

warning "Atualização de backend (outubro de 2026)" A inferência do GitHub Models foi descontinuada em 30 de julho de 2026. As referências abaixo são mantidas apenas como contexto histórico enquanto este lab é migrado. Use o Lab 013 ? Backends de LLM gratuitos para configurar o Foundry Local ou o Ollama antes de continuar.

Lab 044: Phi-4 + Ollama em Produção

Nível: L400 Trilha: Pro Code Tempo: ~90 min ? Custo: Gratuito (local) ? 8GB+ RAM para Phi-4-mini

O que Você Vai Aprender

Executar o Microsoft Phi-4 localmente via Ollama (custo zero, sem necessidade de internet)

Fazer benchmark do Phi-4-mini vs GPT-4o-mini em tarefas de raciocínio Construir um agente local-first com fallback automático para a nuvem Adicionar cache de respostas para evitar chamadas redundantes ao LLM Processamento em lote para casos de uso de alto throughput

Introdução

O Phi-4 é a família de modelos de linguagem pequenos de última geração da Microsoft. O Phi-4-mini (3.8B parâmetros) roda em um laptop com 8GB de RAM e iguala o GPT-4o-mini em muitos benchmarks de raciocínio ? gratuitamente, localmente, sem preocupações com privacidade.

Este lab constrói uma configuração de inferência local de nível de produção: cache, processamento em lote, fallback e benchmarking.

Pré-requisitos

Ollama instalado 8GB+ RAM (16GB recomendado para melhor desempenho) Python 3.11+ pip install openai httpx Opcional: GITHUB_TOKEN para fallback na nuvem

Exercício do Lab

Passo 1: Baixar os modelos Phi-4

[diagram or code sample]

Passo 2: Testar o Phi-4 via API compatível com OpenAI do Ollama

O Ollama expõe uma API compatível com OpenAI em http://localhost:11434/v1, então você pode usar o mesmo cliente Python openai:

[diagram or code sample]

Passo 3: Construir um cliente local-first com fallback para a nuvem

[diagram or code sample]

Passo 4: Benchmark do Phi-4-mini vs GPT-4o-mini

[diagram or code sample]

[diagram or code sample]

Passo 5: Processamento em lote para alto throughput

[diagram or code sample]

Comparação de Desempenho (resultados típicos)

Modelo	Latência Média	Qualidade	Custo	Privacidade
phi4-mini (local)	800?2000ms	?????	Gratuito	? Local
phi4 (local)	3000?8000ms	?????	Gratuito	? Local
gpt-4o-mini (nuvem)	300?800ms	?????	~\$0.00015/1k tok	?? Nuvem

tip "Quando escolher local" Use Phi-4-mini para: tarefas de alto volume, dados sensíveis, cenários offline e desenvolvimento. Use nuvem para: raciocínio complexo, SLAs de produção e tarefas que exigem as capacidades mais recentes.

Próximos Passos

Containerize sua configuração de IA local: Adicione o Ollama ao devcontainer (já pré-configurado em .devcontainer/devcontainer.json) Incorpore o Phi-4 em RAG: ? Lab 022 ?

RAG with pgvector

Kit de apoio

O build do site gera downloads estáveis em PDF e ZIP. O ZIP contém o guia do lab e todos os arquivos de apoio declarados para este lab.

Baixar PDF gerado Baixar ZIP gerado

Supporting file: CHEATSHEET.md

Arquivos de apoio: Lab 044: Phi-4 + Ollama em Produção

Comece pelo caminho local ou gratuito documentado. Registre premissas e evidências de validação. Mantenha serviços hospedados opcionais, salvo quando o lab os exigir.

Supporting file: README.md

Lab 044: Phi-4 + Ollama em Produção

Use estes arquivos com o guia do lab e o kit gerado.