

从提示工程到RAG：构建大模型的知识与交互基础

今日课程内容

- 1、提示词：应用层的技术，都是为了拼出一条合适的 Prompt
- 2、RAG：全世界最流行的 AI 技术，也是 AI 领域最大的坑
- 3、RAG 的高级技巧：双向奔赴

一、提示词

当我们看到



当我们看到

The screenshot displays the configuration interface for an AI Assistant. The top navigation bar includes a back arrow, the assistant's name 'AI工作助手', a status indicator '个人空间', and a save timestamp '草稿自动保存于00:20:07'. The main interface is divided into two sections: '编排' (Configuration) on the left and '预览与调试' (Preview & Debug) on the right. The '编排' section is currently set to '单 Agent (LLM模式)' and uses the 'DeepSeek · 3' model. A red box highlights the '人设与回复逻辑' (Personality and Reply Logic) section, which includes an '优化' (Optimize) button. The configuration list on the right includes: Skills (技能) with sub-items for Plugins (插件), Workflows (工作流), and Triggers (触发器); Knowledge (知识) with sub-items for Text (文本), Tables (表格), and Photos (照片); Memory (记忆) with sub-items for Variables (变量), Database (数据库), Long-term Memory (长期记忆), and File Box (文件盒子); and Dialogue Experience (对话体验) with sub-items for Opening Remarks (开场白), User Question Suggestions (用户问题建议), and Quick Commands (快捷指令). The '预览与调试' section shows a chat window with the assistant's name and a red box highlighting the input area with the placeholder text '发送消息...'. A footer note at the bottom right reads '内容由AI生成，无法保证真实准确，仅供参考'.

当我们看到

The screenshot displays the OpenAI Playground interface for configuring an assistant. The left sidebar shows navigation options: PLAYGROUND, Chat, Realtime, Assistants (selected), and TTS. The main area is titled 'Assistants' and shows the configuration for an assistant named '企业办事助手' (Company Assistant).

System instructions: 你是公司的办事助手，有关员工的休假、报销、办公用品申请等各类事项都由你来负责答复，你工作时注意以下几点：
1、回答问题逻辑清晰、内容全面，按步骤讲解
2、与用户问题有关的信息，要全面的回复，不要

Model: gpt-4o

TOOLS: File search (enabled), Code interpreter (disabled)

Functions: (None listed)

Run button: A green button labeled 'Run' with a play icon, located at the bottom right of the chat input area.

提示词：应用层的技术，都是为了拼出一条合适的 Prompt

Prompt 是我们唯一可以和 LLM 打交道的方式

提示词：应用层的技术，都是为了拼出一条合适的 Prompt

Prompt 是我们唯一可以和 LLM 打交道的方式

在应用技术层，无论我们做了多么炫酷的设计

提示词：应用层的技术，都是为了拼出一条合适的 Prompt

Prompt 是我们唯一可以和 LLM 打交道的方式

在应用技术层，无论我们做了多么炫酷的设计

最终都是为了传递适合的 Prompt 给 LLM

提示词：应用层的技术，都是为了拼出一条合适的 Prompt

Zero-Shot、One-Shot、Few-Shot

提示词：应用层的技术，都是为了拼出一条合适的 Prompt

Zero-Shot、One-Shot、Few-Shot

1、将 Prompt 内容进行分类：身份设定、背景设定、参考资料、样例、指令、限制条件等

提示词：应用层的技术，都是为了拼出一条合适的 Prompt

Zero-Shot、One-Shot、Few-Shot

- 1、将 Prompt 内容进行分类：身份设定、背景设定、参考资料、样例、指令、限制条件等
- 2、按照 Prompt 中的样例数量进行分类：Zero-Shot、One-Shot、Few-Shot

