Manual QI160 — OCI Generative AI Service

Compilado em 28/10/2025 • Conteúdo oficial Oracle (docs + páginas comerciais) • pt-BR

Resumo do Manual

Este manual foi projetado para **engenheiros** e **novos usuários** do **OCI Generative AI**. Cobre: *catálogo de modelos*, *playground*, *Dedicated AI Clusters*, *fine-tuning*, *endpoints (públicos/privados)*, *SDK Python*, *CLI*, *precificação*, *limites/quotas*, *IAM*, *operações*, *RAG/Agentes*, *boas práticas* e *troubleshooting*. Todos os links relevantes para verificação estão no final de cada seção e no Apêndice.

1. Visão Geral do Serviço

O **OCI Generative AI** é um serviço gerenciado que provê LLMs e embeddings para chat, geração de texto, sumarização e criação de embeddings. Você pode usar o **Playground**, **APIs** e **CLI** para testar modelos pré-treinados ou **criar/hostear** modelos customizados em **Dedicated AI Clusters**. docs: https://docs.oracle.com/en-us/iaas/Content/generative-ai/overview.htm

2. Catálogo de Modelos Pré-treinados (Completo por Família)

A **lista oficial** de modelos é dinâmica e inclui famílias como **Meta Llama**, **Cohere (Command/Embed/Rerank)** e **xAI Grok**, entre outras. Cada cartão do catálogo indica: *regiões*, *oferta* (on-demand vs dedicado), *benchmarks* e *ciclo de vida* (deprecação/aposentadoria). Para uma **lista completa e atualizada**, consulte: https://docs.oracle.com/en-us/iaas/Content/generative-ai/pretrained-models.htm.

Como usar o catálogo: acesse a página acima, filtre por família/região e abra cada cartão para ver: *Dedicated AI Cluster para o modelo* (tamanho de unidade e número de unidades), *limites*, *modos de consumo* e *observações de versão*.

3. Arquitetura do Serviço

Model Endpoint: ponto em um *Dedicated AI Cluster* onde o modelo recebe requisições e retorna respostas.
Dedicated AI Cluster: recurso isolado para *servir* modelos pré-treinados/customizados e para *fine-tuning*.
Playground: interface para experimentação e testes. docs: https://docs.oracle.com/en-us/iaas/Content/generative-ai/concepts.htm

4. Walkthrough — Console (Playground, Cluster, Endpoint, Fine-tuning)

4.1 Playground

1. Console → Analytics & AI → Generative AI → Playground.
2. Selecione um modelo da lista (pré-treinado) e teste prompts/respostas.
3. Salve experimentos e anote parâmetros (máx. tokens, temperatura, top_p, seed, etc.).

docs: https://docs.oracle.com/en-us/iaas/Content/generative-ai/home.htm

4.2 Criar Dedicated AI Cluster (Hosting)

1. Console → Generative AI → Dedicated AI clusters → Create dedicated AI cluster.
2. Defina nome/descrição, tipo (hosting ou fine-tuning), tags e compartimento.
3. Para hosting: adicione *units* (réplicas) conforme o modelo escolhido.

docs: https://docs.oracle.com/en-us/iaas/Content/generative-ai/create-ai-cluster-hosting.htm

4.3 Fine-tuning — Cluster e Dataset

1. Crie um cluster de fine-tuning ou utilize modo suportado no cluster.
2. Prepare dataset no formato JSONL com pares prompt/completion (mín. 32 pares).
3. Uma base de treinamento por modelo customizado; split 80:20 automático (train/val).

docs: https://docs.oracle.com/en-us/iaas/Content/generative-ai/create-ai-cluster-fine-tuning.htm

requisitos dataset: https://docs.oracle.com/en-us/iaas/Content/generative-ai/limitations.htm

4.4 Criar Endpoint (Público/Privado)

1. Console → Generative AI → Endpoints → Create endpoint.
2. Escolha Público ou Privado. Para privado, crie primeiro o endpoint e depois associe o modelo.
3. Selecione o modelo (pré-treinado ou customizado) e o Dedicated AI Cluster de hosting.

docs: https://docs.oracle.com/en-us/iaas/Content/generative-ai/create-endpoint.htm

