Inside an Enterprise RAG Architecture

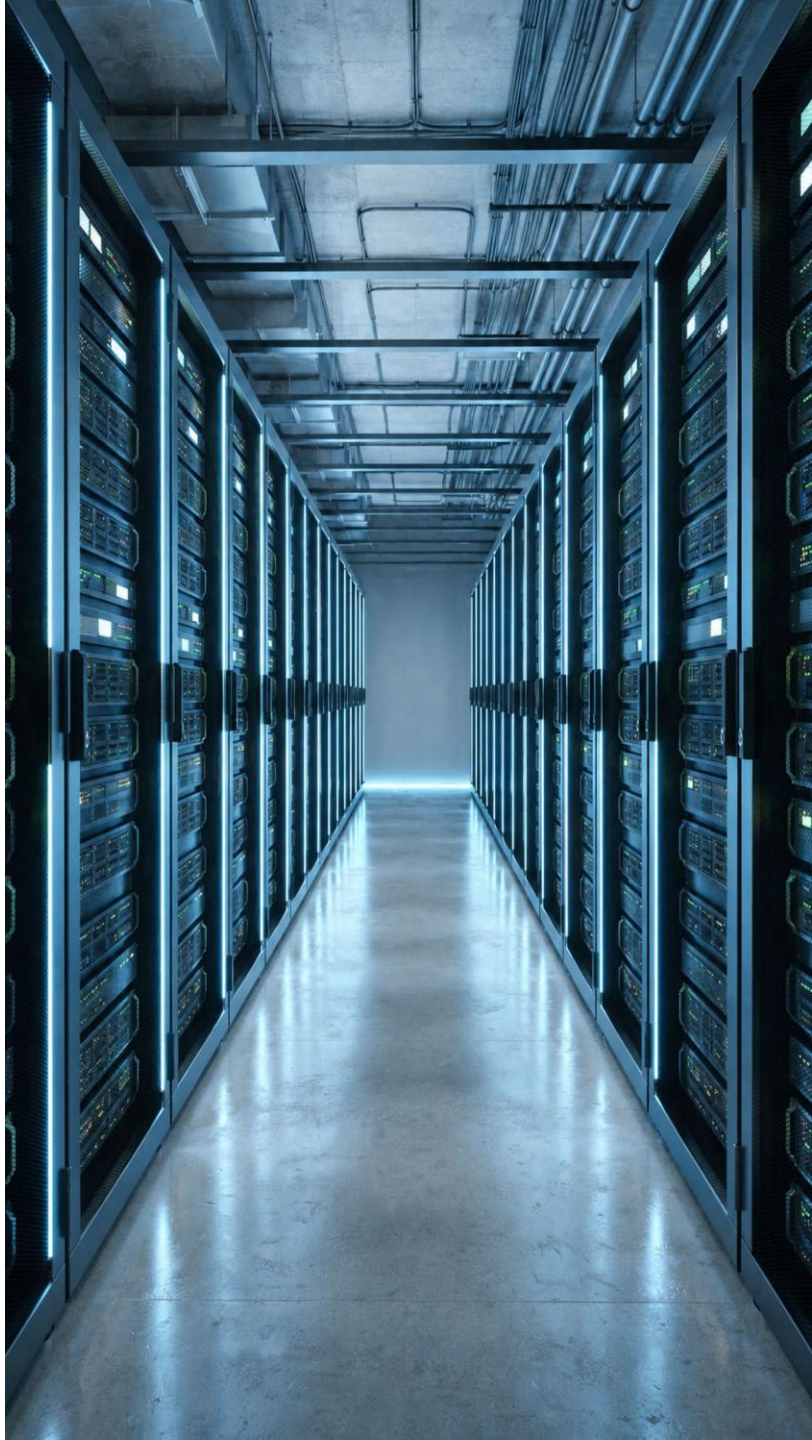
How LLMs, Embeddings, and Vector Databases Work Together

Enterprise AI requires more than the capabilities of general-purpose Large Language Models (LLMs). Retrieval-Augmented Generation (RAG) bridges this gap by grounding AI responses in an organization's trusted data through a pipeline of chunking, embeddings, vector databases, and context-aware retrieval.

