**Kbot 3.0**

文档修订

| 版本 | 修改日期 | 作者 | 备注 |
| --- | --- | --- | --- |
| 0.1 | 2015-06-09 | Justin Zou | 建立大纲 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

目录

[文档修订 2](#_Toc200551955)

[1 关于Kbot 3.0 4](#_Toc200551956)

[2 需求 4](#_Toc200551957)

[2.1 整体功能 4](#_Toc200551958)

[2.2 整体工作流程 5](#_Toc200551959)

[2.3 后台服务 6](#_Toc200551960)

[2.3.1 知识库管理 6](#_Toc200551961)

[2.3.1.1 本地数据上传 7](#_Toc200551962)

[2.3.1.2 对象存储数据上传 7](#_Toc200551963)

[2.3.1.3 WebPageURL 7](#_Toc200551964)

[2.3.1.4 数据审批 7](#_Toc200551965)

[2.3.1.5 数据解析状态查询 7](#_Toc200551966)

[2.3.1.6 离线大批量数据上传接口 7](#_Toc200551967)

[2.3.1.7 数据同步接口 7](#_Toc200551968)

[2.3.2 数据解析 7](#_Toc200551969)

[2.3.2.1 语义解析引擎 8](#_Toc200551970)

[2.3.2.2 Markdown格式引擎 9](#_Toc200551971)

[2.3.2.3 QA引擎 9](#_Toc200551972)

[2.3.2.4 SelfSplitChunk引擎 9](#_Toc200551973)

[2.3.2.5 Excel解析 10](#_Toc200551974)

[2.3.2.6 文本摘要引擎 10](#_Toc200551975)

[2.3.3 数据向量化 10](#_Toc200551976)

[2.3.4 模型管理 10](#_Toc200551977)

[2.3.5 Query理解 10](#_Toc200551978)

[2.3.6 检索模块 10](#_Toc200551979)

[2.3.7 接口管理 11](#_Toc200551980)

[2.3.8 Prompt管理 11](#_Toc200551981)

[2.3.9 GraphRAG 11](#_Toc200551982)

[2.3.10 LightRAG 11](#_Toc200551983)

[2.3.11 日志管理 11](#_Toc200551984)

[2.3.12 系统配置 11](#_Toc200551985)

[2.4 前端部分 11](#_Toc200551986)

[2.4.1 登录模块 12](#_Toc200551987)

[2.4.2 首页Dashboard 12](#_Toc200551988)

[2.4.3 多语言支持 12](#_Toc200551989)

[2.4.4 Chat模块 13](#_Toc200551990)

[2.4.5 知识库管理 13](#_Toc200551991)

[2.4.5.1 数据集管理 13](#_Toc200551992)

[2.4.5.2 数据审批 13](#_Toc200551993)

[2.4.5.3 数据解析 13](#_Toc200551994)

[2.4.6 同义词管理 13](#_Toc200551995)

[2.4.7 提示词管理 13](#_Toc200551996)

[2.4.8 用户交互 13](#_Toc200551997)

[2.4.9 用户管理 14](#_Toc200551998)

[2.4.10 系统管理 14](#_Toc200551999)

[2.4.11 Release Note 14](#_Toc200552000)

[3 设计 14](#_Toc200552001)

[4 开发 14](#_Toc200552002)

[5 运维 14](#_Toc200552003)

[6 参考 14](#_Toc200552004)

