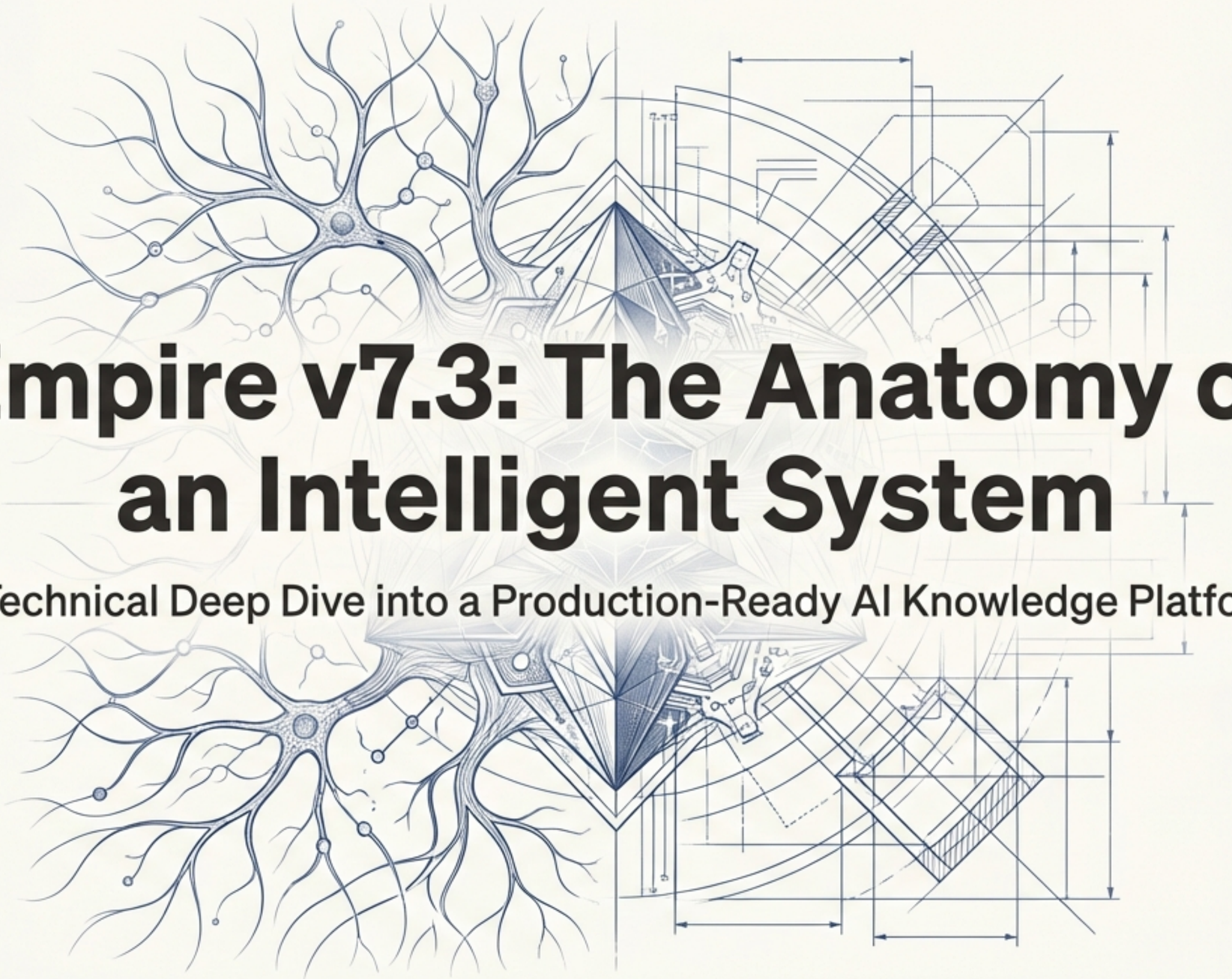


Empire v7.3: The Anatomy of an Intelligent System

A Technical Deep Dive into a Production-Ready AI Knowledge Platform

Version 7.3.0 | Status: Production Ready | Last Updated: 2025-11-30

A Complete, Production-Ready AI Knowledge Platform

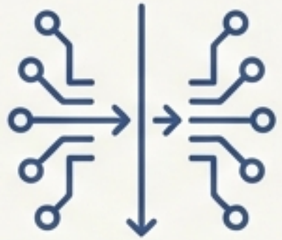
Empire v7.3 is a fully operational knowledge management system designed for enterprise-grade performance, security, and intelligence. This presentation deconstructs its architecture layer by layer.

15 Specialized AI Agents

Powered by a central AI core for tasks ranging from summarization to multi-agent orchestration.

Hybrid Database Architecture

Combining PostgreSQL (pgvector) for efficient vector search and Neo4j for complex relationship discovery in a knowledge graph.



26 REST API Endpoints

A comprehensive interface for document processing, intelligent chat, agent access, and system management.

Production Status: 46/46 Tasks Complete

The system has completed all development phases and is fully production-ready.

The System's Core Intelligence is Powered by Claude Sonnet 4.5

All primary AI reasoning—including summarization, classification, chat responses, and agent logic—is performerbrmed by Anthropic's Claude Sonnet 4.5 via its cloud API. This ensures a state-of-the-art, consistent, and scalable intelligence layer without local LLM inference.

