ENTERPRISE RETRIEVAL-AUGMENTED GENERATION WITH C#

BUILDING PRODUCTION-GRADE AI APPLICATIONS IN THE .NET ECOSYSTEM

REAL-WORLD PATTERNS.
PROVEN ARCHITECTURES.
PRODUCTION READY.

- HYBRID SEARCH & RETRIEVAL
- VECTOR DATABASES
- INGESTION PIPELINES
- SECURITY & SAFETY
- EVALUATION & OBSERVABILITY
- SCALABILITY & DEPLOYMENT

STEVE T.

Enterprise Retrieval-Augmented Generation with C#

Building Production-Grade AI Applications in the .NET Ecosystem

Steve T. Team Publications

This book is available at
https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc

This version was published on 2026-07-03

This is a Leanpub book. Leanpub empowers authors and publishers with the Lean Publishing process. Lean Publishing is the act of publishing an in-progress ebook using lightweight tools and many iterations to get reader feedback, pivot until you have the right book and build traction once you do.

Contents

Building Production-Grade AI Applications in the .NET Ecosystem . . . **1**

Introduction: The Enterprise RAG Challenge **2**

Chapter 1: Foundations of RAG and the .NET AI Ecosystem **5**
What RAG Actually Solves (and What It Doesn't) 5
The Evolution from Prompt Engineering to Retrieval-Augmented Systems 5
The Modern .NET AI Stack: Microsoft.Extensions.AI, Semantic Kernel, and Agent Framework 5
Choosing Your LLM Provider: Azure OpenAI, OpenAI API, and Local Models 5
A Quick-Start RAG App in .NET 6
From Prototype to Production: Refactoring into a Layered Architecture 6
Key Takeaways 6

Chapter 2: Embeddings and Vector Representations **7**
How Embeddings Encode Meaning 7
Choosing an Embedding Model: Accuracy vs. Speed vs. Cost 7
Domain-Specific Embeddings and Fine-Tuning 7
Dimensionality, Quantization, and Storage Efficiency 7
Evaluating Embedding Quality 7
Key Takeaways 8

Chapter 3: Vector Databases for .NET Developers **9**
The Landscape of Vector Stores in 2026 9
Qdrant: Performance and Operations 9
PostgreSQL + pgvector: Consolidation and Simplicity 9
Azure AI Search: Managed Hybrid Search 9
Microsoft.Extensions.VectorData Abstraction Layer 9
Making the Choice: A Decision Framework 10

Key Takeaways 10

Chapter 4: Building Robust Ingestion Pipelines 11
The Ingestion Pipeline Architecture 11
Document Parsing and Layout-Aware Extraction 11
Chunking Strategies: Fixed, Recursive, Semantic, and Structural 11
Metadata Enrichment and Classification Tags 11
Incremental Ingestion and Change Detection 11
Key Takeaways 12

Chapter 5: Hybrid Search and Retrieval Quality 13
Why Dense Vector Search Is Not Enough 13
Hybrid Search with RRF in Azure AI Search 13
Cross-Encoder Reranking 13
Query Transformation Techniques 13
Designing a Production Retriever 13
Key Takeaways 14

Chapter 6: Prompt Engineering for Grounded Generation 15
The Anatomy of a RAG Prompt 15
Context Assembly Strategies: Write, Select, Compress, Isolate 15
Citation and Grounding Techniques 15
Anti-Hallucination Guardrails in Prompts 15
Prompt Versioning and Experimentation 15
Key Takeaways 16

Chapter 7: Agentic RAG with the Microsoft Agent Framework 17
From RAG to Agentic RAG 17
Microsoft Agent Framework 1.0: Architecture and Primitives 17
Sequential and Concurrent Workflows 17
Agentic Retrieval Patterns 17
Human-in-the-Loop and Approval Gates 17
Key Takeaways 18

Chapter 8: Security, Privacy, and Governance 19
The RAG Security Threat Model 19
Document-Level Access Control at Retrieval Time 19
PII and PHI Redaction Pipelines 19
Prompt Injection and Content Guardrails 19
Audit Trails and Compliance 19

Key Takeaways 20

Chapter 9: Evaluation Frameworks for RAG Systems 21
What to Evaluate in a RAG System 21
Retrieval Metrics: Precision, Recall, MRR, NDCG 21
Generation Metrics: Faithfulness, Groundedness, Answer Relevance . 21
Microsoft.Extensions.AI.Evaluation in Practice 21
Building Custom Evaluators and CI/CD Gates 21
Key Takeaways 22

Chapter 10: Observability and Debugging RAG Systems 23
The Observability Challenge in RAG 23
End-to-End Tracing with OpenTelemetry 23
Metrics and Dashboards 23
Root-Cause Analysis Playbook 23
Debugging Tools and Techniques 23
Key Takeaways 24

Chapter 11: Scalability and Performance Optimization 25
The Latency Budget Problem 25
Semantic Caching Strategies 25
Async Pipelines and Concurrent Retrieval 25
Model Routing and Tiered Inference 25
Infrastructure Scaling Patterns 25
Key Takeaways 26

