浙江大学 ZHEJIANG UNIVERSITY | 图书馆

知识库构建
及其经验分享

从RAG技术原理调研到应用实践

探索AI时代图书馆的知识管理新范式

刘礼国

内容概览

01 调研分析

RAG技术原理、向量化、检索召回

02 实操过程

数据预处理、测试集构建、平台使用

03 心得分享

UX体验、数据质量、经验总结

基于RAG的知识库构建

RAG

检索增强生成

01 调研分析

RAG技术深度解析

为什么需要知识库（RAG）？

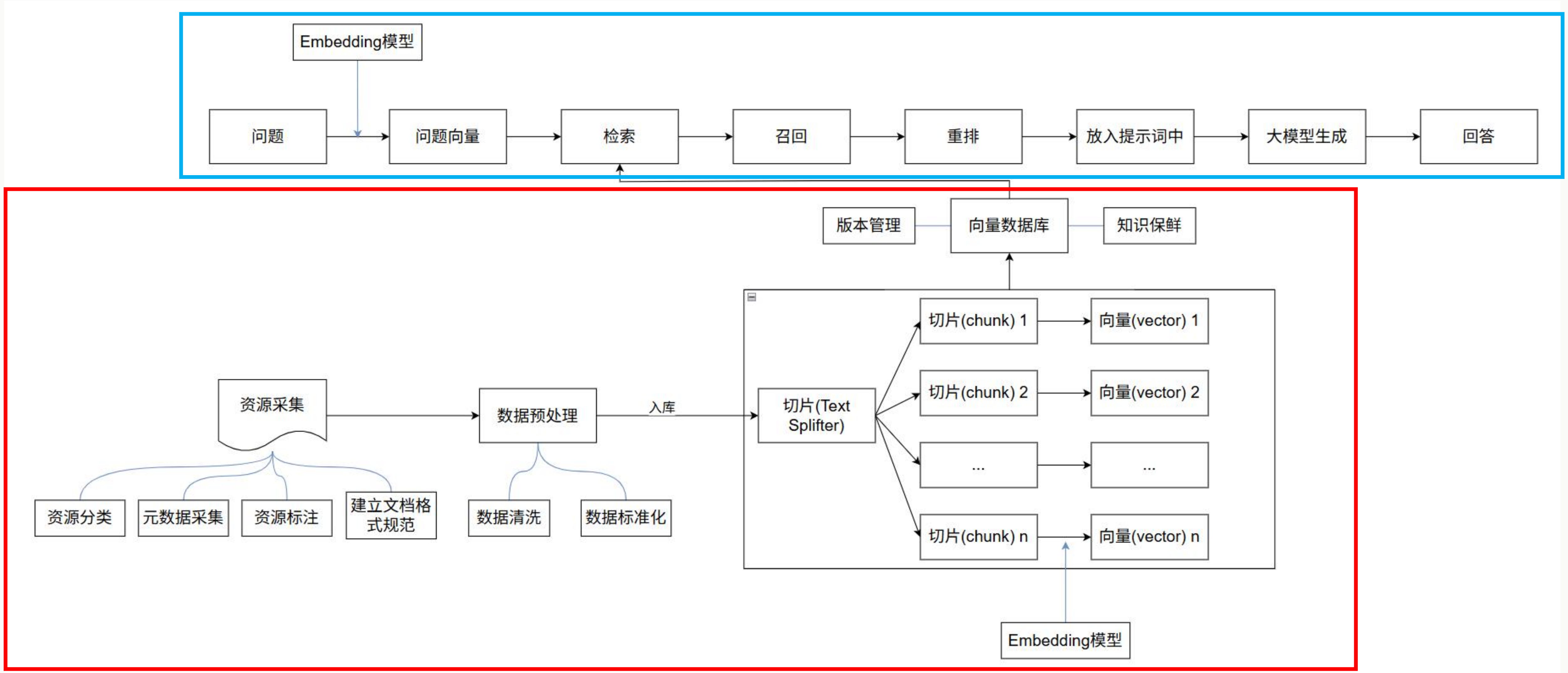
LLMs的固有问题

- 知识更新不及时，存在时效性限制
- 预训练成本过高，难以频繁更新
- 上下文输入长度有限，无法处理长文档

核心优势

- 改善LLM幻觉问题，提升回答准确性
- 无需预训练即可适配特定领域知识

知识库-RAG技术原理与流程



1 资源采集

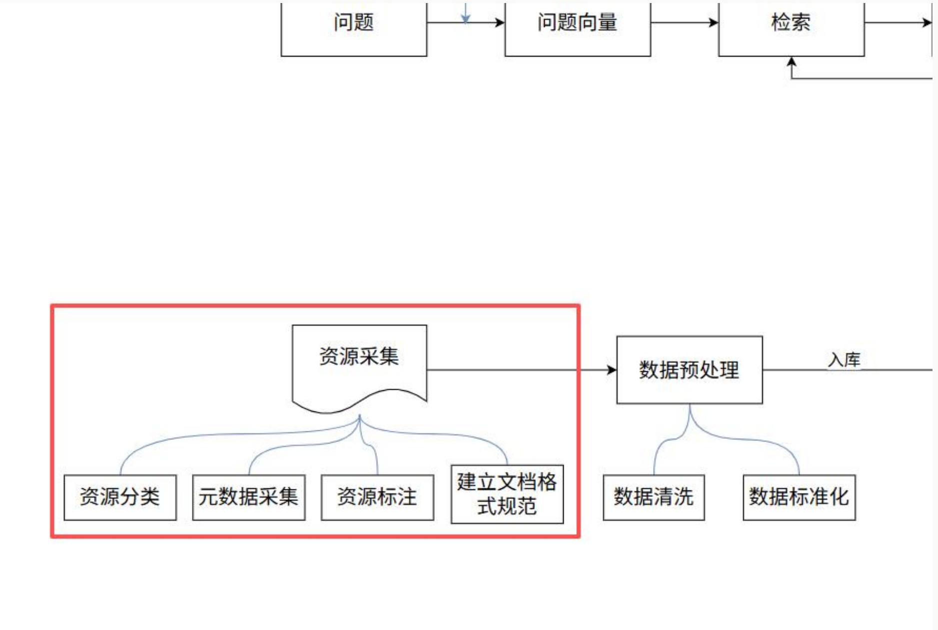
2 数据预处理

3 文本切分

4 向量编码

5 检索生成

数据处理策略



文档规范与标注

- 建立统一的文档分类标准
- 实施结构化的内容标注体系
- 文档格式尽量简洁
- 确保元数据完整性
- 例如：采用国标（GB/T 13745—2009）的学科分类，或是根据资源类型进行分类
- 例如：标题标注、关键词标注等
- 收集原则：统一格式、记录来源与时间、处理隐私、去重、支持动态更新

