RCMOS – Revista Científica Multidisciplinar O Saber.
ISSN: 2675-9128. São Paulo-SP.
Ano IV, v.1, n.2, ago./ dez 2024. | submissão: 08/11/2024 | aceito: 10/11/2024 | publicação:12/11/2024

Protótipo de Pesquisa Documental para a Polícia Militar do Paraná com Retrieval Augmented Generation e Gemini AI em Ambiente Dockerizado

Prototype for Document Retrieval for the Military Police of Paraná with Retrieval Augmented Generation and Gemini AI in a Dockerized Environment

Cleiton Giacomelli da Silva
1º Tenente, Pós-Graduado em Ciências Jurídicas
Instituição: QOPM da Polícia Militar do Paraná
Endereço: Av. Mal. Floriano Peixoto, 1401, Rebouças, Curitiba - PR, CEP: 80230-110
E-mail: cleiton.giacomelli@pm.pr.gov.br

RESUMO
Este trabalho apresenta o desenvolvimento de um protótipo de pesquisa documental para as diretrizes e doutrinas da Polícia Militar do Paraná (PMPR) utilizando a estratégia Retrieval Augmented Generation (RAG) e o modelo de linguagem Gemini AI Flash 1.5. O protótipo foi implementado em um ambiente containerizado com Docker, visando garantir portabilidade e reprodutibilidade. A estratégia RAG combina a busca tradicional com modelos de linguagem avançados para gerar respostas mais precisas e completas às consultas dos usuários. O protótipo foi testado com perguntas reais e os resultados preliminares demonstram a capacidade do sistema em compreender as perguntas e fornecer respostas relevantes, com base nas informações contidas nos documentos da PMPR. O trabalho discute o potencial do protótipo para auxiliar os policiais militares no acesso às informações relevantes, superando as limitações da atual forma de pesquisa documental na instituição.
Palavras-chave: Pesquisa documental, Polícia Militar do Paraná, Retrieval Augmented Generation, Gemini AI, Docker, Processamento de Linguagem Natural.

ABSTRACT
This paper presents the development of a prototype for document retrieval for the Military Police of Paraná (PMPR) using the Retrieval Augmented Generation (RAG) strategy and the Gemini AI Flash 1.5 language model. The prototype was implemented in a containerized environment with Docker, aiming to ensure portability and reproducibility. The RAG strategy combines traditional search with advanced language models to generate more accurate and complete answers to user queries. The prototype was tested with real questions and preliminary results demonstrate the system's ability to understand the questions and provide relevant answers, based on the information contained in the PMPR documents. The paper discusses the potential of the prototype to assist military police officers in accessing relevant information, overcoming the limitations of the current form of document retrieval in the institution.
Keywords: Documental research, Paraná Military Police, Retrieval-Augmented Generation, Gemini AI, Docker, Natural Language Processing.

1 INTRODUÇÃO

A Polícia Militar do Paraná (PMPR), como principal instituição responsável pela segurança pública no estado, possui um extenso conjunto de diretrizes e doutrinas que regem suas ações e procedimentos operacionais. Essas diretrizes e doutrinas, que abrangem temas como uso da força, policiamento ostensivo, direitos humanos, atendimento a ocorrências e procedimentos administrativos, são essenciais para garantir a uniformidade de atuação, a legalidade das ações policiais e o respeito aos direitos humanos (BRASIL, 2023). O acesso rápido e eficiente a essas informações é crucial para o bom desempenho das atividades policiais e, consequentemente, para a qualidade do serviço prestado à população.

No entanto, a pesquisa documental nesse acervo enfrenta desafios significativos. Atualmente, a PMPR não possui um sistema centralizado de pesquisa documental. Os policiais militares dependem da busca manual em pastas na intranet, organizadas por localização física e nome dos arquivos. Esse processo se mostra ineficiente e demorado, especialmente em situações que demandam agilidade, como o atendimento a ocorrências policiais. A busca por palavras-chave, método tradicional de pesquisa, frequentemente se mostra limitada quando aplicada a conjuntos de documentos complexos e extensos como o da PMPR. Essas técnicas

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

podem não levar em consideração o contexto semântico das palavras, resultando em baixa precisão na recuperação de documentos relevantes e na geração de listas extensas de resultados que exigem análise manual, tornando o processo moroso e pouco eficaz (BAUMANN et al., 2023).