This overview examines the key challenges organizations must address—including hallucinations, security and access control, data quality, and knowledge freshness—while highlighting the program manager's critical role in governing, scaling, and evolving RAG from an initial pilot into a secure, reliable, enterprise-grade platform.

Kimberly Wiethoff, MBA, PMP, PMI-ACP — AI Transformation Program Leader

#ArtificialIntelligence #GenerativeAI #RAG #RetrievalAugmentedGeneration #LLM #VectorDatabase #Embeddings #EnterpriseAI #AITransformation #AIGovernance #ProgramManagement #TechnicalProgramManagement #PMO #EnterpriseArchitecture



The Enterprise AI Gap

What General AI Knows

A general-purpose LLM understands industries, domains, and concepts — finance, healthcare, project management, operations. But it does not automatically know what lives inside your organization's systems, documents, or institutional memory.

What Enterprises Actually Need

Every organization generates enormous volumes of knowledge daily: policies, contracts, project plans, technical documentation, regulatory requirements, incident records, and years of lessons learned. The real value of enterprise AI begins when a model can securely access and use *that* knowledge.

The fundamental challenge: How do we let AI use trusted organizational knowledge without retraining the model every time that knowledge changes?

The Answer: Retrieval-Augmented Generation

RAG allows an organization to connect an LLM to trusted enterprise information so that responses are generated using relevant organizational knowledge — rather than relying solely on what the model learned during training.

More Than a Document Connection

RAG is often described as "connecting company documents to an LLM." That captures only part of the idea.

A Full Architecture

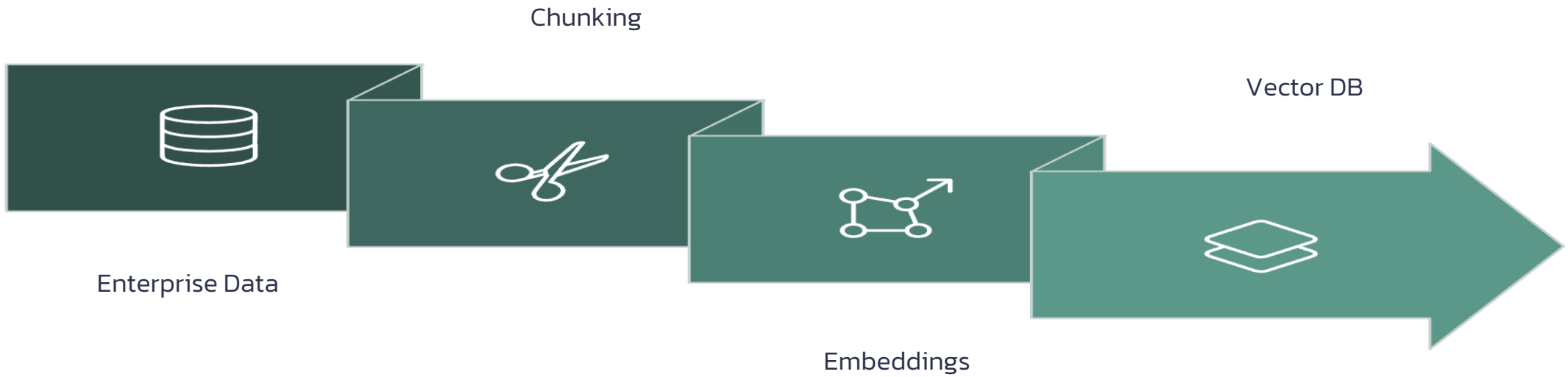
Enterprise RAG involves data pipelines, embeddings, vector databases, retrieval mechanisms, security controls, governance, and operational processes.

