RCMOS – Multidisciplinary Scientific Journal of Knowledge.
ISSN: 2675-9128. São Paulo-SP.
Year IV, v.1, n.2, Aug./Dec. 2024. | submission: 2024-11-08 | accepted: 2024-11-10 | publication: 2024-11-12

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

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
1st Lieutenant, Postgraduate in Legal Sciences
Institution: QOPM of the Military Police of Paraná
Address: Av. Mal. Floriano Peixoto, 1401, Rebouças, Curitiba - PR, Zip Code: 80230-110 E-mail: cleiton.giacomelli@pm.pr.gov.br

SUMMARY
This paper presents the development of a documentary research prototype for the guidelines and doctrines of 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 responses to user queries. The prototype was tested with real questions and the 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 documentary research in the institution. **Keywords:** Documentary research, Military Police of Paraná, Retrieval Augmented Generation, Gemini AI, Docker, Natural Language Processing.

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: Documentary research, Paraná Military Police, Retrieval-Augmented Generation, Gemini AI, Docker, Natural Language Processing.

1 INTRODUCTION

The Military Police of Paraná (PMPR), as the main institution responsible for public security in the state, has an extensive set of guidelines and doctrines that govern its actions and operational procedures. These guidelines and doctrines, which cover topics such as use of force, overt policing, human rights, response to incidents and administrative procedures, are essential to ensure uniformity action, the legality of police actions and respect for human rights (BRAZIL, 2023). The Quick and efficient access to this information is crucial for the good performance of police activities and, consequently, for the quality of the service provided to the population.

However, documentary research in this collection faces significant challenges. Currently, the The PMPR does not have a centralized document search system. Military police officers rely on manual searches in folders on the intranet, organized by physical location and file names. This process is inefficient and time-consuming, especially in situations that require agility, such as responding to police incidents. Keyword searches, a traditional search method, often prove to be limited when applied to complex and extensive sets of documents such as those of the PMPR. These techniques

may not take into account the semantic context of words, resulting in low accuracy in retrieving relevant documents and generating extensive lists of results that require manual analysis, making the process slow and ineffective (BAUMANN et al., 2023).

Faced with this challenge, the Retrieval Augmented Generation (RAG) strategy emerges as a promising alternative. By combining traditional search with advanced language models, RAG seeks to overcome the limitations of conventional techniques, generating more accurate and complete responses to user queries. The strategy is based on the recovery of relevant excerpts from documents (*retrieval*)and in using these excerpts as context for generating responses in natural language (*generation*)through language models (LEWIS et al., 2020). RAG allows the system to understand the semantic meaning of questions and find relevant documents even if they do not contain the exact keywords of the query, making the search more efficient and accurate.

This paper presents the development of a documentary research prototype for PMPR guidelines and doctrines that uses the RAG strategy and the Gemini AI Flash 1.5 language model. In order to ensure the portability and reproducibility of the system, it was decided to implement the prototype in a containerized environment with Docker. Docker allows isolating the execution environment of the prototype, facilitating its deployment on different machines and ensuring that the system works consistently in any environment that supports Docker (BOETTIGER, 2015).

The central hypothesis of this work is that the prototype with RAG and Gemini AI provides **greater accuracy in retrieving relevant documents**compared to traditional keyword search approaches, as well as offering**more complete and informative answers**in natural language.

2. MATERIALS AND METHODS

The documentary research prototype was developed using a modular and integrated architecture. nerized, based on the Docker platform (BOETTIGER, 2015). Containerization provides greater portability and reproducibility of the system, allowing its execution in different environments without the need for complex configurations. The following components were containerized:

- **N8n:**Open source workflow automation platform. In the prototype, N8n acts as an orchestrator, integrating the different modules of the system and managing the data flow between them, including two distinct workflows: one for processing and indexing documents and another for user interaction with the search system (N8N, 2023).
- **Qdrant:**Open-source vector database optimized for storing and searching document embeddings. In the prototype, Qdrant is used to index the embeddings of PMPR documents and to search for semantic similarity between the user's query and the documents (QDRANT, 2023).
- **Postgres:**Open source relational database management system. In the prototype, Postgres is used to store the history of conversations between the user and the system, allowing the retrieval of information from previous searches.

