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

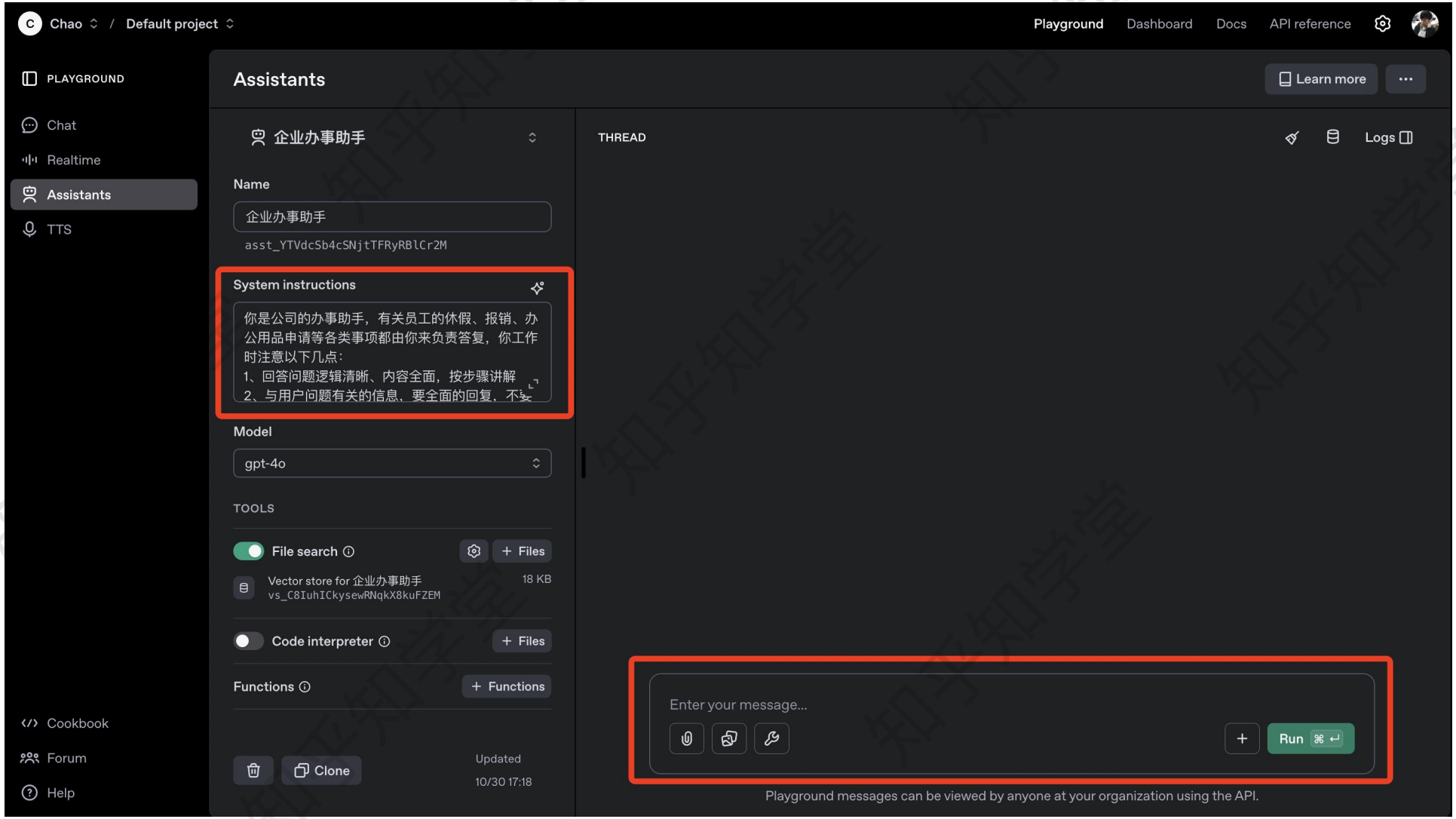
一、提示词

当我们看到

当我们看到


AI工作助手
个人空间
草稿自动保存于00:20:07
编排
分析
发布
编排
单 Agent （LLM模式）
DeepSeek·3
预览与调试
技能
调试
人设与回复逻辑
优化
技能
插件
工作流
触发器
知识
自动调用
文本
表格
照片
记忆
变量
数据库
长期记忆
关闭
文件盒子
关闭
对话体验
开场白
用户问题建议
开启
快捷指令
AI工作助手
发送消息...

