A proposta:

1- O desafio a ser resolvido é o desafio 2.
2- WECANDO NEGOCIOS E INOVACAO LTDA , PAULO.MELO@WECANDO.COM.BR ,(62) 8564-2474 ou (62) 98216-9350.
3-
INOVAÇÃO: A solução proposta consiste em uma plataforma de consulta e análise, desenvolvida para auxiliar os servidores da Controladoria Geral do Estado (CGE) na busca e análise de documentos relacionados a processos de auditoria, legislações e notícias da internet. Com o uso de tecnologias de processamento de linguagem natural (PLN), a plataforma permite a extração automática de conteúdo escaneado e estruturado de documentos em PDF. Além disso, conta com um sistema que faz vetorização de pedaços de documentos e armazena em um banco de dados vetorial, que possibilita a implementação de modelos eficientes de busca semântica entre todos arquivos. Essa combinação permite a recuperação de processos semelhantes no histórico de documentos da CGE, proporcionando uma experiência de consulta precisa e intuitiva utilizando linguagem natural no modelo de chat com os documentos e outras funcionalidades que serão citadas posteriormente neste texto. Com o apoio tecnológico do banco de dados vetorial, os servidores conseguem encontrar rapidamente processos similares aos que estão trabalhando, facilitando a elaboração de seus processos de jurisprudência e tornando a análise mais ágil e precisa.

A plataforma centraliza e organiza informações dispersas, permitindo acesso rápido e eficiente a dados relevantes, o que facilita a análise e o uso de informações pelos servidores de maneira responsiva. Esse ganho de produtividade representa uma economia significativa para a CGE, pois seus servidores podem dedicar mais tempo a atividades de maior valor agregado, reduzindo o custo e o tempo de trabalho envolvidos nas pesquisas manuais de documentos e legislações.

No estágio atual, a solução encontra-se no nível de prontidão tecnológica TRL5, com um protótipo funcional testado em um ambiente relevante. O ambiente de teste utilizou documentos públicos da Controladoria-Geral da União (CGU) – um órgão semelhante à CGE – extraídos do site eaud.cgu.gov.br simulando o ambiente da CGE e validando a eficácia e funcionalidade da solução em um contexto próximo ao de operação real.

A solução incorpora tecnologias robustas, como o padrão RAG (Retrieved Augmented Generation), Inteligência Artificial generativa com arquitetura GPT (Generative Pre-Trained Transformer), modelos de linguagem de grande escala (LLMs) de código aberto, OCR (Optical Character Recognition) e bancos de dados vetoriais. Essas tecnologias aumentam as capacidades dos LLMs, combinando o conhecimento desses modelos com dados externos (contextos específicos dos processos), melhorando a precisão e a credibilidade das respostas através da citação de fontes. Esse conjunto de tecnologias já foi testado e comprovado em projetos como o notebook LM da Google e a ferramenta Louise, utilizada pelo Tribunal de

Contas do Estado de Goiás (TCE-GO) para otimizar buscas em seu portal.

Além disso, a plataforma opera exclusivamente dentro da infraestrutura estatal, sem dependência de APIs externas, garantindo a segurança e a integridade dos dados confidenciais, e conta com um sistema robusto de autorizações que assegura o acesso controlado às informações pelos servidores.

O **PIPELINE** da plataforma de consulta e análise é dividido nas seguintes etapas:

1. **Ingestão de Dados**: Partindo da hipótese de que a CGE vai liberar o acesso à todos os dados dos processos (do SEI e da base de dados dos que foram escaneados), o sistema realiza a integração com as fontes de dados da CGE, permitindo a ingestão inicial de documentos em PDF (ex: legislações, textos específicos ou arquivos de processos da CGE). A plataforma também permite upload direto de documentos em PDF, preparando o ambiente antes da primeira interação do usuário, para otimizar o tempo de processamento.

2.**Parsing de PDFs**: Com os documentos carregados, o sistema realiza a extração de dados dos PDFs por meio de OCR e técnicas de conversão PDF para texto, dividindo o conteúdo em partes (“chunks”) otimizadas para busca. Cada fragmento de texto recebe metadados antes de salvar, como por exemplo a origem e contexto do processo, facilitando futuras referências na construção de respostas.

