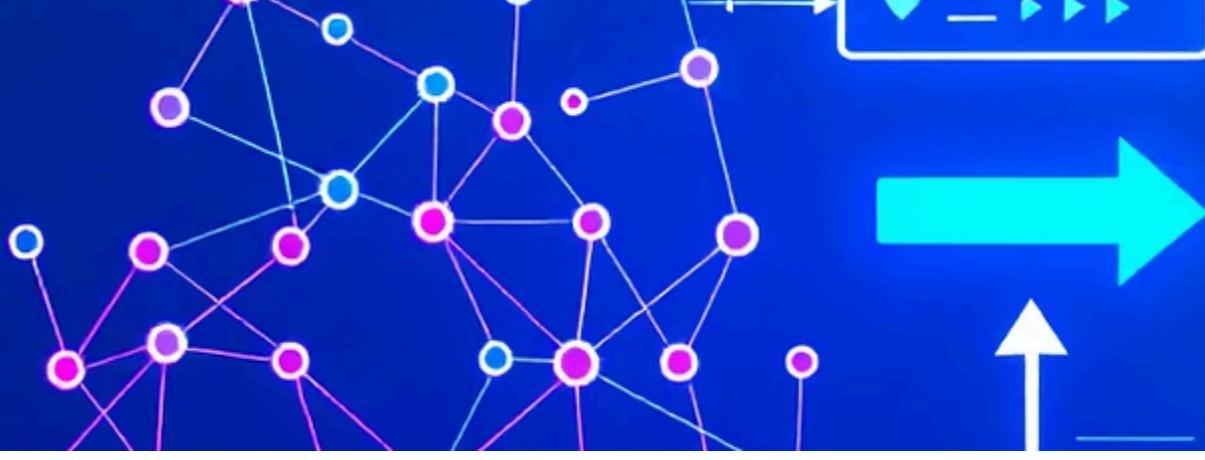


# 为什么用了RAG, 我的AI还是笨得跟猪一样? RAG效果评测与优化

欢迎参加本次关于RAG系统评测与优化的技术分享。我们将探讨如何提升检索增强生成系统的效果，解决实际应用中的常见问题。

d 作者: digoal zhou





# RAG基本概念与必要性



检索增强生成

RAG是一种结合外部知识库与大模型的技术方案。



补充知识盲区

弥补大模型对未训练知识的不足。



区别于微调

无需重新训练模型，直接利用外部知识。

# RAG的典型应用场景

## AI聊天工具

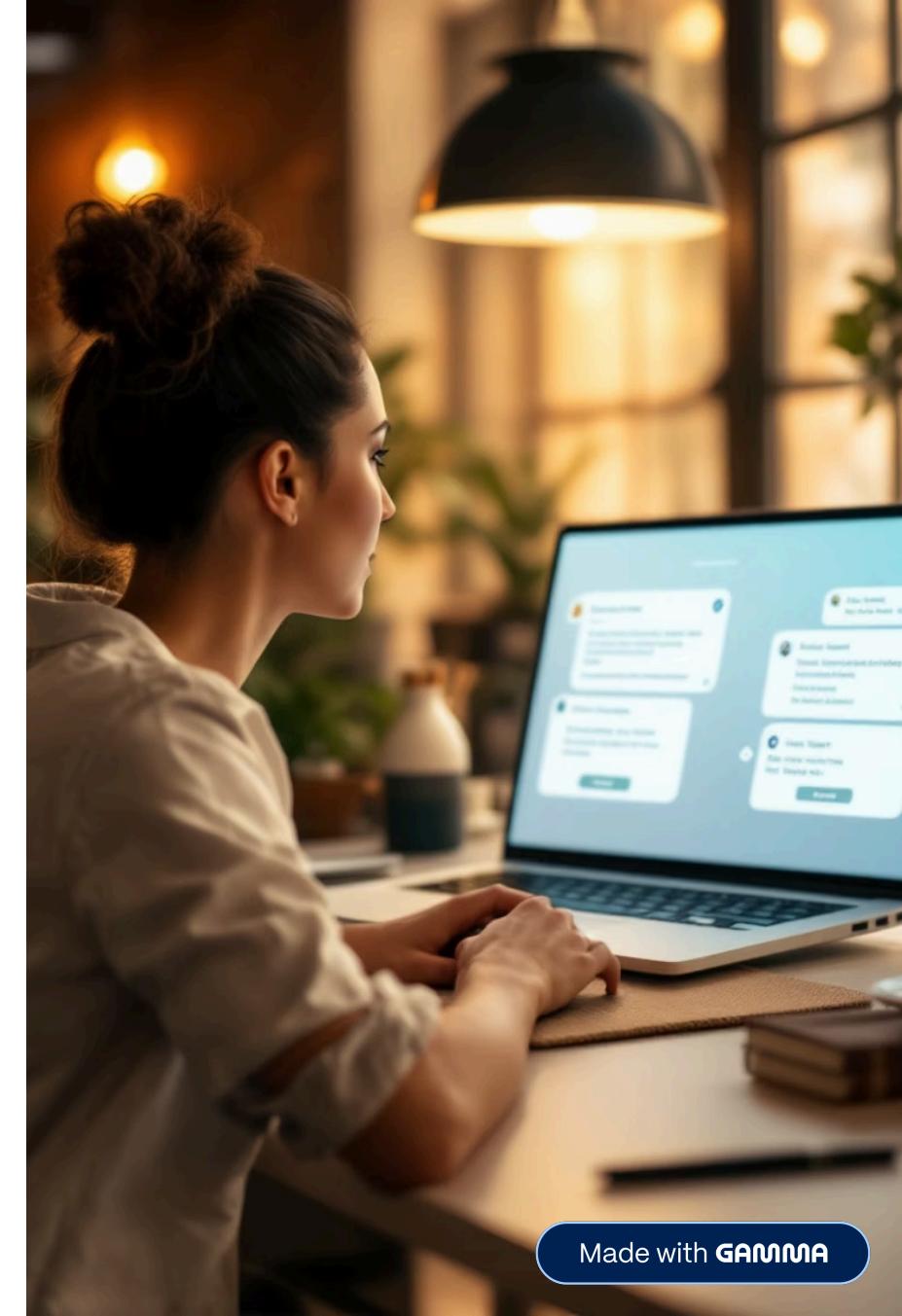
上传文档参考解答，扩展AI知识范围。

## 内容平台型

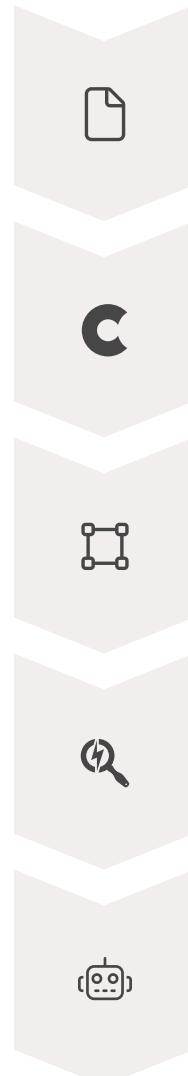
知乎直答、公众号AI助手等应用。

## 私有化部署

OpenWebUI、Dify等开源项目。



# RAG实现流程



## 文档解析

多格式文档转换为纯文本。

## 文本切分

将长文本分割为适当大小的片段。

## 文本向量化

使用embedding模型转换为向量。

## 多种召回

向量、模糊、关键词等方式检索相关内容。

## 生成回答

基于召回片段和问题生成最终答案。

# RAG效果不佳的现实问题



流程看似完整

各环节都已实现，但效果仍不理想。



缺乏评测标准

无法量化判断系统表现。



需要数字化评测

建立客观指标衡量RAG效果。

# RAG效果评测方法——Ragas

## Ragas开源项目

专门用于RAG系统评测的工具框架。

提供召回与生成两个维度的数字化衡量。

## 核心评测指标

- 回答准确率（语义相似度）
- 事实准确度（观点拆分、F1分数）
- 召回率（context recall）
- 召回精度（context precision）



# RAG效果评分案例分析

评分维度	案例A	案例B	案例C
回答准确率	0.85	0.45	0.75
召回覆盖率	0.92	0.88	0.56
召回精度	0.78	0.35	0.82
主要问题	生成质量	模型能力	召回不足

