Documentação do Projeto: Extrator-NFSe-Ollama

- **Título do Projeto:** Extrator-NFSe-Ollama: Automação Inteligente de Processamento de Notas Fiscais
- **Autor:** Kelvin Felipe Dos Santos
- **Disciplina:** Projeto em Computação Aplicada
- **Professor:** Luis Carlos Dos Santos Junior
- **Data:** 15 de Novembro de 2025

---

1. Introdução

O processamento de documentos fiscais é uma tarefa fundamental na gestão financeira de qualquer empresa. No entanto, a natureza não estruturada da maioria desses documentos, como Notas Fiscais de Serviço (NFSe) — frequentemente distribuídas em formatos PDF ou imagem — impõe um desafio operacional significativo. O processo manual de extração de dados é lento, caro e suscetível a erros humanos.

Este projeto propõe o desenvolvimento de um sistema de automação inteligente, o Extrator-NFSe-Ollama, projetado para otimizar esse fluxo de trabalho. O objetivo principal é criar uma solução *end-to-end* que automatiza a leitura, extração, validação e armazenamento de dados de NFSe.

Para alcançar isso, o sistema utiliza uma arquitetura híbrida: emprega o serviço de OCR (Reconhecimento Óptico de Caracteres) de alta precisão **Azure Vision** para converter documentos em texto e um Modelo de Linguagem Grande (**LLM**) executado localmente via **Ollama** para interpretar e estruturar esses dados. A solução é apresentada através de uma interface web construída em **Streamlit** e os dados validados são persistidos em um banco de dados **MySQL** para futuras análises.

---

2. O Problema

Empresas de todos os portes enfrentam um gargalo operacional comum no processamento de notas fiscais recebidas de fornecedores. Este problema se manifesta em quatro áreas principais:

1. **Custo Operacional Elevado:** A necessidade de alocar funcionários para a digitação manual de dados de centenas ou milhares de notas fiscais gera um custo de mão de obra direto e significativo.
2. **Alta Incidência de Erros:** A digitação manual é inerentemente propensa a erros, como a troca de números em um CNPJ, valores incorretos ou datas erradas. Esses erros podem levar a problemas de conformidade fiscal e contabilidade imprecisa.
3. **Dados "Presos" (Não Estruturados):** Quando as informações de uma NFSe estão em um arquivo PDF ou imagem, elas não estão acessíveis para análise de software. Os dados ficam "presos" no documento, inutilizáveis para inteligência de negócios.

4. **Dificuldade na Geração de Relatórios:** A falta de um repositório de dados centralizado e estruturado impede que gestores financeiros gerem relatórios em tempo real, como dashboards de gastos por fornecedor, impostos retidos ou análise por centro de custo.

---

3. A Solução Proposta

A solução desenvolvida é um sistema completo que aborda os problemas mencionados através de quatro componentes integrados:

3.1. Interface Web e Gestão

Uma aplicação web construída em **Streamlit** serve como o *frontend* do sistema. Ela fornece uma interface amigável para:

- Autenticação e gerenciamento de usuários.
- Upload em lote de arquivos NFSe (PDF, PNG, JPG).
- Uma tela de validação humana (`st.data_editor`) para revisar os dados extraídos pela IA.
- Um dashboard financeiro para visualização dos dados já processados.

3.2. OCR de Alta Precisão em Nuvem

Para garantir a qualidade máxima dos dados de entrada, o sistema utiliza a API do **Azure Vision**. Este serviço é responsável por analisar os arquivos de imagem ou PDF e extrair todo o texto bruto contido neles. O uso de uma solução de OCR robusta é vital, pois a precisão do LLM subsequente depende diretamente da qualidade do texto recebido.

3.3. Extração Inteligente e Local (LLM)

Este é o núcleo da inteligência do sistema. Após o OCR, o texto bruto é enviado ao **Ollama**, um gerenciador que executa LLMs (como Phi 3 ou Llama 3) localmente. Os principais benefícios desta abordagem são:

