Utilização de LLM com técnica RAG para Geração de Artefatos de Contratação Pública

Carlos Stucki
Centro Universitário
SENAI CIMATEC
Salvador, Brasil
stucki.carlos@gmail.com

Gerson Yamashita
Centro Universitário
SENAI CIMATEC
Lauro de Freitas, Brasil
gersonyamashita@hotmail.com

Sandro Dantas
Centro Universitário
SENAI CIMATEC
Salvador, Brasil
sandro@mpba.mp.br

Márcio Freire Cruz
Centro Universitário
SENAI CIMATEC
Salvador, Brasil
marciofreire@gmail.com

***Resumo*—** A modernização dos processos de licitação pública, por meio da digitalização e automação, é essencial para aumentar a eficiência, transparência e conformidade com as normas legais. A Lei 14.133/2021 introduziu regulamentações que demandam uma abordagem mais estruturada e detalhada na preparação de artefatos de licitação, como Estudos Técnicos Preliminares (ETP) e Termos de Referência (TR). Este projeto visa utilizar informações atualizadas e precisas na elaboração desses artefatos, aplicando um modelo de Inteligência Artificial (IA) Generativa, baseado em *Large Language Models* (LLMs) e na técnica *Retrieval* Augmented Generation (RAG), para superar os desafios na geração dos artefatos primários das contratações públicas.Os resultados obtidos demonstraram que a aplicação da abordagem proposta resulta em documentos com melhor clareza, coerência e precisão em comparação com textos gerados apenas por LLMs com documentos anexos. Avaliações feitas por servidores responsáveis pela elaboração dos documentos indicaram uma nota final média de 56/60 para RAG+LLM contra 45/60 para LLM. Este trabalho confirma que a automação inteligente pode aumentar a eficiência, precisão e conformidade dos processos licitatórios, beneficiando tanto as instituições públicas quanto a sociedade.

Palavras-chave — Contratações Públicas, LLMs, RAG, IA Generativa, DFD, ETP, TR.

I. Introdução

A ideia para este trabalho se originou da análise e das experiências vividas pelos membros deste grupo em suas respectivas áreas de atuação. Durante nossas atividades profissionais, identificamos uma grande lacuna: a falta de uma ferramenta que pudesse auxiliar os servidores encarregados de iniciar um processo de contratação via licitação. Frequentemente, esses profissionais se deparam com a ausência de um modelo a ser seguido para os serviços ou objetos a serem contratados ou adquiridos, o que dificulta significativamente a elaboração dos documentos necessários.

Os servidores envolvidos no processo de construção dos artefatos de licitação pública enfrentam consideráveis desafios para entender exatamente o que deve ser incluído nos documentos, mesmo quando dispõem de modelos padronizados que especificam todos os campos a serem preenchidos. Essa dificuldade em redigir, elaborar e estruturar os documentos de licitação acarreta um custo elevado para a instituição, tanto em termos de tempo quanto de recursos.

A elaboração inadequada dos documentos leva a atrasos. O processo de licitação, que deveria ser um procedimento eficiente e direto, torna-se um ciclo repetitivo de correções e revisões entre os diferentes setores envolvidos. Cada vez que um documento é devolvido para correção, há um atraso adicional, impactando diretamente o cronograma da contratação prevista.

Além disso, a necessidade de realizar atividades técnicas repetitivamente no mesmo processo consome um tempo valioso dos servidores. Esse tempo poderia ser direcionado para outras atividades estratégicas da instituição, como planejamento e desenvolvimento de projetos e inovação que agreguem mais valor. A ineficiência no processo não só retarda a obtenção dos bens e serviços necessários, mas também sobrecarrega os servidores, que precisam constantemente revisitar e corrigir os mesmos processos.

Outro aspecto crítico é o custo do capital intelectual empregado na análise dos documentos de licitação. Os servidores responsáveis por validar se os documentos estão em conformidade com todas as normas acabam sendo demandados em ciclos de validação até o processo estar apto para licitação. Devido às deficiências na construção dos artefatos de licitação, esses profissionais acabam dedicando uma parte significativa de seu tempo em revisões repetitivas. Como os documentos frequentemente não são bem-produzidos inicialmente, eles precisam ser devolvidos para ajustes, exigindo várias rodadas de análise minuciosa. Esse processo prolongado não só sobrecarrega os servidores, mas também aumenta os custos operacionais e consome tempo valioso que poderia ser utilizado em outras atividades mais produtivas.

A falta de uma ferramenta eficaz que auxilie na construção dos documentos de licitação compromete a eficiência operacional e a qualidade dos documentos produzidos. Documentos mal elaborados podem resultar em processos de licitação falhos, que não atendem aos requisitos legais ou que podem ser contestados, gerando ainda mais atrasos e custos adicionais para a instituição. Portanto, a identificação e a solução desse gargalo são cruciais para otimizar os processos de contratações públicas.

Para solucionar esse problema, foi desenvolvida uma ferramenta que automatiza a elaboração de Documentos de Formalização de Demanda (DFD), Termos de Referência (TR) e Estudos Técnicos Preliminares (ETP). Esta ferramenta utiliza informações atualizadas e precisas para a elaboração desses artefatos, aplicando técnicas de inteligência artificial como *Retrieval Augmented Generation* (RAG) e *Large Language Models* (LLMs.

Segundo Nóbrega e Torres (2024), a modernização dos processos de licitação pública, incluindo a implementação de plataformas eletrônicas de compras públicas (e-marketplaces), é uma tendência. No Brasil, a implementação da nova lei de licitações obrigou a reestruturação dos processos de contratações, uma vez que ela traz como um dos princípios basilares o planejamento e como um dos objetivos, o incentivo à inovação e ao desenvolvimento nacional sustentável (BRASIL, 2021). Na esteira do planejamento, a lei trouxe a obrigatoriedade de construção de artefatos de licitações com maior eficiência, redução de custos, economia de tempo, melhor comunicação entre governos, empresas e cidadãos, acesso online a serviços, transparência e menor burocracia.

A criação desta ferramenta poderá contribuir para agilizar a elaboração dos documentos de licitação, simplificando a busca das informações necessárias pelos servidores e resultando em prazos mais curtos do que os atualmente praticados. É importante ressaltar que a elaboração desses documentos deve estar em conformidade com as diretrizes estabelecidas pela Lei nº 14.133, de 1º de abril de 2021, que institui a Nova Lei de Licitações e Contratos.

Além disso, a adoção de IA generativa para automatizar a criação de artefatos de licitação pode transformar significativamente o processo de licitação pública. A IA pode reduzir o tempo necessário

para a geração de documentos, minimizar erros e inconsistências e garantir que os documentos atendam às exigências legais.

II. Trabalhos Relacionados/ Referencial Teórico

1. Compras Públicas

1.1 Inovações em compras públicas

Segundo Obwegeser e Müller (2018), inovação em compras públicas refere-se à utilização estratégica das compras governamentais para promover melhorias e introduzir novos métodos e tecnologias nos serviços públicos.

A modernização dos processos de licitação pública no Brasil tem sido marcada por diversas inovações significativas que visam aumentar a eficiência, transparência e conformidade com as normativas legais.

A implementação de portais de compras na internet, como o Comprasnet, promoveu a inclusão da administração pública no comércio eletrônico, aumentando a transparência e a participação de fornecedores nas licitações públicas (FERNANDES, 2019).

Essa inovação também facilitou a realização de pregões eletrônicos, permitindo que os fornecedores participem dos processos licitatórios de qualquer local com acesso à internet.

A criação do pregão eletrônico é uma dessas inovações, destacando-se pela simplificação dos procedimentos e pela redução de custos das compras públicas. O pregão eletrônico correspondia em 2014 a 99% do valor dos pregões da administração federal. Os dados atestam o sucesso dessa política indicando uma adaptação eficaz dos dirigentes e técnicos das unidades de compras à forma eletrônica, além de adequada oferta de treinamento dos pregoeiros e de acesso aos sistemas do Comprasnet, que operacionalizam os pregões eletrônicos na administração federal (FERNANDES, 2019).

