

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 026: Agentic RAG Pattern

Nível: L200 Trilha: RAG Tempo: ~40 min ? Custo: GitHub Gratuito + Docker (local)

O que Você Vai Aprender

Por que o RAG ingênuo falha em perguntas complexas Reescrita de consulta ? melhorando a recuperação antes da busca Hypothetical Document Embeddings (HyDE) ? gerar para recuperar RAG multi-hop ? recuperação iterativa para perguntas de múltiplas partes Autorreflexão ? o agente avalia a qualidade de sua própria resposta

Introdução

RAG ingênuo: incorporar consulta ? buscar ? gerar. Funciona bem para perguntas simples.

Falha para:

Perguntas vagas: `"Me fale sobre isso"` (o que é "isso"?) Multi-hop: `"O que é mais barato ? equipamento de camping ou de escalada? Quanto eu economizaria?"` Lacunas de conhecimento: `"Qual é o produto mais novo?"` (pode precisar saber a data atual) Alucinação: o modelo inventa fatos que não estão no contexto

O RAG agêntico adiciona ciclos de raciocínio em torno da recuperação. O agente **decide** como buscar, **avalia** os resultados e **tenta novamente** se necessário.

Pré-requisitos

Ter completado o Lab 022 (pgvector rodando + documentos ingeridos) GITHUB_TOKEN set pgvector container running: `docker start pgvector-rag`

tip "Dados de exemplo já carregados?" Se você executou a etapa de ingestão do Lab 022 com o dataset de exemplo, você já tem 42 documentos no pgvector prontos para este lab. Caso contrário, execute o Passo 3 do Lab 022 primeiro.

Exercício do Lab

Passo 1: Reescrita de consulta

Antes de buscar, peça ao LLM para reescrever a pergunta do usuário em melhores consultas de busca.

[diagram or code sample]

Passo 2: Recuperar com múltiplas consultas e deduplicar

[diagram or code sample]

Passo 3: HyDE ? Hypothetical Document Embeddings

Em vez de incorporar a pergunta, gere uma **resposta hipotética** e incorpore-a. Isso frequentemente corresponde melhor ao documento real.

[diagram or code sample]

Passo 4: RAG Multi-hop

Para perguntas complexas, recupere ? gere resposta parcial ? recupere novamente.

[diagram or code sample]

Passo 5: Autorreflexão (verificação de qualidade da resposta)

[diagram or code sample]

Resumo da Arquitetura de RAG Agêntico

[diagram or code sample]

Próximos Passos

Memória do agente entre sessões: ? Lab 027 ? Agent Memory Patterns Avaliar qualidade do RAG em escala: ? Lab 035 ? Agent Evaluation

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 026: Agentic RAG Pattern

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 026: Agentic RAG Pattern

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