Diante desse desafio, a estratégia Retrieval Augmented Generation (RAG) surge como uma alternativa promissora. Combinando a busca tradicional com modelos de linguagem avançados, o RAG busca superar as limitações das técnicas convencionais, gerando respostas mais precisas e completas às consultas dos usuários. A estratégia se baseia na recuperação de trechos relevantes de documentos (*retrieval*) e na utilização desses trechos como contexto para a geração de respostas em linguagem natural (*generation*) por meio de modelos de linguagem (LEWIS et al., 2020). O RAG permite que o sistema compreenda o significado semântico das perguntas e encontre documentos relevantes mesmo que estes não contenham as palavras-chave exatas da consulta, tornando a busca mais eficiente e precisa.

Este trabalho apresenta o desenvolvimento de um protótipo de pesquisa documental para diretrizes e doutrinas da PMPR que utiliza a estratégia RAG e o modelo de linguagem Gemini AI Flash 1.5. Visando garantir a portabilidade e reprodutibilidade do sistema, optou-se por implementar o protótipo em um ambiente containerizado com Docker. O Docker permite isolar o ambiente de execução do protótipo, facilitando sua implantação em diferentes máquinas e garantindo que o sistema funcione de forma consistente em qualquer ambiente que suporte o Docker (BOETTIGER, 2015).

A hipótese central deste trabalho é que o protótipo com RAG e Gemini AI proporciona **maior precisão na recuperação de documentos relevantes** em comparação com as abordagens tradicionais de busca por palavras-chave, além de oferecer **respostas mais completas e informativas** em linguagem natural.

2. MATERIAIS E MÉTODOS

O protótipo de pesquisa documental foi desenvolvido utilizando uma arquitetura modular e containerizada, baseada na plataforma Docker (BOETTIGER, 2015). A containerização proporciona maior portabilidade e reprodutibilidade do sistema, permitindo sua execução em diferentes ambientes sem a necessidade de configurações complexas. Os seguintes componentes foram containerizados:

- **N8n:** Plataforma de automação de fluxo de trabalho de código aberto. No protótipo, o N8n atua como orquestrador, integrando os diferentes módulos do sistema e gerenciando o fluxo de dados entre eles, incluindo dois fluxos de trabalho distintos: um para processamento e indexação dos documentos e outro para a interação do usuário com o sistema de pesquisa (N8N, 2023).
- **Qdrant:** Banco de dados vetorial de código aberto, otimizado para armazenar e buscar embeddings de documentos. No protótipo, o Qdrant é utilizado para indexar os embeddings dos documentos da PMPR e para realizar a busca por similaridade semântica entre a consulta do usuário e os documentos (QDRANT, 2023).
- **Postgres:** Sistema de gerenciamento de banco de dados relacional de código aberto. No protótipo, o Postgres é utilizado para armazenar o histórico de conversas entre o usuário e o sistema, permitindo a recuperação de informações de pesquisas anteriores.

Tabela 1: Ferramentas e Bibliotecas

Ferramenta/Biblioteca	Versão	Descrição
Docker Desktop	4.35.0	Plataforma para containerização de aplicações
N8n	1.64.3	Plataforma de automação de fluxo de trabalho
Gemini Flash AI API	v1.5	API para o modelo de linguagem Gemini
Gemini Text Embedding	004	API para o modelo de embbedings Gemini
Qdrant	0.9.1	Banco de dados vetorial otimizado para busca semântica.
Postgres	17.0	Sistema de gerenciamento de banco de dados relacional, conhecido por sua robustez e flexibilidade
Python	3.11	Linguagem de programação utilizada para o desenvolvimento de scripts não existentes no N8n.

