面向开源操作系统社区的智能化服务能力构建

齐国强　openEuler社区贡献者、华为2012 LAB 高级工程师
黄河清　华为计算产品线 高级工程师

开放原子开源基金会 OPENATOM FOUNDATION | OpenEuler

目录

- 从需求出发：openEuler社区高度重视用户&开发者服务
- 从能力思考：AIGC从辅助内容生产到主导内容生产
- 以数据为起点：领域数据集及质量成为体验关键因素
- 以模型为主体：基于算力、场景选型开源LLM，SFT、RLHF加强业务属性
- 以RAG、Agent为触点：对接开放知识库，动态知识自动感知
- 以 Instruction Compliance 为终点

从需求出发：openEuler社区高度重视用户&开发者服务

社区规模庞大，协作活动频繁，面向海量用户、开发者、单位成员提供开源服务，效率与体验保障至关重要

16854
贡献者

2126058
社区用户

103
特别兴趣小组

22
商用OSV

145.5K
合并请求

73.3K
需求&问题

2296.2K
评审

1335
单位成员

从能力思考：AIGC从辅助内容生产到主导内容生产

AIGC发展趋势从文本转移到多模态和多媒体，接入生产流程

文本：对话、翻译、编程等，如 ChatGPT,PaLM2；

音频：语音识别、合成、音乐生成、语音风格转换等；

图像：图片编辑、场景化图片生成、3D图像等，如 Midjourney、Stable Diffusion、Adobe Firefly；

视频：视频剪辑、视频内容生成、视频画质增强、视频风格迁移等；

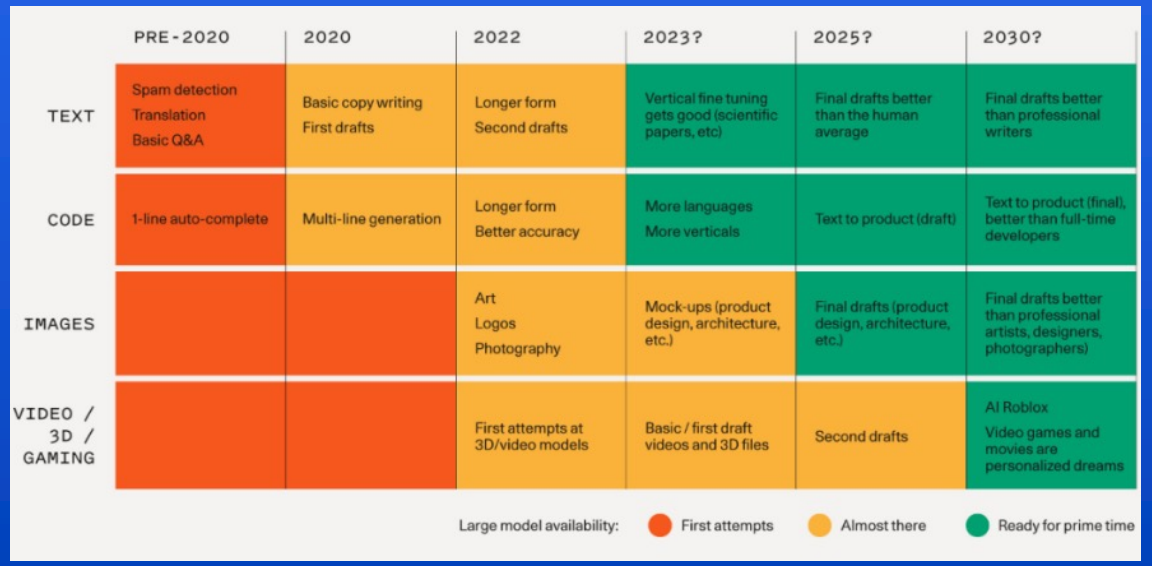
内容生产路径从PGC、UGC到AIGC

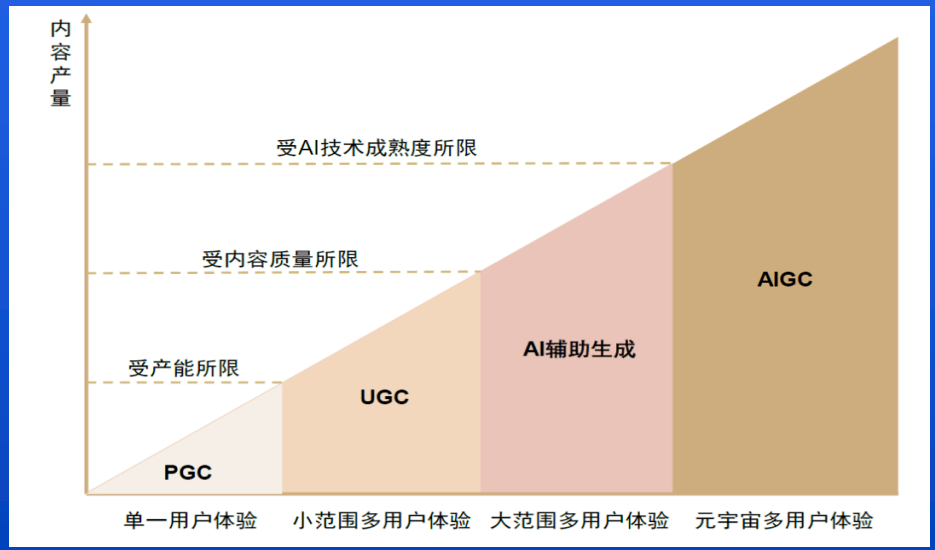
PGC：电影、电视和游戏等；专业团队生产；质量高；效率低，门槛高；