Specialized AI Model Distribution

Purpose	Model	Provider	Type
All Primary AI Processing	**Claude Sonnet 4.5**	Anthropic	Cloud API
Query Expansion	Claude Haiku	Anthropic	Cloud API
Image Analysis	Claude Vision	Anthropic	Cloud API
Embeddings	BGE-M3	Local/Render	1024-dim vectors
Reranking	BGE-Reranker-v2	Local	Search optimization

A 15–Agent System for Specialized Cognitive Tasks

Content Processing Agents

AGENT-002: Content Summarizer
Generates PDF summaries with key points, themes, and actionable insights.

AGENT-008: Department Classifier

Classifies content into 10 predefined business departments.

Document Analysis Agents

AGENT-009: Senior Research Analyst
Extracts topics, entities, and facts.

AGENT-010: Content Strategist
Generates executive summaries and recommendations.

AGENT-011: Fact Checker
Verifies claims and assigns confidence scores with citations.

Multi-Agent Orchestration

AGENT-012: Research Agent
Performs web/academic search and assesses source credibility.

AGENT-013: Analysis Agent
Detects patterns and identifies correlations.

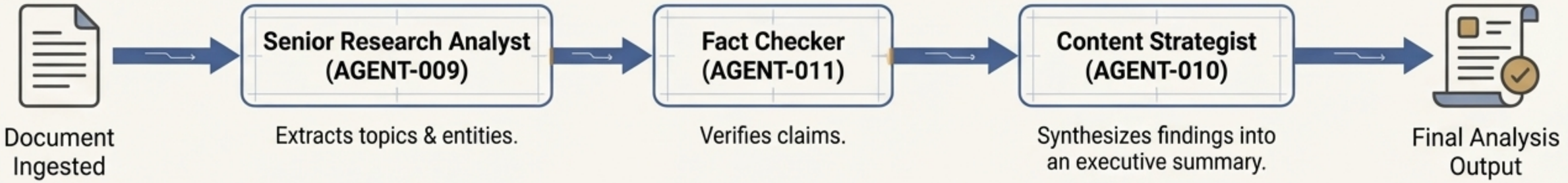
AGENT-014: Writing Agent
Generates reports in multiple formats with citations.

AGENT-015: Review Agent
Provides quality assurance and recommends revisions.

