

大语言模型结合企业知识库的 私有化部署

黄鸿波 | 鼎盛方圆科技

科技生态圈峰会 + 深度研习



—1000+ 技术团队的选择



 **K+峰会**  **敦煌站**

K+ 思考周®研习社

时间: 2025.08.29-30

 **K+峰会**  **上海站**

K+ 金融专场

时间: 2025.09.26-27

 **K+峰会**  **香港站**

K+ 思考周®研习社

时间: 2025.11.17-18



K+峰会详情



 **AiDD峰会**  **上海站**

AI+研发数字峰会

时间: 2025.05.23-24

 **AiDD峰会**  **北京站**

AI+研发数字峰会

时间: 2025.08.08-09

 **AiDD峰会**  **深圳站**

AI+研发数字峰会

时间: 2025.11.14-15



AiDD峰会详情



黄鸿波

鼎盛方圆科技 CEO

鼎盛方圆科技联合创始人

机器学习方向谷歌开发者专家

前西山居AI技术专家

2018年出版《TensorFlow进阶指南 基础、算法与应用》一书

目录

CONTENTS

1. 为什么要用私有化大模型部署
2. 企业知识库的构建思路
3. 实际落地的方案

PART 01

为什么要用私有化大模型部署

▶ 为什么要用私有化大模型部署



给你一个提纲，根据给出的参考文档，请帮我生成一段口岸物流的需求文档，只能用我给你的提纲，不能增加或者减少，我给你的提纲是：整体流程、设备信息、软件信息、设备管理模块

需求文档：

整体流程：

口岸物流整体流程包括货物进口、报关、通关、货物出口等环节。在货物进口环节，需要检验货物信息，完成报关手续；在通关环节，需要对货物进行查验、核对，完成相关清关手续；在货物出口环节，需要与供应商对接，安排货物装运。整体流程需要高效、便捷地完成，确保货物的及时准确流通。

设备信息：



口岸物流需要使用一系列设备，包括货物托盘、装卸设备、货物追踪器等。这些设备需要具备高效、稳定、安全的特点，以确保口岸物流的顺畅运行。

软件信息：

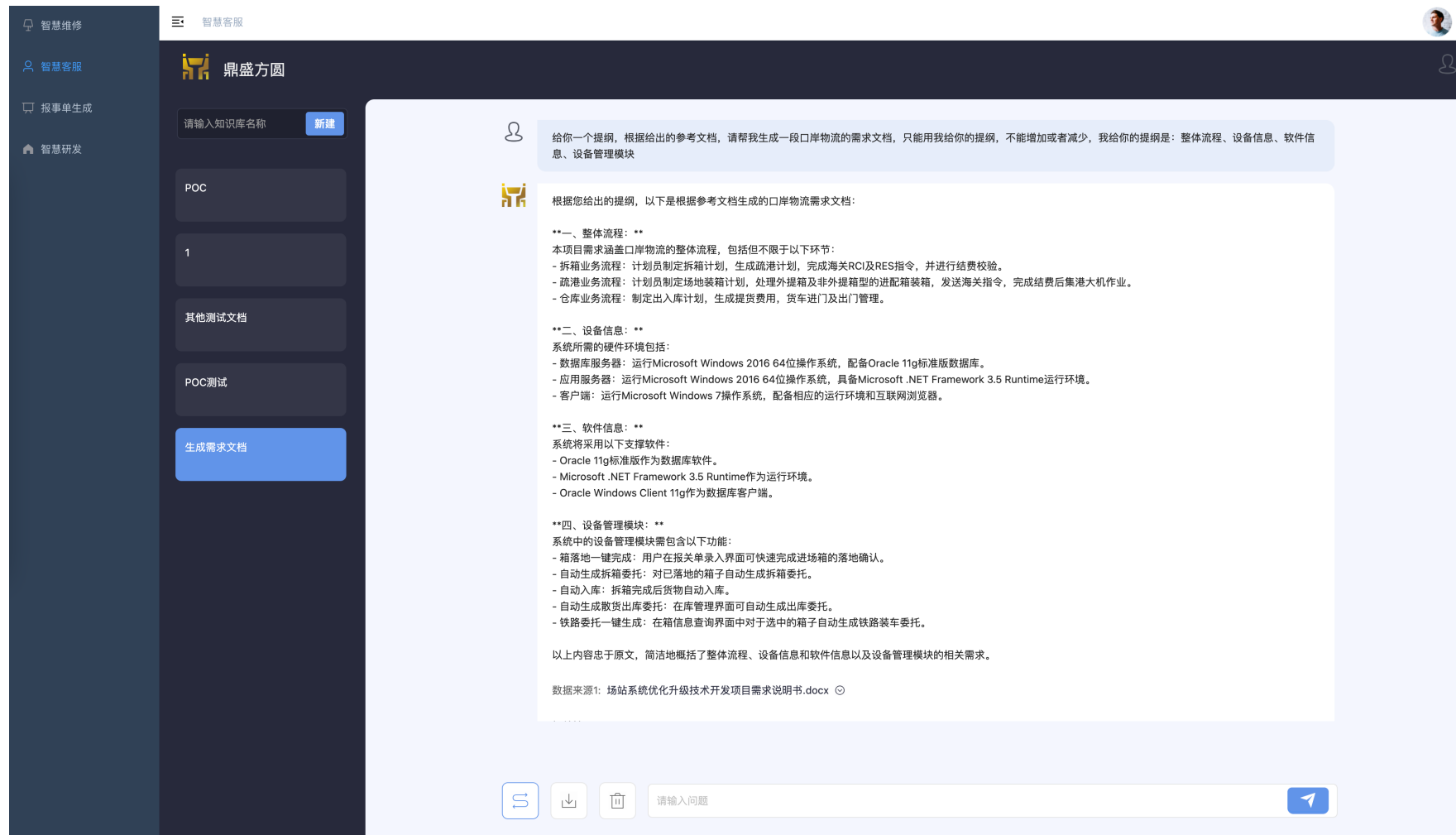
口岸物流需要使用软件对整体流程进行管理和协调。软件需要包括货物信息管理、报关系统、通关系统、货物追踪系统等功能模块，以实现口岸物流的信息化管理和智能化运作。

设备管理模块：

设备管理模块需要实现对口岸物流所需设备的管理和监控。包括设备状态监测、维护保养记录、设备使用情况统计等功能，以保障口岸物流设备的正常运行和高效利用。

口岸物流的需求文档需要综合考虑整体流程、设备信息、软件信息和设备管理模块等方面，以实现口岸物流的高效运作和信息化管理。

为什么要用私有化大模型部署



▶ 为什么要用私有化大模型部署

一、行政赔偿范围的概念	634
二、行政赔偿范围的含义	634
三、行政赔偿范围的制约因素	635
四、行政赔偿范围的确定方式和设定标准	636
第三十八章	
第一节 概述	683
一、行政补偿的含义与特征	684
二、行政补偿的种类	685
第二节 行政补偿的理论基础	687
（一）公共负担平等说	687
（二）结果责任说	687
（三）特别牺牲说	687
（四）社会保险说	688
第三节 行政补偿的范围与方式	688
一、行政补偿的范围	688
二、行政补偿的方式	689
第四节 行政补偿的标准与程序	690
一、行政补偿的标准	690
二、行政补偿的程序	691
附录 参考文献	696
一、最高人民法院公报（1985—2010）和最高人民	696
二、国内行政法与行政诉讼法教科书、专著和案例评析的选读书目 [2]	702
三、国外行政法著作和我国学者编著的关于外国行政法著作的选读书目	714

