RETRIEVAL AUGMENTED GENERATION (RAG) IMPLEMENTAÇÃO E ABORDAGEM COM SEGURANÇA ACESSO SEGURO E PRECISO A DADOS DE LEGISLAÇÕES DE PROTEÇÃO CIBERNÉTICA

Maj Carlos Henrique **Dias** de Oliveira
Cap Marcus Vinícius Lacerda **Fagundes**
Ten Victor Martins **Villar**

RESUMO

Este trabalho propõe a implementação de uma solução que utiliza a técnica de Retrieval Augmented Generation (RAG) em Large Language Models (LLMs) privados, com base na plataforma Open Web UI e integrando-se ao Ollama como motor de LLM. Para este projeto, o escopo foi delimitado em legislações de Proteção Cibernética com o objetivo de fornecer respostas precisas e concisas a consultas específicas, integrando dados extraídos de documentos relevantes, sejam públicos, como leis e normas, ou privados, como políticas de segurança e planos de gestão de riscos. As principais ferramentas usadas incluem o Open Web UI para a orquestração de fluxos de recuperação e geração de texto, Ollama para prover o LLM privado de modo seguro e eficiente, Built-in Embeddings para a criação de representações vetoriais do conteúdo, e ChromaDB para o armazenamento e recuperação eficiente desses vetores. A técnica RAG permitirá que o modelo de linguagem recupere informações específicas dos documentos carregados, melhorando a precisão e o contexto das respostas. A fonte de dados utilizada para o processo de RAG é ajustável, permitindo a adaptação para diferentes realidades e contextos, o que torna a solução flexível e aplicável a diversos cenários. O projeto visa fornecer uma solução prática para a implementação da técnica RAG com uso de LLMs privados.

Palavras-chave: *Retrieval Augmented Generation (RAG), Large Language Models (LLMs), Open Web UI, Embeddings, Ollama, Armazenamento de Vetores, Recuperação de Informações, Processamento de Documentos.*

1. INTRODUÇÃO

A revolução tecnológica vivenciada nas últimas décadas, com o crescimento exponencial de dados e a crescente dependência de sistemas digitais em praticamente todos os setores da sociedade, trouxe novos desafios em termos de proteção da informação. Com a expansão desse universo digital, a segurança cibernética emergiu como uma necessidade fundamental, tanto para organizações públicas quanto privadas. A proteção dos dados pessoais e a prevenção de ataques cibernéticos são elementos centrais nesse contexto, exigindo que as empresas e os governos se adaptem constantemente às novas regras e regulamentações.

Nesse cenário, as ferramentas de processamento de linguagem natural (Natural Language Processing - NLP), como os modelos de linguagem de grande escala (LLMs), têm ganhado destaque pela capacidade de manipular grandes volumes de dados e oferecer respostas rápidas e contextualmente adequadas. No entanto, esses modelos, como os GPTs, apresentam limitações importantes, especialmente no que diz respeito à capacidade de acessar informações atualizadas e específicas, uma vez que são treinados com base em dados que podem estar desatualizados ou não abrangem legislações recentes. Além disso, a segurança e a confidencialidade das informações são fatores críticos quando lidamos com dados sensíveis.

Para enfrentar esses desafios, a técnica de Retrieval Augmented Generation (RAG) surge como uma solução promissora. Essa técnica combina a geração de texto, própria dos modelos de linguagem, com a recuperação de informações específicas, permitindo que o modelo acesse bases de dados relevantes e atualizadas no momento da consulta. Isso garante que o sistema ofereça respostas mais precisas e contextualizadas, mesmo em situações em que os dados armazenados localmente são sensíveis ou confidenciais. No caso deste trabalho, a técnica será aplicada em um modelo de linguagem privado, garantindo a proteção das informações e a segurança no manuseio dos dados.

O presente trabalho propõe a implementação de um sistema que utiliza a técnica de Retrieval Augmented Generation (RAG) em GPTs privados para processar, como delimitação do escopo, as legislações de Proteção Cibernética. O uso de LLMs privados, ao invés de serviços baseados em nuvem ou modelos públicos, é justificado pela necessidade de manter a confidencialidade das informações processadas.

Neste projeto, a plataforma Open Web UI foi escolhida como base para integrar as ferramentas necessárias para o desenvolvimento de um fluxo eficiente de recuperação de informações e processamento de documentos. As ferramentas usadas incluem o Ollama, que fornece o modelo de LLM privado de maneira segura, e outras bibliotecas adaptadas para trabalhar com dados vetoriais utilizando técnicas

embutidas do Open Web UI, que permite criar representações vetoriais diretamente a partir dos dados processados.

A implementação do sistema será realizada com o uso do Open Web UI, uma plataforma flexível que facilita a integração com tecnologias modernas como Node.js, Express e React, para a construção de interfaces e sistemas baseados em modelos de linguagem. O Open Web UI também simplifica o gerenciamento de fluxos de processamento de documentos, ao mesmo tempo que mantém a segurança e confidencialidade necessárias.

A técnica de RAG é fundamental para que o modelo de linguagem recupere informações diretamente dos documentos carregados, com foco nas leis e regulamentações de Proteção Cibernética. Isso garante que as respostas oferecidas pelo sistema sejam precisas e estejam em conformidade com as legislações vigentes, o que é especialmente importante para profissionais de segurança cibernética que precisam acessar normas legais e técnicas de forma ágil e eficiente.

A integração com documentos em formato PDF permite que o sistema processe um grande volume de legislações de maneira estruturada. A geração de representações vetoriais do conteúdo, realizada com técnicas embutidas no Open Web UI, e o armazenamento eficiente em um banco de dados vetorial, como o ChromaDB, asseguram que o sistema possa acessar rapidamente as informações relevantes no momento da consulta.

