

BUILDING PRODUCTION-READY AI AGENTS WITH LLAMA INDEX

FROM FIRST QUERY TO MULTI-AGENT SYSTEMS:
A PRACTICAL GUIDE FOR PYTHON DEVELOPERS



RETRIEVE
WHAT MATTERS



ORCHESTRATE
COMPLEX WORKFLOWS



BUILD
MULTI-AGENT SYSTEMS



OBSERVE, EVALUATE,
AND IMPROVE



DESIGN FOR
RELIABILITY



THREE
END-TO-END
PROJECTS

INCREASING IN COMPLEXITY



PRACTICAL
PYTHON
EXAMPLES

THAT WORK IN PRODUCTION



DEPLOY AND SCALE
WITH CONFIDENCE

REAL-WORLD PATTERNS

KASEY SMITH

Building Production-Ready AI Agents with LlamaIndex

From First Query to Multi-Agent Systems: A Practical Guide for Python Developers

Steve T. Publications

This book is available at

<https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>

This version was published on 2026-07-10



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.

© 2026 Steve T. Publications

Contents

- From First Query to Multi-Agent Systems: A Practical Guide for Python Developers 1**
- Introduction: Why Agents, Why LlamaIndex, Why Now 2**
 - The Agent Moment 2
 - What Production-Ready Really Means 2
 - Why LlamaIndex Over Alternatives 2
 - How to Read This Book 2
- Chapter 1: LLM Fundamentals for Agent Builders 3**
 - How Large Language Models Generate Text 3
 - Tokens, Context Windows, and the Cost Equation 3
 - Prompting Patterns That Matter for Agents 3
 - Model Selection: Open Source vs Proprietary Trade-offs 3
 - Setting Up Your First LLM Connection with LlamaIndex 3
- Chapter 2: Inside LlamaIndex – Architecture and Core Concepts 5**
 - The LlamaIndex Abstraction Stack 5
 - Documents, Nodes, and the Ingestion Pipeline 5
 - Settings and Global Configuration 5
 - The Callback System and Extensibility Points 5
 - Project Setup: Installation, Virtual Environments, and Dependencies 5
- Chapter 3: Data Ingestion and Indexing 7**
 - Loading Data from Multiple Sources 7
 - Node Parsers and Chunking Strategies 7
 - Metadata Extraction and Enrichment 7
 - The Ingestion Pipeline API 7
 - Building Your First Index 7
- Chapter 4: Embeddings and Vector Stores 9**

CONTENTS

How Embeddings Represent Meaning	9
Choosing an Embedding Model	9
Vector Store Landscape and Selection Criteria	9
Implementing Your First Vector Store	9
Index Persistence and Versioning	9
Chapter 5: Retrieval Strategies for Production RAG	11
Top-K Retrieval and Its Limitations	11
Hybrid Search: Combining Semantic and Keyword	11
Reranking and Relevance Filtering	11
Metadata Filtering and Auto-Retrieval	11
Node Post-Processors for Fine-Tuning Results	11
Chapter 6: Building Your First RAG Application	13
Project Architecture and Design Decisions	13
Data Pipeline Implementation	13
Retrieval and Query Engine Setup	13
Adding Conversational Memory	13
Testing the Complete System	13
Chapter 7: Tools, Function Calling, and Structured Outputs	15
The Tool Abstraction in LlamaIndex	15
Defining Custom Tools with Type Signatures	15
Function Calling and Model Support	15
Structured Outputs with Pydantic Programs	15
Error Handling and Retry Logic for Tool Calls	15
Chapter 8: Agent Workflows and Orchestration	17
Event-Driven Architecture Fundamentals	17
Defining Steps and Events	17
State Management with Context Objects	17
Parallel Execution and Conditional Routing	17
Streaming Progress Events to Users	17
Chapter 9: Memory, Planning, and Reasoning	19
Short-Term vs Long-Term Memory Architectures	19
ChatMemoryBuffer and VectorMemoryBlock	19
Fact Extraction and Knowledge Accumulation	19
ReAct: Reasoning and Acting in Loops	19
Self-Reflection and Iterative Improvement	19

