Training Course on Retrieval-Augmented Generation (RAG) Systems for LLMs

Professional Development Training Course Outline

Category	Data Science	**Duration**	5 Days
Base Fee	$1,500 USD	**Delivery Mode**	Online / On-site Hybrid

Course Introduction

Training Course on Retrieval-Augmented Generation (RAG) Systems for LLMs dives deep into Retrieval-Augmented Generation (RAG) systems, a cutting-edge paradigm revolutionizing Large Language Models (LLMs). Participants will master the art of building **robust, knowledge-augmented LLM applications** that overcome the inherent limitations of static pre-trained models. From **foundational concepts** to **advanced RAG techniques**, this course equips professionals with the **practical skills** and **trending methodologies** to deploy intelligent, **factually grounded AI solutions** across diverse industries. We will explore **vector databases, prompt engineering, data orchestration, and evaluation metrics** to ensure optimal performance and mitigate **hallucinations** in real-world scenarios.

The program emphasizes a **hands-on, project-based learning approach**, enabling participants to develop and deploy their own RAG systems. We will navigate the complexities of **data preprocessing, embedding generation, efficient retrieval mechanisms, and intelligent response synthesis**. This course is designed to empower **AI engineers, data scientists, and developers** to unlock the full potential of LLMs by integrating **dynamic external knowledge**, leading to **enhanced accuracy, reduced costs, and accelerated innovation** in the rapidly evolving landscape of **generative AI**.

Programme Curriculum

Training Course on Retrieval-Augmented Generation (RAG) Systems for LLMs

Introduction

Training Course on Retrieval-Augmented Generation (RAG) Systems for LLMs dives deep into Retrieval-Augmented Generation (RAG) systems, a cutting-edge paradigm revolutionizing Large Language Models (LLMs). Participants will master the art of building **robust, knowledge-**

augmented LLM applications that overcome the inherent limitations of static pre-trained models. From **foundational concepts** to **advanced RAG techniques**, this course equips professionals with the **practical skills** and **trending methodologies** to deploy intelligent, **factually grounded AI solutions** across diverse industries. We will explore **vector databases, prompt engineering, data orchestration, and evaluation metrics** to ensure optimal performance and mitigate **hallucinations** in real-world scenarios.

The program emphasizes a **hands-on, project-based learning approach**, enabling participants to develop and deploy their own RAG systems. We will navigate the complexities of **data preprocessing, embedding generation, efficient retrieval mechanisms, and intelligent response synthesis**. This course is designed to empower **AI engineers, data scientists, and developers** to unlock the full potential of LLMs by integrating **dynamic external knowledge**, leading to **enhanced accuracy, reduced costs, and accelerated innovation** in the rapidly evolving landscape of **generative AI**.

Course Duration

5 days

Course Objectives

1. Comprehend the core principles, architecture, and workflow of Retrieval-Augmented Generation (RAG) systems.
2. Develop and deploy both basic and sophisticated RAG pipelines for diverse use cases.
3. Learn robust techniques for **document extraction, cleaning, normalization, chunking, and metadata analysis** to prepare unstructured data.
4. Gain expertise in utilizing **vector databases (e.g., Pinecone, Weaviate, Chroma)** for efficient semantic search and information retrieval.
5. Design and refine **prompt templates** to guide LLMs in generating accurate and contextually relevant responses.
6. Explore and implement **query expansion, re-ranking (cross-encoders, bi-encoders), and dense passage retrieval (DPR)**.
7. Understand and apply methods for building and integrating **knowledge graphs (e.g., Neo4j)** to improve RAG system accuracy and contextuality.
8. Develop strategies and implement guardrails to significantly reduce factual errors and "hallucinations" in LLM outputs.
9. Learn to measure and assess RAG effectiveness using key metrics like **groundedness, relevance, coherence, and answer accuracy**.
10. Acquire skills in designing, optimizing, and deploying scalable RAG systems for real-world enterprise solutions.
11. Implement best practices for managing, monitoring, and maintaining RAG systems in production environments.
12. Delve into designing intelligent agents that leverage RAG for multi-step reasoning and complex task completion.
13. Gain insights into future trends, including the integration of multimodal data (images, audio, video) into RAG systems.