Chapter 12: Testing and Quality Assurance 27
What to Test in a RAG System 27
Unit Testing Retrieval and Generation 27
Integration Testing with Testcontainers 27
Regression Evaluation and Golden Datasets 27
A/B Testing and Prompt Experiments 27
Key Takeaways 28

Chapter 13: Deployment, CI/CD, and Cost Management 29
Containerizing a RAG Application 29
CI/CD Pipelines for AI Applications 29
Kubernetes Deployment Patterns 29
Production Runbooks and Incident Response 29
Cost Management and Token Budgeting 29

The Complete Architecture: Putting It All Together 30
Key Takeaways 30

Conclusion: Building the Next Generation of Enterprise Knowledge Systems **31**

References **32**

Building Production-Grade AI Applications in the .NET Ecosystem

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Introduction: The Enterprise RAG Challenge

In late 2023, a team at a mid-sized financial services company built a proof-of-concept chatbot that answered employee questions about HR policies. It worked beautifully in the demo. Within three weeks of rolling it out to two hundred users, the support desk was flooded with complaints. The system confidently fabricated policy details, cited documents that did not exist, and occasionally regurgitated sensitive salary data from an improperly filtered document index. The engineering team had built a RAG system that looked correct on a handful of test queries but collapsed under the weight of real-world usage.

This is the enterprise RAG challenge. The difference between a demo and a production system is not incremental. It is architectural, operational, and cultural. A demo answers ten queries with clean data. A production system answers ten thousand queries per hour against millions of documents, many of which are contradictory, outdated, or partially redacted, while staying within strict latency budgets, security boundaries, and cost ceilings.

The RAG pattern has become the de facto approach for grounding large language models in enterprise knowledge. The idea is elegant and simple: instead of asking a model to answer from its parametric training data, retrieve relevant documents from your own corpus, feed them as context, and let the model synthesize an answer. It avoids the cost and risk of fine-tuning on proprietary data. It keeps your knowledge base current without retraining. And it gives you citations, which is critical when a wrong answer has real consequences.

But the elegance of the pattern masks a mountain of engineering complexity beneath. Retrieval quality determines everything. A poor retriever feeds the model garbage context, and no amount of prompt engineering can compensate. The chunking strategy you choose at ingestion time locks in retrieval behavior for months or years. The vector database you select affects latency, cost, and operational complexity. The embedding model you pick trades off between accuracy, dimensionality, and inference speed. Hybrid

search, reranking, query transformation, and semantic caching each add layers of sophistication that are easy to sketch on a whiteboard but difficult to get right in production.

Security and governance are not afterthoughts. A RAG system is a data access layer with natural language as the query interface. If you do not enforce document-level access controls at retrieval time, any authenticated user can potentially retrieve and receive answers based on any document in your index, regardless of their authorization. PII and PHI must be redacted before ingestion and again before generation. Prompt injection attacks target the LLM through the retrieved context, bypassing your guardrails entirely. Audit trails must capture every query, retrieval, and generation step to support compliance reviews and incident forensics.

Observability is equally critical. When a RAG response is wrong, you need to know why. Was it a retrieval failure, where the right document was not found? A reranking problem, where relevant documents were ranked too low? A generation issue, where the model ignored or misinterpreted the retrieved context? Without structured tracing and metrics, debugging RAG systems is guesswork.

This book addresses all of these challenges head-on. It is written for .NET developers who are building or planning to build RAG applications in production. You do not need to be an AI researcher. You need to understand the concepts well enough to make informed architectural decisions, write correct C# code, and reason about trade-offs. The modern .NET ecosystem has converged on a set of abstractions that make enterprise RAG both possible and maintainable: Microsoft.Extensions.AI provides provider-agnostic interfaces for chat and embeddings. Microsoft.Extensions.VectorData offers a unified layer for vector store operations. Semantic Kernel and the newer Microsoft Agent Framework provide orchestration primitives, tool calling, and multi-agent workflows. Azure AI services, when you are on Azure, offer managed hybrid search, integrated vectorization, and built-in guardrails.

The book is structured to take you from foundations to mastery. The early chapters establish the conceptual groundwork: what RAG actually solves, how embeddings represent meaning, and the landscape of vector databases available to .NET developers. We then move into the core engineering challenges: building ingestion pipelines that handle messy enterprise documents, designing retrieval systems that combine dense and sparse search, and crafting prompts that produce grounded, citable responses.

The middle chapters cover the advanced patterns that separate prototypes from production systems. Agentic RAG introduces multi-agent workflows where agents plan, retrieve, verify, and act. Security and governance chapters address the threat models, PII filtering, access controls, and compliance requirements that enterprises cannot ignore. Evaluation and observability chapters give you concrete frameworks for measuring quality and debugging failures.

The final chapters focus on scaling, cost management, testing, and deployment. We cover semantic caching, async pipelines, model routing, Kubernetes deployment strategies, CI/CD for AI applications, and production runbooks. The book culminates in a fully deployable enterprise RAG solution that integrates everything we have discussed.