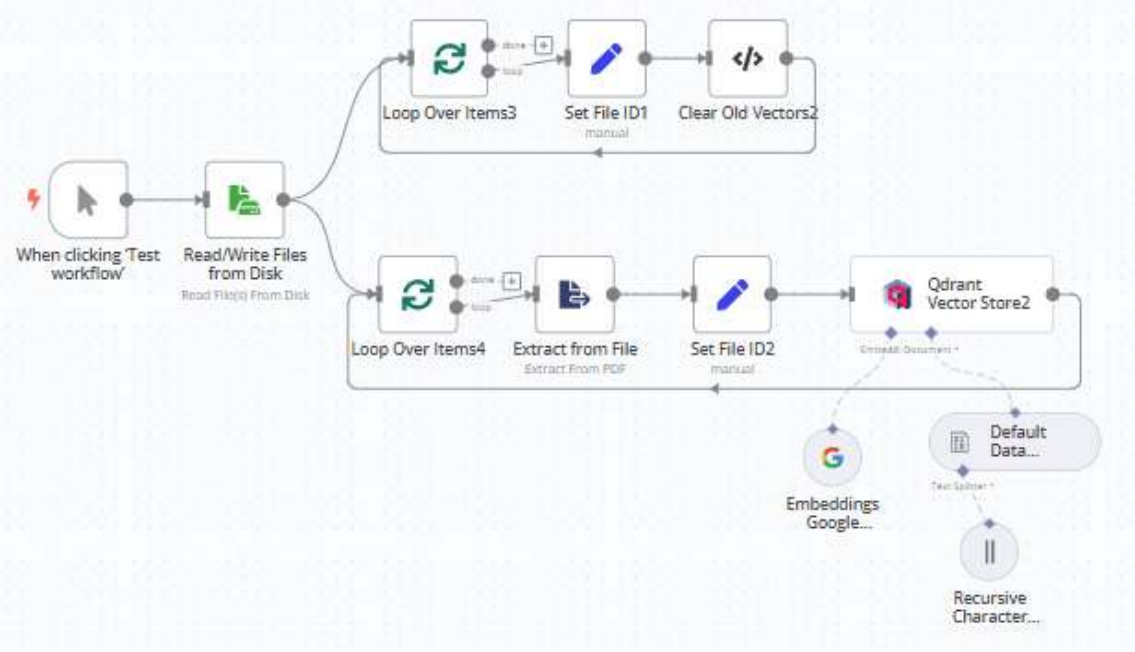
Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Fonte: Elaborado pelo autor.

2.1 WORKFLOW DO N8N: FLUXO DE PROCESSAMENTO E INDEXAÇÃO DE DOCUMENTOS

Este fluxo de trabalho é responsável por processar os documentos em PDF, gerar seus embeddings e indexá-los no banco de dados Qdrant. O processo é automatizado pelo N8n e pode ser dividido nas seguintes etapas:

Figura 1. Tela inicial ao abrir o aplicativo.



Fonte: Elaborado pelo autor.

1. **Leitura dos documentos:** O *node* "Read/Write Files from Disk" no N8n lê os documentos em PDF a partir da pasta designada, onde os arquivos da PMPR foram armazenados.
2. **Exclusão de vetores iguais:** O documento é enviado para dois destinos, o superior verifica se o documento já existe na base de dados do Qdrant e os exclui caso existam.
3. **Conversão para texto:** O *node* "Extract From File" converte os documentos PDF para texto puro.
4. **Pré-processamento:** O *node* "Default Data Retriever" realiza o pré-processamento do texto, dividindo os textos em blocos.
5. **Geração de embeddings:** O *node* "Qdrant Store" envia o texto pré-processado para a API do Gemini AI para gerar os embeddings dos documentos, utilizando o modelo Text Embedding 004.
6. **Indexação no Qdrant:** O *node* "Qdrant" armazena os embeddings e os metadados dos documentos (título, autor, data de publicação, etc.) no banco de dados vetorial Qdrant. Os metadados são extraídos do nome do arquivo PDF ou de informações presentes no próprio documento.

