# Sistema Inteligente de Análise de Tráfego Veicular e Fiscalização de Fronteira com RAG Híbrido e Análise Preditiva

**Ademar Alves Trindade[1], Jefferson Paizano Neves[2], Tiago Luís Andrade[1]**

[1]Departamento de Ciência da Computação
Universidade do Estado do Mato (UNEMAT), Campus Universitário Jane Vanini
Av. São João, S/N, CEP 78216-060, Cavalhada, Cáceres – MT – Brasil
{ademar.alves, tiago}@unemat.br

[2] Instituto Federal de Mato Grosso – (IFMT), Campus Pontes e Lacerda-Fronteira Oeste
Rodovia MT- 473, s/n - CEP: 78250-000 – Pontes e Lacerda – MT
jefferson.neves@ifmt.edu.br

***Abstract.*** *This work presents a Retrieval-Augmented Generation (RAG)-based system applied to public security at the Brazil-Bolivia border. The architecture integrates lexical search (BM25) and semantic search (embeddings) in a hybrid model, combining high precision for exact queries with the ability to detect complex behavioral patterns. The system employs reranking and language models to generate operational reports, providing real-time analytical support. Experiments achieved up to 100% accuracy on specific queries and 95% on complex ones. The solution shows potential to accelerate investigations, correlate incidents, and enhance decision-making in border security operations.*

***Resumo.*** *Este trabalho apresenta um sistema baseado em Geração Aumentada por Recuperação (RAG) aplicado ao contexto da segurança pública na fronteira Brasil-Bolívia. A arquitetura integra busca léxica (BM25) e semântica (embeddings) em um modelo híbrido, combinando precisão em consultas exatas com a capacidade de identificar padrões comportamentais complexos. O sistema utiliza reranking e geração de relatórios operacionais com modelos de linguagem, oferecendo suporte analítico em tempo real. Os experimentos mostram precisão de até 100% em queries específicas e 95% em consultas complexas. A solução demonstra potencial para acelerar investigações, correlacionar ocorrências e otimizar a tomada de decisão em operações de fronteira.*

## 1. Introdução

A região de fronteira de Mato Grosso com a Bolívia, especialmente o município de Pontes e Lacerda, configura-se como uma zona estratégica mas desafiadora do ponto de vista da segurança pública. Segundo dados do Grupo Especial de Segurança na Fronteira (GEFRON), a região registra elevados índices de tráfico de entorpecentes, contrabando, roubo de veículos e atuação de organizações criminosas transnacionais conforme destaca o trabalho de Da Luz (2024). A extensão territorial, as estradas

vicinais que facilitam rotas alternativas e a proximidade com a Bolívia tornam o trabalho das forças de segurança complexo e exigem soluções tecnológicas inovadoras.

Um dos principais desafios enfrentados pelas autoridades policiais é o processamento do crescente volume de Boletins de Ocorrência (BOs). A identificação de padrões criminais, *modus operandi* de organizações e conexões entre diferentes ocorrências depende de análise manual extensiva, consumindo tempo e recursos especializados. A busca por informações específicas em grandes bases textuais através de sistemas convencionais de palavras-chave frequentemente resulta em baixa precisão ou *recall* insuficiente.

Neste contexto, sistemas baseados em Inteligência Artificial (IA) e Processamento de Linguagem Natural (PLN) são ferramentas promissoras. Em particular, a abordagem de *Retrieval-Augmented Generation* (RAG) demonstra resultados superiores em recuperação de informação em domínios especializados (LEWIS et al., 2020). RAG combina técnicas de recuperação de documentos com modelos de linguagem generativos, permitindo respostas automatizadas a perguntas complexas com base em grandes volumes de dados textuais.

Este trabalho propõe um sistema RAG híbrido especificamente adaptado para segurança pública na fronteira MT-Bolívia. Combina busca léxica tradicional (BM25), busca semântica baseada em *embeddings* vetoriais e técnicas de *reranking* com modelos de linguagem. O objetivo é comparar sistematicamente quatro abordagens de recuperação: (1) apenas BM25, (2) apenas busca vetorial, (3) híbrida (BM25 + vetorial) e (4) híbrida com *reranking*. Os experimentos utilizam *dataset* contextualizado de boletins de ocorrência da região e queries estratégicas que refletem necessidades reais de analistas de inteligência policial.