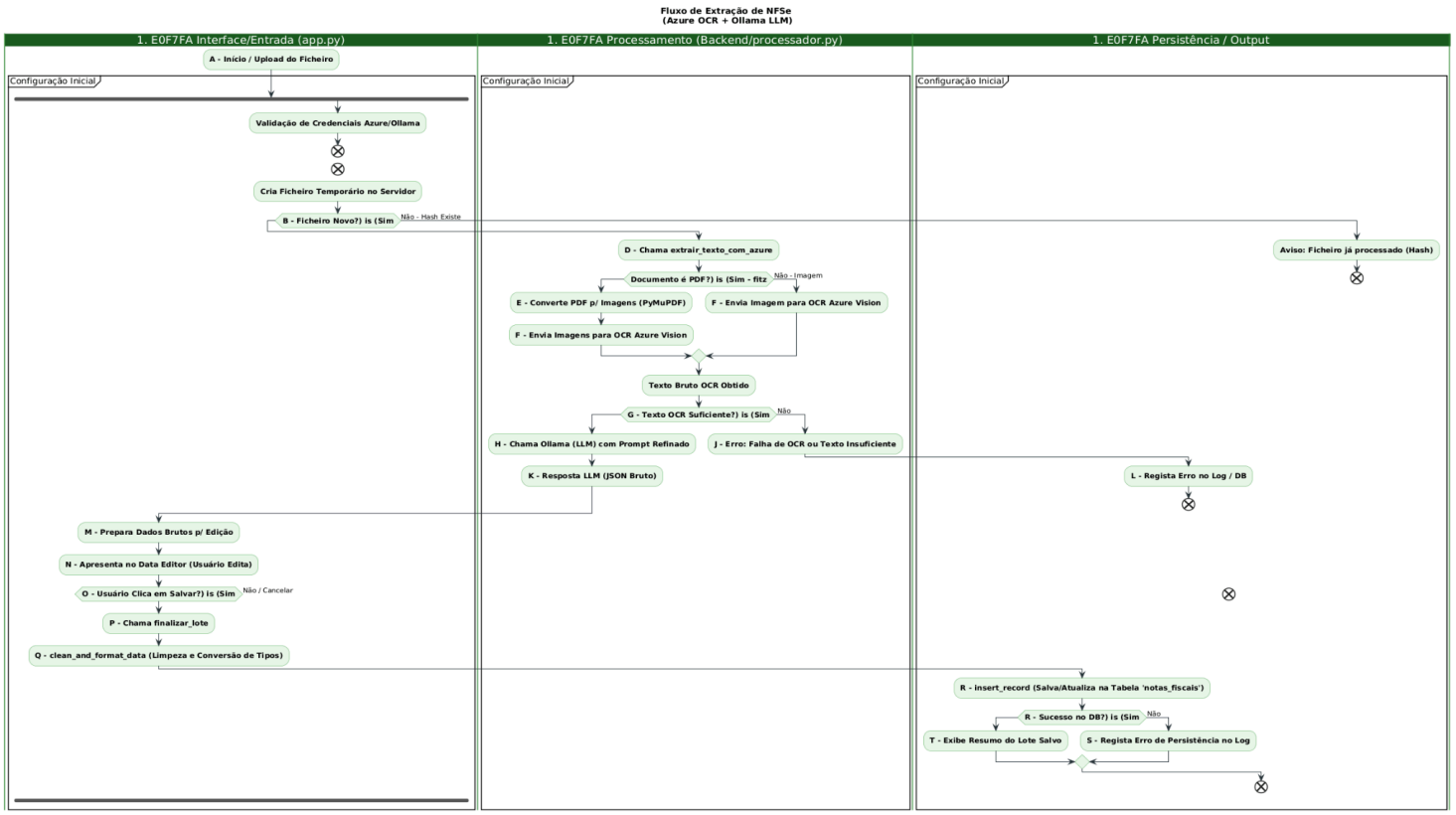
- **Privacidade:** Os dados fiscais sensíveis do cliente (contidos no texto) nunca saem da infraestrutura local para uma API de IA de terceiros.
- Custo: A inferência do LLM é executada em hardware local, eliminando custos por chamada de API.
  O LLM é instruído através de engenharia de prompt para analisar o texto e retornar um objeto JSON perfeitamente estruturado contendo os campos desejados (ex: ocr_prestador_nome, ocr_valor_total, etc.).

3.4. Persistência e Análise

Os dados validados pelo usuário são higienizados (ex: "R$ 1.500,00" é convertido para o tipo `float` `1500.0`) e armazenados em um banco de dados relacional **MySQL**. Esta estruturação permite que os dados sejam imediatamente consumidos pelo dashboard do Streamlit ou por qualquer ferramenta externa de Business Intelligence (BI).

4. Arquitetura e Fluxo de Trabalho

O fluxo de processamento de um documento no sistema segue uma sequência lógica e rastreável. O diagrama abaixo ilustra a arquitetura:



Legenda: Diagrama de fluxo de trabalho do sistema, detalhando a interação entre a interface (Streamlit), o OCR (Azure), o LLM (Ollama) e o Banco de Dados (MySQL).

