Welcome to AI for Telco Part 3: Retrieval Augmented Generation (RAG)

ANI_110d | On-Demand | Automation and Insights |
Course Duration: 1 hour

This course explores Retrieval Augmented Generation (RAG) at a very high level. We cover the what, why and how of RAG along with key applications of RAG. This training also provides the benefits and limitations of RAG to provide a balanced view. The discussion on RAG includes its architecture, specific database considerations, and operations. An overview of the operations of RAG includes tokenization and effective search criteria. The webinar will conclude with a discussion of the practical applications of RAG with LLMs, including web grounding and advanced techniques like LangChain and prompt chaining. This course explores telco use cases specifically focused on network engineering and operations.

Intended Audience

This is an introductory course on RAG and is suitable for beginners to this area.

Objectives

After completing this course, the learner will be able to:

- Define Retrieval Augmented Generation (RAG)
- List RAG benefits and limitations
- Identify scenarios where RAG brings value in Telecom networks
- Sketch RAG architecture and process
- List considerations for RAG

Outline

1. Customizing a Large Language Model (LLM)
1.1 Review of GenAI and LLM
1.2 How to augment an LLM with new data
1.3 Augmenting a LLM with RAG
1.4 Refining a LLM with Fine Tuning

2. Introduction to RAG
2.1 Why do we need RAG?
2.2 RAG architecture
2.3 RAG considerations
2.4 RAG applications within Telco

3. RAG Operations
3.1 Vectordb and tokenization
3.2 Importance of search criteria
3.3 Using RAG with LLM
3.4 Web Grounding
3.5 LangChain and Prompt Chaining

4. Conclusion
4.1 Summary