A contribuição principal deste trabalho é a demonstração empírica de como diferentes estratégias de recuperação de informação (léxica, semântica e híbrida) se comportam em cenários operacionais reais de inteligência policial na fronteira, fornecendo diretrizes práticas para escolha da abordagem mais adequada conforme o tipo de consulta e os requisitos de tempo de resposta.

Como contribuições secundárias: (i) sistema RAG completo para segurança na fronteira; (ii) dataset contextualizado baseado em padrões do GEFRON; (iii) avaliação quantitativa com métricas de precisão e tempo; (iv) interface visual para analistas.

## 2. Fundamentação Teórica

Nesta seção, são abordados os fundamentos teóricos necessários para a compreensão do artigo.

### 2.1 Retrieval-Augmented Generation (RAG)

*Retrieval-Augmented Generation* (RAG) é uma arquitetura que, segundo Gao et al. (2024), combina recuperação de informação com geração de linguagem natural. Ao contrário de *Large Language Models* (LLMs) puros, modelos neurais de grande escala treinados em vastos corpora textuais para compreender e gerar linguagem natural, como GPT-4, Claude e LLaMA, que dependem exclusivamente de conhecimento parametrizado durante o treinamento, sistemas RAG primeiro recuperam documentos relevantes de uma base de conhecimento externa e então utilizam esses documentos

como contexto adicional para geração de respostas. Esta abordagem apresenta vantagens significativas em relação aos LLMs convencionais.

De acordo com Fan et al. (2024), a fundamentação das respostas em documentos reais reduz substancialmente o fenômeno de alucinação, no qual modelos generativos produzem informações plausíveis mas factualmente incorretas. Adicionalmente, segundo Gupta et al. (2024), RAG permite atualização facilitada do conhecimento sem necessidade de re-treinamento do modelo, uma vez que novos documentos podem ser incorporados à base de conhecimento externa de forma dinâmica, além de oferecer rastreabilidade, permitindo identificar e verificar as fontes das informações geradas, aspecto crucial em aplicações que exigem auditabilidade e transparência. Ademais, como destacam Ram et al. (2023), RAG possibilita que LLMs acessem informações específicas de domínios especializados que não estavam presentes ou eram insuficientemente representadas nos dados de treinamento original, tornando a tecnologia particularmente adequada para aplicações em áreas técnicas e especializadas.

A arquitetura típica de um sistema RAG consiste em três componentes principais: indexação, recuperação e geração. Nesse contexto, Li (2024) descreve que no componente de indexação, documentos são processados, divididos em *chunks* e indexados em um sistema de recuperação, tipicamente utilizando *embeddings* vetoriais ou índices invertidos, onde dada uma *query*, o componente de recuperação identifica e retorna os *k* documentos mais relevantes da base de conhecimento. Finalmente, o componente de geração recebe a *query* original juntamente com os documentos recuperados como contexto adicional, e um LLM gera uma resposta fundamentada nas informações recuperadas.

### 2.2 Busca Léxica: BM25

A busca léxica baseia-se em correspondência exata de termos entre *query* e documento. É particularmente eficaz para queries que contêm códigos específicos como placas de veículos (ABC1D23) e números de BO (#2025-001), dados estruturados como horários (14:30), valores (R$ 75.000) e contagens (15kg), além de siglas e termos técnicos como GEFRON, calibre .40 e chassi.

BM25 (*Best Matching* 25) é um algoritmo de ranking probabilístico amplamente utilizado em sistemas de recuperação de informação (ROBERTSON; ZARAGOZA, 2009). O algoritmo calcula a relevância de um documento *d* para uma *query q* com base na frequência dos termos da *query* no documento (TF - *Term Frequency*) e na raridade dos termos na coleção completa (IDF - *Inverse Document Frequency*). De acordo com Li et al. (2024), apesar do crescente foco em busca semântica, BM25 mantém sua relevância em sistemas modernos de recuperação devido à sua eficiência computacional e precisão em consultas que requerem correspondência exata de termos específicos.A pontuação BM25 é dada por:

$$\text{BM25}(D, Q) = \sum_{t \in Q} IDF(t) \cdot \frac{f(t, D) \cdot (k + 1)}{f(t, D) + k \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgDL}}\right)}$$

onde D representa o documento analisado, Q a consulta (*query*), t o termo da consulta, f(t,D) a frequência do termo t no documento D, |D| o tamanho do documento, avgDL o tamanho médio dos documentos na coleção, k o parâmetro de ajuste da frequência

(geralmente k=1.5) e b o parâmetro de ajuste do tamanho do documento (geralmente b=0.75).