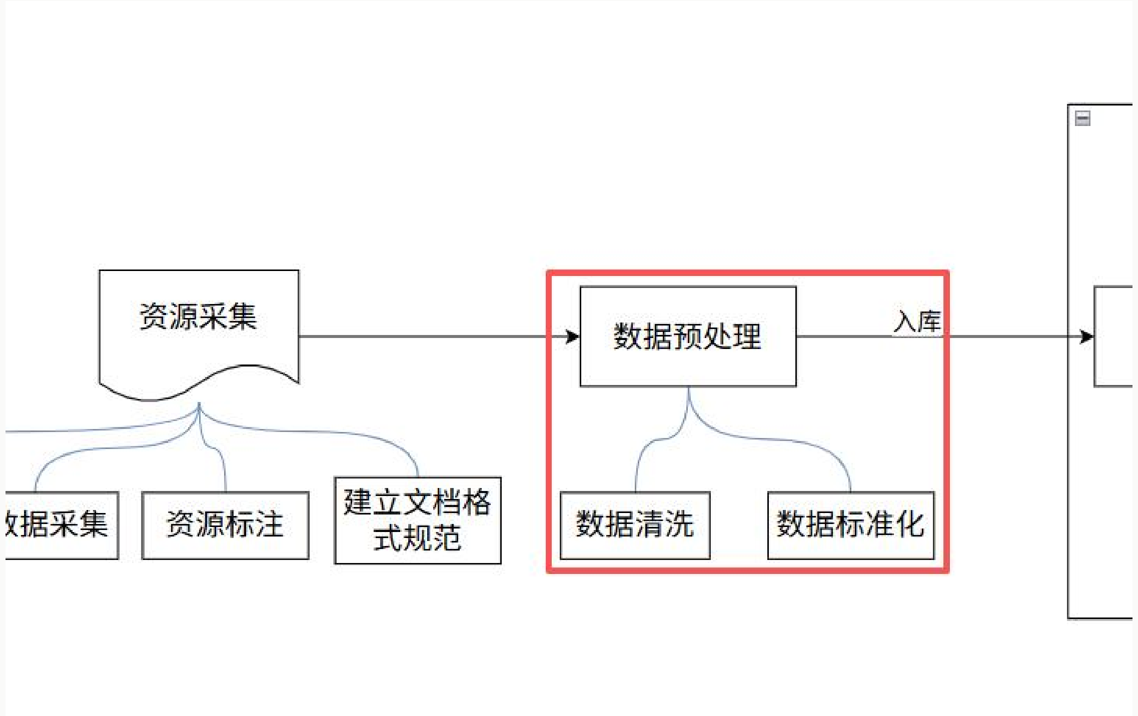
类型	内容	示例
结构标注	标注文档结构层次，包括标题、章节、目录、引言、图表、脚注、参考文献、附录等	“标题“、“绪论”、“正文“、“结尾”
关键词标注	提取反映文献核心内容的主题词，用于索引、聚类与推荐	“人工智能”、“数据隐私”
命名实体标注（Named Entity Recognition，NER）	标注文献中的专有名词，包括：作者名、机构名、地名、时间、数据集名称、标准号等	“7月”、“北京”
引用与引用关系标注	标注文献内部引用、被引情况，以及与其他文献之间的引用关系	“引用文献A”、“更新自文件B”
实体关系标注	标注文中实体之间的逻辑、语义或结构性关系，如“属于”、“作用于”、“依赖于”、“改进了”等	“算法 A 改进了 B”、“模型 M 应用于领域 N”
摘要标注	提取或生成对文献核心内容的简明总结，可用于搜索预览、快速浏览	“本论文介绍了知识库搭建的注意事项。”