《行政法》目录

▶ 为什么要用私有化大模型部署



《行政法》有多少个章节



《行政法》是研究行政机关的组织结构、职权、行使程序、行政行为法律规范。根据我所知，中国《行政法》通常指的是《中华人民共和国行政诉讼法》。《中华人民共和国行政诉讼法》共分为八章，包括总则、行政诉讼的管辖、起诉、审理、证据、审判、执行等内容。每一章节又包括多个条款，具体的章节和条款可以根据具体版本进行查阅。如果您需要详细了解《行政法》的章节情况，建议您参考相关法律条文或咨询法律专业人士。

公有化大模型

▶ 为什么要用私有化大模型部署



《行政法》有多少个章节



根据您给出的参考信息，关于《行政法》的章节数目，最相关的信息是：

“本书的体系在前两部教材的基础上作了较大修改，最后确定为六编三十八章。”

因此，根据这个信息，可以得出《行政法》一书的章节总数是三十八章。

数据来源1: 行政法与行政诉讼法(第5版)-姜明安.docx

相关性: 0.701171875

数据来源2: 行政法与行政诉讼法(第5版)-姜明安.docx

相关性: 0.65576171875

数据来源3: 行政法与行政诉讼法(第5版)-姜明安.docx

相关性: 0.6513671875

重新生成



结合企业知识库的私有化大模型

公有化模型和私有化模型的区别

数据的安全性

私有化部署的大语言模型可以在用户自己的服务器或者私有云上运行，这意味着数据的安全性更高，因为数据不需要离开用户的控制范围。这一点对于那些管理着敏感信息的企业或组织来说无疑是极其重要的。

公有化部署的大语言模型通常运行在开发商提供的云服务上，用户的数据会在使用过程中传输到这些服务器上进行处理。

同时，用户在使用公有化部署的大语言模型时，提供的数据可能被用于改进模型训练和优化，虽然这些数据在大多数情况下都会被匿名化处理，但无法确保绝对不会出现数据泄露的情况。

私有化知识

私有化部署的大语言模型可以更好地针对特定场景进行优化和定制化，可以针对企业或组织的特定需求进行定制和优化。例如，可以针对特定的行业术语进行训练，以提高模型的有效性和准确性。

公有云部署的大语言模型通常是通用型的，模型预训练的数据覆盖广泛，适应各种常见场景，但对于特定行业或特定任务的精确性可能欠缺。因为这种模型的训练数据和参数通常并没有针对特殊需求进行优化调整，它可能无法理解特定领域的复杂术语或符合特定业务需求的工作逻辑。

数据的合规性

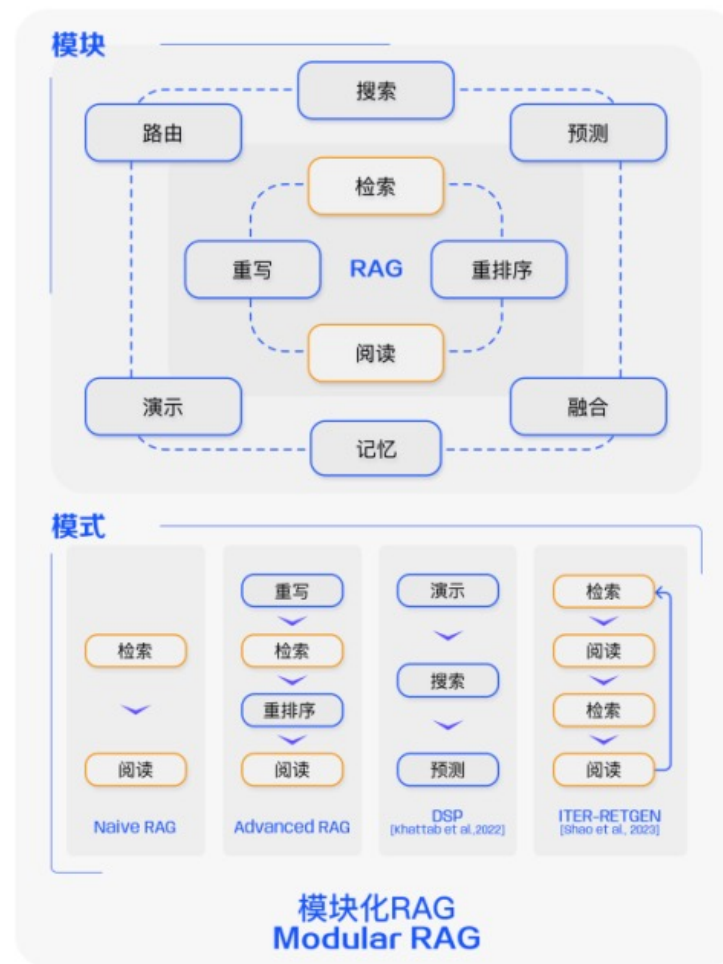
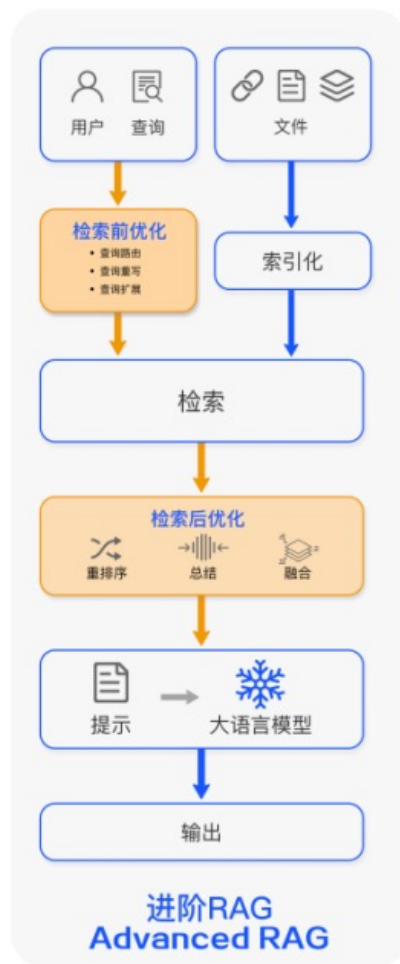
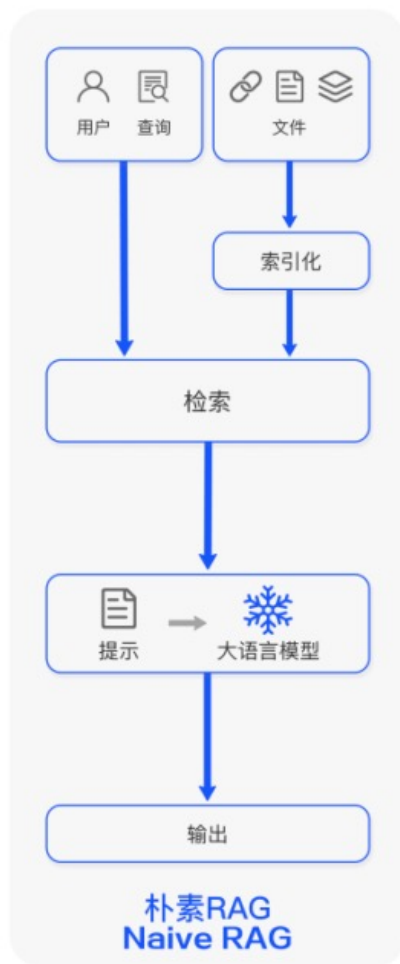
通过私有化部署的方式更好地管理和控制数据获取和使用，以确保符合各类数据隐私和合规性要求。

公有化部署的大语言模型在数据合规性方面可能面临挑战，因为模型运营商通常无法确保其数据处理和存储符合所有地区和行业的特定数据合规要求。在不同国家和地区，对于数据存储、处理和传输的法规要求可能存在显著差异。

