Advanced RAG & Long-Context Private Models

- 3 Days
- Lecture and Hands-on Labs

Course Overview

The ultimate 3-day mastery class that turns programmers into owners of production-grade, long-context RAG systems powered by private models. Using your own fine-tuned LLM (from AI-301) and the latest 2025 retrieval techniques, you will build, debug, and deploy RAG pipelines that reason accurately over entire codebases, 200-page PDFs, or proprietary knowledge bases, all tracked and versioned in ClearML. Every student leaves with a fully functional, high performance private-data RAG endpoint.

Who Should Attend

- Programmers, ML engineers, and solution architects
- Teams moving from basic chatbots to company-brain-scale AI

What You'll Learn

- 30+ tracked ClearML RAG experiments
- Hybrid retrieval, re-ranking, and query rewriting
- Long-context (128k–1M) private model mastery
- Production RAG that survives real enterprise data

Outline

RAG Foundations & Data Engineering

- Lecture: Beyond Naive RAG: The Production Architecture
- Lecture: Strategies for "Messy" Enterprise Data (PDFs, Tables, Code)
- Lecture + Lab: Lab: Advanced Ingestion & Semantic Chunking
- Lecture + Lab: Lab: Host Your Own Embedding Model (BGE-M3 / E5)
- Lecture + Lab: Lab: Versioning Vector Indexes in ClearML

Advanced Retrieval Strategies

- Lecture: Hybrid Search: Combining Dense Vectors & BM25
- Lecture: Query Transformations: HyDE, Rewrite-Retrieve-Read
- Lecture + Lab: Lab: Implementing Hybrid Search (ChromaDB / Qdrant)

- Lecture + Lab: Lab: Building a Query Rewriter Pipeline
- Lecture + Lab: Lab: Visualizing Retrieval Quality in ClearML

Re-Ranking & Precision Filtering

- Lecture: The Cross-Encoder Advantage: Re-Ranking for Relevance
- Lecture: Metadata Filtering & Self-Querying Retrievers
- Lecture + Lab: Lab: Deploying a Private Re-Ranker Service
- Lecture + Lab: Lab: Implementing a Two-Stage Retrieval Pipeline
- Lecture + Lab: Lab: Comparing Reranked vs. Raw Results

Mastering Long-Context Models (128k - 1M)

- Lecture: The Long Context Era: Architecture & KV Caches
- Lecture: "Lost in the Middle" Phenomenon & Mitigation
- Lecture + Lab: Lab: Loading Long-Context Private Models (Yi / Llama-3)
- Lecture + Lab: Lab: Needles in a Haystack: Testing Retrieval Limits
- Lecture + Lab: Lab: Prompt Compression Techniques

Generation & Grounding

- Lecture: Grounding the LLM: Citations & Source Tracking
- Lecture: Response Synthesis Strategies (Refine vs. Map-Reduce)
- Lecture + Lab: Lab: Forcing Citation Formats via Grammars
- Lecture + Lab: Lab: Handling "I Don't Know" Scenarios Gracefully

Evaluation & Observability (RAGOps)

- Lecture: RAG Evaluation Metrics: Faithfulness & Answer Relevancy
- Lecture: Automated Evaluation Pipelines (RAGAS / Aries)
- Lecture + Lab: Lab: Setting up RAGAS with a Local Judge LLM
- Lecture + Lab: Lab: Running a Full RAG Evaluation Suite in ClearML
- Lecture + Lab: Lab: Analyzing Failure Modes (Retrieval vs. Generation Errors)

Capstone: Enterprise Knowledge Base

- Lecture: Capstone Brief: The "Company Brain" Challenge
- Lecture + Lab: Lab: Ingest a Multi-Modal Dataset (Docs + Code)

- Lecture + Lab: Lab: Build the Hybrid Retrieval & Reranking Chain
- Lecture + Lab: Lab: Optimize for Latency & Throughput
- Lecture + Lab: Lab: Deploy the Final RAG Endpoint

Continuing Education

- Lecture + Lab: Continuing Education (Agentic RAG)
- Lecture: Curriculum Path: Fine-Tuning Your Embeddings

Prerequisites

- Completion of AI-301 (Fine-Tuned LLM with ClearML)
- Working knowledge of vector databases