UGC：短视频、播客、社交媒体；普通用户生产，繁荣；内容质量不高；

AIUGC：AIGC 辅助用户生产；人在关键环节依然需要输入指令。

AIGC：完全AI自主文字创作，图片创作。

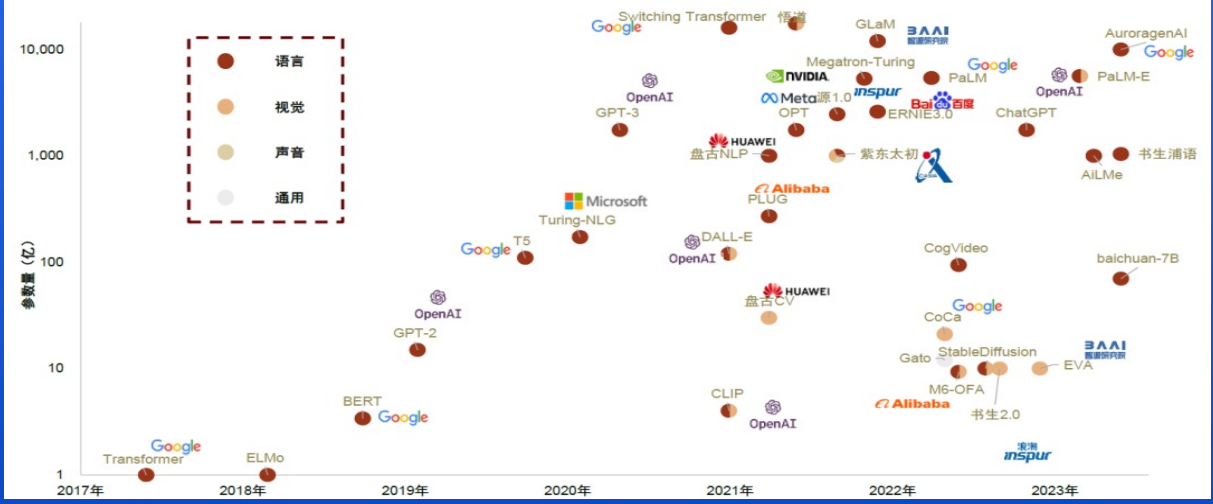


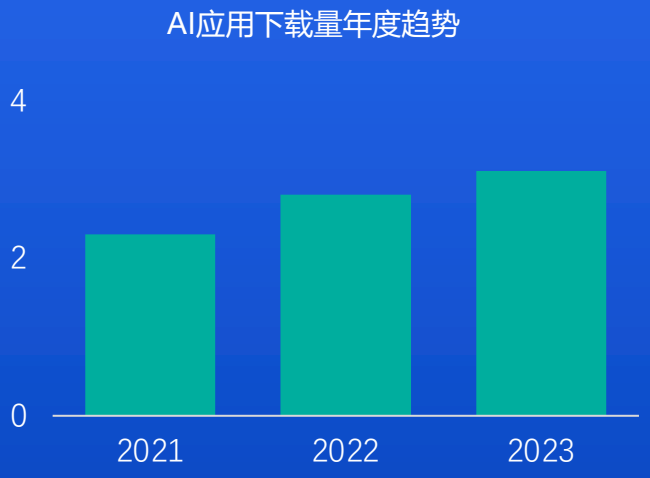


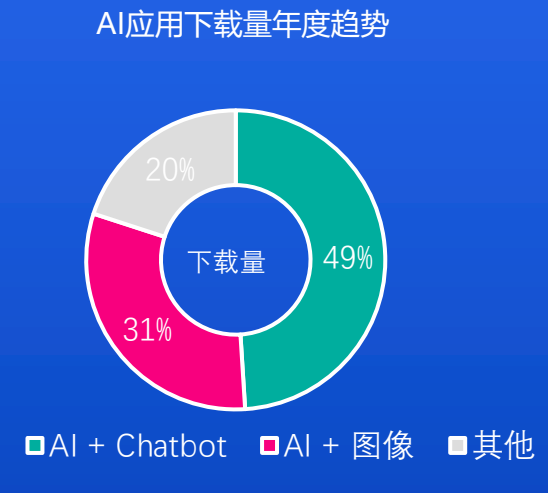
从能力思考：AIGC从辅助内容生产到主导内容生产

聊聊涌现

- **LLM模型涌现：**模型在数量、模型参数、研发投入和模型能力表现上均快速增长；
- **AI应用涌现：**23年上半年AI应用下载量同比增长**114%**，突破三亿次，**超出22年全年水平**；23H1全球下载前100名的AI应用中，**49%** 为AI+ChatBot应用。在ChatBot和图形生成及处理细分领域，AI技术得到最广泛地应用。







从能力思考：AIGC从辅助内容生产到主导内容生产

LLM 能力完全匹配社区需求场景，基于开源模型微调为可行方案

对比维度		FAQ检索型	知识图谱型	LLM型（类chatGPT）	
				自己构建	外部服务
问答能力	语义理解能力	无	弱	强	
	内容生成能力	无	弱	强	
	泛化能力	无	弱	强	
	上下文感知能力	无	无	强	
	能力满足度	0%	50%	100%	
复现难度	专业难度	低	较高，与高校合作	高，与领域顶尖团队合作；开源；	低
	资源需求	低	低	高，预训练难度大，SFT可接受	平台开放程度、定制化成本
	人员投入	少	较高，知识图谱建立需要人力	高，大量人力投入数据制作	少
	业界方案	成熟	成熟	热门领域	缺少行业/领域知识
潜在风险	数据安全	隐私保护		SFT数据来源可靠	数据（跨境）传输、泄露风险
	技术治理			满足国家/公司/社区 AI治理需求	