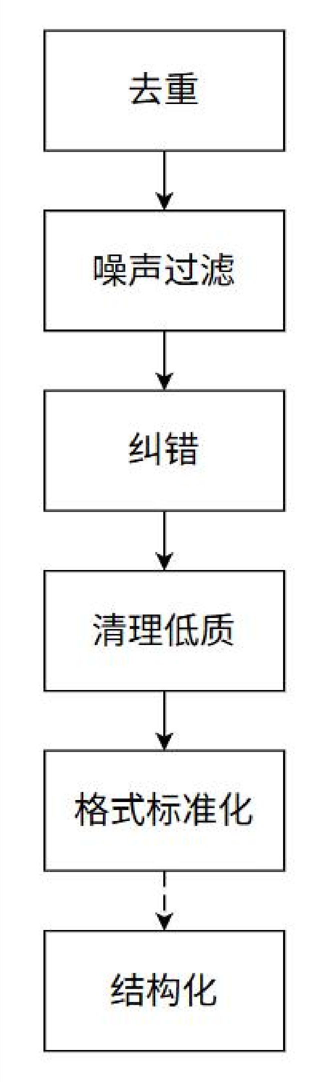
数据处理策略

- 去除重复数据，过滤噪声，提升数据质量。
- 纠正错误，统一编码与文件格式，确保数据规范；对于可以或需要结构化的数据可以采用正则表达式进行操作。

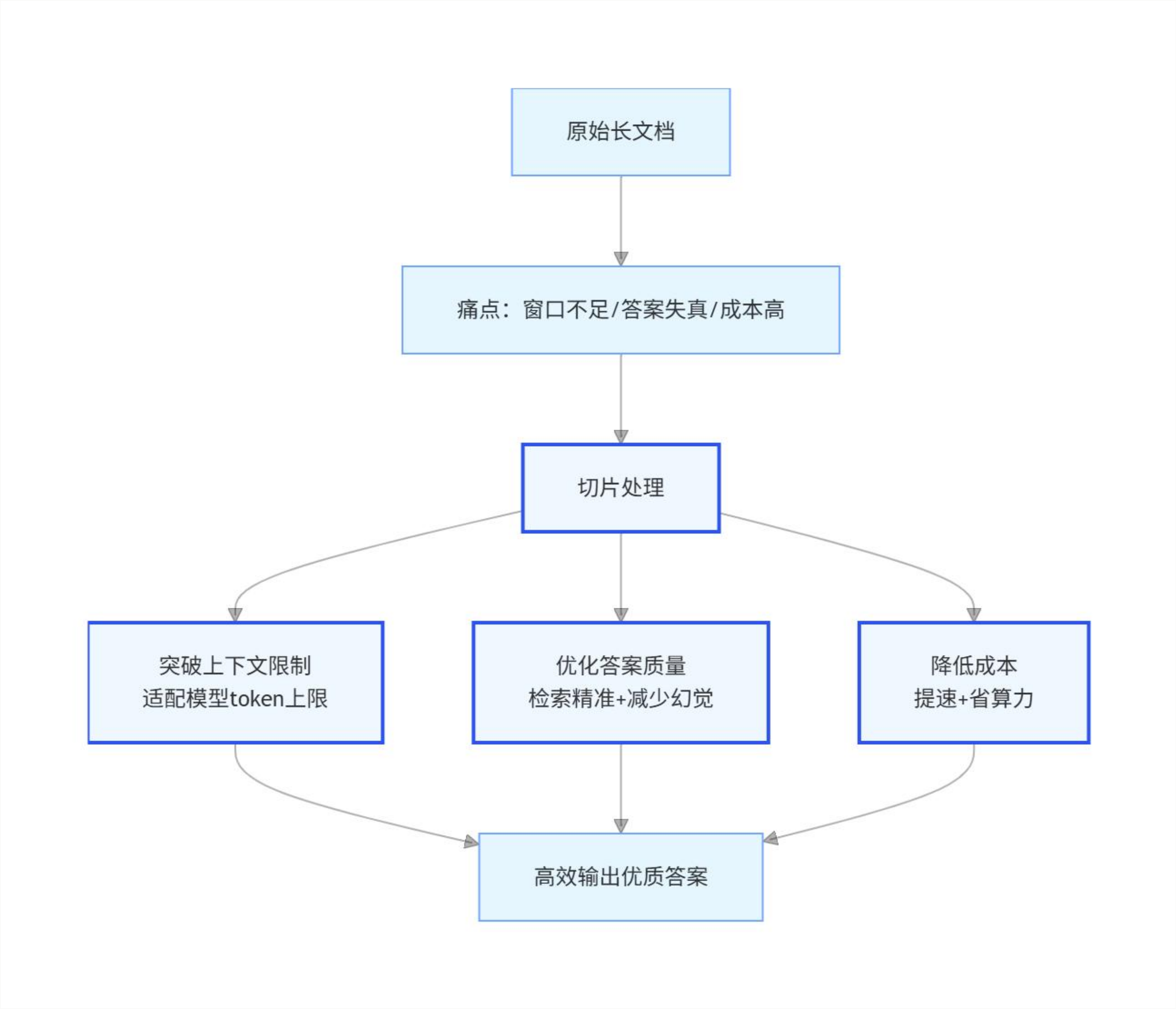
数据预处理

- 数据清洗：去除噪声和冗余
- 数据标准化：统一格式和编码
- 质量检测：确保数据可用性



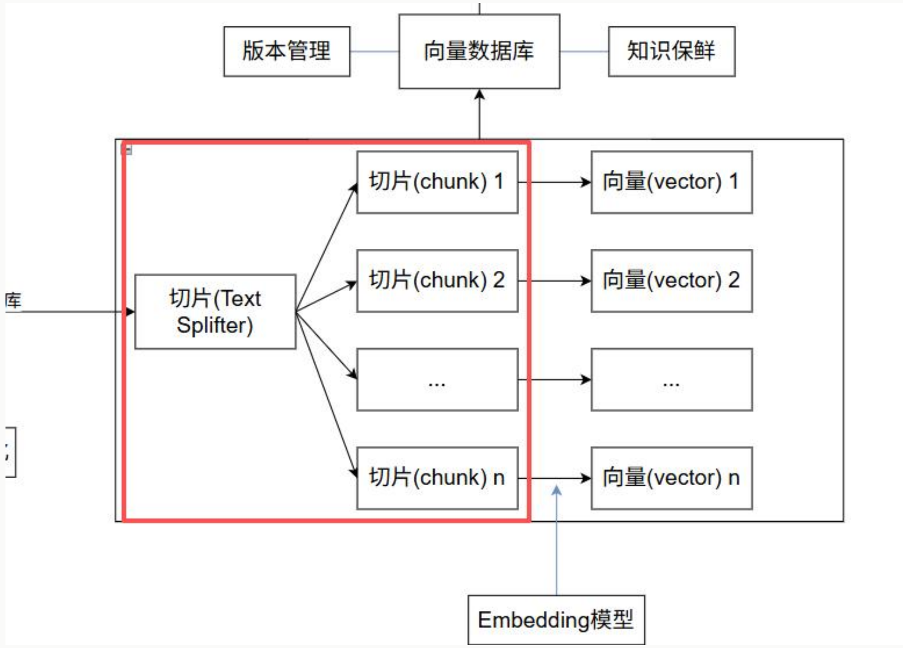


数据处理策略



文本切分策略

- 固定长度切分
- 滑动窗口切分
- 语义单元切分
- …



数据处理策略

ABCDEFGHIJK

Chunk 1:
ABCDE

Chunk 2:
FGHIJ

Chunk 3:
K

ABCDEFGHIJK

Chunk 1:
ABCDE

Chunk 2:
DEFGH

Chunk 3:
GHIJK

ABCDEFGHIJK

Chunk 1:
ABC

Chunk 2:
DEF

Chunk 3:
GHIJK

文本切分策略

- 固定长度切分:
  text="ABCDEFGHIJK"
  5个字符一个切片
- 滑动窗口切分:
  text="ABCDEFGHIJK"
  5个字符一个切片，步长为3
- 语义切分:
  text="ABCDEFGHIJK"
  假定ABC、DEF、GHIJK为三个词语

