

# Hybrid Search Strategy

Current Problem:

Semantic search misses keywords

Keyword search doesn't understand intent

Idea:

1. Vector similarity (semantic)
2. BM25 keyword matching
3. Re-rank combined results using LLM

Benefits:

- Better accuracy
- Catches edge cases
- more relevant results

Challenge: Optimize latency

Solution: Cache popular queries

Critical for the RAG pipeline architecture