3. **Configuração do Banco de Dados Vetorial**: Nesta etapa, é selecionado o banco de dados vetorial mais adequado para a expertise da equipe da infraestrutura estadual (Dentre as tecnologias: Qdrant, Elastic, Oracle ou MongoAtlas), onde são configurados os metadados e a estrutura de indexação. Em seguida, as informações processadas são armazenadas de maneira organizada para facilitar buscas semânticas.

4. **Login do usuário final (o servidor público que faz as auditorias)**: O usuário acessa a plataforma por meio do site (que a TI da CGE fará que ele fique disponível apenas para a rede do órgão) e realiza o login com suas credenciais únicas. O sistema permitirá que seja configurado permissões específicas para que dependendo do perfil do usuário, ele tenha acesso a somente uma parte do sistema. Abriremos a oportunidade um servidor da TI da CGE vir personalizar a maneira que será feita o login para se adequar ao login único já utilizado dentro do órgão(como por exemplo keycloak e LDAP)

5. **Processamento e Resolução de Consultas via chat:** Para informar o contexto do processo que o usuário está trabalhando, ele irá preencher o número de processo e a plataforma buscará nos documentos da base de dados, os documentos relacionados a tal processo. Depois disso, o usuário poderá fazer perguntas aos documentos relacionados ao processo por meio de uma interface de chat e o sistema gerará uma resposta contextualizada citará de onde ele trouxe as informações para respondê-lo.

6. **Padrão RAG**: A plataforma adota o padrão RAG, onde realiza a recuperação de informações antes de gerar respostas com base nos dados e com base nessa recuperação é capaz de

personalizar o sistema para a geração de respostas com o conteúdo específico dos documentos do CGE.

7. **Criação de Processos na aplicação**: O usuário têm a opção de criar novos processos, definindo etapas específicas e vinculando documentos ou dados pertinentes a esses processos.

8. **Finalização e Consultas Futuras**: O usuário pode salvar suas consultas e resultados obtidos, permitindo que novas perguntas sejam feitas com base no histórico de interações na plataforma.

INTEGRAÇÃO: A solução é capaz de realizar a devida integração com os processos de dados da CGE, bem como com serviços privados em nuvem, permitindo a coleta de dados e o tratamento alinhado com as bases de dados estaduais e federais.O OCR será utilizado para extrair informações de documentos digitalizados, convertendo-os em texto editável e correlacionando com os dados como datas e números de processos com registros já existentes na CGE. A técnica de web scraping realizará a coleta de dados de portais governamentais, como sites de transparência e páginas de publicações oficiais, atualizando informações importantes para a CGE. Esses dados serão processados e integrados nas bases da CGE para garantir que novos relatórios e atualizações estejam disponíveis. Já o banco de dados vetorial permitirá que usuários da CGE localizem informações contextualmente relacionadas em grandes volumes de dados, melhorando a eficiência na busca de documentos e a precisão das consultas.

RESILIÊNCIA e ESCALABILIDADE: A solução proposta é projetada para possuir **Resiliência** e ser **Escalável**, pois é capaz de se adaptar a diferentes tipos de processos da CGE, como auditorias de conformidade, avaliações de risco e investigação de fraude. Ou seja, devido às tecnologias utilizadas a plataforma é capaz de recuperar informações de diferentes fontes de dados, como documentos em vários idiomas e formatos distintos aumentando a robustez e adaptabilidade da solução para novos desafios e contextos, sendo capaz de suportar uma expansão eficiente, pois a utilização de tecnologias em nuvem e o armazenamento em bancos de dados vetoriais possibilitam a adição de funcionalidades sem comprometer a performance podendo crescer em termos de capacidade de processamento e volume de dados.

TEMPO DE DESENVOLVIMENTO: O **tempo de desenvolvimento** da solução será em média 10 meses, devido ao tempo das tarefas de Levantamento processos, Infra Busca, Integração dados CGE, Importação de procesos escaneados/OCR, a criação da Interface Usuário, a otimização Modelo linguagem, a disponibilização de Endpoints, a criação de um Sistema Filas e a Especialização do produto final.

TESTES:
Descrição dos Testes da Solução Inovadora

Para assegurar que a solução inovadora atenda plenamente às necessidades dos stakeholders e opere de forma eficaz, será implementado um plano de testes abrangente. Este plano

contempla atividades detalhadas de teste para cada uma das três principais funcionalidades desenvolvidas, bem como entregáveis específicos ao longo do CPSI.

