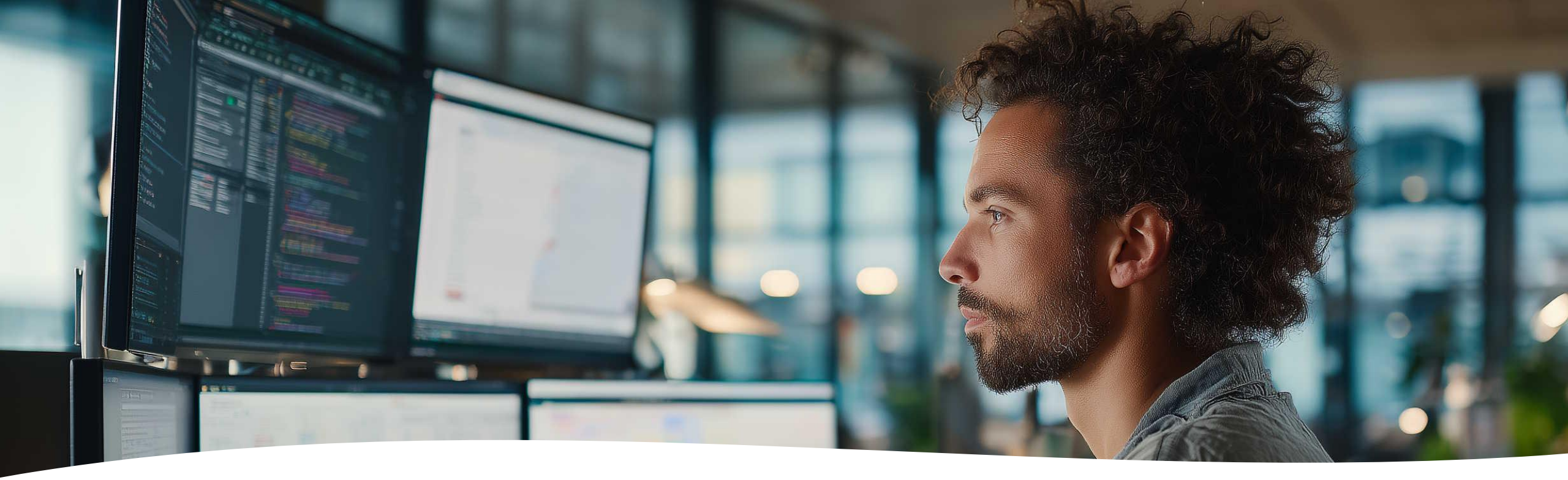
SharpAI: Enterprise Retrieval-Augmented Knowledge Assistant

SharpAI is a public-facing AI knowledge assistant integrated into a large technical content platform containing over 40,000 articles. The system combines semantic retrieval, vector embeddings, and controlled prompt orchestration to deliver context-grounded responses aligned with internal content. Designed and deployed in approximately five days, the project demonstrates how production-grade Retrieval-Augmented Generation systems can be implemented rapidly while remaining extensible to enterprise and domain-specific deployments such as legal research and compliance knowledge systems.

Knowledge Discoverability at Scale

- Organizations often accumulate tens of thousands of documents, articles, policies, or technical guides. Traditional keyword search surfaces links, but not answers. Users must manually scan, interpret, and synthesize information. This slide establishes the core problem SharpAI addresses: the gap between stored knowledge and actionable insight.

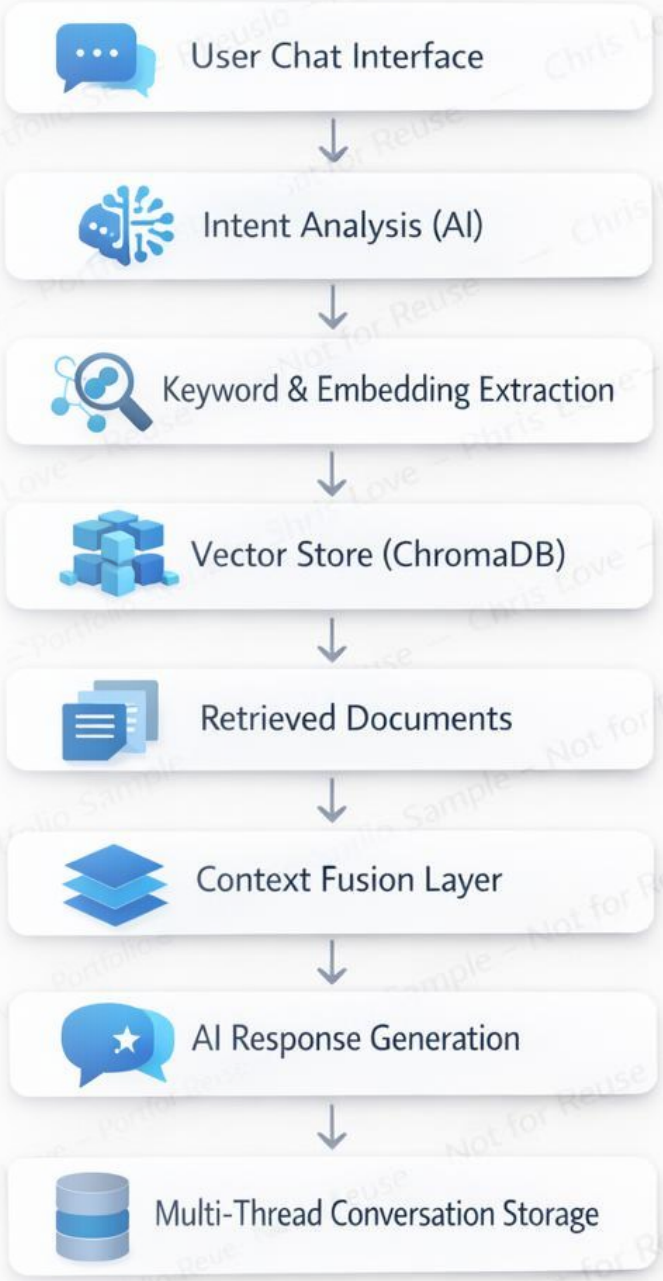


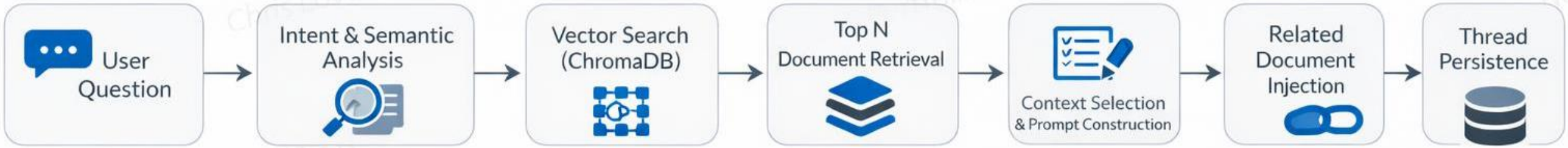
The SharpAI Approach

- SharpAI approaches AI interaction as a grounding and orchestration problem. Instead of generating free-form responses disconnected from internal content, the system retrieves semantically relevant documents using vector embeddings and conditions responses on those materials. This ensures higher specificity, improved credibility, and increased engagement within the platform. The slide introduces Retrieval-Augmented Generation conceptually before transitioning to architecture and workflow diagrams.