A Leadership Concern

Program managers and business leaders must understand how components interact, where failures occur, and what must be governed as AI moves to production.

The RAG Architecture at a Glance

Each stage in the pipeline plays a distinct role in determining the quality, accuracy, and trustworthiness of the final answer. A weakness at any point — from source data governance to retrieval ranking — can degrade the response the user receives. Understanding this end-to-end flow is essential for program leaders responsible for RAG delivery.



STEP 1

Enterprise Data: The Knowledge Foundation

Every RAG system begins with a critical question: **what knowledge should the AI be allowed to access?** Before debating platforms or models, organizations must determine and govern their sources.

Operational Knowledge

Policies, procedures, SOPs, training materials, project documentation, architecture records, incident and problem management records

Regulated Content

Contracts, regulatory documentation, compliance requirements, customer-support information, financial records

Institutional Knowledge

Lessons learned, product specifications, knowledge-base articles, SharePoint content, technical documentation

The First Real Challenge Is Information Governance

Before the first line of AI code is written, organizations must answer governance questions that are entirely non-technical in nature.

Authority & Ownership

Which version of a policy is authoritative? Who owns the document, and who is responsible for keeping it current?

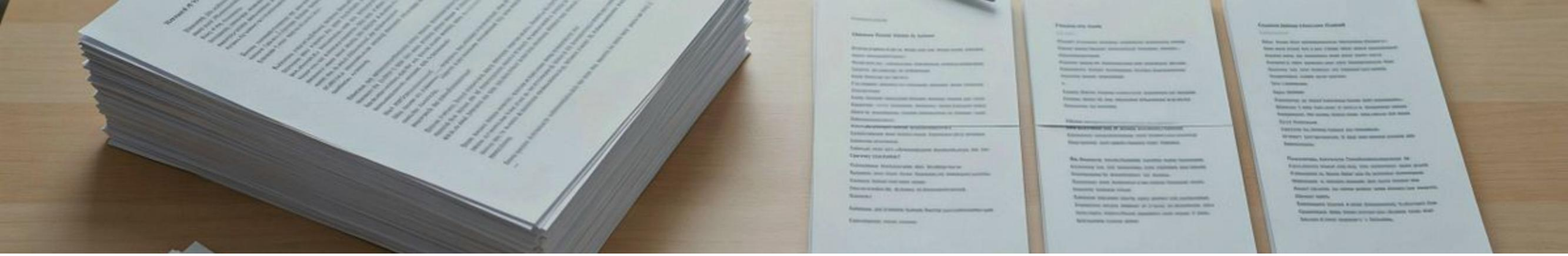
Classification & Access

Does the document contain confidential, regulated, proprietary, or personally identifiable information? Should every employee receive the same answer?

Retention & Currency

When was this document last updated? How long should it remain in the knowledge base? How quickly must changes be reflected in the AI's responses?

RAG cannot create trustworthy organizational knowledge from untrustworthy source information. The quality of the AI experience begins with the quality of the enterprise data behind it.



STEP 2

Chunking: Breaking Documents Into Usable Pieces

Why Chunking Is Necessary

An enterprise may contain millions of pages of information. Sending every relevant document to an LLM for each query would be inefficient, expensive, and often technically impossible — models have hard limits on how much text they can process in a single interaction.

Instead, documents are divided into smaller sections called **chunks**. A 60-page governance manual might be split into passages covering risk management, change control, financial governance, escalation procedures, and project closure.

Why Chunking Is a Design Decision

Too large: Chunks may contain significant amounts of irrelevant information that confuse the retrieval system.

Too small: Critical context may be separated from the information needed to interpret it correctly.

For example, a paragraph describing an approval threshold is meaningless if the heading identifying which business unit it applies to was placed in a different chunk. Chunking directly determines retrieval quality — and ultimately the quality of the AI's answers.

