MICROSOFT TRAINING

Building Rag Agents Llms Nvidia Course

Learn to build RAG agents and develop LLM solutions on Nvidia platforms. Master advanced AI for real-world applications.

DURATION
16 hours

LEVEL
Intermediate

FORMAT
Instructor-Led

- Globally Recognized Certification
- Expert-Led Hands-On Training
- Flexible Scheduling

Computer Pride | 1st Floor, ICEA Building, Kenyatta Avenue, Nairobi | www.computer-pride.co.ke

COURSE OVERVIEW

This intermediate-level "Building RAG Agents LLMs NVIDIA Course" from Microsoft is meticulously designed to equip AI and machine learning professionals with the expertise to construct sophisticated Retrieval Augmented Generation (RAG) agents leveraging the power of Large Language Models (LLMs) and optimized NVIDIA infrastructure. Participants will delve into the fundamental principles of RAG, understanding how to enhance LLM capabilities by integrating external knowledge sources, thereby mitigating hallucinations and improving response accuracy and relevance. The course emphasizes practical application, guiding learners through the entire lifecycle of RAG agent development, from data preparation and vector database integration to prompt engineering and agent orchestration. A core focus of this 16-hour intensive course is the hands-on utilization of NVIDIA's cutting-edge platforms and tools for accelerating LLM inference and RAG pipeline optimization. You will learn to harness NVIDIA's robust ecosystem, including potentially leveraging CUDA for GPU acceleration, NVIDIA TensorRT for model deployment, or NVIDIA NeMo for LLM fine-tuning and deployment, to build highly efficient and scalable RAG solutions. By mastering these advanced techniques, you will be able to develop intelligent agents capable of performing complex tasks, accessing real-time information, and delivering contextually rich responses, positioning you at the forefront of generative AI innovation.

Duration	16 hours
Level	Intermediate
Vendor	Microsoft

COURSE CURRICULUM

01 Module 1: Introduction to LLMs & Retrieval Augmented Generation (RAG)

- Overview of Large Language Models (LLMs) and their limitations
- Understanding the RAG paradigm: Retrieval vs. Generation
- Architecture of RAG systems: components and workflow
- Key use cases and benefits of RAG agents in enterprise applications

02 Module 2: Building the Retrieval Component

- Data ingestion, cleaning, and chunking strategies for RAG
- Text embedding models and techniques (e.g., Sentence Transformers, OpenAI embeddings)
- Vector databases: concepts, selection criteria, and practical implementation (e.g., Pinecone, ChromaDB, Weaviate)
- Advanced retrieval strategies: hybrid search, re-ranking, and contextual retrieval

03 Module 3: Integrating LLMs and Prompt Engineering for Generation

- Interacting with diverse LLMs (e.g., OpenAI, Hugging Face models) via APIs and local deployments
- Effective prompt engineering techniques specifically for RAG systems
- Managing conversation history and multi-turn interactions
- Evaluating RAG system output: relevance, coherence, and factual accuracy

04 Module 4: Leveraging NVIDIA for RAG Agent Acceleration

- Introduction to NVIDIA's AI ecosystem for LLMs (e.g., CUDA, Triton Inference Server, TensorRT-LLM, NeMo Framework)
- Optimizing embedding generation and vector search with NVIDIA GPUs
- Accelerating LLM inference for the generation component using NVIDIA tools
- Strategies for deploying RAG components on NVIDIA accelerated infrastructure

05 Module 5: Advanced RAG Agent Architectures & Deployment

- Building multi-agent systems and orchestrating tool usage with LLMs
- Utilizing agent orchestration frameworks (e.g., LangChain, LlamaIndex)
- Monitoring, logging, and maintaining RAG agents in production environments
- Security, privacy, and ethical considerations for RAG agent deployment

06 Module 6: Project: Developing a Custom RAG-Powered AI Agent

- Defining a real-world problem and agent requirements specification
- Designing and implementing a comprehensive RAG agent solution from scratch
- Rigorous testing, evaluation, and iterative refinement of the agent
- Presenting the developed project and discussing future enhancements and scaling strategies

CAREER PATHWAYS

Graduates of this program are prepared for the following roles:

> **AI/ML Engineer (Generative AI & RAG Specialist)**

> **Conversational AI Architect**

> **NLP Research Scientist (Agentic Systems & Knowledge Retrieval)**

FREQUENTLY ASKED QUESTIONS

Q: What foundational knowledge or technical background is recommended before enrolling in this "Building RAG Agents LLMs NVIDIA Course"?

This intermediate-level course assumes a solid understanding of Python programming, including familiarity with common data science libraries like NumPy and Pandas. Prior experience with machine learning concepts, natural language processing (NLP) fundamentals, and basic exposure to deep learning frameworks (e.g., PyTorch or TensorFlow) would be highly beneficial, though not strictly required. A conceptual understanding of how LLMs work will also aid your learning.

Q: How does this course specifically leverage NVIDIA technologies, and what tangible benefits will I gain from this integration for RAG agent development?

This course is unique in its focus on integrating NVIDIA's powerful hardware and software stack to significantly enhance RAG agent performance. You will gain hands-on experience optimizing LLM inference and embedding generation using NVIDIA GPUs, potentially leveraging tools like NVIDIA TensorRT-LLM for faster deployment and the NeMo Framework for efficient model adaptation. This specialized knowledge will enable you to build and deploy RAG agents that are not only accurate but also highly performant and scalable, crucial for enterprise-grade generative AI applications.

Q: What types of practical RAG agent applications will I be capable of building upon completing this certification, and how will these skills differentiate me in the job market?

Upon completion, you will possess the practical skills to develop advanced RAG agents capable of tasks such as sophisticated, knowledge-grounded question-answering systems for specific domains (e.g., legal, medical, technical documentation), intelligent chatbots that retrieve real-time information, and personalized content generation tools that minimize hallucinations. These capabilities are highly sought after, differentiating you as a specialist in building robust, knowledge-aware generative AI solutions, making you a valuable asset in the rapidly evolving AI landscape.

ABOUT COMPUTER PRIDE

East Africa's Premier IT Training Center

With over 15 years of excellence, Computer Pride delivers globally recognized certification programs from leading technology vendors. Located on Kenyatta Avenue in Nairobi's CBD, our state-of-the-art facilities and expert instructors have empowered thousands of professionals across East Africa.

15+	10,000+	50+
Years of Excellence	Professionals Trained	Vendor Certifications

READY TO ENROL?

Take the next step in your technology career.

Location
1st Flr, ICEA Building
Kenyatta Ave, Nairobi

Email
info@computer-pride.co.ke

Phone
+254 705 900 900

www.computer-pride.co.ke