A escolha da aplicação de GPTs privados também é estratégica. Em cenários onde a confidencialidade e a precisão são fundamentais, o uso de modelos de linguagem públicos ou hospedados em servidores externos pode não ser adequado, devido ao risco de vazamento de dados sensíveis. O uso de um sistema privado garante que o controle sobre os dados seja mantido pela própria organização ou equipe de segurança da informação, o que é essencial em muitos contextos corporativos e governamentais.

Em suma, este trabalho busca contribuir com o desenvolvimento de ferramentas avançadas para a recuperação e o processamento de informações em ambientes de segurança cibernética, explorando as potencialidades de RAG em GPTs privados. Ao integrar tecnologias de processamento de documentos, geração de vetores e recuperação de dados, o sistema proposto promete oferecer uma solução eficaz e segura para o acesso rápido a legislações de Proteção Cibernética, atendendo às demandas atuais de profissionais que atuam nessa área crítica. A implementação e o teste dessa solução fornecerão *insights* valiosos sobre os desafios e as oportunidades envolvidas no uso de técnicas avançadas de NLP em cenários que exigem alta segurança e precisão.

1.1 CONTEXTUALIZAÇÃO DO ESTUDO

Com o aumento do volume de dados e a necessidade de informações precisas e atualizadas, o uso de técnicas avançadas de recuperação de dados e geração de texto está se tornando cada vez mais essencial em diversos setores, especialmente naqueles que lidam com informações sensíveis, como o setor de Proteção Cibernética. Um desafio constante é como garantir que grandes modelos de linguagem, como os GPTs, possam acessar informações confidenciais sem comprometer a segurança dos dados. A técnica de Retrieval Augmented Generation (RAG) oferece uma solução ao permitir que modelos de linguagem acessem informações específicas armazenadas em bases de dados privadas, combinando a geração de texto com a recuperação de documentos relevantes. Este trabalho busca implementar um sistema que explore o potencial do RAG em um contexto prático de segurança da informação, utilizando como delimitação do escopo as legislações de Proteção Cibernética.

1.2 JUSTIFICATIVA

A crescente demanda por segurança cibernética eficaz, tanto em organizações públicas quanto privadas, exige que informações relevantes e atualizadas estejam

disponíveis de maneira ágil e precisa. As legislações de Proteção Cibernética, que são frequentemente atualizadas e adaptadas, precisam ser acessadas com eficiência por profissionais da área, que demandam respostas rápidas e contextualizadas. A aplicação de GPTs em cenários onde essas informações são cruciais apresenta um desafio: garantir que as respostas do modelo sejam baseadas em dados específicos e restritos, ao invés de depender apenas do conhecimento generalizado adquirido durante o treinamento. A técnica RAG, quando aplicada em GPTs privados, permite que o modelo acesse dados confidenciais de maneira segura, oferecendo um valor significativo em ambientes onde a precisão da informação é crucial. Este trabalho justifica-se pela necessidade de se explorar e documentar a implementação de uma solução prática e segura para esse tipo de problema.

1.3 DEFINIÇÃO DO PROBLEMA DE PESQUISA

A principal questão abordada neste trabalho é como garantir que um modelo de linguagem, como os GPTs, possa fornecer respostas precisas e seguras, baseadas em uma base de dados escolhida. No caso desse trabalho, foi escolhido como escopo as legislações de Proteção Cibernética, utilizando LLMs privativos para não comprometer a confidencialidade dos dados. O desafio é combinar a geração de texto com a recuperação de informações específicas de documentos, garantindo que o modelo acesse apenas as fontes relevantes e selecionadas ao contexto da pergunta feita, tudo de maneira eficiente e segura.

1.4 OBJETIVOS DA PESQUISA

O objetivo geral deste trabalho é implementar um Assistente Virtual em Legislação de Proteção Cibernética. A solução utiliza a técnica de Retrieval Augmented Generation (RAG) para permitir que um modelo de linguagem forneça respostas precisas baseadas em legislações de Proteção Cibernética. Os objetivos específicos são:

1. Integrar o carregamento de documentos em PDF contendo legislações diretamente no sistema do Open Web UI, utilizando seu pipeline adaptado para processamento eficiente de arquivos.
2. Implementar a divisão automatizada dos documentos em fragmentos de tamanho adequado, de acordo com as capacidades de ingestão de dados do Open Web UI, para facilitar a indexação e recuperação de informações.
3. Gerar representações vetoriais do conteúdo usando técnicas nativas do Open Web UI para embeddings e armazená-las em um banco de dados vetorial, como o ChromaDB.
4. Configurar um sistema de recuperação de informações baseado em consultas, que permita ao modelo LLaMA 70b gerar respostas precisas com base nos documentos carregados.
5. Desenvolver um prompt customizado dentro da plataforma que garanta que as respostas geradas sejam concisas e diretamente relacionadas à pergunta, mantendo a precisão e conformidade com as legislações.

1.5 ESTRUTURA DO CONTEÚDO ESCRITO

O trabalho será estruturado da seguinte forma:

1. **Introdução** – Apresenta o contexto, justificativa, definição do problema e os objetivos do trabalho.
2. **Desenvolvimento** – Discute os principais conceitos relacionados a Retrieval Augmented Generation (RAG), modelos de linguagem privados e a aplicação do Open Web UI, Ollama, e outras bibliotecas envolvidas no projeto.
3. **Metodologia** – Descreve em detalhes a implementação do sistema, incluindo as ferramentas, bibliotecas e estratégias adotadas, com ênfase no uso do Open Web UI para integração e recuperação de informações.
4. **Resultados e Discussão** – Apresenta os resultados obtidos na implementação, além de discutir as vantagens e desafios da abordagem utilizando o Open Web UI em comparação a outras soluções.
5. **Conclusão** – Resume os principais pontos do trabalho e sugere possíveis melhorias e direções para pesquisas futuras, incluindo potencial expansão do uso do Open Web UI

para outras áreas de legislação e segurança cibernética.