System Architecture Overview

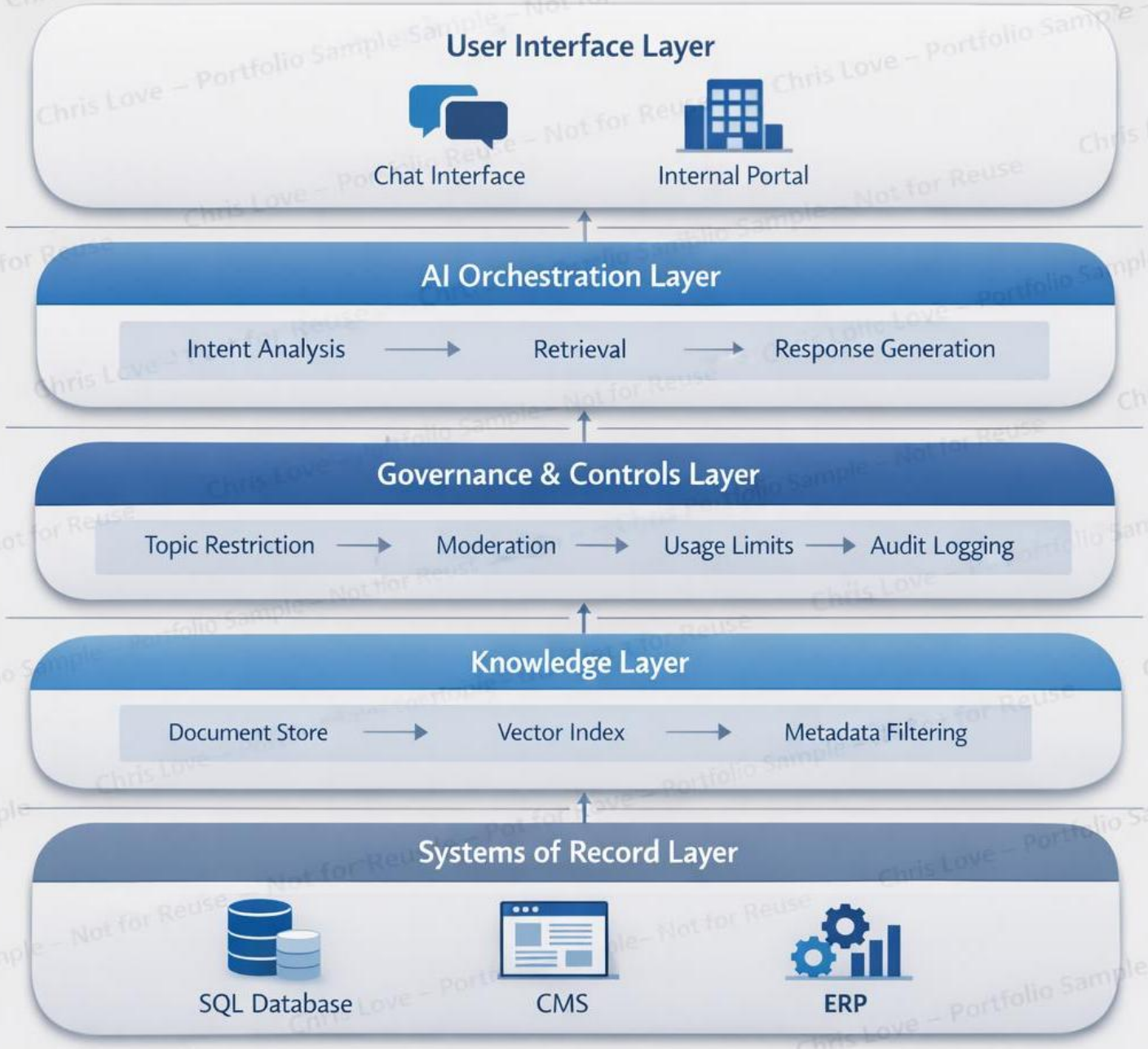
- User input enters through a conversational interface and is processed by an intent analysis layer.
- Extracted semantic signals are used to query a vector index stored in ChromaDB, retrieving relevant internal articles.
- Retrieved documents are fused into a structured prompt context before response generation.
- The system also persists multi-thread conversation state for ongoing research workflows.
- The diagram emphasizes separation of concerns, retrieval grounding, and structured orchestration.





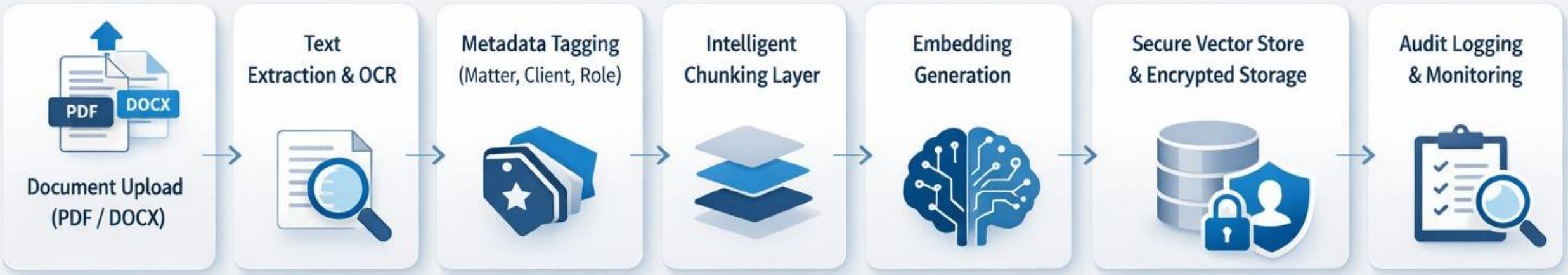
Retrieval-Augmented AI Workflow

- A user question is first analyzed for semantic intent and keyword signals.
- Those signals are used to query the vector store, retrieving the top relevant internal documents.
- The highest-value content is selected and fused into a structured prompt context.
- The AI then generates a response grounded in retrieved material.
- Finally, related article links are injected into the interface and the conversation thread is persisted.
- The diagram emphasizes staged orchestration, grounding, and traceability.