STEP 3

Embeddings: Turning Meaning Into Numbers

Once documents are chunked, the system needs a way to determine which chunks are *conceptually related* to a user's question — even when the exact words differ. That is where embeddings become essential.

What an Embedding Is

A numerical representation of meaning. Rather than encoding only literal words, embeddings capture relationships between concepts in a multidimensional mathematical space.

Why It Matters

"Project risk management" and "Identifying and mitigating delivery threats" use different words but share the same meaning. Traditional keyword search misses this. Embeddings recognize it.

Semantic Search Power

The user's question is also converted into an embedding. The system then compares it to stored chunk embeddings to find which pieces of organizational knowledge are closest in meaning.

STEP 4

Vector Databases: Where Meaning Is Stored and Searched

Traditional Databases

Exceptionally good at answering exact, structured queries:

- Find customer ID 10452
- Show all projects with status = Red
- Return invoices created after August 1

These systems match precise values. They are not designed to understand conceptual similarity.

Vector Databases

Designed for a fundamentally different question:

"Which pieces of information are most similar in meaning to this question?"

When the user's question is converted into an embedding vector, the system compares it to all stored vectors and surfaces the chunks that are semantically closest — even when the user's terminology doesn't match the source document's exact wording.

STEP 5

Retrieval: Finding the Most Relevant Information

When a program manager asks *"What is our escalation process for a critical production issue?"* — a RAG system doesn't send that question directly to the LLM. Something important happens first.

The system searches the approved enterprise knowledge base and retrieves the most relevant passages: severity classifications, escalation procedures, required stakeholders, response-time expectations, communication requirements, and executive-notification thresholds.

Retrieval quality is one of the most important factors determining whether a RAG application succeeds. A sophisticated LLM cannot compensate for consistently retrieving the wrong information.

- Wrong document retrieved
- Wrong section of the right document retrieved
- Important section not retrieved at all
- Outdated version retrieved

STEPS 6 & 7

From Context to Grounded Response

Step 6: Retrieved Information Becomes Context

The retrieved passages are added to the prompt supplied to the LLM. The model is not asked *"What do you know about this subject?"* — it is asked *"Based on this trusted information, what is the answer?"*

This distinction is at the heart of enterprise RAG. Organization-specific knowledge retrieved at the time of the request is combined with the model's general reasoning capabilities.

It also means organizations can update the knowledge available to the AI **without retraining the underlying LLM** every time a policy or procedure changes.

Step 7: The LLM Generates the Answer

The model receives the user's question and the enterprise context, then generates a grounded response — for example:

> *"Critical production incidents must be escalated to the Incident Manager and application owner within 15 minutes. Executive notification begins for Severity 1 incidents following the initial assessment."*

A mature RAG application also provides **citations** linking back to the source material — creating the traceability that regulated industries, legal workflows, and audit environments require.

RAG Reduces Hallucinations — But Doesn't Eliminate Them

One of the most common misconceptions about RAG is that it eliminates hallucinations. It does not. RAG substantially improves grounding, but multiple failure points remain across the entire pipeline.

Wrong document retrieved

The retrieval system surfaces the wrong source entirely

Outdated source information

The knowledge base reflects a policy that has since changed

Chunking context loss

Critical context was separated during document splitting

Model misinterpretation

The LLM incorrectly summarizes or misreads the retrieved material

Organizations must evaluate the **entire RAG pipeline** — not just the LLM. Quality is determined by: **Source Quality × Chunking × Embedding × Retrieval × Prompt × Model × Governance.**

Security and Access Control Cannot Be an Afterthought

A technically successful retrieval system may be capable of finding HR policies, financial records, legal documents, customer contracts, and executive strategy documents. That does not mean every employee should receive all of them.

Are permissions applied at retrieval time?

Security must be enforced where information is surfaced — not just at the application layer.

What happens when an employee changes roles?

Access changes must propagate quickly to prevent unauthorized retrieval.

Are prompts and responses logged and auditable?

Administrators must be able to audit which sources generated any given response.

