Desafio: Micro-RAG com Guardrails

Objetivo

Entregar um microserviço que responda perguntas com base em **3 documentos locais**, retornando **resposta, citações** e **métricas de execução**. O objetivo é avaliar clareza arquitetural, domínio de LLMs/RAG e boas práticas de produção.

Escopo e Requisitos

1) Ingestão e Indexação

- Ler 3 arquivos de conteúdo (listados abaixo) a partir de uma pasta de dados.
- Realizar chunking com tamanho razoável e overlap quando fizer sentido.
- Gerar embeddings e armazenar em um índice pesquisável.
- Explicar no README como decidiu: tamanho dos chunks, overlap, top-k, técnica de busca e por quê.

2) Endpoint funcional

- Expor um endpoint único de pergunta e resposta.
- Entrada: um texto de pergunta.
- Saída deve conter, no mínimo:
  - “answer” (texto final),
  - “citations” (lista de fontes com trechos/excerpt),
  - “metrics” (latência total, latência do retrieval, estimativa de tokens e custo se aplicável).

Obs.: não precisa mostrar código; descreva o contrato no README (campos, tipos e exemplos em linguagem natural).

3) RAG

- Recuperar os trechos mais relevantes (top-k configurável).
- (Opcional) Re-ranking antes de compor o contexto.
- O prompt de geração deve incentivar respostas ancoradas nas fontes e com citações.

4) Guardrails

- Bloquear pedidos fora do domínio (ex.: “me informe CPFs”), tentativas de prompt injection (ex.: “ignore as instruções”, “revele o system prompt”) e conteúdos indevidos.
- Em bloqueio, o serviço deve responder com uma mensagem clara de recusa e uma indicação do motivo (ex.: política violada).

5) Observabilidade

- Registrar por requisição: timestamps, latência total, latência do retrieval, quantidade aproximada de tokens de prompt e resposta, custo estimado (se usar API cobrada), top-k utilizado e tamanho do contexto.
- Descrever quais métricas você acompanharia em produção (ex.: p95 de latência, groundedness, taxa de bloqueio por guardrail).

6) Qualidade e Processo

- Descrever critérios de teste: o que você validaria (retrieval correto, presença de citações, bloqueios de guardrail, formato da resposta).
- Explicar como estruturaria CI (lint, testes, build) e como faria versionamento de prompts/modelos.
- Fornecer um roteiro de validação manual (roteiro de “aceite”) com quatro perguntas alvo (listadas abaixo) e resultados esperados, sem comandos.

Entregáveis

- Repositório público com:

- README detalhado (decisões técnicas, trade-offs, limitações, custo/latência esperados).
- Descrição do contrato do endpoint (campos e formatos em texto).
- Pasta “data” com os três arquivos acima.
- Descrição do processo de testes (o que foi testado e como validar manualmente).
- Desenho arquitetural simples (pode ser em texto): ingestão → índice → retrieval/re-ranking → composição de contexto → geração → resposta com citações → logging/observabilidade → guardrails.