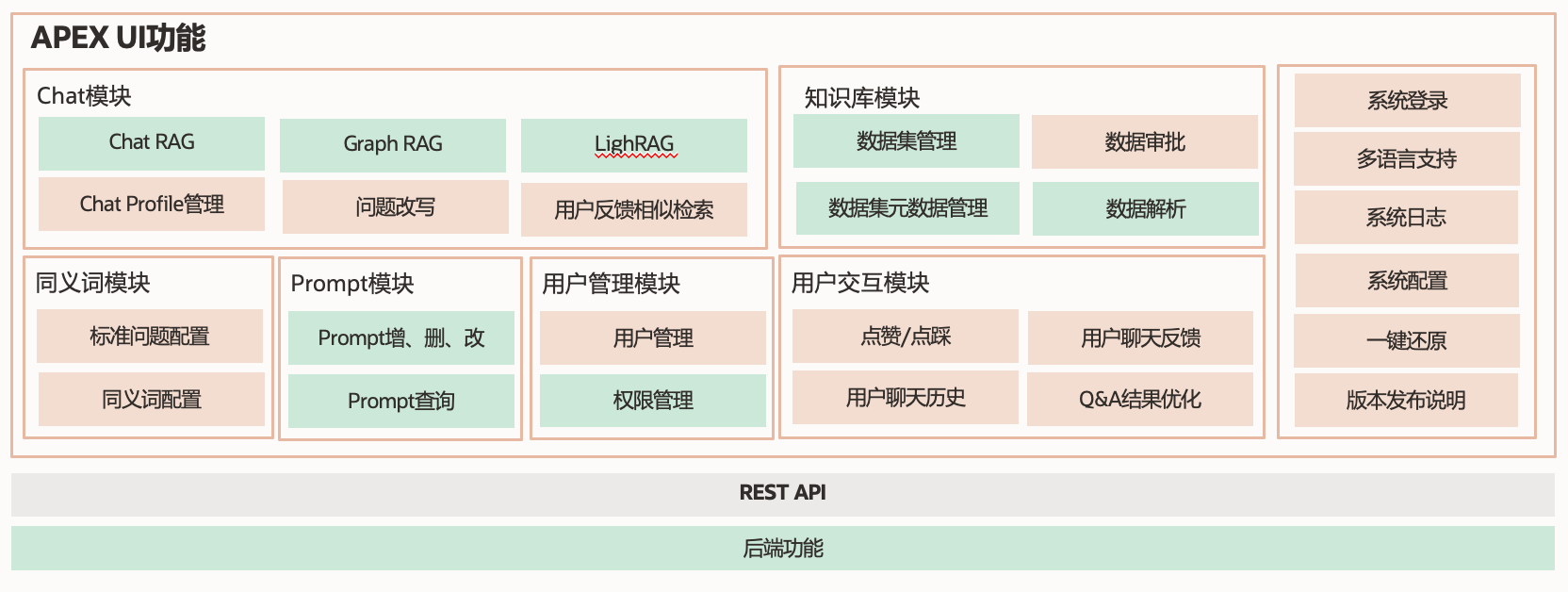
# 关于Kbot 3.0

Kbot 3.0，是基于大模型的知识构建平台，能够结合企业专属数据，助力企业更快更高效地搭建RAG。

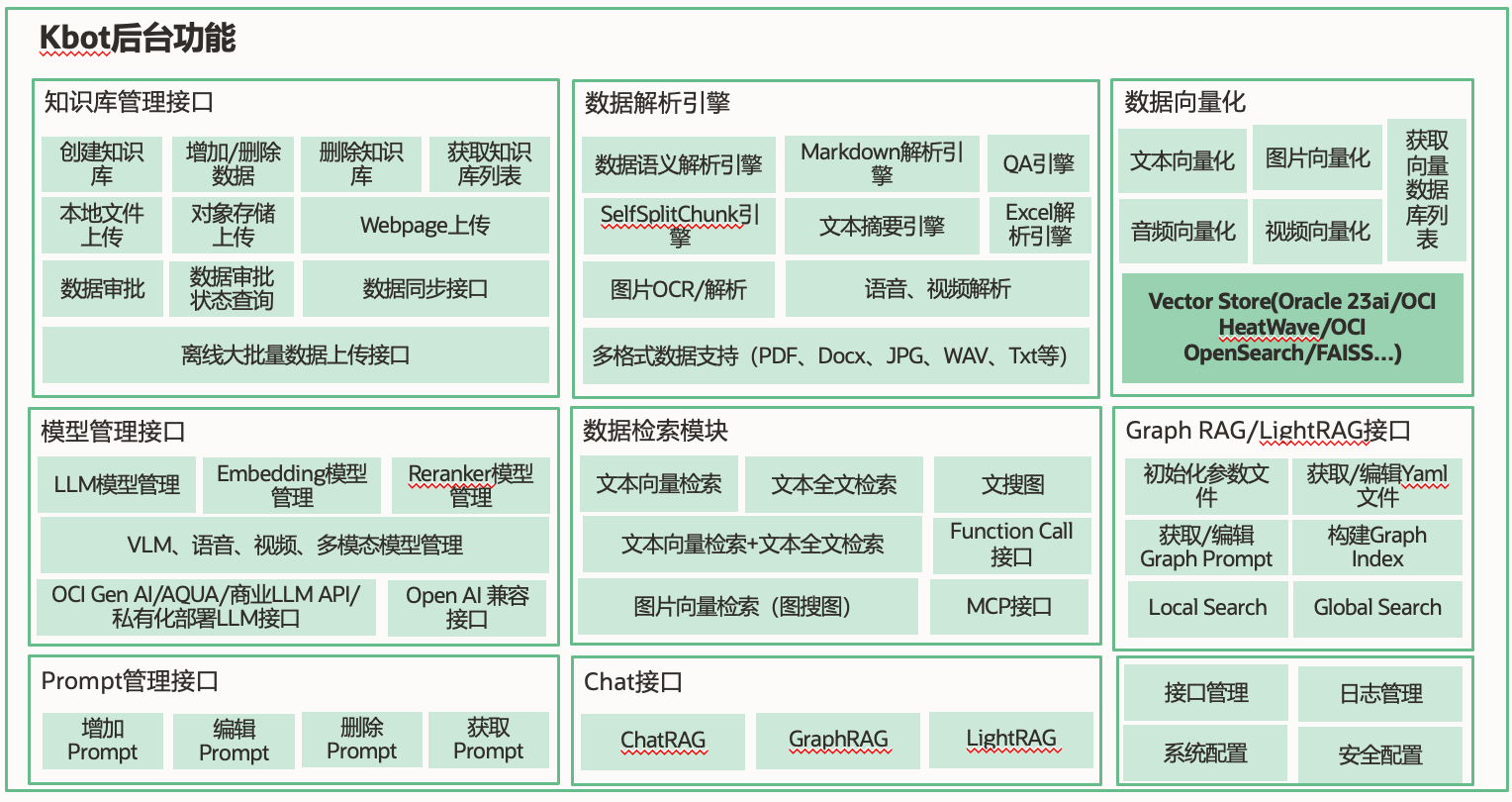
# 需求

## 整体功能

前端整体功能：

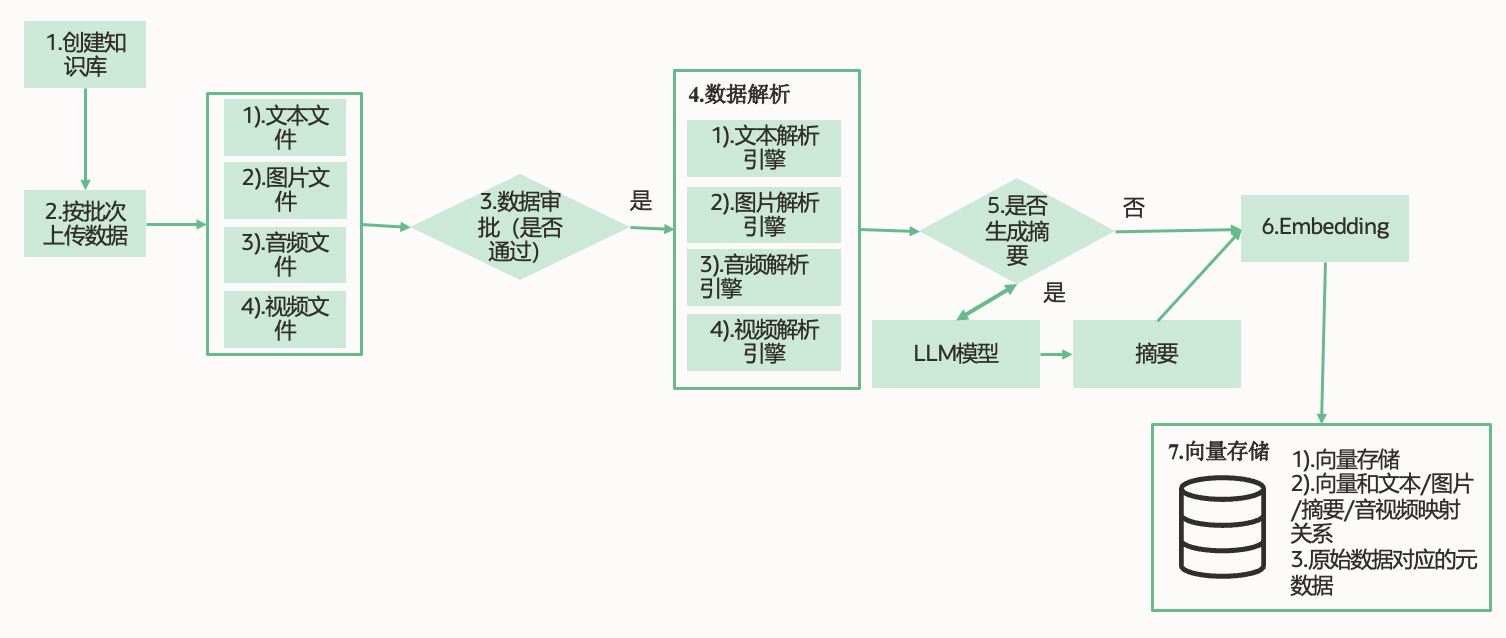


后台整体功能：

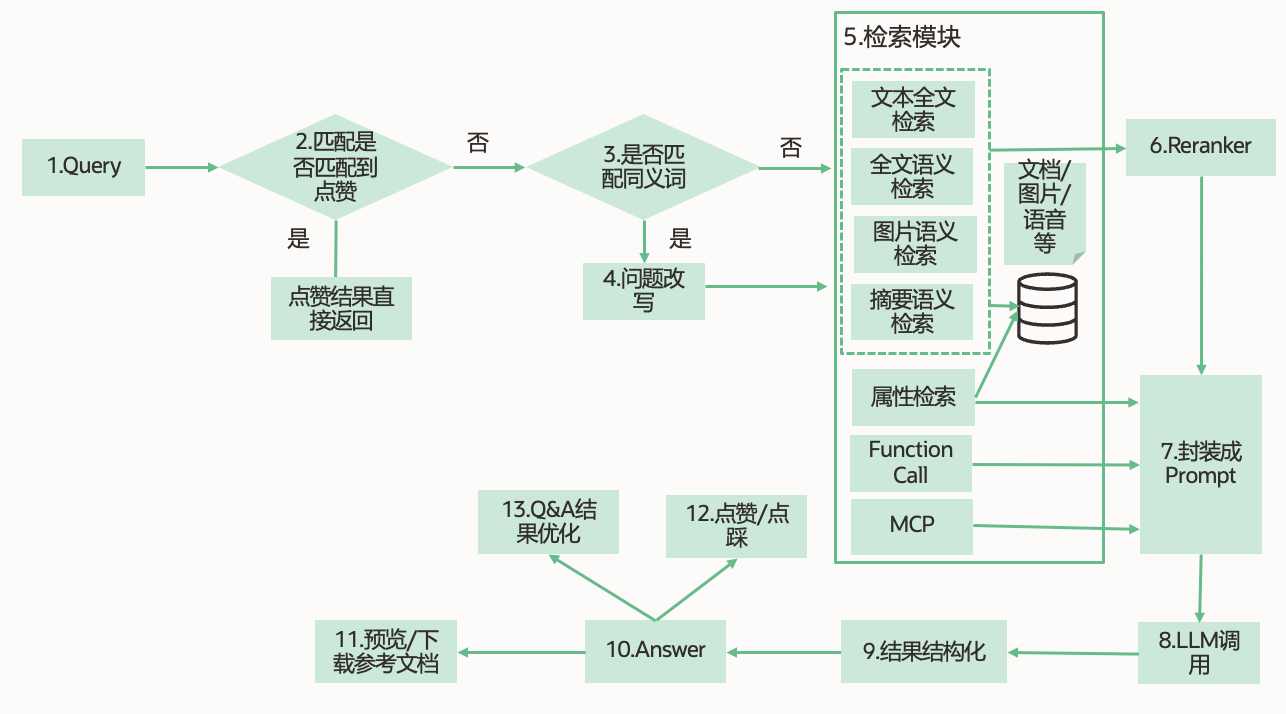


## 整体工作流程

**创建知识库流程：**



**RAG流程（文搜文）：**



## 后台服务

### 知识库管理