Enterprise Deployment & Governance Model

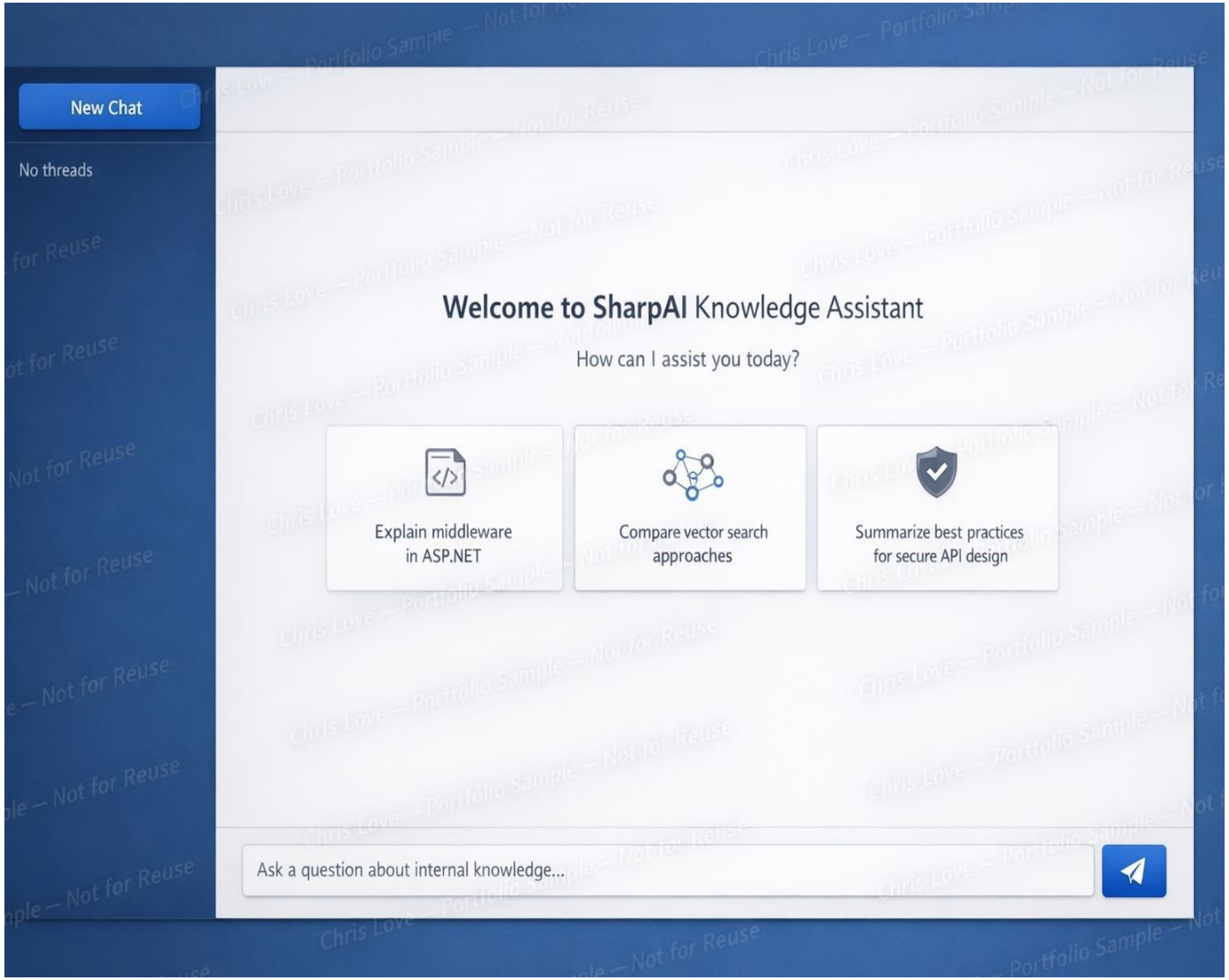
- SharpAI can operate within a governed enterprise environment.
- The top layer represents user interfaces such as chat or internal portals.
- Below that sits the AI orchestration layer, responsible for retrieval and generation.
- A governance layer enforces topic restrictions, moderation, usage controls, and audit logging. Beneath this is the knowledge layer, including vector indexes and document stores.
- Finally, underlying systems of record such as CMS, ERP, or SQL databases remain separated as authoritative sources.
- The diagram reinforces configurability, safety, and enterprise readiness.



Document Handling Pipeline

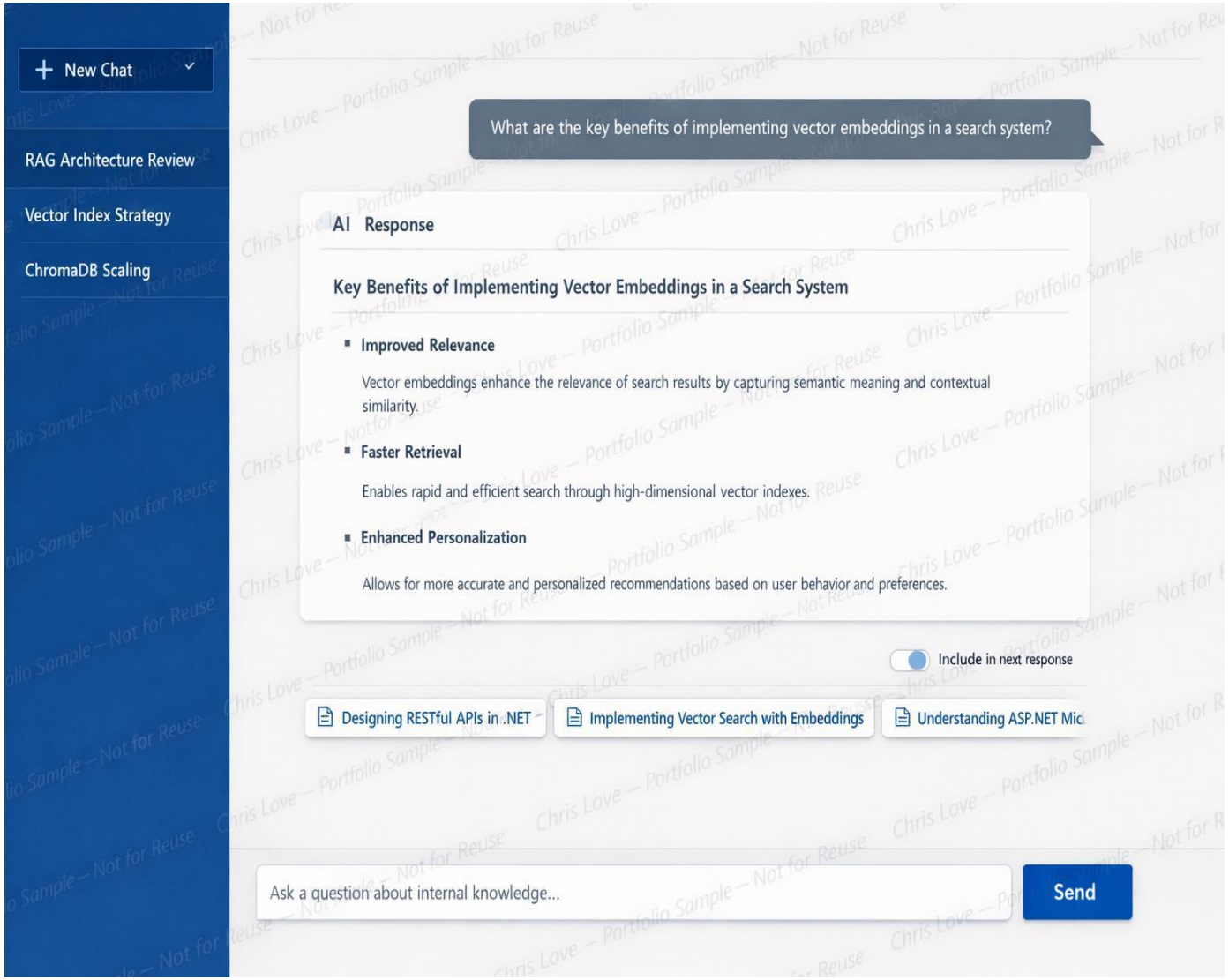
- Controlled document ingestion (PDF, DOCX, structured uploads)
- Intelligent text extraction with validation safeguards
- Matter-based metadata tagging at entry point
- Deterministic chunking designed for legal document structure
- Encrypted storage (vector + source separation)
- Permission-aware indexing for role-based retrieval
- Full audit trace for document access and AI citation