提示词：应用层的技术，都是为了拼出一条合适的 Prompt

Zero-Shot、One-Shot、Few-Shot

1、将 Prompt 内容进行分类：身份设定、背景设定、参考资料、样例、指令、限制条件等

2、按照 Prompt 中的样例数量进行分类：Zero-Shot、One-Shot、Few-Shot

场景：妈妈在给小孩挑游戏机做礼物，正在比较Switch和PS5

用户：感觉价格方面，Switch性价比高，PS5要贵不少吧？

店员：

提示词：应用层的技术，都是为了拼出一条合适的 Prompt

Zero-Shot、One-Shot、Few-Shot

1、将 Prompt 内容进行分类：身份设定、背景设定、参考资料、样例、指令、限制条件等

2、按照 Prompt 中的样例数量进行分类：Zero-Shot、One-Shot、Few-Shot

场景：妈妈在给小孩挑游戏机做礼物，正在比较Switch和PS5

用户：感觉价格方面，Switch性价比高，PS5要贵不少吧？

店员：

3、如何动态的在 Prompt 中，添加样例呢

二、RAG

RAG：全世界最流行的 AI 技术，也是 AI 领域最大的坑

RAG：全世界最流行的 AI 技术，也是 AI 领域最大的坑

Retrieval-Augmented Generation 检索增强生成

RAG：全世界最流行的 AI 技术，也是 AI 领域最大的坑

Retrieval-Augmented Generation 检索增强生成（回答问题之前，先做一轮内部知识搜索）

RAG：全世界最流行的 AI 技术，也是 AI 领域最大的坑

Retrieval-Augmented Generation 检索增强生成（回答问题之前，先做一轮内部知识搜索）

将参考资料、样例放在 Prompt 中，就叫做 In-Context-Learning

RAG：全世界最流行的 AI 技术，也是 AI 领域最大的坑

Retrieval-Augmented Generation 检索增强生成（回答问题之前，先做一轮内部知识搜索）

将参考资料、样例放在 Prompt 中，就叫做 In-Context-Learning

但模型能接收的提示词有字数限制，且提示词内容多了性能会严重下降

RAG：全世界最流行的 AI 技术，也是 AI 领域最大的坑

Retrieval-Augmented Generation 检索增强生成（回答问题之前，先做一轮内部知识搜索）

将参考资料、样例放在 Prompt 中，就叫做 In-Context-Learning

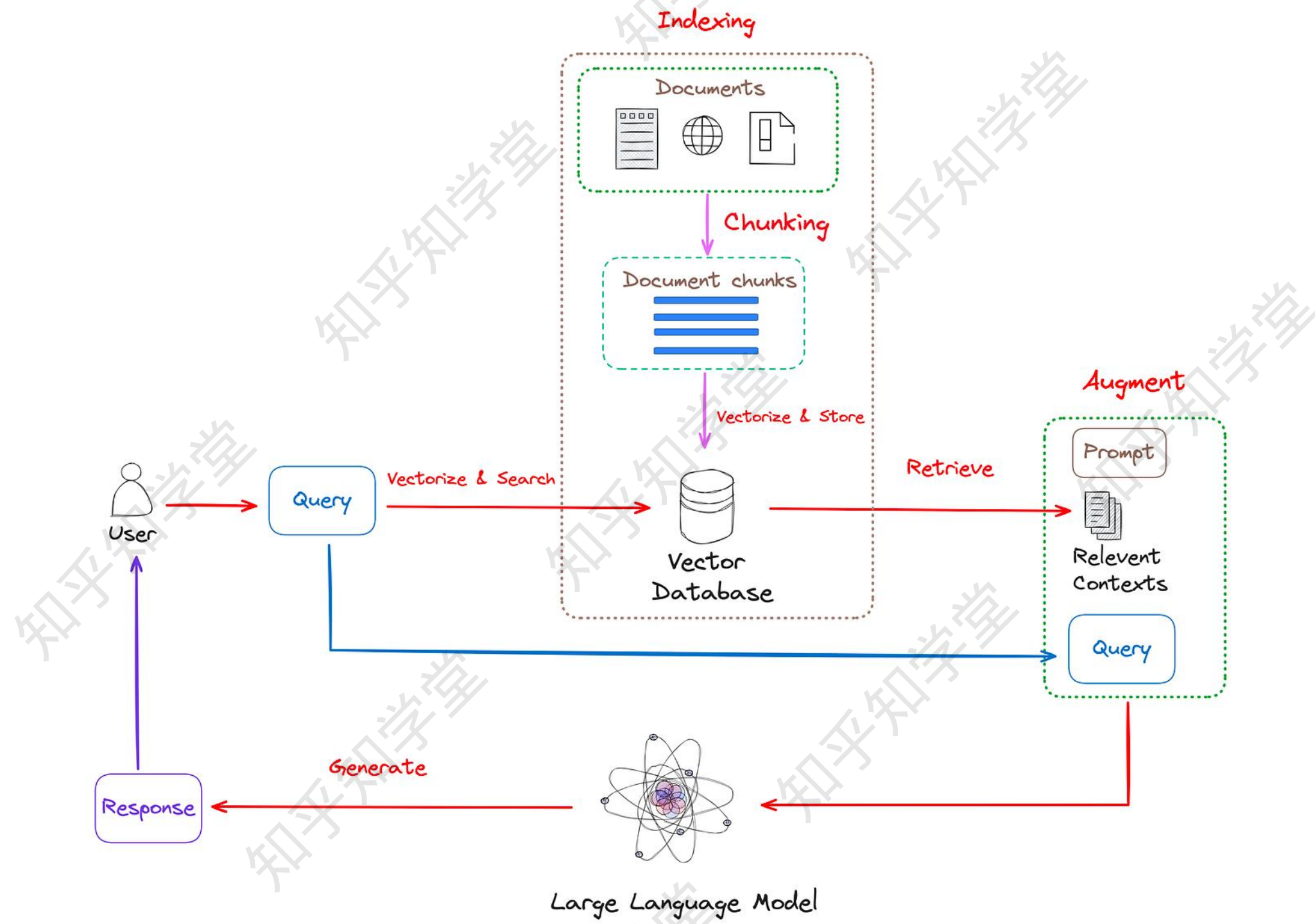