知识库管理是可以支持用户自助构建一个知识库，并按批次上传数据，在上传的数据中，可以支持增加元数据，对上传的数据进行审批，最后对审批之后的数据选择数据解析引擎进行后台异步解析，获取数据解析结果。数据管理的流程图下：

数据管理包括如下接口：

* 本地数据上传
* 对象存储数据上传
* WebPage URL
* 数据审批
* 数据解析状态查询

数据支持分批次上传，同一个批次共享如下参数：

* Embedding模型
* 解析引擎

支持的数据格式：

* 文本类：PDF、Docx、Doc、PPT、Excel、Markdown、Txt
* 图片类：JPG、PNG、JPEG、BMP
* 语音类：wav、mp3
* 视频类：mp4、wmv

#### 本地数据上传

#### 对象存储数据上传

#### WebPageURL

#### 数据审批

#### 数据解析状态查询

#### 离线大批量数据上传接口

#### 数据同步接口

### 数据解析

数据解析整体逻辑：

1.同一批次数据自动根据文件后缀分类，分成如下几类，并放到对应类型的文件夹子目录：

* 文本类
* 图片类
* 语音类
* 视频类

2.选择对应的数据解析引擎对不同数据类型转成文本。

文本类可选的数据引擎有：

* + 语义解析引擎：自动抽取文本/表格/图片，并分别对文本/表格/图片和文档建立mapping关系。
  + Markdown：遍历段落层次关系，对每一个层级段落，生成一个chunk。
  + QA：一个文件一个chunk
  + SelfSplitChunk：自定义chunk size和chunk over lap
  + Excel解析

