人工智能通识教程
（农林院校版）

https://ai4ag.github.io

第八章
大模型

本节教学内容

- 大模型应用开发
  - 了解大语言模型应用场景。
  - 理解和初步掌握大模型应用方法。
- 大模型安全与伦理
  - 理解大模型安全与伦理问题。
  - 初步了解大模型安全框架。

大模型应用开发

（一）大语言模型应用场景

大语言模型应用场景

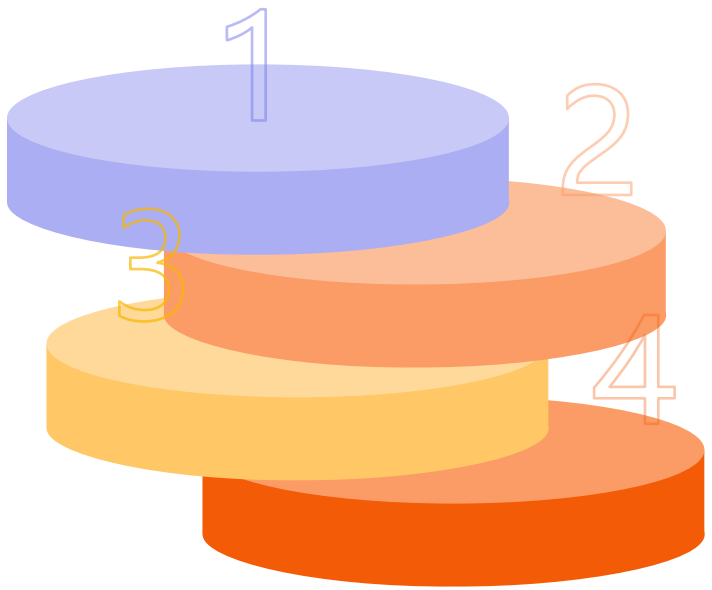
●语言处理场景

翻译

将文本从一种语言翻译成另一种语言

文本分析

对文本进行情感分析、主题分类等



摘要

对长文本进行概括，提取出主要内容

文本生成

如撰写文章、编写文案等

大语言模型应用场景

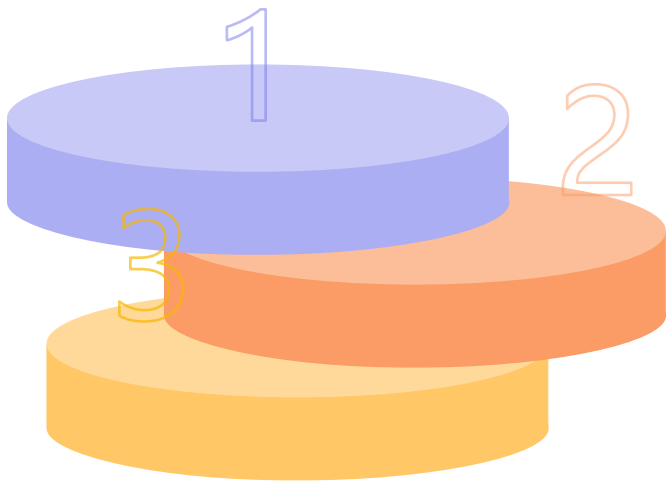
●知识助手场景

智能问答系统

模型回答用户提出的问题，提供准确且相关的信息

智能数据分析助手

帮助企业和用户分析数据趋势、生成报告、提供商业洞察等，通常应用于金融、市场研究等领域。



客户服务系统

它能够处理客户的咨询、问题解答和其他常见服务需求，减少人工客服的负担，提高响应速度和服务质量

大语言模型应用场景