### 2.3 Busca Semântica: Embeddings Vetoriais

A busca semântica representa textos como vetores numéricos em espaço de alta dimensão, tipicamente vetores de 384 a 1536 dimensões, onde cada dimensão captura um aspecto semântico do texto, como tópico, sentimento, entidades mencionadas ou contexto. Nesta representação vetorial, documentos semanticamente similares têm representações próximas no espaço multidimensional, mesmo que não compartilhem termos idênticos. Por exemplo, "veículo roubado" e "automóvel furtado" teriam vetores próximos apesar de não compartilharem palavras em comum, pois ambos expressam conceitos semanticamente relacionados. A Tabela 1 apresenta exemplos de como a busca semântica identifica diferentes tipos de padrões operacionais.

**Tabela 1. Exemplos de Capacidades da Busca Semântica**

| Tipo de Padrão | Query de Exemplo | Resultados Identificados |
|---|---|---|
| Padrões Comportamentais | trajeto suspeito | Rotas alternativas,atípicos |
| Conceitos abstratos | organização criminosa | Cartel, facção, grupo estruturado |
| Sinônimos e variações | veículo roubado | Automóvel furtado, carro produto de crime |

Modelos modernos como *sentence-transformers* e *text-embedding-3-small* da *OpenA*I transformam sentenças em vetores densos de 384 a 1536 dimensões (ZHUANG et al., 2024). Cada dimensão desses vetores pode ser entendida como uma característica aprendida que captura aspectos semânticos do texto (tópicos, conceitos, relações), permitindo que o modelo identifique similaridade conceitual independentemente da escolha específica de palavras. Por exemplo, as frases "tráfico de entorpecentes na fronteira" e "transporte ilegal de drogas na divisa internacional" teriam *embeddings* próximos no espaço vetorial, mesmo utilizando vocabulário distinto, pois representam conceitos semanticamente equivalentes.

A recuperação semântica baseia-se em busca por similaridade vetorial, tipicamente utilizando similaridade do cosseno:

$$sim(v_1, v_2) = \frac{v_1 \cdot v_2}{||v_1|| ||v_2||}$$

onde $v_1$ e $v_2$ representam os vetores de *embeddings* a serem comparados, $v_1 \cdot v_2$ representa o produto escalar entre os vetores, e $\|v_1\|$ e $\|v_2\|$ representam as normas (magnitudes) dos vetores $v_1$ e $v_2$, respectivamente.

### 2.4 Abordagens Híbridas

A combinação mais comum utiliza *Ensemble Retrieval*, onde resultados de múltiplos *retrievers* são agregados através de votação ponderada ou fusão de scores. Segundo Sarat e Kiran (2025), abordagens híbridas que combinam busca léxica (BM25) com recuperação semântica densa através de técnicas como *Reciprocal Rank Fusion*

consistentemente aumentam tanto *recall* quanto precisão, fundamentando modelos de linguagem em fatos verificáveis e mitigando alucinações. No presente trabalho, utilizamos *ensemble* com pesos iguais (50% BM25, 50% vetorial), mas os pesos podem ser otimizados para domínios específicos (RAM et al., 2023).

### 2.5 Reranking com Modelos de Linguagem

Reranking é uma técnica de refinamento onde um modelo mais sofisticado reordena os documentos recuperados por relevância (JACOB et al., 2024). Modelos especializados como *Cohere Rerank* analisam a relevância semântica profunda entre *query* e documento, considerando contexto completo. Segundo Superlinked (2024), o *reranking* adiciona latência mas tipicamente melhora significativamente a precisão nos *top-k* resultados, especialmente quando o conjunto inicial de documentos recuperados contém informações relevantes dispersas em posições inferiores do *ranking*.