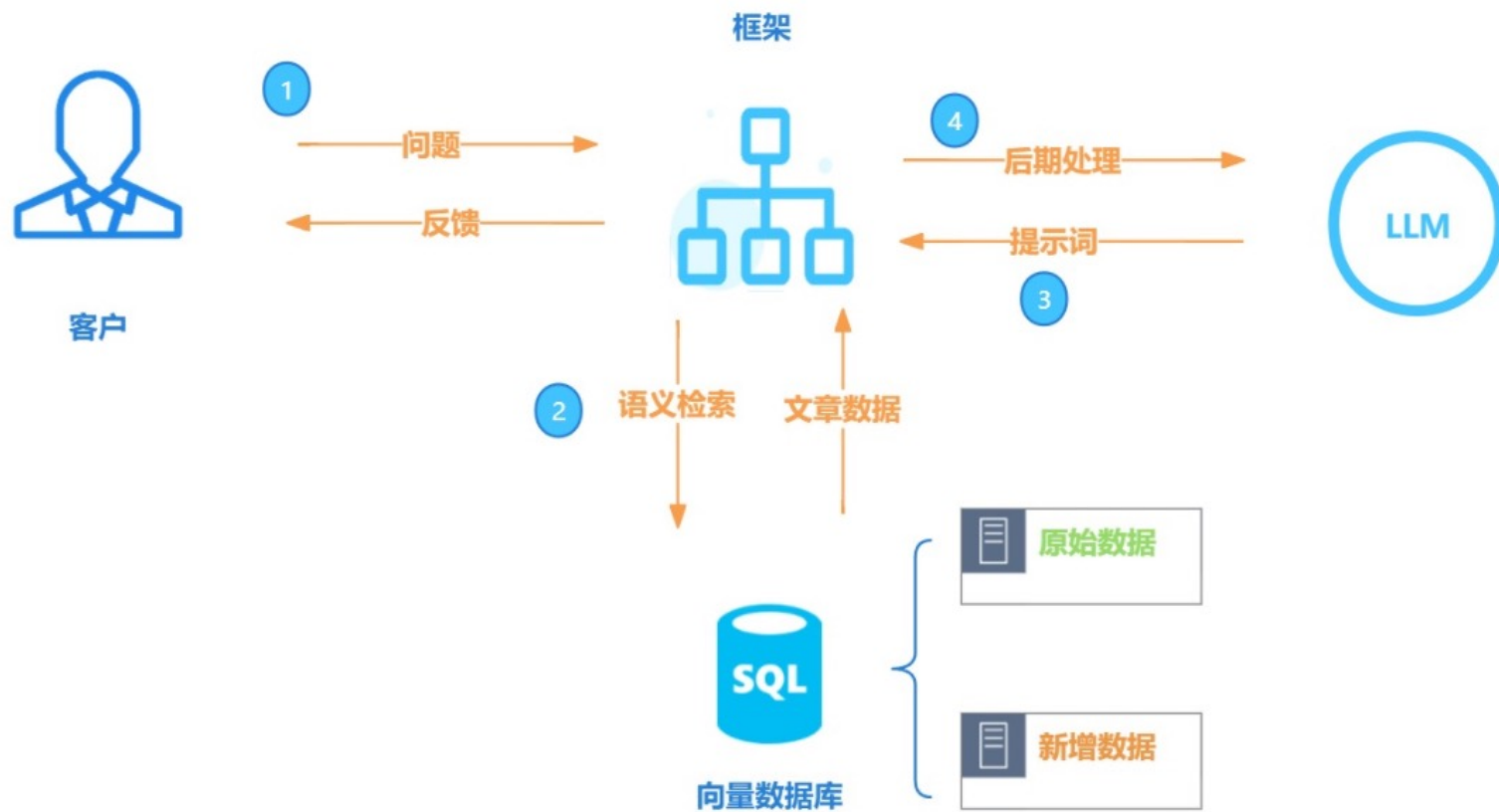
PART 02

企业知识库的构建思路

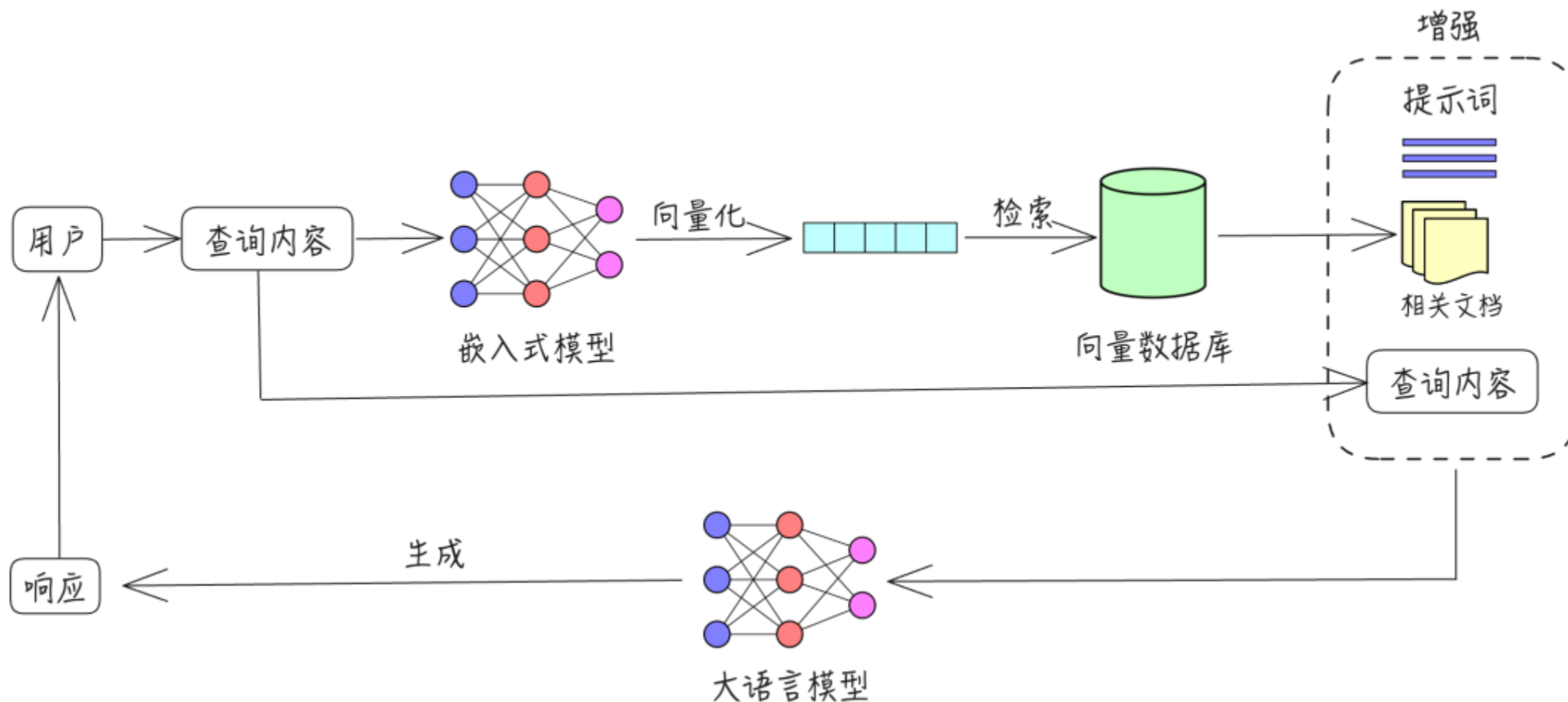
▶ 企业私有化大模型部署方案



▶ 企业私有化大模型部署方案

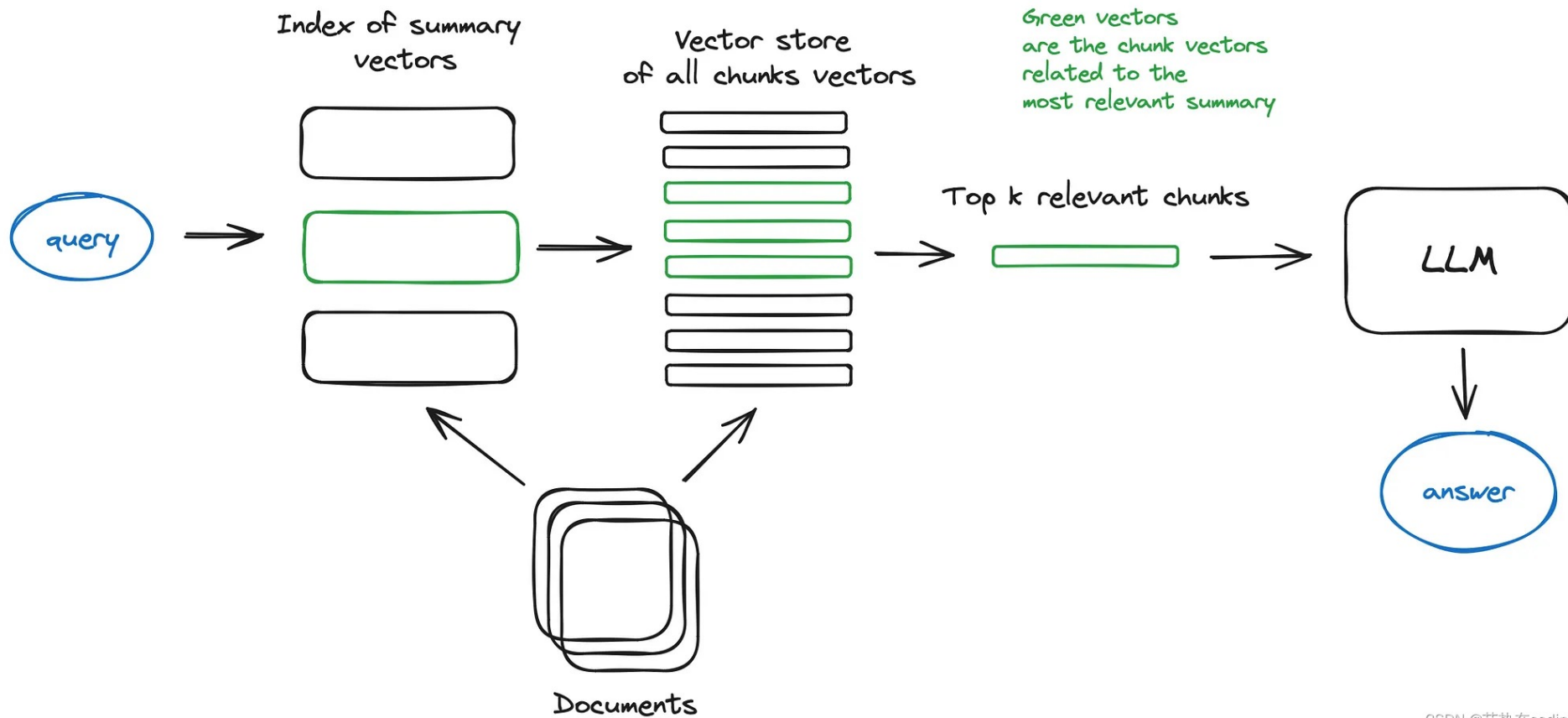


企业私有化大模型部署方案



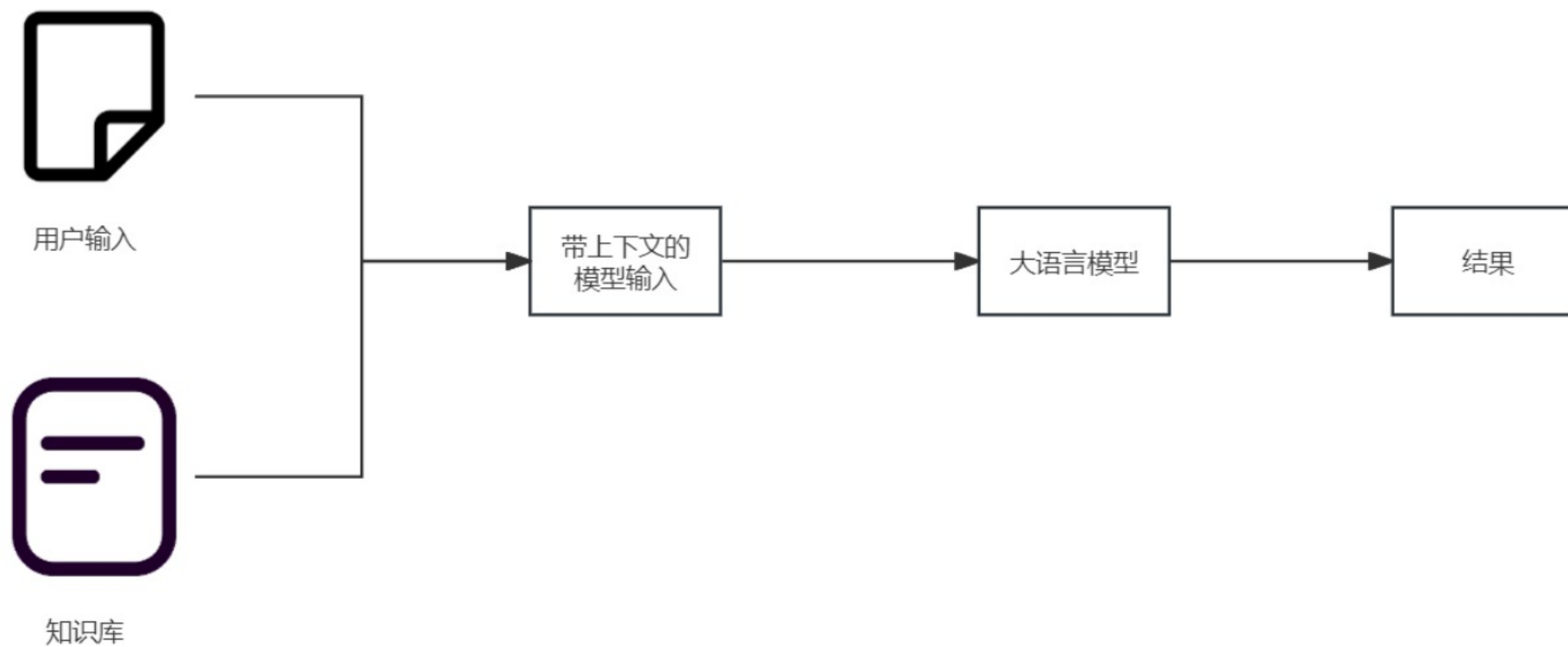
▶ 企业私有化大模型部署方案

Hierarchical index retrieval



CSDN @艾热在coding

▶▶ 企业私有化大模型部署方案



▶▶ 大语言模型企业落地目前存在的问题

- 1、只支持纯文本的输出
- 2、输出的内容有“幻觉”
- 3、对于特定领域的知识无法很好的回答

▶ 大语言模型微调 and 改造方案——只支持纯文本输出

workflow 管道

我们设置了一个 workflow 管道，在模型生成文本描述图像的时候，令一个不同的服务根据描述来索引、搜索和返回适合的图片。

同时，模型可以产生对于数据库中图像的详细描述。这个描述可以被用来查找并返回相应的图片。