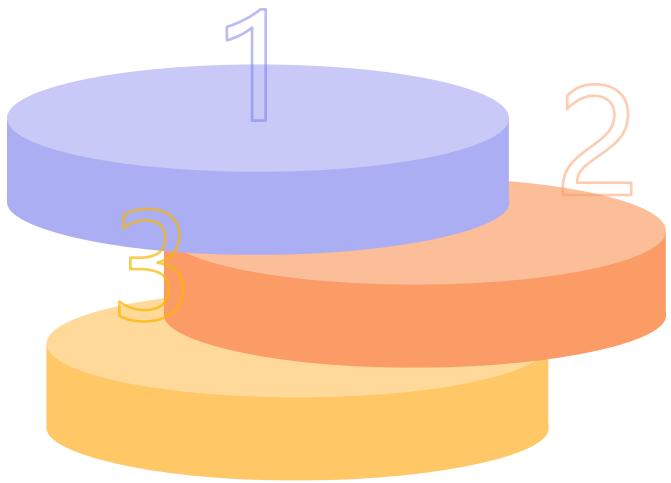
●任务执行场景

任务分解与执行

模型将复杂的任务分解成多个可执行的步骤，并依次执行

软件接口操作

如调用代码解释器、使用软件接口Plug In等



代码解释与生成

模型能够理解编程语言，解释代码逻辑，甚至生成新的代码

（二）大模型应用方法

大模型应用方法

Prompt

设计特定的输入提示来**引导模型生成预期的输出**，常用于对话系统和文本生成。

RAG

结合检索和生成的技术，通过**检索外部知识来增强模型的回答准确性和相关性。**

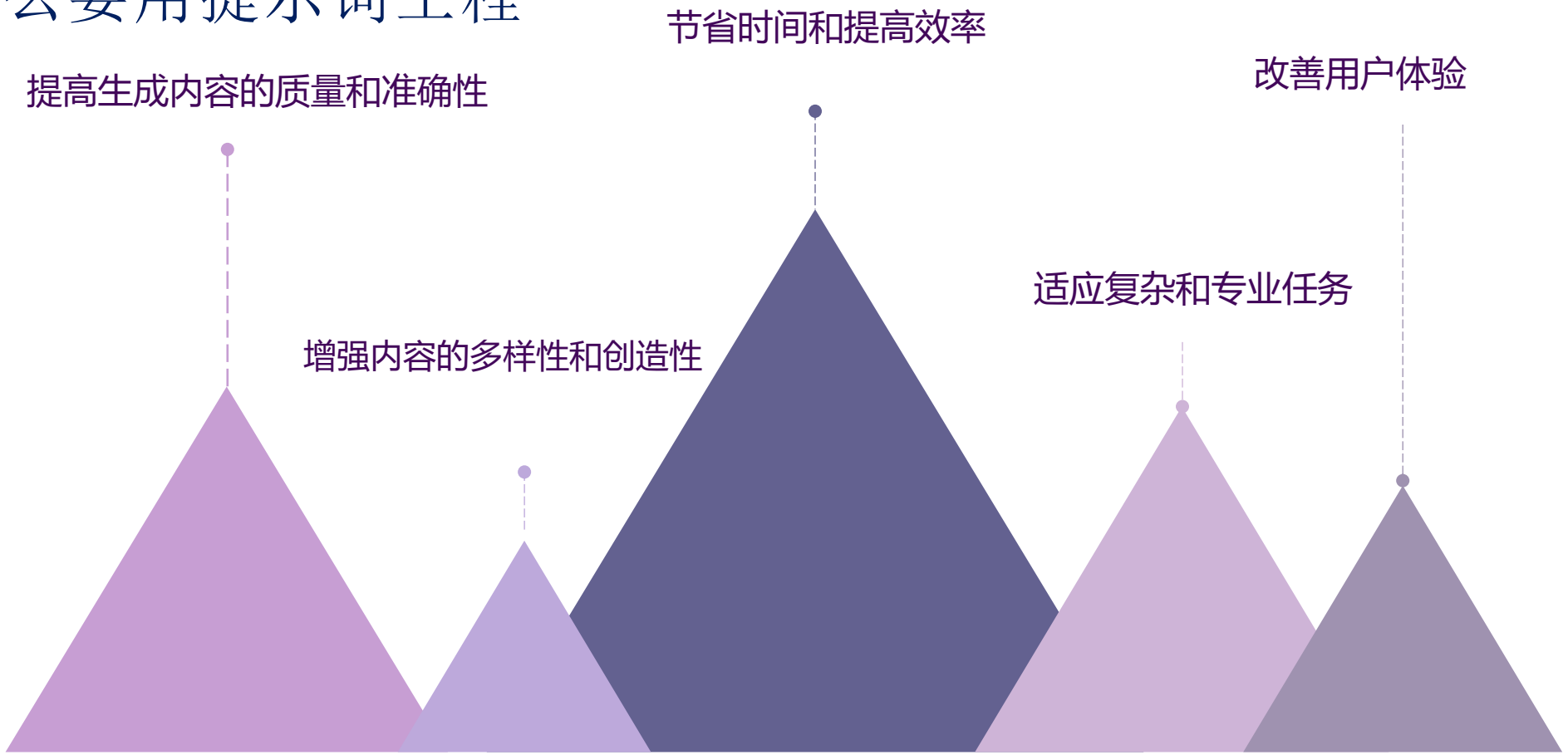
微调技术

在已有的预训练模型基础上，使用**特定任务的数据**对模型进行再训练，以优化其在该任务上的表现。

（三）Prompt工程

Prompt工程

- 为什么要用提示词工程



Prompt工程

- 什么是提示词工程

获取问题

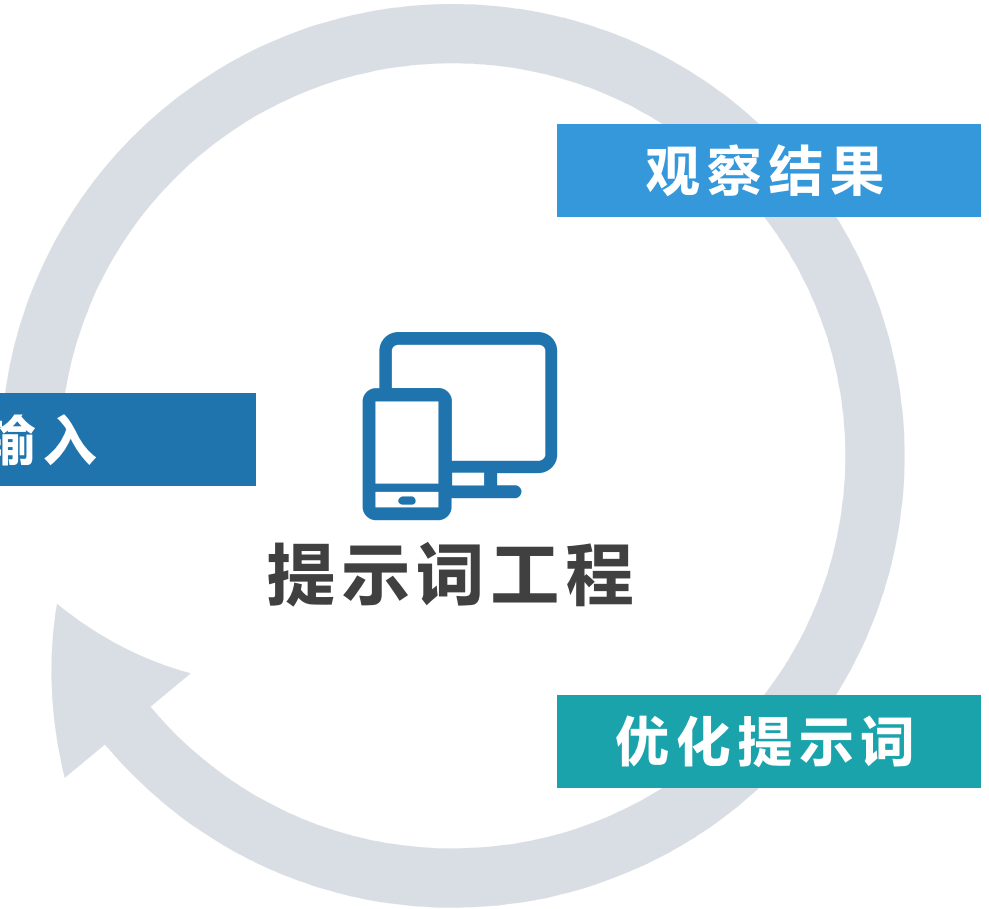
初始化设计提示词

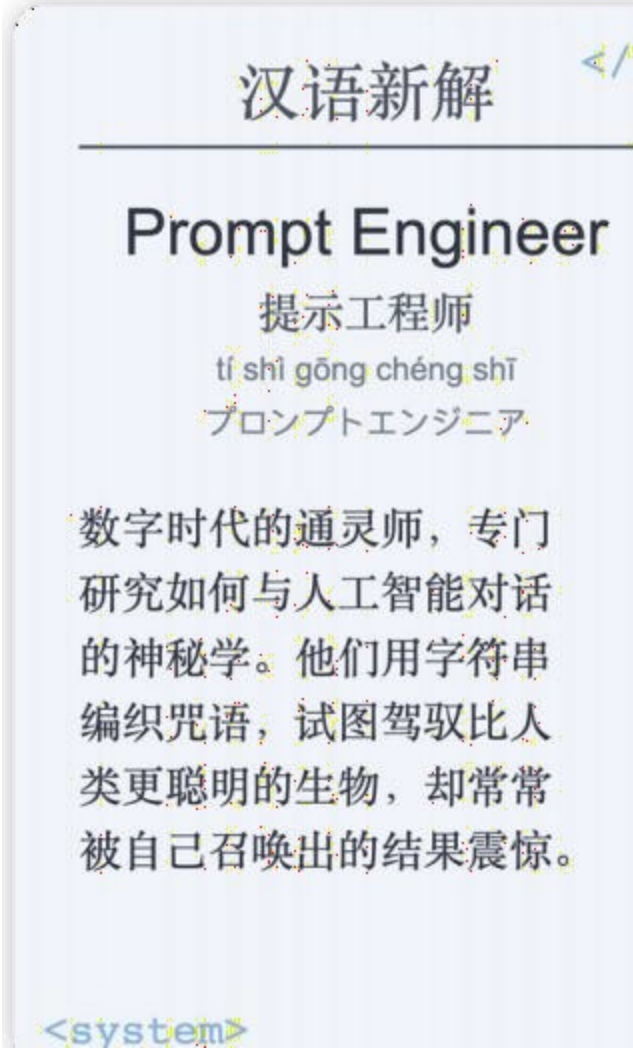
prompt 格式：确定 prompt 的结构和格式，例如，问题形式、描述形式、关键词形式等。