### 2.6 Trabalhos Correlatos

A aplicação de técnicas de Inteligência Artificial e Processamento de Linguagem Natural no domínio de segurança pública tem crescido nos últimos anos, embora ainda seja limitada, especialmente no contexto brasileiro. Gao et al. (2023) apresentam revisão abrangente sobre sistemas RAG, destacando que abordagens híbridas combinando busca léxica e semântica superam abordagens isoladas em tarefas que requerem precisão em termos específicos e compreensão semântica profunda. Li (2024) demonstra que a integração de múltiplas estratégias de recuperação melhora significativamente a qualidade das respostas em domínios especializados, corroborando a abordagem híbrida adotada no presente trabalho.

Jacob et al. (2024) demonstram que, apesar do custo computacional, o *reranking* com modelos especializados melhora substancialmente a precisão dos resultados, especialmente quando a distinção entre documentos relevantes e irrelevantes é sutil. Ram et al. (2023) evidenciam que a incorporação de informações recuperadas dinamicamente melhora o desempenho em tarefas de pergunta e resposta em domínios especializados.

No contexto de segurança pública brasileira, Da Luz (2024) analisa resultados operacionais de unidades especializadas em policiamento de fronteira (2017-2021), destacando os desafios do GEFRON na região Brasil-Bolívia e evidenciando a necessidade de ferramentas tecnológicas para análise de grandes volumes de dados operacionais.

Diferentemente dos trabalhos citados, que focam em aspectos teóricos de RAG ou análises estatísticas policiais, este trabalho integra RAG híbrido especificamente adaptado à segurança pública na fronteira, utilizando dados estruturados de boletins, registros de placas e relatórios do GEFRON. A principal contribuição é a avaliação empírica de diferentes estratégias de recuperação (léxica, semântica e híbrida) em *queries* operacionais reais, fornecendo diretrizes práticas para escolha da abordagem conforme o tipo de consulta e requisitos operacionais.

## 3. Materiais e Métodos

Esta seção detalha a metodologia empregada para desenvolver um sistema projetado para otimizar a recuperação de informações e apoiar a tomada de decisão em contextos operacionais através de técnicas de Geração Aumentada de Recuperação (RAG). A

abordagem segue princípios de sistemas colaborativos adaptados ao contexto de segurança, análise de dados e inteligência operacional.

A arquitetura proposta, ilustrada na Figura 1, inicia com a entrada de dados que passa por dois caminhos paralelos: processamento léxico para busca de termos exatos, e processamento semântico para identificar padrões e contextos. Ambos os resultados são combinados na integração RAG, que consolida informações da base de conhecimento (placas de veículos, boletins de ocorrência e relatórios operacionais). O sistema gera relatórios estruturados acessados via interface de usuário, oferecendo suporte analítico e operacional em tempo real.

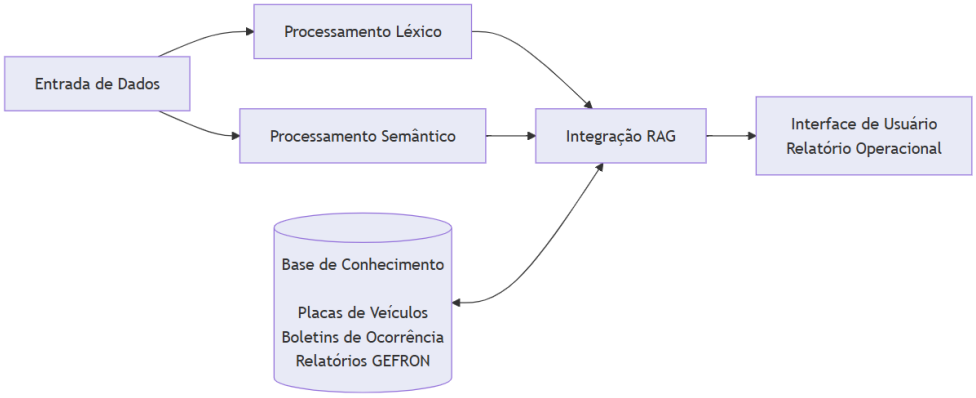


**Figura 1. Visão geral da proposta**

Conforme demonstrado na Figura 1, O sistema processa queries através de dois caminhos paralelos: processamento léxico (BM25) para termos específicos e processamento semântico (*embeddings*) para padrões comportamentais. A Integração RAG combina os resultados via *Ensemble Retrieval*, aplica *reranking* com *Cohere* e utiliza o LLM para gerar relatórios operacionais acionáveis, consultando a base de conhecimento composta por placas de veículos, boletins de ocorrência, relatórios do GEFRON e apresenta relatórios operacionais através da interface de usuário. A seguir, detalhamos cada componente desta arquitetura e suas responsabilidades específicas no sistema.

