




# EcoLex RAG: Sistema HyPA-RAG com Legal-BERTimbau para Suporte Explicável à Decisão em Gestão Ambiental Brasileira

*EcoLex RAG: HyPA-RAG System with Legal-BERTimbau for Explainable Decision Support in Brazilian Environmental Management*

Waldemiro Negreiros <sup>1</sup>, Delvek Sousa<sup>1</sup>, Leonardo Estevam <sup>1</sup>, Suellen Negreiros<sup>1</sup>, Vitor Melo<sup>1</sup>, Filipe Corrêa<sup>1</sup>, Saulo Oliveira<sup>1</sup>, Carlos Motta<sup>1</sup>, Paulo Pontes<sup>2</sup>, Marcos Seruffo <sup>1</sup>,

<sup>1</sup>Universidade Federal do Pará (UFPA)  
Caixa Postal 66075-110 – Belém – Pará – Brasil

<sup>2</sup>Instituto Federal do Pará (IFPA)  
Caixa Postal 66093-020 – Belém – Pará – Brasil

waldemiro.negreiros@ifpa.edu.br, suellen\_dionisio@hotmail.com  
{leonardoestevam0605, filipecorreal3976, vitorhugomelo9, }@gmail.com  
{Delveck29, saulomattheus, amotta.eti.br, paulo9pontes}@gmail.com  
seruffo@ufpa.br

**Abstract.** *Post-COP30 environmental governance demands agile navigation through Brazil’s complex legislative framework. This paper introduces EcoLex RAG, a triple-retrieval HyPA-RAG system (BM25, Legal-BERTimbau, and Knowledge Graph) designed for explainable decision-making. Driven by an adaptive complexity classifier, the architecture achieves a 65% Law Retrieval Rate (LRR) and a 48% Citation Article Rate (CAR). Notably, it reduces context consumption by 32% (from 10.0 to 6.8 average sources) compared to static baselines without sacrificing accuracy. The system delivers a highly efficient, auditable, and impactful AI tool for public administration.*

**Keywords** *Environmental Law, RAG, Legal-BERTimbau, Explainable AI, HyPA-RAG.*

**Resumo.** *A governança ambiental pós-COP30 exige agilidade na consulta ao complexo arcabouço legislativo brasileiro. Este artigo apresenta o EcoLex RAG, um sistema HyPA-RAG triplo (BM25, Legal-BERTimbau e Knowledge Graph) focado no suporte explicável à decisão. Controlado por um classificador adaptativo de complexidade, o sistema alcança 65% de Law Retrieval Rate (LRR) e 48% de Citation Article Rate (CAR). A abordagem reduz o consumo de contexto em 32% (de 10,0 para 6,8 fontes em média), superando métodos estáticos. O sistema viabiliza uma inteligência artificial eficiente, estritamente auditável e de alto impacto estratégico para a administração pública.*

**Palavras-Chave** *Direito Ambiental, RAG, Legal-BERTimbau, IA Explicável, HyPA-RAG.*

## 1. Introdução

A eficácia da gestão ambiental no Brasil depende intrinsecamente da correta aplicação de um arcabouço normativo extenso e de alta complexidade hierárquica. Diplomas como o Código Florestal, a Política Nacional do Meio Ambiente e a Lei de Crimes Ambientais formam uma teia jurídica intrincada. A navegação manual por artigos, parágrafos, incisos e dispositivos remissivos impõe uma carga cognitiva severa aos gestores públicos e analistas, frequentemente resultando em morosidade na emissão de pareceres [Robertson e Zaragoza 2009].

Este desafio operacional adquire relevância estratégica e urgência inquestionáveis no contexto pós-COP30 (*Conferência das Nações Unidas sobre Mudanças Climáticas*), sediada em Belém do Pará. As metas de governança climática e bioeconomia assumidas pelo país exigem que os órgãos ambientais operem com agilidade sem precedentes no licenciamento, fiscalização e enquadramento legal de atividades econômicas. Nesse cenário, a ausência de ferramentas de ecoinformática capazes de traduzir a complexidade legal em suporte à decisão acionável e transparente representa um gargalo crítico.

