Sistema RAG Híbrido para Inteligência Policial e Apoio à Tomada de Decisão na Fronteira de Mato Grosso

Ademar Alves Trindade

Universidade do Estado do Mato (UNEMAT), Campus Universitário Jane Vanini
Av. São João, S/N, CEP 78216-060, Cavalhada, Cáceres, MT - Brasil

ademar.alves@unemat.br

Abstract. *This paper presents a hybrid RAG system for border intelligence at the Brazil–Bolivia frontier, combining BM25 and semantic retrieval via Reciprocal Rank Fusion with Cohere reranking. Evaluation on a 1,000-record synthetic corpus over 20 queries shows BM25 dominating precision (P@10 = 0.515; MRR = 0.652) due to structured police vocabulary, while semantic retrieval performs poorly (P@10 = 0.065) due to embedding-domain mismatch. Geographic and cross-database queries achieve the strongest hybrid performance (P@10 = 0.675 and 0.625). A deterministic layer resolves license-plate tokenization limitations.*

Resumo. *Este trabalho apresenta um sistema RAG híbrido para inteligência de fronteira Brasil-Bolívia, combinando BM25 e busca semântica via Reciprocal Rank Fusion com reranking da Cohere. A avaliação sobre corpus sintético de 1.000 registros em 20 consultas revela que o BM25 domina em precisão (P@10 = 0,515; MRR = 0,652) pelo vocabulário policial estruturado, enquanto a busca semântica isolada apresenta desempenho reduzido (P@10 = 0,065) por incompatibilidade de domínio. Consultas por localização e cruzadas entre bases alcançam o melhor desempenho híbrido (P@10 = 0,675 e 0,625). Uma camada determinística resolve limitações de tokenização de placas.*

1. Introdução

A região de fronteira entre Mato Grosso e Bolívia, especialmente o município de Pontes e Lacerda, apresenta importância estratégica para a segurança pública, mas também desafios significativos. Segundo dados do Grupo Especial de Segurança na Fronteira (GEFRON), a região registra elevados índices de tráfico de entorpecentes, contrabando e roubo de veículos (DA LUZ, 2024). Somente em 2025, o GEFRON apreendeu 21,5 toneladas de drogas ilícitas, 224 veículos e 10 aeronaves, acumulando, desde 2019, prejuízos de R$ 442,7 milhões às organizações criminosas.

Um dos principais desafios é o processamento de grandes volumes de Boletins de Ocorrência (BOs). A identificação de padrões criminais, *modus operandi* e correlações depende de análise manual extensiva, consumindo tempo e recursos especializados. Nesse contexto, a abordagem Retrieval-Augmented Generation (RAG) tem demonstrado resultados superiores em domínios especializados (GAO et al., 2023), integrando busca léxica e semântica para fundamentar a geração em fatos verificáveis.

A principal contribuição deste trabalho é uma demonstração empírica, com protocolo formal e métricas reproduzíveis, de como estratégias léxica, semântica e

híbrida se comportam em cenários operacionais de inteligência policial na fronteira Brasil-Bolívia, fornecendo diretrizes práticas para escolha da abordagem mais adequada conforme o tipo de consulta.

2. Fundamentação Teórica

2.1 Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) combina técnicas de recuperação de informação com geração de linguagem natural, permitindo que modelos de linguagem utilizem conhecimento externo durante a produção de respostas (GAO et al., 2024). Diferentemente de LLMs puros, sistemas RAG recuperam documentos relevantes de uma base de conhecimento e os utilizam como contexto para geração textual. Essa abordagem reduz alucinações, permite atualização dinâmica de informações e oferece maior rastreabilidade das respostas, características particularmente relevantes em aplicações de segurança pública (FAN et al., 2024; GUPTA et al., 2024). Em sua arquitetura típica, o pipeline opera em três etapas: indexação de documentos em segmentos com representação vetorial ou léxica, recuperação dos *k* documentos mais relevantes e geração da resposta com base nas evidências recuperadas, possibilitando o acesso a conhecimento especializado ausente nos dados de treinamento do modelo.

2.2 Fundamentos de Recuperação de Informação

Abordagens de recuperação de informação podem ser classificadas em léxica, semântica ou híbrida. A busca léxica BM25 baseia-se em correspondência exata de termos, sendo eficaz para *queries* com códigos específicos, siglas técnicas e dados estruturados. Sua pontuação é dada por:

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

onde *k*=1,5 e *b*=0,75 são parâmetros padrão de ajuste de frequência e comprimento do documento (ROBERTSON et al., 2009).