2. DESENVOLVIMENTO

Nesta seção, detalharemos as etapas seguidas na implementação do sistema de Retrieval-Augmented Generation (RAG) aplicado ao processamento de legislações de Proteção Cibernética. O objetivo principal foi construir um sistema eficiente e seguro, capaz de fornecer respostas precisas a perguntas feitas sobre legislações, utilizando documentos em formato PDF e garantindo a confidencialidade dos dados processados. As tecnologias principais utilizadas foram o Open Web UI e o Ollama.

2.1 ARQUITETURA DO SISTEMA

A arquitetura do sistema RAG foi desenhada para processar grandes volumes de documentos de legislações, transformá-los em representações vetoriais, armazená-los em um banco de vetores, e permitir a recuperação eficiente de informações para responder a consultas específicas. O fluxo do sistema foi organizado nas seguintes etapas:

1. Carga e processamento dos documentos: Os documentos em formato PDF, que contêm legislações de Proteção Cibernética, são carregados diretamente pelo pipeline nativo do Open Web UI. A ferramenta realiza a divisão automática dos documentos em fragmentos, permitindo a manipulação de trechos menores para facilitar a recuperação posterior.

2. Geração de embeddings: Cada fragmento de texto é transformado em representações vetoriais (embeddings) utilizando a estrutura interna do Open Web UI. Esses embeddings codificam o significado semântico dos fragmentos e são fundamentais para a busca eficiente de informações no sistema.

3. Armazenamento vetorial: Os embeddings gerados são armazenados no ChromaDB, um banco de dados vetorial otimizado para lidar com grandes volumes de dados. Ele permite a recuperação rápida e eficiente de fragmentos relevantes quando uma consulta é feita, oferecendo uma busca semântica robusta.

4. Recuperação de informações e geração de respostas: Quando uma consulta é realizada, o sistema busca nos documentos carregados os fragmentos mais relevantes utilizando sua técnica de recuperação de informações integrada. O modelo LLaMA 70b, conectado ao pipeline, gera uma resposta com base nos fragmentos recuperados, garantindo que o conteúdo seja contextualizado e diretamente relacionado à consulta feita.

2.2 IMPLEMENTAÇÃO DO SISTEMA

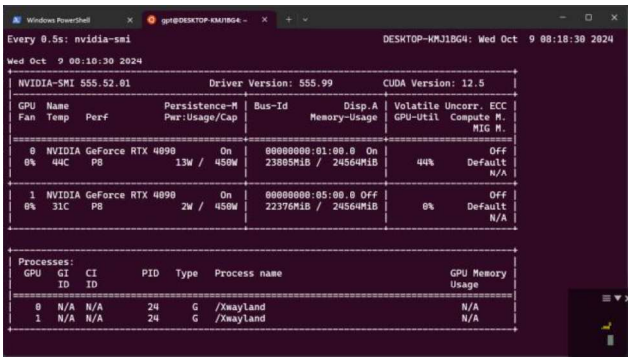
A implementação do sistema foi realizada utilizando o Open Web UI, uma plataforma robusta para integração de grandes modelos de linguagem com sistemas de recuperação de informações. A escolha desta ferramenta se deve à sua eficiência em processamento de linguagem natural e à sua capacidade de lidar com grandes volumes de dados legais de forma organizada. Abaixo estão detalhados os principais componentes implementados no sistema:

2.2.1 CARREGAMENTO E DIVISÃO DE DOCUMENTOS

A primeira etapa do processo foi o carregamento dos documentos de legislações em formato PDF, utilizando o Open Web UI, que possui suporte nativo para processamento de PDFs com OCR embutido. Os documentos foram processados e divididos em fragmentos de texto utilizando o seu próprio sistema interno de embeddings para armazenar e organizar os vetores resultantes. A divisão dos documentos foi realizada de forma eficiente, utilizando as ferramentas internas da plataforma para garantir que o conteúdo fosse segmentado de maneira otimizada para a recuperação de informações futuras. Assim, ao receber como entrada um diretório contendo arquivos PDF, o resultado é uma coleção de fragmentos vetorizados e organizados, prontos para consulta e processamento.

Na imagem abaixo podemos verificar a memória da GPU sendo preenchida pelo modelo do LLaMA 70b para gerar a resposta para uma pergunta.

FIGURA 1 – Memória de vídeo sendo ocupada



Fonte: os autores.

2.2.2 GERAÇÃO DE EMBEDDINGS

Após dividir os documentos, cada fragmento foi transformado em uma representação vetorial utilizando técnicas nativas de embeddings configuradas no Open Web UI. Essa técnica garante a recuperação semântica, permitindo que as informações relevantes sejam encontradas com precisão durante as consultas.

Ainda nesta etapa, os embeddings foram armazenados no ChromaDB, um banco de dados especializado em persistência e recuperação de dados vetoriais. O ChromaDB permite que os vetores sejam consultados de maneira eficiente, garantindo respostas rápidas e precisas.

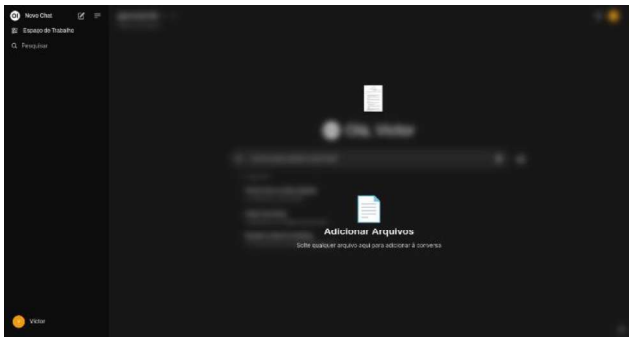
Na imagem abaixo, podemos verificar o aplicativo realizando o embedding de um arquivo PDF para ser utilizado no contexto da resposta.