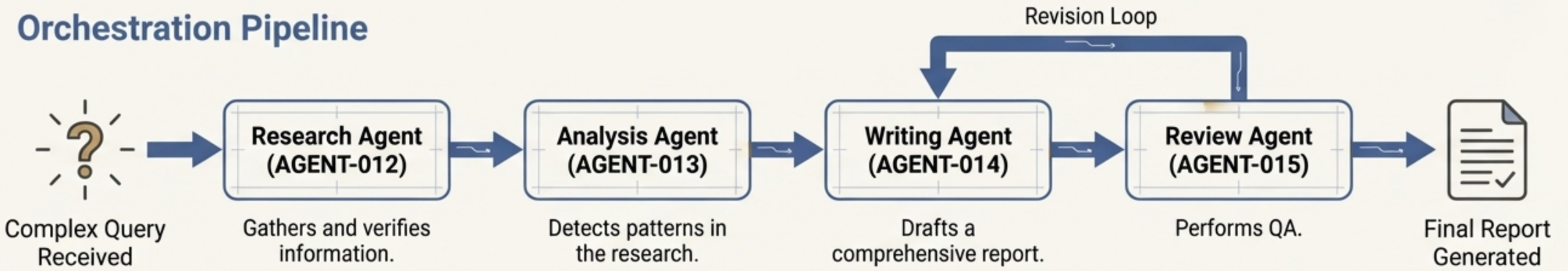
Multi-Agent Workflows Enable Complex Analysis

Individual agents are coordinated into intelligent pipelines to handle sophisticated, multi-step tasks with built-in quality control.

Document Analysis Pipeline



Orchestration Pipeline



A Hybrid Database Forms the System's Skeletal Structure and Memory

Empire utilizes a hybrid database approach to leverage the distinct advantages of relational, vector, and graph data models, ensuring both rapid semantic retrieval and deep relational understanding.

Database Architecture Components

Database	Provider	Purpose	Monthly Cost
PostgreSQL	Supabase	Vectors (pgvector), user data, sessions, audit logs	$25 |
Neo4j	Mac Studio Docker	Knowledge graphs, entity relationships, graph traversal	$0 (self-hosted)
Redis	Upstash	Caching, Celery task broker, rate limiting	Free Tier |

The Hybrid Approach: Purpose-Built for Comprehensive Knowledge

Why use both a relational/vector database and a graph database?

PostgreSQL with pgvector: The Power of Semantic Search and Structure

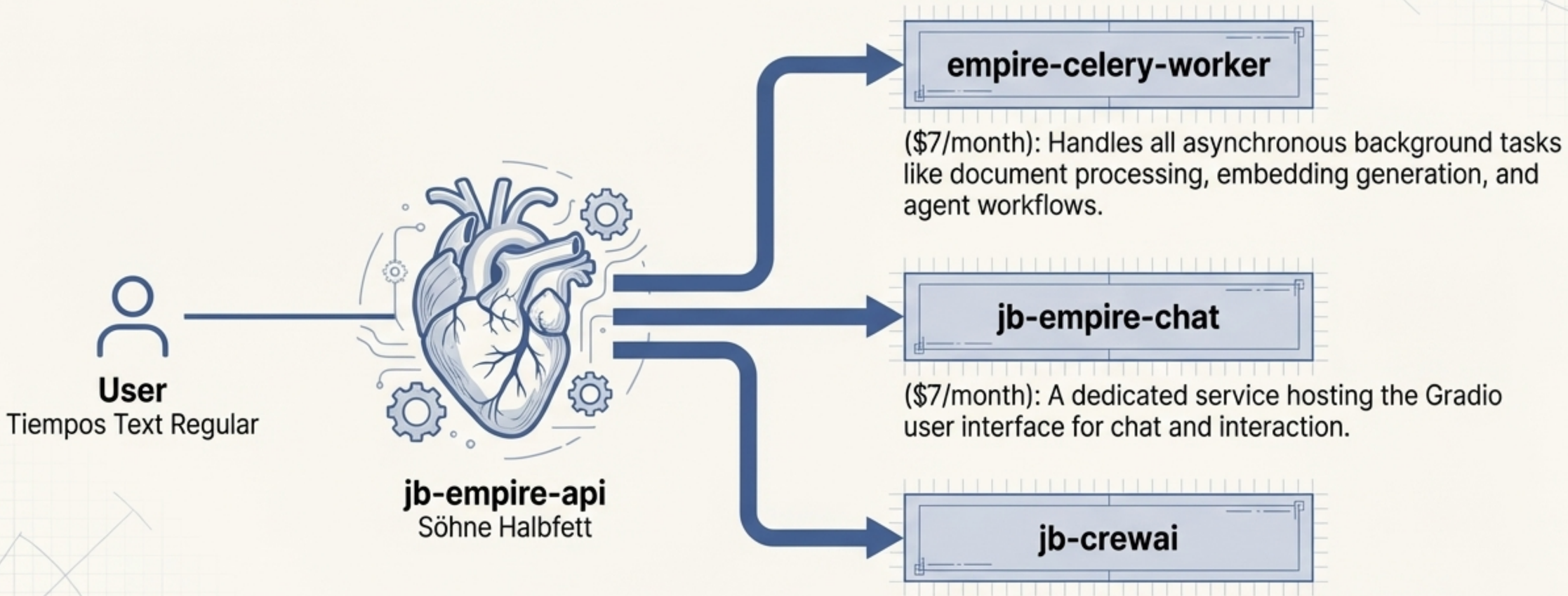
- Optimized for fast vector similarity search.
- Enforces data integrity with ACID transactions.
- Provides robust security with Row-Level Security (RLS).
- Ideal for storing document chunks, metadata, and user data.