Table 1: Tools and Libraries

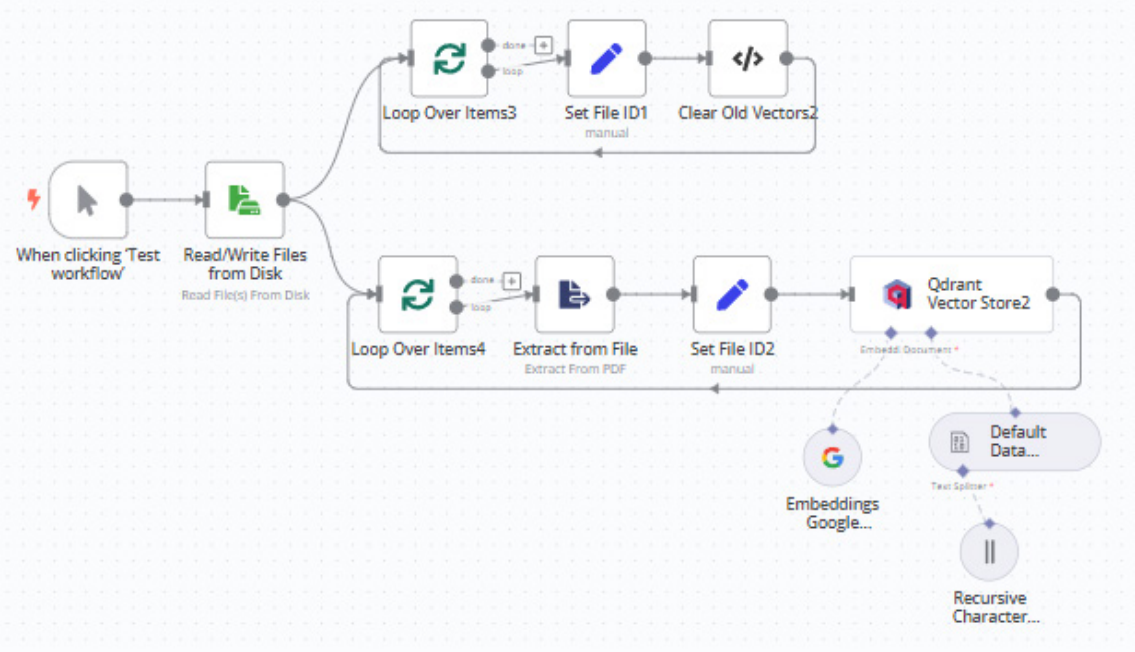
Tool/Library	Version	Description
Docker Desktop	4.35.0	Platform for application containerization
N8n	1.64.3	Workflow Automation Platform
Gemini Flash AI API	v1.5	API for the Gemini Language Model
Gemini Text Embedding	004	API for Gemini Embeddings Model
Qdrant	0.9.1	Vector database optimized for semantic search.
Postgres	17.0	Relational database management system, known for its robustness and flexibility
Python	3.11	Programming language used for the development of scripts that do not exist in N8n.

Source: Prepared by the author.

2.1 N8N WORKFLOW: DOCUMENT PROCESSING AND INDEXING FLOW

This workflow is responsible for processing PDF documents, generating their embeddings and indexing them. in the Qdrant database. The process is automated by N8n and can be divided into the following steps:

Figure 1. Home screen when opening the application.



Source: Prepared by the author.