Enterprise AI governance must extend beyond the model itself to the data and retrieval layers surrounding it.

Knowledge Freshness: The Hidden Operational Challenge

Enterprise Knowledge Is Always Changing

Policies are revised. Contracts are amended. Technical documentation changes after releases. Product specifications evolve. Projects close. Employees change roles. If the RAG knowledge base is not kept current, the AI can confidently surface information that was correct six months ago — and is incorrect today.

Questions Every RAG Program Must Answer

- When a source document changes, when is it reprocessed?
- When are chunks regenerated and embeddings refreshed?
- How is outdated information removed or superseded?
- How does the system determine which version is authoritative?

These are not operational details — at enterprise scale, they become **governance and service-management concerns**. A RAG system requires an ongoing operating model, not a one-time deployment.

RAG Is a Program, Not a Feature

A production RAG environment is an ecosystem — not simply "connecting documents to ChatGPT." Program leaders must coordinate across multiple technical and business disciplines simultaneously.

Data Engineering

Ingestion pipelines, document processing, metadata management, data quality

AI/ML Engineering

Embeddings, retrieval strategies, model selection, evaluation and optimization

Cybersecurity

Identity, authorization, encryption, data protection and monitoring

Compliance & Legal

Regulatory requirements, privacy, retention, and acceptable-use policies

Business Teams

Authoritative content ownership, use cases, acceptance criteria, business outcomes

Operations

Monitoring, incident management, change control, continuous improvement

The Program Manager's Role in Enterprise RAG

Many of the hardest enterprise AI problems are not purely technical. They involve ownership, governance, dependencies, risk, operating models, adoption, and organizational change — familiar program-management challenges in a new technological environment.

01

Define Data Ownership

Who owns the source data? Who determines which documents are authoritative? Who approves new knowledge sources?

02

Establish Governance Processes

How are user permissions enforced? How frequently are embeddings refreshed? How are models, prompts, and retrieval configurations versioned?

03

Define Quality Standards

Who defines acceptable retrieval quality? How are incorrect answers reported and investigated? What metrics determine production readiness?

04

Own the Operating Model

What happens when a source document changes? Who owns the solution after implementation? How will continuous improvement be managed?

What Program Managers Should Measure

Measuring whether employees "like the chatbot" is not enough. Organizations need metrics across technical, operational, user, and business layers.

Retrieval Accuracy

Did the system retrieve the information required to answer the question?

Groundedness

Are the claims in the answer supported by retrieved enterprise information — with accurate citations?

Latency & Cost

How quickly does the system respond? What does each query cost at enterprise scale?

