Integração do LiteLLM Proxy com OCI Generative AI

Contexto

O proxy oficial do LiteLLM não possui compatibilidade nativa com o OCI Generative AI, o que inviabiliza o uso de clusters dedicados e causa falhas em requisições para modelos on-demand. Para contornar essa limitação, foi desenvolvida uma versão modificada do proxy que corrige essas falhas e estabiliza a integração.

1. Execução via Docker (Recomendado)

Para evitar configurações manuais de ambientes Python, recomenda-se utilizar a imagem Docker. O container abaixo foi preparado para **ARM64** e **AMD64**:

```
docker run -v $(pwd)/oci.yaml:/app/oci.yaml -p 4000:4000 speglich/litellm:latest --config oci.yaml
```

2. Instalação via Pip

Certifique-se de que o **git** está instalado no ambiente. Em seguida, execute:

```
# Clone do repositório com a versão corrigida
git clone https://github.com/speglich/litellm.git

# Checkout da branch com suporte à OCI
git checkout fix/oci-support

# Instalação local do pacote
pip install .
```

3. Configuração do Arquivo oci.yaml

Crie o arquivo **oci.yaml** no diretório de trabalho. Este arquivo contém as credenciais necessárias para autenticação e execução dos modelos na OCI. Substitua os valores de exemplo pelos dados corretos do seu ambiente:

```
model_list:
  - model_name: meu-modelo-dedicado
    litellm_params:
      model: "oci/OCID_DO_SEU_MODELO"
      oci_region: "SUA_REGIAO"
      oci_user: "OCID_DO_USUARIO"
      oci_fingerprint: "FINGERPRINT_DA_CHAVE_API"
      oci_tenancy: "OCID_DA_TENANCY"
      oci_compartment_id: "OCID_DO_COMPARTIMENTO"
      oci_serving_mode: "DEDICATED"   # ou "ON_DEMAND"
      oci_key: |
        -----BEGIN PRIVATE KEY-----
        COLE_SUA_CHAVE_PRIVADA_COMPLETA_AQUI
        -----END PRIVATE KEY-----
```

Nota de segurança: garanta que o arquivo *oci.yaml* seja protegido, limitando permissões de leitura apenas ao usuário responsável.

4. Execução do Proxy Manual

Caso tenha optado pela instalação via pip, inicie o proxy com:

```
litellm --config oci.yaml
```

O serviço será iniciado na **porta 4000** e estará pronto para receber requisições.

Observações Importantes

- Esta solução é **customizada** para atender uma limitação atual do projeto LiteLLM.
- O **projeto oficial** está em constante evolução. Futuras atualizações podem exigir ajustes ou revisão dessa implementação.
- O código-fonte da modificação está disponível no repositório GitHub para consulta e manutenção.