Neo4j: The Power of Relationships and Discovery

- Superior performance for graph traversal and relationship queries.
- Uncovers hidden connections between entities (people, organizations, concepts).
- Enables graph-native algorithms for community detection and influence analysis.

Together, they create a system that can both find what you're looking for (semantic search) and discover what you didn't know to ask for (relationship discovery).

Cloud Services on Render.com Provide the System's Circulatory Flow



26 API Endpoints Offer Comprehensive System Control and Interaction

The system exposes a well-structured REST API, allowing for deep integration and control over its core functions.

Document Processing

`POST /api/v1/upload`

`GET /api/status/*`

Chat & Query

`POST /api/query/auto`
(Intelligent Routing)

`POST /api/chat/*`

`WS /ws/*`
(Real-time Updates)

AI Agents

`POST /api/summarizer/summarize`

`POST /api/classifier/classify`

`POST /api/document-analysis/analyze`

`POST /api/orchestration/workflow`

Management & Administration

`/api/documents/*`

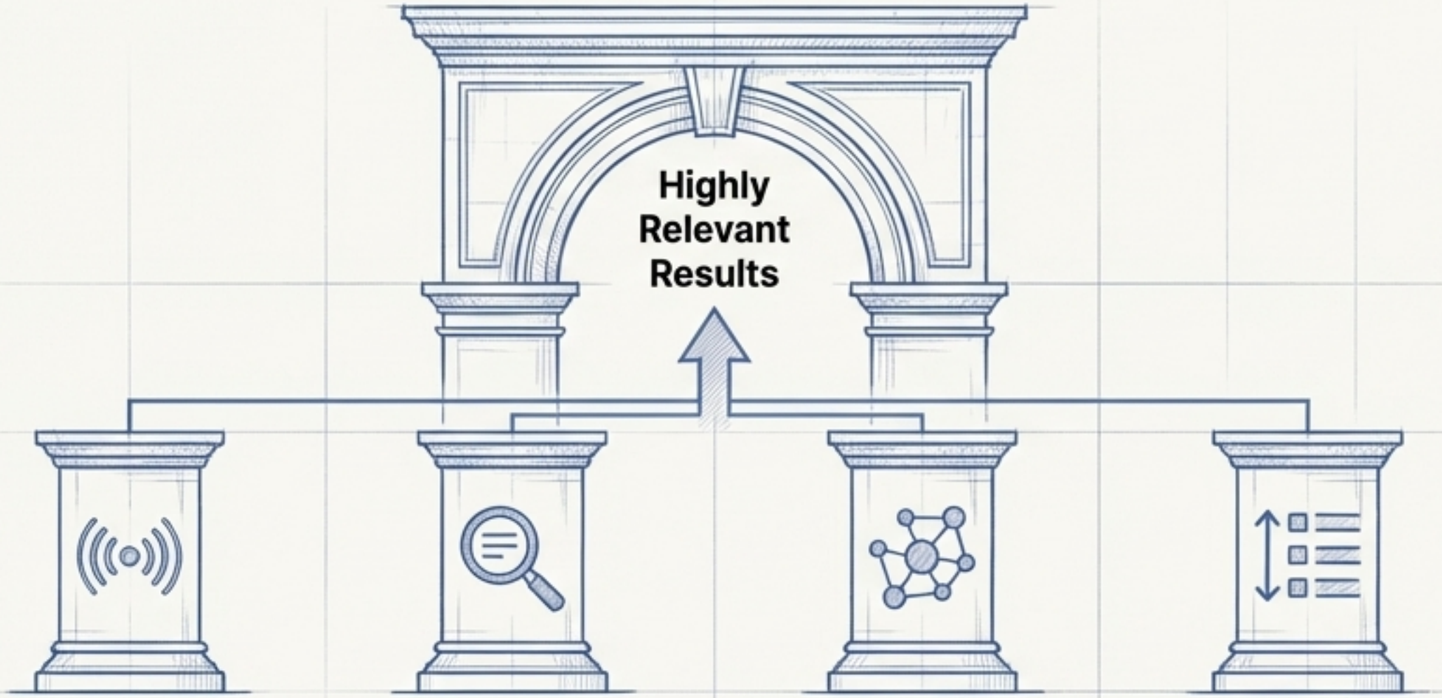
`/api/users/*`

`/api/rbac/*`
(Role-Based Access Control)