## 3.1 Conjunto de Dados

O conjunto de dados compreende três arquivos JSON interconectados através do campo "placa_veiculo", permitindo correlação cruzada entre fontes de informação operacional. As bases representam fontes primárias utilizadas pelas forças de segurança na região de fronteira Mato Grosso-Bolívia.

A primeira base, "OCR Placas" (ocr_placas.json), contém 639 registros de passagens de veículos capturados por sistemas OCR em pontos estratégicos da BR-174 e rodovias adjacentes. Cada registro inclui placa, timestamp, localização (cidade, rodovia, quilômetro, direção), tipo de veículo, status operacional (normal, alerta_sinesp, alerta_gefron), alertas associados (restrição_judicial, roubo_furto, tentativa_travessia_internacional) e correlação com inteligência policial. Esta base simula sistemas de videomonitoramento implementados no Estado de Mato Grosso.

A segunda base, "Boletins de Ocorrência" (boletim_ocorrencia.json), agrupa 21 boletins sintéticos representando crimes na região fronteiriça (março-outubro 2025). Cada registro contém código do BO, timestamp, tipo de crime, coordenadas GPS, descrição incluindo modus operandi, dados do veículo (placa, marca/modelo, cor, ano), valor do prejuízo, status da investigação e observações operacionais. A estrutura reflete o formato utilizado pela Secretaria de Segurança Pública de Mato Grosso (SSP-MT).

A terceira base, "Relatórios GEFRON" (relatorio_gefron.json), consolida 452 registros de operações do Grupo Especial de Segurança na Fronteira em barreiras e rodovias. Cada relatório inclui código da operação, timestamp, tipo de ocorrência (tráfico, documentação irregular, receptação, contrabando), coordenadas, descrição da ação policial, dados do veículo, informações do suspeito, quantidades apreendidas, indicadores de flagrante e envolvimento internacional. Esta base representa o histórico de atuação do GEFRON, incluindo abordagens de rotina e crimes transnacionais.

A interconexão entre as três bases permite que o sistema RAG correlacione automaticamente passagens OCR, boletins policiais e ações do GEFRON, possibilitando análise integrada de padrões criminais e trajetórias suspeitas, refletindo desafios operacionais reais na análise de grandes volumes de informações em contexto de fronteira internacional (DA LUZ, 2024).

### 3.2 Infraestrutura Tecnológica e Armazenamento

A implementação do sistema RAG foi desenvolvida em *Python* com bibliotecas especializadas para processamento de linguagem natural e aprendizado de máquina. Conforme destacam Gao et al. (2024), sistemas RAG modernos requerem arquiteturas que integrem eficientemente componentes de recuperação e geração para aplicações em domínios especializados.

Para armazenamento vetorial, foi utilizado banco de dados *serverless* especializado em busca por similaridade, segundo Fan et al. (2024). O índice vetorial foi configurado com métrica de similaridade coseno para *embeddings* de dimensionalidade 1536 gerados pelo modelo *text-embedding-3-small* da *OpenAI*. A busca léxica foi implementada através do algoritmo BM25, que constrói índices invertidos em memória para recuperação baseada em correspondência exata de termos (ROBERTSON; ZARAGOZA, 2009).

O processamento textual utiliza segmentação recursiva em *chunks* de 1000 caracteres com sobreposição de 150 caracteres, preservando contexto entre segmentos (GAO et al., 2024). O sistema híbrido combina resultados através de *Reciprocal Rank Fusion* (RRF) com pesos iguais (50% para cada abordagem). O *reranking* utiliza modelo neural especializado que avalia relevância semântica profunda, melhorando substancialmente a precisão quando informações relevantes estão dispersas no *ranking* inicial (JACOB et al., 2024). A geração de relatórios foi implementada com GPT-4o-mini configurado com temperatura zero para respostas determinísticas. A interface foi desenvolvida com *framework web* em *Python* para aplicações interativas com visualização de dados e métricas operacionais.