Entre os benefícios das compras eletrônicas é possível citar: menores custos; lucros potenciais maiores (para o setor privado); maior quantidade de recursos para a execução das políticas públicas; maior alcance e liquidez de mercado; maior transparência; maior organização do mercado; eliminação de barreiras geográficas e remoção de bloqueios e canais de distribuição. As compras eletrônicas também ajudam a coordenar e organizar o processo de compras. Facilitam obter o histórico da formação de preços para cada produto, guardar informações essenciais sobre vendedores e compradores e, até mesmo, informações pós-negociação de gestão de compras que colaboram com a logística (RIBEIRO, 2009)

Ao longo do tempo, o processo de digitalização vem trazendo grandes avanços nas compras públicas. Seguindo esta linha, o objetivo deste trabalho está alinhado em promover inovação para os usuários internos de órgãos públicos, que desenvolvem o trabalho de estruturar as necessidades de contratações, diferente do que vem acontecendo ao longo dos anos, que são as inovações voltadas para integração, intermédio e interação entre órgãos públicos e fornecedores, fomentando a transparência para sociedade.

1.2 Documentos Essenciais nas Contratação Públicas

A Lei nº 14.133/2021 estabelece um conjunto de documentos essenciais que devem ser produzidos na fase interna da licitação. Esses documentos são fundamentais para assegurar a legalidade, transparência e eficiência do processo licitatório. Abaixo, apresentamos uma explicação detalhada sobre cada um desses documentos, com base nos artigos relevantes da lei.

a) Documento de Formalização da Demanda

O documento de formalização da demanda é o ponto de partida para todo o processo licitatório. Ele é elaborado pelos órgãos responsáveis pelo planejamento das contratações e tem como objetivo identificar e justificar a necessidade de determinada contratação. Esse documento é fundamental, pois dele derivam todos os outros documentos preparatórios, assegurando que a contratação atenda uma necessidade real da administração pública. Ele deve conter a descrição clara e detalhada da necessidade a ser atendida, os objetivos que se pretendem alcançar e a justificativa para a realização da contratação.

b) Plano Anual de Contratações

O plano anual de contratações é um documento que visa racionalizar e organizar as aquisições e contratações da administração pública. Conforme o Art. 12, VII, da Lei nº 14.133/2021, este plano deve ser elaborado a partir de documentos de formalização de demandas e ser alinhado ao planejamento estratégico e às leis orçamentárias dos órgãos e entidades. O objetivo é garantir que as contratações sejam feitas de maneira ordenada e eficiente, evitando desperdícios e sobreposições de gastos. Além disso, o plano anual de contratações deve ser divulgado publicamente em um sítio eletrônico oficial, assegurando a transparência do processo.

c) Estudo Técnico Preliminar

O estudo técnico preliminar é um documento que fundamenta a necessidade da contratação, caracterizando o interesse público envolvido e propondo a melhor solução para o problema identificado. Segundo o Art. 18, I, da Lei nº 14.133/2021, esse estudo deve evidenciar a viabilidade técnica e econômica da contratação. Ele deve incluir a descrição da necessidade da contratação, detalhando o problema a ser resolvido sob a perspectiva do interesse público, e demonstrar a previsão da contratação no plano de contratações anual, indicando o alinhamento com o planejamento da administração. Além disso, deve especificar os requisitos da contratação, os quais envolvem as necessidades técnicas e funcionais do objeto a ser contratado, e fornecer estimativas das quantidades necessárias, acompanhadas de memórias de cálculo e documentos justificativos. Também é essencial que o estudo contenha um levantamento de mercado, que consiste na análise das alternativas disponíveis e na justificativa técnica e econômica da escolha da solução a ser contratada.

d) Termo de Referência

O termo de referência é um documento essencial para a contratação de bens e serviços, elaborado com base no estudo técnico preliminar. De acordo com o Art. 6º, XXIII, da Lei nº 14.133/2021, ele deve conter elementos que permitam a caracterização completa do objeto a ser contratado. Este documento deve definir detalhadamente o que será contratado e fundamentar a contratação, apresentando os motivos que a justificam. Além disso, deve descrever a solução proposta, incluindo todos os aspectos relacionados ao ciclo de vida do objeto, e definir os requisitos técnicos e operacionais necessários, incluindo acessibilidade. O termo de referência também deve estabelecer os critérios de medição e pagamento, especificando como o cumprimento do contrato será medido e como os pagamentos serão realizados. Outro elemento essencial é a estimativa de preços, que deve ser baseada em pesquisa de mercado, garantindo que os valores sejam compatíveis com os praticados no mercado. A adequação orçamentária deve ser assegurada, confirmando a disponibilidade de recursos financeiros para a

contratação. Finalmente, o termo de referência deve identificar as interdependências com outras contratações que possam afetar ou ser afetadas pelo objeto contratado.

A caracterização dos documentos essenciais na fase interna da licitação, conforme estabelecido pela Lei nº 14.133/2021, fornece uma base teórica sólida para a compreensão das exigências e procedimentos necessários para a realização de contratações públicas de forma eficiente, transparente e legal. No entanto, para aprimorar ainda mais esse processo e torná-lo mais ágil e preciso, queremos incorporar novas tecnologias que possam auxiliar os usuários na construção desses documentos complexos.

2. Inovação com Inteligência Artificial

Neste contexto, segundo JEONG (2023), a utilização de tecnologias avançadas, como a Inteligência Artificial, especificamente Modelos de Linguagem de Grande Escala (LLM), como o GPT-4, combinados com a Geração Aumentada por Recuperação (RAG) , apresenta-se como uma solução inovadora e promissora. Essas tecnologias têm o potencial de transformar o modo como os documentos de licitação são elaborados, proporcionando automação, precisão e eficiência.

Ainda conforme JEONG (2023), os modelos de LLM, como o GPT-4, são capazes de entender e gerar textos complexos, oferecendo suporte na redação e revisão dos documentos, garantindo que eles atendam às exigências legais e técnicas. A tecnologia RAG permite que esses modelos de linguagem sejam alimentados com volumes de dados relevantes, melhorando a qualidade e a relevância das informações utilizadas na elaboração dos documentos.

A adoção dessas tecnologias pode revolucionar a forma como os DFDs, ETPs e TRs são produzidos, aliviando a carga de trabalho dos profissionais envolvidos e minimizando erros e inconsistências. Com essa abordagem inovadora, é possível garantir que os documentos sejam elaborados com maior rapidez e precisão, promovendo uma gestão pública mais eficiente e transparente.

Segundo Timpone e Guidi (2023), os modelos de IA generativa são capazes de criar dados que se assemelham a um conjunto de dados de treinamento. Eles são amplamente utilizados em tarefas como geração de texto, imagens e música, entre outras. O lançamento do GPT-4 pela OpenAI, do Bard pelo Google, e de ferramentas de geração de imagens como DALL-E 2, Midjourney e Craiyon, juntamente com as recentes descobertas que permitem carregar e treinar modelos de linguagem em máquinas pessoais, como o LLaMA e Alpaca usando Dalai, ainda que com algumas limitações, exemplificam a vastidão e a competitividade que impulsionarão os avanços tecnológicos dessa área.

Nesse contexto, os citados autores relatam que há dois ramos principais para a apresentação e aplicação da IA:

- **IA Analítica** é usada para tarefas como análise preditiva e reconhecimento de imagem e fala. Utilizamos a IA para gerar insights, mudar a forma como interagimos com os entrevistados e para a automação de processos. Utilizamos ferramentas específicas, como processamento de linguagem natural (PNL), transcrição de fala para texto, análise de imagens etc., em conjunto com vídeo desde a medição de audiência até a escuta social e o suporte à crises.

- **IA Generativa** é uma extensão mais recente que pode criar coisas novas em todas as mídias que, até então, eram vistas como exclusivas da inteligência e da criatividade humana: texto, vídeo, áudio, imagens - toda mídia digital pode ser alimentada pela IA generativa.

Ainda segundo Timpone e Guidi (2023), a IA Generativa já está sendo utilizada para acelerar a criação de diversos tipos de documentos, podendo, dentro do contexto deste artigo, ser utilizada para criar documentos relativos à contratação pública, como ETPs e Termos de Referência. O advento da IA logo poderá ser uma ferramenta útil para gestores e servidores designados para atuarem nos processos de contratações.