链接和参照文本

如果图片存储在网页或在线数据库中，模型可以生成包含图片链接的文本。用户可以通过链接来查看图片。

采用这种方式，我们可以将图片存储在 oss 上，然后再在对模型的输出结构进行修改，使其输出相应的链接，然后再让前端渲染。

多模态模型

训练一个多模态模型（处理和产生不同类型数据的模型，例如文本和图片），然后使用多模态模型直接进行图像的输出。

▶ 大语言模型微调和改造方案——对于特定领域的知识无法很好的回答

与特定领域知识库建立联系

将模型与特定领域的知识库（例如医疗、计算机科学、法律等）进行链接，让模型在需要回答相关问题时可以从检索到必要的信息。

预处理和清洗数据

数据预处理是任何AI项目的关键步骤。需要清洗和整理你的知识库，使其能为模型提供清晰、丰富、准确的信息。同时，确保知识库数据和模型可以兼容。

精准检索

在模型生成回答之前，先对相关知识片段进行检索。这可以通过BM25或向量搜索等技术实现。这一步会从知识库中提取有助于回答问题的文档或段落。

微调和优化

为了让模型更好地理解并生成特定领域的内容，我们需要大量的训练和微调。这里需要注意的是，我们需要在确保模型不会做出伤害性言论或生成不准确信息的前提下进行训练。

大语言模型微调和改造方案——输出的内容有“幻觉”