数据处理策略

文本切分策略

- 固定切分的局限性：语义断裂
- 滑动窗口：存在边界效应
- 语义单元：噪声干扰、语义漂移

```
Chunk 2:
和染色技术 年德国病理学家 创立了细胞病理学。 世纪 随着电子显微镜的问世和 年代 变组织和细胞的形态变化 随着分子生物学、免疫学、细胞生物 产生了超微病理
学。近三十年来 超薄切片技术的发展 免疫组织化学、流式细胞术、图像分析技术等新技 学、细胞遗传学、计算机科学等学科的发展 又相继出现了分子病理学、免疫病理
学等新兴病理学学科。 术的应用 病理学技术的不断完善和创新人类认识疾病、研究疾病 相信随着科学技术的不断发展 的能力和水平将会得到不断提高。 第一节 病理检
验与病理检验技术 病理学是运用科学的方法研究患病机体的形态结构、功能和代谢等方面的改变,揭示疾病的 病因、发病机制和转归的医学基础学科。研究患病机体形态
结构的变化及其发生机制是病理学 的主要任务。 作为一门医学基础学科 ，病理学一方面研究疾病的病因、发病机制 ,为人类认识和掌握疾病 发生发展的规律 ,为人类
战胜疾病提供重要的理论基础 ；另一方面 ，病理学从学科创建开始就在 诊断疾病、治疗疾病、分析预后等临床医疗实践中发挥重要的作用。病理学在临床医疗实践中
的 具体应用就是开展病理检验。 一、病理检验的主要任务 病理检验也称病理学检查。在临床医疗实践
```

数据处理策略

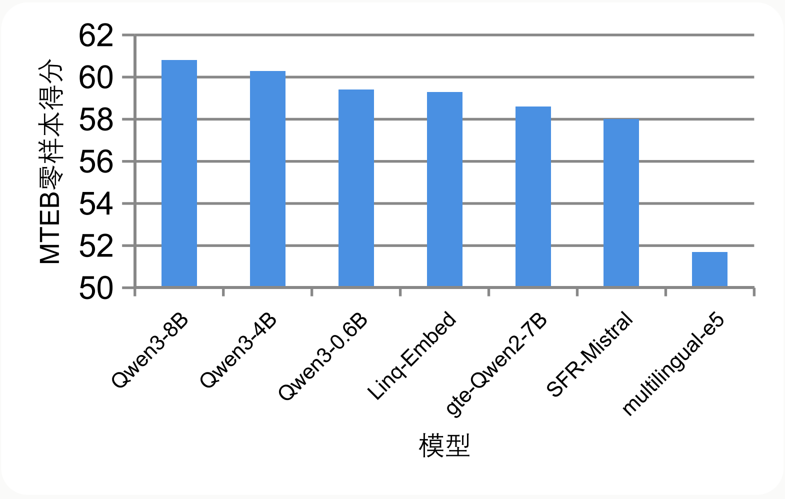
文本切分策略

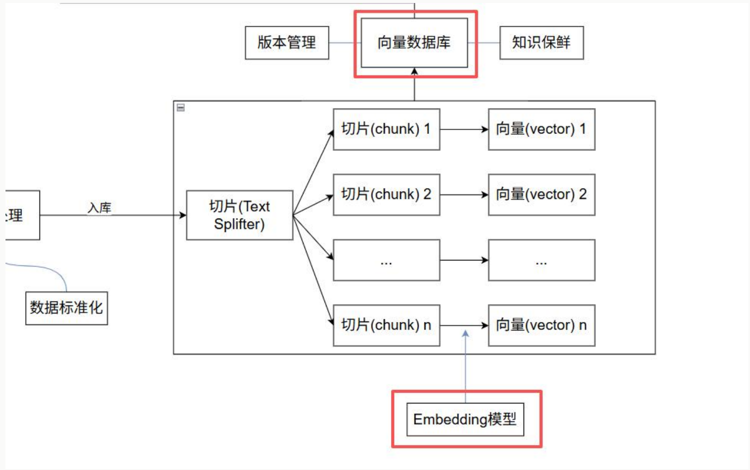
Chunk 2:
和染色技术 年德国病理学家 创立了细胞病理学。 世纪 随着电子显微镜的问世和 年代 变组织和细胞的形态变化 随着分子生物学、免疫学、细胞生物 产生了超微病理
学。近三十年来 超薄切片技术的发展 免疫组织化学、流式细胞术、图像分析技术等新技 学、细胞遗传学、计算机科学等学科的发展 又相继出现了分子病理学、免疫病理
学等新兴病理学学科。 术的应用 病理学技术的不断完善和创新人类认识疾病、研究疾病 相信随着科学技术的不断发展 的能力和水平将会得到不断提高。 第一节 病理检
验与病理检验技术 病理学是运用科学的方法研究患病机体的形态结构、功能和代谢等方面的改变,揭示疾病的 病因、发病机制和转归的医学基础学科。研究患病机体形态
结构的变化及其发生机制是病理学 的主要任务。 作为一门医学基础学科 ，病理学一方面研究疾病的病因、发病机制 ,为人类认识和掌握疾病 发生发展的规律 ,为人类
战胜疾病提供重要的理论基础 ；另一方面 ，病理学从学科创建开始就在 诊断疾病、治疗疾病、分析预后等临床医疗实践中发挥重要的作用。病理学在临床医疗实践中
的 具体应用就是开展病理检验。 一、病理检验的主要任务 病理检验也称病理学检查。在临床医疗实践