5. CLI — Comandos Essenciais (conceituais)

Criar endpoint:

```
oci generative-ai endpoint create \ --compartment-id \ --dedicated-ai-cluster-id \ --model-id \
--name "endpoint-llm-prod" \ --endpoint-type PUBLIC
```

ref CLI: https://docs.oracle.com/iaas/tools/oci-cli/latest/oci_cli_docs/cmdref/generative-ai/endpoint/create.html

Listar Dedicated AI Clusters:

```
oci generative-ai dedicated-ai-cluster list --compartment-id
```

docs: https://docs.oracle.com/en-us/iaas/Content/generative-ai/list-ai-cluster.htm

6. SDK Python — Exemplos Concretos

Inicialização do cliente:

```
from oci.generative_ai import GenerativeAiClient from oci.config import from_file config =
from_file() # ~/.oci/config client = GenerativeAiClient(config)
```

API Python: https://docs.oracle.com/en-us/iaas/tools/python/api/generative_ai.html

Criar endpoint (exemplo conceitual via SDK):

```
from oci.generative_ai.models import CreateEndpointDetails create_details =
CreateEndpointDetails( compartment_id="ocid1.compartment.oc1..xxxx",
dedicated_ai_cluster_id="ocid1.aicluster.oc1..xxxx",
model_id="ocid1.generativeaimodel.oc1..xxxx", name="endpoint-llm-prod", endpoint_type="PUBLIC"
) resp = client.create_endpoint(create_details) print(resp.data.id)
```

Chamada de inferência (conceitual):

```
from oci.generative_ai_inference import GenerativeAiInferenceClient from
oci.generative_ai_inference.models import ChatDetails, Message inf =
GenerativeAiInferenceClient(config) chat = ChatDetails(messages=[Message(role="USER",
content="Explique RAG no OCI")]) # Endpoint OCID/URI conforme provisionado: result =
```

```
inf.chat(chat, endpoint_id="ocid1.generativeaiendpoint.oc1..xxxx")
print(result.data.output_text)
```

API Inference: https://docs.oracle.com/en-us/iaas/tools/python/api/generative_ai_inference.html

7. Padrões de Uso: Chat, Geração, Embeddings, Rerank

Chat/Geração: envie mensagens com papéis (user/system/tool), ajuste *max_output_tokens*, *temperature*, *top_p*, *seed*. **Embeddings:** passe lista de textos, receba vetores; verifique dimensões por modelo. **Rerank:** forneça query + documentos; receba pontuações/ordem para reordenação de resultados.

8. Precificação — Itens, Fórmulas e Exemplos

On-demand (inferência): geralmente medido por transações/caracteres (ou tokens) conforme modelo. Veja exemplos específicos para *xAI Grok*. **Dedicated AI Clusters (hosting):** cobrado por *AI Unit/hora*. Cada modelo/família pede um número de *unidades* (réplicas) e um *tamanho de unidade*. Use a página do modelo (catálogo) para saber a exigência. Fórmula de hosting mensal típica: *744 unit-hours x $preço_unitário*. **Fine-tuning:** também possui métrica em *AI Unit-hora* e requisitos de dataset. Links oficiais com valores e itens por família: https://www.oracle.com/artificial-intelligence/generative-ai/generative-ai-service/pricing/ https://www.oracle.com/cloud/price-list/ https://docs.oracle.com/en-us/iaas/Content/generative-ai/calculate-cost.htm https://docs.oracle.com/en-us/iaas/Content/generative-ai/pay-dedicated.htm https://docs.oracle.com/en-us/iaas/Content/generative-ai/pay-on-demand.htm

9. Limites, Quotas e IAM

Clusters dedicados: por padrão o limite inicial por tenancy é *0*; solicite aumento conforme o modelo e região. **Alguns modelos** são *apenas on-demand* (sem opção de cluster dedicado). Veja na página do modelo. **IAM:** use o recurso *generative-ai-family* para conceder permissões (listar, criar, gerenciar endpoints/cluster/modelos). **Quotas:** defina quotas por compartimento/serviço para controlar consumo. docs: https://docs.oracle.com/en-us/iaas/Content/generative-ai/limits.htm docs quotas (geral): https://docs.oracle.com/en-us/iaas/Content/General/Concepts/servicelimits.htm

