Desafio Técnico: Pesquisador de IA Generativa (P&D)

1. Boas-vindas ao Desafio

Olá, candidato(a)!
Bem-vindo(a) ao desafio técnico para a posição de Pesquisador(a) em IA Generativa.
Nosso time de P&D em IA Generativa tem como missão explorar o "estado da arte" em IA, validar novas tecnologias e prototipar soluções que possam se tornar features estratégicas em nosso produto. O entregável principal do nosso time não é apenas um código, mas um **Dossiê de Oportunidade**: um pacote que inclui uma Prova de Conceito (POC) funcional e um Relatório de Viabilidade e Integração (analisando custos, arquitetura e desempenho).
Este desafio foi desenhado para ser uma simulação de alta fidelidade de um ciclo de pesquisa real em nossa equipe. Queremos ver como você aborda um problema de negócio, pesquisa soluções, experimenta e comunica seus resultados.

2. O Contexto do Problema

Atualmente, nosso produto possui uma API de extração de dados de documentos . A documentação de referência pode ser encontrada aqui.
Hoje, esta API utiliza um LLM em um processo de duas etapas:

1. **Interpretação e Extração:** O modelo "lê" o documento e extrai as informações solicitadas.
2. **Formatação:** Um segundo *completion(inferência)* é usado para formatar os dados extraídos no schema JSON de saída esperado pelo cliente.

O Problema: Vamos supor um cenário fictício no qual a api apresentasse os seguintes problemas:

- **Latência:** O processo de duas etapas é lento, impactando a experiência do usuário.
- **Custo:** O uso de modelos de ponta para cada chamada tem um custo operacional elevado.
- **Complexidade:** A API sofre para lidar com documentos extensos (muitas páginas) e layouts complexos (ex: tabelas aninhadas, formulários densos, múltiplos blocos de texto).
- **Limitações:** A interpretação de gráficos e imagens não-textuais é limitada.

3. Sua Missão

Sua missão é conduzir uma pesquisa para identificar e validar uma abordagem alternativa

para nossa api que resolva (ou minimize) os problemas de **latência** e **custo**, ao mesmo tempo que busca manter ou aumentar a **acurácia** da extração.
Buscamos soluções que priorizem modelos *open-source* (SLMs ou LLMs) que possam ser auto-hospedados (self-hosted) em nossa própria infraestrutura, oferecendo maior controle sobre custos e desempenho.

4. Entregáveis Esperados

Você deverá nos entregar dois artefatos principais:

Entregável 1: O Dossiê de Pesquisa (Documento Técnico)

Este é o entregável mais importante. Esperamos um documento técnico (em formato .pdf) que siga uma metodologia científica clara, contendo as seguintes seções:

1. **Introdução:** Breve descrição do problema de negócio (baseado na Seção 2) e o objetivo da sua pesquisa.
2. **Metodologia:**
    - Como você conduziu sua pesquisa? Quais foram suas fontes (artigos, fóruns, repositórios)?
    - Quais abordagens você considerou e quais descartou rapidamente (e por quê)?
    - **Uso de Ferramentas de IA:** Incentivamos o uso de IA como ferramenta de apoio à pesquisa e ao desenvolvimento. Caso tenha utilizado, descreva brevemente quais ferramentas usou e como elas auxiliaram no seu processo. Sua autenticidade, criatividade e pensamento crítico na análise final são fundamentais.
3. **Técnicas e Modelos Avaliados:** Uma análise "shortlist" das 2-3 técnicas/modelos mais viáveis que você identificou.
4. **Resultados e Experimentos:**
    - **Nota Importante:** Entendemos que pode ser inviável testar empiricamente *todas* as técnicas (especialmente modelos muito grandes). Nesses casos, **sinta-se à vontade para usar benchmarks, dados de artigos e pesquisas de terceiros** para fundamentar sua análise comparativa. Estamos interessados na sua capacidade de análise e curadoria, mesmo que um experimento prático não seja possível.
    - Quais métricas você usou para comparar as soluções (ex: Acurácia/F1-Score para extração, Latência média por página, Custo de inferência estimado)?
    - Apresente uma tabela ou gráficos comparando os *trade-offs* de cada abordagem (ex: Modelo A é 50% mais rápido, mas 10% menos preciso em layouts complexos).
5. **Conclusão e Recomendação:** Qual técnica/modelo você recomenda? Justifique sua escolha com base nos *trade-offs* e no contexto do problema.
6. **Análise de Viabilidade de Integração:**
    - **Custo:** Qual o custo estimado da infraestrutura para rodar o modelo recomendado (ex: tipo de GPU/CPU, consumo de RAM, custo de apis)?