No domínio da Inteligência Artificial, a técnica de *Geração Aumentada por Recuperação* (RAG) [Lewis et al. 2020] consolidou-se como o padrão-ouro para fundamentar Modelos de Linguagem de Grande Escala (*Large Language Models* – LLMs) em bases de conhecimento externas, mitigando alucinações. Contudo, a transposição direta de arquiteturas RAG convencionais para o cenário jurídico-ambiental brasileiro esbarra em três obstáculos fundamentais: (i) a incapacidade de *embeddings* genéricos capturarem o jargão técnico nacional (e.g., “Área de Preservação Permanente”); (ii) a não exploração da rica topologia estrutural das leis brasileiras; e (iii) a rigidez computacional de sistemas estáticos, que desperdiçam processamento e janela de contexto ao tratar consultas factuais simples com o mesmo peso de análises inter-regulatórias complexas.

Inspirado pelos recentes avanços na adaptação paramétrica [Kalra et al. 2024], o presente trabalho propõe o *EcoLex RAG*, uma solução computacional aplicada e estruturada para transformar a forma como o setor público interage com a legislação ambiental. As principais contribuições deste artigo para a comunidade do WCAMA incluem:

1. **Arquitetura de Recuperação Tripla Híbrida:** Integração otimizada para o direito ambiental brasileiro, unindo BM25 (*sparse*), representações vetoriais densas com Legal-BERTimbau [Rufimelo et al. 2022] e um *Knowledge Graph* que espelha a hierarquia das leis, unificados via *Reciprocal Rank Fusion* (RRF) [Cormack et al. 2009];
2. **Modulação Inteligente de Contexto:** Um classificador de complexidade baseado em heurísticas jurídico-linguísticas que adapta dinamicamente os parâmetros de busca (parâmetro de recuperação e reescritas de consulta). O objetivo desta estratégia é duplo: mitigar a recuperação excessiva de informações irrelevantes (*over-retrieval*) e, simultaneamente, reduzir o custo computacional associado ao processamento de contexto nos LLMs;
3. **Suporte à Decisão Explicável (XAI):** Um pipeline de geração com Mistral 7B Instruct [Jiang et al. 2023] configurado para produzir pareceres auditáveis, exigindo estruturação lógica e citação exata da fonte legal;

4. **Validação Empírica no Domínio Ambiental:** Um estudo de ablação detalhado com 100 consultas reais sobre as principais legislações ambientais do país, comprovando a eficácia e a economia de processamento da abordagem.

O restante deste artigo estrutura-se da seguinte forma: a Seção 2 revisa a fundamentação teórica; a Seção 3 detalha a metodologia do sistema *EcoLex RAG*; a Seção 4 apresenta o protocolo experimental; a Seção 5 discute o impacto prático dos resultados; e a Seção 6 apresenta as conclusões e os passos futuros rumo a uma adoção em escala governamental.

## 2. Trabalhos Relacionados

Esta seção posiciona o *EcoLex RAG* na interseção entre processamento de linguagem natural focado no domínio jurídico e o desenvolvimento de sistemas de *Ecoinformática* para apoio à governança.

### 2.1. Geração Aumentada por Recuperação (RAG) e Limitações

Desde a formalização do paradigma RAG por [Lewis et al. 2020], a arquitetura tornou-se essencial para reduzir a alucinação de LLMs em domínios críticos. A técnica tradicionalmente contrapõe métodos de correspondência léxica (*sparse*) como o BM25 [Robertson e Zaragoza 2009] — altamente eficazes para jargões específicos como “Resolução CONAMA 357” — a métodos de *embedding* (*dense*) que capturam intenção semântica. No entanto, aplicados isoladamente ou de forma estática ao direito normativo, ambos os métodos frequentemente falham em encadear o raciocínio necessário para resolver litígios ambientais que dependem de múltiplas leis interconectadas.

