SEC Filings QA System - Technical Summary – Ya-Wei (Jeremy) Tsai

System Architecture

Core Pipeline：

The system follows a 5-stage pipeline optimized for financial research workflows:

1. Query Analysis: OpenAI-powered extraction of tickers, time periods, document types, and query classification
2. Company Resolution: Ticker validation and intelligent fallback suggestions using SEC company data
3. Document Retrieval: Multi-strategy retrieval combining targeted and general approaches
4. Financial Analysis: LLM synthesis with inline citation tracking
5. Citation Normalization: Automatic renumbering and formatting of source references

Key Components

EnhancedSECQASystem (qa_system.py): Main orchestrator handling the complete pipeline with sophisticated citation management. Implements citation parsing with regex extraction, compact renumbering, and metadata preservation.

EnhancedSECDataRetriever (data_retriever.py): Handles SEC API integration with three retrieval modes:

- Targeted retrieval for specific forms/sections (e.g., 10-K Item 1A for risk factors)
- General fallback across form types
- Low-traffic probe mode for edge cases

Query Analysis Agents (query_agent.py, analysis_agent.py): Specialized LLM agents using pydantic-ai for structured output and type safety. Query agent extracts structured metadata; analysis agent synthesizes findings with citations.

CompanyResolver (company_resolver.py): Resolves tickers/company names using SEC's company_tickers.json with local caching and fuzzy name matching.

Technical Approach

Query Processing Strategy:

The system implements intelligent query classification supporting six query types:

- Single ticker analysis
- Multi-ticker comparisons
- Temporal trend analysis
- Multi-dimensional queries (ticker + time + document type)

- Industry-wide analysis
- Thematic research

Query analysis extracts structured metadata including suggested tickers when none are specified, enabling proactive company recommendations based on industry context.

Document Retrieval Architecture

Three-Tier Retrieval Strategy:

1. Targeted Retrieval: Maps query intent to specific SEC forms and sections
    - Insider trading queries → Forms 3/4/5
    - Risk analysis → 10-K Item 1A
    - Compensation research → DEF 14A proxy statements
2. Structured Section Extraction: Uses SEC API's extractor service for precise section retrieval from 10-K/10-Q filings with Item-level granularity
3. General Fallback: Broad search with filing summaries when targeted approaches fail

Citation Management System

Implements a sophisticated citation tracking system addressing the challenge of maintaining source attribution across document chunks:

- Inline Citation Tracking: Uses [C#] tags throughout analysis
- Usage Detection: Regex parsing to identify actually used citations
- Compact Renumbering: Sequential renumbering of citations in order of appearance
- Metadata Preservation: Maps citations back to original DocumentChunk objects

Performance Optimizations

- Document Caching: Prevents redundant API calls with cache keys combining company, document types, and time periods
- Rate Limiting: 0.3-0.7 second delays between API calls to respect sec-api.io limits
- Chunk Limiting: Processes maximum 6-8 document chunks to balance thoroughness with latency
- Fallback Strategies: Graceful degradation when primary retrieval methods fail
- Direct Processing Architecture: Chose local data processing over LLM tool calls after discovering that tool-based approaches led to repetitive calling patterns, increased latency, and inefficient token consumption during financial document analysis

Challenges Addressed

1. Complex SEC Data Integration

Challenge: SEC filings exist in multiple formats (HTML, text) with inconsistent structures across different form types.

Solution: Implemented form-specific processing logic with specialized handlers for insider trading (Forms 3/4/5), structured filings (10-K/10-Q), and event-driven forms (8-K). Uses both search and extractor APIs for optimal data quality.

2. Multi-Dimensional Query Support

Challenge: Financial research requires simultaneous filtering by company, time period, and document type.

Solution: Developed structured query analysis that extracts and validates multiple dimensions, with intelligent defaults and suggestions when information is missing.

3. Source Attribution at Scale

Challenge: Maintaining accurate citations across multiple document chunks while keeping answers concise.

Solution: Built citation normalization system that tracks usage, renumbers sequentially, and provides detailed source mapping without overwhelming the user.

4. Missing Data Handling

Challenge: Not all companies have all filing types available for requested time periods.

Solution: Implemented graceful degradation with missing company tracking, alternative suggestion systems, and transparent limitation reporting.

Capabilities and Limitations

Capabilities

- Multi-Ticker Analysis: Simultaneous analysis of multiple companies with comparative insights
- Temporal Analysis: Trend identification across filing periods
- Precise Source Attribution: Every claim linked to specific SEC filings with URLs
- Intelligent Fallbacks: Automatic company suggestions and alternative retrieval strategies
- Robust Error Handling: Graceful failure modes with informative error messages

Current Limitations

- API Rate Limits: SEC API constraints limit concurrent processing speed
- Recent Filing Dependency: Analysis quality depends on recent filing availability
- Section Extraction Limits: Some complex filing structures may not extract perfectly
- Token Usage: Large document analysis can approach LLM context limits
- Numerical Precision: Financial calculations rely on LLM interpretation rather than structured data parsing

Performance Characteristics

- Average Query Latency: 15-30 seconds for multi-company analysis
- Token Efficiency: ~2,000-5,000 tokens per query through targeted retrieval
- Cache Hit Rate: ~60-80% for repeated company/timeframe combinations
- Success Rate: ~90% for well-formed queries with valid tickers

Trade-offs and Design Decisions

Retrieval Strategy Trade-offs

Targeted vs. Comprehensive: Chose targeted retrieval prioritizing precision over recall. This improves answer quality and reduces token usage but may miss relevant information in unexpected sections.

Caching vs. Real-time: Implemented document caching to improve performance, trading some real-time accuracy for significant speed improvements on repeated queries.

LLM Integration Trade-offs

Two-Agent Architecture: Separated query analysis from financial analysis to improve modularity and debugging capability, at the cost of additional API calls.

Structured Output: Used Pydantic models for type safety and validation, trading some flexibility for reliability and maintainability.

Local Processing Over Tool Calls: Deliberately chose local data processing and analysis over LLM tool integration. During development, tool-based approaches resulted in repetitive calling patterns and inefficient token usage when processing financial data. Local processing with structured prompts proved more reliable and cost-effective for SEC document analysis.

Citation Strategy Trade-offs

Inline vs. Endpoint Citations: Chose inline citations for better readability during analysis, with post-processing normalization adding complexity but improving user experience.

Citation Granularity: Balanced between section-level and sentence-level citations, settling on chunk-level attribution that provides sufficient detail without overwhelming users.