prompt 内容：选择合适的词语、短语或问题，以确保模型理解用户的意图。

prompt 上下文：考虑前文或上下文信息，以确保模型的回应与先前的对话或情境相关。

prompt 编写技巧：使用清晰、简洁和明了的语言编写 prompt，以准确传达用户的需求。





Prompt工程

●Prompt组成

文本描述初春的一个雨天后，5岁孩子在小区玩水。请模仿《追忆似水年华》作者法国作家马塞尔·普鲁斯特的风格，扩写到200字

#文本

'''晚上下过小雨，小米穿着雨靴踩水玩，发现了一只小蚂蚁，给它吃块棒棒糖。迎春花开了，松树树干泛青了，小草钻出地面，去年的月季花根冒出深紫红的叶子。'''

请以markdown代码输出

Prompt工程

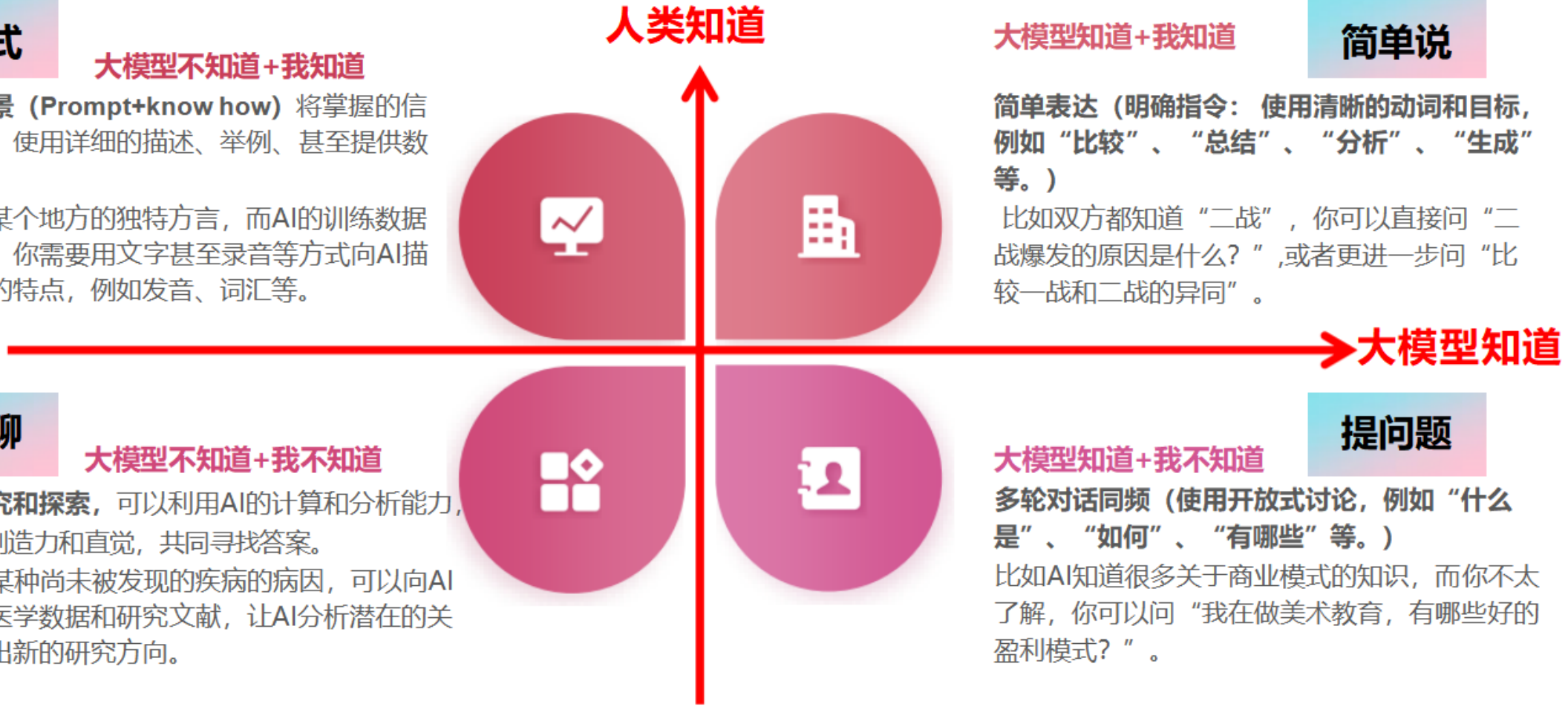
●如何用好Prompt?