### 2.2. HyPA-RAG e a Otimização de Recursos

Para solucionar a rigidez do RAG tradicional, [Kalra et al. 2024] propuseram o HyPA-RAG (*Hybrid Parameter-Adaptive RAG*). O modelo resolve as limitações integrando grafos de conhecimento e adaptando parâmetros com base na complexidade da consulta. Ao modular a busca, o sistema previne a sub-recuperação em análises profundas e o desperdício computacional em perguntas diretas. A unificação dos diferentes mecanismos é feita através da Fusão de Classificação Recíproca (RRF) [Cormack et al. 2009], um método que dispensa calibração de pesos e opera diretamente sobre o ranqueamento dos trechos, tornando-se ideal para arquiteturas heterogêneas. A adoção desta otimização tem impacto direto na sustentabilidade da IA, alinhando-se aos princípios da computação verde ao reduzir ciclos de inferência desnecessários.

### 2.3. Linguagem para o Português Jurídico

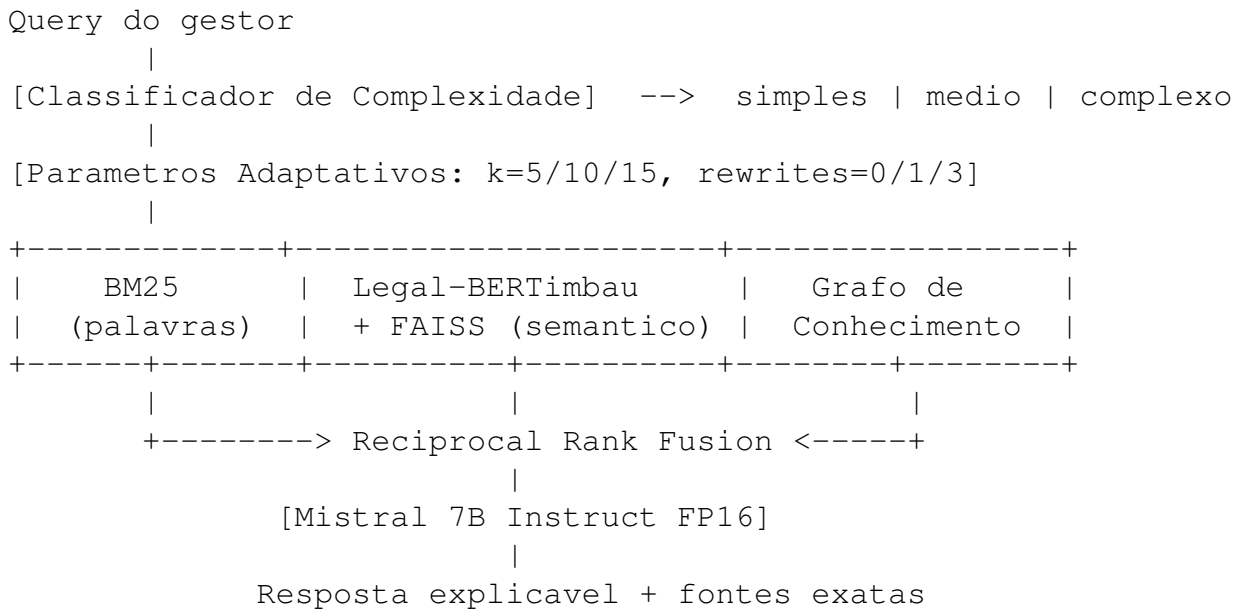
O desempenho de sistemas de busca semântica depende da qualidade da vetorização. O BERTimbau [Souza e Nogueira 2020] revolucionou o processamento do português, e sua posterior especialização contínua gerou o Legal-BERTimbau [Rufimelo et al. 2022]. O modelo `rufimelo/Legal-BERTimbau-sts-base-ma` captura com precisão semânticas estritamente jurídicas. Aliado a modelos generativos eficientes como o Mistral 7B [Jiang et al. 2023], que utiliza atenção de janela deslizante, torna-se viável processar contextos jurídicos densos utilizando hardware de menor custo, viabilizando implantações locais em secretarias de meio ambiente.