以数据为起点：领域数据集及质量成为体验关键因素

建立数据到语料自动化能力，基础人工标注平台，助力业务数据回流和领域模型落地

5分钟生成您自己的语料

1. 选择Loader
覆盖10+类型

2. 设置切片
自定义切片策略

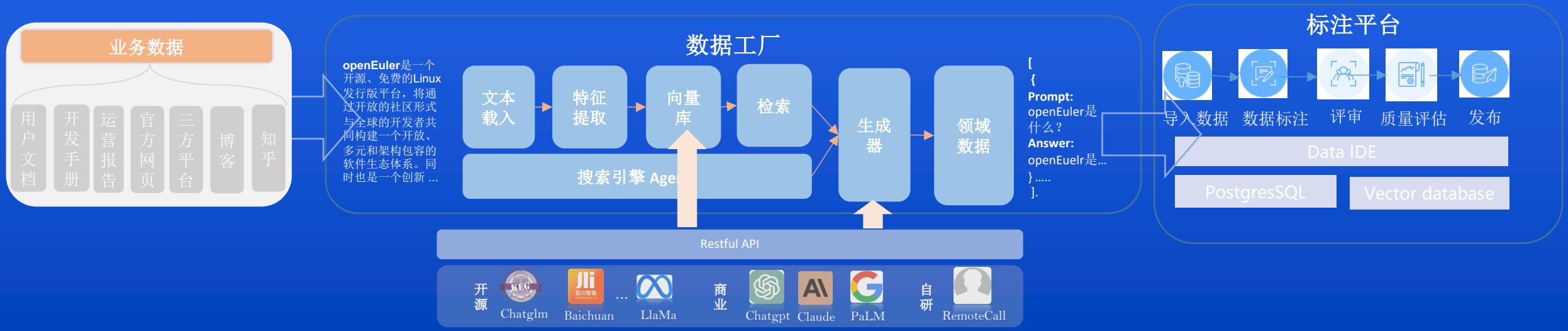
3. 选择embedding模型
支持5+主流模型

4. 问题生成
批量生成，在线审核

5. 检索器
多种检索策略支持

6. 对话模型选择
50+开源模型及商用API

7. 答案生成
批量生成，在线确认



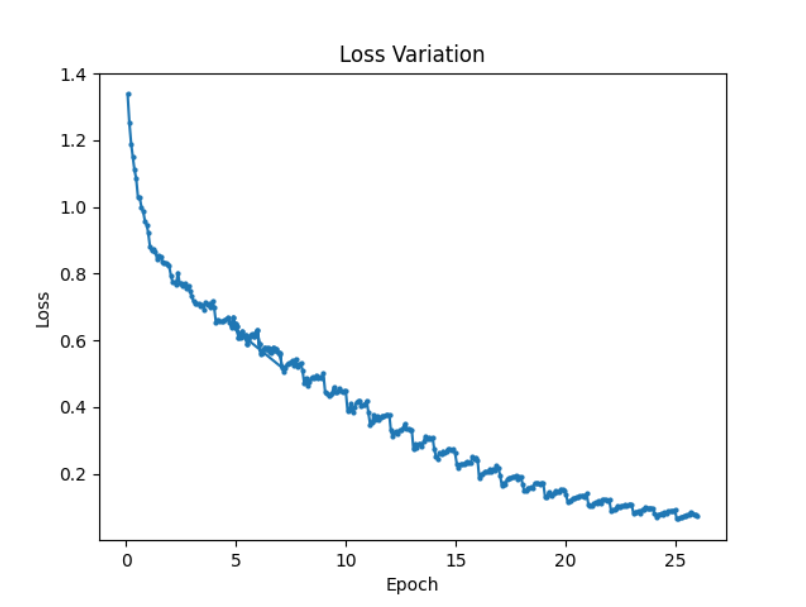
以模型为主体：基于算力、场景选型开源LLM，SFT、RLHF加强业务属性

自动化指标计算与人工判断结合的模型能力评测，结合算力选型开源基础LLM

对比维度	Baichuan2 7B	ChatGLM3 6B	LLaMA2 7B	Baichuan2 13B	Gpt3.5 turbo
上下文感知长度	4K	8K	4K	4K	4K
语义理解能力	59.6	61.7	50.9	**63.1**	64.6
逻辑推理能力	45.7	**53.5**	41.4	50.1	64
通用知识能力	49	43.6	**54.3**	51.9	64.5
领域知识能力	26	20.3	18.7	**27.5**	32.6

以模型为主体：基于算力、场景选型开源LLM，SFT、RLHF加强业务属性

SFT训练： baichuan2 + lora + 10W QA对 (模型收敛效果好)； 知识“蒸馏”能缓解灾难性遗忘现象；



对比维度	Baichuan2 13B	Lora SFT	Lora SFT + “Distillation”
语义理解能力	63.1	↓	≈
逻辑推理能力	50.1	↓	≈
通用知识能力	51.9	↓ ↓	≈
领域知识能力	27.5	↑	↑

以模型为主体：基于算力、场景选型开源LLM，SFT、RLHF加强业务属性

SFT模型能有效感知新数据，在无预训练算力下为可行方案；RLHF暂未实施，是否是敏感信息规避方案?

openEule是什么?
openEuler是一个开源项目，由华为发布并由开放原子开源基金会孵化和运营。它是面向数字基础设施的操作系统，支持多种应用场景，包括服务器、云计算、边缘计算和嵌入式系统。openEuler旨在提供安全、稳定和易用的操作系统，并通过为应用提供确定性保障能力，支持OT（运营技术）领域应用和OT与ICT（信息通信技术）的融合。它提供了一个平台，让企业级用户获得安全、稳定和可靠的操作系统。每两年推出一个LTS（长期支持）版本，为企业用户提供持久的支持。华为于2021年11月将openEuler的全量代码等捐赠给开放原子开源基金会，标志着openEuler逐步发展为一个产业共建和社区自治的开源项目。

isula如何启动容器?
isula是一个开源的容器运行时解决方案，它可以与docker兼容，并支持运行docker镜像。要启动一个容器，可以使用以下命令：
```
Isula create [OPTIONS] IMAGE [COMMAND] [ARG...]
```
其中，[OPTIONS]是一些可选参数，IMAGE是要启动的容器镜像，[COMMAND]是容器启动后要执行的命令，[ARG...]是命令的参数。例如，要启动一个基于ubuntu 20.04镜像的容器，并在容器内执行"/bin/echo 'Hello World'"命令，可以使用以下命令：
```
Isula create -tid --network=host --name testubuntu ubuntu:20.04 /bin/echo 'Hello World'
```
这将在当前主机的网络中创建一个名为testubuntu的容器，并在容器内执行指定的命令。\
请注意，isula的具体命令和选项可能会有所不同，具体取决于你使用的isula版本和配置。
建议参考isula的官方文档或运行"isula help"命令来获取更详细的信息。