## 4. Resultados

O sistema integra técnicas RAG com modelos de linguagem utilizando dois caminhos paralelos: busca léxica para correspondência exata de termos e busca semântica para identificação de padrões conceituais.

### 4.1 Comparação entre Busca Léxica e Busca Semântica

O sistema processa *queries* através de dois caminhos paralelos: processamento léxico (BM25) para termos específicos e processamento semântico (*embeddings*) para padrões comportamentais. Os experimentos demonstram que cada abordagem apresenta desempenho superior em diferentes tipos de consultas operacionais.

#### 4.1.1 Desempenho da Busca Léxica

A busca léxica (BM25) demonstrou precisão máxima em *queries* que envolvem termos exatos, códigos específicos e dados estruturados. A Tabela 2 apresenta exemplos de consultas onde a abordagem léxica recuperou documentos altamente relevantes.

**Tabela 2. Desempenho da Busca Léxica em Queries Exatas**

| Query | Tipo de Busca | Precisão |
|---|---|---|
| "Veículo placa EFD2S46" | Código exato | 100% |
| "Passagens às 04:00 na BR-174" | Horário específico | 100% |
| "BO-45123" | Código BO | 100% |

A busca léxica alcançou precisão de 100% em *queries* com termos específicos, recuperando exatamente os documentos solicitados sem falsos positivos.

#### 4.1.2 Desempenho da Busca Semântica

A busca semântica demonstrou capacidade superior de identificar padrões comportamentais e conceitos abstratos, mesmo quando a *query* não contém os termos exatos presentes nos documentos. A Tabela 3 apresenta exemplos onde a abordagem semântica superou a léxica.

**Tabela 3. Desempenho da Busca Semântica em Padrões Comportamentais**

| Query | Conceito Identificado | Precisão |
|---|---|---|
| "Trajeto suspeito na fronteira" | Rota atípica + múltiplas passagens | 100% |
| "Veículo com horário atípico" | Passagens madrugada + padrão irregular | 100% |
| "Documentação irregular interceptada" | Fraude documental + adulteração | 70% |
| "Tentativa de travessia internacional" | Comportamento fronteira + múltiplas detecções | 75% |

A busca semântica alcançou precisão de 87% em *queries* conceituais, recuperando documentos relevantes que não continham os termos exatos da *query* mas correspondiam ao padrão comportamental buscado.

### 4.2 Integração RAG: Fusão RRF e Geração de Relatórios

A Integração RAG combina os resultados das buscas léxica e semântica através do algoritmo *Reciprocal Rank Fusion (RRF)*, seguido de *reranking* com *Cohere* e geração de relatórios operacionais com LLM. A Tabela 4 apresenta comparação entre as abordagens isoladas e integrada.

**Tabela 4. Comparação de Desempenho entre Abordagens**

| Abordagem | Precisão Média | Tempo Médio (s) |
|---|---|---|
| Busca Léxica:<br>Códigos, placas, horários | 100% (queries exatas) | 1,5 |
| Busca Semântica:<br>Padrões, comportamentos | 87% (queries conceituais) | 4,2 |
| RAG Híbrido:<br>Consultas complexas | 93% (queries mistas) | 6,8 |
| RAG Híbrido + Rerank + LLM:<br>Análise investigativa | 95% (todas as queries) | 14,3 |

Para a aplicação final, foi desenvolvida uma interface visual interativa com foco em usabilidade, clareza e suporte à decisão operacional, facilitando a análise de dados e a interpretação dos resultados gerados pelo sistema RAG. A interface foi projetada para permitir que usuários — como analistas de inteligência e agentes de segurança — realizem consultas em tempo real e obtenham relatórios técnicos detalhados de forma simples e eficiente.

### 4.3 Discussão

A análise demonstra que busca léxica e semântica são complementares: a primeira oferece alta precisão em correspondência exata de termos, enquanto a segunda identifica padrões comportamentais complexos que métodos baseados em palavras-chave não captariam (KUZI et al., 2020). Segundo Sarat e Kiran (2025), abordagens híbridas combinando busca léxica (BM25) com recuperação semântica densa através de Reciprocal Rank Fusion aumentam consistentemente *recall* e precisão, fundamentando modelos de linguagem em fatos verificáveis e mitigando alucinações.