当我们看到


Chao / Default project
Playground
Dashboard
Docs
API reference
PLAYGROUND
Chat
Realtime
Assistants
TTS
Cookbook
Forum
Help
Assistants
Learn more
企业办事助手
Name
企业办事助手
asst_YTVdcSb4cSNjtTFRyRBlCr2M
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
+ Files
Vector store for 企业办事助手
vs_C8IuhICkysewRNqkX8kuFZEM
18 KB
Code interpreter
+ Files
Functions
+ Functions
Clone
Updated
10/30 17:18
THREAD
Logs
Enter your message...
Run
Playground messages can be viewed by anyone at your organization using the API.

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

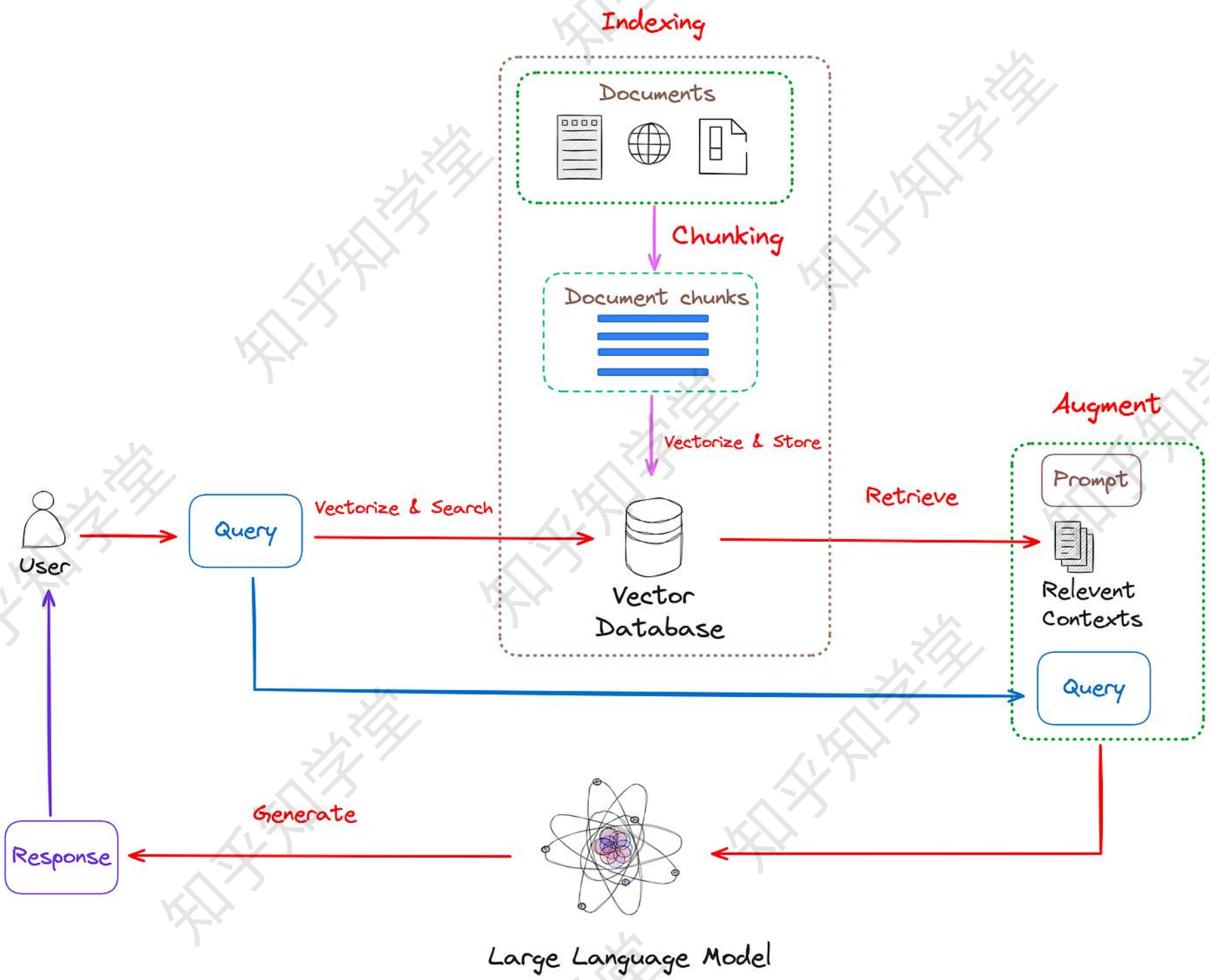
所以需要一个知识库，需要的时候就去知识库里找一些有用的信息

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


Indexing
Documents
Chunking
Document chunks
Vectorize & Store
Augment
Prompt
User
Query
Vectorize & Search
Vector Database
Retrieve
Relevent Contexts
Query
Response
Generate
Large Language Model

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