








# 工作流实现-Dify召回精度提高





## 摘要

本期视频详细介绍了一种提升知识库召回精度与效率的实用方法，基于作者对RG Flow与DeFi两大知识库平台的深入测试和理解，强调了知识库平台本身只是工具，关键在于对文档的预处理和分段策略。视频首先比较了RG Flow和DeFi在文档切分和召回机制上的不同侧重点，指出RG Flow擅长多样化的文章切片和格式识别，而DeFi更注重召回的稳定性和精度提升。作者提出了基于“父子分段”的文本处理工作流，通过人工标识段落、合理切块、利用大模型进行语义分析，显著提升知识库的召回质量。针对复杂格式的企业内部文档（如Excel、PPT等），作者分享了将内容转为纯文本（txt）格式、图片URL代理和批量转换的实操技巧。最终，经过细致语义分段和格式清理后，文档被高效植入知识库，实现了高召回率和准确度。视频强调，提升知识库效果的核心在于合理的文档预处理和语义分段，而非平台本身，推荐用户结合自身实际场景，采用类似的工作流进行知识库构建。

## 亮点

-  知识库平台只是工具，关键是文档的预处理和分段策略
-  RG Flow擅长文章格式识别，DeFi注重召回精度和稳定性
-  父子分段工作流有效提升文本切块的语义连贯性和召回质量
-  复杂格式文档处理技巧：转txt格式，图片URL代理与批量转换
-  利用大模型语义分析辅助段落划分，避免机械式切分的精度不足
-  人工粗打标记辅助分段，保证关键内容完整且不被拆分
-  实操案例展示，适用于企业内部多样化文档的知识库构建

## 关键洞察

-  工具定位与核心价值：知识库平台如RG Flow和DeFi本质是向量存储和检索入口，真正的召回效果依赖于数据预处理和嵌入模型，用户应重点关注如何优化数据切分和语义标注，而非盲目在平台间对比。
-  分段策略的重要性：机械化的基于长度和换行符的切分方法限制了召回精度，通过父子分段结合语义标记，可以实现更自然、更准确的内容划分，极大提升召回的相关性和覆盖度。
-  图文混排文档处理难点：企业内部文档格式多样且复杂，图片与文字信息紧密结合，需通过截图、URL代理等方式实现内容的完整表达，保证知识库中信息的多维度呈现。
-  大模型辅助文本处理：利用大模型处理切分后的文本块，进行语义分析和段落标识，弥补了传统机械切分在理解内容逻辑上的欠缺，增强了知识库的智能化水平。

⌚ 切块大小与模型限制：合理设置文本切块大小（如4000字节）符合大模型的token限制，避免一次性输入过多信息导致处理失败，同时通过人工标记避免关键内容被切断。

🔄 自动化与人工结合：纯自动切分虽便捷，但人工粗标记可显著提升分块质量，体现了自动化与人工经验结合的必要性，提升了知识库构建的精准度。

🎯 面向实际应用场景的优化：方法不仅适用于标准格式文本，对企业内部非结构化、离散内容同样有效，提升了知识库在真实业务环境中的应用价值和适应性。

总结来看，视频内容系统且实用，强调知识库建设的本质在于对原始文档的深度理解和合理分段，辅以智能工具和人工标记，帮助用户打造高效、精准的知识库检索体系。