图片类，解析引擎有：

* OCR引擎：PaddleOCR
* VLM引擎：QwenVL

语音类，解析引擎有：

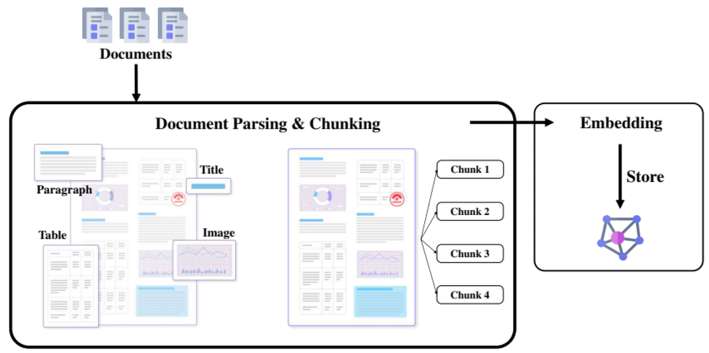
* Whisper

3.第二步骤生成的所有文本数据，进行embedding。  
4.Embedding之后，持久化到Vector DB中。

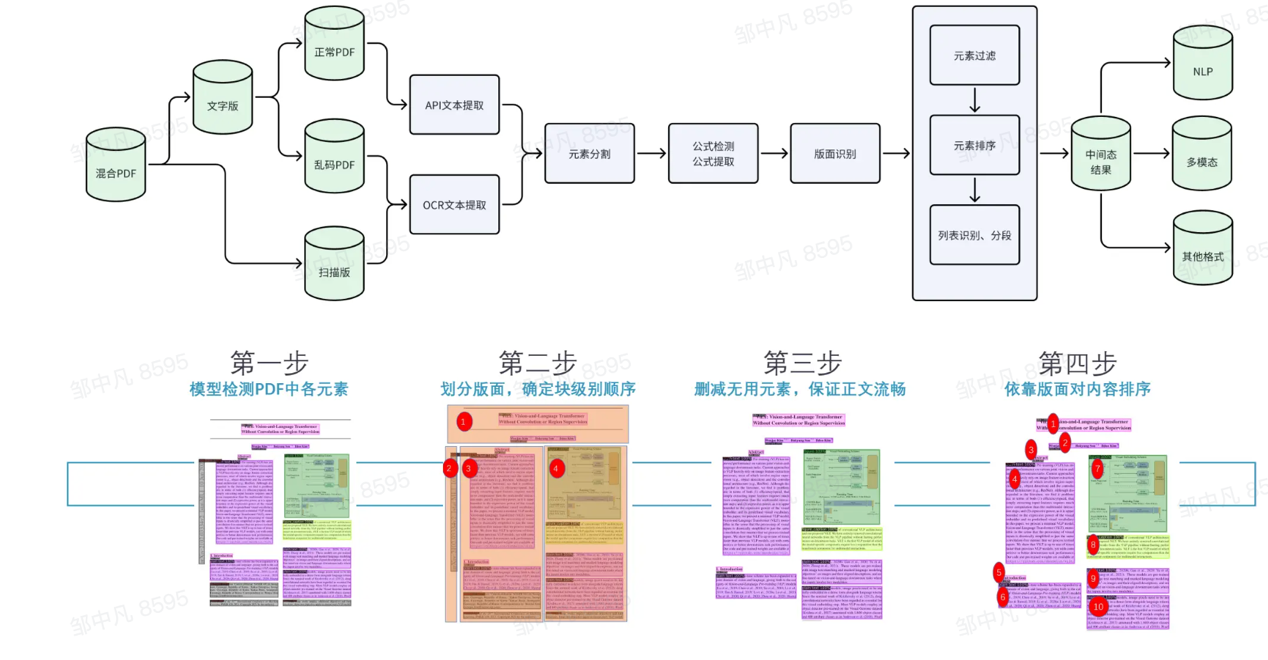
这里的Embedding包括对文本数据的Embedding，也包括对图片的Embedding（用于支持图搜图/或者文搜图）。