生成用于创建新用户并配置root权的shell指令限
创建一个新用户并配置其root权限，可以使用以下命令：
1. 创建新用户：
```
sudo useradd <username>
```
其中，`<username>`是要创建的用户名。
2. 设置用户密码：
```
sudo passwd <username>
```
系统会提示您输入新用户的密码并进行确认。
3.设置用户的root权限：
```
sudo usermod -aG root <username>
```
这将把新用户添加到root组中，从而授予其root权限。
请注意，创建新用户并配置其root权限可能需要管理员权限（使用`sudo`命令）。确保在执行这些命令时小心谨慎，并仅在必要的情况下使用。

以RAG、Agent为触点：对接开放知识库，动态知识自动感知

RAG(Retrieval Augmented Generation)是结合知识搜索和模型生成内容的框架，Agent为动作执行单元：

- **模型迭代周期长，知识时效性差，且数据全面性保障难度大，RAG和Agent成为领域模型配套能力；**
- **现有Agent相关CoT / ReAct 设计中文适配差，且小模型能力不匹配，当前只能实现单轮agent；**

```
template = """Complete the objective as best you can. You have access to the following tools:

{tools}

Use the following format:

Question: the input question you must answer
Thought: you should always think about what to do
Action: the action to take, should be one of [{tool_names}]
Action Input: the input to the action
Observation: the result of the action
... (this Thought/Action/Action Input/Observation can repeat N times)
Thought: I now know the final answer
Final Answer: the final answer to the original input question

These were previous tasks you completed:


Begin!

Question: {input}
{agent_scratchpad}"""
```

```
模板 = """尽力完成目标。您可以使用以下工具：

{tools}

请使用以下格式：

问题：您必须回答的输入问题
思考：您应该始终考虑要做什么
行动：需要采取的行动，应为[{tool_names}]之一
行动输入：行动的输入
观察结果：行动的结果
...（这个思考/行动/行动输入/观察结果可以重复N次）
思考：我现在知道最终答案
最终答案：原始输入问题的最终答案

以下是您已完成的先前任务：

开始！

问题：{input}
{agent_scratchpad}"""
```

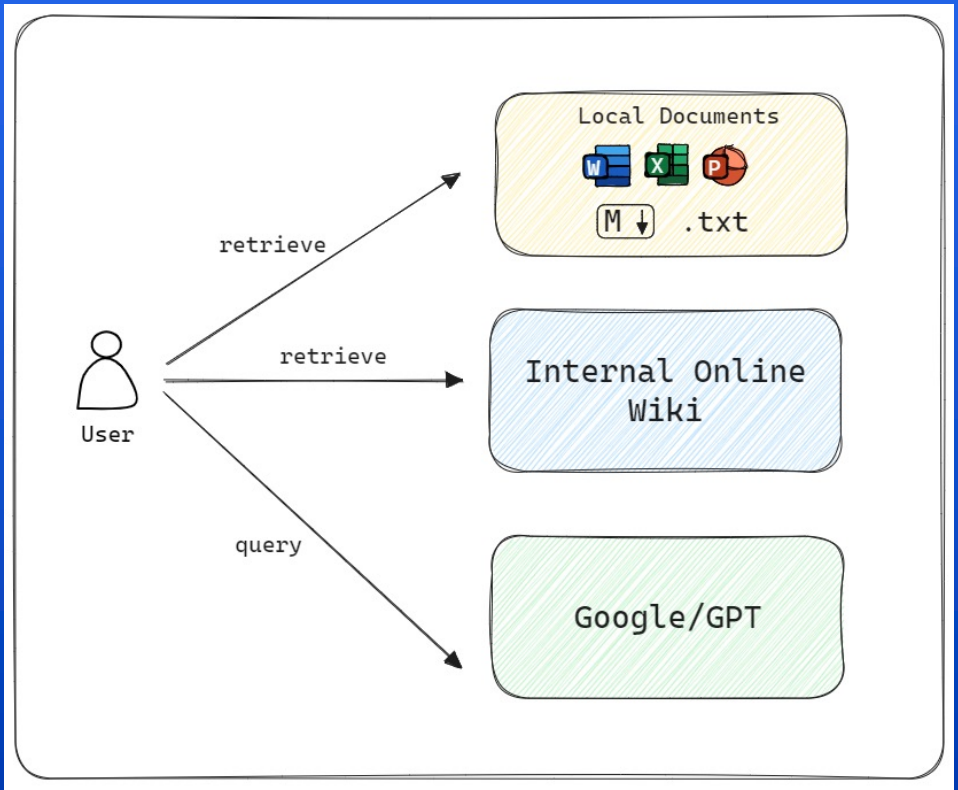
```
模板 = “”“你需要尽可能准确的回答用户问题或者完成任务。

你可以利用搜索引擎工具来获得更多的相关信息帮助更精准的回答问题和完成任务。

当你处理有关事件、时间和新闻等问题时需要调用搜索引擎工具获得参考信息。

请回答或者完成 {prompt} , 如果你需要调用搜索引擎工具获得更多信息请输出“指令：调用搜索工具”，如果不用调用搜索引擎请直接响应问题或者任务。
"""
```