Chunk 3:
疗实践中的 具体应用就是开展病理检验。 一、病理检验的主要任务 病理检验也称病理学检查。在临床医疗实践中,通过对患者病变组织或细胞进行检查,以协 助临床诊断
疾病的方法称为病理检验。病理检验的主要任务包括以下几个方面。 (一)确定疾病的诊断 临床上,虽然有不少疾病在经过临床有关检查后就能得出初步的临床诊断结果 ，
某些以功能 或代谢紊乱为主要改变的疾病也不需要进行病理学检查,但是对于大多数有明确器质性病变的 疾病而言,病理学检查和诊断仍然是最正确、最可靠和最后的诊
断。如临床检查所发现的各种肿 块或影像学检查出的占位性病变、内窥镜检查见到的各种溃疡、局部僵硬及肿块等,都需要进行 病理学检查才能对病变性质做出正确的判
断,并在此基础上确立诊断。 (二)为临床选择治疗方案提供依据 通常情况下,只有当疾病的诊断明确之后,临床治疗方案的制定才具有针对性、合理性和有 效性。例如颈

- 优化策略：
  添加前后缓冲区，保留更多的上下文语义
  针对不同文本选择不同策略（需实践测试得出）

数据处理策略





▪ Embedding模型选择时，不同模型可能在不用任务上的表现也大不相同，例如：

- text-embedding-ada-002 适合英文、通用任务
- BGE/M3E 在中文任务上召回效果更好
- SimCSE 在短文本匹配上有优势

这部分的调优也往往比较玄学，所以在实际部署之前可以参考下Embedding模型榜单或论坛再行部署

Rank (Bo…	Model	Zero-shot	Memory U…	Number of P…	Embedding D…	Max Tokens	Mean (T…	Mean (TaskT…	Bitext …	Classification	Clusterin
1	llama-embed-nemotron-8b	99%	28629	7B	4096	32768	69.46	61.09	**81.72**	73.21	54.35
2	Qwen3-Embedding-8B	99%	28866	7B	4096	32768	**70.58**	**61.69**	80.89	**74.00**	**57.65**
3	gemini-embedding-001	99%	Unknown	Unknown	3072	2048	68.37	59.59	79.28	71.82	54.59
4	Qwen3-Embedding-4B	99%	15341	4B	2560	32768	69.45	60.86	79.36	72.33	57.15
5	Qwen3-Embedding-0.6B	99%	2272	595M	1024	32768	64.34	56.01	72.23	66.83	52.33
6	gte-Qwen2-7B-instruct	⚠ NA	29040	7B	3584	32768	62.51	55.93	73.92	61.55	52.77
7	Linq-Embed-Mistral	99%	13563	7B	4096	32768	61.47	54.14	70.34	62.24	50.60
8	multilingual-e5-large-instruct	99%	1068	560M	1024	514	63.22	55.08	80.13	64.94	50.75
9	embeddinggemma-300m	99%	578	307M	768	2048	61.15	54.31	64.40	60.90	51.17
10	SFR-Embedding-Mistral	96%	13563	7B	4096	32768	60.90	53.92	70.00	60.02	51.84
11	text-multilingual-embedding-002	99%	Unknown	Unknown	768	2048	62.16	54.25	70.73	64.64	47.84

向量化技术

▪ Embedding模型选择：MTEB

▪ 向量数据库

▪ 性能优化：索引策略与缓存

数据处理策略

数据库	开源/托管	性能	易用性	适用场景
Pinecone	托管	高	高	实时推荐、语义搜索
Milvus	开源	高	中	大规模图像/NLP 处理
Weaviate	开源/托管	中	高	混合搜索、知识图谱
Elasticsearch	开源/托管	中	中	企业级搜索、日志分析
PgVector	开源	中	高	混合数据查询、中小规模 AI 应用
Redis	开源/托管	高	高	实时推荐、缓存与搜索结合
ChromaDB	开源	中	高	轻量级 AI 应用、RAG 系统
LanceDB	开源	高	中	边缘设备、数据湖分析

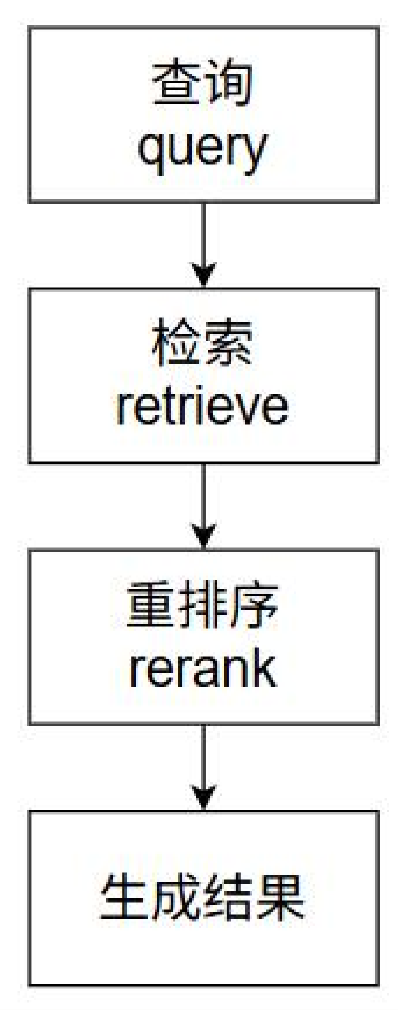
- 根据场景选择商用托管或开源自建数据库，满足不同需求。

示例：
- 商用托管: Pinecone
- 开源自建: Milvus, Weaviate
- 组合场景: ES / PgVector / Redis
- 轻量嵌入: Chroma, Lance