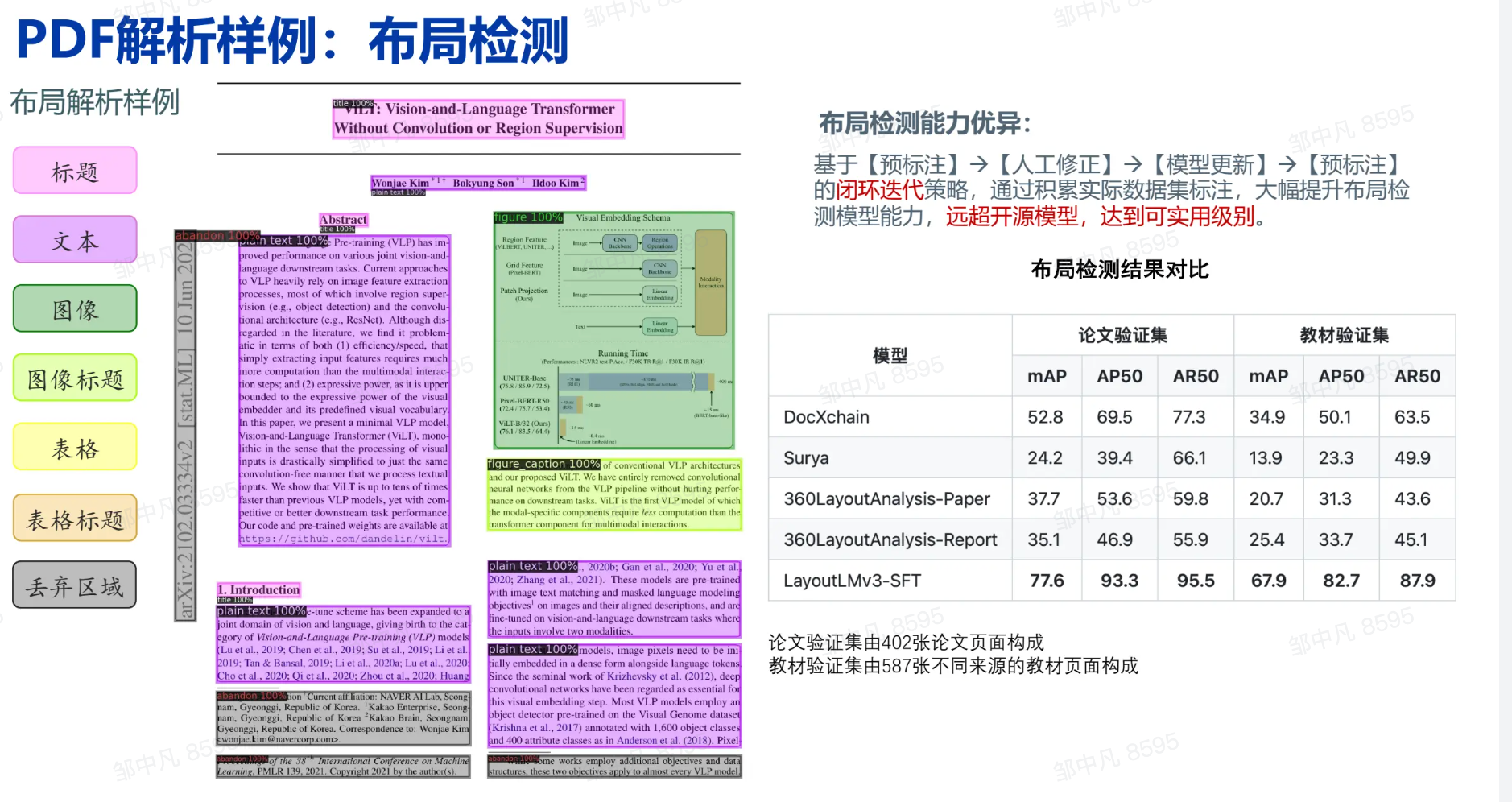
5.这个数据解析对外是一个原子接口，且是异步的。需要另一个配套接口，获取这个数据解析的状态信息。这个状态信息，能知道数据解析的所处阶段以及处理情况（阶段、开始/处理中/成功/失败，处理的数量等）。