1. **Reading the documents:** THE *node* "Read/Write Files from Disk" on the N8n reads documents PDF files from the designated folder where the PMPR files were stored.
2. **Exclusion of equal vectors:** The document is sent to two destinations, the upper one checks if the document already exists in the Qdrant database and deletes it if it does.
3. **Convert to text:** THE *node* "Extract From File" converts PDF documents to plain text.
4. **Pre-processing:** THE *node* "Default Data Retriever" performs text preprocessing, dividing texts into blocks.
5. **Embedding generation:** THE *node* "Qdrant Store" sends the preprocessed text to the Gemini AI API to generate document embeddings using the Text Embedding 004 model.
6. **Indexing in Qdrant:** THE *node* "Qdrant" stores document embeddings and metadata (title, author, publication date, etc.) in the Qdrant vector database. Metadata is extracted from the PDF file name or from information present in the PDF itself.

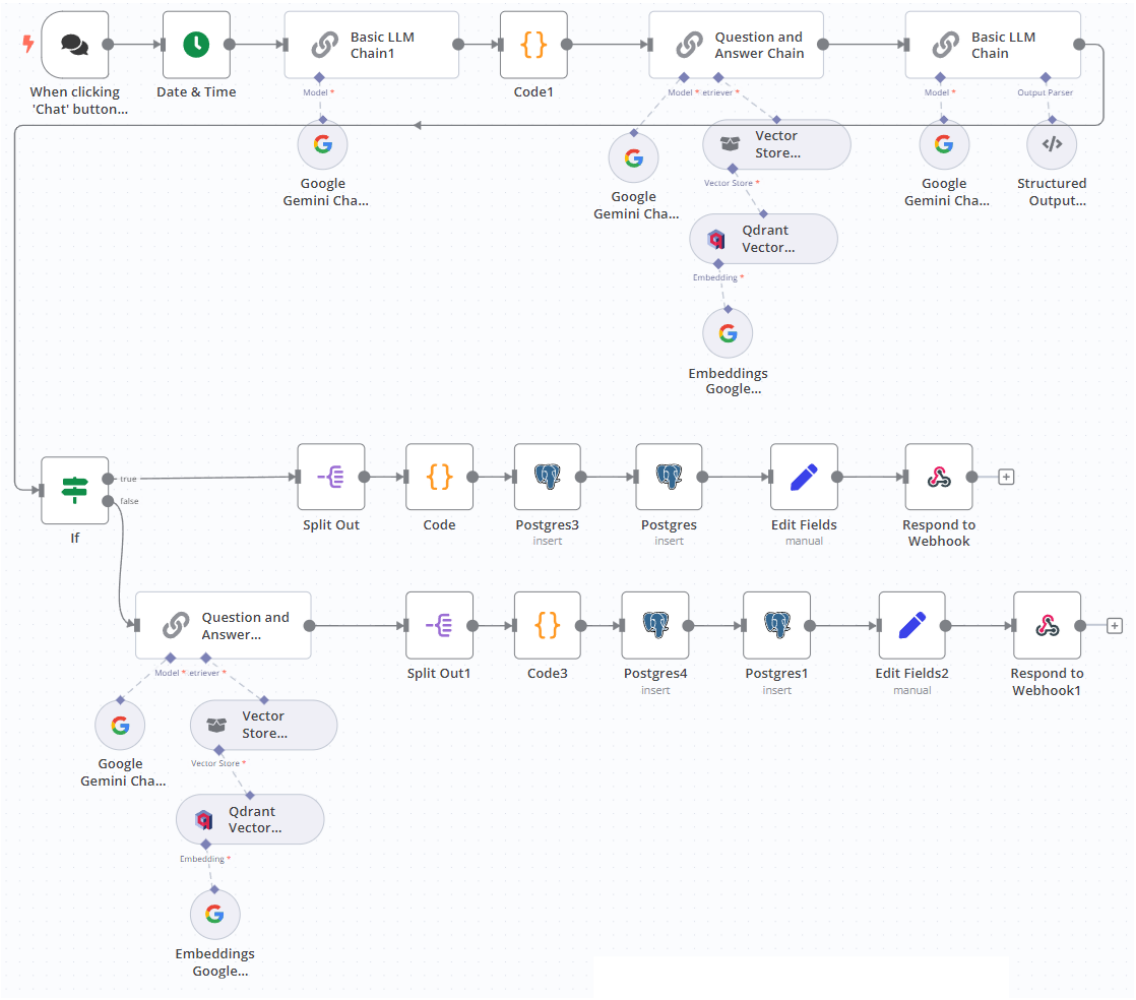
document.