User Adoption

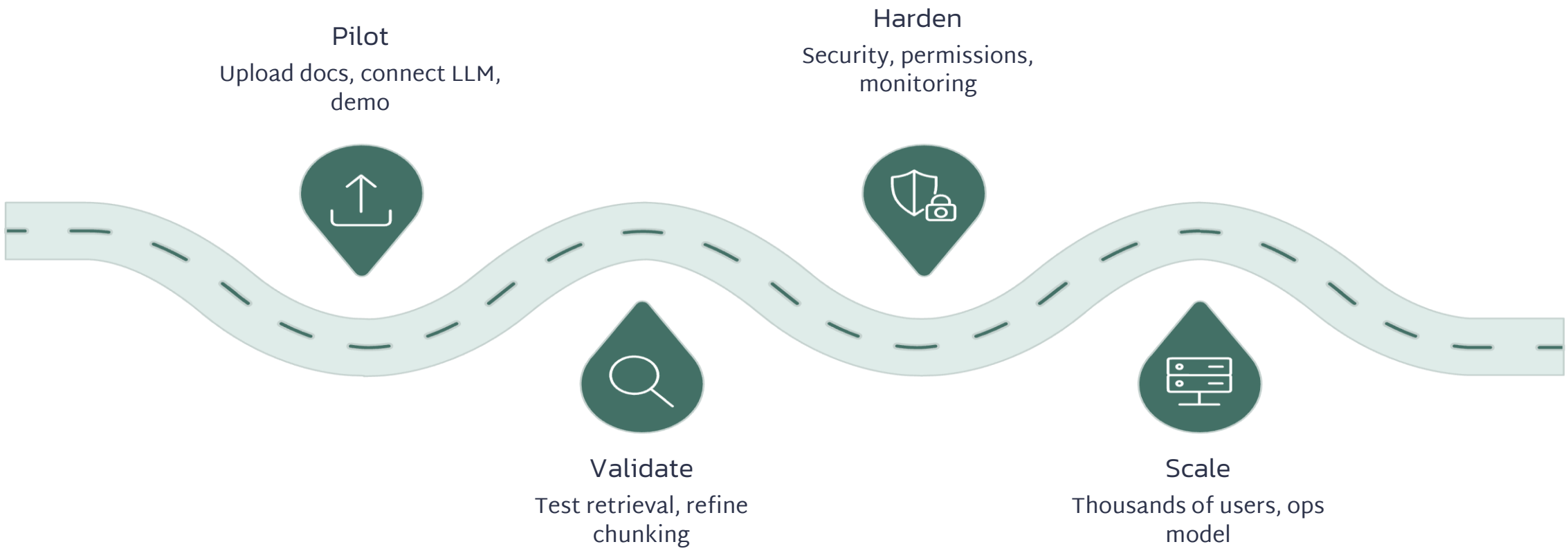
Are employees using the solution — and returning to it? Are they completing the tasks they came to accomplish?

Business Impact

Is the system improving a measurable outcome? Support resolution time, onboarding speed, policy compliance, analyst productivity?

A technically impressive system that does not improve a measurable business outcome is not a successful transformation.

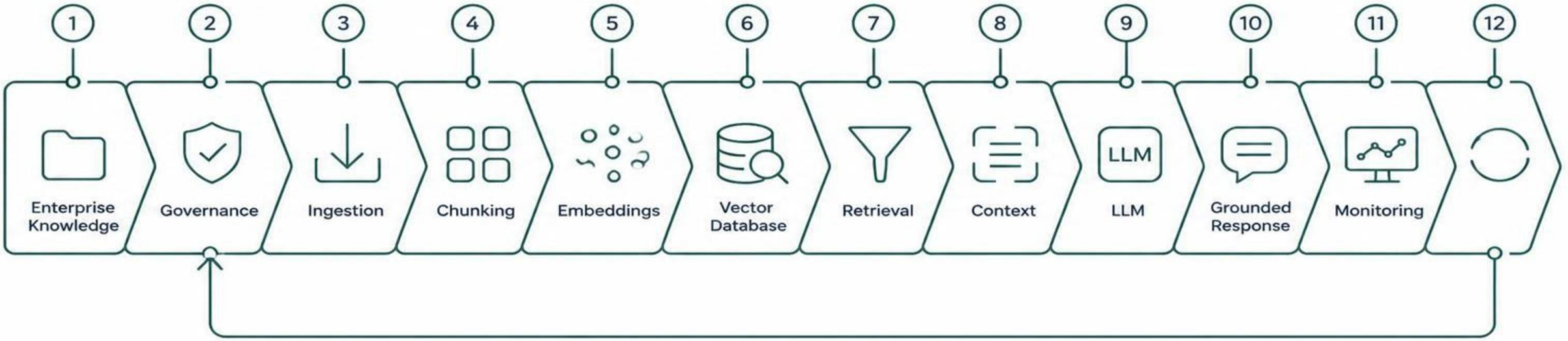
From RAG Pilot to Enterprise Platform



A proof of concept proves technical feasibility. Production introduces an entirely different set of demands: scaling to thousands of users and millions of documents, synchronizing permissions, monitoring retrieval quality, controlling costs, handling incidents, versioning models and prompts, and enabling employees to report incorrect answers. This is where many AI initiatives encounter the gap between successful demonstration and sustainable enterprise capability. The technology is only one part of crossing that gap — the rest is program execution.