RAG： 领域知识与大模型协同，为组织打造专属AI助手

Retrieval Augmented Generation

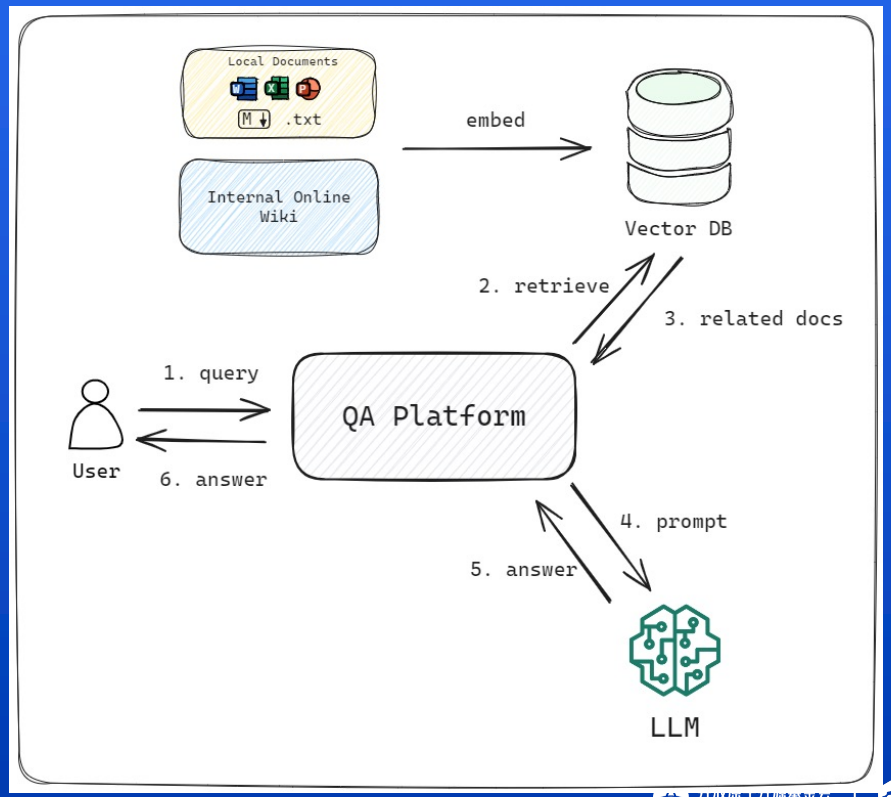
As-is

1. 大模型**不了解领域知识**
2. 大模型训练/微调成本高、时间周期长，知识**更新不及时**
3. 需要**人工检索**多个知识源



To-be

1. 构建**领域知识库**充当大模型 “海马体”
2. 领域知识库**更新便捷**，**及时**为大模型提供**最新**知识
3. 检索知识库，通过大模型**归纳汇总**