2.2 N8N WORKFLOW: USER INTERACTION FLOW

This workflow is responsible for receiving the user's request, performing the search in the documents

indexed and generate the response. The process is orchestrated by N8n and can be described in the following steps:

Figure 2: User Interaction Flow Diagram



Source: Prepared by the author.

1. Receipt of request: THE *node* initial receives the query in natural language sent by the user through the interface *web* of *chat* N8n standard.
2. Request refinement: THE *node* "Basic LLM Chain" uses Gemini AI to perform user request refinement. The following tasks can be performed: a. Spelling correction;

b. Expansion of synonyms;
c. Identification of named entities (people, places, organizations, etc.).
d. Rephrasing the question to make it clearer and more precise.
3. Generating request embeddings: THE *node* "Question and Answer Chain" sends the refined request to the Gemini AI API, which generates a vector of *embeddings* representing the semantic meaning of the request, using the Text Embedding 004 model.
4. Search by similarity: THE *node* "Qdrant" sends the vector of *embeddings* of the request to Qdrant, which searches for the most relevant documents based on the similarity between them. *embeddings*.
5. Response generation: THE *node* "Question and Answer Chain" sends the retrieved documents and the original request to the Gemini AI API, which generates a concise and informative response using the documents as context and the Flash 1.5 model.
6. Response validation: A new one *node* "Basic LLM Chain" is used to validate the response generated by Gemini, using Gemini itself with different configurations.

7. Response to user or retry:

and. If the response is validated (or if validation is not used): You *nodes* "Postgres" stores the user request and the generated response in the Postgres database. The *node* "Webhook Response" sends the response generated by Gemini to the user through the interface *web*.

f. If the response is not validated: Other *node* "Question and Answer Chain" with Gemini AI is used to reformulate the original question based on the context of the previous answer and the retrieved documents, seeking to obtain a more appropriate answer. *nodes* "Postgres" store the user request and the *response* generated in the Postgres database. The *node* "Webhook Response" sends the response generated by Gemini to the user through the interface *web*.

2.3 IMPLEMENTATION OF THE RAG STRATEGY

Implementing the RAG strategy in the prototype involves the steps described in the N8n workflows, including preprocessing the documents, generating embeddings for the documents and the user request, searching for semantic similarity in Qdrant, and generating the response by Gemini AI, with the addition of refining the user request and validating the response (optional), as detailed in section 2.2.

3. RESULTS

The documentary research prototype for PMPR guidelines and doctrines was subjected to a series of tests with the aim of demonstrating its functionality and ability to answer questions relevant to police activity. It is important to emphasize that, in this initial phase of development, no formal tests were carried out with precision, recall and response time metrics, nor were systematic qualitative evaluations of the responses. The main objective of this section is to present examples of interaction with the prototype, allowing the reader to draw their own conclusions about the potential of the tool.

The questions used in the tests were randomly designed to cover different topics and levels of complexity present in the PMPR guidelines and doctrines. Below are some examples of questions and their respective answers generated by the prototype:

3.1 QUESTION 1

Question:

What documents deal with reinstatement of possession?