向量化技术

- Embedding模型选择：MTEB
- 向量数据库
- 性能优化：索引策略与缓存

检索优化策略



检索效率提升

1.问题改写：优化查询表达

2.意图识别：理解用户真实需求

3.多路召回：提升召回率

4.上下文补充：增强语义理解

Re-call参数

- top-k：控制召回数量
- 相似度阈值：过滤低质量结果
- 平衡召回率与精确率

Re-rank参数

- top-k'：重排序候选集
- 排序阈值：质量控制
- 提升最终结果相关性

检索优化策略

Temperature（温度）	控制输出随机性，数值越高（0-2）越发散创新，越低越聚焦确定。
Top-K	限制生成时仅从概率前 K 个候选词中选择，K 越小输出越集中，越大越灵活。
Top-P（核采样）	按概率累积选候选词，累积概率达 P（0-1）即停止，平衡多样性与相关性。
Frequency Penalty（频率惩罚）	抑制重复词出现，数值越高（-2-2），越减少已多次出现的词汇。
Presence Penalty（存在惩罚）	鼓励引入新主题，数值越高（-2-2），越倾向于生成未提及过的内容。

LLM配置：Temperature、Top-K、Top-P、Frequency Penalty、Presence Penalty等参数来控制生成随机性；Prompt来控制输出风格

02

实操过程

从理论到实践的完整路径

文件处理与测试集构建

文件处理流程

1. JSON → TXT 格式转换
2. 数据清洗：去换行、HTML标签、统一编码
3. 噪声过滤与去重处理

- 基础文本清洗：
  - 去除 HTML 标签
  - 空白规范化：将换行符、制表符等各类空白字符统一替换为空格
  - 编码处理：将编码格式转变成utf-8
- 噪声过滤：
  - 去除空文本、过短文本、无有效字符、纯标点符号、低可读性
- 文件间去重：
  - 剔除近似重复文本，避免重复文本入库

文件处理与测试集构建

测试集构建方法

人工构建

- 从文档中选择代表性内容
- 提炼问题，构建问答对

序号	问题	答案
1	微生物学的分支有哪些	普通微生物学:研究微生物一般形态和生理的规律性、分类法
2	糖酵解酶系主要存在于哪	细胞质基质
3	与动脉粥样硬化发病关系	LDL
4	细菌的基本形态有哪些	球菌、杆菌、螺形菌
5	细菌的一般构造包含哪些	细胞膜、原生质、内含物、核质
6	在显微镜检查时，常用的	革兰氏染色及耐酸染色等
7	巴甫洛夫关于偏执给出的	神经类型异常与大脑皮层功能失调
8	以玉米为主食的地区易发	癞皮病
9	免疫球蛋白的分类有哪些	IgG、IgM、IgA
10	胸腺的功能有哪些	胸腺是T淋巴细胞的发源地、胸腺制造和分泌胸腺素、胸腺在
11	简谐振动方程的公式是什	x=Acos(wt+Φ)
12	在高频电疗时，为避免刺	1兆赫
13	多少剂量的伦琴射线会影	照射剂量如50,100,250及500琴也使酵母菌的出芽生殖受到
14	静脉的构造与其功能上的	静脉的特殊构造是与它們的功能有关系的。静脉把血液带到
15	呼吸运动的肌群有哪些	分吸气与呼气的肌群。主要靠肋间肌和膈肌起作用。
16	活动头的肌群有哪些	活动头的肌群分屈、伸、侧屈及回旋四个肌群。1.前俯头(屈
17	消化管壁的构造有几层，	四层:由内向外分别为粘膜层、粘膜下层、肌层和外膜
18	神經系肺吸虫病的分类有	一、脑型:1.脑膜脑炎型2.扩張型--多房性脓肿3.萎縮型--
19	肝脏對各种刺激物作用的	糖代谢变化
20	海绵状脑病是什么	1.海绵状脑病:spongiorm encephalopathies,一组慢性病

