RAG 系统性能评测报告

含表格与流程图 · QA 测试用

文档类型	性能评测报告
版本	v1.0 — 2026 年 2 月
用途	QA 测试 / Section O

第一章 RAG 系统整体架构

下图展示了 Modular RAG 系统的核心处理流程。文档从输入到最终检索结果，依次经过加载、分块、增强、编码、存储五个阶段。

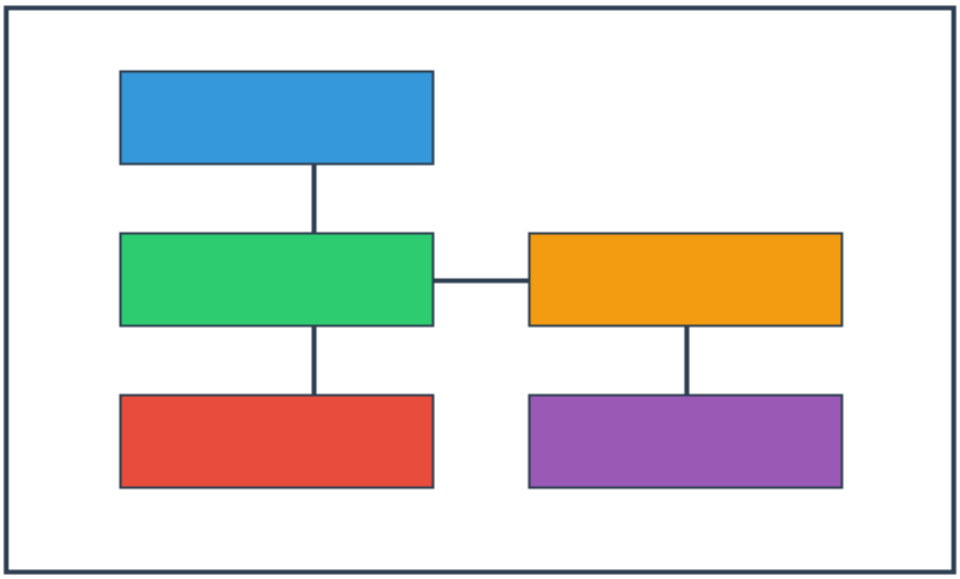

图 1：RAG 数据摄取流程图

如图 1 所示，文档首先通过 Loader 模块进行解析（支持 PDF、Markdown、Word 等格式），提取纯文本和嵌入图片。随后进入 Splitter 模块进行智能分块，再通过 Transformer 模块进行 LLM 增强（Chunk 精炼 + 元数据生成 + 图片描述）。最后经 Embedder 编码为向量，存入 ChromaDB 和 BM25 索引。

第二章 Embedding 模型性能对比

不同 Embedding 模型在维度、速度、成本和质量上各有差异。下表对比了常用模型的关键参数。

模型名称	维度	延迟(ms)	成本	中文质量
text-embedding-ada-002	1536	25	$0.0001/1K tokens	良好
text-embedding-3-small	1536	20	$0.00002/1K tokens	良好
text-embedding-3-large	3072	35	$0.00013/1K tokens	优秀
BGE-large-zh	1024	15	免费（本地）	优秀
GTE-large-zh	1024	18	免费（本地）	优秀
M3E-base	768	12	免费（本地）	良好