RAG Is Really Knowledge Architecture

RAG represents a fundamental shift: instead of relying entirely on what an LLM learned during training, enterprises provide controlled access to relevant organizational knowledge at the moment a user asks a question. The complete picture is far richer than a simple document connection.



Every component introduces decisions involving architecture, security, governance, ownership, performance, cost, compliance, and operational support. The organizations that succeed will not simply have the largest LLM — they will have the most trusted knowledge architecture surrounding it.

The Competitive Advantage Is Trusted Knowledge

Right knowledge.
Right time.
Right user.
Right controls.

The Question Is Shifting

As generative AI moves deeper into enterprise operations, the strategic question will increasingly shift from *"Which AI model should we use?"* to *"How do we build the trusted enterprise architecture around it?"*

Where Program Leadership Becomes Critical

The organizations that win with RAG will know which information is authoritative, control who can access it, continuously evaluate retrieval and response quality, monitor cost and performance, establish clear ownership — and connect everything to measurable business outcomes.

For Program Managers

Understanding RAG architecture may soon be as essential as understanding cloud migrations, Agile delivery, cybersecurity, or DevOps. And that is where experienced program leadership becomes the decisive factor.

INSIDE AN ENTERPRISE RAG ARCHITECTURE:

How LLMs, Embeddings, and Vector Databases Work Together

Retrieval-Augmented Generation (RAG) connects your organization's trusted knowledge to a Large Language Model—delivering accurate, relevant, and traceable AI responses.

1 ENTERPRISE DATA

Policies, procedures, documents, knowledge bases, emails, tickets, contracts, manuals, and more.

2 CHUNKING

Documents are split into smaller, meaningful chunks to preserve context and improve retrieval accuracy.

3 EMBEDDINGS

Each chunk is converted into an embedding—a numerical representation of meaning in a multi-dimensional space.

4 VECTOR DATABASE

Embeddings are stored in a vector database designed for similarity search and semantic retrieval.

5 RETRIEVAL

When a user asks a question, the system retrieves the most relevant chunks based on semantic similarity.

6 AUGMENTED PROMPT

The retrieved content is added to the user's question to provide context for the LLM.

7 LLM GENERATES A GROUNDED RESPONSE

The LLM uses the provided context to generate an accurate, relevant, and traceable response—with citations.

KEY CONSIDERATIONS FOR ENTERPRISE RAG SUCCESS

GOVERNANCE & SECURITY
Classify data, enforce access controls, and protect sensitive information.

DATA QUALITY
Clean, current, and authoritative data leads to better retrieval and better answers.

RETRIEVAL QUALITY
The right chunks, metadata, and filtering drive accuracy and relevance.

MONITOR & MEASURE
Track retrieval accuracy, user adoption, performance, and business impact.

CROSS-FUNCTIONAL COLLABORATION
IT, data, security, legal, and business teams must work together to deliver trusted AI at scale.

The right information.
The right context.
The right answer.
That's the power of RAG.

THE BUSINESS IMPACT

MORE ACCURATE RESPONSES
Grounded in your trusted, organization-specific information.

REDUCED RISK
Less chance of hallucinations and incorrect or out-of-date information.

TRACEABILITY
Citations and source links create transparency and build user trust.

GREATER PRODUCTIVITY
Employees get faster answers so they can focus on higher-value work.

SCALABLE VALUE
Consistent, governed AI that grows with your data and your business.

Kimberly Wiethoff, MBA, PMP, PMI-ACP
Senior Program Manager | AI-Enabled Delivery Leader
Managing Projects the Agile Way

> RAG turns your organization's knowledge into a strategic advantage—delivering the right answer, at the right time, every time.

Read the full article in my latest newsletter at *Managing Projects the Agile Way.*