Porém, conforme Zeichick (2023), a utilização das IA generativas de forma exclusiva, especificamente dos LLMs pode trazer respostas menos precisas e contextuais, já que eles se utilizam de dados que inicialmente foram treinados, o que limita a relevância e a atualidade das informações recebidas. Para solucionar esse problema, a integração da RAG com as LLMs propõe melhorar a precisão e a relevância das respostas.

2.1 Large Language Models (LLM) ou Modelos de Linguagem de Grande Escala

De acordo com MALINEN (2024), os grandes modelos de linguagens representam uma das inovações mais significativas no campo da inteligência artificial (IA) e do processamento de linguagem natural (NLP). Esses modelos são baseados em arquiteturas de redes que permitem a análise e a geração de texto com elevada coerência e contexto. O avanço das LLMs tem sido possível graças à combinação de poder computacional aumentado e grandes volumes de dados de treinamento. Modelos como o GPT-4o, utilizados neste trabalho, possuem bilhões de parâmetros e são capazes de executar uma variedade de tarefas de linguagem natural, incluindo sumarização e geração de texto. Mas, ainda com toda essa evolução existem limitações conforme já relatado acima. Por isso, neste trabalho, resolvemos adotar o RAG como um potencializador da LLM no contexto das contratações públicas.

Segundo Fatehkia, Lucas e Chawla (2024) os LLMs representam os avanços mais recentes no PNL, demonstrando uma ampla gama de capacidades em processamento de linguagem. Eles ganharam destaque após o ChatGPT. Isso impulsionou tentativas de usar LLMs para uma variedade de aplicações, que vão desde escrita criativa, programação até domínios legais que exigem maior precisão factual.

Uma área promissora de aplicação para LLMs é a resposta à perguntas sobre documentos organizacionais proprietários, como manuais de governança/política. Tais documentos são frequentemente consultados, pois orientam as operações e tomadas de decisão diárias dentro de uma organização. Ainda para Fatehkia, Lucas e Chawla (2024), isso resulta em referências frequentes a esses documentos ou a especialistas dentro da organização que respondem a consultas sobre essas informações. Portanto, há potencial para aumentar a eficiência com uma aplicação que possa responder à uma ampla gama de consultas de usuários com base em documentos organizacionais.

Porém, mesmo com toda essa evolução, há limitações. De acordo com Fatehkia, Lucas e Chawla (2024), os modelos necessitam de grandes volumes de dados para treinamento, resultando em altos custos e tempo investido. Além disso, eles têm dificuldade em se adaptar a novos dados, o que compromete a precisão das respostas a perguntas fora do escopo dos dados originais de treinamento. Segundo a pesquisa de MAGESH (2024), trabalhos anteriores descobriram que LLMs de propósito geral alucinam na criação de artefatos jurídicos, em média, entre 58% e 82% das vezes. No entanto, a utilização de RAG, testada em diversas LLMs

disponíveis, evidenciou que as alucinações tiveram uma redução significativa, ocorrendo em média entre 17% e 33% das vezes.

2.2 Retrieval Augmented Generation – RAG ou Geração Aumentada por Recuperação

Conforme Jeong (2023), o modelo RAG (Retrieval Augmented Generation) combina técnicas de recuperação de informações e geração de texto para melhorar a precisão e relevância das respostas fornecidas por modelos de linguagem de grande escala (LLMs). O modelo de RAG melhora a capacidade dos LLMs de fornecer respostas precisas e relevantes ao buscar informações de bancos de dados previamente construídos e utilizá-las para complementar a geração de texto. Isso ajuda a mitigar problemas como a falta de dados e a tendência dos modelos a gerar respostas imprecisas ou "alucinações".

Na visão de Fatehkia, Lucas e Chawla (2024), a Geração Aumentada por Recuperação (RAG) melhora significativamente o desempenho dos Modelos de Linguagem de Grande Escala (LLM) em tarefas que necessitam de respostas precisas e contextualmente relevantes baseadas em um conjunto específico de informações ou dados especializados. Isso é feito fornecendo ao modelo uma fonte externa de informações, o que aumenta a precisão e a relevância das respostas geradas, destacando-se particularmente pela redução das alucinações do modelo. Além disso, ao recuperar informações de uma base de dados externa e usá-las como contexto para o modelo LLM, o RAG possibilita que o sistema acesse dados atualizados e específicos do tema abordado sem a necessidade de treinar novamente o modelo inteiro, tornando-o mais eficiente e robusto em contextos em que a precisão e a segurança dos dados são cruciais.

O RAG envolve duas etapas principais para responder a uma consulta: a recuperação de documentos relevantes e a geração da resposta com base nesses documentos. A recuperação seleciona documentos relevantes de um grande universo de dados, utilizando técnicas que variam de pesquisas simples por palavras-chave a algoritmos complexos de aprendizado de máquina para captar o significado semântico das consultas. Em seguida, esses documentos são fornecidos ao modelo de linguagem (LLM) junto com o texto original da consulta, permitindo que o LLM utilize essas informações para gerar uma resposta. A vantagem do RAG é que ele permite que o modelo responda de forma mais precisa ao incluir informações recuperadas diretamente no prompt, mitigando muitas alucinações comuns em LLMs genéricos (MAGESH et al., 2024).

Segundo Jeong (2023) o RAG utiliza bancos de dados vetoriais para melhorar a geração de texto em LLMs. Os dados são transformados em vetores (embeddings) e armazenados nesses bancos de dados. Quando uma consulta é feita, o RAG recupera os vetores relevantes e os utiliza como contexto para gerar respostas mais precisas e confiáveis. Isso ajuda a mitigar problemas de "alucinação" e limitações de memória de longo prazo nos LLMs, aprimorando sua utilidade prática.

Jeong (2023) descreve os bancos de dados vetoriais como uma solução moderna para a deficiência de memória de longo prazo em modelos de linguagem grande (LLMs). Esses bancos de dados são especializados em armazenar e gerenciar índices vetoriais de alta dimensão, otimizados para consultas por similaridade de vetores. A melhora na memória de longo prazo é fundamental para superar as limitações desses modelos em reter e recuperar informações relevantes de forma eficiente. Isso permite que os LLMs forneçam respostas mais precisas e contextualmente apropriadas, melhorando significativamente sua utilidade em aplicações práticas, como atendimento ao cliente e consultas baseadas em dados internos de empresas.A Figura 1 abaixo exemplifica o modelo de RAG, de uma plataforma de serviços de IA generativa. Primeiro, dados de várias fontes (como PDF, TXT, sites e YouTube) são extraídos de um banco de dados. Esses dados são processados e divididos em partes menores (chunks). Cada uma dessas partes é convertida em embeddings, que são representações numéricas de informações (como palavras, frases, ou documentos) em um espaço vetorial, usadas para que algoritmos de aprendizado de máquina ou inteligência artificial possam processar e comparar essas informações de forma eficiente.. Esses embeddings são então armazenados em um banco de vetores usando tecnologias como Chroma ou FAISS.

Quando um usuário faz uma pergunta, a plataforma busca os embeddings relevantes no banco de vetores e utiliza um modelo de linguagem (LLM) para gerar uma resposta baseada nesses embeddings. A resposta é então fornecida ao usuário através de uma interface de chatbot. Todo esse processo é orquestrado pelo LangChain, uma estrutura (framework) desenvolvida para facilitar a construção de aplicações que integram grandes modelos de linguagem (LLMs, como o GPT) com fontes externas de dados e lógica personalizada. Ele permite que você combine modelos de linguagem com pipelines complexos de processamento de dados, possibilitando a criação de sistemas que respondam a perguntas, realizem buscas em documentos e executem tarefas de maneira mais eficaz e integrada. A geração de texto é realizada por meio da interação com modelos de linguagem, como os da OpenAI, que utilizam redes neurais avançadas para compreender o contexto e gerar respostas relevantes e coerentes com base nos dados fornecidos.