Response:

Regarding your question regarding the documents that deal with reinstatement of possession, the PMPR has specific rules for the use of troops to support Court Officers in situations of reinstatement of possession. According to Guideline No. 006/2022 – PM3, the PMPR's actions in these cases are governed by several documents, including:

- State Decree 2,484, of August 17, 1993: provides for the police authority that receives a judicial request to deploy troops to carry out collective evictions and provides other guidelines.
- State Law No. 16,575, of September 28, 2010: Basic Organization Law of PMPR.
- State Decree 1,238, of May 4, 2015: Normalizes and standardizes the use of instruments with less offensive potential by public security operators.

Furthermore, Guideline No. 006/2022 also mentions the need for prior planning to deal with situations of eviction, repossession and possession of real estate. Therefore, the specific PMPR legislation, the General Command guidelines and state decrees are the documents that guide the PMPR's actions in situations of repossession.

3.2 QUESTION 2

Question 2:

What should be the vehicle's travel speed?

Response:

I understand your doubt about the vehicle's speed. Instruction Note No. 001/2001-PM3, which regulates the speed, movement and parking of PMPR operational vehicles, defines the maximum and minimum speed for different roads and types of vehicles, based on the Brazilian Traffic Code (CTB).

When patrolling, the speed at which the vehicle moves must be slightly lower than that of the normal flow of vehicles, allowing the military police officer to observe and be observed (Instruction Note No. 001/2001-PM3, item 4.a.1). When responding to police incidents or emergencies, the speed of operational vehicles may not exceed the maximum speed permitted for the road, highway or road (Instruction Note No. 001/2001-PM3, item 4.a.2). In relation to convoys (Instruction Note No. 001/2001-PM3, item 4.a.4), the speed must be maintained in a safe manner for all members of the convoy and third parties. When responding to incidents, such as robberies with escape in vehicles, pursuit is prohibited under any circumstances (Instruction Note No. 001/2001-PM3, item 4.a.2).

It is important to note that, in any situation, the speed must be compatible with the binomial of speed and safety (POP nº 006 – VEHICLE RETENTION AND REMOVAL, item 4), that is, the team will pay attention to the aspect of traffic safety and the need to reach the destination safely. In order for PMPR vehicles to move safely and efficiently, it is essential that military police officers pay attention to the CTB standards and the Corporation's guidelines, always seeking the best way to guarantee the safety of everyone involved.

3.3 QUESTION 3

Question 3:

How to respond to an incident involving CAC?

Response:

I understand your doubt about how to respond to an incident involving a Collector, Shooter and Hunter (CAC). According to Standard Operating Procedure (SOP) No. 100.10 - Incident Involving a Collector, Shooter and Hunter (CAC), the military police team must follow the following steps:

1. Approach with caution: Approach the person with reasonable suspicion of carrying a firearm. fire, observing PMPR protocols and techniques. Remove the weapon that the person being approached is carrying, even if he identifies himself as a CAC, until his identity is confirmed and all documentation is checked.
2. Search safely: Conduct a personal and vehicle search if there is reasonable suspicion, to check that the person does not have other firearms or irregularities. The search must be carried out with attention to the safety of the team and the person being approached, who must remain at a safe distance and with their hands visible.
3. Request documentation: Request the following documents: Collector, Shooter and Hunter Registration Certificate (CRCAC) Traffic Guide (GT) - for the transportation of weapons, ammunition and accessories, in accordance with Art. 82 of Decree No. 10,030, of September 30, 2019.
4. Check compliance and validity: Check the validity and compliance of the documents with the models attached to the POP. The absence or inadequacy of any document implies the arrest and referral of the individual to the competent police authority (Art. 14 or Art. 16 of Law 10.826/2003).
5. Check transportation conditions: The CAC is not authorized to carry firearms, and the weapons being transported must be unloaded and safely stored. If these requirements are not met, the team must refer the individual to the police authorities (Art. 14 or Art. 16 of Law 10.826/2003).
6. Report to SFPC5: Report to the Controlled Products Inspection Service of the 5th Military Region (SFPC5) any occurrence that results in the CAC being forwarded to the Police Station for administrative measures to be taken.