`/api/monitoring/*`

Demonstrated Capability: Advanced Search and Retrieval

The system's primary function is a sophisticated, four-part hybrid search that delivers highly relevant results by combining multiple retrieval strategies.



Vector Similarity

For semantic understanding and finding conceptually similar content (via BGE-M3 embeddings)

Full-Text Search

For precise keyword matching with ranking

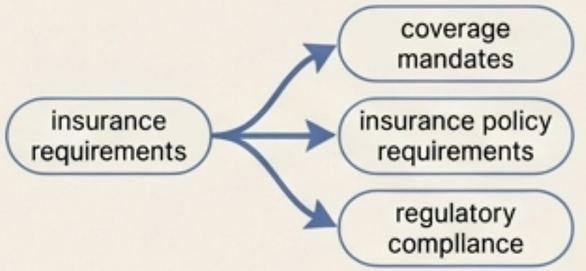
Graph Traversal

For discovering entity relationships and related concepts (via Neo4J)

Reranking

For ultimate precision, using BGE-Reranker-v2 to re-order the combined results for maximum relevance

Query Expansion with Claude Haiku



Automatically generates query variations, resulting in **15-30% better recall**.

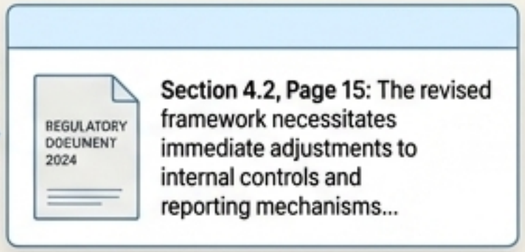
Demonstrated Capability: Core Features of a Production-Ready System

1. Source Attribution & Citations

What it is: Every AI response includes inline citations [1], [2] with page numbers and click-to-expand source context.

How it works: This is enabled by the meticulous document processing pipeline that maps every text chunk back to its original source document and location.