## 2.4. Ecoinformática e Inteligência Artificial Explicável (XAI)

A aplicação de técnicas de IA no domínio ambiental exige transparência e auditabilidade. Sistemas de suporte à decisão utilizados por órgãos públicos devem fornecer justificativas rastreáveis, alinhando-se aos princípios de governança algorítmica. O EcoLex RAG incorpora requisitos de XAI desde a etapa de recuperação até a geração final, obrigando o LLM a estruturar respostas com citações explícitas.

## 3. Metodologia: Arquitetura do EcoLex RAG

A arquitetura do EcoLex RAG foi desenhada para operar como uma ponte robusta entre o arcabouço normativo ambiental e o tomador de decisão. O fluxo de processamento HyPA-RAG é delineado na Figura 1.



**Figura 1. Arquitetura do EcoLex RAG baseada em HyPA-RAG.**

### 3.1. Classificador de Complexidade e Otimização de Recursos

Diferenciando-se das abordagens estáticas, o EcoLex RAG examina o peso informacional da consulta. Perguntas como “Qual a reserva legal no bioma Amazônia?” divergem substancialmente de cenários cruzados como “Compare as sanções do SNUC com as da Lei de Crimes Ambientais”. O classificador heurístico mitiga o desperdício energético operando um escore  $s(q)$ , definido na Equação 1:

$$s(q) = \min(\lambda_L, 3) + 2 \cdot \mathcal{K}_C(q) + \mathcal{K}_D(q) + 2 \cdot \mathcal{K}_M(q) + \mathcal{K}_?(q) + \mathcal{K}_\wedge(q) + \mathcal{K}_{|q|>30}. \quad (1)$$

Na Equação 1, cujo domínio resulta em uma pontuação  $s(q) \in \mathbb{N}$ , o termo  $\lambda_L \in \mathbb{N}$  representa a contagem de termos do léxico ambiental presentes na consulta, enquanto as funções indicadoras  $\mathcal{K}_x(q) : Q \rightarrow \{0, 1\}$  assumem valor 1 se o critério for atendido, e 0 caso contrário.

A taxonomia final segue a distribuição:

$$\text{complexidade}(q) = \begin{cases} \text{simples} & \text{se } s(q) \leq 2 \\ \text{média} & \text{se } 3 \leq s(q) \leq 5 \\ \text{complexa} & \text{se } s(q) \geq 6 \end{cases}$$

### 3.2. Recuperação Tripla: Maximizando o Espectro de Busca

A robustez informacional do sistema baseia-se em três pilares paralelos:

1. **BM25 Okapi (esparsa (léxica))**: Garante a preservação de referências normativas exatas (ex: “Art. 4º”) via indexação léxica [Robertson e Zaragoza 2009], imprescindível na redação governamental.
2. **Legal-BERTimbau + FAISS (densa (vetorial))**: Lida com o abismo semântico gerado pelas abstrações da linguagem natural de usuários não especialistas. *Embeddings* de 768 dimensões gerados pelo Legal-BERTimbau-sts-base-ma [Rufimelo et al. 2022] são normalizados via L2 e indexados via FAISS (Facebook AI Similarity Search) *IndexFlatIP* para máxima acurácia cossenoide.
3. **Knowledge Graph Legal (graph-based)**: Mapeia a genealogia normativa da lei brasileira através de regras de extração baseadas em expressões regulares (REGEX). O motor gera sub-redes formatadas em triplets do tipo (Art. X, detalha, §Y) ou (Art. Z, estabelece\_medida, 30\_metros).