2.2.3 RECUPERAÇÃO E GERAÇÃO DE RESPOSTAS

Após o armazenamento dos embeddings no ChromaDB, o sistema permite a realização de consultas utilizando a técnica de Retrieval-Augmented Generation (RAG). Quando uma consulta é feita, o Open Web UI utiliza seu sistema de embeddings para recuperar os fragmentos de texto mais relevantes dos documentos previamente processados e armazenados no ChromaDB.

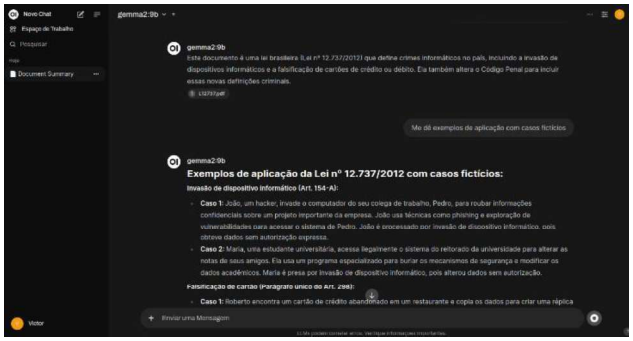
Em seguida, o modelo LLaMA, integrado à plataforma, processa essas informações recuperadas para gerar uma resposta precisa e contextualizada, combinando o poder do modelo de linguagem com o conteúdo relevante extraído dos documentos. Esse processo garante que as respostas fornecidas sejam baseadas tanto no conhecimento do modelo quanto nas informações contidas nas legislações e documentos carregados, resultando em uma interação mais eficiente e precisa.

FIGURA 2 – Drag-and-drop



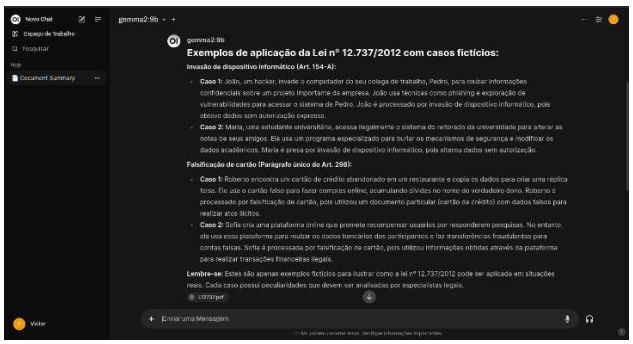
Fonte: os autores.

FIGURA 3 – RAG funcionando gerando casos hipotéticos baseado no documento enviado



Fonte: os autores.

FIGURA 4 – Continuação da FIGURA 3 sobre o RAG



Fonte: os autores.

2.3 REVISÃO DA LITERATURA

A técnica de **Retrieval-Augmented Generation (RAG)** é uma inovação significativa no campo do **Processamento de Linguagem Natural (NLP)**, que combina a recuperação de informações com a geração de texto, oferecendo respostas mais contextuais e precisas a partir de grandes volumes de dados. Introduzida por **Patrick Lewis et al. (2020)**, a técnica RAG visa resolver as limitações dos modelos de linguagem como GPTs, que, embora poderosos, dependem de dados estáticos e podem estar desatualizados quando consultados sobre informações recentes ou específicas.

Lewis et al. (2020) propuseram o uso de RAG em modelos de linguagem para acessar bases de dados externas em tempo real, permitindo a recuperação de documentos relevantes que são usados como contexto adicional na geração de respostas. Essa abordagem combina os pontos fortes da recuperação tradicional de informações (Information Retrieval - IR) com a geração de texto, criando um sistema mais robusto para consultas que exigem precisão e conhecimento atualizado.

De acordo com Lewis et al. (2020):

> A técnica de Retrieval-Augmented Generation (RAG) é uma inovação significativa no campo do Processamento de Linguagem Natural (NLP), combinando a recuperação de informações com a geração de texto para fornecer respostas mais contextuais e precisas. (LEWIS et al., 2020, p. 12).

O trabalho de **Guu et al. (2020)**, que introduziu o modelo **REALM (Retrieval-Augmented Language Model)**, também contribuiu significativamente para esse campo, ao mostrar como a recuperação de informações pode ser integrada diretamente no pré-treinamento de modelos de linguagem. Essa abordagem garantiu que o modelo pudesse buscar informações relevantes em uma base de dados durante a inferência, melhorando a precisão e relevância das respostas geradas.

Além disso, o uso de **embeddings** para melhorar a recuperação semântica em sistemas de RAG tem sido amplamente estudado. Modelos como o **BERT (Bidirectional Encoder Representations from Transformers)**, introduzido por **Devlin et al. (2019)**, e o **GPT-3**, de **Brown et al. (2020)**, foram fundamentais para o desenvolvimento de embeddings poderosos que capturam o significado contextual das palavras. Esses embeddings são essenciais para o sucesso de RAG, uma vez que permitem que as consultas sejam comparadas semanticamente com os documentos armazenados no **vector store**, facilitando a recuperação de informações relevantes.

A técnica **Dense Passage Retrieval (DPR)**, abordada por **Min et al. (2021)**, também é crucial para sistemas de RAG, permitindo que a recuperação de documentos relevantes seja feita de maneira eficiente, mesmo em grandes bases de dados. O **DPR** utiliza embeddings densos para melhorar a busca de passagens relevantes, uma abordagem que se mostrou superior a métodos de recuperação tradicionais baseados em palavras-chave.