First-Time User Empty State



The initial state of SharpAI before any conversation threads have been created.
The interface introduces the assistant's purpose and provides example prompts aligned with permitted domains. This reinforces clarity and controlled scope.

Primary AI Chat Interface



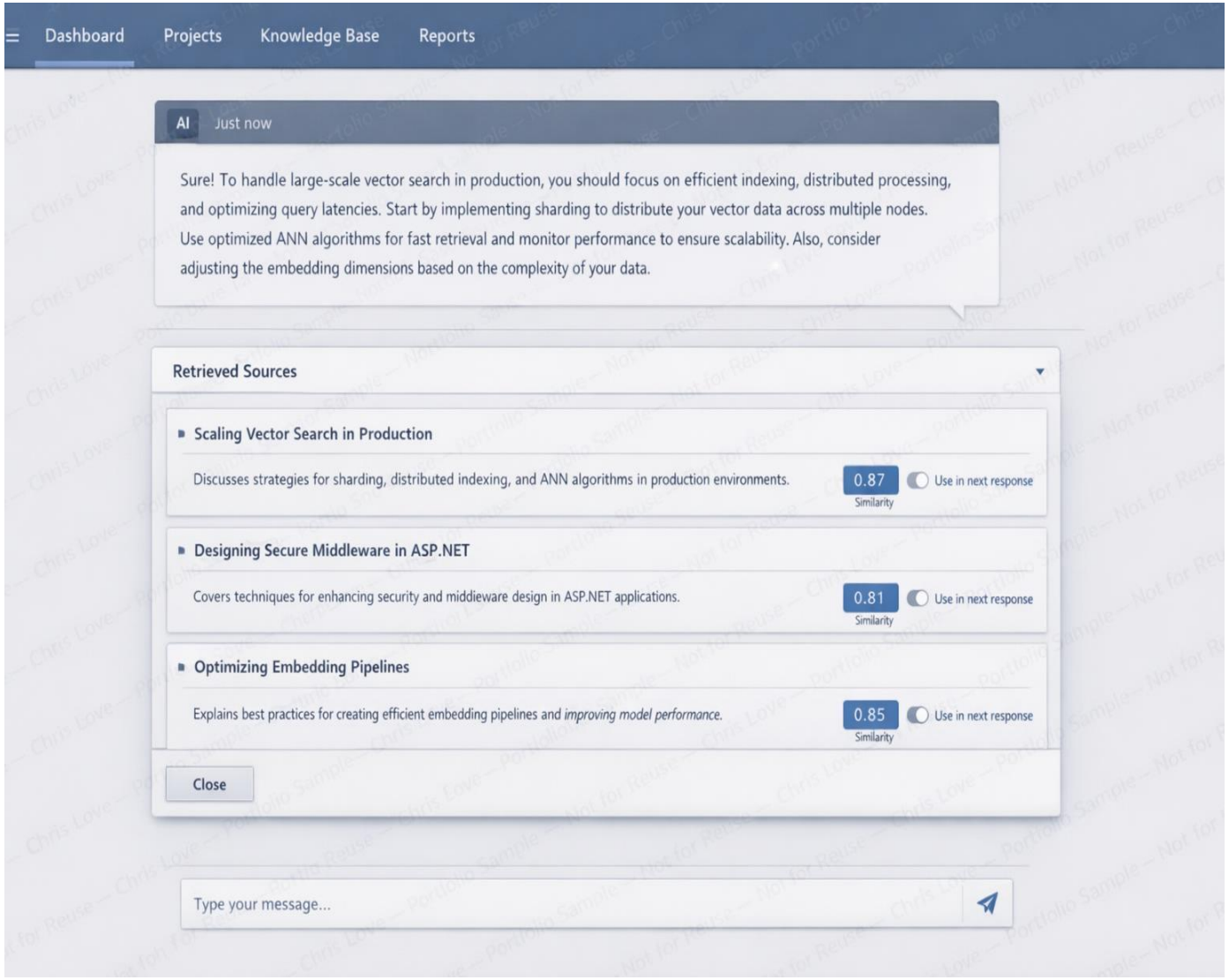
Users interact through a structured conversational interface while the system performs semantic retrieval behind the scenes.

AI responses are grounded in internally retrieved documents, and relevant articles are displayed as interactive pill-style links beneath each response.

The left-hand sidebar enables persistent multi-thread research sessions, allowing users to revisit, rename, and continue prior discussions.

The interface is intentionally restrained and enterprise-oriented, emphasizing clarity, credibility, and sustained research workflows.

Retrieval Transparency Panel



Users can expand a panel beneath an AI response to view which internal documents were retrieved through vector search.

Each document includes a title, short excerpt, and a relevance indicator.

Users may optionally select specific documents to seed the next response, reinforcing the system's grounded and controllable behavior.

Thread Management Workspace

Users can maintain separate research conversations, rename them for clarity, archive completed investigations, and revisit prior discussions.

The workspace reinforces that SharpAI is designed for sustained knowledge work rather than one-off chatbot interactions.