O processo pode ser descrito em 7 etapas:

1. **Upload e Autenticação:** O usuário faz login no sistema Streamlit e faz o upload de um ou mais arquivos NFSe.
2. **Verificação de Duplicidade (Hash):** O sistema calcula o *hash* MD5 do arquivo. Uma consulta é feita ao MySQL para verificar se este *hash* já existe. Se sim, o arquivo é ignorado para evitar duplicidade.
3. **Pré-processamento e OCR:** Se o arquivo for um PDF, a biblioteca **PyMuPDF (fitz)** é usada para converter cada página em uma imagem de alta resolução. As imagens são então enviadas ao **Azure Vision**, que retorna o texto bruto.
4. **Extração de Dados (LLM):** O texto bruto é formatado dentro de um *prompt* de sistema detalhado, que instrui o **Ollama** a extrair os campos de interesse e retornar exclusivamente um objeto JSON.

5. **Validação Humana:** O JSON retornado pelo LLM é usado para popular uma tabela editável (`st.data_editor`) na interface. O usuário tem a oportunidade de revisar e corrigir rapidamente qualquer campo que a IA possa ter interpretado erroneamente.
6. **Higienização e Persistência:** Após a aprovação do usuário, os dados validados passam pela função `clean_and_format_data`. Esta função crucial converte strings em tipos numéricos (float) e datas (datetime), garantindo a integridade dos dados antes de serem salvos no **MySQL**.
7. **Visualização:** Os dados recém-salvos no MySQL são imediatamente disponibilizados para o dashboard financeiro da aplicação.

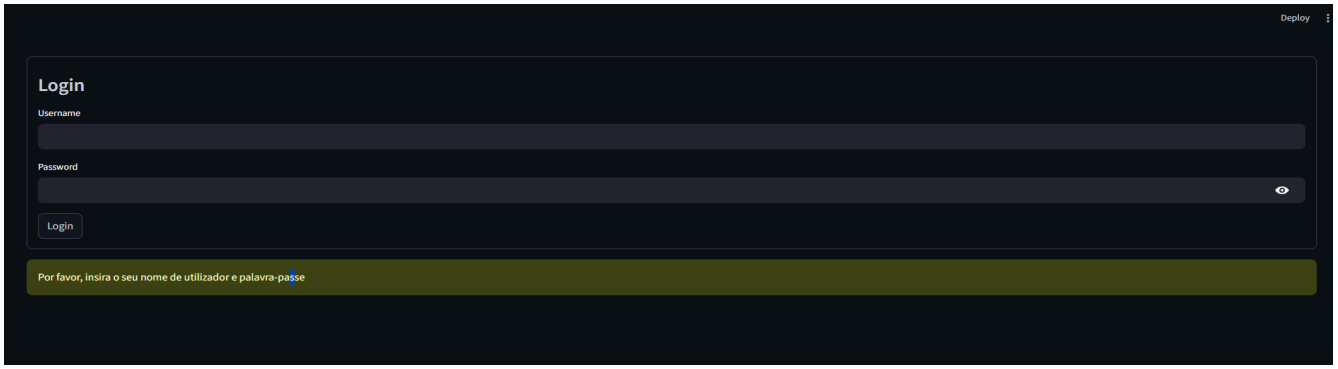
---

5. Apresentação da Solução (Funcionalidades)

A interface do sistema foi desenvolvida em Streamlit para garantir uma experiência de usuário intuitiva e direta. Abaixo estão as principais funcionalidades da aplicação:

5.1. Autenticação e Gestão de Usuários

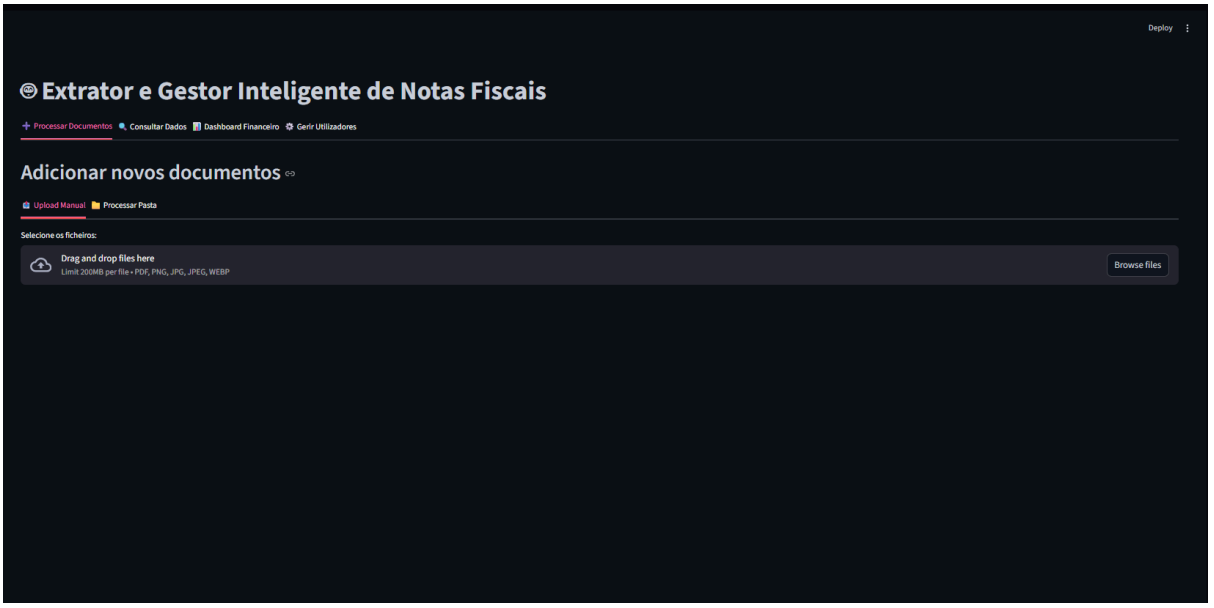
A segurança é tratada através do módulo `streamlit-authenticator`. O sistema exige login e senha, com os *hashes* das senhas armazenados no banco de dados. Usuários administradores têm acesso a um painel (⚙️) onde podem criar novos usuários, redefinir senhas e gerenciar permissões.



5.2. Processamento de Documentos (Upload)

Na aba "Processar Documentos" (➕), o usuário pode submeter as notas fiscais. A aplicação oferece duas modalidades:

1. **Upload Manual:** Permite o upload de múltiplos arquivos (PDF, PNG, JPG) diretamente do computador do usuário.
2. **Processar Pasta:** Permite que o sistema leia e processe todos os arquivos de uma pasta específica já existente no servidor, ideal para fluxos de automação em lote.



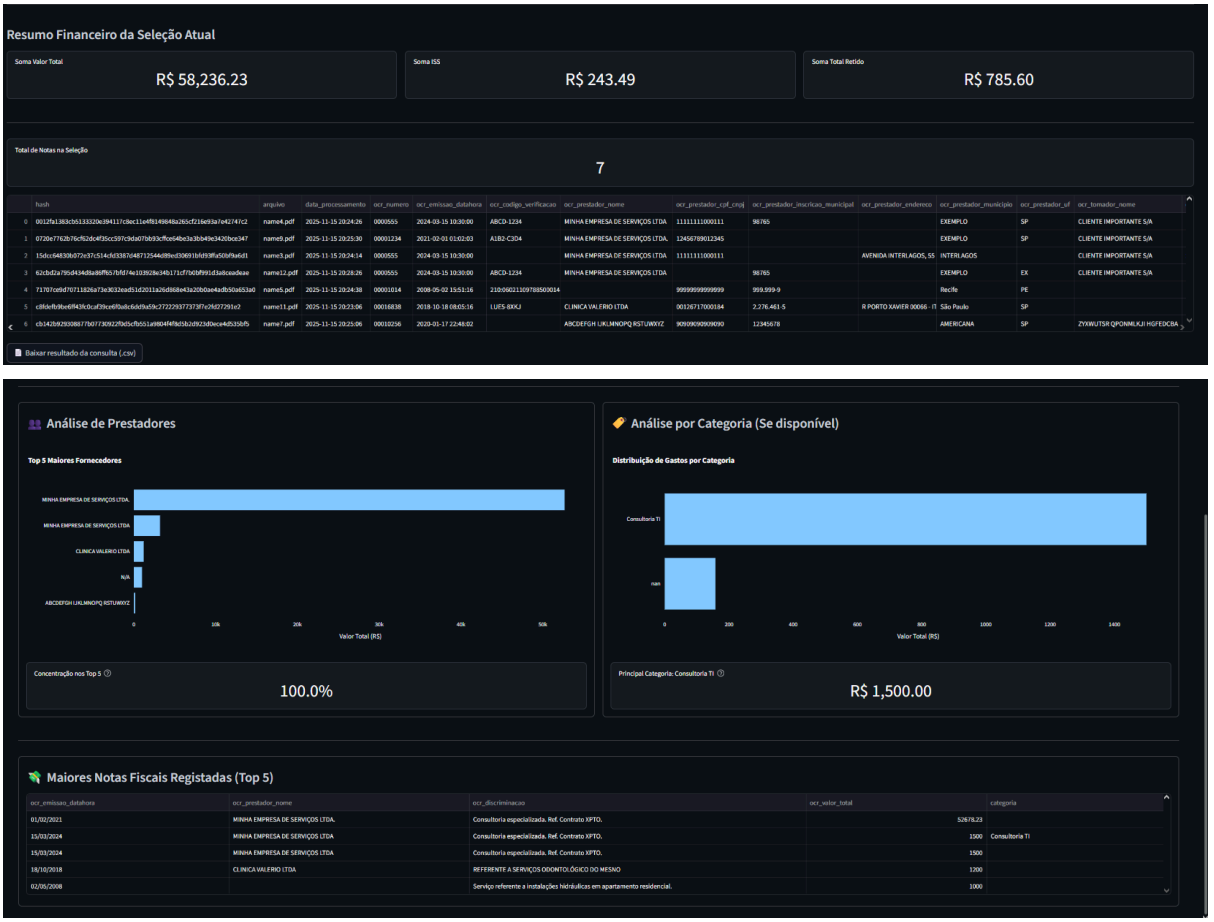
5.3. Validação Humana (Etapa Crítica)