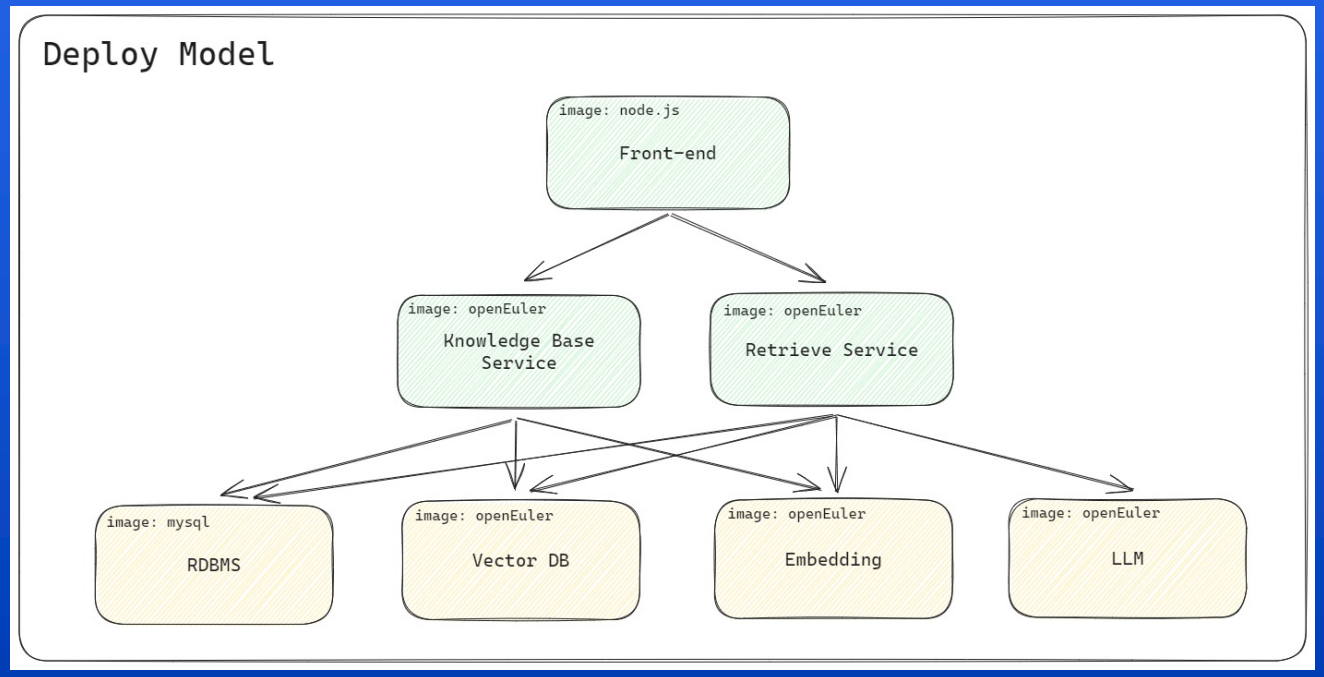


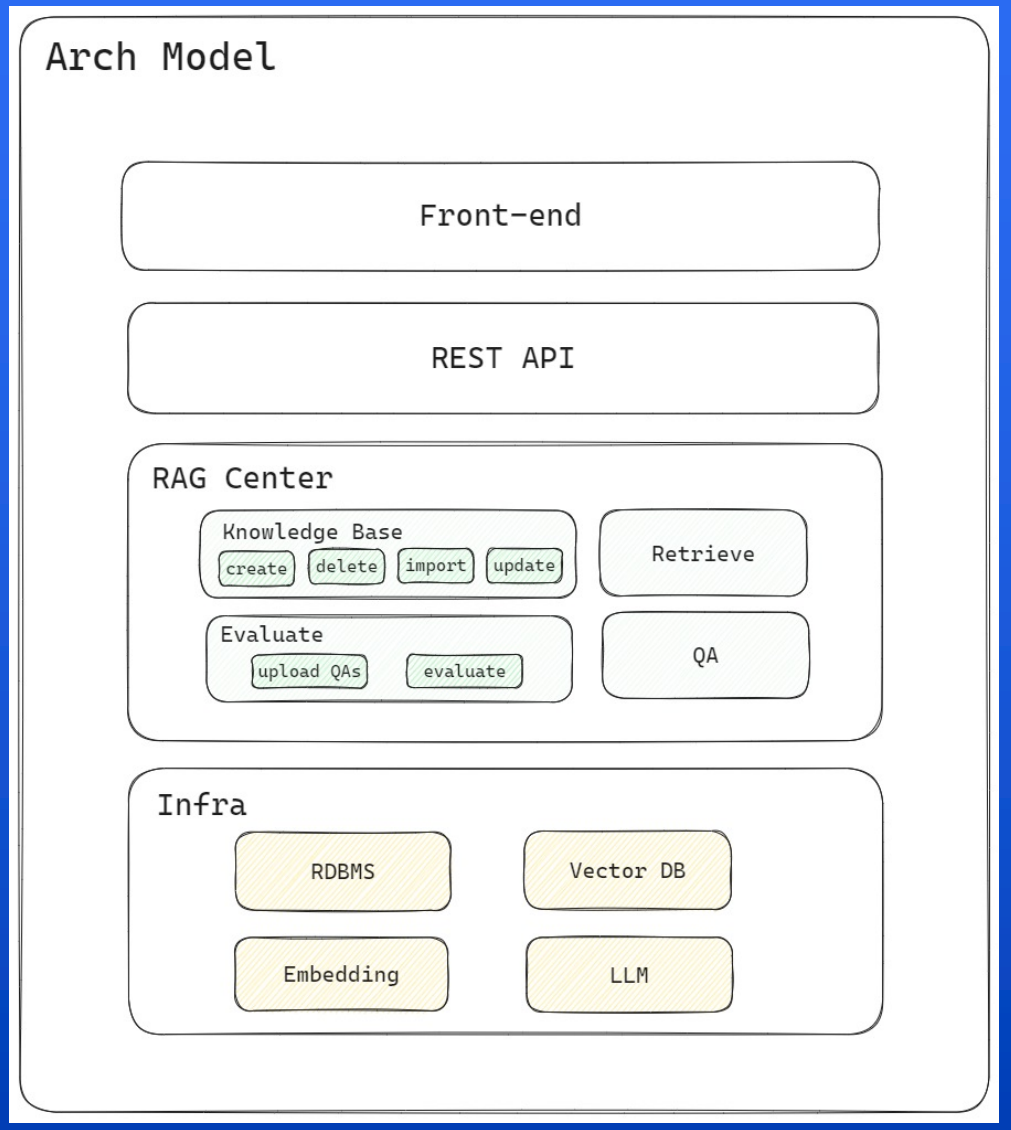
RAG：整体方案

一站式提供知识导入、效果评测、智能问答能力，支持快速构建智能问答系统

✓ 微服务架构

✓ 容器化部署



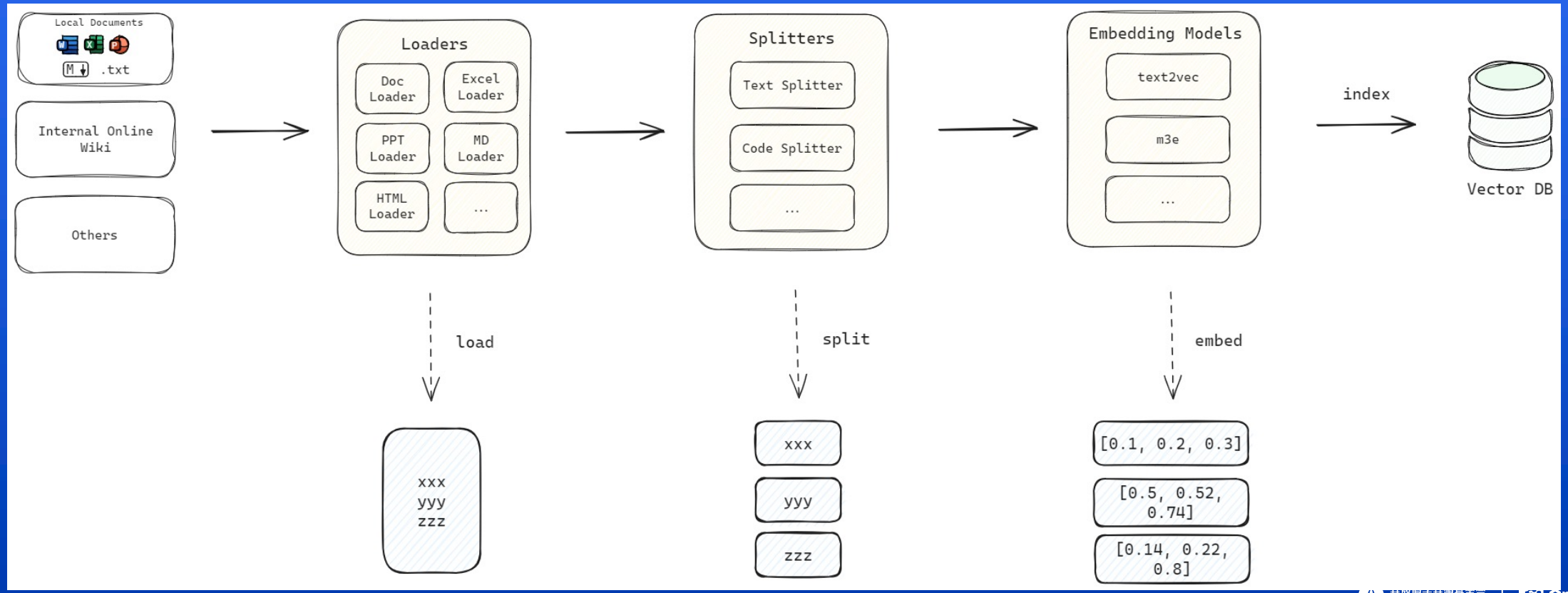


RAG：知识导入

知识来源、文件格式多样化
- 本地文件/网页/...
- word/ppt/excel/markdown/txt/html/...