A busca semântica, por sua vez, representa textos como vetores densos de alta dimensão e identifica similaridade conceitual independentemente de correspondência léxica exata. Abordagens híbridas combinam ambas as modalidades, integrando resultados via técnicas de fusão de *rankings*, tal como a *Reciprocal Rank Fusion* (RRF), o que tende a aumentar *recall* e precisão mitigando alucinações (SAWARKAR et al., 2024). Etapas adicionais de *reranking* baseadas em modelos neurais podem refinar a ordenação inicial, priorizando documentos mais alinhados à intenção semântica da consulta (JACOB et al., 2024).

2.3 Arquitetura do Sistema e Conjunto de Dados

O pipeline de recuperação implementa uma abordagem híbrida combinando busca léxica baseada em BM25 (k_1 = 1,5; b = 0,75) e busca vetorial com *embeddings text-embedding-3-small* (1.536 dimensões). A integração dos resultados é realizada por fusão de *rankings* via *Reciprocal Rank Fusion* (RRF), seguida por uma etapa de *reranking* semântico utilizando o modelo *rerank-multilingual-v3.0* da *Cohere*, mitigando o posicionamento subótimo de documentos semanticamente relevantes pelo

BM25. A geração final das respostas é realizada pelo modelo GPT-4o-mini com temperatura zero, visando respostas determinísticas (GAO et al., 2024; JACOB et al., 2024).

Para avaliação do sistema, foram construídas três bases sintéticas reprodutíveis (seed = 42) em conformidade com a LGPD: OCR Placas (500 registros de passagens nas BR-174, BR-364 e MT-246), Boletins de Ocorrência (250 boletins, 10 tipologias criminais) e Relatórios GEFRON (250 operações, 16 ilícitos). Os 1.000 registros foram indexados no *Pinecone* com particionamento por *namespace* e *pool* de placas compartilhado.

3. Protocolo Experimental e Resultados

3.1 Protocolo de Avaliação

A avaliação do sistema utilizou 20 consultas em linguagem natural, distribuídas em cinco categorias operacionais: placa específica, tipologia criminal, localização geográfica, recorte temporal e cruzamento entre bases. O gabarito de relevância (ground truth) foi estabelecido pela contagem direta dos documentos relevantes no corpus. O desempenho da etapa de recuperação foi avaliado por meio das métricas Precision@k, Recall@k, F1@k e Mean Reciprocal Rank (MRR), amplamente empregadas em sistemas de recuperação de informação (AGNIHOTRAM; SARKAR, 2025).

3.2 Desempenho Comparativo

A Tabela 1 apresenta as métricas médias sobre as 20 consultas. O BM25 alcançou o melhor desempenho em precisão (P@10 = 0,515; MRR = 0,652), confirmando a superioridade da busca léxica em corpora com vocabulário técnico estruturado (ROBERTSON et al., 2009). O MRR = 0,652 indica que, em mais de 65% das consultas, o documento mais relevante aparece entre os dois primeiros resultados, reduzindo o esforço de triagem do analista.

Tabela 1. Comparativo de métricas de recuperação (média — 20 queries, corpus de 1.000 registros , $k \in \{5, 10\}$).

Modo	P@5	R@5	F1@5	P@10	R@10	F1@10	MRR
BM25	0.530	0.074	0.114	0.515	0.139	0.180	0.652
Semântico	0.130	0.004	0.008	0.065	0.004	0.007	0.300
Híbrido	0.350	0.042	0.063	0.435	0.104	0.137	0.494

A abordagem semântica apresentou desempenho inferior em todas as métricas (P@10 = 0,065; MRR = 0,300), refletindo incompatibilidade estrutural entre o modelo *text-embedding-3-small*, treinado em texto geral, e o vocabulário técnico policial do corpus (GAO et al., 2023; KUZI et al., 2020). O MRR = 0,300 permaneceu estável em todas as configurações testadas, confirmando que o fenômeno é intrínseco ao domínio.