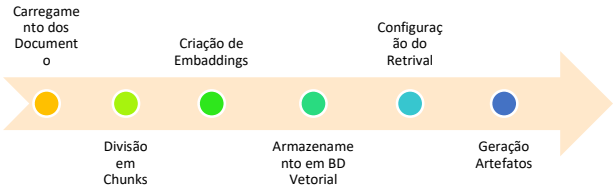


Figura 1. Fluxograma da execução do código

O modelo de RAG representa um avanço significativo na capacidade dos Modelos de Linguagem de Grande Escala (LLMs) de fornecer respostas precisas e relevantes. Ao combinar técnicas de recuperação de informações com geração de texto, o RAG aborda desafios como a falta de dados e as alucinações comuns em LLMs tradicionais. Utilizando bancos de dados vetoriais, o RAG permite a incorporação de informações contextualmente adequadas e atualizadas sem a necessidade de re-treinamento extensivo dos modelos. Isso não só melhora a precisão das respostas, mas também aumenta a eficiência e robustez dos sistemas de IA em aplicações práticas, como atendimento ao cliente e consultas empresariais. Em suma, o RAG proporciona um meio eficaz de integrar dados externos no processo de geração de texto, aprimorando a utilidade e a confiabilidade dos LLMs.

III. MÉTODOLOGIA

1. Introdução

A base epistemológica escolhida para este estudo foi a Design Science (DS), utilizando especificamente os métodos de Design Science Research (DSR). A DSR segue um processo interativo e

adaptativo que inclui a identificação de um problema, definição de objetivos, concepção e desenvolvimento de soluções, avaliação da solução e comunicação dos resultados (PEFFERS et al., 2007). A escolha da DSR se justifica por sua relevância em estudos científicos, promovendo a troca de conhecimento entre profissionais e academia e estabelecendo um processo estruturado para produzir instrumentos que resolvam problemas, visando resultados práticos e aplicáveis.

2. Identificação do Problema

Neste estudo, a DSR é aplicada para analisar o problema enfrentado pelas instituições públicas em seus processos de contratações, especialmente os relacionados à criação dos artefatos iniciais das contratações públicas. As instituições públicas, de maneira geral, enfrentam um grande problema nas contratações públicas, sendo o principal deles a fragilidade e a demora na construção dos artefatos iniciais dos processos de contratação.

3. Definição de Objetivos

O objetivo é buscar uma solução utilizando tecnologias emergentes e desenvolver um protótipo capaz de construir esses artefatos de maneira clara e concisa. A solução baseia-se em um conjunto de dados iniciais, utilizando a tecnologia RAG e grandes modelos de linguagens de inteligência artificial, como o modelo ChatGPT da OpenAI, a biblioteca LangChain, o banco de dados vetoriais Chroma e tecnologia de recuperação avançada.

4. Concepção e Desenvolvimento

O desenvolvimento do protótipo envolveu as seguintes etapas técnicas:

4.1 Configuração do Ambiente de Desenvolvimento

Inicialmente, foi realizada a configuração do ambiente de desenvolvimento, que incluiu a instalação das bibliotecas necessárias para o projeto, como langchain, openai, langchain-chroma, python-docx, pymupdf, faiss-gpu, entre outras. Além disso, foi realizada a montagem do Google Drive para permitir o acesso aos documentos armazenados.

4.2 LangChain

A biblioteca LangChain foi utilizada neste estudo como uma ferramenta central para o desenvolvimento do protótipo. LangChain permite a construção e orquestração de fluxos de trabalho envolvendo grandes modelos de linguagem (LLMs), facilitando a integração e utilização desses modelos para diversas tarefas, incluindo a geração de textos e o processamento de linguagem natural. A escolha desta biblioteca se deu por sua robustez e flexibilidade, proporcionando um ambiente ideal para o desenvolvimento do algoritmo necessário para a criação dos artefatos de contratação pública.

4.3 Carregamento e Processamento de Documentos

A segunda etapa consistiu no carregamento e processamento dos documentos. Para isso, foi desenvolvida uma função que carrega arquivos PDF e DOCX diretamente do Google Drive. A biblioteca PyMuPDF foi utilizada para a leitura e extração de texto dos arquivos PDF, enquanto a biblioteca DirectoryLoader foi utilizada para processar documentos DOCX.

4.4 Divisão de Texto, Geração de Embeddings e busca semântica

Após o carregamento dos documentos, foi necessário dividi-los em partes menores (chunks) para facilitar o processamento. Utilizou-se o RecursiveCharacterTextSplitter, uma classe disponível na biblioteca LangChain, usada especificamente para dividir os documentos em chunks menores, garantindo que cada chunk não excedesse um determinado tamanho. Em seguida, foram gerados embeddings para esses chunks utilizando o modelo OpenAIEmbeddings, o que permitiu a criação de representações vetoriais dos textos.

Essas representações são fundamentais para implementação da busca léxica permitindo que o sistema encontre informações com base em palavras-chave e correspondências exatas, enquanto a busca semântica utiliza embeddings para captar o significado contextual das consultas, garantindo que até mesmo informações relacionadas, mas não explicitamente mencionadas, sejam recuperadas. No código, o retriever exemplifica essa combinação, pois permite que acesse tanto a semântica quanto a léxica dos documentos, retornando os mais relevantes. Isso é essencial para melhorar a qualidade e relevância dos documentos gerados pelo modelo, assegurando que as respostas sejam não apenas precisas, mas também contextualmente apropriadas, mitigando problemas comuns de LLMs como "alucinações" e falta de dados.

4.5 Configuração do Banco de Dados Vetorial Chroma

Os chunks de texto gerados foram então carregados em um banco de dados vetorial Chroma. O Chroma foi escolhido devido à sua capacidade de integrar facilmente com o LangChain e oferecer uma recuperação eficiente de informações (chunks), essencial para a fase de construção dos artefatos. Embora tenhamos testado o FAISS, ele não atendeu às nossas expectativas de desempenho e integração, tornando o Chroma a melhor opção.

4.6 Desenvolvimento do Modelo de Chat e Cadeia de Recuperação

Para interagir com o usuário e recuperar informações relevantes, foi configurado um modelo de chat baseado no GPT-4, juntamente com uma memória de conversação para armazenar o histórico dos diálogos. A cadeia de recuperação conversacional foi criada para permitir a recuperação de documentos relevantes a partir do banco de dados vetorial Chroma.

4.7 Geração e Preenchimento de Documentos

A última etapa envolveu a geração e preenchimento dos documentos de contratação pública. Foram desenvolvidas funções específicas para preencher os campos dos documentos de acordo com templates pré-definidos. Além disso, foram implementadas técnicas de anonimização para proteger dados sensíveis presentes nos documentos. Desta forma, a anonimização dos dados sensíveis presentes nos documentos gerados, tem o objetivo de garantir a proteção das informações confidenciais relacionadas aos órgãos públicos e seus servidores. A anonimização foi realizada por meio de um conjunto de funções que identificam e substituem automaticamente nomes próprios, endereços e e-mails nos textos processados. O código utiliza expressões regulares (Regex) para detectar esses dados, substituindo-os por marcadores genéricos como "[NOME]", "[EMAIL]" e "[ENDEREÇO]". Dessa forma, assegura-se que as informações pessoais sejam removidas dos documentos gerados, mantendo a conformidade com os princípios de privacidade e segurança de dados. Essa abordagem é fundamental ao lidar com documentos de diferentes órgãos públicos, garantindo que nenhum dado sensível seja exposto inadvertidamente. O conteúdo gerado foi então salvo em formato DOCX, garantindo a conformidade com as diretrizes e regulamentações aplicáveis.

Essas etapas foram essenciais para a concepção e desenvolvimento de um protótipo eficiente e funcional, capaz de automatizar a geração de artefatos de contratação pública, atendendo às necessidades das instituições e garantindo a conformidade com as legislações vigentes.

4.8 Testes de Soluções e Refinamento do Código

4.8.1 Carregamento de Documentos

Originalmente, o carregamento de documentos utilizava o DirectoryLoader para suportar arquivos em formato genérico. Entretanto, foi implementada uma função personalizada load_documents, que passou a suportar arquivos PDF, utilizando a biblioteca PyMuPDF e arquivos DOCX. Essa mudança proporcionou uma maior flexibilidade no carregamento de diferentes tipos de documentos, ampliando a aplicabilidade do código em diversos contextos documentais e garantindo um suporte mais robusto e específico para os formatos necessários.

