MICROSOFT TRAINING

Generative Ai Rag Systems Nvidia Course

Learn to build advanced Generative AI RAG systems using Nvidia tools. Master cutting-edge techniques for robust AI development.

DURATION
2 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 course dives into the rapidly evolving world of Generative AI, with a specific focus on Retrieval-Augmented Generation (RAG) systems. Participants will gain a comprehensive understanding of how RAG architectures overcome common limitations of standalone Large Language Models (LLMs), such as factual inaccuracies and outdated information. The course highlights the critical role RAG plays in building reliable, context-aware, and enterprise-ready generative AI applications by integrating real-time or proprietary data sources, thereby significantly reducing hallucinations and improving output relevance.

Duration	2 hours
Level	Intermediate
Vendor	Microsoft

COURSE CURRICULUM

01 Module 1: Foundations of Generative AI and the RAG Paradigm

- Overview of Generative AI and its transformative potential
- Limitations of standalone Large Language Models (LLMs): Hallucinations and knowledge cutoffs
- Introduction to Retrieval-Augmented Generation (RAG): Concept, architecture, and benefits
- Why RAG is essential for building accurate and enterprise-grade LLM applications

02 Module 2: Core Components of a RAG System

- The 'Retrieval' aspect: Data indexing, chunking, and vector databases
- Generating and utilizing vector embeddings for semantic search
- The 'Generation' aspect: Integrating LLMs for context-aware response synthesis
- Understanding the data flow and overall system architecture within RAG

03 Module 3: Leveraging NVIDIA for RAG System Acceleration

- Introduction to NVIDIA's role in accelerating AI and LLM inference
- Conceptual understanding of NVIDIA GPU-accelerated computing for vector embeddings and LLMs
- Principles of optimizing RAG system performance using NVIDIA hardware capabilities
- Overview of relevant NVIDIA tools and platforms that enhance RAG deployments (e.g., CUDA, Triton Inference Server concepts)

04 Module 4: Implementing and Evaluating RAG Systems

- High-level workflow for conceptualizing and building a basic RAG application
- Key considerations for data preparation, source integration, and prompt engineering in RAG
- Introduction to metrics and strategies for evaluating RAG system performance and accuracy
- Discussion on common challenges and best practices for RAG deployment

CAREER PATHWAYS

Graduates of this program are prepared for the following roles:

> **AI Solutions Architect specializing in LLM and RAG deployments**

> **Machine Learning Engineer focused on Generative AI MLOps**

> **Prompt Engineer and LLM Application Developer**

FREQUENTLY ASKED QUESTIONS

Q: What specific NVIDIA technologies or frameworks will I gain practical experience with in this 2-hour course?

While this 2-hour course focuses on the conceptual and architectural understanding of Generative AI RAG systems, it will highlight how NVIDIA's GPU-accelerated computing and associated software platforms (like CUDA and Triton Inference Server) are leveraged for optimizing performance in real-world RAG deployments. You will learn the *principles* of how NVIDIA technologies contribute to efficiency rather than hands-on coding with specific NVIDIA SDKs.

Q: How does mastering RAG systems with an NVIDIA focus enhance my ability to deploy accurate and reliable AI solutions in an enterprise setting?

By understanding RAG systems, you gain the crucial ability to build Generative AI applications that are factually grounded, leverage proprietary enterprise data, and reduce LLM hallucinations. The NVIDIA focus ensures you comprehend how high-performance computing enables the scalability and efficiency required for these robust, production-grade AI solutions, making your deployments more reliable and effective for business needs.

Q: What are the prerequisite skills or knowledge recommended to get the most out of this Intermediate-level Generative AI RAG Systems course?

To maximize your learning from this course, a foundational understanding of core Artificial Intelligence (AI) and Machine Learning (ML) concepts is highly recommended. Familiarity with the basic principles of Large Language Models (LLMs) and general programming concepts, ideally in Python, would also be beneficial, although no deep coding experience is required for the course content itself.

ABOUT COMPUTER PRIDE

computerpride

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	Email	Phone
1st Flr, ICEA Building Kenyatta Ave, Nairobi	info@computer-pride.co.ke	+254 705 900 900

www.computer-pride.co.ke