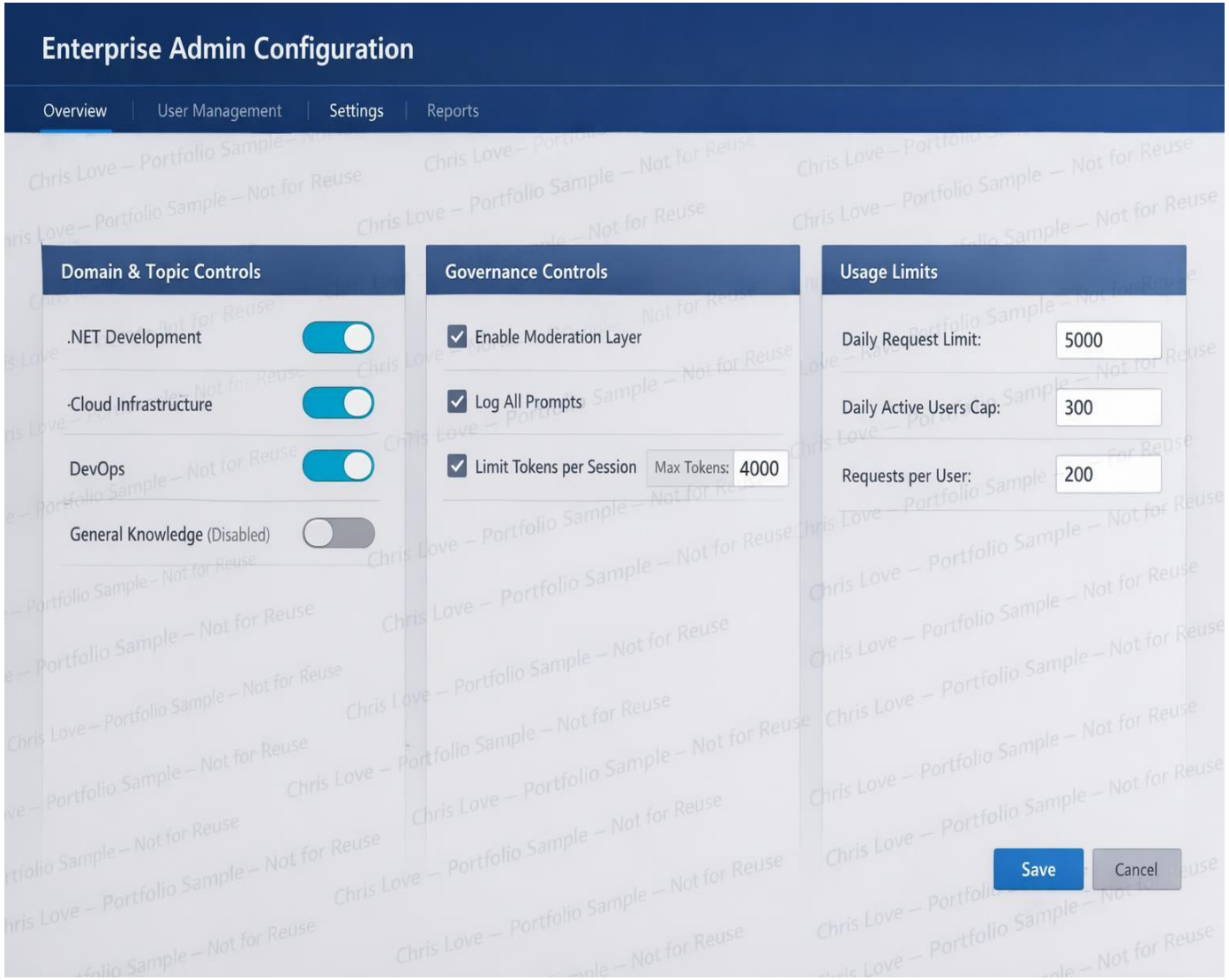
Content Ingestion & Embedding Pipeline

Administrators can monitor the

- Embedding of large document sets
- Review indexing status
- Validate synchronization with the vector store

The interface demonstrates production-level operational awareness rather than experimental AI usage.

Topic Restriction & Governance Configuration

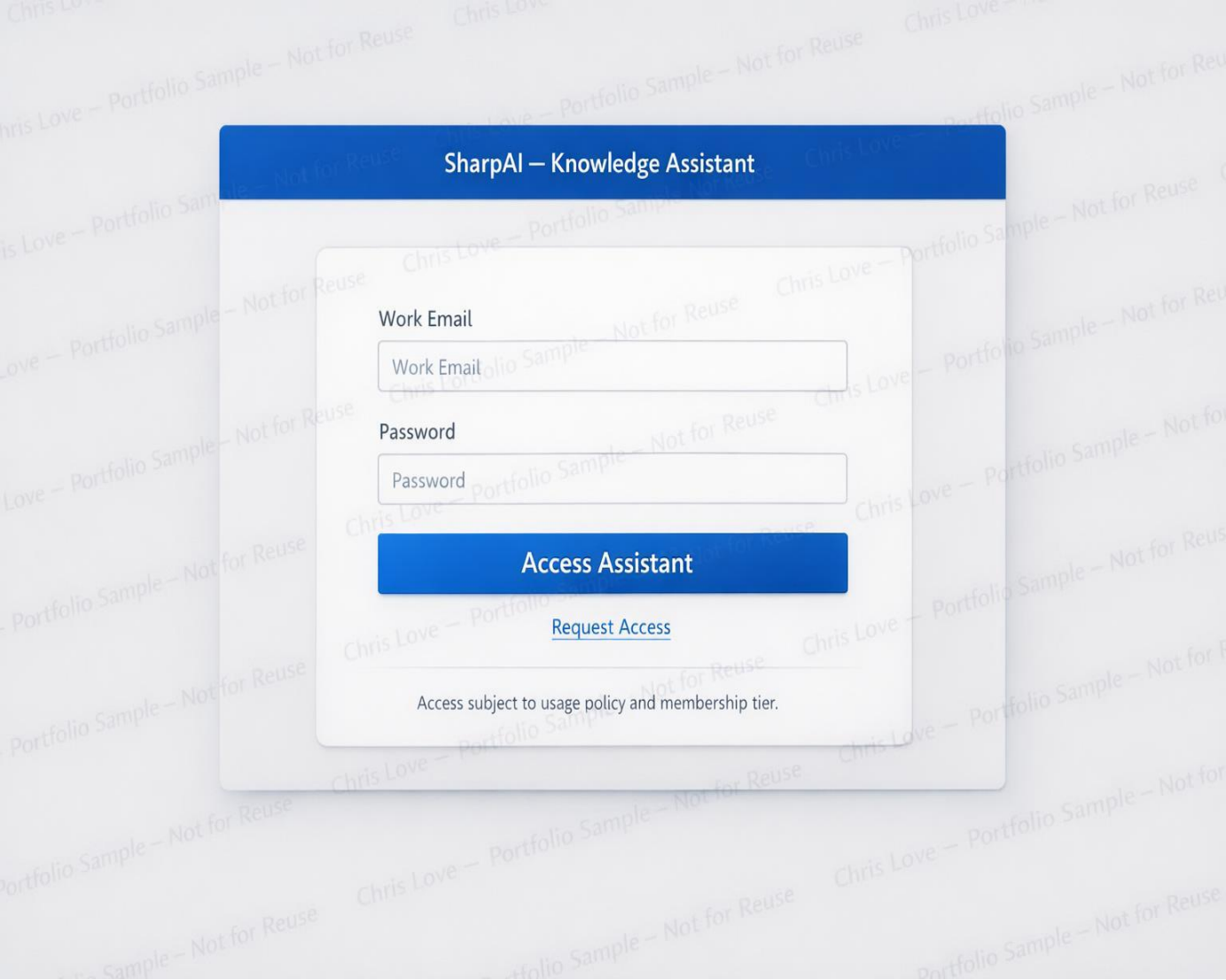


This screen illustrates SharpAI's governance capabilities for enterprise environments.