4.8.2 Divisão de Texto, Overlap e Embeddings

O processo de divisão de textos foi outra área de evolução significativa. Inicialmente, o código utilizava o CharacterTextSplitter com um chunk_size de 1500. Durante o processo de melhoria do código, foi introduzido o RecursiveCharacterTextSplitter com um chunk_size aumentado para 1900, e o overlap foi ajustado de 50 para 250. Essa evolução começou com testes em tamanhos menores de chunks e progrediu até encontrar o equilíbrio ideal em 1900, com um overlap maior, garantindo uma maior assertividade na geração de documentos. O aumento do overlap permitiu que as informações fossem melhor preservadas na continuidade do texto, evitando a perda de dados críticos, especialmente em documentos longos e complexos, como Termos de Referência (TRs) com muitos itens.

Além disso, para melhorar ainda mais a precisão na divisão e recuperação de textos, foram testados diferentes embeddings, e optamos por utilizar o 'text-embedding-3-large', por ser a opção mais avançada e eficaz disponível. Os embeddings permitem que o sistema capture melhor as nuances semânticas do texto, assegurando que as divisões de texto e as subsequentes recuperações de informações sejam mais relevantes e alinhadas ao conteúdo original.

4.8.3 Configuração do Modelo de Chat e Recuperação de Informações

Outra mudança importante foi a configuração do modelo de chat. Inicialmente, o modelo ChatOpenAI era utilizado com uma temperatura de 1.0, mas durante os testes, a temperatura foi reduzida para 0.5, e o modelo começou a ser testando o com o GPT-3.5 Turbo, avançando para o GPT-4 e, finalmente, evoluindo para o GPT-4o, que se mostrou o mais adequado para nossas necessidades. Essas modificações foram feitas para melhorar o equilíbrio entre criatividade e precisão, essencial para a geração de documentos legais, garantindo que o contexto extraído do repositório fosse mais relevante e preciso.

A cadeia de recuperação de informações também foi otimizada. Inicialmente, utilizávamos o tipo de cadeia "stuff", mas durante os testes passamos a utilizar "map_reduce", um método mais sofisticado e eficiente para combinar respostas em grandes volumes de texto. Essa mudança resultou em uma melhoria na eficiência e na precisão da recuperação de informações, especialmente útil na análise de textos extensos.

4.8.4 Funções de Anonimização e Preenchimento de Documentos

As funções de anonimização não tiveram muitas mudanças, garantindo a conformidade com normas de privacidade e segurança de dados. No entanto, o preenchimento de documentos foi ampliado e refinado. Ao final dos testes, os templates foram expandidos para incluir novos tipos de documentos, como "TR_AQUISIÇÃO" e "TR_AQUISIÇÃO_SERVIÇOS_COMUNS", com adaptações específicas para estar em conformidade com a demanda de cada tipo de documento.

Essa evolução permitiu uma maior adaptação às necessidades organizacionais e regulatórias, resultando em uma maior aplicabilidade e conformidade dos documentos gerados. Além disso, foi introduzida uma função de formatação de documentos DOCX, que garante que os documentos sejam bem formatados, facilitando o manuseio posterior do usuário.

4.8.5 Desafios e Soluções

Ao longo do desenvolvimento, enfrentamos vários desafios que moldaram a evolução do código. Um exemplo notável foi a dificuldade inicial em utilizar a biblioteca Faiss para leitura de documentos em PDF. Essa abordagem mostrou-se inadequada, levando à substituição pelo LangChain, que se demonstrou mais eficaz. Além disso, identificamos a necessidade de zerar as variáveis do sistema ao gerar novos tipos de documentos para evitar erros na recuperação de informações do ChromaDB, um problema que ainda requer solução definitiva.

Outro aspecto importante foi o ajuste da temperatura do modelo, que foi reduzida para tornar as respostas mais alinhadas ao repositório, reduzindo a criatividade em prol da precisão. Embora isso possa resultar em uma redução de 80% a 90% no trabalho manual necessário no futuro para a conclusão do tipo de documento, ainda assim será necessário ajustar o documento final para garantir sua completa adequação. Foi identificada a necessidade de criar templates específicos para cada tipo de documento, bem como instruções mais elaboradas, para ter um resultado mais aprimorado nos objetos mais complexos. Já os objetos mais simples foram resolvidos de forma eficiente pelo código, com a instrução padrão.

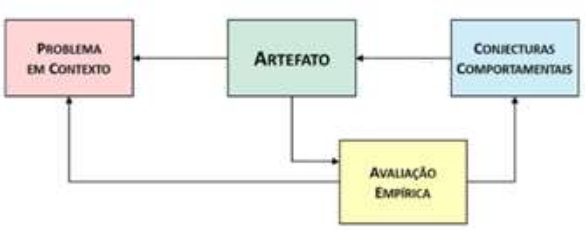
4.8.6 Conclusão

A versão final do código apresenta melhorias significativas em termos de organização, eficiência e aplicabilidade. As mudanças implementadas não apenas refinaram as funcionalidades existentes, mas também introduziram novas capacidades, tornando o sistema mais robusto, escalável e alinhado com as necessidades específicas do projeto e das regulamentações em questão. Essas alterações demonstram uma evolução do código em direção a um grau de maturidade mais elevado, comparável ao Nível 9 da TRL (Technology Readiness Level), tornando-o adequado ao contexto legal e técnico de sua aplicação. Isso reflete a complexidade e o rigor exigidos em um projeto voltado para a automação de processos em licitações e contratos administrativos.

5. Avaliação da Solução

No modelo DSR, destacam-se os **critérios de verificação** para validação do artefato e os **critérios de aceitação** para indicar se o problema foi resolvido ou mitigado. Esses critérios são importantes para avaliar se o produto da pesquisa resolve satisfatoriamente o problema em questão. Todavia devido ao tempo previsto não houve a formalização do critério de aceitação.
A Figura 2 revela os elementos centrais envolvidos no percurso metodológico da DSR.

Figura 2 – Elementos Centrais do Modelo DSR



Fonte: PIMENTEL, FILIPO, SANTOS, 2020.

6. Avaliação Empírica

Outra etapa importante deste modelo é a avaliação empírica, que é decomposta em duas vertentes – questões e hipóteses de pesquisa para aceitação do artefato e questões e hipóteses para análise relacionada às conjecturas comportamentais.

5.1.1 Comunicação dos Resultados

Os elementos centrais da DSR são definidos a seguir:

- **Artefato** – Um protótipo de uma solução que recebe um conjunto de documentos referenciais e constrói, baseados em template previamente definidos, os seguintes artefatos: Documento de Formalização da Demanda (DFD), Estudo Técnico Preliminar (ETP) e Termo de Referência.
- **Problema de contexto** – As instituições públicas de maneira geral enfrentam grandes problemas nas contratações públicas e os principais deles são a fragilidade e a demora na construção dos artefatos iniciais dos processos de contratação.
- **Problema de pesquisa** – Verificar como o conjunto de Tecnologia RAG e LLM combinados podem construir artefatos robustos que reduzam o processo de contratação e evitem erros recorrentes.
- **Conjecturas comportamentais** – A aplicação do protótipo nas instituições públicas levará a: redução de tempo e esforço gastos na construção dos artefatos iniciais das contratações públicas, sejam eles DFD, ETP e TR; aumento da eficácia e eficiência; maior eficiência na utilização de recursos e na entrega de resultados e diminuição do ruido nos processos de comunicação entre as diversas áreas envolvidas.

7. Campo

Inicialmente o objeto de estudo escolhido para este trabalho foi constituído pelas contratações realizadas pelo Ministério Público da Bahia (MPBA), mas verificamos, em virtude da atuação em rede no MPBA, a possibilidade de fazer pequenas avaliações com outros Ministérios Públicos.

A escolha do MPBA, entidade como foco da investigação, deveu-se a vários fatores. Em um segundo momento, verificamos que poderiam ser ampliados os testes para além do MPBA. Pelo fato do Ministério Público brasileiro atuar em redes de cooperação através do seu Conselho Nacional, em especial pelos seus comitês temáticos, foi possível a partir de uma interlocução com o Comitê de Políticas de Gestão Administrativa, testar o protótipo no Ministério Público do Trabalho (MPT), Ministério Público Federal (MPF) e no Ministério Público do Distrito Federal e Territórios (MPDFT).