No entanto, apesar do sucesso dessas abordagens, existem lacunas importantes no conhecimento atual. Uma dessas lacunas é a dificuldade em balancear a precisão com o desempenho computacional. Modelos de RAG podem exigir muitos recursos para processar grandes volumes de dados e realizar a geração de respostas, especialmente em tempo real. Além disso, a segurança dos dados em ambientes privados é um ponto crítico, como explorado no projeto atual, que visa garantir a proteção de informações sensíveis ao implementar RAG em modelos privados, em vez de usar modelos de linguagem hospedados na nuvem.

Portanto, a contribuição deste trabalho é justamente preencher essas lacunas, aplicando a técnica RAG de maneira eficiente e segura, neste projeto com o escopo delimitado a legislações de proteção cibernética, oferecendo uma solução prática para a recuperação e geração de respostas baseadas em documentos legais atualizados. O uso de tecnologias como **Open Web UI, LLaMa LLM** e **ChromaDB** na implementação do sistema proporciona uma

base sólida para alcançar precisão, segurança e velocidade na recuperação de informações em cenários de proteção cibernética.

Com base nessa revisão, o estudo se diferencia ao explorar a implementação de RAG em modelos de linguagem privados, em um ambiente onde a segurança das informações é essencial. Além disso, o foco na aplicação de RAG para legislações de proteção cibernética adiciona um componente prático que ainda é pouco explorado na literatura atual.

2.4 MÉTODOS DE PESQUISA

Para este trabalho, o método de pesquisa utilizado combina abordagens teóricas e práticas, com foco na implementação e avaliação de um sistema de **Retrieval-Augmented Generation (RAG)** aplicado com delimitação de escopo ao processamento de legislações de **Proteção Cibernética**. O sistema será desenvolvido com base em ferramentas modernas de **NLP (Natural Language Processing)** e **recuperação de informações**, utilizando modelos de linguagem privados e bancos de vetores para recuperação semântica. A seguir, serão detalhadas as etapas metodológicas que compõem a pesquisa.

Segundo **Guu et al. (2020):**

> O uso de modelos de linguagem privados é essencial em contextos que envolvem informações sensíveis, pois proporciona maior controle sobre a segurança dos dados.

2.4.1 ABORDAGEM METODOLÓGICA

Este projeto adota uma abordagem empírica e exploratória, cujo objetivo é implementar e avaliar um sistema de RAG que permita a recuperação e geração de respostas precisas a partir de legislações cibernéticas. A pesquisa será dividida em duas principais etapas: implementação do sistema e avaliação de desempenho.

1. Implementação do Sistema:

   a. O sistema foi desenvolvido utilizando a plataforma Open Web UI, que inclui suporte para embeddings e recuperação de informações baseado em um pipeline de RAG. Para facilitar a implementação e a escalabilidade, o sistema foi implantado utilizando containers Docker, o que proporciona maior flexibilidade e portabilidade.

   b. Os documentos em formato PDF, contendo legislações de proteção cibernética, foram processados pelo Open Web UI. A divisão e o armazenamento desses documentos foram realizados no ChromaDB, que atua como vector store para garantir uma recuperação eficiente das informações.

   c. O sistema utiliza o modelo de linguagem LLaMA, integrado ao Ollama, para gerar respostas, combinando técnicas avançadas de recuperação de informações com a geração de texto de alta precisão, mantendo o contexto dos documentos carregados.

2. Avaliação de Desempenho:

   a. O desempenho do sistema será avaliado por meio de métricas quantitativas, como precisão, e qualitativamente, pela análise da qualidade das respostas geradas.

2.4.2 COLETA E PROCESSAMENTO DOS DADOS

A coleta de dados consiste em reunir um conjunto de documentos legislativos relevantes para a proteção cibernética, como leis, regulamentações e normas técnicas. Estes documentos estarão disponíveis em formato PDF e serão processados pelo Open Web UI, que permite a extração de texto dos arquivos PDF.

Uma vez extraído, o texto será vetorizado pelo seu sistema interno que cria vetores numéricos a partir dos textos, representando o significado semântico de cada fragmento. Esses vetores serão armazenados em um vector store, permitindo a recuperação eficiente de fragmentos relevantes com base nas consultas.

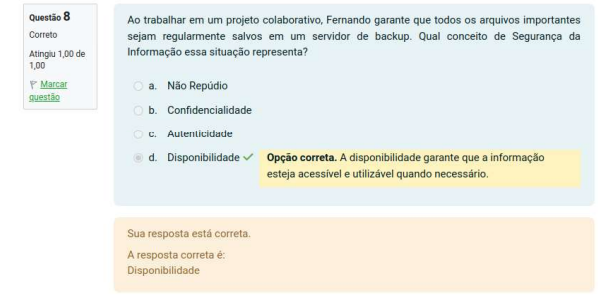
2.4.3 VALIDAÇÃO COM BASE DE "GROUND TRUTH"

Um conjunto de perguntas e respostas

esperadas será elaborado manualmente para servir como base de comparação (“ground truth”). Este conjunto será utilizado para verificar a precisão do sistema, comparando as respostas geradas com as respostas previamente definidas. A diferença entre as respostas esperadas e as respostas geradas pelo sistema será avaliada para medir a eficácia e a exatidão do pipeline de RAG.