Chapter 10: Multi-Agent Systems 21

 Why Multi-Agent Architectures 21

 AgentWorkflow: Linear Swarm Pattern 21

 Orchestrator Agent: Sub-Agents as Tools 21

 Custom Planner: Maximum Flexibility 21

 Building a Multi-Agent Research Assistant (Project #2) 21

Chapter 11: Evaluation, Testing, and Debugging 23

 Defining What Good Looks Like 23

 Retrieval Evaluation Metrics 23

 Generation Quality Assessment 23

 Automated Testing Strategies 23

 Debugging Common Failure Modes 23

Chapter 12: Observability and Monitoring 25

 The Three Pillars of Observability 25

 OpenTelemetry Integration with LlamaIndex 25

 Tracing Agent Execution Paths 25

 Metrics, Dashboards, and Alerting 25

 Cost Tracking and Token Accounting 25

Chapter 13: Performance Optimization and Scalability 27

 Caching Strategies for Embeddings and LLM Calls 27

 Async Programming Patterns in Production 27

 Query Optimization Techniques 27

 Scaling Infrastructure: From Single Server to Distributed 27

 Cost Optimization Without Sacrificing Quality 27

Chapter 14: Security, Deployment, and Production Patterns 29

 Security Considerations for AI Agent Systems 29

 Building Production APIs with FastAPI 29

 Containerization with Docker 29

 Cloud Deployment Strategies 29

 Real-World Production Patterns and Anti-Patterns 29

Conclusion: The Road Ahead 31

 What We Built 31

 The Evolving Agent Landscape 31

 Principles That Will Endure 31

 Your Next Steps 31

References 32

From First Query to Multi-Agent Systems: A Practical Guide for Python Developers

This book takes you from foundational concepts through advanced multi-agent architectures, teaching you to design, build, deploy, and maintain sophisticated AI agents using LlamaIndex in real-world production environments. Through three end-to-end projects of increasing complexity, you will learn data ingestion and retrieval at depth, master structured workflows and orchestration patterns, implement robust observability and evaluation, and design systems that handle failure gracefully. Every concept is paired with working code, architectural trade-offs are surfaced honestly, and the emphasis throughout is on what actually works in production.

Introduction: Why Agents, Why LlamaIndex, Why Now

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

The Agent Moment

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

What Production-Ready Really Means

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Why LlamaIndex Over Alternatives

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

How to Read This Book

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Chapter 1: LLM Fundamentals for Agent Builders

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

How Large Language Models Generate Text

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Tokens, Context Windows, and the Cost Equation

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Prompting Patterns That Matter for Agents

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Model Selection: Open Source vs Proprietary Trade-offs

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Setting Up Your First LLM Connection with LlamaIndex

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Chapter 2: Inside LlamaIndex – Architecture and Core Concepts

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

The LlamaIndex Abstraction Stack

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Documents, Nodes, and the Ingestion Pipeline

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Settings and Global Configuration

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

The Callback System and Extensibility Points

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Project Setup: Installation, Virtual Environments, and Dependencies

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Chapter 3: Data Ingestion and Indexing

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Loading Data from Multiple Sources

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Node Parsers and Chunking Strategies

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Metadata Extraction and Enrichment

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

The Ingestion Pipeline API

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Building Your First Index

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Chapter 4: Embeddings and Vector Stores

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

How Embeddings Represent Meaning

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Choosing an Embedding Model

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Vector Store Landscape and Selection Criteria

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Implementing Your First Vector Store

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Index Persistence and Versioning

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Chapter 5: Retrieval Strategies for Production RAG

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Top-K Retrieval and Its Limitations

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Hybrid Search: Combining Semantic and Keyword

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Reranking and Relevance Filtering

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Metadata Filtering and Auto-Retrieval

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Node Post-Processors for Fine-Tuning Results

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Chapter 6: Building Your First RAG Application

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Project Architecture and Design Decisions

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Data Pipeline Implementation

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Retrieval and Query Engine Setup

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Adding Conversational Memory

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Testing the Complete System

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Chapter 7: Tools, Function Calling, and Structured Outputs

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

The Tool Abstraction in LlamaIndex

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Defining Custom Tools with Type Signatures

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Function Calling and Model Support

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Structured Outputs with Pydantic Programs

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Error Handling and Retry Logic for Tool Calls

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Chapter 8: Agent Workflows and Orchestration

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Event-Driven Architecture Fundamentals

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Defining Steps and Events

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

State Management with Context Objects

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Parallel Execution and Conditional Routing

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Streaming Progress Events to Users

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Chapter 9: Memory, Planning, and Reasoning

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Short-Term vs Long-Term Memory Architectures

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

ChatMemoryBuffer and VectorMemoryBlock

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Fact Extraction and Knowledge Accumulation

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

ReAct: Reasoning and Acting in Loops

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Self-Reflection and Iterative Improvement

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Chapter 10: Multi-Agent Systems

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Why Multi-Agent Architectures

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

AgentWorkflow: Linear Swarm Pattern

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Orchestrator Agent: Sub-Agents as Tools

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Custom Planner: Maximum Flexibility

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Building a Multi-Agent Research Assistant (Project #2)

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Chapter 11: Evaluation, Testing, and Debugging

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Defining What Good Looks Like

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Retrieval Evaluation Metrics

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Generation Quality Assessment

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Automated Testing Strategies

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Debugging Common Failure Modes

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Chapter 12: Observability and Monitoring

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

The Three Pillars of Observability

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

OpenTelemetry Integration with LlamaIndex

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Tracing Agent Execution Paths

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Metrics, Dashboards, and Alerting

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Cost Tracking and Token Accounting

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Chapter 13: Performance Optimization and Scalability

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Caching Strategies for Embeddings and LLM Calls

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Async Programming Patterns in Production

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Query Optimization Techniques

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Scaling Infrastructure: From Single Server to Distributed

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Cost Optimization Without Sacrificing Quality

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Chapter 14: Security, Deployment, and Production Patterns

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Security Considerations for AI Agent Systems

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Building Production APIs with FastAPI

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Containerization with Docker

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Cloud Deployment Strategies

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Real-World Production Patterns and Anti-Patterns

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Conclusion: The Road Ahead

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

What We Built

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

The Evolving Agent Landscape

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Principles That Will Endure

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

Your Next Steps

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.

References

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/buildingproduction-readyaiagentswithllamaindex>.