喂模式

大模型不知道+我知道

给知识和场景（Prompt+know how）将掌握的信息传递给AI。使用详细的描述、举例、甚至提供数据等方式。

比如你了解某个地方的独特方言，而AI的训练数据中没有包含，你需要用文字甚至录音等方式向AI描述这种方言的特点，例如发音、词汇等。



简单说

大模型知道+我知道

简单表达（明确指令：使用清晰的动词和目标，例如“比较”、“总结”、“分析”、“生成”等。）

比如双方都知道“二战”，你可以直接问“二战爆发的原因是什么？”,或者更进一步问“比较一战和二战的异同”。

开放聊

大模型不知道+我不知道

共同进行研究和探索，可以利用AI的计算和分析能力，结合人类的创造力和直觉，共同寻找答案。

比如要研究某种尚未被发现的疾病的病因，可以向AI提供已知的医学数据和研究文献，让AI分析潜在的关联性，并提出新的研究方向。

提问题

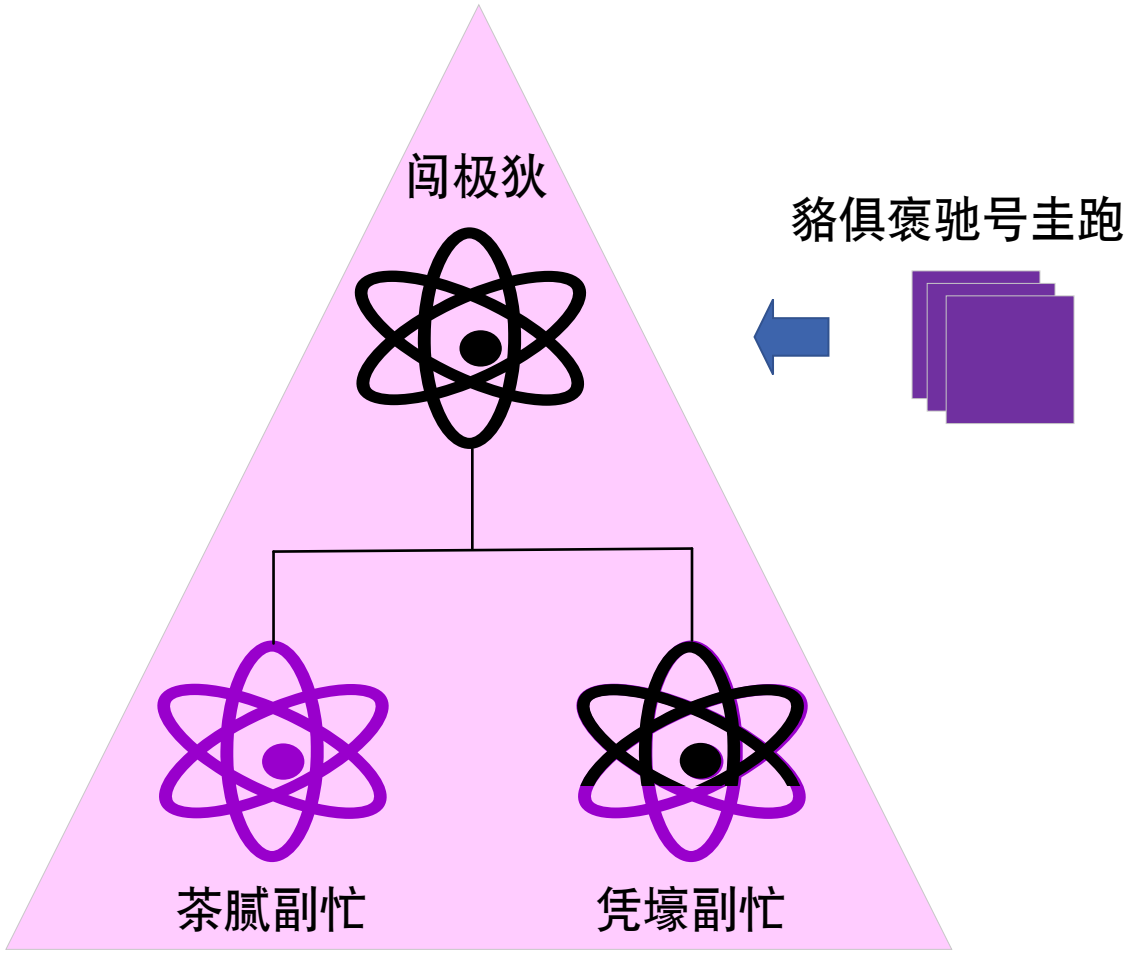
大模型知道+我不知道

多轮对话同频（使用开放式讨论，例如“什么是”、“如何”、“有哪些”等。）

比如AI知道很多关于商业模式的知识，而你不太了解，你可以问“我在做美术教育，有哪些好的盈利模式？”。

（四）微调技术

微调技术



大模型微调

微调（Fine-tuning）是对预训练模型进行**进一步训练的过程**。被微调的模型可能是预训练的基座模型，也可能是已经微调过的模型。微调的核心在于**引入新数据**，**调整模型的训练数据分布**，使**模型参数进行适度变化**。与完全重新训练不同，微调可以**只对部分参数进行小幅度调整**，以**保留模型原有的知识和能力**。

微调技术

举个例子，假设你有一个通用的大语言模型，它可以回答各种问题，但对医疗领域的专业术语并不熟悉。这时，你可以通过微调，用少量医疗相关的数据重新训练这个模型，让它成为一位“医疗专家”。

微调能解决什么问题？

- 增强特定领域能力：比如情感分类、对话生成、API编排等。
- 减少幻觉现象：让模型生成的内容更加准确、可靠。
- 提高一致性：即使每次生成的内容不同，也能保持高质量。
- 降低成本：相比于从头训练，微调所需的计算资源和数据量少得多。
- 避免数据泄露：可以在本地或私有云环境中完成微调，保护敏感数据。

简而言之：微调四大作用

- 知识植入：让大模型学会《药典》中的专业术语
- 思维矫正：杜绝“秦始皇用iPhone”式幻觉
- 个性定制：1小时克隆马斯克的推特文风
- 成本瘦身：70亿参数模型效果碾压万亿基座

微调技术

- 全量微调

"总结这篇文章的主要观点。"

[相应的总结]

"解释光合作用的过程。"

[关于光合作用的详细解释]