- Administrators can enable or restrict
- topic domains
- apply moderation controls
- configure usage limits

The configuration layer ensures that the AI operates within defined boundaries aligned with organizational policies.

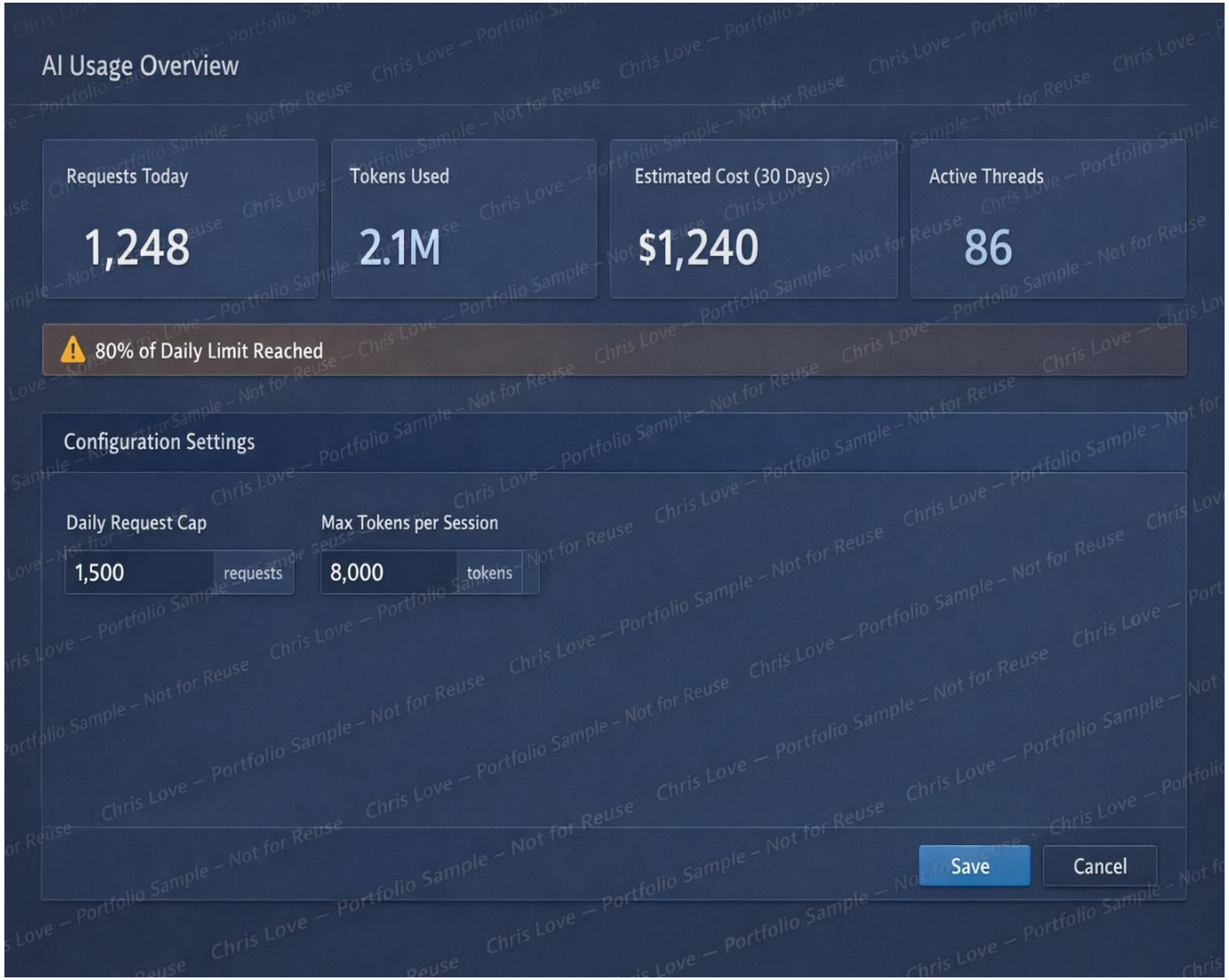
Authenticated Access Gate



Because AI usage incurs operational costs and potential liability, the assistant is positioned as a gated enterprise capability.

The interface emphasizes professional access rather than public experimentation.