生成之前

LLM-Augmenter: 当接收到用户查询，该框架首先从外部知识中检索证据并构建证据链。然后，使用整合证据的提示向GLM4提问，生成基于外部知识的候选响应。LLM-Augmenter验证候选响应的真实性，若产生虚假证据，则生成反馈信息用于修订提示并重新提问。此过程持续迭代，直至候选响应通过验证并发送给用户。

FreshPrompt: 该方法使用搜索引擎将相关的和最新信息整合到提示中，增强了动态问答，特别适合用在快速变化的知识场景中。

生成过程中

知识检索:在生成句子前首先利用模型的logit输出值识别可能的幻觉，并验证其准确性，然后纠正已发现的幻觉。

D&Q框架: 该方法指导模型在利用外部知识的同时限制推理时使用的可靠信息，从而减少幻觉的风险。该框架包括一个无需工具调用的监督微调阶段，并在预测阶段让模型使用外部工具查询可靠的问答库，根据需要进行回溯和启动新的搜索。

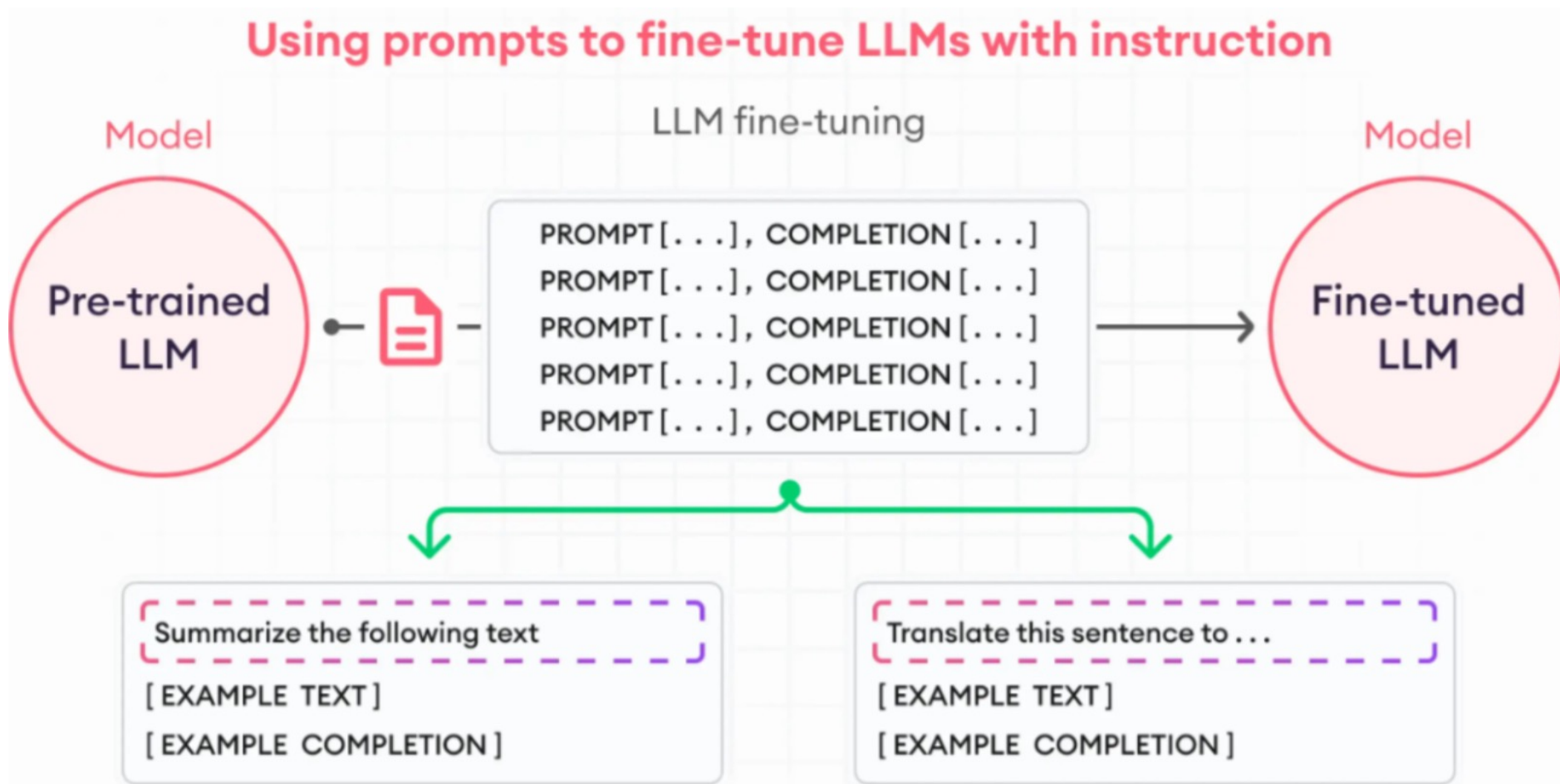
生成之后

RARR: 该框架可为任何文本生成模型自动化归因过程，弥补生成内容缺乏适当的支持的缺陷。通过研究和后期编辑来使内容与检索到的证据保持一致，同时保留原始特性，在LLM生成之后无缝运作。