### 3.3. Adaptação Paramétrica e Fusão (RRF)

As saídas disjuntas são congregadas via *Reciprocal Rank Fusion* [Cormack et al. 2009], conforme definido na Equação 2:

$$\text{RRF}_{\text{score}}(d) = \sum_{r \in \{\text{BM25}, \text{dense}, \text{KG}\}} \frac{1}{k + \text{rank}_r(d)}. \quad (2)$$

Onde a função de fusão  $\text{RRF} : D \rightarrow \mathbb{R}^+$  mapeia o conjunto de documentos recuperados  $D$  para um valor real positivo. O parâmetro constante  $k \in \mathbb{R}_+^*$  (usualmente  $k = 60$ ) garante que a expressão no denominador seja estritamente maior que zero, impossibilitando indefinições matemáticas matemáticas.

O diferencial do EcoLex RAG é a injeção adaptativa na fase de recuperação (Tabela 1). Em vez de afogar o sistema com uma busca larga constante, o *top\_k* e a quantidade de reescritas de ampliação de escopo geradas pelo LLM (*reescritas de consulta*) flutuam, economizando *tokens* preciosos na janela de contexto e garantindo eficiência para ambientes operacionais governamentais.

### 3.4. Geração XAI com Mistral 7B Instruct

O sistema delega a inferência final ao Mistral 7B Instruct v0.3 [Jiang et al. 2023] operando em precisão FP16. Para garantir a Inteligência Artificial Explicável (XAI), essencial na esfera pública, o *prompt* submetido impõe regras rígidas de formatação. O modelo deve categorizar sua resposta em **FUNDAMENTAÇÃO LEGAL**, **ANÁLISE** e **CONCLUSÃO**, citando minuciosamente artigos, parágrafos ou incisos, tornando o aconselhamento tecnicamente inviolável e rastreável.

**Tabela 1. Parâmetros adaptativos conforme complexidade da consulta**

Complexidade	top_k	reescritas de consulta	Exemplo de consulta
Simple	5	0	“O que é APP?”
Média	10	1	“Qual a faixa de APP para rio de 10 m?”
Complexa	15	3	“Compare as exigências de APP e Reserva Legal para propriedade na Amazônia Legal.”

#### 4. Configuração Experimental

Para avaliar o grau de proficiência do EcoLex RAG diante de situações práticas vivenciadas por analistas ambientais, estruturou-se uma avaliação orientada à resolução de *cases* normativos.

A base de conhecimento cobre o núcleo duro do licenciamento nacional: Código Florestal (Lei n. 12.651), Política Nacional do Meio Ambiente (Lei n. 6.938), SNUC (Lei n. 9.985), Lei de Crimes Ambientais (Lei n. 9.605) e resoluções do CONAMA (357 e 430). Este material resultou em 473 fragmentos textuais e 1.714 relações no *Knowledge Graph*. O conjunto de testes compreendeu 100 consultas desafiadoras classificadas e distribuídas nas referidas leis (Tabela 2). Para fomentar a reprodutibilidade e a criação de *benchmarks* em ecoinformática brasileira, o dataset completo de 100 consultas e as respostas de referência estão disponíveis publicamente<sup>1</sup>.

**Tabela 2. Distribuição empírica das consultas de avaliação por complexidade legislativa**

Legislação	Simple	Média	Complexa	Total
Código Florestal	6	12	12	30
PNMA	6	5	5	16
SNUC	6	5	5	16
Crimes Ambientais	10	3	3	16
CONAMA 357/2005	6	6	0	12
CONAMA 430/2011	6	4	0	10
<b>Total</b>	<b>40</b>	<b>35</b>	<b>25</b>	<b>100</b>

Foi desenhado um estudo de ablação com 6 configurações: (1) BM25-only; (2) Semantic-only; (3) KG-only; (4) BM25+Semantic; (5) híbrido fixo (estático,  $k = 10$ ); e (6) o modelo pleno HyPA-RAG adaptativo. A avaliação das métricas foi conduzida de forma automatizada por meio de scripts de correspondência de strings e padrões de citação. O sistema comparou as respostas geradas pelo LLM contra um conjunto de referência (padrão-ouro), cujas respostas ideais foram elaboradas e validadas previamente por especialistas humanos no domínio do Direito Ambiental, garantindo o rigor técnico dos critérios de acerto.