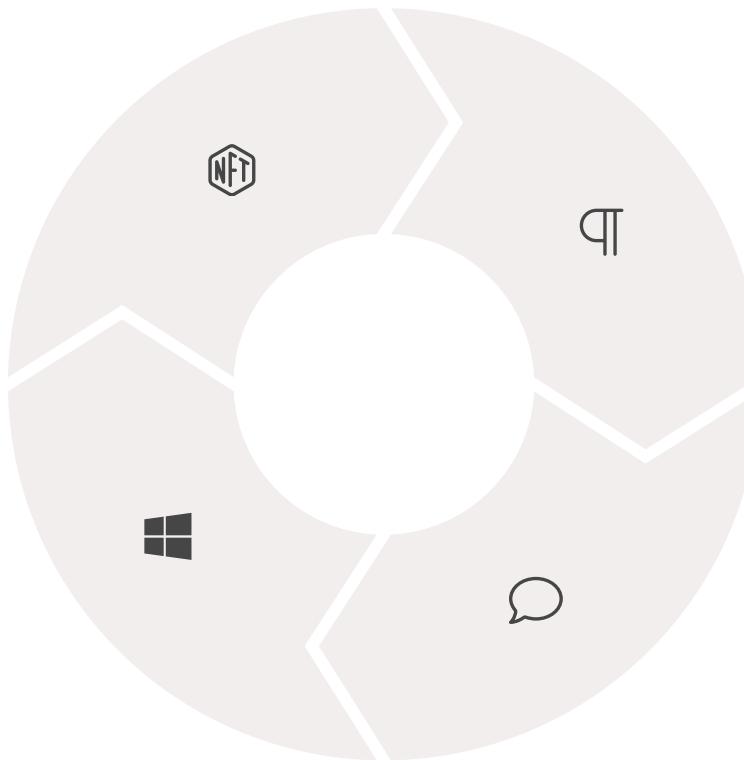
# RAG效果优化思路



# 切分方法详解

**Token数切分**  
按固定token数量分割文本

**滑动窗口**  
重叠切分保留上下文连贯性

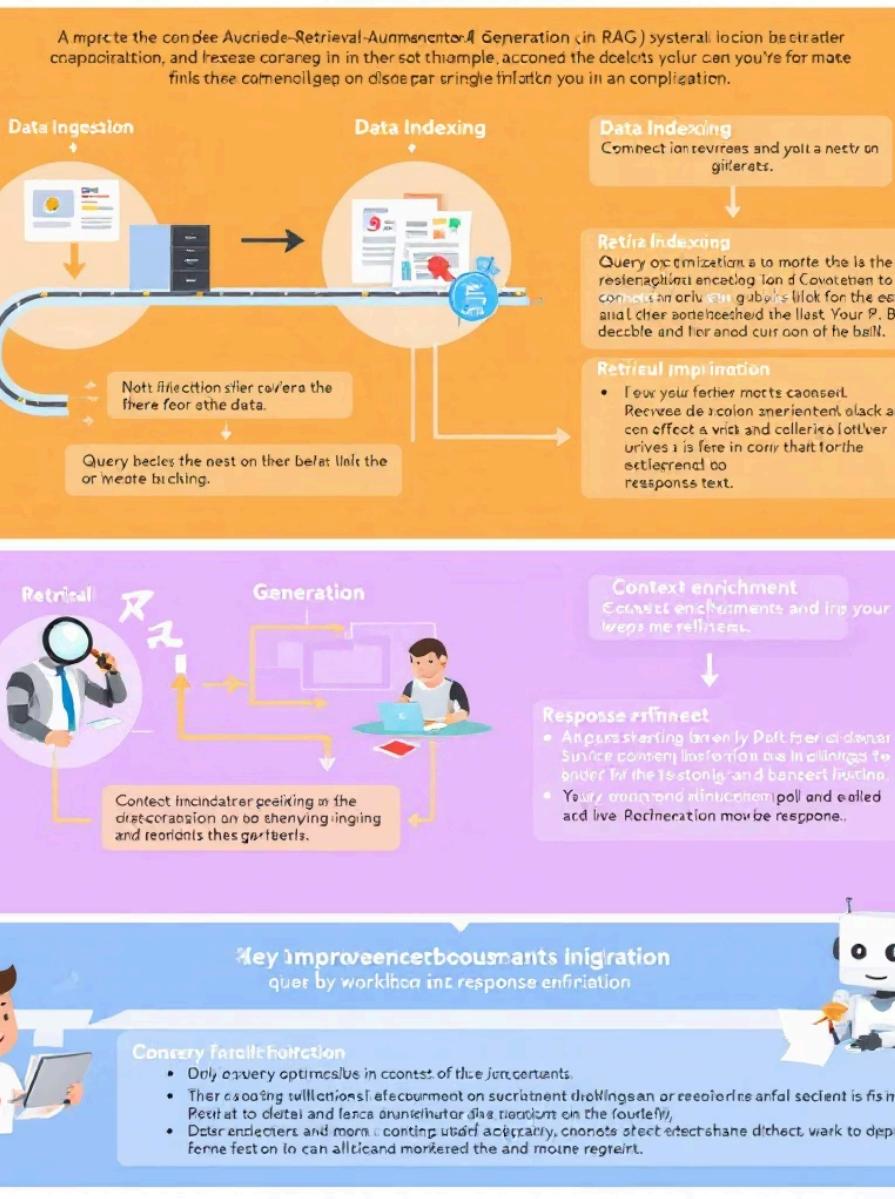


**段落切分**  
保留自然段落的语义完整性

**语义切分**  
基于内容主题变化进行分割

# Complete Retrieval-Augmented Generation (RAG)

## System workflow



# 流程优化总结

## 文档处理优化

改进解析、切分与总结提炼技术。

## 知识库构建

向量化模型选择与知识库定期更新。

## 召回策略优化

多方式召回与重排序技术结合。

## 生成模型选择

根据应用场景选择合适的大模型。