Em resumo, o MPBA é apropriado para esta pesquisa porque oferece elementos necessários para avaliação do protótipo, uma gama diversificada de artefatos referências e uma clara estrutura de contratação bem instituída e de fácil acesso. Isso permite analisar e entender melhor as necessidades e os desafios enfrentados por organizações similares e como elas podem melhorar suas operações por meio do uso mais eficaz nos seus processos de contratação.

8. Sujeitos e Coleta de Dados

Os sujeitos da pesquisa foram os servidores responsáveis pela elaboração e validação dos documentos de licitação nas instituições participantes. Para a coleta de dados, foi realizado um levantamento junto aos Ministérios Públicos (MPs) sobre os artefatos que eles gostariam que fossem gerados pelo protótipo.

9. Análise de Dados

Para a análise de dados, foi realizado um levantamento junto aos Ministérios Públicos (MPs) para identificar os artefatos que eles gostariam que fossem gerados pelo protótipo. Após esse levantamento, foi aplicado um questionário para avaliação dos artefatos gerados, considerando as seguintes variáveis:

1. Clareza2. Coerência3. Precisão
4. Completude
5. Relevância
6. Gramática/Estilo

A partir dessa análise, foi possível compreender que a utilização das tecnologias propostas é viável para o objetivo pretendido. Constatamos que a aplicação das tecnologias poderia ser imediata nos órgãos, embora sejam necessários ajustes futuros, especialmente nos documentos referenciais necessários.

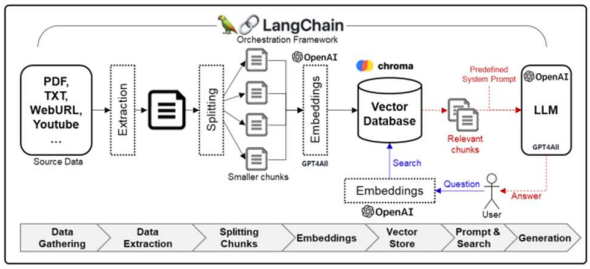
Vale ressaltar que, por questão metodológica e de tempo da pesquisa, a classificação dos itens acima foi avaliada de 0 a 10 por cada órgão, sendo discricionário de cada organização avaliar esses resultados, uma vez que já existem estudos que analisam textos gerados pelas LLMs que não foram considerados para elaboração deste trabalho.

9. Implementação do Código

Do ponto de vista da implementação do código, o protótipo segue um fluxo detalhado que integra diversas tecnologias e bibliotecas para alcançar os objetivos propostos. A implementação começa com a configuração do ambiente de desenvolvimento, seguida pelo carregamento e processamento de documentos em diferentes formatos. Esses documentos são então divididos em partes menores e transformados em embeddings para facilitar a recuperação e análise. A biblioteca LangChain desempenha um papel central na construção e orquestração dos fluxos de trabalho envolvendo grandes modelos de linguagem. Utilizando a recuperação conversacional e a geração de textos, o protótipo preenche automaticamente os templates dos artefatos de contratação pública. Todo o processo é coordenado e monitorado para garantir conformidade com as diretrizes legais e eficiência na geração dos documentos necessários. A Figura 3 ilustra o fluxograma de execução do código, demonstrando o fluxo de trabalho do LangChain para o processamento de dados e geração de respostas com modelos de linguagem. Inicialmente, os dados são extraídos de diferentes fontes, como PDFs, arquivos TXT, URLs da web, entre outros. Esses dados são então divididos em pedaços menores, chamados *chunks*. Cada um desses chunks é convertido em embeddings, ou seja, representações vetoriais que capturam o significado dos textos, utilizando modelos como os da OpenAI.

Esses embeddings são armazenados em um banco de dados vetorial, como o Chroma, permitindo que as informações sejam recuperadas de forma eficiente. Quando uma pergunta é feita pelo usuário, o sistema busca os chunks mais relevantes no banco de dados e utiliza um **prompt de sistema predefinido** (*predefined system prompt*) – uma estrutura de comando ou instrução previamente configurada para guiar o modelo – para fornecer contexto ao modelo de linguagem. Com base nisso, o modelo (como o GPT da OpenAI) gera uma resposta coerente e precisa, detalhando assim cada etapa do processo de implementação do protótipo:

Figura 3. Modelo RAG



Fonte: JEONG (2023)

IV. RESULTADOS E DISCUSSÃO

Uma das análises interessantes foi estruturar com os mesmos parâmetros, isto é, o mesmo documento, o mesmo modelo da LLM, e posteriormente submeter a uma avaliação com base nos mesmos parâmetros de clareza, coerência, precisão, completude, relevância e gramática/estilo.

Na Tabela 1 há um comparativo do texto de um dos tópicos de um documento obrigatório nas contratações públicas, gerado pela solução proposta neste trabalho (Texto 1), com o texto gerado diretamente pelo ChatGPT 4o (Texto 2). Na Coluna 1 da Tabela 1, temos a combinação da técnica de Geração Aumentada por Recuperação (RAG) + LLM (API Model gpt-4º, *Temperature* = 0.5). Nesta abordagem, a API do modelo GPT-4 foi configurada com um parâmetro de temperatura de 0.5, o que determina a aleatoriedade/criatividade das respostas geradas, buscando um equilíbrio entre criatividade e precisão. Na Coluna 2, utilizamos o modelo GPT-4o "Ommi" disponível na interface do ChatGPT, fornecendo documentos anexos como referência para a geração de texto. Embora ainda se baseie no poderoso modelo de linguagem GPT-4o, essa técnica depende da capacidade do modelo de interpretar e integrar os documentos fornecidos no contexto das respostas geradas, sem o benefício de uma recuperação estruturada de informações como no método RAG.

TABELA 1 COMPARATIVO PARA GERAÇÃO DE MODELOS.

Texto 1	Texto 2
1. DESCRIÇÃO DA NECESSIDADE DA CONTRATAÇÃO A presente contratação é necessária para atender às demandas urgentes do Ministério Público do Estado da Bahia (MPBA) no que tange à modernização e fortalecimento das suas capacidades operacionais e tecnológicas. Atualmente, o MPBA enfrenta uma série de desafios significativos devido à defasagem tecnológica e à insuficiência de recursos estruturais que impactam diretamente na eficiência e eficácia dos serviços prestados à população. A situação atual é marcada por sistemas de informação obsoletos, equipamentos de informática inadequados e infraestrutura física que não atendem as necessidades contemporâneas de um órgão que lida com questões complexas de justiça e cidadania. Esses problemas resultam em atrasos processuais, dificuldades na gestão de casos e, consequentemente, comprometem o bem-estar e o interesse público. Quantitativamente, o déficit tecnológico afeta diretamente mais de 300 unidades do MPBA espalhadas pelo estado, incluindo Promotorias e Procuradorias de Justiça, que enfrentam dificuldades operacionais diárias. Qualitativamente, a falta de modernização compromete a celeridade e a transparência dos processos, essenciais para a confiança da sociedade no Sistema de Justiça. Esforços anteriores para resolver essas questões incluíram tentativas de atualização de sistemas e aquisição de novos equipamentos, mas tais medidas foram insuficientes devido à limitação de recursos e a fragmentação das iniciativas. Portanto, a contratação proposta visa implementar uma solução integrada e abrangente, alinhada com o Plano Nacional de Segurança Pública (Lei 11.530/2007) e em conformidade com a Lei 14.133/2021, garantindo a estruturação adequada e o reaparelhamento das unidades do MPBA, além da valorização dos profissionais que atuam na instituição.	**1. DESCRIÇÃO DA NECESSIDADE DA CONTRATAÇÃO** A contratação de empresa especializada para fornecimento e atualização de licença perpétua do ArcGIS Enterprise é essencial para atender às necessidades do Ministério Público do Estado da Bahia (MPBA). Atualmente, a falta de atualização e suporte técnico do software compromete a eficiência na análise e gerenciamento de dados geoespaciais, impactando negativamente o apoio às investigações ambientais realizadas pelo Centro de Apoio Operacional às Promotorias de Justiça do Meio Ambiente (Ceama). A aquisição visa solucionar problemas de manutenção e atualização do software, proporcionando uma ferramenta robusta para análises espaciais avançadas, suporte técnico especializado e treinamento. Esta ação está alinhada com a Iniciativa Estratégica do PCA de 2024, código 55-CEAMA, que visa o aperfeiçoamento de medidas de estruturação do geoprocessamento.

