RAG & Vector Search for LLM Apps

2027 □□□□ 21 – 17
Paris (France)



RAG & Vector Search for LLM Apps

Ref: 20220746_217023 Location: Paris (France) Fees: 5900 GBP

Why This Course

The rise of Large Language Models LLMs such as GPT and Claude has transformed how businesses use AI.

However, to make LLMs truly useful in enterprise settings, organizations must integrate them with Retrieval-Augmented Generation RAG and vector search technologies.

This program provides participants with the knowledge and practical skills to design, build, and deploy RAG-powered applications.

By mastering vector databases, embeddings, and retrieval pipelines, participants will learn how to make LLMs smarter, more accurate, and context-aware for real-world use cases such as chatbots, knowledge assistants, and document intelligence systems.

What You'll Learn and Practice

By joining this program, you will:

- Understand the architecture and principles of Retrieval-Augmented Generation RAG.
- Learn how embeddings and vector databases enhance LLM contextual reasoning.
- Gain hands-on experience in building RAG pipelines with real-world data.
- Explore leading vector search tools and frameworks FAISS, Pinecone, Weaviate, Chroma.
- Develop strategies for deploying and scaling RAG-based AI applications securely and efficiently.

The Program Flow

Day 1: Foundations of LLMs and Retrieval-Augmented Generation

- Overview of LLMs: capabilities, limitations, and architecture.
- Why retrieval is essential: addressing hallucinations and data freshness.
- Core RAG concepts: retriever, generator, and data orchestration.
- Workflow: how RAG integrates with prompt engineering and context injection.
- Case study: enterprise knowledge assistant powered by RAG.



Day 2: Understanding Embeddings and Vector Representations

- The concept of embeddings — turning text into numerical vectors.
- Types of embeddings: semantic, sentence, and domain-specific.
- Measuring similarity — cosine similarity, Euclidean, and dot product.
- Practical exercise: generating and comparing embeddings using open-source tools.
- Workshop: embedding a corporate document corpus for retrieval.

Day 3: Vector Databases and Search Architectures

- Introduction to vector databases: FAISS, Pinecone, Weaviate, and Chroma.
- Indexing, searching, and ranking in high-dimensional vector space.
- Combining metadata filters with vector similarity search.
- Building hybrid search systems keyword + vector retrieval.
- Simulation: implementing a document retrieval system with a vector database.

Day 4: Building and Deploying RAG Applications

- Constructing an end-to-end RAG pipeline: ingestion, retrieval, and generation.
- Integrating vector search with LLM APIs OpenAI, Anthropic, Hugging Face.
- Context window management and response optimization.
- Evaluation metrics for RAG performance — accuracy, relevance, and latency.
- Hands-on project: building a domain-specific chatbot with RAG and vector search.

Day 5: Scaling, Security, and Real-World Implementation

- Scaling RAG architectures for enterprise use.
- Data security, privacy, and compliance in RAG systems.
- Improving reliability with caching, reranking, and hybrid retrieval.
- Future trends: multimodal retrieval, agentic RAG, and fine-tuning with feedback loops.
- Action workshop: designing a RAG deployment plan for your organization.

Individual Impact

- Gain technical mastery of RAG architecture and vector search fundamentals.
- Strengthen the ability to design and deploy LLM-based intelligent applications.
- Develop problem-solving skills for context-aware AI retrieval challenges.
- Build practical experience integrating data and AI in real business workflows.
- Position yourself at the forefront of next-generation AI application development.



Work Impact

- Enhance AI application accuracy, reliability, and compliance.
- Leverage existing data assets securely through retrieval-augmented systems.
- Improve decision-making and efficiency with intelligent, context-aware tools.
- Reduce costs by optimizing LLM usage and minimizing unnecessary queries.
- Drive digital innovation through scalable, enterprise-ready AI solutions.

Training Methodology

This program blends hands-on coding, real-world examples, and strategic insights to deliver both technical depth and business relevance.

Learning methods include:

- Practical labs using Python and popular vector database frameworks.
- Step-by-step implementation of a RAG pipeline.
- Group discussions and design challenges.
- Evaluation exercises for optimizing retrieval accuracy.
- Access to reusable templates, sample datasets, and architecture blueprints.

Beyond the Course

Upon completion, participants will be equipped to design, implement, and deploy Retrieval-Augmented Generation systems using vector search and embeddings.

They will leave ready to build smarter, scalable, and secure AI applications — turning data into intelligent, context-driven insights powered by LLMs.