Usage & Cost Governance

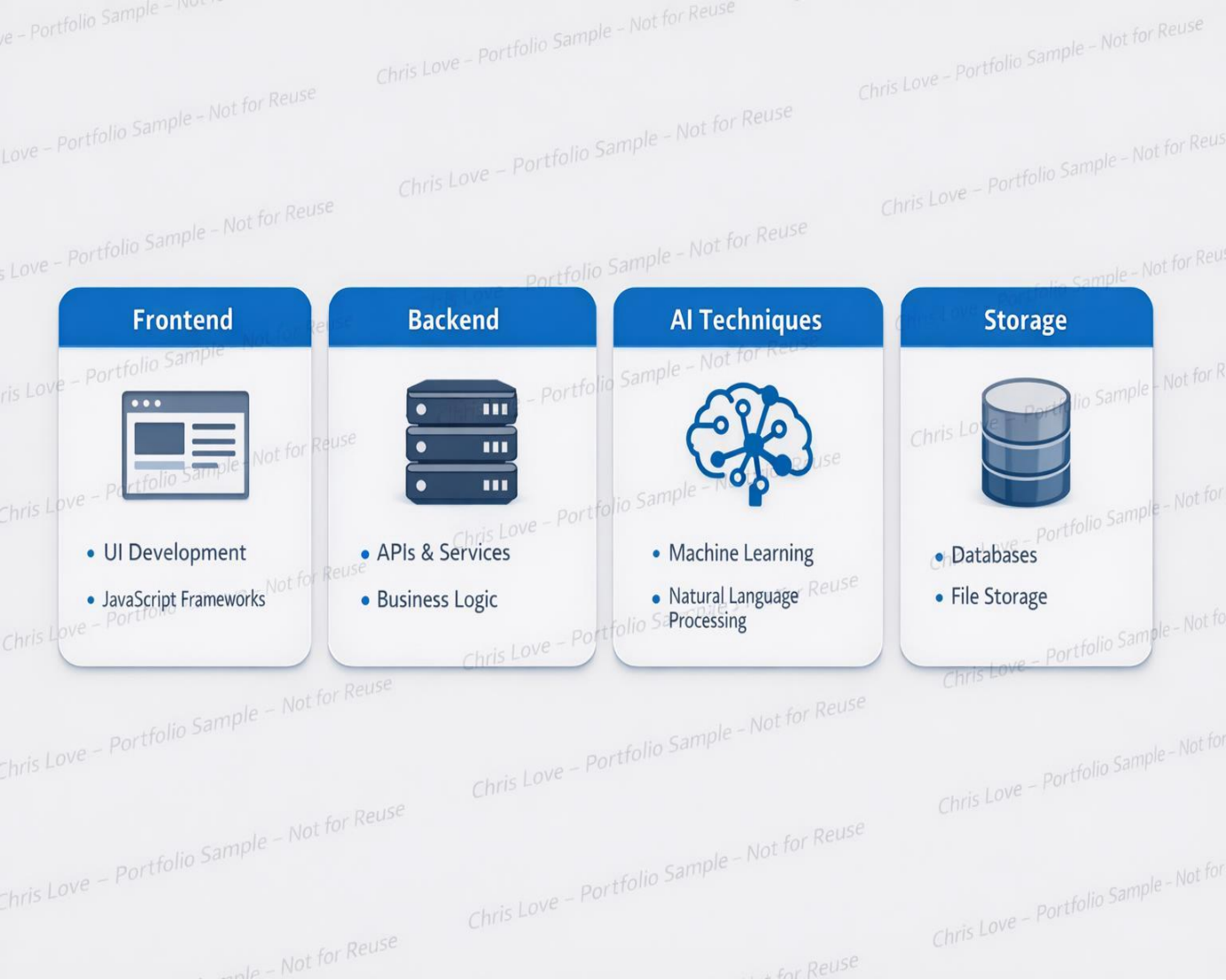


Administrators can monitor

- request counts
- token consumption
- estimated cost exposure.

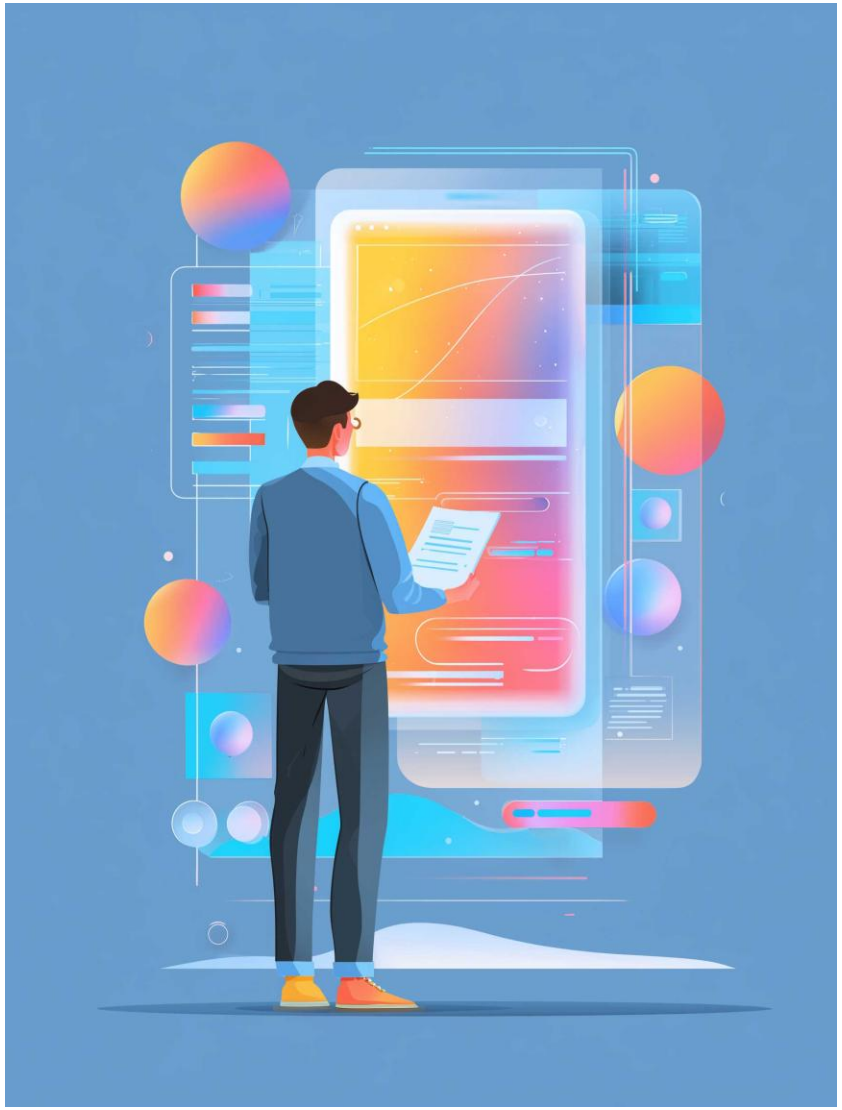
The interface reinforces responsible deployment and enterprise readiness.

Technical Stack Overview



- The system combines a Node.js backend with OpenAI APIs for semantic analysis and response generation.
- A ChromaDB vector store supports semantic retrieval across more than 40,000 embedded documents.
- The frontend implements a multi-thread conversational interface designed for enterprise research workflows.