数据质量与测试策略

测试集构建策略

LLM辅助构建

- 使用LLM生成QA对
- 质量评估与筛选

1 **采样维度考量**

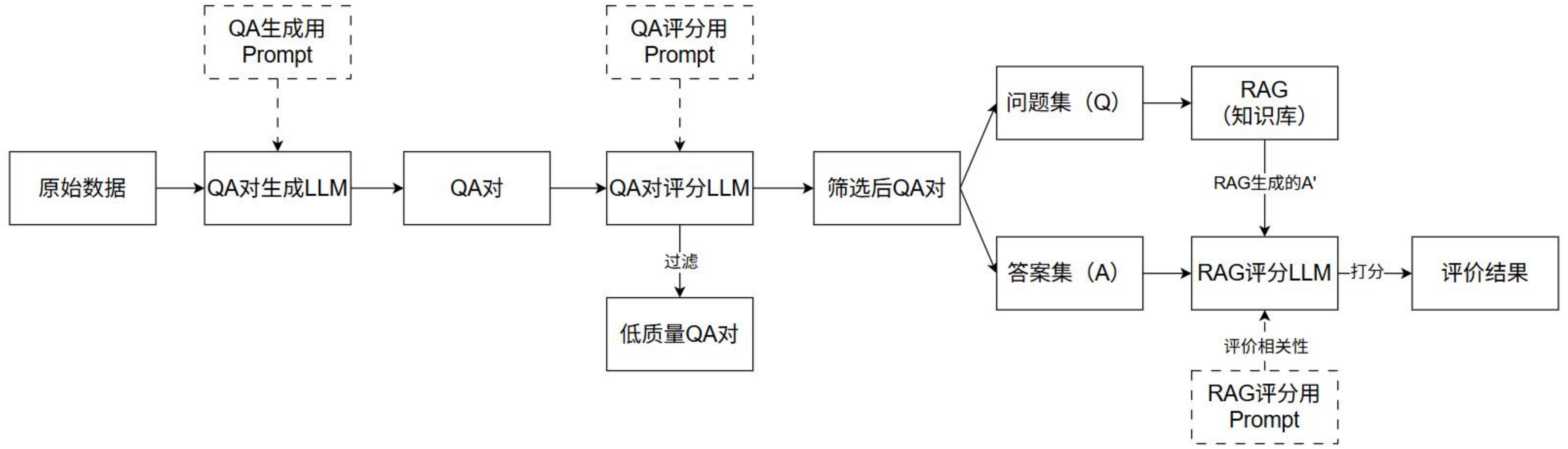
确保问题普遍性，考虑时间有效性

2 **LLM辅助构建**

生成QA对，批判评分，过滤优化

3 **自动化评估**

LLM评分RAG答案，持续迭代优化



知识库构建平台

文件上传

支持多模态内容：docx、pdf、excel等

切片配置

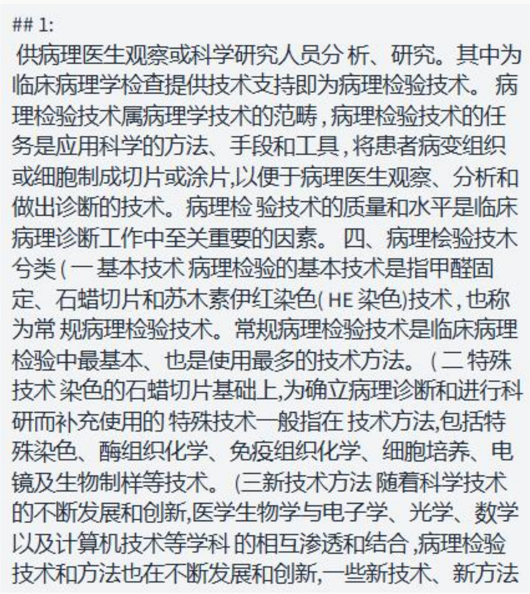
配置策略，预览切片质量，支持修改

Agent构建

配置提示词，设置recall/rerank参数

效果评估

基于测试集评估RAG表现，持续优化



1:
供病理医生观察或科学研究人员分析、研究。其中为临床病理学检查提供技术支持即为病理检验技术。病理检验技术属病理学技术的范畴，病理检验技术的任务是应用科学的方法、手段和工具，将患者病变组织或细胞制成切片或涂片，以便于病理医生观察、分析和做出诊断的技术。病理检验技术的质量和水平是临床病理诊断工作中至关重要的因素。四、病理检验技术分类（一基本技术 病理检验的基本技术是指甲醛固定、石蜡切片和苏木素伊红染色（HE 染色）技术，也称为常规病理检验技术。常规病理检验技术是临床病理检验中最基本、也是使用最多的技术方法。（二特殊技术 染色的石蜡切片基础上，为确立病理诊断和进行科研而补充使用的特殊技术一般指在技术方法，包括特殊染色、酶组织化学、免疫组织化学、细胞培养、电镜及生物制样等技术。（三新技术方法 随着科学技术的不断发展和创新，医学生物学与电子学、光学、数学以及计算机技术等学科的相互渗透和结合，病理检验技术和方法也在不断发展和创新，一些新技术、新方法

03

心得分享

实践经验与未来展望

数据质量与测试策略

数据质量优化

文档格式优选

尽量选能保留结构，简洁明了的格式

知识梳理整合

文档整合、过期内容清理、精准性保证

手动优化

添加标识符/分割符，提升切片效果

```
- LAION-SG has 6.39 objects per sample, excluding abstract proper nouns and focusing on specific nouns that reflect
true semantic relationships. LAION-SG contains 20% more object information than the original LAION-Aesthetics dataset,
and this advantage increases to 216% when excluding proper nouns.
- The average annotation length for our scene graphs and original captions is 32.2 and 19.0, which reflecting SGs
contain richer information in a more compact form.
- The annotation accuracy of the scene graph is also higher than that of the original captions. For details, please
refer to the paper.

### Data Splits

- A total of 540,005 SG-image pairs annotated with objects, attributes, and relationships.
- 480,005 samples for taining
- 10,000 samples for validation
- 50,000 samples for test

## Uses
 1. **Download the annotation files and processing code.**

    Download the `dataset` folder and the `code` folder to your local machine.

 2. **Download images for LAION-SG dataset.**

    Due to copyright issues, we are unable to provide the image files of the dataset. Please download the required
    images for the LAION-SG dataset using the URLs provided in the three JSON files under the `dataset/` directory. All
    images should be stored in the `data` folder, as specified by `args.image_dir`.


    *A reference download method:*
```

成为 CADAL 项目共建共享单位申请流程：

如有进一步合作意向，希望获取更多资源服务权限，请联系 CADAL 项目管理中心秘书处进行咨询，以便根据所属地区情况，签订相应服务协议。

参与 CADAL 项目子项目建设的申请流程如下：(在申报前需成为 CADAL 项目的共建共享单位)

1. 项目申报

CADAL 项目管理中心将通过邮件或官网公告等方式，向各共建单位发布项目申报通知，明确申报时间与具体要求。

2. 材料提交

各院校可根据自身意愿，针对资源类、服务推广类和技术（平台）类三种子项目类型，按申报要求准备并提交申请材料。

3. 评审与立项

项目材料经专家评审通过后，院校与 CADAL 项目管理中心正式签订协议，完成立项。

4. 项目执行

院校应按照协议要求，全面开展项目实施工作。如申报资源类项目，院校须

CADAL 项目常见问题

么是 CADAL?

字图书馆国际合作计划"China Academic Digital Associative Library,简称 CADAL 由
建设，作为教育部"211"重点工程，由浙江大学联合国内外的高等院校、科研机构
，项目负责人为浙江大学潘云鹤院士。CADAL 项目一期（2001-2006）完成 100
数字化，提供便捷的全球可访问的图书浏览服务。CADAL 项目二期（2007-2012）
万册图书数字化，构建了较完善的项目标准规范体系，初步建设分布全国的服务
ADAL 项目从单纯的数据收集向技术与服务升级发展转变。2013 年以后，CADAL
运维保障期，继续在资源，服务，技术，对外交流合作等方面推进工作。

么是共建单位?

建设、技术合作、设备支持等的单位。

么是共享单位?

方或者直接与项目管理中心签订服务协议的单位。

什么要注册?