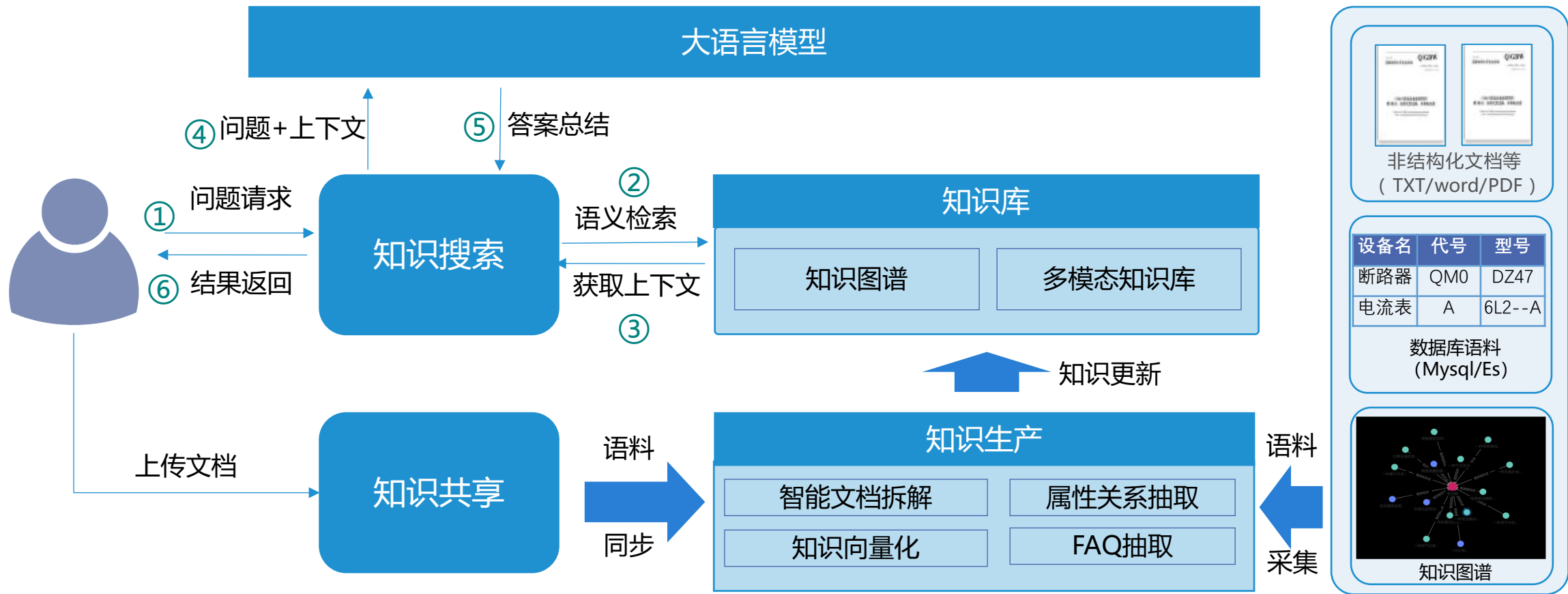
PART 03

实际落地的方案

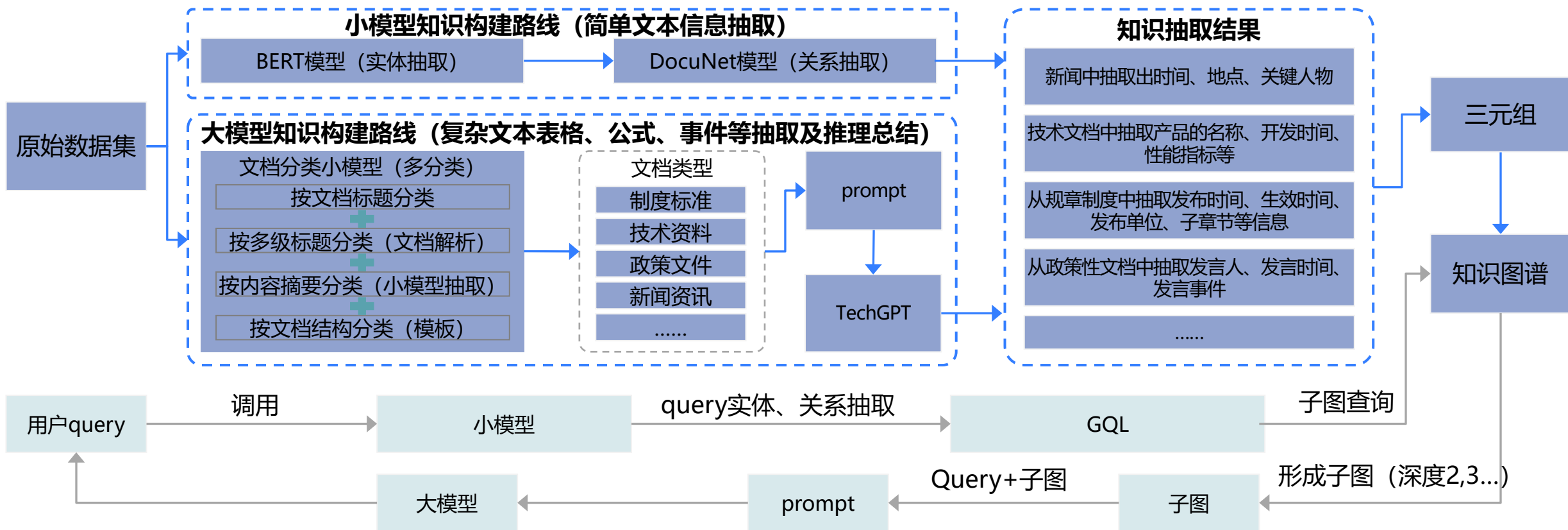
▶ 大语言模型微调 and 改造方案——指令微调



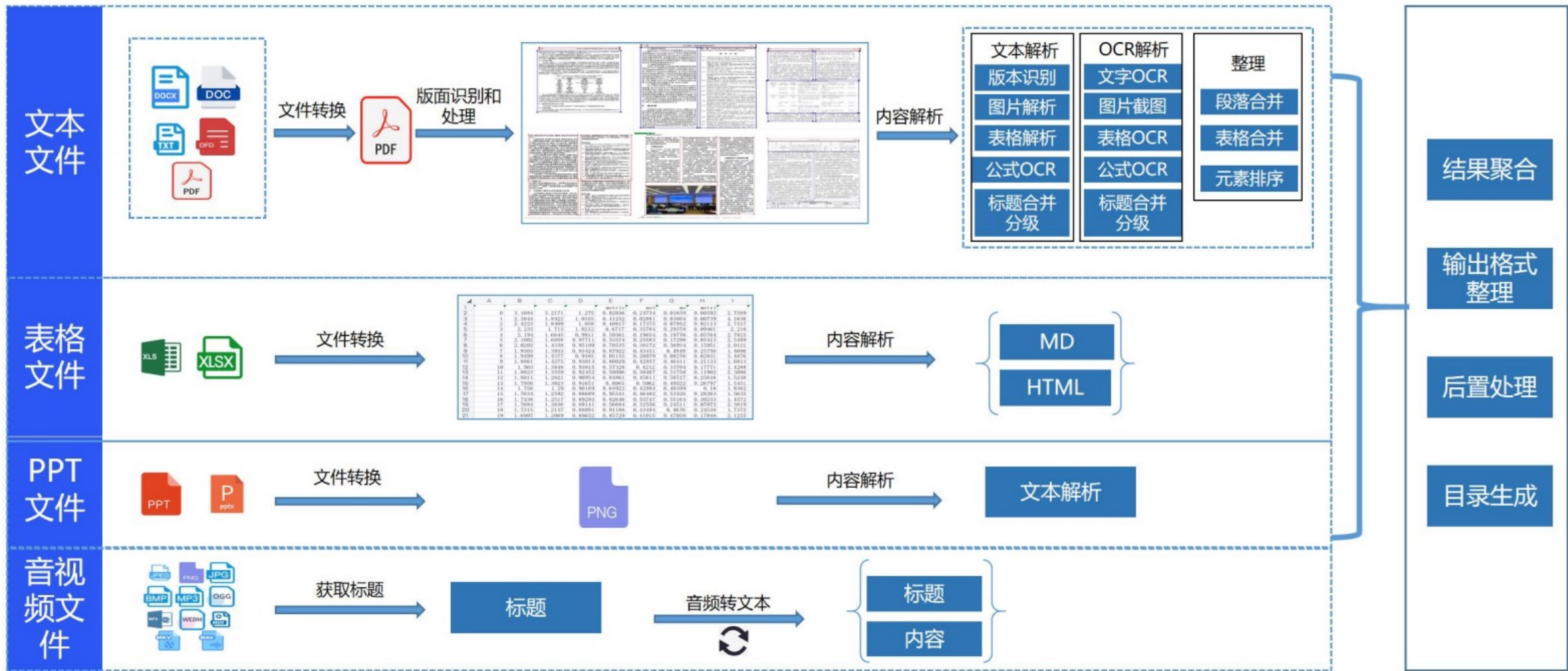
▶ AI赋能关键节点



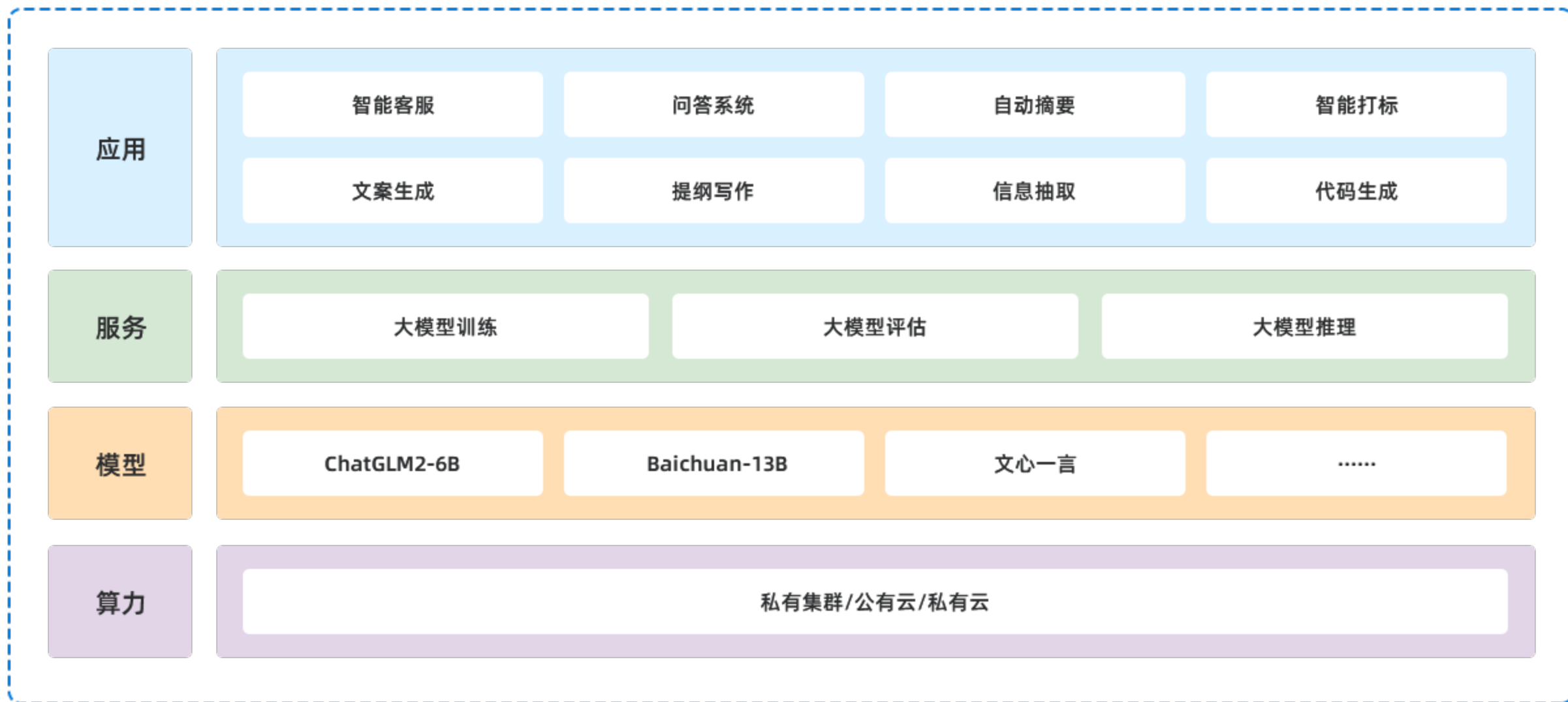
知识图谱与大模型相辅相成



企业私有化大模型部署方案



▶ 企业私有化大模型部署方案



实际落地案例

AI助手

深圳市前海深港现代服务业合作区管理局关于支持人工智能高质量发展高水平应用的若干措施中有多少条

根据您提供的知识库，深圳市前海深港现代服务业合作区管理局发布的《关于支持人工智能高质量发展高水平应用的若干措施》共有16条具体支持措施。这些措施分为五个部分：

1. 第一部分“推动人工智能产业集聚”包含4条措施。
2. 第二部分“强化人工智能关键要素供给”包含3条措施。
3. 第三部分“打造人工智能应用示范”包含4条措施。
4. 第四部分“完善人工智能产业生态”包含5条措施。
5. 第五部分为“附则”。

因此，总共有16条具体措施支持人工智能的高质量发展和高水平应用。

第一部分有哪四条

根据知识库的内容，第一部分“推动人工智能产业集聚”共有4条措施。这些措施包括：

1. 租金补贴
2. 所得税优惠
3. 营收增长奖励
4. 投融资支持

这些措施旨在支持优质人工智能企业在前海合作区集聚，做大做强。

深圳市鼓励跨国公司设立总部企业办法中申报地区总部应具备的条件有哪些

根据知识库的内容，申报地区总部应具备以下条件：

1. 具有独立法人资格的外商投资企业；
2. 申报企业实缴注册资本不低于200万美元；
3. 申报企业被授权管理的境内外独立法人企业不少于2家；
4. 母公司资产总额不低于1亿美元（服务业领域企业设立地区总部的，母公司资产总额不低于5000万美元）；