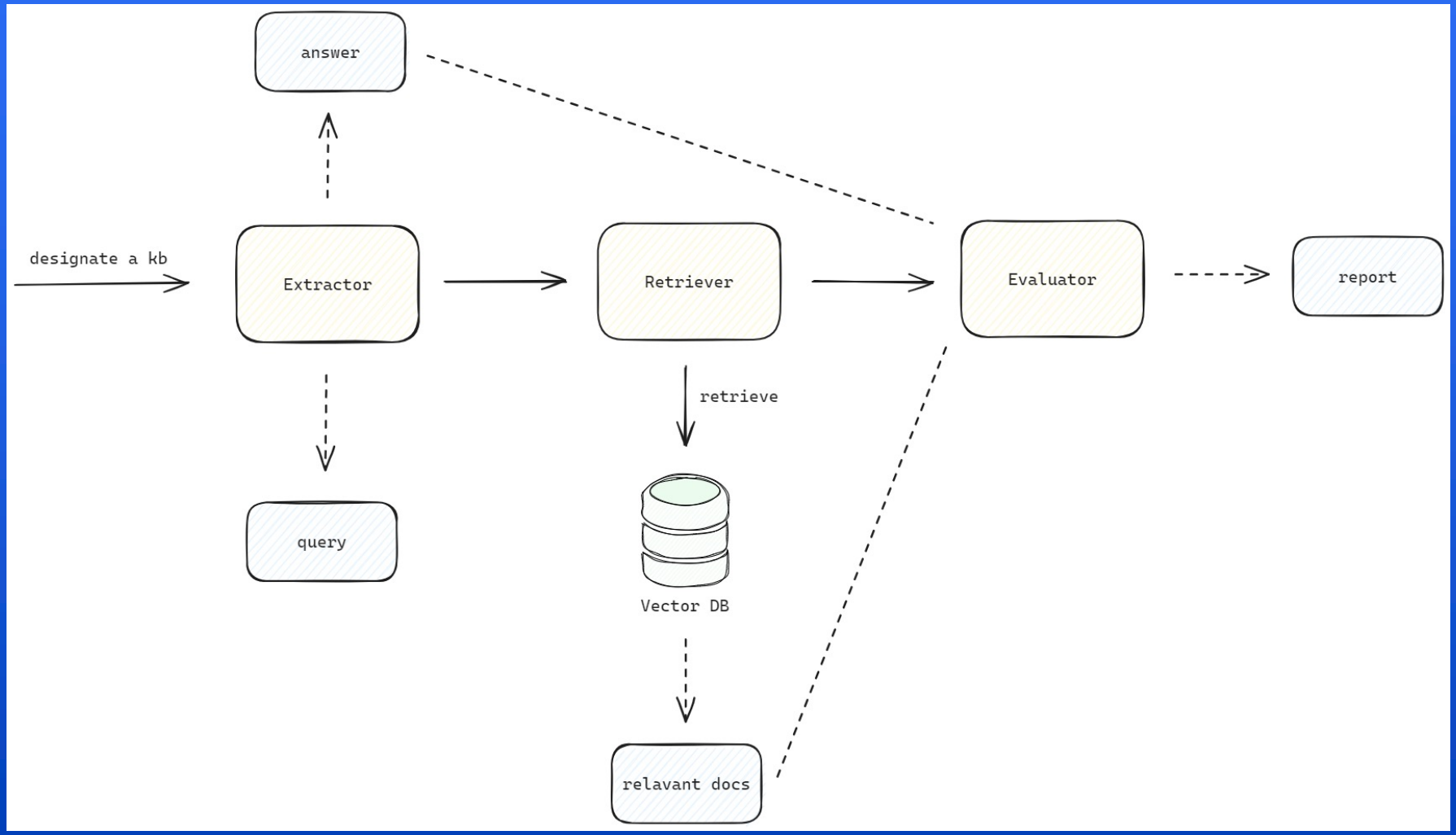
知识导入配置化
- Loader
- Splitter
- Embedding Model

任务流水线
- 任务编排
- 自动重试
- 幂等



RAG：效果评测

- ✓ 基于指定知识库**自动生成QA对**进行**检索效果评测**
  - Precision
  - Recall
  - F1 Score
- ✓ 历史评测结果展示，方便对向量化任务进行**参数调优**