Every chapter includes complete, production-ready C# examples using the modern .NET ecosystem. We use Microsoft.Extensions.AI for provider-agnostic chat and embedding abstractions, Qdrant and PostgreSQL with pgvector as vector database options, Azure AI Search for managed hybrid search, and the Microsoft Agent Framework for agentic workflows. Code examples are tested against real libraries and follow current best practices as of 2026.

This is not a book about theory alone. It is a book about building things that work in production, where data is messy, users are impatient, budgets are finite, and compliance teams are watching. Let us get to work.

Chapter 1: Foundations of RAG and the .NET AI Ecosystem

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

What RAG Actually Solves (and What It Doesn't)

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

The Evolution from Prompt Engineering to Retrieval-Augmented Systems

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

The Modern .NET AI Stack: Microsoft.Extensions.AI, Semantic Kernel, and Agent Framework

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Choosing Your LLM Provider: Azure OpenAI, OpenAI API, and Local Models

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

A Quick-Start RAG App in .NET

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

From Prototype to Production: Refactoring into a Layered Architecture

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Key Takeaways

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Chapter 2: Embeddings and Vector Representations

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

How Embeddings Encode Meaning

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Choosing an Embedding Model: Accuracy vs. Speed vs. Cost

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Domain-Specific Embeddings and Fine-Tuning

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Dimensionality, Quantization, and Storage Efficiency

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Evaluating Embedding Quality

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Key Takeaways

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Chapter 3: Vector Databases for .NET Developers

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

The Landscape of Vector Stores in 2026

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Qdrant: Performance and Operations

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

PostgreSQL + pgvector: Consolidation and Simplicity

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Azure AI Search: Managed Hybrid Search

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Microsoft.Extensions.VectorData Abstraction Layer

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Making the Choice: A Decision Framework

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Key Takeaways

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Chapter 4: Building Robust Ingestion Pipelines

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

The Ingestion Pipeline Architecture

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Document Parsing and Layout-Aware Extraction

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Chunking Strategies: Fixed, Recursive, Semantic, and Structural

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Metadata Enrichment and Classification Tags

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Incremental Ingestion and Change Detection

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Key Takeaways

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Chapter 5: Hybrid Search and Retrieval Quality

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Why Dense Vector Search Is Not Enough

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Hybrid Search with RRF in Azure AI Search

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Cross-Encoder Reranking

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Query Transformation Techniques

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Designing a Production Retriever

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Key Takeaways

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Chapter 6: Prompt Engineering for Grounded Generation

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

The Anatomy of a RAG Prompt

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Context Assembly Strategies: Write, Select, Compress, Isolate

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Citation and Grounding Techniques

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Anti-Hallucination Guardrails in Prompts

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Prompt Versioning and Experimentation

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Key Takeaways

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Chapter 7: Agentic RAG with the Microsoft Agent Framework

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

From RAG to Agentic RAG

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Microsoft Agent Framework 1.0: Architecture and Primitives

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Sequential and Concurrent Workflows

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Agentic Retrieval Patterns

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Human-in-the-Loop and Approval Gates

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Key Takeaways

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Chapter 8: Security, Privacy, and Governance

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

The RAG Security Threat Model

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Document-Level Access Control at Retrieval Time

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

PII and PHI Redaction Pipelines

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Prompt Injection and Content Guardrails

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Audit Trails and Compliance

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Key Takeaways

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Chapter 9: Evaluation Frameworks for RAG Systems

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

What to Evaluate in a RAG System

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Retrieval Metrics: Precision, Recall, MRR, NDCG

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Generation Metrics: Faithfulness, Groundedness, Answer Relevance

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Microsoft.Extensions.AI.Evaluation in Practice

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Building Custom Evaluators and CI/CD Gates

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Key Takeaways

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Chapter 10: Observability and Debugging RAG Systems

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

The Observability Challenge in RAG

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

End-to-End Tracing with OpenTelemetry

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Metrics and Dashboards

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Root-Cause Analysis Playbook

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Debugging Tools and Techniques

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Key Takeaways

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Chapter 11: Scalability and Performance Optimization

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

The Latency Budget Problem

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Semantic Caching Strategies

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Async Pipelines and Concurrent Retrieval

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Model Routing and Tiered Inference

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Infrastructure Scaling Patterns

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Key Takeaways

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Chapter 12: Testing and Quality Assurance

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

What to Test in a RAG System

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Unit Testing Retrieval and Generation

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Integration Testing with Testcontainers

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Regression Evaluation and Golden Datasets

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

A/B Testing and Prompt Experiments

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Key Takeaways

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Chapter 13: Deployment, CI/CD, and Cost Management

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Containerizing a RAG Application

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

CI/CD Pipelines for AI Applications

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Kubernetes Deployment Patterns

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Production Runbooks and Incident Response

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Cost Management and Token Budgeting

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

The Complete Architecture: Putting It All Together

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Key Takeaways

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

Conclusion: Building the Next Generation of Enterprise Knowledge Systems

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.

References

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/enterpriseretrieval-augmentedgenerationwithc.