Advances in Retrieval-Augmented Generation

A Comprehensive Study

Abstract

Retrieval-Augmented Generation (RAG) has emerged as a powerful paradigm for enhancing large language models with external knowledge. This paper examines recent advances in RAG systems, focusing on embedding models, chunking strategies, and reranking techniques. We present experimental results showing that multi-stage retrieval with cross-encoder reranking improves accuracy by 15-20% compared to single-stage vector search. Our findings demonstrate that unlimited chunking with 25% overlap provides optimal context preservation while maintaining reasonable computational costs.

1. Introduction

Large Language Models (LLMs) have revolutionized natural language processing, but they suffer from knowledge cutoff dates and inability to access proprietary or recent information. RAG addresses these limitations by augmenting LLM prompts with relevant documents retrieved from external knowledge bases. The effectiveness of RAG systems depends critically on three components: the embedding model used for vectorization, the chunking strategy for document segmentation, and the retrieval mechanism for finding relevant content.

2. Methodology

2.1 Embedding Models

We evaluated three embedding approaches: OpenAI's text-embedding-3-small (1536 dimensions), all-MiniLM-L6-v2 (384 dimensions), and BAAI BGE-M3 (1024 dimensions). Each model was tested on a corpus of 2000 technical documents with known question-answer pairs. Performance metrics included Mean Reciprocal Rank (MRR), Precision@10, and Recall@10.

2.2 Chunking Strategies

Document chunking was performed with three strategies: fixed-size (800 characters), semantic paragraph-based, and hybrid paragraph-sentence splitting. We tested chunk limits of 4, 50, and unlimited (with 500 safety limit). Overlap ratios of 0%, 25%, and 50% were compared for context preservation.

2.3 Reranking

Two-stage retrieval was implemented using fast vector search (k=50) followed by cross-encoder reranking (ms-marco-MiniLM-L-6-v2) to select top-10 results. We compared this against single-stage retrieval and reciprocal rank fusion (RRF) with multiple query variants.

3. Results

Experimental results show that BGE-M3 embeddings achieved MRR of 0.863, significantly outperforming all-MiniLM-L6-v2 (MRR: 0.710) while OpenAI embeddings reached 0.945. Unlimited chunking with 25% overlap improved recall by 23% compared to 4-chunk limit. Reranking provided consistent 15-18% accuracy improvements across all embedding models. The combination of BGE-M3, unlimited chunking, and reranking achieved 89.2% accuracy on our test set, approaching proprietary API performance while remaining fully local.

4. Conclusion

This study demonstrates that high-quality RAG systems can be built using open-source components. The key findings are: (1) Unlimited chunking with overlap prevents information loss, (2) Reranking significantly improves relevance, (3) BGE-M3 provides excellent quality for multilingual scenarios. Future work should explore dynamic chunking based on document structure and hybrid retrieval combining dense and sparse representations.

References

• Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS.

• Karpukhin, V., et al. (2020). Dense Passage Retrieval for Open-Domain Question Answering. EMNLP.

• Xiao, S., et al. (2023). C-Pack: Packaged Resources for General Chinese Embeddings. arXiv.