Essa contratação é essencial não só para a melhoria das condições de trabalho dos servidores, mas também para assegurar que o MPBA continue a cumprir seu papel constitucional de defender os direitos dos cidadãos e promover a justiça de forma eficiente e eficaz. A proposta está em conformidade com o art. 40 da Lei nº 14.133/2021, que estabelece que o planejamento de compras deve considerar a expectativa de consumo anual e atender aos princípios da padronização, garantindo compatibilidade de especificações estéticas, técnicas ou de desempenho. Além disso, alinha-se com as diretrizes do Plano de Contratações Anual (PCA) do MPBA, visando à eficiência, economicidade e melhoria contínua dos serviços prestados à sociedade.	

Os textos gerados foram avaliados com base em seis critérios principais: clareza, coerência, precisão, completude, relevância e gramática/estilo. A avaliação foi realizada pelo servidor responsável pela elaboração dos documentos de licitação nas instituições participantes. Os resultados das avaliações foram resumidos na Tabela 2 destacando as diferenças de desempenho entre as duas abordagens.

TABELA 2 COMPARATIVO DO RESULTADO DAS AVALIAÇÕES.

Texto 1	Texto 2
Clareza: O texto é claro e detalhado, explicando a necessidade da contratação de forma precisa. Nota: 9/10 **Coerência:** A argumentação é bem estruturada e os parágrafos seguem uma sequência lógica. Nota: 9/10 **Precisão:** O texto fornece dados específicos sobre os problemas enfrentados pelo MPBA e a quantidade de unidades afetadas. Nota: 9/10 **Completude:** A descrição cobre todos os aspectos importantes da necessidade da contratação, incluindo desafios atuais, tentativas anteriores e benefícios esperados. Nota: 10/10 **Relevância**: Todo o conteúdo é relevante e diretamente relacionado ao tema da contratação e às necessidades do MPBA. Nota: 10/10 **Gramática/Estilo:** A gramática é correta e o estilo é formal e adequado para o contexto institucional. Nota: 9/10 Nota Final: 56/60	**Clareza:** O texto é claro, mas menos detalhado comparado ao primeiro. Nota: 7/10 **Coerência:** A estrutura do texto é coerente, mas poderia ser mais elaborada. Nota: 8/10 **Precisão**: O texto menciona a falta de atualização e suporte técnico, mas não oferece a mesma quantidade de detalhes específicos como o primeiro. Nota: 7/10 **Completude**: A descrição é mais breve e não cobre todos os aspectos da necessidade de forma tão abrangente quanto o primeiro texto. Nota: 7/10 **Relevância**: O conteúdo é relevante, mas não tão detalhado quanto o primeiro texto. Nota: 8/10 **Gramática/Estilo**: A gramática é correta e o estilo é formal, porém, há algumas redundâncias que poderiam ser evitadas. Nota: 8/10 Nota Final: 45/60

Em um segundo teste com arquivos diferentes e objetos de contratação mais complexos, reduzimos a temperatura do modelo para 0,3. Esse ajuste foi realizado para aumentar a precisão do modelo na geração de documentos, garantindo que as respostas fossem mais alinhadas ao conteúdo do repositório e menos influenciadas por variações criativas.

Durante esses testes, é importante destacar que, devido ao uso de modelos e dados provenientes de outros órgãos, não foi possível apresentar os documentos gerados, pois foi necessário manter a confidencialidade desses materiais.

Além disso, verificamos que o protótipo conseguiu gerar os documentos completos previstos, como os Documentos de Formalização da Demanda (DFDs), Estudos Técnicos Preliminares (ETPs) e Termos de Referência (TRs). Os resultados indicaram que o protótipo foi capaz de gerar ETPs e TRs em sua totalidade, conforme esperado. Nesse sentido, conforme a Tabela 3, a avaliação foi a seguinte:

TABELA 3 COMPARATIVO DO RESULTADO DAS AVALIAÇÕES COM ARQUIVOS DIFERENTS.

Texto 1	Texto 2
-Clareza:9/10 -Coerência:9/10 -Precisão:9/10 -Completude:10/10 -Relevância:10/10 -Gramática/Estilo:9/10 Nota Final: 56/60	-Clareza:8/10 -Coerência:8/10 -Precisão:8/10 -Completude: 8/10 -Relevância: 9/10 -Gramática/Estilo: 8/10 Nota Final: 49/60

O texto gerado pela combinação de RAG+LLM (Texto 1) é mais detalhado, completo e preciso, atendendo melhor aos parâmetros de avaliação dos textos gerados por LLMs. O segundo texto, gerado apenas com LLM e documentos anexos no ChatGPT (Web), é mais breve e menos detalhado, resultando em uma avaliação inferior.

Além dos resultados dos textos, submetemos à avaliação dos gestores públicos do Ministério Público Federal (MPF), Ministério Público do Distrito Federal e Territórios (MPDFT) e consultores especialistas na área.

O secretário de administração do MPF do Piauí relatou a seguinte conclusão:

- Clareza: 10
- Coerência: 10
- Precisão: 6
- Completude: 4
- Relevância: 6
- Gramática/Estilo: 7

O mesmo disse: “De maneira geral, deu um bom início, mas ainda tem um bocado de trabalho para fazer. Mas acho que esse trabalho dado é realmente mais desafiador que o normal. Para um ser humano já não é fácil sintetizar uma contratação com DEMO. Para um algoritmo, então... Foi interessante a busca dele pelos normativos de sustentabilidade do MPF. Mas o posicionamento não ficou o melhor possível (e até errado). O uso desses normativos tem condicionantes e aplicações certas. Então, é uma parte que vai dar um pouco mais de trabalho para acertar... Nos deu um bom início para o TR”.

Já o secretário de administração do MPDFT fez o seguinte relato: Vou te passar o feedback sobre seu robô: ele faz muito bem os documentos, porém a impressão é de que viaja em algumas informações técnicas (ele fala muito mais do que o esperado). Eu não sei se o objeto que te passei foi mais complexo e confundiu o robô ou se o robô é prolixo mesmo, mas ele ajudou em 80% do trabalho. Eu acho uma solução espetacular! Melhora muito o serviço”.

V. CONCLUSÃO

A implementação de tecnologias avançadas, como a combinação de Modelos de Linguagem de Grande Escala (LLMs) e a técnica de Geração Aumentada por Recuperação (RAG), mostrou-se uma solução promissora para a automação e otimização da criação de artefatos de contratação pública. Este estudo destacou diversas conclusões, desafios e possibilidades futuras que podem guiar novas pesquisas e desenvolvimentos.

A criação de bases de dados auxiliares contendo pareceres saneadores é fundamental. Esses pareceres podem identificar e resolver os principais problemas enfrentados durante a produção manual de artefatos, oferecendo um banco de conhecimento consolidado para referência futura. Essas bases de dados devem ser constantemente atualizadas para refletir as melhores práticas e mudanças regulatórias. Além disso, a criação de uma base de dados robusta contendo artefatos referenciais organizados por temática pode ser extremamente útil. Isso facilitaria a geração de novos documentos ao proporcionar exemplos claros e contextualmente relevantes, diminuindo o tempo necessário para a elaboração e aumentando a precisão e conformidade dos documentos. A implementação de mecanismos de busca léxica e semântica é crucial no modelo de recuperação de informações. Esses mecanismos permitem uma recuperação mais eficiente e precisa dos dados necessários, melhorando significativamente a qualidade e relevância dos documentos gerados.