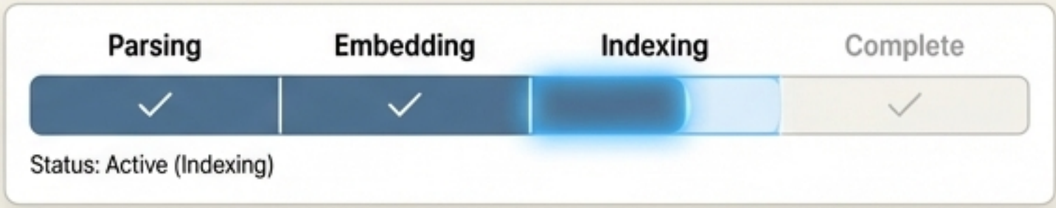
The new regulations require a significant overhaul of existing compliance procedures [1]. This will directly impact operational workflows across departments.



2. Real-Time Processing Status

What it is: Users see live progress bars and stage-by-stage updates (parsing → embedding → indexing) for document processing.

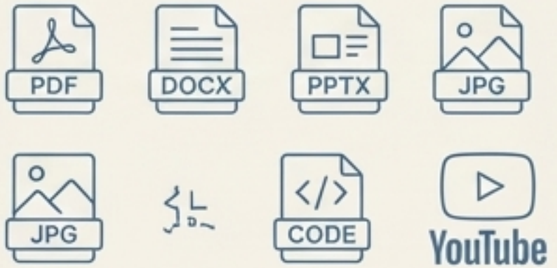
How it works: Leverages WebSocket (/ws/*) endpoints for real-time, bidirectional communication between the frontend and the Celery background workers.



3. Broad File and URL Support

What it is: Ingests and processes a wide range of formats including PDF, DOCX, PPTX, images, code files, YouTube videos, and web articles.

How it works: A flexible ingestion service detects content type and routes it to the appropriate parser before the standardized analysis pipeline begins.



The Immune System: A Multi-Layer Security Architecture

Transport: All services secured with **TLS 1.2+** encryption.

Authentication: User identity verified via **Clerk Auth + JWT tokens**.

Authorization: Data access is strictly controlled by **Row-Level Security (RLS)** on 14 critical database tables.

Rate Limiting: Protects against abuse with **Redis-backed, tiered limits** for different API endpoints.

Encryption at Rest: All data is secured with **AES-256** encryption in Supabase and Backblaze B2.

Audit Logging: A comprehensive audit trail tracks all significant events for security and compliance.

Key HTTP Security Headers (HSTS, CSP, X-Frame-Options) are enforced.