Organizational Benefits

- Significantly reduce LLM hallucinations by grounding responses in verified, up-to-date knowledge bases.

- Enable quicker and more precise access to internal company data, policies, and product information.
- Minimize the need for expensive and time-consuming LLM retraining by dynamically updating external knowledge sources.
- Provide verifiable sources for LLM-generated content, fostering user confidence and accountability.
- Empower teams to rapidly build and deploy sophisticated, domain-specific AI solutions.
- Leverage cutting-edge AI techniques to create differentiated products and services.
- Design systems that can adapt to evolving information needs and handle increasing data volumes.
- Gain greater control over the data used by LLMs, ensuring compliance with internal policies and regulations.

Target Audience

1. AI Engineers.
2. Data Scientists
3. Machine Learning Engineers
4. Software Developers
5. NLP Practitioners
6. Product Managers (AI/ML focus)
7. Solutions Architects
8. Researchers in AI/ML

Course Outline

Module 1: Introduction to RAG and LLM Fundamentals

- Understanding the LLM Landscape: Overview of Large Language Models, their capabilities, and inherent limitations (e.g., knowledge cutoff, hallucinations).
- The Rise of RAG: Definition, motivation, and the core concept of augmenting LLMs with external knowledge.
- RAG Architecture Overview: Components of a RAG system: Retriever, Generator, and Knowledge Base.
- Setting Up Your RAG Development Environment: Essential tools, libraries (Python, Hugging Face, etc.), and framework choices (LangChain, LlamaIndex).
- **Case Study: Customer Support Chatbot (Initial)** - Demonstrating how a naive LLM struggles with specific, un-trained knowledge.

Module 2: Data Preparation and Knowledge Base Creation

- Sourcing and Ingesting Diverse Data: Techniques for collecting data from various sources (documents, databases, APIs, web).
- Document Processing and Chunking Strategies: Best practices for breaking down large texts into manageable chunks for effective retrieval.
- Metadata Extraction and Enrichment: Utilizing metadata to improve search relevance and contextual understanding.
- Embedding Models and Vectorization: Deep dive into different embedding models (e.g., Sentence Transformers, OpenAI Embeddings) and generating vector representations.
- **Case Study: Enterprise Document Q&A** - Preparing a corpus of internal company policy documents for a RAG system.

Module 3: Deep Dive into Vector Databases

- Introduction to Vector Databases: Understanding their role in storing and querying high-dimensional vectors.
- Popular Vector Database Options: Exploring key features and trade-offs of Pinecone, Weaviate, Chroma, FAISS, etc.
- Indexing Strategies: Efficiently adding and updating document embeddings in the vector store.
- Similarity Search Algorithms: Techniques for finding relevant documents based on semantic similarity (e.g., Cosine Similarity, ANN).
- **Case Study: Medical Research Assistant** - Storing and retrieving medical journal articles for relevant insights.

Module 4: Building the Retriever Component

- Basic Retrieval Techniques: Implementing keyword-based search and simple vector similarity search.
- Advanced Retrieval Strategies: Exploring query expansion, re-ranking with cross-encoders and bi-encoders.
- Hybrid Search Approaches: Combining lexical and semantic search for improved recall and precision.
- Optimizing Retriever Performance: Strategies for reducing latency and improving retrieval accuracy.
- **Case Study: Legal Document Analysis** - Retrieving relevant legal precedents from a large legal database.