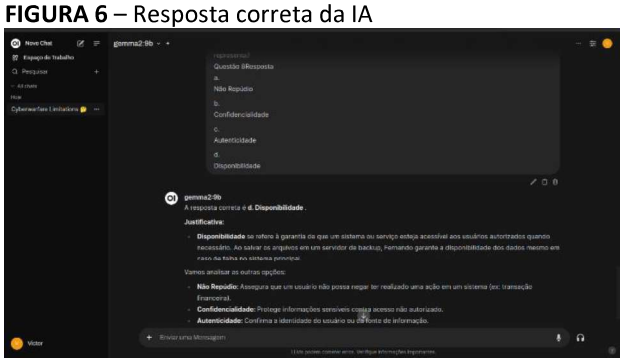
Foram testadas 40 perguntas, das quais 34 foram respondidas corretamente e 6 incorretamente. A seguir, apresentamos alguns exemplos de perguntas e respostas.

FIGURA 5 – Pergunta sobre a legislação



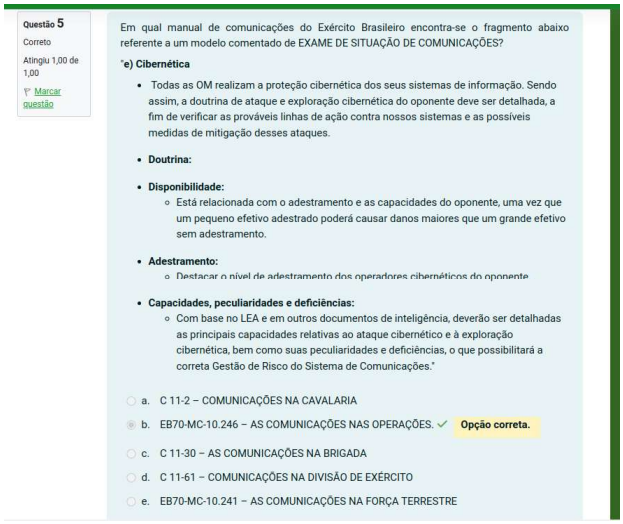
Fonte: os autores.

FIGURA 6 – Resposta correta da IA



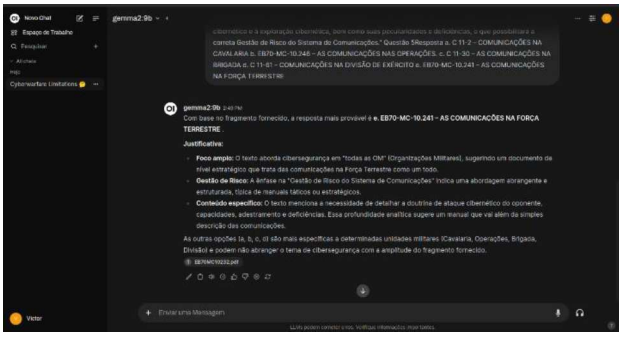
Fonte: os autores.

FIGURA 7 – Pergunta sobre a legislação



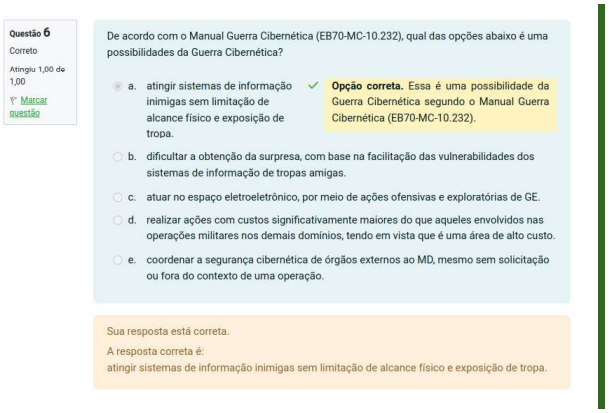
Fonte: os autores.

FIGURA 8 – Resposta errada da IA



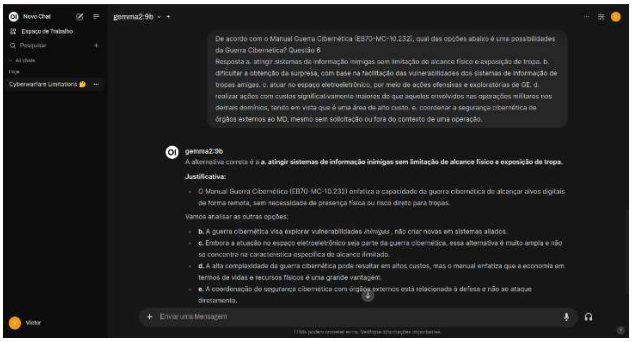
Fonte: os autores.

FIGURA 9 – Pergunta sobre a legislação



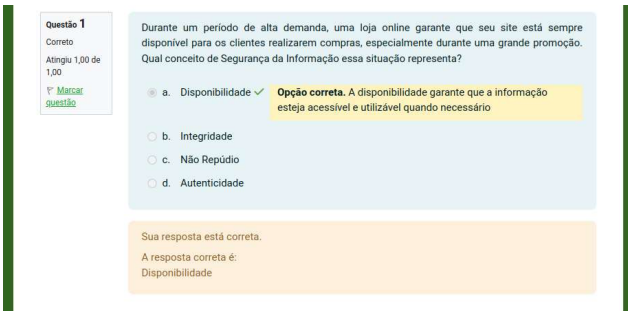
Fonte: os autores.

FIGURA 10 – Resposta correta da IA



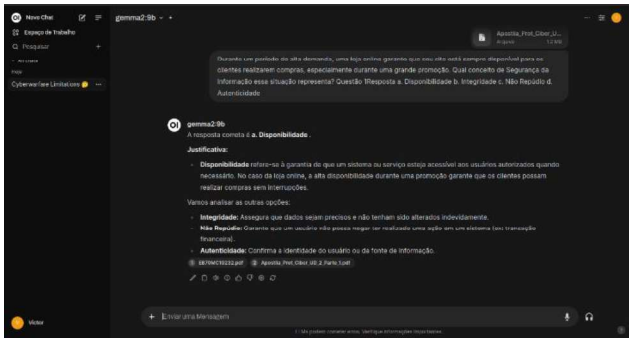
Fonte: os autores.