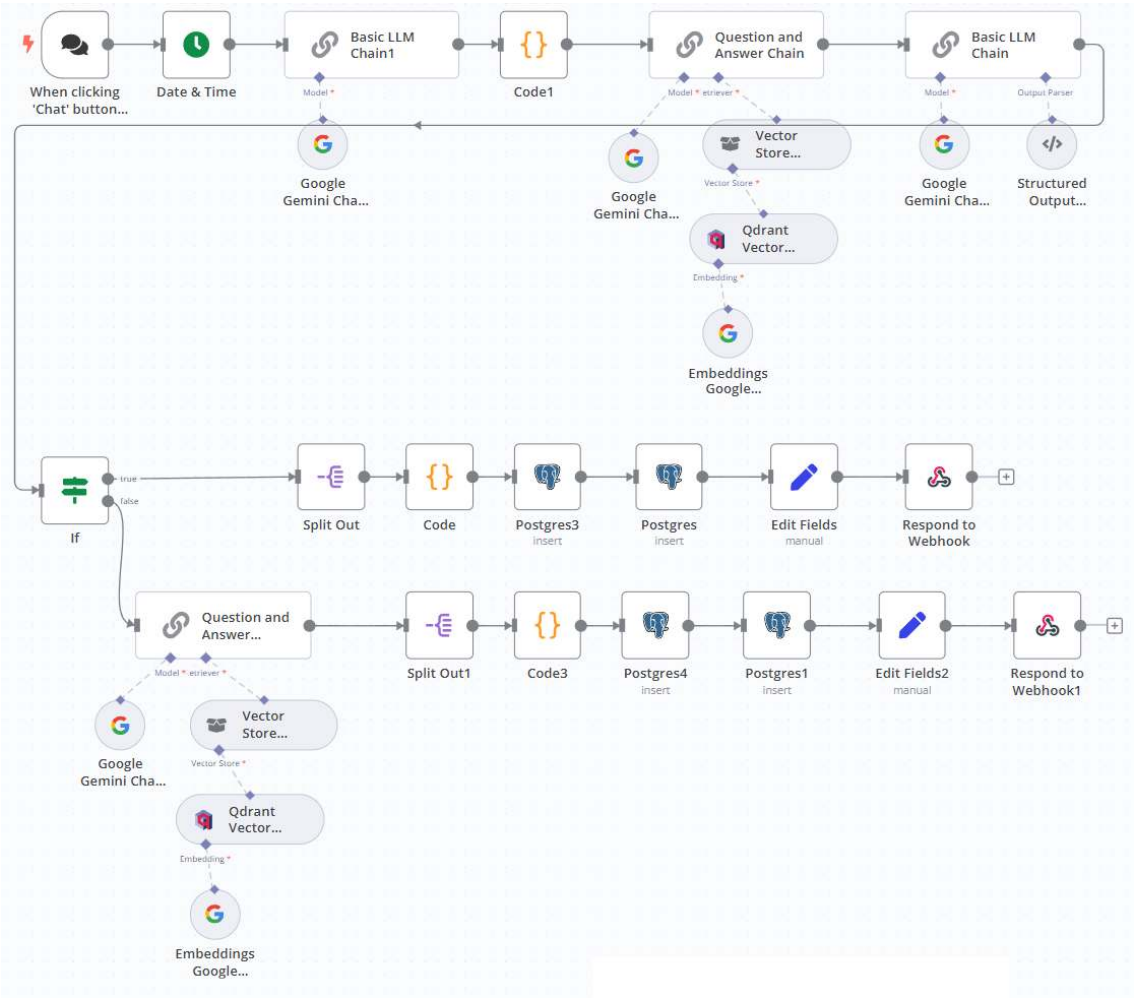
2.2 WORKFLOW DO N8N: FLUXO DE INTERAÇÃO COM O USUÁRIO

Este fluxo de trabalho é responsável por receber a solicitação do usuário, realizar a busca nos documentos

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

indexados e gerar a resposta. O processo é orquestrado pelo N8n e pode ser descrito nas seguintes etapas:

Figura 2: Diagrama do Fluxo de Interação com o Usuário



Fonte: Elaborado pelo autor.

1. **Recebimento da solicitação:** O *node* inicial recebe a consulta em linguagem natural enviada pelo usuário através da interface *web* de *chat* padrão do N8n.
2. **Refinamento da solicitação:** O *node* "Basic LLM Chain" utiliza o Gemini AI para realizar o refinamento da solicitação do usuário. As seguintes tarefas podem ser realizadas:
   a. Correção ortográfica;
   b. Expansão de sinônimos;
   c. Identificação de entidades nomeadas (pessoas, lugares, organizações, etc.).
   d. Reformulação da pergunta para torná-la mais clara e precisa.
3. **Geração de embeddings da solicitação:** O *node* "Question and Answer Chain" envia a solicitação refinada para a API do Gemini AI, que gera um vetor de *embeddings* representando o significado semântico da solicitação, utilizando o modelo Text Embedding 004.
4. **Busca por similaridade:** O *node* "Qdrant" envia o vetor de *embeddings* da solicitação para o Qdrant, que realiza a busca pelos documentos mais relevantes com base na similaridade entre os *embeddings*.
5. **Geração da resposta:** O *node* "Question and Answer Chain" envia os documentos recuperados e a solicitação original para a API do Gemini AI, que gera uma resposta concisa e informativa, utilizando os documentos como contexto e o modelo Flash 1.5.
6. **Validação da resposta:** Um novo *node* "Basic LLM Chain" é utilizado para realizar a validação da resposta gerada pelo Gemini, utilizando o próprio Gemini com configurações diferentes.
7. **Resposta ao usuário ou nova tentativa:**

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

e. Caso a resposta seja validada (ou se a validação não for utilizada): Os *nodes* "Postgres" armazenam a solicitação do usuário e a resposta gerada no banco de dados Postgres. O *node* "Webhook Response" envia a resposta gerada pelo Gemini para o usuário através da interface *web*.

f. Caso a resposta não seja validada: Outro *node* "Question and Answer Chain" com o Gemini AI é utilizado para reformular a pergunta original com base no contexto da resposta anterior e dos documentos recuperados, buscando obter uma resposta mais adequada. Os *nodes* "Postgres" armazenam a solicitação do usuário e a *resposta* gerada no banco de dados Postgres. O *node* "Webhook Response" envia a resposta gerada pelo Gemini para o usuário através da interface *web*.