Module 5: The Generative Component and Prompt Engineering for RAG

- Integrating LLMs with Retrieved Context: How to effectively combine retrieved information with the user query.
- Designing Effective Prompt Templates: Techniques for structuring prompts to guide LLMs and leverage context.
- In-context Learning and Few-shot Prompting: Maximizing LLM performance with relevant examples from the retrieved data.
- Controlling LLM Output: Techniques for steering responses towards desired formats and factual adherence.
- **Case Study: Technical Support AI Agent** - Generating accurate and helpful responses based on retrieved product manuals and troubleshooting guides.

Module 6: Advanced RAG Techniques and Architectures

- Multi-hop Retrieval: Handling complex queries requiring information from multiple retrieved documents.
- Agentic RAG Workflows: Building AI agents that can reason, plan, and execute multi-step tasks using RAG.
- Knowledge Graphs Integration: Using knowledge graphs to provide structured context and enhance reasoning.
- Iterative Retrieval and Generation: Refining responses through multiple retrieval and generation steps.
- **Case Study: Financial Analyst Assistant** - Answering complex financial questions by combining data from reports and market news.

Module 7: Evaluation, Optimization, and MLOps for RAG Systems

- Metrics for RAG Evaluation: Quantifying performance (retrieval quality, groundedness, answer relevance, latency).
- Debugging and Troubleshooting RAG Systems: Identifying and resolving common issues in RAG pipelines.
- Fine-tuning Components (Optional/Advanced): Briefly covering fine-tuning retrievers or generators for specific domains.
- Deployment Strategies for RAG: Best practices for deploying RAG systems to production environments (cloud, on-premise).
- **Case Study: E-commerce Product Recommender** - Optimizing RAG for personalized product recommendations and handling evolving product catalogs.

Module 8: Ethical Considerations, Future Trends, and Real-World Applications

- Bias and Fairness in RAG: Addressing potential biases in retrieved data and LLM outputs.
- Security and Privacy Concerns: Protecting sensitive information within RAG systems.
- Multimodal RAG: Exploring the integration of images, audio, and video into RAG workflows.
- Emerging Trends in Generative AI: Discussing the future of RAG, including self-correcting RAG and real-time knowledge integration.
- **Case Study: Personalized Learning Platform** - Delivering tailored educational content and answering student queries using RAG.

Training Methodology

This course employs a participatory and hands-on approach to ensure practical learning, including:

- Interactive lectures and presentations.
- Group discussions and brainstorming sessions.
- Hands-on exercises using real-world datasets.
- Role-playing and scenario-based simulations.
- Analysis of case studies to bridge theory and practice.
- Peer-to-peer learning and networking.
- Expert-led Q&A sessions.
- Continuous feedback and personalized guidance.

Register as a group from 3 participants for a Discount

Send us an email: info@fineskilltrainingcenter.org or call *+254769199797*

Certification

Upon successful completion of this training, participants will be issued with a globally- recognized certificate.

Tailor-Made Course

We also offer tailor-made courses based on your needs.

Key Notes

a. The participant must be conversant with English.

b. Upon completion of training the participant will be issued with an Authorized Training Certificate

c. Course duration is flexible and the contents can be modified to fit any number of days.

d. The course fee includes facilitation training materials, 2 coffee breaks, buffet lunch and A Certificate upon successful completion of Training.

e. One-year post-training support Consultation and Coaching provided after the course.

f. Payment should be done at least a week before commence of the training, to Fineskill Training Center account, as indicated in the invoice so as to enable us prepare better for you.

Upcoming Scheduled Sessions

Start Date	End Date	Location	Price
23 Nov 2026	27 Nov 2026	Online	$1,000
23 Nov 2026	27 Nov 2026	Physical	$1,500
23 Nov 2026	27 Nov 2026	Physical	$3,000
23 Nov 2026	27 Nov 2026	Physical	$3,000
23 Nov 2026	27 Nov 2026	Physical	$6,000
23 Nov 2026	27 Nov 2026	Physical	$4,500
23 Nov 2026	27 Nov 2026	Physical	$6,000
23 Nov 2026	27 Nov 2026	Physical	$6,000

Ready to enroll or request a customized program? Book online or contact us:

Register Online Now