System Health: The Monitoring and Observability Stack

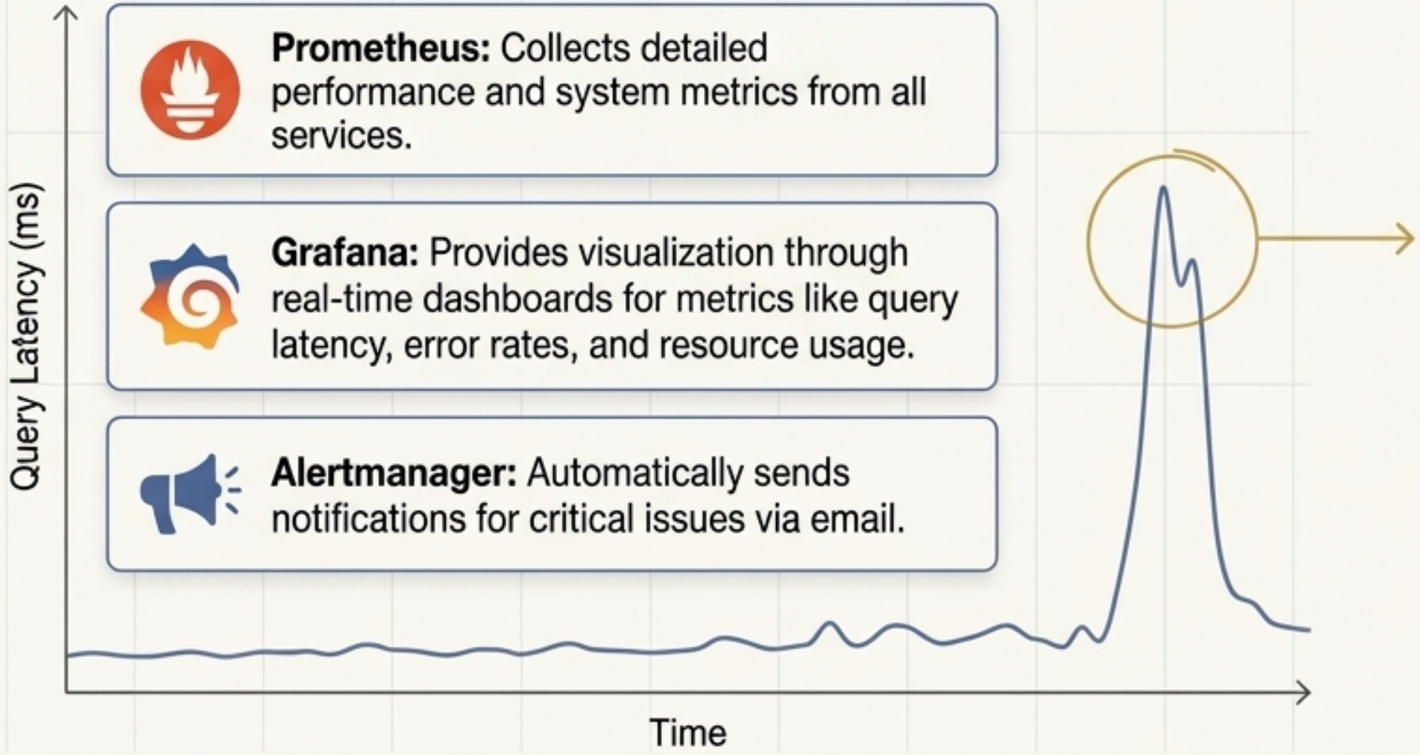
The system's health is continuously tracked through a robust monitoring stack, enabling proactive issue detection and performance analysis.

Core Components

Prometheus: Collects detailed performance and system metrics from all services.

Grafana: Provides visualization through real-time dashboards for metrics like query latency, error rates, and resource usage.

Alertmanager: Automatically sends notifications for critical issues via email.



39 Pre-configured Alert Rules

Rules are in place, categorized by severity to ensure the right level of response:

Critical: Service down, high error rate, extreme latency.

Warning: Elevated errors, slow processing, high resource usage.

Info: System health summaries and trends.

The Complete Architectural Blueprint

Layer	Technology
AI Processing	Claude Sonnet 4.5 (Anthropic API)
Backend Framework	FastAPI (Python 3.11+)
Task Processing	Celery
Vector Database	PostgreSQL + pgvector (on Supabase)
Graph Database	Neo4j (self-hosted)
Cache / Message Broker	Redis (on Upstash)
File Storage	Backblaze B2
Frontend UI	Gradio
Monitoring	Prometheus + Grafana + Alertmanager
Deployment	Render.com

Efficient, Complete, and Production-Ready

A lean infrastructure designed for efficiency.

Category	Monthly Cost
Render Services (4)	$28
Supabase PostgreSQL	$25
Backblaze B2 Storage	~$5
Upstash Redis	Free Tier
Neo4j (self-hosted)	$0
Total Infrastructure Cost	**~$60**

**Anthropic API costs are usage-based and separate from infrastructure.*

All 46 development tasks are complete.

46/46 TASKS COMPLETE

- ✓ **Phase 0: Foundation** (8 tasks) ✓
- ✓ **Phase 1: Sprint 1** (19 tasks) ✓
- ✓ **Phase 2: Sprint 2** (14 tasks) ✓
- ✓ **Phase 3: Sprint 3 (AI Agent System)** (5 tasks) ✓