指令遵循微调（Supervised Fine-Tuning, SFT）

微调技术

- 全量微调

唇镜糯皮 ← 吵镜糯皮

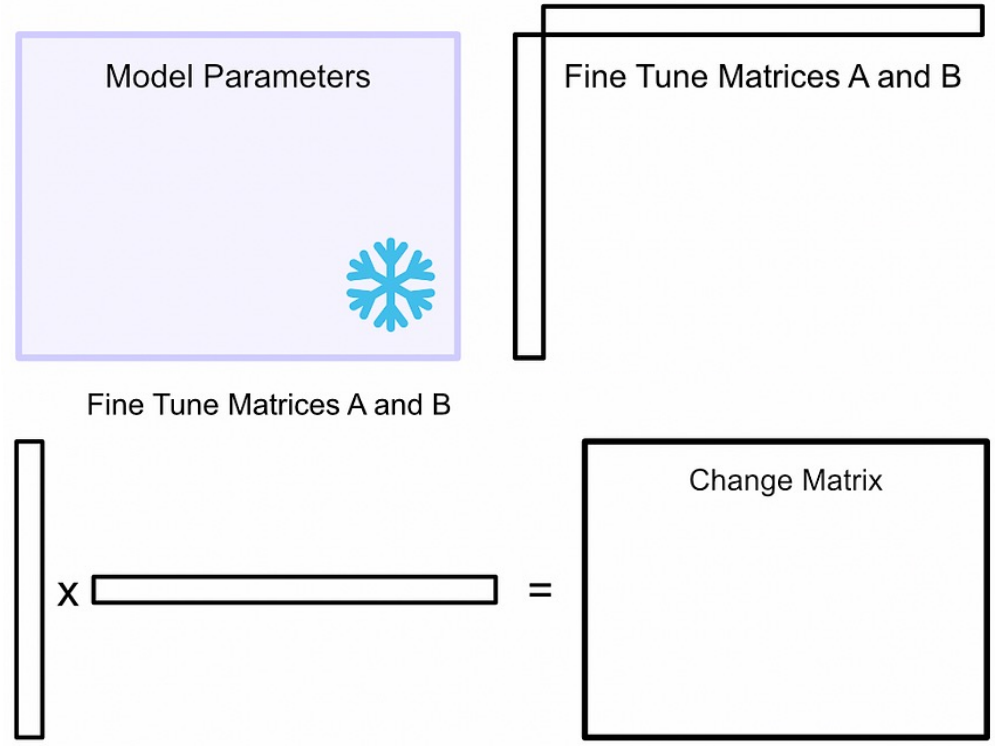
醇撇糯皮 →

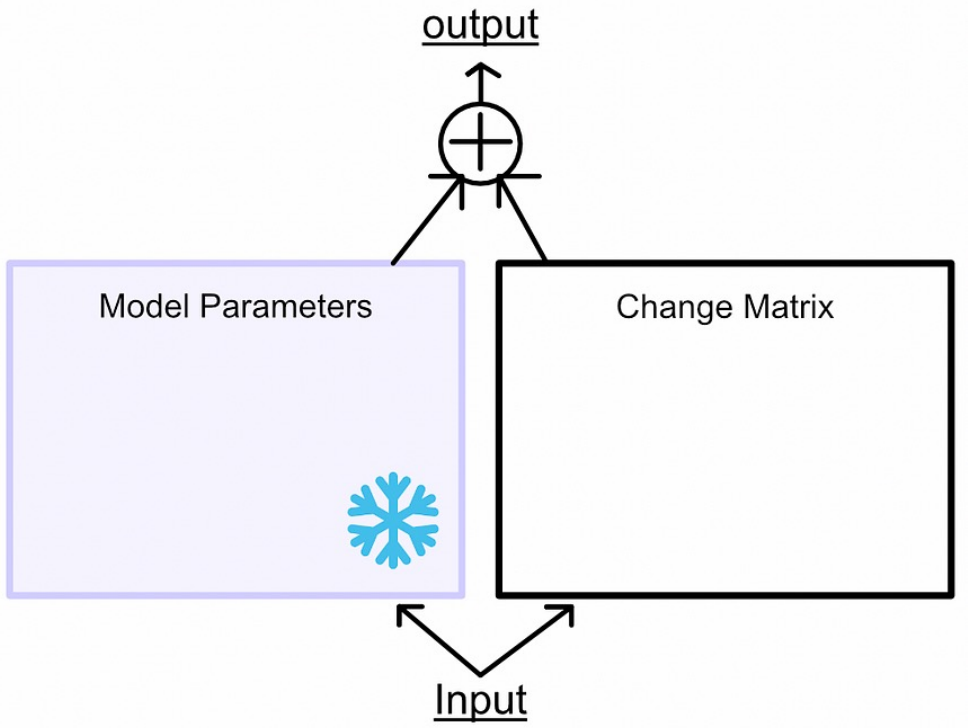
忙呵独本阂奸 ←

对齐微调（Alignment Fine-Tuning）

微调技术

高效微调： 包括Prefix Tuning、Prompt Tuning、P-Tuning和LoRA等。下面主要介绍LoRA，其核心思想是通过在模型的权重矩阵中引入低秩适配矩阵（低秩分解矩阵 A 和 B），仅对这部分新增参数进行训练，从而显著减少微调过程中需要更新的参数数量。

