SEC Filings QA Agent: Technical Report

Executive Summary

This technical report documents a production-grade question-answering system designed for complex financial research across SEC filings. The system addresses key challenges in processing heterogeneous financial documents through multi-dimensional query routing, semantic retrieval, and section-aware chunking. Core innovations include a concept mapping system for financial terminology, quantized embeddings reducing storage by 4x, and intelligent routing that extracts ticker, temporal, and document context from natural language queries.

Approach

The system implements a domain-specific design through six core layers optimized for financial document analysis:

Document Processing Layer: Form-specific regex parsers handle 10-K, 10-Q, 8-K, DEF14A , 3, 4, 5 structures with section-aware chunking. XBRL parsing extracts fiscal periods and metadata through the edgartools API. The FormChunk model preserves complete filing context including CIK, ticker, fiscal year/quarter, section identifiers, and document URLs.

Multi-Dimensional Query Routing: The routing system solves complex financial query understanding through specialized extractors for ticker identification, temporal context parsing (quarters, fiscal years, trends), and document type inference. Financial concepts map to specific SEC sections with confidence scoring.

Semantic Retrieval System: Dense vector search using intfloat/e5-small-v2 embeddings with quantized storage. The system was initially designed for hybrid retrieval combining dense and sparse (BM25) methods, but CockroachDB's lack of production-level vector search support forced simplification to semantic-only retrieval.

Storage Architecture: CockroachDB provides ACID transactions with async connection pooling. Vector embeddings use int8 quantization with zlib compression achieving 4x storage reduction while preserving semantic accuracy.

Challenges Addressed

Heterogeneous Document Structures: SEC forms vary significantly in structure (10-K vs 8-K vs proxy statements). The system addresses this through form-specific parsers preserving regulatory structure while enabling cross-document queries.

Database Vector Search Limitations: Initial investigation into CockroachDB's vector capabilities revealed that while the database advertises vector support, production-level vector search functionality was not available. This limitation required several days of development time to identify and forced a significant architectural change. The planned hybrid retrieval system combining dense vector search with sparse BM25 retrieval had to be abandoned in favor of a simplified semantic-only approach, reducing the system's ability to handle queries requiring precise lexical matching.

Complex Financial Terminology: Domain-specific language and nested relationships in financial documents required specialized concept mapping beyond general-purpose embeddings.

Temporal and Company Context: Financial queries often require understanding of time periods, company comparisons, and specific filing types - addressed through multi-dimensional routing intelligence.

Scale and Performance: Processing large volumes of SEC filings while maintaining response times required quantized embeddings, intelligent caching, and async processing architectures.

Capabilities and Limitations

Capabilities:

- Comprehensive SEC form support (10-K, 10-Q, 8-K, DEF14A, Forms 3-5) with section-aware processing
- Multi-dimensional query routing understanding ticker, temporal, and document constraints simultaneously
- Semantic similarity search with quantized embeddings for efficient storage and retrieval
- 4x storage efficiency through quantized embeddings with minimal accuracy loss
- Production-ready architecture with async processing, connection pooling, and error handling

Limitations:

- Single-node vector storage limits horizontal scaling potential
- Semantic-only retrieval may miss queries requiring precise lexical matching due to abandonment of hybrid approach
- Limited to English-language documents and US SEC filing formats
- Reduced precision on financial terminology queries that would benefit from BM25 sparse retrieval

Performance

Memory Efficiency: Quantized embeddings reduce storage requirements by 4x with <2% accuracy degradation. Three-tier TTL caching system prevents memory leaks while optimizing response times.

Processing Speed: Async architecture with connection pooling handles concurrent queries effectively. Smart ingestion based on routing intelligence reduces unnecessary document processing by targeting relevant filings.

Accuracy Trade-off: The forced move from hybrid to semantic-only retrieval reduced accuracy on queries requiring exact term matching, particularly for specific SEC section references and financial terminology that would have benefited from sparse retrieval methods.

Trade-offs

Storage vs. Accuracy: Quantized embeddings provide significant storage savings at minimal accuracy cost, enabling production deployment within resource constraints.

Retrieval Complexity vs. Database Constraints: Originally designed hybrid retrieval system was simplified to semantic-only due to CockroachDB vector search limitations, trading query precision for implementation feasibility.

Complexity vs. Functionality: Multi-dimensional routing and semantic retrieval enable sophisticated query understanding, though less precise than the originally planned hybrid approach.

Technology Choices: CockroachDB selected for storage limits and ACID compliance, but vector search limitations forced architectural compromises. Gemini 2.5 Flash chosen for free tier availability (1000 requests/day). Sentence Transformers (384 dimensions) balanced speed and quality for embedding generation.

Conclusions

The SEC QA system demonstrates production-ready financial document analysis through domain-specific innovations, despite architectural limitations imposed by database constraints. Multi-dimensional routing enables sophisticated query understanding while semantic retrieval provides reasonable performance for most financial research queries.

Key Technical Achievements: 4x storage reduction through quantization, comprehensive SEC form support spanning multiple document types, and intelligent query routing with financial concept mappings enabling precise section targeting.

Production Considerations: The system's async architecture, connection pooling, TTL caching, and multi-provider LLM support enable deployment at scale. However, the database vector search limitations highlight the critical importance of thorough technology evaluation in system design, as they forced abandonment of hybrid retrieval and reduced overall system precision for certain query types.

Lessons Learned: The gap between advertised database features and production-ready implementations can significantly impact system architecture. Future iterations should prioritize vector database solutions with proven production vector search capabilities, such as supabase, to enable the originally planned and semi-implemented hybrid retrieval approach.