10. Integrações: RAG e Generative AI Agents

Para **RAG**, combine embeddings + banco vetorial (ex.: Oracle 23ai) + rerank para melhorar precisão. O serviço **Generative AI Agents** permite orquestrar agentes com *knowledge bases*, ferramentas (ex.: API Endpoint Calling), guardrails e endpoints. Links: Overview Agents: https://docs.oracle.com/en-us/iaas/Content/generative-ai-agents/overview.htm Limits Agents: https://docs.oracle.com/en-us/iaas/Content/generative-ai-agents/limits.htm Pricing Agents: https://www.oracle.com/artificial-intelligence/generative-ai/agents/pricing/ API Tools: https://docs.oracle.com/en-us/iaas/Content/generative-ai-agents/api-calling-tool-add.htm

11. Troubleshooting, Operações e Boas Práticas

- Sem quota para Dedicated AI Clusters: solicite aumento referenciando o modelo exato (unidades/tamanho).
- Erros de endpoint privado: valide VCN/Subnet/rotas/SGs e políticas IAM.
- SDK/CLI desatualizados: atualize e verifique mudanças na API.
- Observabilidade: monitore chamadas, caracteres/tokens, erros; ajuste parâmetros de inferência.

Apêndice A — Links Oficiais (verificação e atualização)

• **Overview do Generative AI**: https://docs.oracle.com/en-us/iaas/Content/generative-ai/overview.htm

• **Home/What's New**: https://docs.oracle.com/en-us/iaas/Content/generative-ai/home.htm

• **Conceitos (Model Endpoint, etc.)**: https://docs.oracle.com/en-us/iaas/Content/generative-ai/concepts.htm

• **Dedicated AI Cluster (hosting)**:
https://docs.oracle.com/en-us/iaas/Content/generative-ai/create-ai-cluster-hosting.htm

• **Dedicated AI Cluster (fine-tuning)**:
https://docs.oracle.com/en-us/iaas/Content/generative-ai/create-ai-cluster-fine-tuning.htm

• **Managing Dedicated AI Clusters**: https://docs.oracle.com/en-us/iaas/Content/generative-ai/ai-cluster.htm

• **Criar endpoint (console/CLI/API)**: https://docs.oracle.com/en-us/iaas/Content/generative-ai/create-endpoint.htm

• **Managing Endpoints**: https://docs.oracle.com/en-us/iaas/Content/generative-ai/endpoint.htm

• **Private Endpoints**: https://docs.oracle.com/en-us/iaas/Content/generative-ai/private-endpoint.htm

• **Listar clusters (console/CLI/API)**: https://docs.oracle.com/en-us/iaas/Content/generative-ai/list-ai-cluster.htm

• **Catálogo de Modelos (completo e dinâmico)**:
https://docs.oracle.com/en-us/iaas/Content/generative-ai/pretrained-models.htm

• **Cálculo de custo (mecânica)**: https://docs.oracle.com/en-us/iaas/Content/generative-ai/calculate-cost.htm

• **On-demand (pagamento)**: https://docs.oracle.com/en-us/iaas/Content/generative-ai/pay-on-demand.htm

• **Dedicated (pagamento)**: https://docs.oracle.com/en-us/iaas/Content/generative-ai/pay-dedicated.htm

• **Price List (geral)**: https://www.oracle.com/cloud/price-list/

• **Pricing (página comercial do serviço)**:
https://www.oracle.com/artificial-intelligence/generative-ai/generative-ai-service/pricing/

• **API Python — GenerativeAiClient**: https://docs.oracle.com/en-us/iaas/tools/python/api/generative_ai.html

• **API Python — Inference**: https://docs.oracle.com/en-us/iaas/tools/python/api/generative_ai_inference.html

• **Agents — Overview**: https://docs.oracle.com/en-us/iaas/Content/generative-ai-agents/overview.htm

• **Agents — Limits**: https://docs.oracle.com/en-us/iaas/Content/generative-ai-agents/limits.htm

• **Agents — Pricing**: https://www.oracle.com/artificial-intelligence/generative-ai/agents/pricing/

• **Agents — API Endpoint Calling Tool**:
https://docs.oracle.com/en-us/iaas/Content/generative-ai-agents/api-calling-tool-add.htm

Aviso: valores, limites e disponibilidade variam por região e podem mudar. Este manual prioriza fórmulas e procedimentos; confira sempre as páginas oficiais para números atuais e lista completa de modelos.