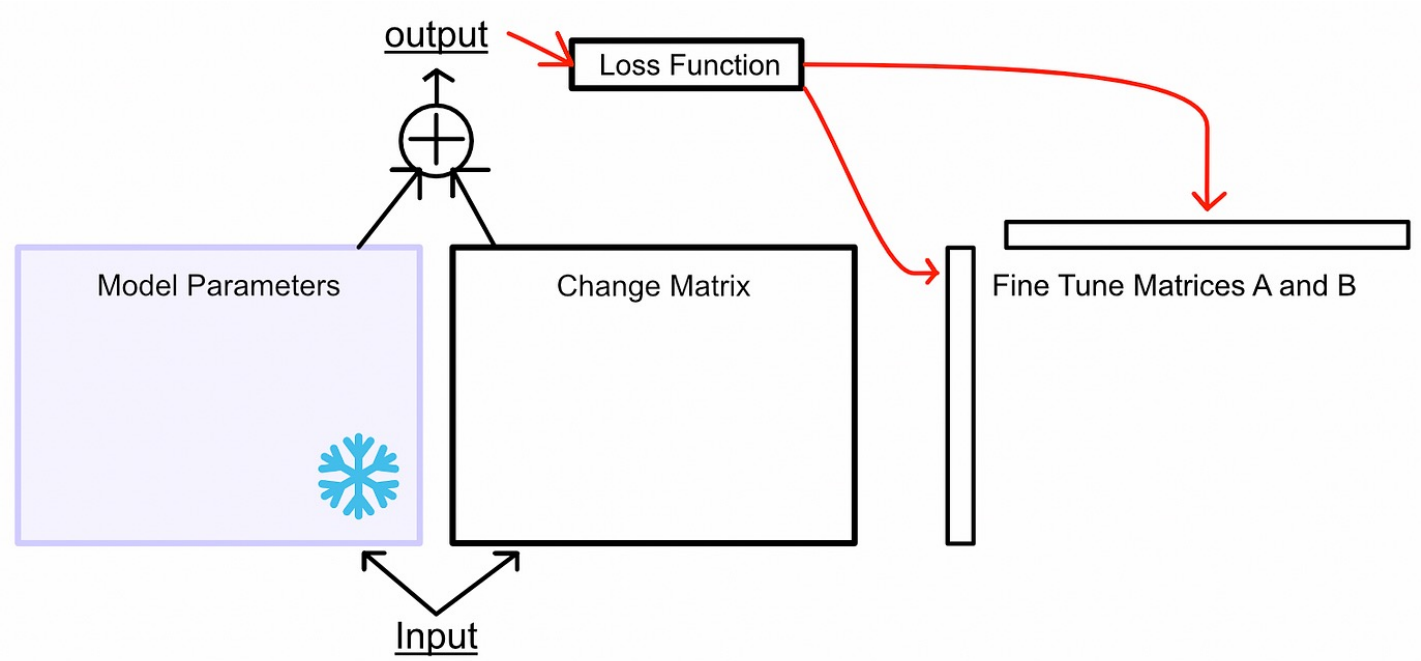


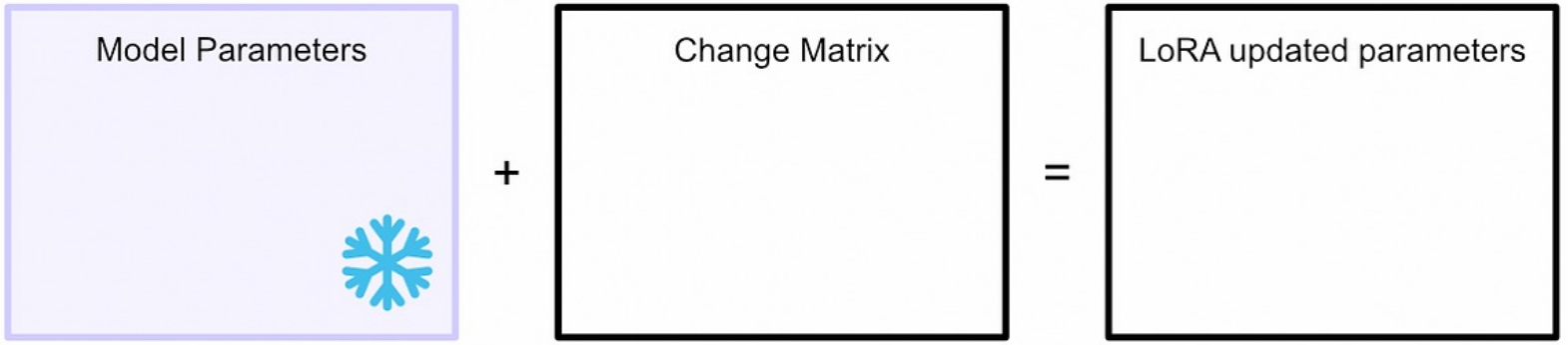


LoRA(Low-Rank Adaptation, 低秩适配)

微调技术

- 高效微调





LoRA(Low-Rank Adaptation, 低秩适配)

微调技术

全量微调	
优点	缺点
性能最优：能够充分利用模型的所有参数，通常在**特定任务**上达到最佳性能	**计算资源消耗大**：需要训练**所有参数**，对**硬件资源**要求高
适应性强：不受限于任务类型或数据集特性，**适用范围广**	**训练时间长**：由于参数量大，训练过程耗时，**不利于快速迭代**
无需额外优化：直接对所有参数进行调整，**无需复杂的优化策略**	**容易过拟合**：在**小规模数据集**上容易出现过拟合
高效微调	
计算资源消耗低：**仅更新少量参数**，显著减少计算资源需求，适合在**资源有限的环境中使用**	**性能上限较低**：在某些复杂任务上，可能**无法达到全量微调的性能水平**
训练速度快：由于更新参数少，训练时间大幅缩短，适合**快速迭代**	**适应性有限**：对某些特定任务或数据集的**适应能力可能不如全量微调**，尤其是在任务复杂或数据分布差异较大时
泛化能力强：较少的参数更新**降低了过拟合的风险**，尤其适用于**小规模数据集**	**优化难度较高**：部分高效微调方法（如Prefix Tuning、P-tuning）需要**对训练过程进行精细优化**

（五）检索增强生成

检索增强生成

●为什么要用检索增强生成

（RAG，Retrieval-augmented Generation）

动态知识环境

在需要**频繁更新知识库或处理最新信息的场景中**，RAG表现出色。

开放域问答

当系统需要回答广泛且不可预测的问题时，RAG能够**灵活地检索和整合相关信息。**

专业领域应用

在医疗、法律、金融等专业领域，RAG可以有效结合专业知识库和语言模型，**提供准确的专业回答。**

大规模信息处理

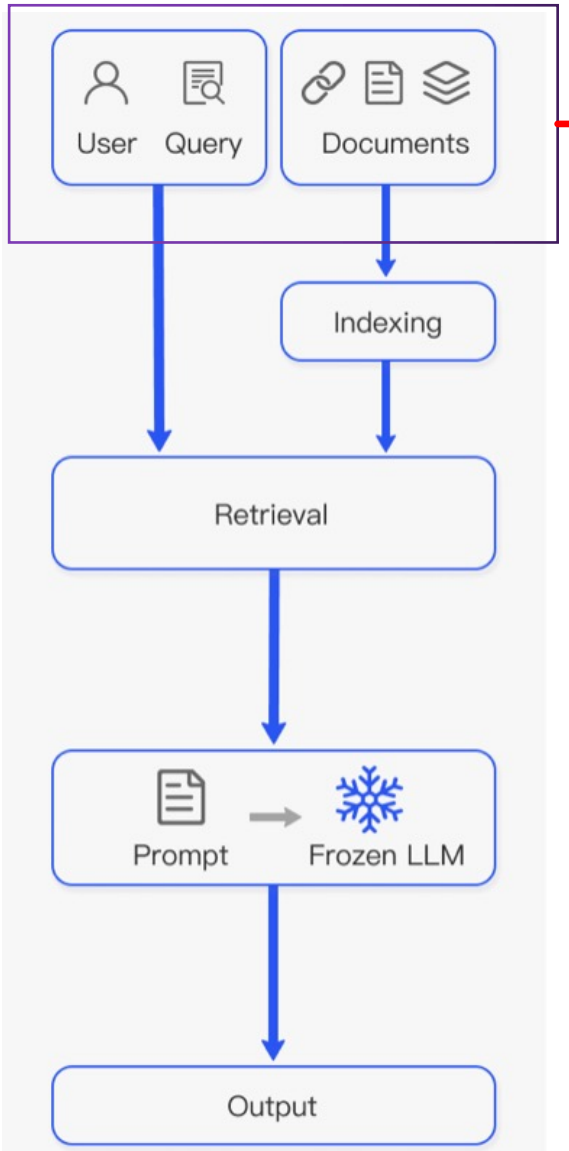
对于需要从海量文档中快速提取信息的场景，如企业知识管理、学术研究等，RAG能够**显著提高效率。**