1. Planejamento dos Testes

- **Análise de Requisitos**: Revisão minuciosa dos requisitos funcionais e não funcionais para garantir que todos os aspectos críticos sejam abordados nos testes.
- **Desenvolvimento do Plano de Testes**: Elaboração de um documento que delineia a estratégia de testes, incluindo casos de teste, critérios de aceitação e cronograma das atividades.

2. Testes das Funcionalidades Principais

a) Chat por Processo com RAG (Geração Aumentada por Recuperação)

- **Testes Unitários**: Verificação individual dos componentes do chat, assegurando que cada elemento funcione corretamente.
- **Testes de Integração**: Avaliação da interação entre o chat, o LLM e os documentos associados ao processo, garantindo respostas precisas e relevantes.
- **Testes de Usabilidade**: Sessões com usuários finais para avaliar a interface e a experiência de uso, identificando possíveis melhorias.
- **Testes de Precisão Semântica**: Verificação da capacidade do sistema em compreender e responder adequadamente às consultas em linguagem natural.
- **Avaliação com Métricas RAGAS**: Implementação do framework RAGAS para medir a eficácia do RAG, utilizando métricas como precisão, recall, relevância e qualidade das respostas geradas.

b) Card de Notícias Relacionadas

- **Testes de Relevância**: Avaliação da pertinência das notícias apresentadas em relação aos documentos do processo, ajustando os critérios de busca por palavras-chave conforme necessário.
- **Testes de Atualização**: Verificação da frequência e precisão com que novas notícias são incorporadas ao card.

c) Card de Documentos Relacionados

- **Testes de Busca Semântica**: Validação da eficácia das buscas semânticas na base indexada, incluindo a recuperação de outros processos e legislações relacionadas.
- **Testes de Consistência**: Garantia de que os documentos relacionados estão corretos e atualizados.

3. Testes de Escalabilidade e Desempenho

- **Testes de Carga (Load Testing)**: Simulação de múltiplos usuários acessando simultaneamente a solução para avaliar o comportamento sob alta demanda e identificar possíveis gargalos.
- **Testes de Stress (Stress Testing)**: Submissão do sistema a cargas além dos níveis operacionais esperados para determinar sua capacidade máxima e pontos de falha.
- **Testes de Volume (Volume Testing)**: Avaliação do desempenho ao manipular grandes volumes de dados, garantindo que a solução mantenha tempos de resposta aceitáveis.
- **Testes de Performance (Performance Testing)**: Medição do tempo de resposta, taxa de transferência e uso de recursos do sistema durante operações normais e de pico.
- **Análise e Otimização**: Identificação de áreas de melhoria e otimização do código, consultas e infraestrutura para aprimorar a eficiência e a escalabilidade.

4. Avaliação do RAG com Métricas RAGAS

- **Implementação do Framework RAGAS**: Utilização do RAGAS (Retrieval-Augmented Generation Assessment Scores) para avaliar sistematicamente o desempenho do componente RAG.
- **Métricas de Precisão e Recall**: Medição da capacidade do sistema em recuperar informações relevantes e a proporção de resultados relevantes recuperados.
- **Métricas de Relevância e Cobertura**: Avaliação da relevância das respostas em relação às consultas dos usuários e da cobertura dos documentos disponíveis.
- **Métricas de Qualidade das Respostas**: Análise da coerência, fluência e utilidade das respostas geradas pelo sistema.
- **Feedback Iterativo**: Uso dos resultados das métricas RAGAS para refinar e melhorar o componente RAG, assegurando alinhamento com as necessidades dos usuários.

5. Testes de Aceitação pelos Usuários (UAT)

- **Sessões de Feedback com Stakeholders**: Realização de workshops e sessões práticas onde os usuários podem interagir com a solução e fornecer feedback direto.
- **Avaliação de Satisfação**: Aplicação de questionários e entrevistas para medir o grau de satisfação dos usuários com as funcionalidades implementadas.
- **Ajustes Iterativos**: Implementação de melhorias baseadas no feedback recebido, garantindo que a solução atenda às expectativas dos stakeholders.

6. Testes de Segurança e Conformidade

- **Análise de Segurança**: Identificação e correção de possíveis vulnerabilidades, assegurando a proteção dos dados e a conformidade com as políticas de segurança.