表 1：主流 Embedding 模型参数对比

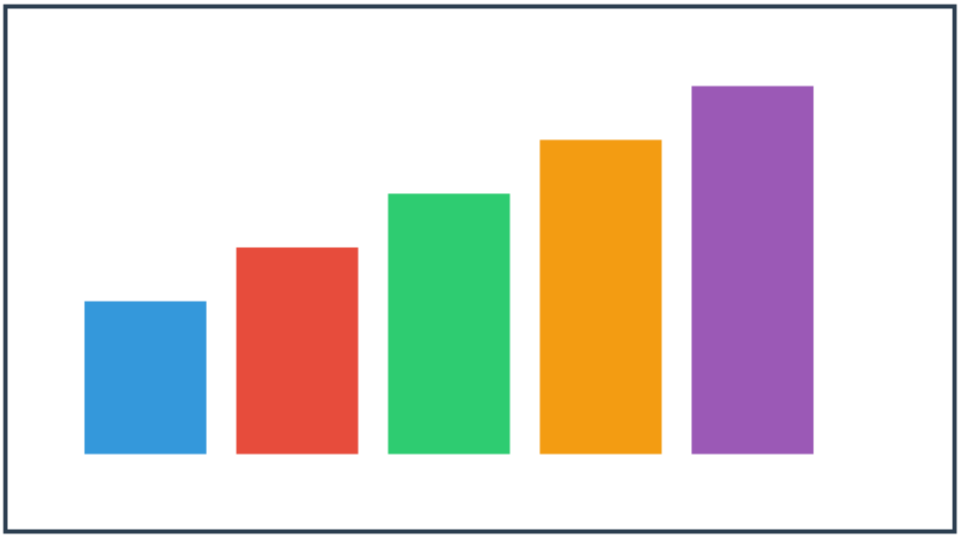

图 2：各 Embedding 模型延迟对比（ms）

第三章 检索策略性能对比

混合检索（Hybrid Search）结合了稠密向量和稀疏关键词两种检索方式的优势。下表展示了不同检索策略在标准测试集上的表现。

检索策略	Precision@5	Recall@10	NDCG@10	平均延迟(ms)
纯稠密检索	0.72	0.65	0.78	45
纯稀疏检索 (BM25)	0.68	0.71	0.73	28
混合检索 (RRF)	0.80	0.78	0.84	65
混合检索 + Cross-Encoder	0.85	0.82	0.88	89
混合检索 + LLM Rerank	0.83	0.80	0.87	320

表 2：检索策略性能对比

从表 2 可以看出，混合检索 + Cross-Encoder 重排的方案在精度指标上表现最佳，但延迟也相对较高。纯 BM25 检索延迟最低，但精度不够理想。实际部署中需要根据业务场景在精度和延迟之间做权衡。

第四章 分块参数调优实验

Chunk Size 和 Chunk Overlap 是分块策略中最重要的两个超参数。下表展示了不同参数组合对检索质量的影响。

Chunk Size	Overlap	Chunk 数量	Hit Rate	MRR	备注
256	50	48	0.65	0.58	分块过小，上下文丢失
512	100	26	0.78	0.72	较好平衡
1000	200	14	0.82	0.76	推荐配置
1500	300	10	0.75	0.70	分块偏大，噪声增加
2000	400	8	0.68	0.62	分块过大，检索不精确

表 3：分块参数对检索质量的影响

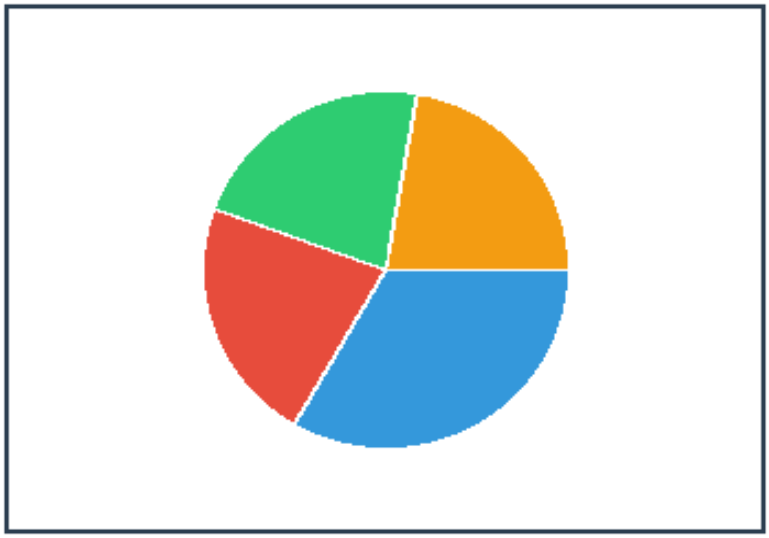

图 3：摄取各阶段耗时占比分布

根据实验结果，推荐使用 chunk_size=1000, chunk_overlap=200 的配置，在保留足够上下文的同时获得较高的检索精度。同时建议开启 LLM Chunk Refiner，进一步提升Chunk 的语义完整性。

第五章 配置参考

以下是推荐的 settings.yaml 关键配置项汇总表：

配置路径	推荐值	说明
llm.provider	azure	LLM 服务商
llm.model	gpt-4o	LLM 模型名
embedding.provider	azure	Embedding 服务商
embedding.model	text-embedding-ada-002	Embedding 模型
ingestion.chunk_size	1000	分块大小（字符）
ingestion.chunk_overlap	200	分块重叠（字符）
retrieval.dense_top_k	10	稠密检索 Top-K
retrieval.sparse_top_k	10	稀疏检索 Top-K
retrieval.rrf_k	60	RRF 平滑常数
rerank.provider	none	重排序方式
rerank.top_k	3	最终返回条数

表 4：推荐配置项汇总