Esta é a tela mais importante do fluxo de trabalho. Após o Ollama extrair os dados, eles não são salvos automaticamente. Em vez disso, são apresentados em uma tabela interativa (st.data_editor). O usuário pode revisar rapidamente os dados extraídos (JSON bruto) e clicar duas vezes em qualquer célula para fazer correções manuais. Somente após clicar em "Salvar Dados", os dados são higienizados e enviados ao banco.

	hash	arquivo	data_processamento	ocr_numero	ocr_emissao_datahora	ocr_codigo_verificacao	ocr_prestador_nome	ocr_prestador_cpf_cnpj	ocr_prestador_inscricao_municipa
0	0012fa1383cb5133320e394117c8ec11e4f8149848a265cf216e93a7e42747c2	name1.pdf	2025-11-15 20:05:10	00000001	1/6/2006 12:01:40	200606011202	FACULDADE DE MEDICINA APLICADA LTDA	64.167.648/0001-33	1.311.306-8
1	0720e7762b76cf62dc4f35cc597c9da07bb93cffce64be3a3bb49e3420bce347	name10.pdf	2025-11-15 20:05:24	00001234	01/02/2021 01:02:03	A1B2-C3D4	PRESTADOR DE SERVIÇOS	12.456.789/0123-45	
2	c8fdefb9be6ff43fc0caf39ce6f0a8c6dd9a59c272229377373f7e2fd27291e2	name11.pdf	2025-11-15 20:05:39	00016838	18/10/2018 08:05:16	LUE5-8XKJ	CLINICA VALERIO LTDA	00.126.717/0001-84	2.276.461-5
3	62cbd2a795d434d8a86ff657bfd74e103928e34b171cf7b0bf991d3a8ceadeae	name12.pdf	2025-11-15 20:05:52	20210	29/01/2021 12:10:44	R. Dom Luiz Maria de Santana	COMERCIO DE METAIS LTDA	37.05	
4	dd64f386c9c63e6092fa2f54861a747a6b7eeebaea89e6fb9ad7211ebe9df7f1	name13.pdf	2025-11-15 20:06:05	00002701	03/08/2011 17:18:22		INSCRICAO PARA TESTE NFE - PJ/0001	99.999.999.0001-00	3.961.999-4
5	dd64f386c9c63e6092fa2f54861a747a6b7eeebaea89e6fb9ad7211ebe9df7f1	name14.pdf	2025-11-15 20:06:18	00002701	03/08/2011 17:18:22		INSCRICAO PARA TESTE NFE - PJ/0001	99.999.999.0001-00	3.961.999-4
6	0012fa1383cb5133320e394117c8ec11e4f8149848a265cf216e93a7e42747c2	name2.pdf	2025-11-15 20:06:33	0000555	15/03/2024 10:30:00	ABCD-1234	MINHA EMPRESA DE SERVIÇOS LTDA	11.111.111/0001-11	98765
7	15dcc64830b072e37c514cfd3387d48712544d89ed30691bfd93ffa50bf9a6d1	name3.pdf	2025-11-15 20:06:47	0000555	15/03/2024 10:30:00		MINHA EMPRESA DE SERVIÇOS LTDA	11.111.111/0001-11	
8	0012fa1383cb5133320e394117c8ec11e4f8149848a265cf216e93a7e42747c2	name4.pdf	2025-11-15 20:07:00	0000555	15/03/2024 10:30:00	ABCD-1234	MINHA EMPRESA DE SERVIÇOS LTDA	11.111.111/0001-11	98765
9	71707ce9d7071182 6a73e3032ead51d2011a26d868e43a20b0ae4adb50a653a	name5.pdf	2025-11-15 20:07:11	00001014	02/05/2008 15:51:16	210:060211097885000148		99.999.999/9999-99	999.999-9