Durante o desenvolvimento deste projeto, algumas dificuldades foram observadas. A integração de diferentes tecnologias, como LLMs e RAG, apresentou desafios técnicos significativos. A necessidade de harmonizar a recuperação de dados com a geração de texto coerente e contextualmente relevante demandou ajustes constantes e uma compreensão profunda dos algoritmos subjacentes. Garantir a precisão e confiabilidade das informações geradas foi um desafio constante. As "alucinações" das LLMs, nas quais o modelo gera respostas factualmente incorretas, precisaram ser mitigadas por meio da combinação com técnicas de recuperação de informações específicas e supervisão humana. A conformidade com a Lei 14.133/2021 e outras regulamentações específicas foi uma preocupação central. A complexidade e o detalhamento exigidos pelos artefatos de licitação pública demandaram uma adaptação cuidadosa das tecnologias para garantir a aderência às normas legais.

Para avançar nas pesquisas e melhorar as soluções propostas, algumas direções podem ser exploradas. Criar um modelo de avaliação robusto para soluções que combinam LLMs e RAG é essencial. Esse modelo deve medir a eficácia, precisão e eficiência na criação de artefatos de licitação, fornecendo dados quantificáveis para aprimoramento contínuo. Integrar novas funcionalidades, como a análise preditiva de riscos e sugestões automatizadas de melhorias para os documentos gerados, pode aumentar ainda mais a utilidade das ferramentas desenvolvidas. Trabalhar na interface do usuário para tornar as ferramentas mais intuitivas e acessíveis aos servidores públicos pode melhorar significativamente a adoção e eficácia das tecnologias implementadas. Ampliar os testes e validações para além do Ministério Público da Bahia (MPBA) pode fornecer insights valiosos sobre a adaptabilidade e escalabilidade da solução desenvolvida.

Este projeto demonstrou que a aplicação de LLMs combinada com a técnica RAG tem um grande potencial para revolucionar a criação de artefatos de contratação pública. Apesar dos desafios enfrentados, os resultados são promissores e apontam para um futuro em que a automação inteligente pode aumentar a eficiência, precisão e conformidade dos processos licitatórios, beneficiando não apenas as instituições públicas, mas também a sociedade como um todo.

VI. REFERÊNCIAS

BAYAZIT, Nigan. Investigating design: A review of forty years of design research. Massachusetts Institute of Technology: **Design Issues**, v. 20, n. 1, p. 16-29, 2004.

BRASIL. Lei nº 14.133, de 1º de abril de 2021. Institui a nova Lei de Licitações e Contratos Administrativos. Diário Oficial da União: seção 1, Brasília, DF, 1º abr. 2021.

DRESCH, A.; LACERDA, D. P.; MIGUEL, P. A. C. Uma Análise Distintiva entre o Estudo de Caso, A Pesquisa-Ação e a Design Science Research. Revista Brasileira de Gestão de Negócios, v. 17, n. 56, p. 1116-1133, 2015. Disponível em: https://doi.org/10.7819/rbgn. v17i56.2069. Acessado em: 8 jul 2024.

FATEHKIA, Masoomali; LUCAS, Ji Kim; CHAWLA, Sanjay. T-RAG: lessons from the LLM trenches. **arXiv preprint arXiv:2402.07483**, 2024.

FERNANDES, Ciro Campos Christo. Compras Públicas no Brasil: Tendências de inovação, avanços e dificuldades no período recente. **Administração Pública e Gestão Social**, v. 11, n. 4, 2019.

JEONG, Cheonsu. Generative AI service implementation using LLM application architecture: based on RAG model and LangChain framework. **Journal of Intelligence and Information Systems**, v. 29, n. 4, p. 129-164, 2023.

MALINEN, Esko. Interactive document summarizer using LLM technology. 2024.

MAGESH, Varun et al. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. arXiv preprint arXiv:2405.20362, 2024.

MENON, Kunal. Utilizing Open-Source AI to Navigate and Interpret Technical Documents: leveraging RAG models for enhanced analysis and solutions in product documentation. 2024.

NÓBREGA, Marcos; DE TORRES, Ronny Charles L. THE NEW PROCUREMENT IN BRAZIL AND E-MARKETPLACE: THE TURNING POINT OF INNOVATION IN PUBLIC ACQUISITIONS, 2024.

OBWEGESER, Nikolaus; MÜLLER, Sune Dueholm. Innovation and public procurement: Terminology, concepts, and applications. **Technovation**, v. 74, p. 1-17, 2018.

PEFFERS, Ken et al. A design science research methodology for information systems research. **Journal of management information systems**, v. 24, n. 3, p. 45-77, 2007.

PIMENTEL, Mariano; FILIPPO, Denise; SANTOS, Thiago Marcondes. Design Science Research: pesquisa científica atrelada ao design de artefatos. **RE@D - Revista de Educação a Distância e eLearning**, v. 3, n. 1, p. 37-61, 2020.

RIBEIRO, Manuella Maia**. Como os estados brasileiros promovem a transparência nos portais de compras eletrônicas?.** 2009.

TIMPONE, GUIDI; GUIDI, Michel. EXPLORANDO A MUDANÇA DE CENÁRIO DA IA.Da IA Analítica à IA Generativa, p. 2023-05, 2023.

ZEICHICK, A. What Is Retrieval-Augmented Generation (RAG)? 2023. Disponível em: https://www.oracle.com/artificial-intelligence/generative-ai/retrieval-augmented-generation-rag/. Acesso em: 7 jun. 2024.

Sistema FIEB | SENAI CIMATEC
PELO FUTURO DA INOVAÇÃO

CENTRO UNIVERSITÁRIO SENAI CIMATEC
ESPECIALIZAÇÃO EM DATA SCIENCE & ANALYTICS

ATA DE APRESENTAÇÃO DE PROJETO FINAL DE CURSO

Ata de apresentação do Projeto Final de Curso "**Utilização de LLM com técnica RAG para Geração de Artefatos de Contratação Pública"**, submetido pelo aluno **Gerson Adriano Yamashita**, como parte dos requisitos para obtenção do Certificado de **Especialista em Data Science & Analytics** pelo Centro Universitário SENAI CIMATEC, às 14h30 do dia 29 de agosto de 2024. Reuniu-se remotamente pela plataforma Microsoft Meet, a Banca Examinadora designada pelo Prof. Dr. Márcio Freire Cruz - Orientador, constituída pelo Prof. Dr. Márcio Freire Cruz e pelo Prof. Dr. Oberdan Rocha Pinheiro. O orientador deu início aos trabalhos com as devidas orientações, e a exposição foi realizada pelo estudante dentro do prazo de tempo estabelecido. Ao final da apresentação a banca reuniu-se atribuindo a seguinte nota: 9,6 (nove pontos e seis décimos).

A banca de avaliadores decidiu pela:

(X) Aprovação do trabalho
Caberá ao aluno apresentar em no máximo em 30 (trinta) dias a contar da data de assinatura desta Ata, uma cópia do trabalho em PDF, constando as considerações pontuadas pela banca. A Ata de Apresentação do Projeto Final de Curso deve ser digitalizada e inserida na terceira página do TCC ou como anexo do artigo.

( ) Reprovação do trabalho
O aluno terá que se matricular novamente no TCC – Trabalho de Conclusão de Curso e ser submetido a uma banca avaliadora no semestre seguinte**.**

As ações consequentes ao status de Aprovação deverão obedecer ao prazo proposto acima sob pena do parecer final ser modificado para o status de Reprovado automaticamente e sem possibilidade de recurso.

Para constar, lavrou-se a presente ata que vai assinada por todos os membros da Banca. Por estarem cientes de suas obrigações estão de acordo com os termos desse documento:

Salvador, 29 de agosto de 2024

Assinado eletronicamente por:
Marcio Freire
CPF: ***.235.375-**
Data: 30/08/2024 18:38:57 -03:00

Márcio Freire Cruz
Professor Orientador

Assinado eletronicamente por:
Oberdan Rocha Pinheiro
CPF: ***.073.705-**
Data: 31/08/2024 07:13:47 -03:00

Oberdan Rocha Pinheiro
Membro da banca

Assinado eletronicamente por:
Patricia Freitas Tourinho
CPF: ***.733.265-**
Data: 30/08/2024 17:27:33 -03:00

Patrícia Freitas Tourinho
Coordenadora do Pós-Graduação Lato Sensu