实行图书借阅模式，要求认证读者身份;
户才允许浏览和借阅图书;
户可以享受系统提供的个性化服务;
户可以向系统管理员提交使用中遇到的问题。

那些个性化服务?

阅历史，快速定位已读图书;
源，查看及评论其他读者评论;
兴趣资源分门别类，快速定位资源;
他读者对自己评论的回复;
献信息，推荐相关资源;
他读者相互评论、点赞、方便知识的分享与交流;

UX设计与平台灵活性

UX设计原则

1. **功能完整性**
   批量上传、简洁界面、直观操作
2. **解析效果**
   PDF/Excel重点优化，保留语义完整性
3. **切片可视化**
   多种策略选择，直观预览修改

平台灵活性

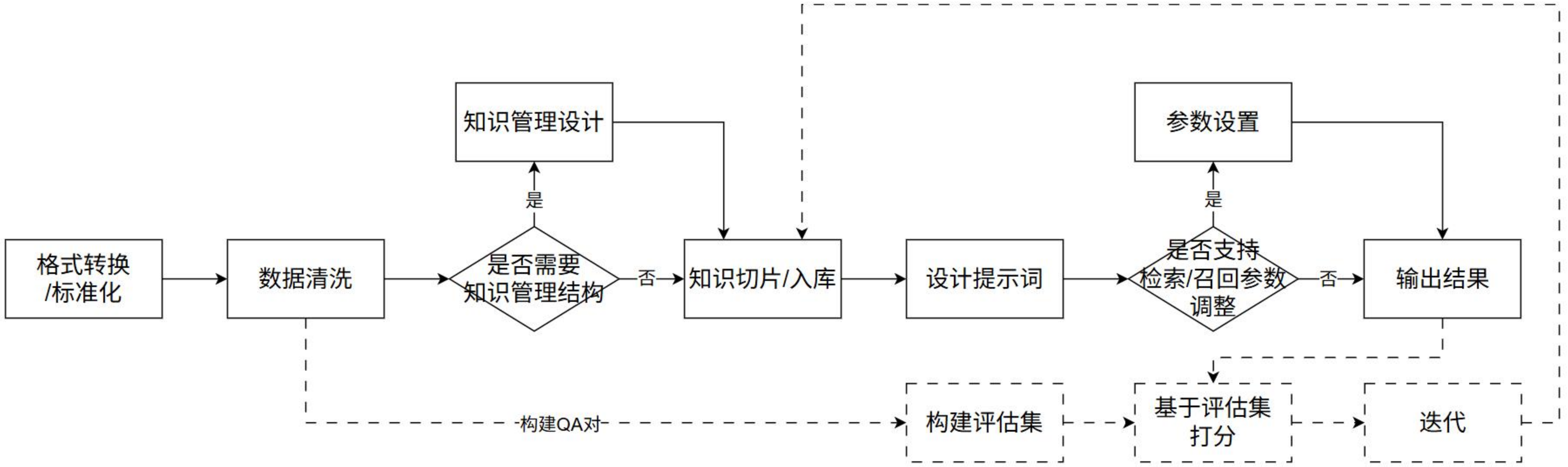
LLM选择

支持多模型适配，根据场景灵活选择

自定义配置

切片策略、召回参数、工作流定制

知识库构建规范



知识管理设计
参数设置
是
是
格式转换
/标准化
数据清洗
是否需要
知识管理结构
否
知识切片/入库
设计提示词
是否支持
检索/召回参数
调整
否
输出结果
构建QA对
构建评估集
基于评估集
打分
迭代

知识库构建规范

知识库建设流程与评价白皮

White Paper on Knowledge Base Development Process and Evalu

4.5 向量化与索引建立

4.5.1 基本概述

在该章节介绍了 Embedding 的主要流程与向量数据库的建设流程，提供向量化及索引建立的技术参考建议,可为后续的检索与应用奠定基础.

4.5.2 Embedding

Embedding 阶段是指将预处理后的数据进行向量化处理，该阶段中一般会涉及 Embedding 模型的选择，通过 Embedding 模型实现数据的向量化，选择时一般考虑如下几个方面:

a) 向量表示精度（语义质量）：通过检索或语义相似度任务的基准测试来衡量.

b) 推理效率与部署：结合具体业务衡量是否需要进行本地部署，以及对 Embedding 模型推理效率的需求.

c) 跨语言与领域适应：结合具体业务综合选择 Embedding 模型是否需要多语言支持，以及跨领域适应.

目前由 hugging face 所发布的 MTEB 评测受到业内的广泛认可，可用作选型时的参考，下述为截至至 2025 年 8 月 5 日的部分开源 Embedding 模型在该评测上的排名，由高到低排序，仅考虑了中文与英文:

表 5 MTEB 评测模型排名（部分）

Rank (Borda)	Model	Zero-shot	Memory Usage (MB)	Number of Parameters	Embedding Dimensions
1	[Qwen3-Embedding-8B](https://huggingface.co/Qwen/Qwen3-Embedding-8B)	98%	28866	7B	4096
2	[Qwen3-Embedding-4B](https://huggingface.co/Qwen/Qwen3-Embedding-4B)	98%	15341	4B	2560
3	[Qwen3-Embedding-0.6B](https://huggingface.co/Qwen/Qwen3-Embedding-0.6B)	98%	2272	595M	1024
4	[Linq-Embed-Mistral](https://	98%	13563	7B	4096

资源类型”双轴分类体系，综合考虑知识的内容主题（学
）），目标是构建多维复合知识索引结构，以满足知识

的国家标准《学科分类与代码》（GB/T 13745—2009）

~190)
~240)
~360)

RAG三要素

构建高效知识库的核心支柱

数据质量

Garbage in, garbage out

检索精度

找到最相关的信息

生成质量

自然流畅的回答

RAG不仅是技术，更是知识管理的艺术

浙江大学 ZHEJIANG UNIVERSITY | 图书馆

感谢垂听

知识库构建：从理论到实践的思考

RAG	实践	未来
技术核心	经验分享	持续优化