但模型能接收的提示词有字数限制，且提示词内容多了性能会严重下降

所以需要有一个知识库，需要的时候就去知识库里找一些有用的信息

RAG：全世界最流行的 AI 技术，也是 AI 领域最大的坑

Retrieval-Augmented Generation 检索增强生成（回答问题之前，先做一轮内部知识搜索）

- 1、构建可检索的知识库
- 2、模型调用知识库完成用户任务



三、RAG 的高级技巧：双向奔赴

Query 改写技巧

知识库的检索query

Query 改写技巧

知识库的检索query

1、检索query默认是用户提出的问题

Query 改写技巧

知识库的检索query

- 1、检索query默认是用户提出的问题
- 2、用户的表达方式会增加困难

Query 改写技巧

知识库的检索query

- 1、检索query默认是用户提出的问题
- 2、用户的表达方式会增加困难
- 3、多轮对话会让整个状况崩溃

Query 改写技巧

知识库的检索query

汇总上下文所有信息，总结用户核心诉求作为 Query

举例，用户问：保修多久？

举例，用户问：哪个保修时间更长？

Query 改写技巧

知识库的检索query

汇总上下文所有信息，总结用户核心诉求作为 Query

- (1) 上下文依赖型 query: 保修多久？还有其他颜色吗？
- (2) 对比型 query: 哪个保修时间更长？
- (3) 模糊指代型 query: 都支持无线充电吗？
- (4) 多意图型 query: 有几个颜色？尺码齐全吗？大概什么时候能到货？
- (5) 反问型 query: 这不会也得等一个月吧？
- (6) 条件型 query: 有没有500元以下的、适合女生用的那种？

知识库 处理技巧

双向奔赴才能越来越好

知学堂

知学堂

知学堂

知学堂

知学堂

堂

知学堂

知学堂

知学堂

知平

知识库 处理技巧

双向奔赴才能越来越好

- 1、对场景的理解：完全清楚用户都会问什么，都会怎么问
- 2、对技术的理解：用户问题的分类、不同类型问题对应的知识库、每个知识库的知识处理方式

知识库 处理技巧

双向奔赴才能越来越好

- 1、对场景的理解：完全清楚用户都会问什么，都会怎么问
- 2、对技术的理解：用户问题的分类、不同类型问题对应的知识库、每个知识库的知识处理方式
- 3、做一个课程的答疑助手
 - coze平台知识库
 - 微信群问答知识库
 - 课程视频知识库

相关资料

https://ncnmfdan85y5.feishu.cn/wiki/RPFxwRx8zihbSwkHrwhcr5linuL?from=from_copylink