ORACLE
University

Oracle Cloud Infrastructure Generative AI Professional

Student Guide

S1107455GC10

Learn more from Oracle University at **education.oracle.com**

Copyright © 2024, 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.

1011112024

Table of Contents

OCI 2024 Generative AI Professional **12**

For whom is this course intended? 13

Course Outline #1: Fundamentals of Large Language Models 15

Course Outline #2: Dive-deep on OCI Generative AI Service 16

Course Outline #3: Build an LLM App using OCI Generative AI Service 17

Meet your instructors 18

Measuring Your Progress: Take the Skill Checks to Test Your Knowledge 19

Get the Answers You Need: Use our "Ask Your Instructor" Form or Join the OU Community 20

Best practices and retention tips 21

Keep Progressing: You're on Your Way to Success! 22

Get the Answers You Need: Use our "Ask Your Instructor" Form or Join the OU Community 20

Best practices and retention tips 21

Keep Progressing: You're on Your Way to Success! 22

Introduction to Large Language Models **23**

What is a Large Language Model? 24

This Module 32

LLM Architectures **33**

Encoders and Decoders 34

Model Ontology 35

Encoders 36

Decoders 37

Encoders -Decoders 38

Architectures at a glance 39

Prompting and Prompt Engineering 40
Affecting the distribution over Vocabulary 41
Affecting the distribution over Vocabulary 42
Prompting 43
Prompt Engineering 46
In-context Learning and Few-shot Prompting 47
Example Prompts 49
Advanced Prompting Strategies 50
Issues with Prompting 53
Prompt Injection 54
Memorization 56
Training 57
Training 58
Hardware Costs 60
Decoding 61
Decoding 62
Greedy Decoding 64
Non-Deterministic Decoding 67
Temperature 72
Hallucination 76
Hallucination 77
Groundedness and Attributability 79
LLM Applications 80
Retrieval Augmented Generation 81
Code Models 82

Multi-Modal 83
Language Agents 84
OCI Generative AI Introduction 85
OCI Generative AI Service 86
How does OCI Generative AI service work? 87
Pretrained Foundational Models 88
Fine-tuning 89
Dedicated AI Clusters 90
Demo Generative AI service Walkthrough 91
Chat Models 92
Tokens 93
Pretrained Chat Models 94
Chat Model Parameters 95
Preamble Override 96
Temperature 97
Chat Model Parameters 98
Top k 99
Top p 100
Frequency and Presence Penalties 101
Demo Chat Models 102
Demo OCI Generative AI Service Inference API 103
Demo Setting up OCI Config for Generative AI API 104
Embedding Models 105
Embeddings 106
Word Embeddings 107

Semantic Similarity 108
Sentence Embeddings 109
Embeddings use case 110
Embedding Models in Generative AI 111
Embedding Models in Generative AI 112
Demo Embedding Model 113
Prompt Engineering 114
Prompt & Prompt Engineering 115
LLMs as next word predictors 116
Aligning LLMs to follow instructions 117
In-context Learning and Few-shot Prompting 118
Prompt Formats 119
Advanced Prompting Strategies 120
Customize LLMs with your data 121
Training LLMs from scratch with my data? 122
In-context Learning/☐Few shot Prompting 123
Fine-tuning a pretrained model 124
Fine-tuning Benefits 125
Retrieval Augmented Generation (RAG) 126
Customize LLMs with your data 127
Fine-tuning and Inference in OCI Generative AI 130
Fine-tuning and Inference 131
Fine-tuning workflow in OCI Generative AI 132
Inference workflow in OCI Generative AI 133
Dedicated AI Clusters 134
T-Few Fine-tuning 135

- T-Few fine-tuning process 136
- Reducing Inference costs 137
- Inference serving with minimal overhead 138

Dedicated AI Clusters Sizing and Pricing 139

- Dedicated AI Cluster Units 140
- Dedicated AI Cluster Units Sizing 141
- Dedicated AI Clusters Sizing 142
- Example Pricing 143

Demo Dedicated AI Clusters 144

Generative AI Fine-tuning Configuration 145

- Fine-tuning Configuration 146
- Fine-tuning Parameters (T-Few) 147
- Understanding Fine-tuning Results 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

Retrieval Augmented Generation 156

- Retrieval Augmented Generation 157
- RAG Framework 158
- RAG Techniques 159
- RAG Pipeline 160

RAG Application 161
RAG Evaluation 162
Vector Databases 163
LLM Versus LLM + RAG 164
Vector 165
Embedding Distance 167
Similar Vectors 168
Vector Database Workflow 169
Vector Databases 170
Role of Vector Databases with LLMs 171
Keyword Search 173
Keyword Search 174
Semantic Search 176
Semantic Search 177
Embeddings 178
Dense Retrieval 179
Rerank 180
Hybrid Search --> Sparse + Dense 181
Oracle 23ai 182
AI Vector Search 183
Use Case 184
Benefits 185
Oracle AI Vector Search 186
Oracle AI Vector Search Workflow 187
Vector Datatype 188
Vector Distance Function 189

ONNX compatible 190
Database-Native Vector Embedding Generation 191
Vector Index 192
Vector Index Syntax 193
Similarity Searches in Oracle 23ai 194
Vector Search SQL 195
Vector Search 196
AI Vector Search powers Gen AI pipelines 197
Application□Development 198
Chatbot Introduction 199
Chatbot Introduction 200
Demo Chtbot 201
Chatbot Architecture & Basic Components 202
Chatbot Architecture 203
OCI Generative AI and LangChain Integration 204
LangChain Components 205
Models, Prompts and Chains 206
LangChain Prompt, Model and Chain Interaction 207
LangChain Prompt Templates 208
String Prompt Template and PromptValue 209
Chat Prompt Template and PromptValue 210
LangChain Models 211
LangChain Models – OCI Chat Models 212
LangChain Models – OCI Embedding Models 213
LangChain Chains 214
LangChain Chains 215

Setting Up a Development Environment 216
Demo Setup Development Environment 217
Demo Prompts, Chains, and LLMs 218
Extending Chatbot by Adding Memory 219
LangChain Memory 220
Memory 221
Memory Chat Messages 222
LangChain Memory Per User 223
Demo Memory 224
Demo Streamlitand Memory 225
Extending Chatbot by Adding RAG 226
RAG with LangChain 227
Retrieval Augmented Generation (RAG) with LangChain 228
Read and Split Documents 229
Embed documents and store in the vector database 230
Retrieve documents and send as a context to the llm 231
Demo RAG -Indexing 232
Demo RAG – Retrieval and Generation 233
Extending Chatbot by Adding RAG + Memory 234
RAG Plus Memory 235
Adding chat history as context 236
Print of Response 237
Demo RAG Plus Memory and Tracing with LangSmith 238
Demo Evaluate Model using LangSmith 239
Recap Chatbot Architecture 240
Chatbot Technical Architecture 241

Deploy Chatbot to OCI Compute Instance **242**
Deploy Chatbot to OCI Compute Instance (Virtual Machine) 243
Demo Deploy Chatbot to VM 244
Deploy Chatbot to OCI Data Science **245**
Deploy LangChain Application to Data Science as Model 246