5.4. Consulta e Dashboard Financeiro

Uma vez que os dados estão no MySQL, as abas "Consultar Dados" (🔍) e "Dashboard Financeiro" (📊) tornam-se funcionais. A tela de consulta permite pesquisar notas específicas (por CNPJ, número, etc.), enquanto o dashboard apresenta visualizações gráficas, como a evolução mensal de gastos, os principais fornecedores (prestadores) e a distribuição de impostos retidos.



---

6. Desafios e Lições Aprendidas

Durante o desenvolvimento, vários desafios técnicos foram encontrados, fornecendo aprendizados valiosos:

- **Desafio: A "Alucinação" do LLM.** O principal desafio foi garantir que o LLM retornasse *sempre* um JSON válido e *apenas* o JSON, sem adicionar texto introdutório (ex: "Aqui está o JSON que você pediu: ...").
    - **Solução:** Foi necessário um trabalho intenso de **engenharia de prompt**. O *prompt* final é altamente restritivo, utiliza a técnica de *few-shot learning* (fornecendo um ou dois exemplos de entrada e saída) e especifica rigorosamente o formato de saída. A escolha de modelos com boa capacidade de seguir instruções (como o Phi 3) foi essencial.
- **Desafio: Variedade de Layouts de NFSe.** Diferente da Nota Fiscal de Produto (NFe), que possui um padrão XML único (DANFE), a Nota Fiscal de Serviço (NFSe) varia drasticamente de uma prefeitura para outra.
    - **Solução:** A arquitetura de dois estágios (OCR + LLM) foi a decisão correta. Tentar extrair dados com base em coordenadas (templates) seria inviável. O **Azure Vision** provou ser robusto na extração de texto de layouts complexos, e o **LLM** demonstrou capacidade de generalização, encontrando os campos corretos (ex: "Valor Total") independentemente de sua posição no documento.

- **Desafio: Higienização de Dados (ETL).** O LLM retorna apenas texto. Um valor "1.500,00" não é um número, é uma string.
    - **Solução:** A criação da função `clean_and_format_data` foi uma etapa crítica de mini-ETL (Extract, Transform, Load). Esta função usa expressões regulares e lógica de conversão para garantir que os dados que chegam ao MySQL estejam limpos, tipados e prontos para operações matemáticas (soma, média).

---

7. Conclusão e Trabalhos Futuros

Este projeto demonstrou com sucesso a viabilidade de construir uma solução de automação de documentos de baixo custo operacional e alta segurança, combinando o poder de serviços de nuvem (Azure) com a privacidade da IA local (Ollama). O sistema transforma dados "mortos", presos em PDFs, em dados "vivos" e acionáveis, armazenados em um banco de dados SQL.

Como trabalhos futuros, várias melhorias são possíveis:

1. **Fine-Tuning (Ajuste Fino):** Utilizar os dados validados pelo usuário (o "antes" do LLM e o "depois" da correção humana) para realizar o *fine-tuning* do modelo LLM local. Isso aumentará drasticamente a precisão da extração, podendo, no futuro, reduzir a necessidade de validação humana para perto de zero.
2. **Suporte a XML:** Expandir o sistema para aceitar o upload de arquivos `.xml` de NFSe. Como estes arquivos já são estruturados, eles podem pular as etapas de OCR e LLM, indo diretamente para a validação e persistência, tornando o processo instantâneo.
3. **Dashboards Avançados:** Com os dados estruturados no MySQL, criar novas visualizações e análises, como gastos por centro de custo, impostos retidos por fornecedor, ou alertas automáticos de despesas.