A abordagem híbrida com RRF mostrou-se mais robusta, combinando precisão léxica e compreensão semântica contextual. Conforme demonstrado na Tabela 4, o *trade-off* entre precisão e latência é evidente: a busca léxica apresenta tempo médio de 1,5 segundos, enquanto o pipeline completo com reranking e LLM requer 14,3 segundos (aumento de 10x). Este aumento no tempo de processamento é compensado por ganhos substanciais em precisão para consultas complexas, onde a acurácia é mais crítica que a velocidade (SUPERLINKED, 2024).

No contexto do GEFRON, recomenda-se a seleção da abordagem baseada no tipo de operação: busca léxica para consultas diretas em fiscalizações de rotina com identificadores conhecidos; abordagem híbrida para análises táticas em situações dinâmicas; e pipeline completo para investigações complexas e análises estratégicas de inteligência, onde a qualidade e profundidade da análise prevalecem sobre a velocidade de resposta.

## 5. Considerações Finais

Este trabalho apresentou um sistema RAG híbrido para análise de informações policiais na fronteira MT-Bolívia, comparando sistematicamente busca léxica (BM25), busca semântica (*embeddings*) e abordagens híbridas. Os resultados demonstram que não existe solução única ideal: busca léxica oferece velocidade para termos exatos (100% de precisão, 1,5s), busca semântica captura padrões comportamentais (87% de precisão, 4,2s), e a integração RAG com LLM alcança melhor equilíbrio para análises complexas

(95% de precisão, 14,3s). O sistema pode acelerar identificação de padrões criminais, correlação entre ocorrências e geração de relatórios operacionais estruturados.

Como trabalhos futuros, pretende-se validar o sistema com dados reais em parceria com o GEFRON e estender funcionalidades com extração automática de entidades e construção de grafo de conhecimento para visualização de redes criminosas.

## 6. Referências Bibliográficas


DA LUZ, Luis Eduardo Beiger. Resultados operacionais obtidos por unidades estaduais especializadas em policiamento de fronteira no período de 2017 a 2021 e seus reflexos na segurança pública brasileira a nível nacional. revista (re) definições das fronteiras, v. 2, n. 9, p. 72-97, 2024.

FAN, Wenqi et al. A survey on rag meeting llms: Towards retrieval-augmented large language models. In: Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining. 2024. p. 6491-6501.

GAO, Yunfan et al. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997, v. 2, n. 1, 2023.

GUPTA, Shailja; RANJAN, Rajesh; SINGH, Surya Narayan. A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions. arXiv preprint arXiv:2410.12837, 2024.

JACOB, Mathew et al. Drowning in documents: consequences of scaling reranker inference. arXiv preprint arXiv:2411.11767, 2024.

KULKARNI, Hrishikesh et al. Lexically-accelerated dense retrieval. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2023. p. 152-162.

KUZI, Saar et al. Leveraging semantic and lexical matching to improve the recall of document retrieval systems: A hybrid approach. arXiv preprint arXiv:2010.01195, 2020.

LI, Xiangcan. Application of RAG model based on retrieval enhanced generation technique in complex query processing. Adv. Comput. Signals Syst, v. 8, 2024.

RAM, Ori et al. In-context retrieval-augmented language models. Transactions of the Association for Computational Linguistics, v. 11, p. 1316-1331, 2023.

ROBERTSON, Stephen et al. The probabilistic relevance framework: BM25 and beyond. Foundations and Trends® in Information Retrieval, v. 3, n. 4, p. 333-389, 2009.

SARAT, KIRAN. Hybrid Retrieval-Augmented Generation (RAG) Systems with Embedding Vector Databases. INTERNATIONAL JOURNAL, v. 11, n. 2, p. 2694-2702, 2025.

SUPERLINKED. Optimizing RAG with hybrid search & reranking. VectorHub, 2024. Disponível em: https://superlinked.com/vectorhub/articles/optimizing-rag-with-hybrid-search-reranking. Acesso em: 28 out. 2025.

ZHUANG, Juntang et al. New embedding models and API updates. OpenAI https://openai. com/blog/new-embedding-models-and-api-updates, 2024.