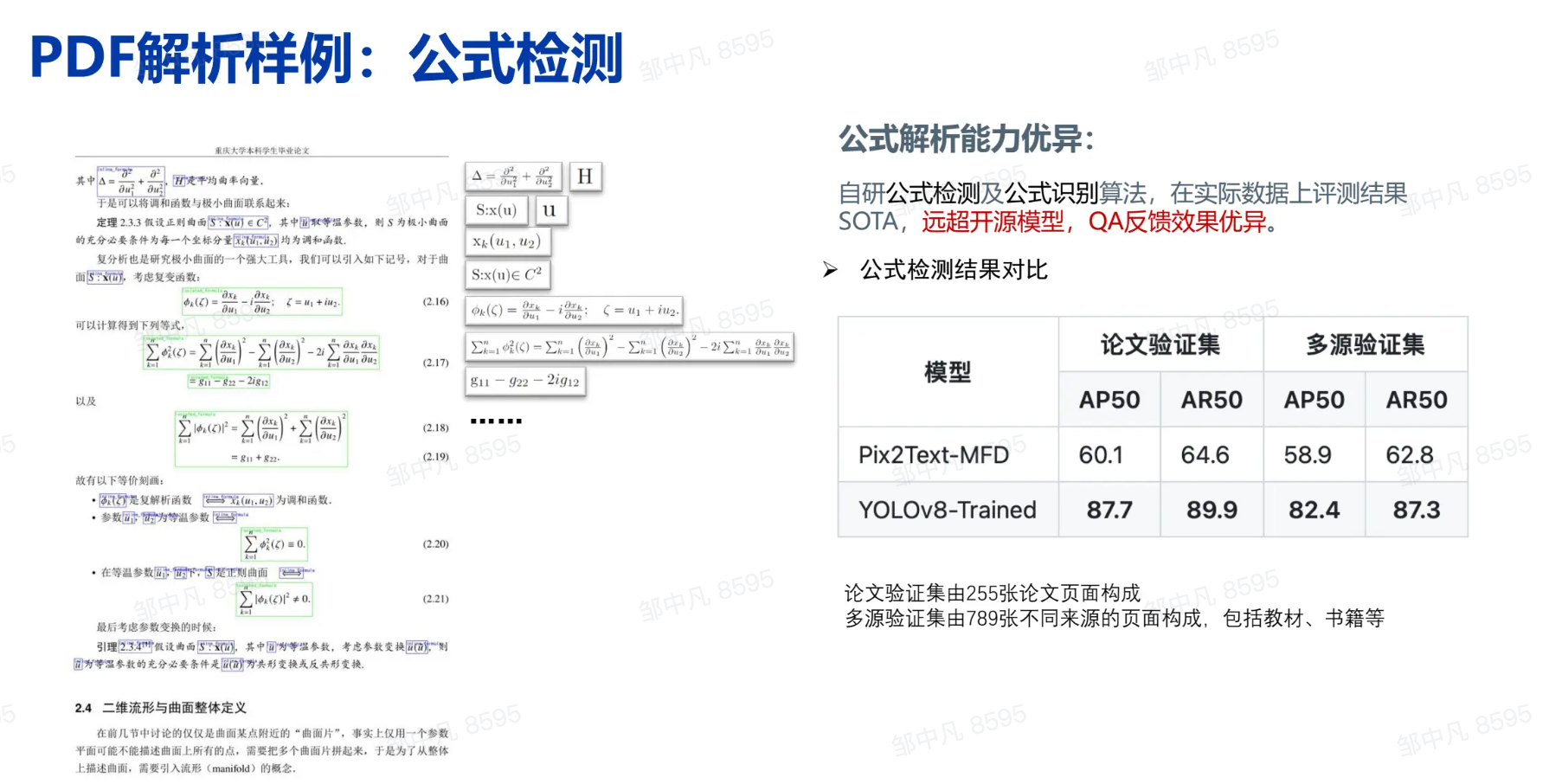
#### 语义解析引擎



参考：<https://aicarrier.feishu.cn/file/SknGbA2nqoYodbxNjYRcqeUsngf>







#### Markdown格式引擎

遍历段落层次关系，对每一个层级段落，生成一个chunk。

#### QA引擎

一个文件一个chunk

#### SelfSplitChunk引擎

自定义chunk size和chunk over lap

#### Excel解析

#### 文本摘要引擎

对文本chunk数据生成摘要，并对摘要进行Embedding，持久化。

### 数据向量化

支持对不用数据类型，进行向量化：

1.文本编码：bge-m3/cohere embed/e5-larger-v2

2.图像编码：RestNet、ViT、CLIP ViT

3.音频编码：Whisper、CLAP

4.视频编码：CMVC

向量化后的数据持久化到Vector DB中

### 模型管理

提供模型管理接口，对模型进行配置管理，支持对模型进行增加、修改、删除、查询等基本功能。支持如下类型的模型管理：

* LLM模型管理
* Embedding模型管理
* VLM模型管理
* ReRanker模型管理
* 语音模型
* 视频模型
* 多模态模型

### Query理解

意图识别

Query改写

同义词管理

### 检索模块

检索模块包括如下子功能：

* 文本向量检索
* 文本全文检索
* 文本向量检索+文本全文检索
* 图片向量检索（图搜图）
* 文搜图
* 调用Function Call
* 调用MCP
* 摘要检索