5. 申报企业未被列入严重失信主体名单。

这些条件是根据《深圳市鼓励跨国公司设立总部企业办法（修订版）》所规定的。

请输入您要咨询的问题，最多500个字

《深圳市鼓励跨国公司设立总部企业办法（修订版）》政策解读.txt - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

二、主要修订内容

本次修订内容主要包括新增认定类型、优化认定条件、增加便利化措施等方面。

一是增加跨国公司总部企业类型。在原有的地区总部、总部型机构基础上增加事业部总部类型，促进跨国公司

二是优化认定条件。在申报主体方面，增加确认投资公司再投资企业可以申报，进一步扩大政策覆盖面；在

三是增加便利化措施。聚焦跨国公司关注度集中的人才、出入境、知识产权等领域进行优化。

三、修订后的适用范围

在本市范围内设立的地区总部、总部型机构及事业部总部适用本办法。

香港、澳门、台湾地区符合条件的公司在本市设立总部企业，参照本办法执行。

四、修订后的总部定义

本办法中跨国公司总部企业，是指跨国公司设立的地区总部、总部型机构及事业部总部。具体来说：

跨国公司地区总部（以下简称地区总部），是指在境外注册的母公司或在国内注册的外商投资性公司在本市设

跨国公司总部型机构（以下简称总部型机构），是指在境外注册的母公司或在国内注册的外商投资性公司在本

跨国公司事业部总部（以下简称事业部总部），是指在境外注册的母公司具有以功能、业务、产品、品牌、服

五、修订后申报主体所需要具备的条件

(一) 申报地区总部应具备以下条件：

1. 具有独立法人资格的外商投资企业；
2. 申报企业实缴注册资本不低于200万美元；
3. 申报企业被授权管理的境内外独立法人企业不少于2家；
4. 母公司资产总额不低于1亿美元（服务业领域企业设立地区总部的，母公司资产总额不低于5000万美元）；
5. 申报企业未被列入严重失信主体名单。