2.3 IMPLEMENTAÇÃO DA ESTRATÉGIA RAG

A implementação da estratégia RAG no protótipo envolve as etapas descritas nos fluxos de trabalho do N8n, incluindo o pré-processamento dos documentos, a geração de embeddings para os documentos e para a solicitação do usuário, a busca por similaridade semântica no Qdrant e a geração da resposta pelo Gemini AI, com a adição do refinamento da solicitação do usuário e da validação da resposta (opcional), conforme detalhado na seção 2.2.

3. RESULTADOS

O protótipo de pesquisa documental para diretrizes e doutrinas da PMPR foi submetido a uma série de testes com o intuito de demonstrar sua funcionalidade e capacidade de responder a perguntas relevantes para a atividade policial. É importante ressaltar que, nesta fase inicial de desenvolvimento, não foram realizados testes formais com métricas de precisão, revocação e tempo de resposta, nem avaliações qualitativas sistemáticas das respostas. O objetivo principal desta seção é apresentar exemplos de interação com o protótipo, permitindo que o leitor tire suas próprias conclusões sobre o potencial da ferramenta.

As perguntas utilizadas nos testes foram elaboradas de forma aleatória, buscando abranger diferentes temas e níveis de complexidade presentes nas diretrizes e doutrinas da PMPR. A seguir, são apresentados alguns exemplos de perguntas e as respectivas respostas geradas pelo protótipo:

3.1 PERGUNTA 1

Pergunta:
Quais documentos tratam sobre reintegração de posse?
Resposta:
Sobre sua dúvida referente aos documentos que tratam sobre reintegração de posse, a PMPR possui normas específicas para o emprego de tropas em apoio aos Oficiais de Justiça em situações de reintegração de posse. De acordo com a Diretriz nº 006/2022 – PM3, a atuação da PMPR nesses casos é regida por diversos documentos, dentre eles:

- Decreto Estadual 2.484, de 17 de agosto de 1993: dispõe sobre a autoridade policial que receber requisição judicial de deslocamento de tropa para realizar despejos coletivos e dá outras diretrizes.
- Lei Estadual nº 16.575, de 28 de setembro de 2010: Lei de Organização Básica da PMPR.
- Decreto Estadual 1.238, de 04 de maio de 2015: Normaliza e padroniza o uso de instrumentos de menor potencial ofensivo pelos operadores de segurança pública.

Além disso, a Diretriz nº 006/2022 também menciona a necessidade de planejamento prévio para o atendimento a situações de despejo, reintegração e imissão de posse de bens imóveis.
Portanto, a legislação específica da PMPR, as diretrizes do Comando Geral e os decretos estaduais são os documentos que orientam a atuação da PMPR em situações de reintegração de posse.

3.2 PERGUNTA 2

Pergunta 2:
Qual deve ser a velocidade de deslocamento da viatura?
Resposta:

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Entendo sua dúvida sobre a velocidade de deslocamento da viatura. A Nota de Instrução nº 001/2001-PM3, que regulamenta a velocidade, deslocamento e estacionamento das viaturas operacionais da PMPR, define a velocidade máxima e mínima para diferentes vias e tipos de veículos, com base no Código de Trânsito Brasileiro (CTB).

Em patrulhamento, a velocidade de deslocamento da viatura deverá ser um pouco inferior à do fluxo normal de veículos, permitindo ao policial militar observar e ser observado (Nota de Instrução nº 001/2001-PM3, item 4.a.1). Em atendimento de ocorrência policial ou em emergências, a velocidade das viaturas operacionais não poderá ser superior à velocidade máxima permitida para a via, rodovia ou estrada (Nota de Instrução nº 001/2001-PM3, item 4.a.2). Em relação a comboios (Nota de Instrução nº 001/2001-PM3, item 4.a.4), a velocidade deve ser mantida de forma segura para todos os componentes do comboio e terceiros. No contexto de atendimento de ocorrências, como assaltos com fuga em veículos, é proibido, em qualquer circunstância, a perseguição (Nota de Instrução nº 001/2001-PM3, item 4.a.2).

É importante observar que, em qualquer situação, a velocidade deve ser compatível com o binômio rapidez e segurança (POP nº 006 – RETENÇÃO E REMOÇÃO DE VEÍCULOS, item 4), ou seja, a equipe atentará para o aspecto da segurança no trânsito e para a necessidade de chegar ao destino com segurança. Para que as viaturas da PMPR se desloquem com segurança e eficiência, é fundamental que os policiais militares se atentem às normas do CTB e às diretrizes da Corporação, buscando sempre a melhor forma de garantir a segurança de todos os envolvidos.