- **Testes de Penetração (Penetration Testing)**: Simulação de ataques para identificar pontos fracos na segurança do sistema.
- **Verificação de Conformidade Legal**: Garantia de que todas as funcionalidades estão em conformidade com as legislações e regulamentações aplicáveis, especialmente no que tange ao tratamento de informações sensíveis.

7. Entregáveis ao Longo do CPSI

- **Relatórios de Progresso**: Documentação regular detalhando o andamento dos testes, resultados obtidos e ações tomadas.
- **Protótipos Funcionais**: Entrega de versões preliminares da solução para avaliação e teste por parte dos stakeholders.
- **Documentação Técnica e de Usuário**: Fornecimento de manuais e guias que facilitem a compreensão e utilização das novas funcionalidades.
- **Relatório Final de Testes**: Compilação completa dos resultados dos testes, conclusões e recomendações para etapas futuras.

Conclusão

O plano de testes proposto assegurará que a solução inovadora seja rigorosamente avaliada em todas as suas funcionalidades, garantindo desempenho, escalabilidade, segurança e aderência às necessidades dos usuários. A implementação das métricas RAGAS permitirá uma avaliação detalhada do componente RAG, assegurando a qualidade e relevância das respostas geradas. Ao longo do CPSI, o foco será mantido na qualidade e na satisfação dos stakeholders, entregando uma solução robusta e eficaz.

4 - O Modelo de Negócios para CSPI é de que após cada etapa ser finalizada será entregue o pagamento e assim por diante para as etapas posteriores.

Levantamento de processos da CGE: Esta etapa visa organizar documentos e fluxos de trabalho, resultando em um relatório de levantamento. Com início em 1º de fevereiro de 2025 e conclusão prevista para 1º de abril de 2025, o custo estimado é de R$ 25.200,00.

Desenvolvimento da infraestrutura de busca: Responsável por implementar a infraestrutura de busca funcional, essa etapa inicia-se em 1º de março de 2025 e vai até 1º de agosto de 2025. Com um custo estimado de R$ 79.800,00, a entrega inclui uma solução funcional para busca eficiente de dados.

Integração com o banco de dados da CGE: A integração com o banco de dados existente, essencial para a funcionalidade completa da solução, ocorre entre 1º de fevereiro de 2025 e 3 de junho de 2025. O valor destinado a esta etapa é de R$ 63.000,00.

Importação de dados dos PDFs com OCR: Esta fase assegura a importação de dados com precisão, utilizando tecnologia OCR. Prevista para começar em 4 de fevereiro de 2025 e finalizar em 4 de fevereiro de 2025, possui um custo de R$ 47.600,00.

Desenvolvimento da interface do usuário: Para garantir uma experiência de uso eficiente, a interface será desenvolvida e testada entre 5 de abril de 2025 e 5 de julho de 2025, com um custo de R$ 49.000,00.

Otimização do Modelo de Linguagem: Essa etapa é crucial para o desempenho do modelo de linguagem, com início em 6 de abril de 2025 e conclusão em 6 de agosto de 2025. O custo é de R$ 56.000,00.

Disponibilização de endpoints: Para acesso às funcionalidades do sistema, os endpoints estarão disponíveis entre 7 de junho de 2025 e 7 de outubro de 2025, com um custo de R$ 11.200,00.

Implementação de um sistema de filas de requisições: Com foco na gestão de requisições, essa etapa ocorrerá de 8 de julho de 2025 a 8 de agosto de 2025, ao custo de R$ 16.800,00.

Especialização do produto final: A última etapa, que inclui a especialização e entrega do produto final, será realizada entre 9 de julho de 2025 e 9 de outubro de 2025, com um orçamento de R$ 51.800,00.

Custo Total: R$440.400,00 devido aos custos de cada etapa mais um valor de R$40.000,00 destinado para cobrir eventuais imprevistos.

Mas contamos com um pagamento antecipado inicial para executar a primeira etapa e dar prosseguimento às demais etapas.

Modelo de Negócios para Contrato de Fornecimento:

Será decidido durante a rodada de negociação com o contratante.

5- https://www.loom.com/share/04b40220b59b43d1af45b8f2e1934dcd?sid=d259c84c-e072-4ba9-95e4-bb05b72831ae

6-Esboço de plano de trabalho : Plano de Trabalho