RAG System Evaluation: Findings & Recommendations

Key Findings

1. Performance Metrics Overview

- F1 Score: 0.621 (62.1%) across all configurations
- Precision: 0.45 (45%)
- Recall: 1.0 (100%)
- Accuracy: 0.45 (45%)

2. Chunking Method Comparison

Method	Avg F1 Score	Avg Latency (ms)
Fixed Size	0.621	59.29
Recursive Character	0.621	39.24

3. Similarity Algorithm Comparison

Algorithm	Avg F1 Score	Avg Latency (ms)
Cosine	0.621	44.81
Dot Product	0.621	53.72

4. Performance Analysis

- Perfect Recall (1.0): The system successfully retrieves all relevant documents for every query.
- Moderate Precision (0.45): For every relevant document returned, the system also returns approximately 1.2 irrelevant documents.
- Identical Effectiveness Metrics: All configurations show the same precision, recall, and F1 score, suggesting the performance differences are primarily in latency rather than retrieval quality.
- Significant Latency Differences: Recursive character chunking is approximately 33% faster than fixed-size chunking.

Recommendations

1. Optimize for Latency

Finding: Recursive character chunking consistently outperforms fixed-size chunking in terms of speed while maintaining identical effectiveness metrics.
Recommendation:

- Adopt recursive character chunking as the default chunking strategy
- If using cosine similarity, the recursive_character + cosine combination provides the best latency (33.60ms)
- Consider the combination of recursive_character + cosine as the production configuration

2. Address Precision Issues

Finding: While recall is perfect (1.0), precision is only moderate (0.45), indicating the system retrieves all relevant documents but also many irrelevant ones.
Recommendations:

- Implement a relevance threshold to filter out lower-scoring matches
- Consider a two-stage retrieval process where:
    1. Initial broad retrieval ensures high recall
    2. Re-ranking step improves precision by more carefully analyzing the top results
- Experiment with different chunk sizes to find an optimal balance between context preservation and specificity

3. Further Testing Recommendations

Finding: The identical effectiveness metrics across different configurations suggest current tests may not fully differentiate between methods.
Recommendations:

- Test with different chunk size parameters:
    - For fixed-size: Try 256, 512, 1024 characters
    - For recursive: Adjust min/max chunk sizes
- Experiment with different overlap settings (10%, 20%, 50%)
- Implement a more nuanced relevance scoring system (e.g., 0-5 scale instead of binary relevance)
- Add more complex, ambiguous queries that might better differentiate between methods

4. Production Implementation

Recommendations:

- Use recursive character chunking with cosine similarity as the default configuration
- Consider a hybrid approach where:
    - Short, specific documents use fixed-size chunking
    - Longer, structured documents use recursive chunking

- Implement monitoring to track real-world precision and recall metrics
- Set up A/B testing to continue comparing configurations with real user queries

5. Advanced Features to Consider

- Hybrid Search: Combine vector search with keyword-based filtering
- Re-ranking: Implement a secondary ranking phase that uses a more computationally expensive but accurate model
- Chunk Metadata: Store metadata about chunk position and relationships to enable context-aware retrieval
- Query Expansion: Implement techniques to expand or rewrite queries to improve retrieval performance

Conclusion

The RAG system demonstrates perfect recall performance but shows room for improvement in precision. The evaluation clearly indicates that recursive character chunking offers better efficiency without compromising effectiveness. By implementing the recommendations above, particularly adopting recursive character chunking and focusing on precision improvements, the system can achieve enhanced speed and accuracy in information retrieval operations.