3.3 PERGUNTA 3

Pergunta 3:
Como atender uma ocorrência envolvendo CAC?
Resposta:
Compreendo sua dúvida sobre como atender uma ocorrência envolvendo um Colecionador, Atirador e Caçador (CAC). De acordo com o Procedimento Operacional Padrão (POP) nº 100.10 - Ocorrência Envolvendo Colecionador, Atirador e Caçador (CAC), a equipe policial militar deve seguir os seguintes passos:

1. Abordagem com cautela: Abordar a pessoa com fundada suspeita de portar arma de fogo, observando os protocolos e técnicas da PMPR. Retirar a arma que o abordado estiver portando, mesmo que ele se identifique como CAC, até que sua identidade seja confirmada e toda a documentação seja conferida.
2. Revistar com segurança: Realizar revista pessoal e do veículo, caso haja fundada suspeita, para verificar se a pessoa não possui outras armas de fogo ou irregularidades. A revista deve ser feita com atenção à segurança da equipe e do abordado, que deve permanecer em distância segura e com as mãos visíveis.
3. Solicitar documentação: Solicitar os seguintes documentos: Certificado de Registro de Colecionador, Atirador e Caçador (CRCAC) Guia de Tráfego (GT) - para o transporte de armas, munições e acessórios, conforme o Art. 82 do Decreto n.º 10.030, de 30 de setembro de 2019.
4. Verificar conformidade e validade: Conferir a vigência e a conformidade dos documentos com os modelos anexados ao POP. A falta ou inadequação de qualquer documento implica a detenção e encaminhamento do indivíduo à autoridade policial competente (Art. 14 ou Art. 16 da Lei 10.826/2003).
5. Verificar condições de transporte: O CAC não tem autorização para portar arma de fogo, e o armamento que esteja transportando deve estar desmuniciado e acondicionado de forma segura. Se essas exigências não forem cumpridas, a equipe deve encaminhar o indivíduo à autoridade policial (Art. 14 ou Art. 16 da Lei 10.826/2003).
6. Comunicar ao SFPC5: Comunicar ao Serviço de Fiscalização de Produtos Controlados da 5.ª Região Militar (SFPC5), qualquer ocorrência que resulte no encaminhamento do CAC à Delegacia de Polícia, para a tomada de medidas administrativas.

Lembre-se de que este procedimento deve ser realizado de acordo com as normas da PMPR e em conformidade com a legislação federal, como a Lei 10.826/2003 e o Decreto 10.030/2019.

Orientações adicionais:

- É crucial manter a segurança da equipe durante a abordagem e revista.
- A equipe deve ter conhecimento das legislações e normas que regulamentam a posse, o porte e o transporte de armas por CACs.
- A comunicação entre os policiais e o acompanhamento da ocorrência são fundamentais para o sucesso do atendimento.

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Importante: Caso tenha dúvidas sobre a aplicação do POP, consulte o seu superior hierárquico.

Observações:

- Os exemplos apresentados sugerem que o protótipo é capaz de compreender as perguntas dos usuários e fornecer respostas relevantes, com base nas informações contidas nas diretrizes e doutrinas da PMPR. As respostas demonstram a capacidade do Gemini AI em gerar textos coerentes, concisos e informativos, além de extrair e sintetizar informações relevantes dos documentos.
- A inclusão de referências aos documentos da PMPR nas respostas aumenta a confiabilidade e a transparência do sistema, permitindo que o usuário verifique a origem das informações.
- É fundamental ressaltar que estes exemplos são apenas ilustrativos e não representam uma avaliação exaustiva do protótipo. Testes mais robustos, com métricas de performance e avaliações qualitativas sistemáticas, com a participação de um grupo representativo de usuários, serão realizados em etapas futuras do desenvolvimento.

4. DISCUSSÃO

Os resultados apresentados na seção anterior, embora preliminares e baseados em um conjunto limitado de testes, sugerem que o protótipo de pesquisa documental com RAG e Gemini AI possui um grande potencial para auxiliar os policiais militares no acesso às informações relevantes das diretrizes e doutrinas da PMPR.

A capacidade do protótipo em compreender perguntas em linguagem natural e fornecer respostas concisas e informativas, com referências aos documentos da PMPR, representa um avanço significativo em relação à forma como a pesquisa documental é realizada atualmente na instituição. Atualmente, a PMPR não possui um sistema centralizado de pesquisa documental. Os policiais militares dependem da busca manual em pastas na intranet, organizadas por localização física e nome dos arquivos, o que torna o processo de encontrar informações específicas ineficiente e demorado. Essa busca manual exige que o policial saiba exatamente onde o documento está armazenado e qual o seu nome, o que nem sempre é possível, especialmente em situações de urgência.