<sup>1</sup>Repositório disponível em <https://github.com/vsousadelvek/ecollex-rag>

## 5. Resultados e Impacto para a Gestão Ambiental

Os resultados do estudo ablativo revelam como a adoção do EcoLex RAG repercute na melhoria da performance de processamento institucional (Tabela 3).

**Tabela 3. Resultados consolidados: impacto da ablação sobre eficiência e precisão legal. A métrica “Fontes” refere-se à média de *chunks* recuperados após a fusão RRF.**

Configuração	Lei Ret.	Art. Ret.	Cit. Lei	Cit. Art.	Fontes	Tempo (s)
BM25-only	65%	70%	63%	45%	10,0	26,2
Semantic-only	61%	53%	61%	40%	10,0	24,6
KG-only	56%	35%	58%	31%	10,0	20,3
BM25+Semantic	65%	66%	65%	48%	10,0	27,6
Híbrido fixo	65%	59%	66%	49%	10,0	23,8
HyPA-RAG	65%	56%	65%	48%	6,8	43,5

### 5.1. A Sustentabilidade do Processamento (Economia de Contexto)

O aspecto mais disruptivo do modelo HyPA-RAG reside na sua redução drástica do custo de contexto sem deteriorar a base referencial. O sistema empatou estatisticamente no índice primordial de Citação de Artigos Exatos (48%) com os modelos de força bruta pesados, mas o fez utilizando uma média enxuta de apenas 6,8 fontes ao invés das 10 requisições estáticas globais. Esta redução de 32% na carga de *input* diminui a diluição semântica no LLM, economiza poder de processamento (VRAM) e alinha o *pipeline* a estratégias de computação de baixo impacto, adequadas aos *data centers* estaduais e municipais. Embora o processo de inferência com *query rewrites* estenda o tempo médio de resposta para 43,5 segundos, este *trade-off* temporal é perfeitamente aceitável e irrisório num cenário governamental assíncrono, como a elaboração de pareceres técnicos que tradicionalmente levam dias para serem concluídos.

### 5.2. Superioridade Direcionada nas Consultas Ambientais Diárias

No dia a dia do licenciamento, as consultas de complexidade média (condicionantes, remissões de parágrafos) formam a espinha dorsal da administração. Conforme atesta a Tabela 4, é exatamente neste segmento que a calibração do HyPA-RAG prospera, atingindo um pico de 54% em Citação Correta de Artigos (CAR). Dada a dimensão da amostra (35 consultas médias), este resultado demonstra um desempenho empírico comparável e uma vantagem competitiva com drástica redução de custo de contexto face à abordagem de Força Bruta Estática (49%).

Nas interações complexas multibioma ou inter-leis (*multi-hop*), representadas na Tabela 5, a fluidez adaptativa do EcoLex RAG, munido de suas 3 reescritas semânticas automáticas, consolidou sólidos 52% de assertividade, mantendo um desempenho empírico comparável ao modelo fixo, mas otimizando substancialmente o uso da janela de contexto.

### 5.3. O Custo da Automação e Oportunidades de Otimização

O mecanismo apresenta contrapartidas compreensíveis. Nas consultas mais diretas e factuais (Tabela 6), o corte intencional do *top\_k* para apenas 5 registros ocasionou uma

**Tabela 4. Acurácia para consultas de complexidade média**

Configuração	Lei Ret.	Art. Ret.	Cit. Lei	Cit. Art.
BM25-only	66%	63%	63%	43%
Semantic-only	60%	43%	60%	34%
KG-only	57%	31%	54%	23%
BM25+Semantic	66%	63%	66%	46%
Híbrido fixo	66%	63%	66%	49%
HyPA-RAG	66%	63%	63%	54%