- **Infraestrutura:** A solução exige alguma arquitetura específica (ex: cluster de GPUs, pipeline de processamento de dados)?

Entregável 2: Prova de Conceito (POC)

Uma implementação mínima (pode ser um Jupyter Notebook, um script Python ou uma API simples) que demonstre o funcionamento da sua **principal recomendação** (o modelo/técnica escolhido na Seção 5 do seu dossiê).

Nota Importante sobre a POC: Entendemos que sua recomendação principal (do Dossiê) pode ser um modelo que exige alta infraestrutura (ex: múltiplas GPUs) e ser inviável de executar em um ambiente local para este desafio. **Se este for o caso, sua POC pode implementar a melhor alternativa viável que você conseguiu testar e executar.** Apenas certifique-se de justificar essa escolha no seu Dossiê.

A POC deve ser capaz de processar 3 casos de uso distintos:

1. **Caso 1: Documento Estruturado (CNH)**
   - **Objetivo:** Extrair campos predeterminados (Nome, CPF, Data de Nascimento, Data de emissão, filiação pai, filiação mãe).
2. **Caso 2: Documento Extenso (The Claude 3 Model Family)**
   - **Objetivo:** Extrair o conteúdo para formato textual mantendo a organização/layout das tabelas e interpretação dos gráficos e imagens.
3. **Caso 3: Documento com Layout Complexo (Fatura de Energia)**
   - **Objetivo:** Extrair o conteúdo da fatura mantendo a organização/layout das informações.

5. Critérios de Avaliação

Nós não estamos buscando uma solução "perfeita", mas sim um processo de pesquisa "excelente". Você será avaliado(a) nos seguintes pontos:

- **Profundidade da Pesquisa:** Você foi além do primeiro resultado do Google? Você demonstrou ter lido documentações técnicas ou artigos? Sua análise sobre modelos relevantes é profunda, **mesmo que baseada em benchmarks de terceiros**?
- **Rigor Metodológico:** Suas métricas são claras? Seus experimentos (práticos ou teóricos) são justos e comparáveis?
- **Clareza e Comunicação:** Seu "Dossiê de Pesquisa" é fácil de ler, bem estruturado e convincente? Ele seria compreendido tanto por um Engenheiro quanto por um Gerente de Produto?
- **Pragmatismo (Visão de Produto):** Sua análise de viabilidade considera os *trade-offs* reais (custo, velocidade, acurácia)? Você equilibrou a "tecnologia mais nova" com a "solução mais viável"?

6. Diretrizes e Prazo

- **Autenticidade:** Esperamos que a análise crítica, os resultados e as conclusões deste trabalho sejam seus. O uso de ferramentas de IA é incentivado como apoio.
- **Foco:** Gaste 80% do seu tempo no Dossiê de Pesquisa (Entregável 1) e 20% na POC

(Entregável 2). A qualidade da sua análise é mais importante que a complexidade do seu código.

- **Envio:** Por favor, envie seus entregáveis (Documento Técnico + Link para um repositório Git com a POC) até terça(04/11) às 23:59 como resposta deste email.

Boa sorte! Estamos ansiosos para ver sua abordagem para este desafio.