

Oracle Cloud Infrastructure Generative AI Professional: Hands-on Workshop (2025)

Student Guide
D1111080GC10



Copyright © 2025, Oracle and/or its affiliates.

Disclaimer

This document contains proprietary information and is protected by copyright and other intellectual property laws. The document may not be modified or altered in any way. Except where your use constitutes "fair use" under copyright law, you may not use, share, download, upload, copy, print, display, perform, reproduce, publish, license, post, transmit, or distribute this document in whole or in part without the express authorization of Oracle.

The information contained in this document is subject to change without notice and is not warranted to be error-free. If you find any errors, please report them to us in writing.

Restricted Rights Notice

If this documentation is delivered to the United States Government or anyone using the documentation on behalf of the United States Government, the following notice is applicable:

U.S. GOVERNMENT END USERS: Oracle programs (including any operating system, integrated software, any programs embedded, installed or activated on delivered hardware, and modifications of such programs) and Oracle computer documentation or other Oracle data delivered to or accessed by U.S. Government end users are "commercial computer software" or "commercial computer software documentation" pursuant to the applicable Federal Acquisition Regulation and agency-specific supplemental regulations. As such, the use, reproduction, duplication, release, display, disclosure, modification, preparation of derivative works, and/or adaptation of i) Oracle programs (including any operating system, integrated software, any programs embedded, installed or activated on delivered hardware, and modifications of such programs), ii) Oracle computer documentation and/or iii) other Oracle data, is subject to the rights and limitations specified in the license contained in the applicable contract. The terms governing the U.S. Government's use of Oracle cloud services are defined by the applicable contract for such services. No other rights are granted to the U.S. Government.

Trademark Notice

Oracle®, Java, MySQL, and NetSuite are registered trademarks of Oracle and/or its affiliates. Other names may be trademarks of their respective owners.

Intel and Intel Inside are trademarks or registered trademarks of Intel Corporation. All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. AMD, Epyc, and the AMD logo are trademarks or registered trademarks of Advanced Micro Devices. UNIX is a registered trademark of The Open Group.

Third-Party Content, Products, and Services Disclaimer

This documentation may provide access to or information about content, products, and services from third parties. Oracle Corporation and its affiliates are not responsible for and expressly disclaim all warranties of any kind with respect to third-party content, products, and services unless otherwise set forth in an applicable agreement between you and Oracle. Oracle Corporation and its affiliates will not be responsible for any loss, costs, or damages incurred due to your access to or use of third-party content, products, or services, except as set forth in an applicable agreement between you and Oracle.

1003062025

Table of Contents

OCI 2025 Generative AI Professional	10
For whom is this course intended?	11
Course Outline - #1: Fundamentals of Large Language Models	13
Course Outline - #2: Deep Dive on OCI Generative AI Service	14
Course Outline - #3: Implement RAG using OCI Gen AI service + Oracle Database 23ai + Langchain	15
Course Outline - #4: Deep Dive on OCI Generative AI Agents Service	16
Meet your instructors	17
Generative AI Labs	18
Measuring Your Progress: Take the Skill Checks to Test Your Knowledge	19
Get the Answers You Need: <input type="checkbox"/> Use our "Ask Your Instructor" Form or Join the OU Community.....	20
Learning and exam tips	21
Introduction to Large Language Models	22
What is a Large Language Model?	23
This Module	31
LLM Architectures	32
Encoders and Decoders	33
Model Ontology	34
Encoders	35
Decoders	36
Encoders -Decoders	37
Architectures at a glance	38
Prompting and Prompt Engineering	39
Affecting the distribution over Vocabulary	40

Affecting the distribution over Vocabulary	41
Prompting	42
Prompt Engineering	45
In-context Learning and Few-shot Prompting	46
Example Prompts	48
Advanced Prompting Strategies	49
Issues with Prompting	52
Prompt Injection	53
Memorization	55
Training	56
Training	57
Hardware Costs	59
Decoding	60
Decoding	61
Greedy Decoding	63
Non-Deterministic Decoding	66
Temperature	71
Hallucination	75
Hallucination	76
Groundedness and Attributability	78
LLM Applications	79
Retrieval Augmented Generation	80
Code Models	81
Multi-Modal	82
Language Agents	83
OCI Generative AI Introduction	84

