S Oracle Cloud Infrastructure Generative AI Professional LVC

Oracle Cloud Infrastructure

DURATION	MODULES	COURSE CODE
1 Days	7 Lectures	—

Course Overview

Currently, the OCI Generative AI service is hosted in the following OCI regions: US Midwest (Chicago) ,Germany Central (Frankfurt) and UK South (London).Please make sure to subscribe to one of these regions to access the OCI Generative AI service

What You Will Learn

Module 1: Course Introduction

- For Whom This Course is Intended
- Course Outline
- Fundamentals of Large Language Models (LLMs)
- Dive Deep on OCI Generative AI Service
- Build an LLM App using OCI Generative AI Service
- Meet Your Instructors
- Measuring Your Progress: Skill Checks
- Getting Answers: "Ask Your Instructor" Form & OU Community
- Best Practices and Retention Tips
- Keep Progressing: Path to Success

Module 2: Introduction to Large Language Models

- What is a Large Language Model?
- Module Overview
- LLM Architectures
- Encoders
- Decoders
- Encoder-Decoder Architectures
- Model Ontology
- Architectures at a Glance

Module 3: Prompting and Prompt Engineering

- Affecting Distribution over Vocabulary
- Prompting and Prompt Engineering
- In-Context Learning and Few-Shot Prompting
- Example Prompts & Advanced Prompting Strategies
- Issues with Prompting:
- Prompt Injection
- Memorization
- Training LLMs
- Hardware Costs
- Decoding Techniques:
- Greedy and Non-Deterministic Decoding
- Temperature
- Hallucination: Groundedness and Attributability
- LLM Applications
- Retrieval Augmented Generation (RAG)
- Code Models
- Multi-Modal Models
- Language Agents

Module 4: OCI Generative AI Service

- OCI Generative AI Introduction & Service Overview
- How OCI Generative AI Service Works
- Pretrained Foundational Models & Fine-Tuning
- Dedicated AI Clusters
- Demo: Generative AI Service Walkthrough
- Chat Models

Tokens, Parameters, Preamble Override, Temperature, Top-k, Top-p, Frequency/Presence Penalties

- Demo: Chat Models & Inference API
- Embedding Models
- Word & Sentence Embeddings
- Semantic Similarity
- Embeddings Use Cases
- Demo: Embedding Models

Module 5: Advanced Prompting, Fine-Tuning, and RAG

- Prompt Engineering Refresher
- LLMs as Next-Word Predictors
- Aligning LLMs to Instructions
- Prompt Formats & Few-Shot Learning
- Customizing LLMs with Your Data
- Training from Scratch
- Fine-Tuning Pretrained Models (T-Few)

- Fine-Tuning Benefits
- Inference Workflow
- Reducing Inference Costs
- Dedicated AI Cluster Sizing and Pricing
- Example Pricing & Demo
- OCI Generative AI Security
- Dedicated GPU & RDMA Network
- Model Endpoints
- Customer Data and Model Isolation
- Retrieval Augmented Generation (RAG)
- Framework, Techniques, and Pipeline
- Application & Evaluation
- Vector Databases
- Embeddings, Distance, Similarity, Dense Retrieval, Hybrid Search
- Oracle 23ai: AI Vector Search

Module 6: Application Development with OCI Generative AI

- Chatbot Introduction & Demo
- Chatbot Architecture & Components
- OCI Generative AI + LangChain Integration
- Models, Prompts, Chains
- Prompt Templates
- Setting Up Development Environment
- Demo: Prompts, Chains, and LLMs
- Extending Chatbot
- Adding Memory (LangChain Memory)
- Demo: Memory & Streamlit Integration
- Adding RAG
- Indexing, Retrieval, Generation
- Combining RAG + Memory
- Chat History as Context
- Demo: Tracing with LangSmith
- Chatbot Recap & Technical Architecture

Module 7: Deploying LLM Applications

- Deploy Chatbot to OCI Compute Instance (VM)
- Demo: Deployment to VM
- Deploy Chatbot to OCI Data Science
- Deploy LangChain Application as Model