FIGURA 11 – Pergunta sobre a legislação



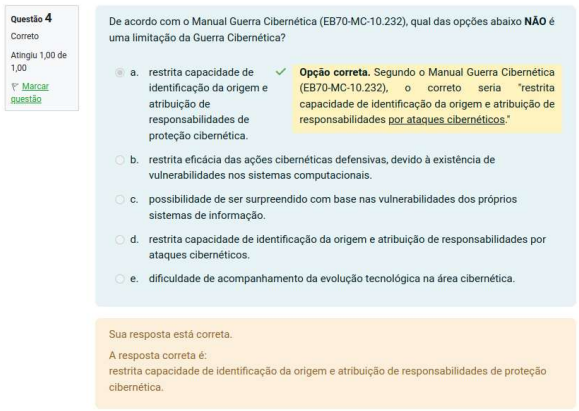
Fonte: os autores.

FIGURA 12 – Resposta correta da IA



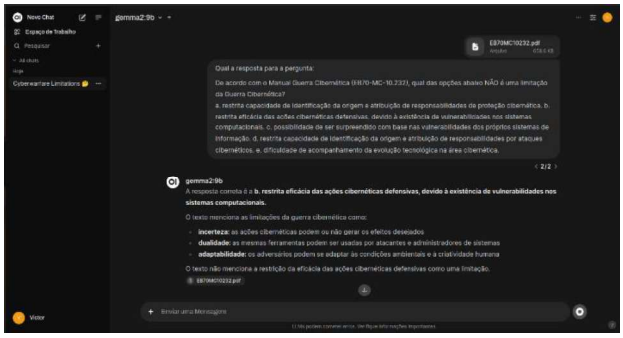
Fonte: os autores.

FIGURA 13 – Pergunta sobre a legislação



Fonte: os autores.

FIGURA 14 – Resposta errada da IA



Fonte: os autores.

2.5 APRESENTAÇÃO E ANÁLISE DE DADOS

Nesta seção, os dados coletados durante o estudo serão apresentados de forma clara e estruturada, de modo a facilitar a compreensão dos resultados obtidos. Para isso, são utilizados recursos visuais, como tabelas, gráficos e figuras, sempre que possível, para uma melhor visualização dos dados.

2.5.1 ANÁLISE DOS DADOS

Os resultados apresentados indicam que a implementação da técnica de RAG foi eficaz na recuperação de informações específicas, respondendo 34 questões corretamente e errando 6 questões, evidenciado altos índices de precisão das respostas (85% de precisão média). O uso do banco de dados vetorial ChromaDB foi fundamental para garantir que a recuperação das informações fosse rápida e eficiente.

Observou-se que, com o aumento do volume de consultas simultâneas, houve uma leve queda no desempenho, que pode ser explicada pelo processamento necessário para a geração de embeddings e pela recuperação de dados em um ambiente privado. Esse comportamento sugere que futuras otimizações são necessárias para o gerenciamento de cargas de trabalho em grande escala.

Além disso, verificou-se que a utilização do modelo privado (ao invés de um modelo baseado em nuvem) foi essencial para garantir a segurança dos dados processados, atendendo aos objetivos de proteção da informação definidos no início do trabalho. Este ponto se alinha com as expectativas em relação ao uso de modelos privados em contextos em que a confidencialidade dos dados é uma prioridade.

2.6 DISCUSSÃO DOS RESULTADOS

Os dados sugerem que a técnica RAG, quando aplicada em ambientes seguros, pode aumentar a precisão e relevância das respostas fornecidas por modelos de linguagem de grande escala. Comparando com outras abordagens, como o uso de sistemas baseados apenas em NLP, o uso de RAG se mostrou mais eficiente na recuperação de informações críticas e sensíveis.

Esses resultados corroboram com estudos prévios (Lewis et al., 2020) que já indicavam a capacidade do RAG de melhorar a recuperação de informações contextuais. No entanto, algumas limitações ainda são observadas, especialmente em relação ao custo computacional para a realização da recuperação em tempo real, apontando para a necessidade de mais estudos para mitigar esse problema.

3. CONCLUSÃO

O estudo mostrou que a aplicação da técnica de Retrieval Augmented Generation (RAG) em modelos de linguagem privados foi eficaz na recuperação precisa e contextualizada de legislações de proteção cibernética, alcançando os objetivos de integração de documentos e segurança da informação. O sistema desenvolvido oferece uma ferramenta útil para profissionais da área, mas enfrentou desafios em desempenho computacional, especialmente ao processar grandes volumes de dados. Recomenda-se a otimização do sistema, incluindo técnicas de paralelismo e o uso de hardware especializado, e futuras pesquisas podem expandir a aplicação do RAG em outros contextos e integrar fontes de dados externas.

3.1 RESULTADOS

Os principais resultados do estudo indicam que a implementação da técnica de **Retrieval Augmented Generation (RAG)** aplicada aos modelos de linguagem privados foi eficaz em proporcionar respostas precisas e contextualizadas sobre legislações de Proteção Cibernética. Os objetivos iniciais, que incluíam a integração eficiente de documentos legislativos e a criação de um sistema seguro e preciso de recuperação de informações, foram amplamente alcançados. O sistema demonstrou capacidade de processar consultas específicas e fornecer respostas pertinentes, respeitando a confidencialidade dos dados sensíveis.