(二) 申报总部型机构的，应具备以下条件：

1. 具有独立法人资格的外商投资企业或其分支机构；
2. 申报企业实缴的注册资本不低于100万美元（如以分支机构形式设立的，总公司拨付的运营资金应不低于100
3. 母公司资产总额不低于5000万美元；
4. 申报企业未被列入严重失信主体名单。

PART 04

总结与思考

总结与思考

简单问候

智能运维客服场景设计

1. 启动智能客服聊天页面
2. 用户语音呼唤：hi, [REDACTED]
3. 智能客服回应：您好！XXX 公司 XX 部门的 XXX，早上好！（根据当前时间，智能选择早上好！上午好或者晚上好）
4. 用户语音：我无法进入 [REDACTED] 维系统了
5. 智能客服回应：请问您目前是使用您的笔记本电脑资产编码 xxxx，还是您的台式机资产编码 xxxx，（调后台数据库，使用人为该用户的资产档案）
6. 用户语音：我是通过我的笔记本电脑在登录 [REDACTED] 维系统
7. 智能客服回应：好的。我目前测试连接 [REDACTED] 系统，可以正常登录。发送登录成功页面。
(通过大模型查询操作手册，获取 [REDACTED] 系统的登录地址，唤起 rpa 机器人进行系统登录测试，并保存登录页面)，因此易运维系统应该是没有问题的。
8. 用户语音：那我为什么无法登录
9. 智能客服回应：应该是你的笔记本的问题，您的笔记本 IP 地址是 192.168.0.111 吗（查询 IT 资产档案中的资产参数获得）
10. 用户回复：是的
11. 智能客服唤起服务器端 PING 192.168.0.111，显示网络没有问题，回复客户：网络没有问题，请把你无法登录的截图发送给我
12. 用户：发送截图
13. 智能客服：通过图像识别，您输入的地址不对，请按如下地址登录系统 https://[REDACTED].com/#/login
14. 用户:好的，我已经可以登录系统了，谢谢 [REDACTED]
15. 智能客服：不客气，祝您工作愉快，记得给我好评哟！

调用外部接口
(function call)

RPA机器人检
查网络环境

通过RAG查找
解决方案

RPA机器人检
查网络环境

OCR内容识别

意图识别

调用外部接口
(function call)

另一种结果

14. 用户：我登录系统时提示账户无法登录
15. 智能客服：好的，我会把您的账号转发给管理员进行处理（生成报事单，报事人，报事资产为 [REDACTED] 维系统，故障描述为：XXXX 账号无法登陆）
16. 用户：好的，谢谢，请尽快帮我解决

➤ 充满复杂的业务逻辑

➤ 需要调用外部接口

➤ 需要使用rpa机器人

总结与思考

结合RAG的文档生成

复杂的解析

复杂的权限管理

整体RAG功能

利用prompt意图识别

用户引导

function call

语音识别

一级功能	二级功能	功能描述
公司知识库构建	知识库构建	构建一个公司内部知识库，该知识库中包含各场景中所需要的知识信息，知识库可按知识类型、所属领域进行分类，分类可人工自定义。
	知识文档上传管理	知识文档可分为普通用户上传和管理用户上传两种：普通用户上传知识文件需要管理用户审核通过后才能进入到知识库中；管理用户上传知识文档，可直接进入到知识库中。上传的文档支持 word 文档、excel 表格、图片（可理解流程图等图片信息）、PDF 及复合文档（文档中包含文字、图片、表格）等文件的解析。
	知识库管理	支持建立多个知识库，支持动态增加知识，支持删除文档/知识库，支持管理权限配置；
	知识库权限管理	同一个知识库，支持基于知识类型划分使用知识的权限，从而做到通过用户绑定知识类型后，可控制该知识库内不同知识的使用权限。
场景知识问答	智能问答	业务人员提问，系统基于知识库或业务流程、Q&A 对生成答案
	多轮问答	业务人员可以对问题进行追问，系统可结合上下文进行语境内问答
	意图识别	可基于多轮问答以及预设流程，对用户的意图进行识别和判断，并通过反问的形式与用户确认，在得到确认后可进行逻辑判断和流程引导
	模型引导及向用户提问	如上方场景中描述的业务所示，系统可基于预设的业务流程引导用户，通过模型向用户提问的方式引导并最终确认用户本次交互的目的，判断和锁定要为用户解决的问题
	关键信息抽取	可基于用户的陈述，在用户的描述话语中抽出关键信息
	链接跳转	可基于在业务流程中的配置，在对话过程中按用户需要，向用户推送可点击的系统跳转链接
	生产业务系统 API 调用	可以基于业务流程，判断是否需要调用业务系统 API，当需要时，可在与用户的对话中抽取关键信息、或直接让用户提供关键信息做为调用 API 的入参，并基于预设模板，将 API 接口返回的数据格式化后友好的展示给用户
	支持文字、语音互转	支持语音或文字方式进行交互

回

文档生成	文档生成	可按要求格式，基于知识库、对话中的关键信息和当前业务及流程，生成指定文档，例如：XX 安全作业管理中的“作业前安全分析表”、清单报告等相关文档，用户可对生成的文档进行评价，评价高的可提交审核，在审核通过后可加入到知识库中
	业务工作流配置	支持个性化场景自定义以及业务工作流的配置。支持自主创建、配置工作流，模型可识别此工作流，并可依据工作流进行问答服务。工作流可自定义增、删、改，并提供工作流与问答服务的配置功能
	插件工具	支持天气、国家节假日、潮汐等相关插件的调用查询
	查看返回答案来源	用户可查看大模型返回结果，查看结果的来源
	系统管理	用户及权限
	生产业务系统 API 管理	记录所有生产业务系统提供的可调用 API，并记录 API 被调用的日志和统计信息
	与客户端对接	支持与 APP 应用、web 客户端的对接，问答时，返回的答案中若涉及图片、文件等附件，需要返回附件链接以及文件类型，前端依据文件类型判断要显示的图片或链接，支持流式问答输出形式
	租户管理	租户管理服务它负责管理所有租户的相关信息和配置。该服务需要提供以下功能： 1、租户注册：允许新的租户注册和创建新的租户账号。注册流程可以包括账号验证、租户信息收集等步骤。 2、租户认证：对已注册的租户进行身份认证，确保只有合法的租户可以访问系统。 3、租户配置：允许租户根据自己的需求进行个性化配置。 4、租户数据管理：提供数据的备份、恢复和迁移等功能，确保租户的数据安全和可用性。

自定义工作流

function call

支持外部系统调用

▶ 总结与思考

理解企业需求

引导企业需求

落地企业需求

科技生态圈峰会 + 深度研习



—1000+ 技术团队的选择



 **K+峰会**  **敦煌站**

K+ 思考周®研习社

时间: 2025.08.29-30

 **K+峰会**  **上海站**

K+ 金融专场

时间: 2025.09.26-27

 **K+峰会**  **香港站**

K+ 思考周®研习社

时间: 2025.11.17-18



K+峰会详情



 **AiDD峰会**  **上海站**

AI+研发数字峰会

时间: 2025.05.23-24

 **AiDD峰会**  **北京站**

AI+研发数字峰会

时间: 2025.08.08-09

 **AiDD峰会**  **深圳站**

AI+研发数字峰会

时间: 2025.11.14-15



AiDD峰会详情



利用AI技术深化计算机对现实世界的理解

推动研发进入智能化时代