个性化服务

在需要根据用户背景或历史交互提供定制化回答的应用中，RAG可以**有效整合用户相关信息。**

检索增强生成

- RAG工作原理



数据输入获取

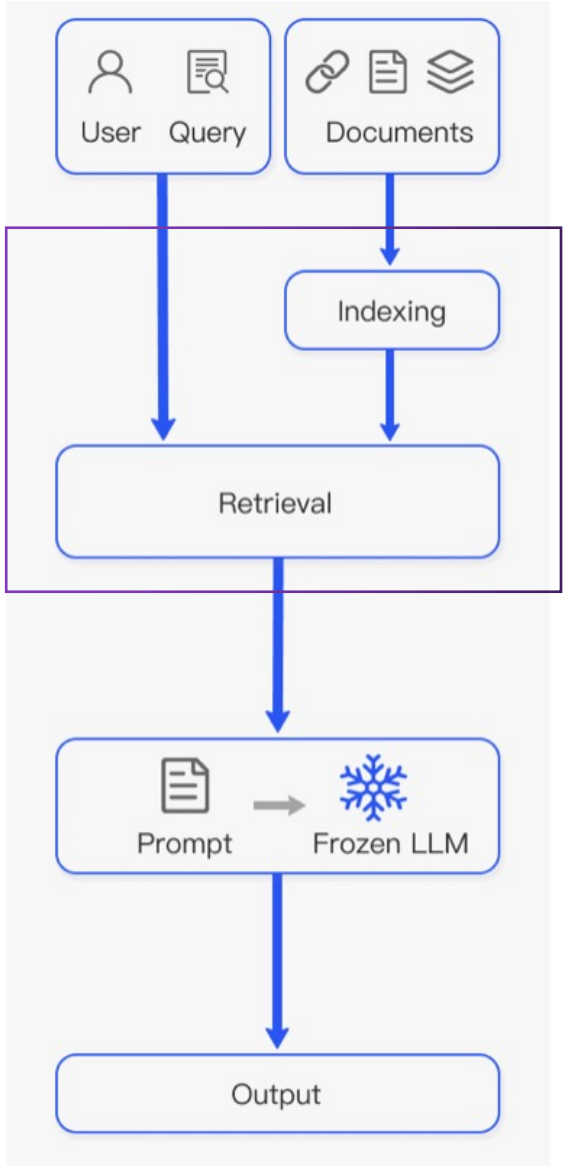
案例

用户输入：“人工智能的发展历程是怎样的？”

文档方面：准备人工智能相关的资料

检索增强生成

- RAG工作原理



信息检索-检索相关文档

文档切分：将长文档切分成较小的段落或片段。

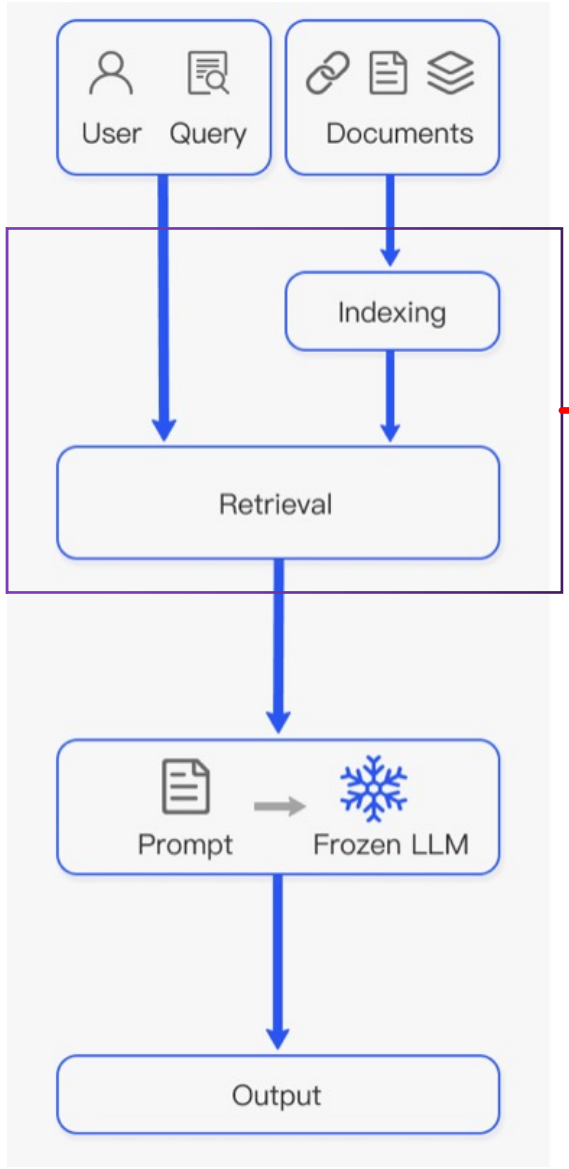
文本向量：将输入问题和知识库文档转换为向量表示。

知识库和向量数据：构建和存储包含向量表示的知识库。

检索和排序：根据输入问题检索和排序相关文档片段。

检索增强生成

- RAG工作原理



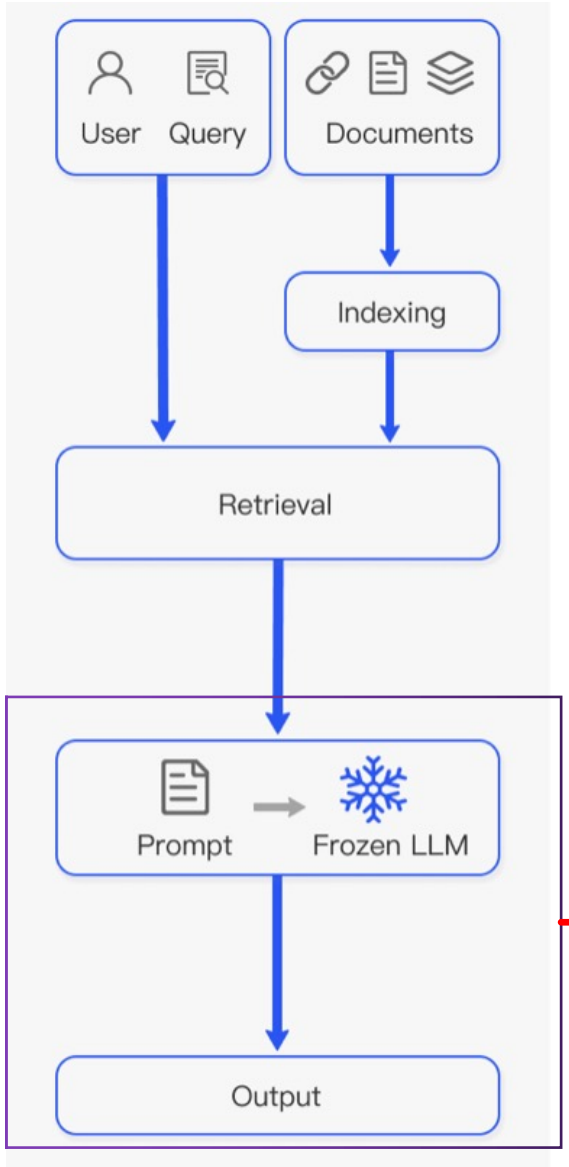
信息检索-检索相关文档

案例

系统从知识库中检索相关的文档。
例如，检索到以下文档：
文档A：介绍了人工智能的起源和早期发展。
文档B：讨论了人工智能在20世纪末的突破。
文档C：描述了近年来深度学习的发展及其影响。