### 接口管理

所有的后台接口，需要增加安全认证。

第二阶段，增加接口的发布/注册和授权管理。

### Prompt管理

支持对Prompt进行管理，包括增加、修改、删除、查询。

### GraphRAG

复用GraphRAG现有接口

### LightRAG

复用LightRAG现有接口

### 日志管理

统一日志管理

### 系统配置

应用配置app\_config.py：配置应用相关，包括如下信息：

* + 接口服务器IP/端口/上下文路径
  + 应用日志级别配置
  + SSL key配置
  + Vector DB 链接信息

环境变量配置app.env

系统配置文件使用YAML？TOML？--考虑YAML

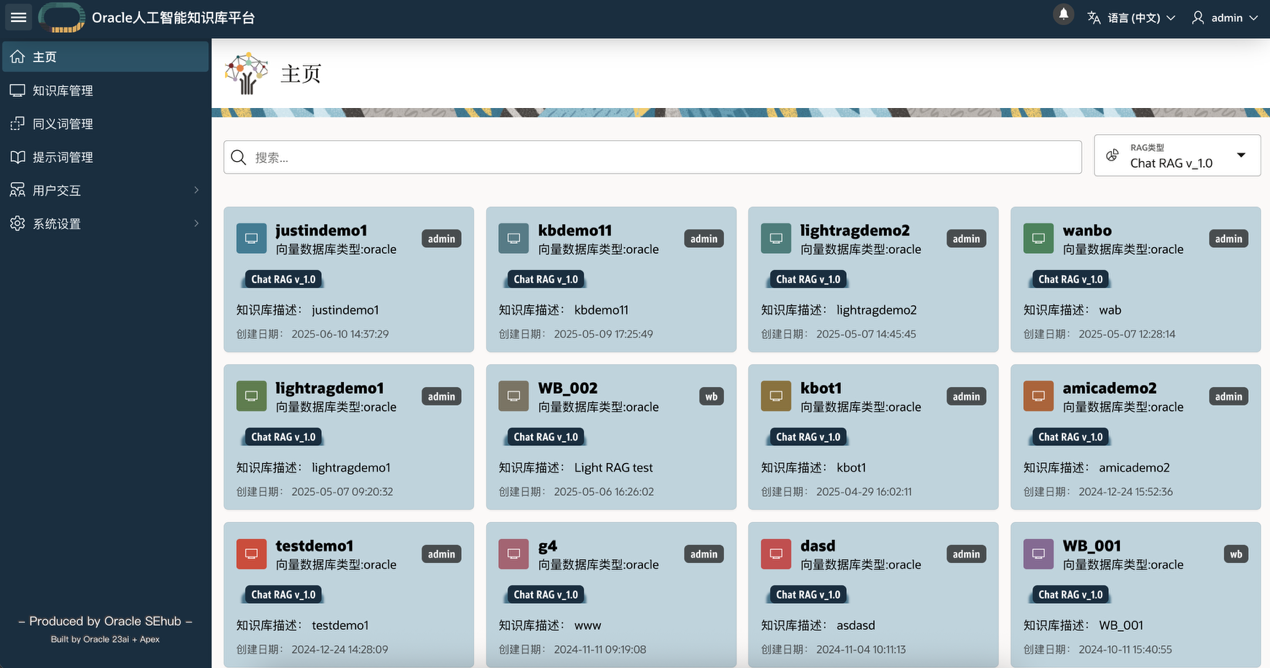
## 前端部分

在现有功能基础上，重新设计知识库管理模块，需要支持知识库的数据上传、审批和数据解析3个子功能，这3个功能是独立的3个子菜单。同时在数据上传功能，需要增加元数据。

### 登录模块



### 首页Dashboard



### 多语言支持

### Chat模块

Chat模块包括：

* ChatRAG
* GraphRAG
* LightRAG

### 知识库管理

知识库管理包括如下子功能：

* 数据集管理（需要支持元数据的追加/编辑）
* 数据审批
* 数据解析

#### 数据集管理

#### 数据审批

#### 数据解析

### 同义词管理

### 提示词管理

### 用户交互

用户交互包括如下3个子菜单：

* + - 用户聊天反馈
    - 用户聊天历史
    - Q&A结果优化

### 用户管理

用户管理

权限管理

### 系统管理

系统管理包括：

* 系统配置
* 一键还原
* 系统日志

### Release Note

# 设计

# 开发

# 运维

# 参考

<https://mp.weixin.qq.com/s/dtOHlposE9J7o6mKiACAsA>

https://aicarrier.feishu.cn/file/SknGbA2nqoYodbxNjYRcqeUsngf