A utilização da estratégia RAG, combinada com o modelo de linguagem Gemini AI, permite que o protótipo vá além da simples busca por palavras-chave. O sistema é capaz de compreender o significado semântico das perguntas e buscar documentos relevantes com base na similaridade de significado, mesmo que estes não contenham as palavras-chave exatas da consulta. Essa capacidade de busca semântica é fundamental para lidar com a complexidade e a variedade de termos e expressões presentes nas diretrizes e doutrinas da PMPR, que podem ser escritas em linguagem jurídica ou técnica, com termos específicos que podem não ser do conhecimento de todos os policiais.

Além disso, a geração de respostas em linguagem natural pelo Gemini AI torna a interação com o sistema mais intuitiva e amigável para o usuário. As respostas são apresentadas de forma clara e concisa, com a inclusão de trechos dos documentos relevantes para fundamentar as informações fornecidas. Essa abordagem contribui para aumentar a confiança e a transparência do sistema, permitindo que o usuário compreenda como a resposta foi gerada e verifique a sua validade, o que é fundamental para garantir a confiabilidade das informações utilizadas na tomada de decisões em situações reais.

4.1 LIMITAÇÕES DO PROTÓTIPO

É importante reconhecer que o protótipo ainda se encontra em fase inicial de desenvolvimento e possui algumas limitações:

- **Conjunto de dados limitado:** O protótipo foi treinado com um conjunto de dados restrito aos documentos disponibilizados pela PMPR, que embora abrangente, não representa a totalidade das informações relevantes para a atividade policial. A inclusão de outras fontes de informação, como legislação federal e estadual, normas de outros órgãos de segurança pública, manuais técnicos e artigos científicos, poderia enriquecer o conhecimento do sistema e melhorar a qualidade das respostas, permitindo ao protótipo responder a perguntas sobre temas mais amplos e específicos.
- **Ausência de testes formais:** A avaliação do protótipo foi realizada com um conjunto limitado de perguntas e sem a utilização de métricas de performance e avaliações qualitativas

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

sistemáticas. A realização de testes mais robustos, com a participação de um grupo representativo de usuários, é fundamental para avaliar o desempenho do sistema em diferentes cenários, identificar as suas deficiências e direcionar os trabalhos futuros. A utilização de métricas como precisão, revocação e F1-score permitiria quantificar a capacidade do sistema em recuperar documentos relevantes e fornecer respostas corretas. As avaliações qualitativas, por sua vez, poderiam coletar a opinião dos usuários sobre a usabilidade do sistema, a clareza das respostas e a utilidade da ferramenta para o seu trabalho.

- **Interface de usuário simples:** A interface de usuário do protótipo ainda é bastante básica, sendo limitada ao envio de perguntas e recebimento de respostas em formato textual. O desenvolvimento de uma interface mais intuitiva e amigável, com funcionalidades como histórico de pesquisas, filtros de busca, visualização dos documentos, realce dos trechos relevantes nas respostas e possibilidade de interação por voz, poderia melhorar significativamente a experiência do usuário e facilitar a utilização do sistema em diferentes dispositivos, como smartphones e tablets.
- **Dependência da API do Gemini AI:** O protótipo depende da disponibilidade e do bom funcionamento da API do Gemini AI, um serviço externo fornecido pelo Google. A utilização de modelos de linguagem alternativos, como modelos de código aberto ou modelos hospedados em servidores próprios da PMPR, ou a implementação de mecanismos de *fallback* para lidar com eventuais falhas na API do Gemini AI, como o uso de um modelo de linguagem menor e mais rápido para responder a perguntas simples, poderiam aumentar a robustez e a autonomia do sistema, garantindo o seu funcionamento mesmo em caso de indisponibilidade da API do Gemini AI.

4.2 TRABALHOS FUTUROS

As limitações identificadas apontam para direções promissoras para trabalhos futuros:

- **Expansão do conjunto de dados:** Integrar outras fontes de informação relevantes para a atividade policial, como legislação federal e estadual, normas de outros órgãos de segurança pública, manuais técnicos e artigos científicos, para enriquecer o conhecimento do sistema e ampliar o escopo de perguntas que o protótipo pode responder.
- **Realização de testes formais:** Avaliar o desempenho do protótipo com um conjunto abrangente de perguntas e métricas de performance, como precisão, revocação e F1-score, além de realizar avaliações qualitativas com a participação de policiais militares, para coletar feedback sobre a usabilidade do sistema e a qualidade das respostas.
- **Desenvolvimento de uma interface de usuário avançada:** Criar uma interface mais intuitiva e amigável, com funcionalidades como histórico de pesquisas, filtros de busca, visualização dos documentos, realce dos trechos relevantes nas respostas e possibilidade de interação por voz, para melhorar a experiência do usuário e facilitar a utilização do sistema em diferentes dispositivos.
- **Exploração de modelos de linguagem alternativos:** Investigar a possibilidade de utilizar outros modelos de linguagem, como modelos de código aberto ou modelos hospedados em servidores próprios da PMPR, para aumentar a flexibilidade, a robustez e a autonomia do sistema, reduzindo a dependência de serviços externos.
- **Integração com outros sistemas da PMPR:** Integrar o protótipo com outros sistemas utilizados pela PMPR, como o sistema de gerenciamento de ocorrências e o sistema de inteligência policial, para oferecer aos policiais militares um acesso unificado às informações relevantes, permitindo que o protótipo acesse informações contextuais sobre a ocorrência em andamento ou o perfil do suspeito, por exemplo, para fornecer respostas mais personalizadas e relevantes.

5. CONCLUSÃO

Este trabalho apresentou o desenvolvimento de um protótipo de pesquisa documental para diretrizes e doutrinas da Polícia Militar do Paraná (PMPR) utilizando a estratégia Retrieval Augmented Generation (RAG) e o modelo de linguagem Gemini AI Flash 1.5, em um ambiente containerizado com Docker. O objetivo principal foi criar uma ferramenta que facilitasse o acesso dos policiais militares às informações relevantes para o desempenho de suas funções, superando as limitações da atual forma de pesquisa documental na instituição, baseada na busca manual em pastas na intranet.

Os resultados obtidos, embora preliminares e baseados em um conjunto limitado de testes, demonstram o potencial do protótipo para auxiliar os policiais militares no acesso às informações relevantes das diretrizes e doutrinas da PMPR. A capacidade do sistema em compreender perguntas em linguagem natural e fornecer

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

respostas concisas e informativas, com referências aos documentos da PMPR, representa um avanço significativo em relação à forma como a pesquisa documental é realizada atualmente.

A utilização da estratégia RAG, combinada com o modelo de linguagem Gemini AI, permite que o protótipo vá além da simples busca por palavras-chave, realizando a busca semântica e gerando respostas em linguagem natural que aumentam a confiança e a transparência do sistema.

Apesar do potencial demonstrado, o protótipo ainda se encontra em fase inicial de desenvolvimento e possui limitações que precisam ser abordadas em trabalhos futuros, como a expansão do conjunto de dados, a realização de testes formais com métricas de performance e avaliações qualitativas, o desenvolvimento de uma interface de usuário mais avançada e a exploração de modelos de linguagem alternativos.

Acredita-se que a continuidade deste trabalho, com a implementação das melhorias propostas, poderá resultar em uma ferramenta valiosa para a PMPR e para outras instituições de segurança pública. A disponibilização de um sistema de pesquisa eficiente e amigável pode contribuir significativamente para a melhoria da qualidade do trabalho policial, facilitando o acesso às informações relevantes, promovendo a uniformidade de procedimentos e auxiliando na tomada de decisões mais eficazes, o que pode ter um impacto positivo na segurança da sociedade paranaense.

Em suma, este trabalho representa um passo inicial promissor no desenvolvimento de soluções inovadoras para a pesquisa documental na área de segurança pública, com o potencial de impactar positivamente a atuação da PMPR e a segurança da sociedade paranaense.

REFERÊNCIAS

BAUMANN, P. et al. **Large Language Models for Retrieval Augmented Generation: A Comprehensive Survey.** arXiv preprint arXiv:2308.01186, 2023.

BOETTIGER, C. **An introduction to Docker for reproducible research.** ACM SIGOPS Operating Systems Review, v. 49, n. 1, p. 71-79, 2015.

BRASIL. Ministério da Justiça e Segurança Pública. **Portaria nº 332, de 27 de abril de 2023.** Dispõe sobre o Manual de Uso da Força Policial. Diário Oficial da União, Brasília, DF, 28 abr. 2023. Seção 1, p. 47.

LEWIS, P. et al. **Retrieval-augmented generation for knowledge-intensive nlp tasks.** Advances in Neural Information Processing Systems, v. 33, p. 9459-9474, 2020.

N8N. **n8n - Workflow Automation.** Disponível em: https://n8n.io/. Acesso em: 06 nov. 2024.

QDRANT. **Qdrant - Vector Database.** Disponível em: https://qdrant.tech/. Acesso em: 06 nov. 2024.

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.