Dify知识库切分

Dify知识库核心配置项信息整理

一、分段设置

1. **通用模式**

- **技术特点**：按自定义规则将文本拆分为独立分段，默认使用`\n`作为分隔符，支持正则表达式自定义

- **参数配置**：

- 最大长度：默认500 tokens（最大4000 tokens）
- 重叠长度：建议设置为总长度的10%-25%（默认50 tokens）
- 预处理规则：支持去除多余空格、URL和电子邮件

- **适用场景**：结构清晰的文档（如公告、FAQ）、段落独立性强的内容

2. **父子模式**

- **技术特点**：双层分段结构，子块用于精确检索，父块提供上下文

- **参数配置**：

- 父分段：默认段落级（`\n\n`分隔），最大500 tokens
- 子分段：默认句子级（`\n`分隔），最大200 tokens
- 支持全文作为父分段（限10000 tokens以内）

- **适用场景**：复杂文档（如合同、技术手册）、需要上下文关联的检索场景

二、索引方式

1. **高质量索引**

- **技术特点**：基于Embedding模型生成向量索引，支持语义检索

- **检索选项**：

- 向量检索：基于语义相似度匹配
- 全文检索：基于关键词匹配
- 混合检索：结合两者并支持Rerank重排序

- **优势**：检索精度高，支持多语言和复杂语义理解

- **成本**：消耗Embedding模型token，需配置模型API

2. **经济索引**

- **技术特点**：基于关键词倒排索引，每个分段提取10个关键词

- **优势**：零token消耗，索引速度快

- **局限**：仅支持关键词匹配，语义理解能力弱

- **适用场景**：预算有限、简单FAQ、关键词明确的检索场景

三、Embedding模型

1. **主流模型对比**

模型	特点	适用场景
Jina Embeddings v2	支持8k上下文，多语言能力强	长文档处理、多语言知识库
Qwen3-Embedding	32k超长上下文，检索精度高	法律文档、技术手册
BGE-M3	平衡性能与效率，支持多粒度检索	通用场景、企业知识库
text-embedding-3-large	OpenAI出品，语义捕捉能力强	英文为主的国际业务

2. **配置建议**

- 中文场景：优先选择Qwen3-8B或BGE-M3
- 资源受限：选择Qwen3-0.6B轻量化模型
- 多语言需求：Jina Embeddings v2或multilingual-e5-large

四、检索设置

1. **核心参数**

- **Top K**：默认3，建议根据模型上下文窗口调整（3-10）

- **Score阈值**：默认0.5，高阈值（>0.7）提升精度，低阈值（<0.5）增加召回率

- **Rerank模型**：推荐Cohere Rerank或bge-reranker，提升排序效果

2. **检索策略**

- **向量检索**：适合语义相似性查询，如"如何申请退款？"

- **全文检索**：适合关键词精确匹配，如"退款政策第3条"

- **混合检索**：权重设置建议语义占70%+关键词30%，需配置Rerank模型

3. **最佳实践**

- 客服场景：启用混合检索+Rerank，Top K=5，Score=0.6
- 技术文档：向量检索为主，Top K=3，Score=0.7
- 多数据集：启用多路召回模式，配置跨库Rerank

高级配置与行业最佳实践

一、自定义分段规则与优化策略

1. **正则表达式分段**

- **技术实现**：通过自定义分隔符（如`[\n。！？]`）实现句子级精准分割

- **代码示例**：

```
代码块
{
  "segmentation": {
    "delimiter": "[\\n。！？]",
    "max_tokens": 300,
    "chunk_overlap": 30
  }
}
```

- **适用场景**：法律文书（条款拆分）、技术手册（步骤分解）

2. **性能优化参数**

- **重叠率设置**：技术文档推荐20%重叠（如500tokens分段保留100tokens重叠）

- **长文档处理**：超过10000tokens文档建议启用"全文父分段+句子子分段"组合

二、Embedding模型深度对比

模型	维度	多语言支持	长文本处理	行业场景
BAAI/bge-m3	1024	100+语言	8k tokens	企业知识库
Jina Embeddings v2	768	200+语言	4k tokens	跨境电商
Qwen3-Embedding	1536	中英日韩	32k tokens	法律/医疗
nomic-embed-text	768	多语言	8k tokens	开源项目

三、检索策略与Rerank模型配置

1. 混合检索权重调优

- **语义权重**：技术术语查询建议70%向量+30%关键词

- **配置示例**：

代码块

```
retrieval_strategy:
  vector_weight: 0.7
  keyword_weight: 0.3
  rerank_model: bge-reranker-v2-m3
```

2. Rerank模型对比

模型	响应速度	准确率	适用场景
BGE-Reranker	300ms	89%	中文场景
Cohere Rerank	150ms	92%	多语言
Jina Reranker	200ms	87%	代码检索

四、行业最佳实践案例

1. **金融风控知识库**

- **分段策略**：父子模式（父段500tokens+子段150tokens）

- **索引配置**：混合检索+Qwen3-Embedding+TopK=5

- **性能指标**：检索延迟<800ms，准确率>95%

2. **医疗文档处理**

- **元数据过滤**：`department:cardiology AND publish_time>2024`

- **模型组合**：bge-m3嵌入 + bge-reranker重排序

- **合规要求**：启用数据加密存储（AES-256）

五、常见问题解决方案

问题场景	解决方案	配置示例
检索结果冗余	启用元数据过滤	`security_level:high`
多语言混淆	按语言拆分知识库	`lang:ja` 专用知识库
响应延迟高	启用Redis缓存	TTL=3600秒
关键词漏检	混合检索+BM25算法	`keyword_boost:["产品型号"]`