A abordagem híbrida (EnsembleRetriever com RRF, pesos 50/50) obteve P@10 = 0,435 e MRR = 0,494, superior ao modo semântico isolado, mas inferior ao BM25. Esse comportamento decorre do mecanismo do RRF: ao combinar rankings de forma uniforme, o sinal semântico sistematicamente inferior reduz a contribuição relativa do BM25, penalizando documentos que este teria ranqueado mais alto (KURU; KESKIN, 2025). Para o corpus GEFRON, composto por registros estruturados com terminologia

técnica policial, a busca léxica representa a estratégia de maior precisão, enquanto a abordagem híbrida preserva utilidade para consultas cujos termos não aparecem literalmente no corpus. O reranking neural (Cohere rerank-multilingual-v3.0), aplicado sobre os resultados híbridos antes da geração pelo LLM, melhora a ordenação final sem alterar as métricas de recuperação reportadas (JACOB et al., 2024).

3.3 Análise por Tipo de Consulta

A Tabela 2 apresenta o desempenho do modo híbrido (P@10) segmentado por tipo de consulta. Essa análise permite identificar diferenças de desempenho entre categorias operacionais relevantes para o contexto do GEFRON.

Tabela 2. P@10 por tipo de consulta no modo híbrido

Tipo	P@10(Híbrido)	*N*	Análise/Desempenho
Localização	0,675	4	Alta
Consulta Cruzada	0,625	4	Alta
Tipo de Crime	0,600	4	Moderada
Período/Data	0,275	4	Baixa
Placa	0,000	4	Nula

Consultas por localização (P@10 = 0,675) apresentaram o melhor desempenho individual, dado o caráter altamente discriminativo de topônimos como nomes de cidades e rodovias ("Pontes e Lacerda", "BR-174"), bem capturados tanto por BM25 quanto por *embeddings*. Consultas cruzadas entre bases (P@10 = 0,625) beneficiam-se de termos operacionais compartilhados entre os três *namespaces*, favorecendo a recuperação híbrida. Consultas por tipo de crime (P@10 = 0,600) refletem variações lexicais entre a linguagem das consultas e a terminologia dos documentos. Consultas por período e data apresentaram desempenho reduzido (P@10 = 0,275), decorrente da distribuição uniforme dos registros e da ausência de filtragem estruturada por data anterior à recuperação vetorial. Consultas por placa específica obtiveram P@10 = 0,000, confirmando limitação estrutural da tokenização de identificadores alfanuméricos, que são segmentados em subpalavras semanticamente vazias pelos modelos de linguagem. O sistema mitiga essa restrição em produção por meio de uma camada determinística de filtragem direta sobre os registros em memória, executada antes do pipeline RAG, cujos resultados são unidos aos documentos recuperados semanticamente, garantindo recuperação exata por igualdade de campo sem duplicação.

4. Conclusões e Trabalhos Futuros

Este trabalho apresentou um sistema RAG híbrido aplicado à análise de registros policiais na fronteira Mato Grosso–Bolívia, comparando busca léxica baseada em BM25, busca semântica com embeddings e abordagem híbrida sobre corpus sintético de 1.000 registros. Os resultados indicam que, em conjuntos de dados estruturados com vocabulário técnico policial, o BM25 apresentou maior precisão (P@10 = 0,515) em comparação ao modo semântico isolado (P@10 = 0,065), evidenciando limitações de embeddings de propósito geral em domínios especializados, achado que constitui diretriz prática para sistemas similares.

A abordagem híbrida amplia a cobertura semântica em consultas investigativas, permitindo recuperar documentos conceitualmente relacionados mesmo sem correspondência léxica direta. Assim, a escolha da estratégia deve considerar a natureza da consulta: BM25 para vocabulário técnico preciso, híbrido com reranking para análises táticas de cobertura ampla e camada determinística obrigatória para identificadores exatos como placas veiculares. Como trabalhos futuros, pretende-se validar o sistema com dados operacionais reais em colaboração com o GEFRON, investigar embeddings especializados para o domínio policial, implementar filtros temporais estruturados pré-recuperação e estender a abordagem para cooperação multi-estadual.

5. Referências Bibliográficas

AGNIHOTRAM, S.; SARKAR, A. Evaluating Precision and Recall at Retrieval Time in Retrieval-Augmented Generation Systems. American Journal of Computer Science and Technology, v. 8, n. 4, p. 1–10, 2025.

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.

KURU, M.; KESKIN, M. Deep Retrieval at CheckThat! 2025: Identifying Scientific Papers from Implicit Social Media Mentions via Hybrid Retrieval and Re-Ranking. In: Conference and Labs of the Evaluation Forum (CLEF 2025). Working Notes Proceedings, 2025.

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.

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.

SAWARKAR, Kunal; MANGAL, Abhilasha; SOLANKI, Shivam Raj. Blended RAG: Improving RAG (retriever-augmented generation) accuracy with semantic search and