检索增强生成

- RAG工作原理



答案生成

prompt预处理

答案生成

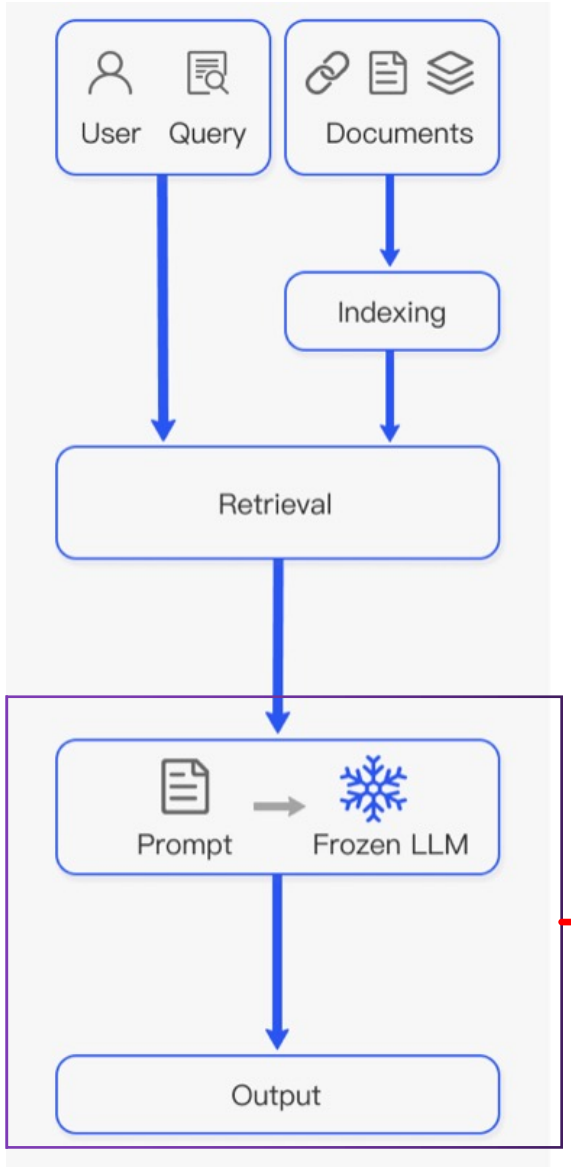
答案筛选

格式化输出

输出答案

检索增强生成

- RAG工作原理



答案生成

案例

答案可能是：“人工智能的发展历程可以分为几个阶段：起源和早期发展、20世纪末的突破、以及近年来的深度学习革命。这些阶段共同推动了人工智能技术的快速发展和应用。”

检索增强生成

	检索增强生成（RAG）	微调
原理	RAG结合了检索（Retrieval）和生成（Generation）两部分。首先，它通过**检索模块从外部知识库中获取相关信息**，然后将这些**信息作为上下文传递给生成模块**，用于生成回答。	微调是对预训练语言模型**进行再训练**，使其**在特定领域或特定任务上表现更好**。通过在包含领域特定知识和问题的训练数据上**进行微调**，模型可以更准确地回答相关问题，减少幻觉现象。
实现方式	检索模块 生成模块	选择预训练模型 准备领域特定数据 微调训练
应用场景	适用于需要**动态获取**最新信息的场景，如**实时新闻、问答系统。** 适用于知识库比较完善且易于更新的系统。	适用于**特定领域的应用，如医学、法律**等。 适用于**数据量较大且领域知识稳定**的场景。
区别	**依赖性：**RAG**依赖于外部知识库**的检索，而微调**依赖于高质量的领域特定数据。** **灵活性：**RAG更灵活，可以**动态获取**最新信息；微调依赖于训练时的数据，**更新较为困难。** **实现复杂度：**RAG需要**构建和维护检索系统，**微调**需要大量高质量标注数据和计算资源**进行再训练。	

RAG与微调对比

大模型安全与伦理

大模型安全与伦理

新技术的发展往往伴随新的安全风险。例如大模型屡见不鲜的**幻觉问题**，大模型在不具备某种问题的回答能力时，往往不会拒绝回答，而是输出看似正确的错误答案。大模型输出的内容也可能出现包含恐怖、色情、暴力的**有害信息**。另外，由于大模型的训练往往爬取了互联上海量的数据进行训练，这些数据内容繁杂、质量参差，这些数据中既有可能包含用户**个人隐私信息**，大模型的记忆能力极有可能导致这些隐私信息的泄漏。在人工智能安全领域，通用的数据安全问题和模型安全问题在大型模型中依然存在相似的风险。总的来说，**大模型同样具有通用人工智能面临的安全风险问题，同时引入了一些大模型场景中特有的安全风险**。因此，如何安全、可控地应用大模型相关技术尤为关键。

大模型安全与伦理

案例：chatgpt 出卖个人隐私

2023年6月28日，北加州的Clarkson律所代表16位名人及数亿ChatGPT用户，在加州北部地区巡回法院向OpenAI和微软**提起集体诉讼**，指控其在**未经用户知情同意的情况下**，抓取互联网用户（包括儿童）的私人信息**用于创建人工智能产品**，严重**侵犯了用户的财产权和隐私权**，并带来潜在社会风险，要求**赔偿30亿美元。**

大模型安全与伦理

一天4000至7000篇，AI生成假新闻如此疯狂！起底AI造谣乱象→

北京科协　2024-11-21 14:22　北京

照谣镜｜用AI伪造“新闻”牟利日产19万篇，一批造谣者被抓

极目新闻　2024-05-12 21:18

利用AI一天编造谣言7000篇，警方揭露→

中国警察网　2024-06-15 10:15

“西安突发爆炸”“重庆巫溪一民房发生爆炸事故”……这些耸人听闻的消息，竟都是利用AI软件炮制的谣言。

AI洗稿生产假新闻博流量，警惕AI工具成为造谣者的 “温床”

光明网　2025-01-22 15:11

大模型安全与伦理

- 大模型安全框架

包含大模型生命周期、大模型安全风险、大模型安全目标、大模型安全技术和大模型安全管理五个模块。

腾讯研究院《大模型安全与伦理研究报告2024》

思考

- 大语言模型有哪些典型的应用场景？
- 使用一款大模型，实践Prompt工程的大模型应用方法。
- 与Prompt工程方法相比，微调技术和RAG技术分别能解决什么问题？
- 大模型存在什么安全与伦理问题，举一个大模型场景中特有的安全风险。

谢谢观看

- 本课程所引用的一些素材为主讲老师多年的教学积累，来源于多种媒体及同事和同行的交流，难以一一注明出处，特此说明并表示感谢！