OCI Generative AI Service	85
How does OCI Generative AI service work?	86
Pretrained Foundational Models	87
Fine-tuning	88
Dedicated AI Clusters	89
Demo Generative AI service Walkthrough	90
Chat Models	91
Tokens	92
Pretrained Chat Models	93
Chat Model Parameters	94
Preamble Override	95
Temperature	96
Chat Model Parameters	97
Top k	98
Top p	99
Frequency and Presence Penalties	100
Demo Chat Models	101
Demo OCI Generative AI Service Inference API	102
Demo Setting up OCI Config for Generative AI API	103
Embedding Models	104
Embeddings	105
Word Embeddings	106
Semantic Similarity	107
Sentence Embeddings	108
Embeddings use case	109

Embedding Models in Generative AI	110
Embedding Models in Generative AI	111
Demo Embedding Model	112
Prompt Engineering	113
Prompt & Prompt Engineering	114
LLMs as next word predictors	115
Aligning LLMs to follow instructions	116
In-context Learning and Few-shot Prompting	117
Prompt Formats	118
Advanced Prompting Strategies	119
Customize LLMs with your data	120
Training LLMs from scratch with my data?	121
In-context Learning/□Few shot Prompting	122
Fine-tuning a pretrained model	123
Fine-tuning Benefits	124
Retrieval Augmented Generation (RAG)	125
Customize LLMs with your data	126
Fine-tuning and Inference in OCI Generative AI	129
Fine-tuning and Inference	130
Fine-tuning workflow in OCI Generative AI	131
Inference workflow in OCI Generative AI	132
Dedicated AI Clusters	133
T-Few Fine-tuning	134
T-Few fine-tuning process	135
Reducing Inference costs	136
Inference serving with minimal overhead	137

Dedicated AI Clusters Sizing and Pricing	138
Dedicated AI Cluster Unit Types	139
Dedicated AI Cluster Unit Sizing	140
Dedicated AI Cluster Pricing Example	141
Demo Dedicated AI Clusters	142
Generative AI Fine-tuning Configuration	143
Fine-tuning Configuration	144
Fine-tuning Parameters (T-Few)	145
Understanding Fine-tuning Results	146
Accuracy	147
Loss	148
Demo Fine-tuning and Custom Models	149
Demo Inference using Endpoint	150
OCI Generative AI Security	151
Dedicated GPU and RDMA Network	152
Model Endpoints	153
Customer Data and Model Isolation	154
Generative AI leverages OCI Security Services	155
OCI Generative AI Integrations	156
OCI Generative AI and LangChain Integration	157
LangChain Components	158
LangChain Models	159
LangChain Prompt Templates	160
LangChain Chains	161
LangChain Prompt, Model and Chain Interaction	162

LangChain Memory	163
OCI Generative AI and Oracle 23 ai Integration	164
Retrieval Augmented Generation	165
Retrieval Augmented Generation	166
RAG Pipeline	167
Process Documents	168
Document loading	169
Chunking	170
Chunk size	171
Chunk overlap	172
Splitting method	173
Read and Split Documents	174
Embed and Store Documents	175
Embeddings capture semantic relationship	176
Vector Embeddings	177
Generate Vector Embeddings	178
Vector Data Type	179
Connect to the vector database	180
Add metadata to chunks and create documents	181
Embed documents and store in the vector database	182
Retrieve Documents and Generate Response	183
<input type="checkbox"/> Retrieval works like this	184
Comparing vectors - Embedding Distance	185
Faster Search - Vector Indexes	186
<input type="checkbox"/> Augmented Generation works like this	187
Retrieve documents and get response from LLM	188

Conversational RAG	189
Chatbot – Conversational RAG	190
Generative AI Agents	191
OCI Generative AI Agents - Overview	192
OCI Generative AI Agents – Architecture	193
Agents Concepts	194
Object Storage Guideline	196
Oracle Database Guideline: □ Database Requirements	198
Knowledge Base – Object Storage	200
Knowledge Base – Oracle 23ai	201
Agent	202
Endpoint	203
Chat	204
Limitations	205