Remember that this procedure must be carried out in accordance with PMPR standards and in compliance with federal legislation, such as Law 10.826/2003 and Decree 10.030/2019.

Additional guidelines:

- It is crucial to maintain the safety of the team during the approach and search.
- The team must be aware of the laws and regulations that regulate the possession, carrying and transportation of weapons by CACs.
- Communication between police officers and monitoring of the incident are essential for successful response.

Important: If you have any questions about the application of the POP, consult your superior.

Notes:

- The examples presented suggest that the prototype is capable of understanding user questions and providing relevant answers based on information contained in the PMPR guidelines and doctrines. The responses demonstrate Gemini AI's ability to generate coherent, concise, and informative texts, as well as extract and synthesize relevant information from documents.
- Including references to PMPR documents in responses increases the reliability and transparency of the system, allowing the user to verify the origin of the information.
- It is important to emphasize that these examples are merely illustrative and do not represent an exhaustive evaluation of the prototype. More robust tests, with performance metrics and systematic qualitative evaluations, with the participation of a representative group of users, will be carried out in future stages of development.

4. DISCUSSION

The results presented in the previous section, although preliminary and based on a limited set of tests, suggest that the documentary research prototype with RAG and Gemini AI has great potential to assist military police officers in accessing relevant information from PMPR guidelines and doctrines.

The prototype's ability to understand natural language questions and provide concise and informative answers, with references to PMPR documents, represents a significant advance in relation to the way in which documentary research is currently conducted at the institution. Currently, PMPR does not have a centralized documentary research system. Military police officers rely on manual searches in folders on the intranet, organized by physical location and file name, which makes the process of finding specific information inefficient and time-consuming. This manual search requires the officer to know exactly where the document is stored and what its name is, which is not always possible, especially in urgent situations.

The use of the RAG strategy, combined with the Gemini AI language model, allows the prototype to go beyond simple keyword searches. The system is able to understand the semantic meaning of queries and search for relevant documents based on similarity in meaning, even if they do not contain the exact keywords of the query. This semantic search capability is essential to deal with the complexity and variety of terms and expressions present in PMPR guidelines and doctrines, which can be written in legal or technical language, with specific terms that may not be known to all police officers.

Furthermore, Gemini AI's natural language response generation makes interaction with the system more intuitive and user-friendly. Responses are presented in a clear and concise manner, with excerpts from relevant documents included to support the information provided. This approach helps to increase the system's reliability and transparency, allowing the user to understand how the response was generated and to verify its validity, which is essential to ensuring the reliability of the information used in decision-making in real-world situations.

4.1 PROTOTYPE LIMITATIONS

It is important to recognize that the prototype is still in the early stages of development and has some limitations:

- **Limited dataset:**The prototype was trained with a dataset restricted to documents made available by the PMPR, which, although comprehensive, does not represent all the information relevant to police activity. The inclusion of other sources of information, such as federal and state legislation, standards from other public security agencies, technical manuals and scientific articles, could enrich the system's knowledge and improve the quality of responses, allowing the prototype to answer questions on broader and more specific topics.
- **Absence of formal testing:**The prototype evaluation was carried out with a limited set of questions and without the use of performance metrics and qualitative assessments.

systematic. Conducting more robust tests, with the participation of a representative group of users, is essential to evaluate the system's performance in different scenarios, identify its deficiencies and direct future work. The use of metrics such as precision, recall and F1-score would allow quantifying the system's ability to retrieve relevant documents and provide correct answers. Qualitative evaluations, in turn, could collect users' opinions on the system's usability, the clarity of the answers and the usefulness of the tool for their work.