Production Considerations

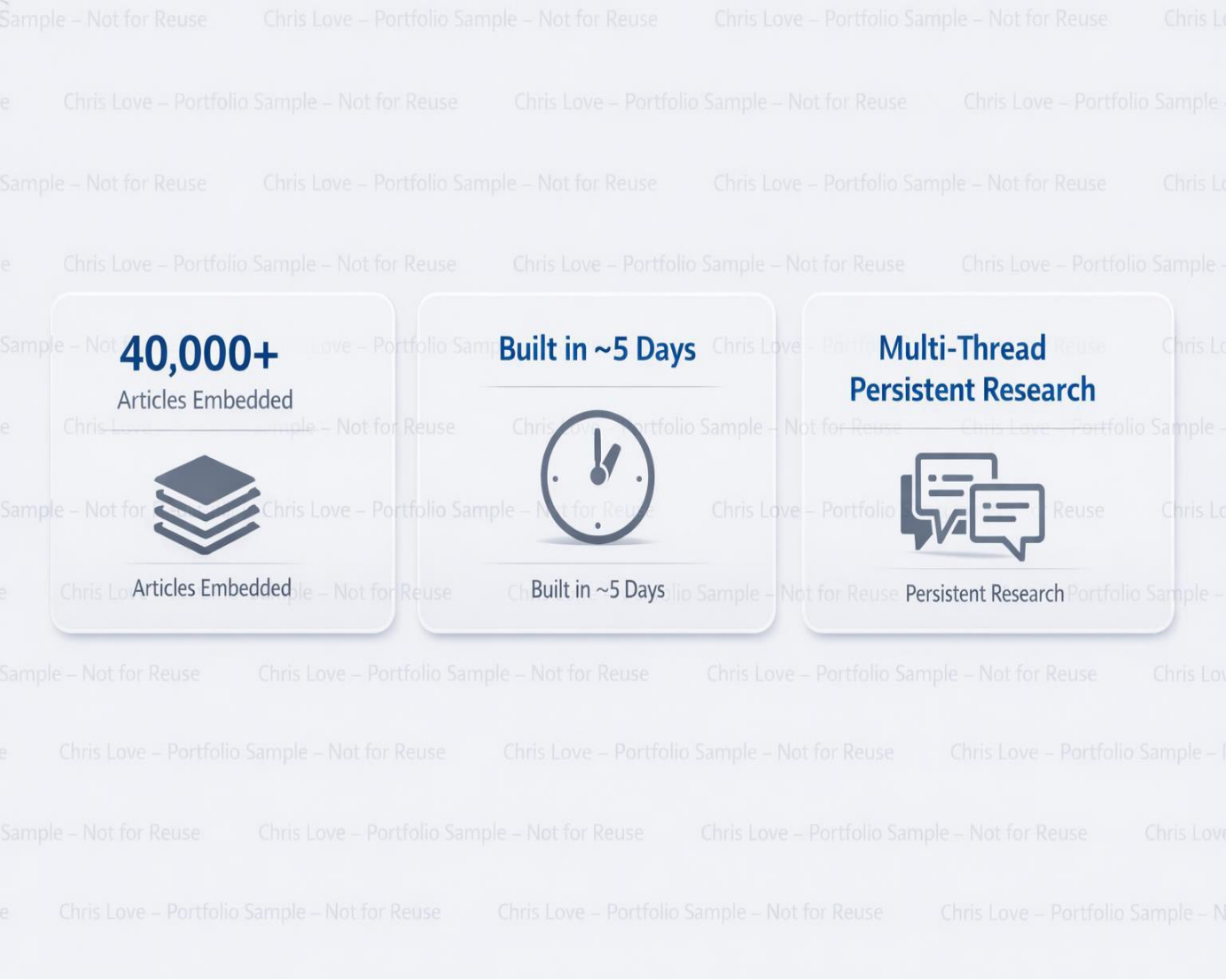


- Controlled prompt templates
- Retrieval augmentation (not raw LLM output)
- Response enrichment with site-specific references
- Modular provider abstraction (OpenAI swappable)

Security & Governance Extension Layer

- Role-based access control (RBAC)
- Matter-level document isolation
- Encrypted storage
- Audit logging
- Retrieval-time permission filtering
- Retrieval-time ACL enforcement
- No client-side document persistence
- Encrypted vector index
- Environment isolation
- Zero-trust API boundary

Key Technical Achievements



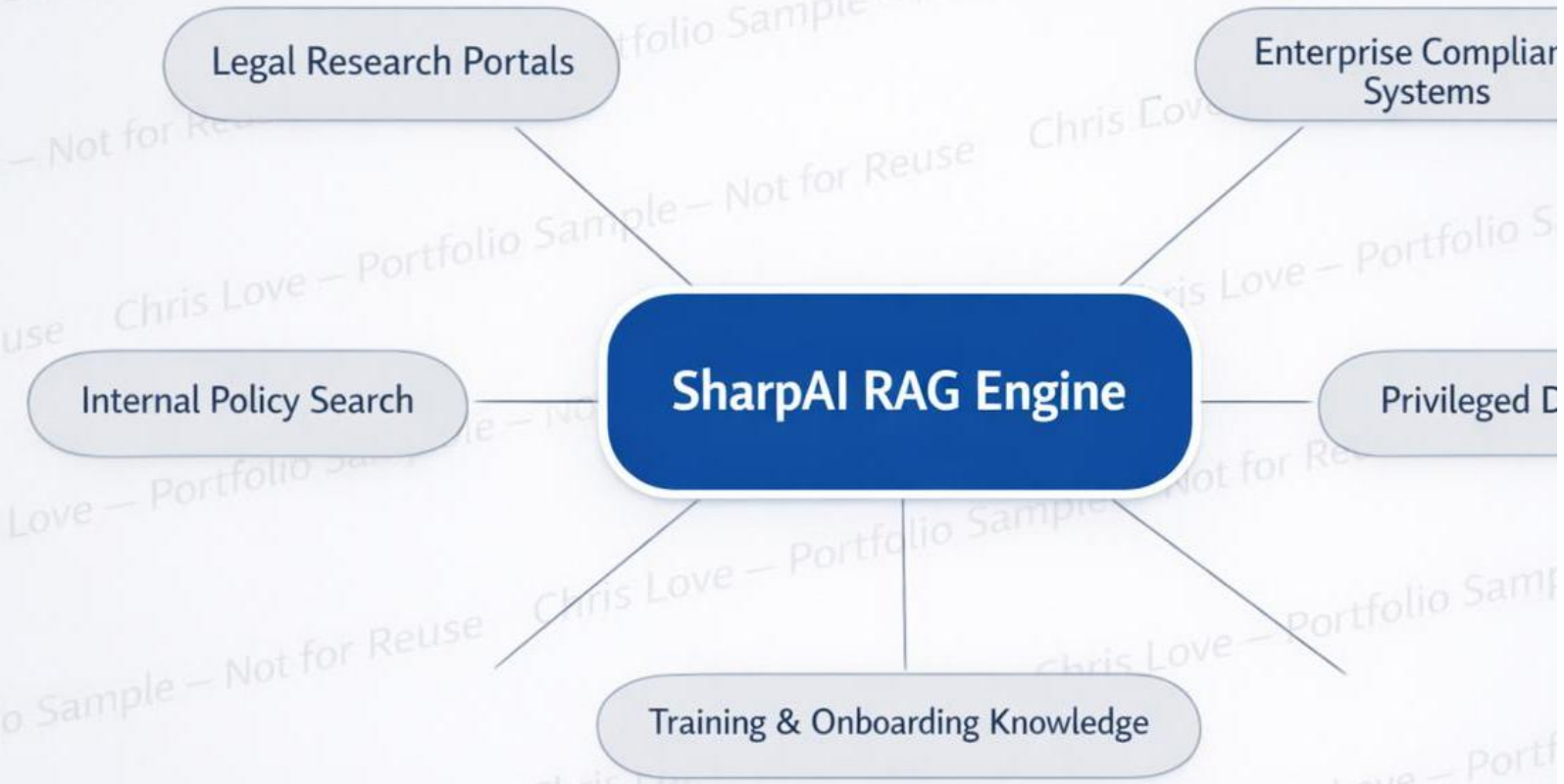
- Over 40,000 articles were embedded and indexed in under one week.
- A custom RAG pipeline was designed and deployed rapidly.
- Multi-thread persistence was implemented for sustained research workflows.
- The system reduced generic AI responses by grounding output in internal content and introduced contextual recommendation injection directly within the chat interface.

Security & Design Considerations

- Key risks addressed in SharpAI's design, including prompt injection exposure, context window management, retrieval precision tuning, and cost control. It also highlights architectural pathways for moderation, topic restriction, and audit logging. The purpose is to demonstrate engineering maturity and enterprise readiness rather than marketing claims.

Strategic Value of SharpAI

- The system demonstrates that enterprise-grade Retrieval-Augmented
- Generation can be implemented rapidly without sacrificing architectural clarity.
- It proves that conversational interfaces can increase engagement while grounding AI output in trusted internal knowledge sources.



Relevance to Legal & Enterprise Knowledge Systems

The RAG engine can be configured for internal legal research portals, compliance documentation search, enterprise policy systems, training knowledge bases, and privilege-preserving document AI.

The core transferable capability is secure, context-aware retrieval at scale.