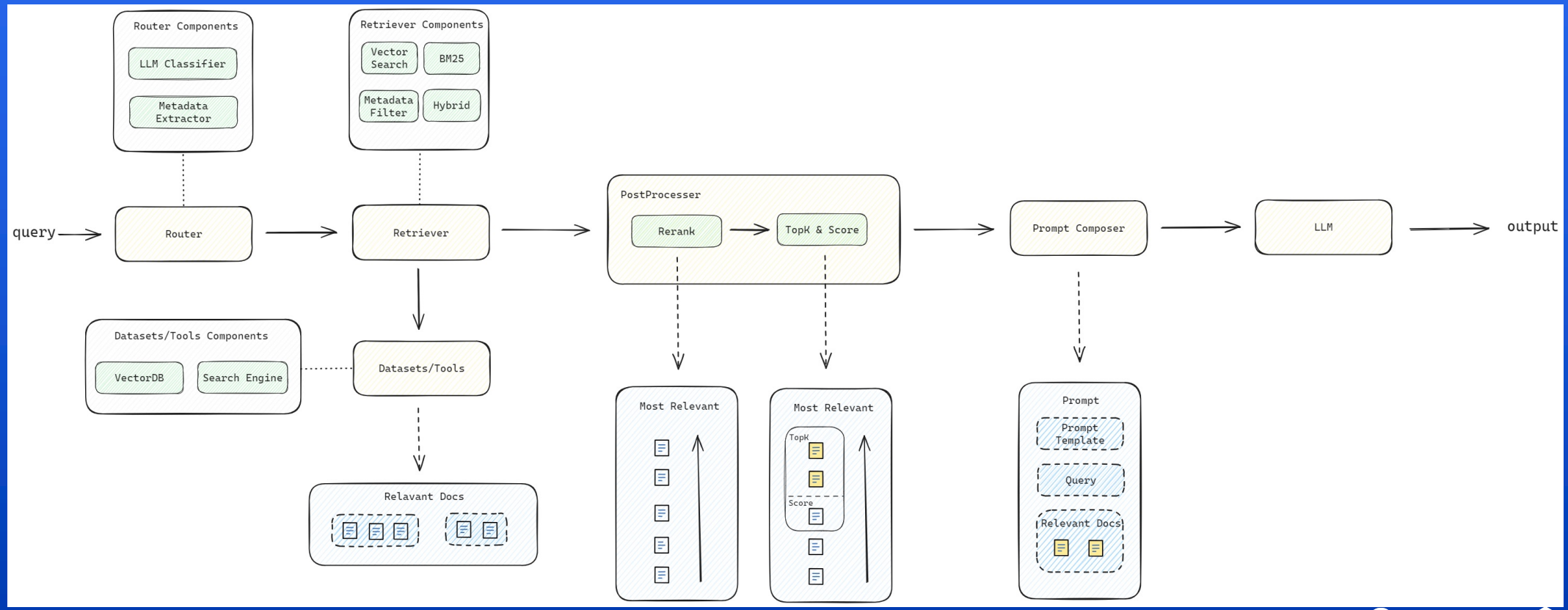
RAG：智能问答

关键技术点

- 关键词识别
- 文档切片
- 召回结果重排序
- Prompt工程

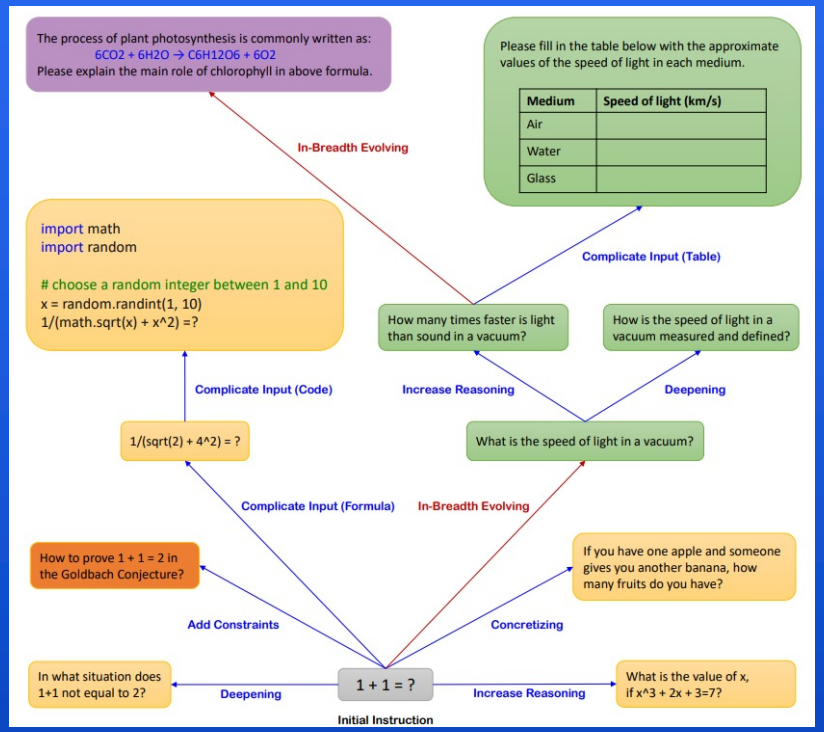
回答准确率影响要素

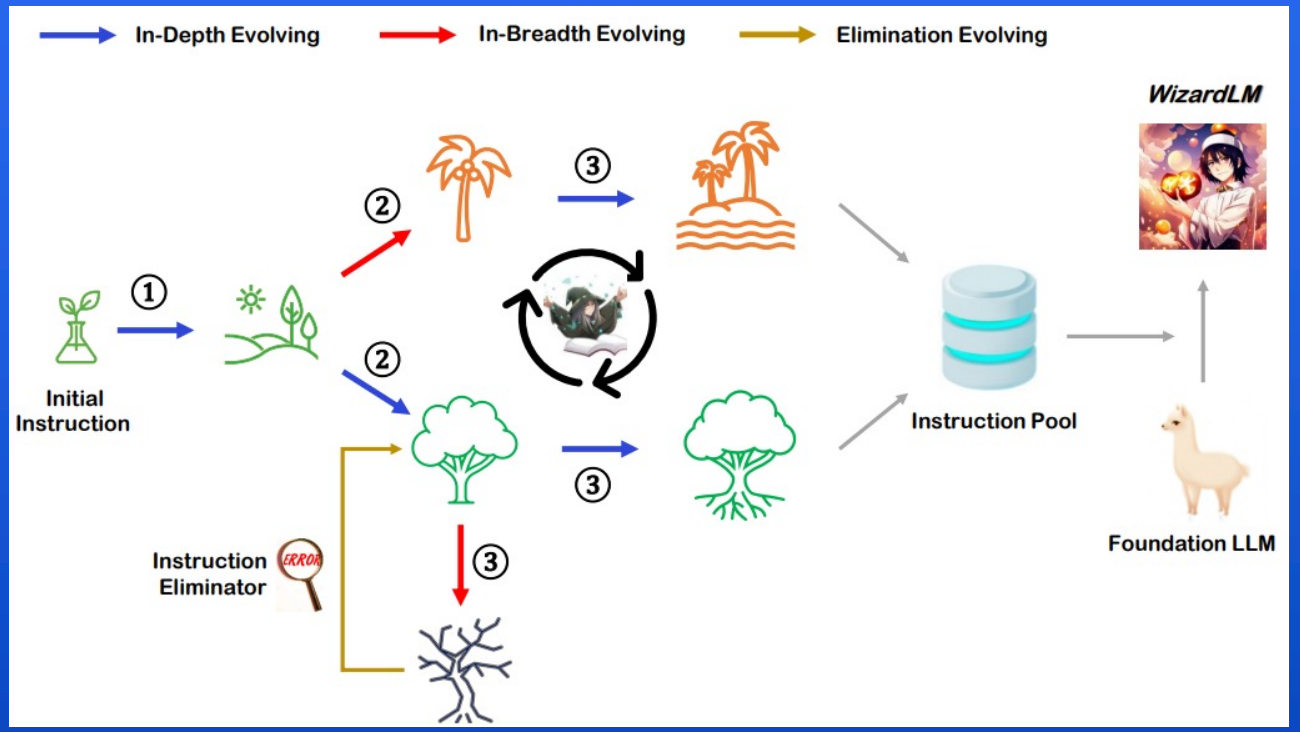
- 原始文档质量
- 文档切片算法：既要保留语义，又要兼顾切片长度
- Embedding模型
- Prompt
- LLM能力



以 Instruction Compliance 为终点：

Instruction Compliance





WizardLM: Empowering Large Language Models to Follow Complex Instructions

THANKS

THANKS

THANKS

THANKS