- **Simple User Interface:**The prototype's user interface is still quite basic, limited to sending questions and receiving answers in text format. The development of a more intuitive and user-friendly interface, with features such as search history, search filters, document viewing, highlighting of relevant sections in answers and the possibility of voice interaction, could significantly improve the user experience and facilitate the use of the system on different devices, such as smartphones and tablets.
- **Gemini AI API Dependency:**The prototype depends on the availability and proper functioning of the Gemini AI API, an external service provided by Google. The use of alternative language models, such as open source models or models hosted on PMPR's own servers, or the implementation of *fallback* to deal with potential failures in the Gemini AI API, such as using a smaller and faster language model to answer simple questions, could increase the robustness and autonomy of the system, ensuring its operation even in the event of unavailability of the Gemini AI API.

4.2 FUTURE WORK

The identified limitations point to promising directions for future work:

- **Dataset Expansion:**Integrate other sources of information relevant to police activity, such as federal and state legislation, standards from other public security agencies, technical manuals and scientific articles, to enrich the system's knowledge and expand the scope of questions that the prototype can answer.
- **Conducting formal tests:**Evaluate the prototype's performance with a comprehensive set of questions and performance metrics, such as precision, recall and F1-score, in addition to carrying out qualitative evaluations with the participation of military police officers, to collect feedback on the system's usability and the quality of the responses.
- **Developing an advanced user interface:**Create a more intuitive and user-friendly interface, with features such as search history, search filters, document viewing, highlighting of relevant sections in responses and the possibility of voice interaction, to improve the user experience and facilitate the use of the system on different devices.
- **Exploring alternative language models:**Investigate the possibility of using other language models, such as open source models or models hosted on PMPR's own servers, to increase the flexibility, robustness and autonomy of the system, reducing dependence on external services.
- **Integration with other PMPR systems:**Integrate the prototype with other systems used by the PMPR, such as the incident management system and the police intelligence system, to offer military police officers unified access to relevant information, allowing the prototype to access contextual information about the ongoing incident or the suspect's profile, for example, to provide more personalized and relevant responses.

.

5. CONCLUSION

This work presented the development of a documentary research prototype for guidelines and doctrines of the Military Police of Paraná (PMPR) using the Retrieval Augmented Generation (RAG) strategy and the Gemini AI Flash 1.5 language model, in a containerized environment with Docker. The main objective was to create a tool that would facilitate the access of military police officers to information relevant to the performance of their functions, overcoming the limitations of the current form of documentary research in the institution, based on manual search in folders on the intranet.

The results obtained, although preliminary and based on a limited set of tests, demonstrate the prototype's potential to assist military police officers in accessing relevant information from PMPR guidelines and doctrines. The system's ability to understand questions in natural language and provide

concise and informative responses, with references to PMPR documents, represents a significant advance in relation to the way documentary research is currently carried out.

The use of the RAG strategy, combined with the Gemini AI language model, allows the prototype to go beyond simple keyword searching, performing semantic search and generating responses in natural language that increase the trust and transparency of the system.

Despite the demonstrated potential, the prototype is still in the early stages of development and has limitations that need to be addressed in future work, such as expanding the data set, carrying out formal tests with performance metrics and qualitative evaluations, developing a more advanced user interface, and exploring alternative language models.

It is believed that the continuation of this work, with the implementation of the proposed improvements, could result in a valuable tool for the PMPR and other public security institutions. The provision of an efficient and user-friendly search system could significantly contribute to improving the quality of police work, facilitating access to relevant information, promoting uniformity of procedures and assisting in more effective decision-making, which could have a positive impact on the security of Paraná society.

In short, this work represents a promising initial step in the development of innovative solutions for documentary research in the area of public security, with the potential to positively impact the performance of the PMPR and the security of Paraná society.

REFERAND**NCIES**

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.

BRAZIL. Ministry of Justice and Public Security.**Ordinance No. 332, of April 27, 2023.**Provides for the Police Force Use Manual. Official Gazette of the Union, Brasília, DF, April 28, 2023. Section 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.**Available at: https://n8n.io/. Accessed on: November 6, 2024.

QDRANT.**Qdrant - Vector Database.**Available at: https://qdrant.tech/. Accessed on: November 6, 2024.