**Tabela 5. Acurácia do Raciocínio Normativo Cruzado (Complexidade Alta)**

Configuração	Lei Ret.	Art. Ret.	Cit. Lei	Cit. Art.
BM25-only	68%	60%	68%	40%
Semantic-only	60%	48%	64%	44%
KG-only	60%	20%	64%	36%
BM25+Semantic	64%	52%	68%	52%
Híbrido fixo	64%	36%	72%	48%
HyPA-RAG	64%	52%	68%	52%

perda de *recall* em relação à busca de texto bruto (BM25), diminuindo o índice CAR (40% vs. 50%). Reconhecemos criticamente que este limite de  $k = 5$  foi excessivamente agressivo. Para implementações reais em produção, sugere-se formalmente que o limiar mínimo seja recalibrado (e.g.,  $k = 7$  ou  $k = 8$ ) de modo a equilibrar adequadamente a economia computacional e a taxa de *recall*.

**Tabela 6. Acurácia para consultas de complexidade simples**

Configuração	Lei Ret.	Art. Ret.	Cit. Lei	Cit. Art.
BM25-only	62%	82%	60%	50%
Semantic-only	62%	65%	60%	42%
KG-only	52%	48%	57%	35%
BM25+Semantic	65%	78%	62%	48%
Híbrido fixo	65%	70%	62%	50%
HyPA-RAG	65%	52%	65%	40%

## 6. Conclusão

O cenário de transição climática impõe que as estruturas estatais reajam com técnica e agilidade diante das demandas de conservação e desenvolvimento. Este trabalho concebeu e validou o EcoLex RAG, demonstrando como a modulação paramétrica híbrida aliada ao processamento de linguagem natural focado no jargão brasileiro transcende o mero laboratório computacional para se tornar uma aliada tangível na formulação de respostas governamentais.

Comprovou-se empiricamente que o modelo assegura um desempenho empírico equivalente em acurácia argumentativa para os casos operacionais medianos, entregando citações diretas confiáveis com uma drástica economia de recursos contextuais (32%

menos *input*). A obrigatoriedade estrutural na resposta XAI mitiga a alucinação e pavimentam o caminho da transparência administrativa, preceito inalienável no Brasil pós-COP30.

Em consonância com as ambições de fomento em Ecoinformática visadas pelo WCAMA 2026, trabalhos futuros abordarão a afinação fina (*fine-tuning*) direta dos LLMs utilizando os pareceres históricos do IBAMA, a substituição da heurística de complexidade por aprendizado supervisionado de máquina, e a integração massiva de resoluções de conselhos estaduais no Grafo de Conhecimento, maximizando o escrutínio sobre particularidades regionais. Cabe ressaltar como limitação que o sistema de reescritas de consulta (*query rewrites*) e a respectiva economia de contexto foram validados num único modelo instrucional (Mistral 7B FP16). A validação cruzada com outros *Large Language Models* (LLMs) recentes (como Llama 3 ou GPT-4) será mandatória em trabalhos futuros para garantir que os ganhos de eficiência do HyPA-RAG não estão enviesados pelas particularidades de um único modelo.

## Referências

- Cormack, G. V., Clarke, C. L. A., e Buettcher, S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09)*, pages 758–759. ACM.
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., e Sayed, W. E. (2023). Mistral 7B. *arXiv preprint arXiv:2310.06825*.
- Kalra, R., Beshaj, L., Massey, T., e Grossman, R. (2024). HyPA-RAG: A hybrid parameter adaptive retrieval-augmented generation system for AI legal and policy applications. *arXiv preprint arXiv:2409.09046*.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., e Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, volume 33, pages 9459–9474.
- Robertson, S. e Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.
- Rufimelo, M., Abrantes, A., Santos, A. F., e Ribeiro, B. (2022). Legal-BERTimbau: Pretrained BERT models for semantic textual similarity in the Portuguese legal domain. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC 2022)*, pages 1334–1340. European Language Resources Association.
- Souza, F. e Nogueira, R. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. In *Proceedings of the 9th Brazilian Conference on Intelligent Systems (BRACIS 2020)*, volume 12319 of *Lecture Notes in Computer Science*, pages 403–417. Springer.