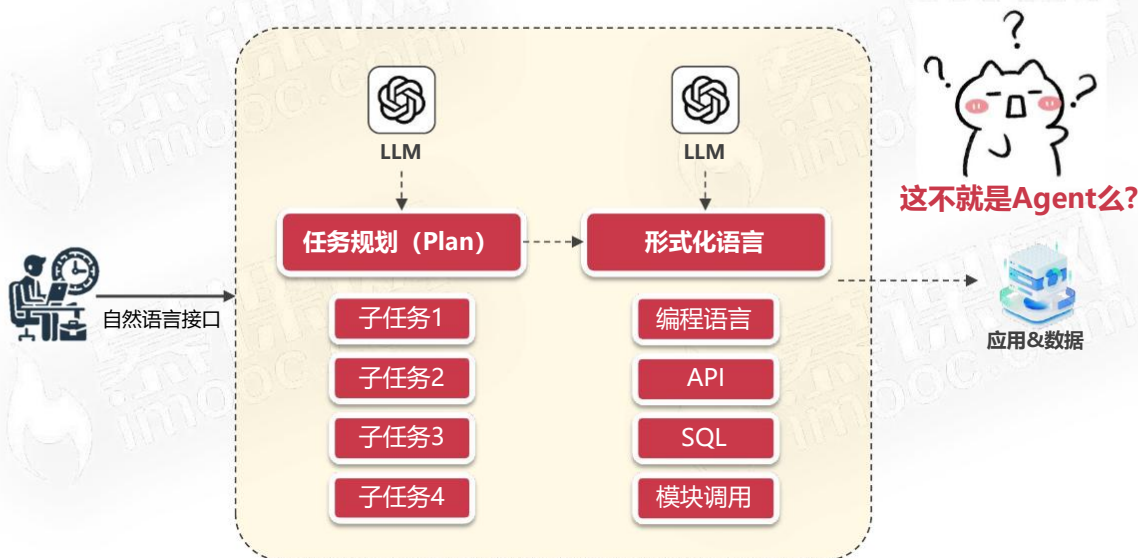
大模型时代的人机交互范式



大模型时代新交互的特点

- ◆ 过去通过某个应用软件与某种数据进行交互，现在变成人和大模型交互，即大语言模型站到了人机交互的**中心位置**。
- ◆ 本质上还是人和数据的关系，只是由于大模型的出现，应用软件被屏蔽到了幕后。
- ◆ 短期来看，LLM可以代替一些应用软件，比如多模态大模型对PhotoShop的取代；长期来看，大模型可能会逐步替代各种功能的软件。

新交互范式下背后发生了什么



结合大模型的新一代应用交互方式

- ◆ **嵌入 (Embedding) 模式**: 用户通过与AI交流, AI协助完成, 如创作小说、音乐、3D内容等, 此模式下, AI是执行工具, 人类是决策者和指挥者。
- ◆ **副驾驶 (Copilot) 模式**: 在这种模式下, 人类与AI作为合作伙伴, 共同完成任务, AI提供建议并协助任务, 二者互补, AI更像知识丰富的伙伴而非工具。
- ◆ **智能体 (Agent) 模式**: 人类设定目标并提供资源, AI独立完成大部分工作, 最后人类监督和评估结果。

课程总结

- ◆ 剖析了传统的人机交互与大模型时代的交互范式的一些差异，并探索了新交互方式的背后发生了什么。
- ◆ 在交互范式的基础上引出Agent的概念，并学习当前主流的3种交互方式：嵌入、副驾驶、智能体等。

章节总结