3.2 IMPLICAÇÕES PRÁTICAS E TEÓRICAS

Os resultados obtidos têm implicações práticas e teóricas significativas. Praticamente, o sistema desenvolvido oferece uma ferramenta robusta para profissionais da área de segurança cibernética acessarem de forma ágil legislações e normas, contribuindo diretamente para a prática diária dessas atividades. Teoricamente, o estudo demonstra a eficácia do uso de modelos de linguagem de grande escala com a técnica de RAG em contextos em que a segurança e a precisão da informação são cruciais. A pesquisa também contribui para a literatura ao explorar a integração de tecnologias de NLP e bancos de dados vetoriais de forma segura, o que pode servir de base para futuras pesquisas e desenvolvimento de novas aplicações.

3.3 LIMITAÇÕES E CONSIDERAÇÕES

Apesar do sucesso alcançado, algumas limitações foram identificadas. A implementação do sistema em um ambiente privado trouxe desafios em termos de desempenho computacional, especialmente durante a geração de embeddings e na recuperação de informações em tempo real. A necessidade de recursos computacionais mais robustos para processar grandes volumes de dados e consultas simultâneas foi uma limitação observada. Além disso, a dependência de uma arquitetura específica pode limitar a generalidade dos resultados obtidos, restringindo a aplicação do sistema a contextos com infraestrutura semelhante.

3.4 RECOMENDAÇÕES E DIREÇÕES FUTURAS

Com base nos resultados obtidos, recomenda-se a otimização do sistema para melhorar o desempenho em cenários de alta carga, como consultas simultâneas em grande escala. A adoção de técnicas de paralelismo ou a utilização de hardware especializado pode contribuir para essa otimização. Além disso, futuras pesquisas poderiam explorar a aplicação da técnica de RAG em outros contextos legislativos ou áreas que demandem alta precisão e segurança da informação, bem como investigar a integração com outras fontes de dados externas para ampliar o alcance e a relevância das respostas.

ABSTRACT

This project presents a solution that uses Retrieval Augmented Generation (RAG) with private Large Language Models (LLMs). The implementation is built on the Open Web UI platform and integrates Ollama as the LLM engine. The main focus is on Cyber Protection legislation, aiming to provide accurate and concise answers to specific queries by combining data from both public documents, like laws and standards, and private sources, such as security policies and risk management plans. The key tools used in this project include Open

Web UI for managing retrieval and text generation workflows, Ollama for secure and efficient LLM provision, Built-in Embeddings for generating vector representations of content, and ChromaDB for efficient vector storage and retrieval. The RAG approach enables the language model to pull specific information from the uploaded documents, which improves the precision and relevance of the responses. The data source for the RAG process is flexible, allowing adaptation to different situations, which makes the solution versatile and suitable for various applications. This project aims to offer a practical approach to implementing the RAG technique with private LLMs.

Keywords: *Retrieval Augmented Generation (RAG), Large Language Models (LLMs), Open Web UI, Embeddings, Ollama, Vector Storage, Information Retrieval, Document Processing.*

REFERÊNCIAS

MICROSOFT. Overview of Retrieval-Augmented Generation (RAG) for NLP. *Microsoft Learn*, 2024. Disponível em: https://learn.microsoft.com/en-us/azure/applied-ai-services/generative-ai-overview. Acesso em: 22 set. 2024.

LEWIS, Patrick; OGUZ, Barlas; RINOTT, Rachel; RIEDEL, Sebastian; STOYANOV, Veselin. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. Disponível em: https://arxiv.org/abs/2005.11401. Acesso em: 28 set. 2024.

ORACLE. Best Practices for Implementing Private Large Language Models. *Oracle Cloud Blog*, 2024. Disponível em: https://blogs.oracle.com/cloud/post/implementing-private-llms-best-practices. Acesso em: 29 set. 2024.

DEVLIN, Jacob; CHANG, Ming-Wei; LEE, Kenton; TOUTANOVA, Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019. Disponível em: https://arxiv.org/abs/1810.04805. Acesso em: 15 out. 2024.

BROWN, Tom; MANN, Benjamin; RYDER, Nick; et al. Language Models are Few-Shot Learners. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. Disponível em: https://arxiv.org/abs/2005.14165. Acesso em: 25 out. 2024.

IBM. Natural Language Processing: An Introduction. *IBM Cloud Education*, 2024. Disponível em: https://www.ibm.com/cloud/learn/natural-language-processing. Acesso em: 3 out. 2024.

GUU, Kelvin; LEE, Kenton; TURTLE, Zora; YU, Yi; FINE, Jacob. REALM: Retrieval-Augmented Language Model. *Proceedings of the International Conference on Machine Learning (ICML)*, 2020. Disponível em: https://arxiv.org/abs/2002.08909. Acesso em: 6 out. 2024.

GOOGLE CLOUD. Introduction to Vector Databases for NLP Applications. *Google Cloud Documentation*, 2024. Disponível em: https://cloud.google.com/vertex-ai/docs/feature-overview/vector-database. Acesso em: 10 out. 2024.

CISCO. Security Considerations for Large Language Models and Chatbots. *Cisco Cybersecurity Insights*, 2024. Disponível em: https://www.cisco.com/c/en/us/solutions/security/cybersecurity-insights.html. Acesso em: 11 out. 2024.

AMAZON WEB SERVICES (AWS). Building Secure NLP Applications with Amazon SageMaker. *AWS Documentation*, 2024. Disponível em: https://docs.aws.amazon.com/sagemaker/latest/dg/nlp-security.html. Acesso em: 18 out. 2024.

MIN, Sewon; LEWIS, Patrick; HAKKANI-TÜR, Dilek; YIH, Wen-tau. Dense Passage Retrieval (DPR). *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. Disponível em: https://arxiv.org/abs/2010.03759. Acesso em: 30 out. 2024.