**Resumo Expandido — Retrieval‑Augmented Generation aplicado ao Direito**

Arthur Bueno Vitola

Evandro Felipe

**Resumo:** Este resumo expandido explora de forma acessível e prática os conceitos, aplicações, vantagens e riscos associados à técnica Retrieval‑Augmented Generation (RAG) no contexto jurídico. RAG combina modelos de linguagem pré‑treinados com mecanismos de recuperação de documentos, permitindo que respostas geradas sejam explicitamente fundamentadas em textos legais, jurisprudência e doutrina. Além de descrever a arquitetura técnica, este texto apresenta recomendações operacionais para implementação responsável em escritórios de advocacia, departamentos jurídicos e plataformas de acesso à justiça, abordando também governança de dados e responsabilidade profissional.

**Abstract:** This expanded abstract presents key concepts, practical applications, benefits and limitations of Retrieval‑Augmented Generation (RAG) when applied to the legal domain. RAG architectures integrate pretrained language models with document retrieval modules so that generated outputs are grounded in retrieved legal texts, case law, and doctrinal sources. We discuss technical foundations, typical pipelines, governance and privacy concerns (e.g., LGPD/GDPR), and propose best practices for corpus curation, provenance tracking, and human-in-the-loop review. Concrete recommendations are provided for legal practitioners who wish to deploy RAG solutions with traceability and compliance.

**1. Introdução**

A adoção crescente de modelos de linguagem de grande porte (LLMs) em ambientes profissionais tem impulsionado a busca por arquiteturas que conciliem capacidade generativa e acesso confiável a conhecimento externo. No Direito, a necessidade de precisão, citação de fontes e atualização constante torna a integração de bases documentais essenciais. Retrieval‑Augmented Generation (RAG) apresenta‑se como uma solução pragmática: o gerador produz texto condicionado por trechos recuperados de um corpus legal cuidadosamente indexado, reduzindo o risco de afirmações infundadas e facilitando a comprovação das assertivas por meio das fontes originais.

**2. Fundamentos técnicos**

Tecnicamente, um sistema RAG típico envolve múltiplos estágios coordenados: (a) pré‑processamento e segmentação dos documentos (chunking) para preservar unidades argumentativas e referências; (b) geração de embeddings das passagens e construção de um índice vetorial; (c) codificação da consulta do utilizador; (d) recuperação das passagens mais relevantes via nearest neighbours; (e) reranking e filtragem para priorizar precisão e reduzir ruído; e (f) execução do gerador condicional sobre as passagens selecionadas. Algumas variantes adotam estratégias híbridas — por exemplo, usar o retriever apenas para sugerir evidências que são depois verificadas por um modelo de fact‑checking ou por regras heurísticas. Técnicas modernas de embeddings, como modelos ajustados ao domínio jurídico (legal‑embeddings), tendem a melhorar a qualidade da recuperação.

**3. Aplicações no Direito**

As aplicações práticas abrangem desde pesquisa jurisprudencial até suporte à redação. Em pesquisa, RAG permite localizar e reunir precedentes análogos, consolidando trechos relevantes e indicando a linha argumentativa predominante. Para redação de peças, o sistema pode propor esboços estruturados, cláusulas padrão e comparativos de jurisprudência, sempre com as fontes anexadas. No contexto de acesso à justiça, chatbots que empregam RAG podem orientar cidadãos sobre direitos, prazos processuais e documentos necessários, encaminhando casos mais complexos para profissionais humanos. Em compliance, RAG acelera auditorias contratuais ao destacar cláusulas com risco potencial e relacioná‑las a normas aplicáveis.

**4. Vantagens**

Além de reduzir a frequência de alucinações quando bem implementado, RAG facilita atualizações do conhecimento sem re‑treinamento do LLM, simplifica a incorporação de novos acórdãos e legislação, e melhora a explicabilidade ao fornecer trechos de apoio. Para escritórios, isso traduz‑se em ganhos de produtividade e maior capacidade de escalonar pesquisas complexas. Em aplicações públicas, torna possível oferecer serviços básicos de orientação jurídica com controles de transparência e logs auditáveis.

**5. Riscos e desafios**

Os principais riscos decorrem da qualidade do corpus e da governança de dados. Fontes incompletas, desatualizadas ou enviesadas podem induzir conclusões equivocadas; portanto, curadoria e versionamento do índice são obrigatórios. A presença de dados pessoais em documentos exige avaliação sob a LGPD (no Brasil) e GDPR (na UE), podendo demandar anonimização, minimização e políticas rígidas de acesso. Do ponto de vista ético e profissional, é fundamental que o advogado mantenha responsabilidade sobre a utilização das saídas geradas — RAG deve funcionar como ferramenta auxiliar, com revisão e validação humana sistemática. Há também riscos técnicos: custos de infraestrutura (armazenamento e consulta em índices grandes), latência em consultas complexas e necessidade de monitoramento contínuo da performance.

**6. Implementação prática — recomendações detalhadas**

Para implementar um sistema RAG com segurança jurídica recomenda‑se um roadmap prático: (i) levantar fontes prioritárias e estabelecer critérios de inclusão; (ii) normalizar metadados (tribunal, data, número do processo, ementa); (iii) aplicar chunking com limites que preservem citações; (iv) treinar ou selecionar embeddings adaptados ao domínio jurídico; (v) utilizar um índice robusto (FAISS, Milvus ou soluções gerenciadas) com estratégia de reranking; (vi) implantar mecanismos de provenance que exibam, lado a lado, a passagem fonte e o trecho gerado; (vii) integrar logs de auditoria e controles de acesso; e (viii) definir políticas de revisão humana e de escalonamento para casos de alto risco. Adicionalmente, recomenda‑se testar o sistema em projetos‑piloto e coletar métricas operacionais (tempo economizado, taxa de correção humana, recall/precision) antes de uma adoção em larga escala.

**7. Exemplo prático e estudos de caso**

Um caso de implementação pode envolver um escritório que indexa suas bases internas e fontes públicas para a área trabalhista. Após validação inicial, o sistema passa a fornecer resumos de precedentes, cláusulas sugeridas para peças e listagens de jurisprudência correlata, com links diretos aos documentos completos. Estudos comparativos mostram redução significativa no tempo de pesquisa e maior consistência nas citações quando workflows de revisão são respeitados. Importante: resultados variam conforme a curadoria e o ajuste do retriever.

**8. Limitações e direções futuras**

Limitações atuais incluem dependência de índices atualizados, dificuldade em interpretar nuances fáticas complexas e a necessidade de interoperabilidade entre sistemas jurídicos distintos. Pesquisas futuras apontam para: melhores embeddings jurídicos multilingues, mecanismos de reranking que incorporem sinais jurídicos (p.ex. autoridade do tribunal), e ferramentas automáticas de detecção de mudanças legislativas que disparem atualizações no índice. A integração com sistemas de gestão de casos pode também ampliar o valor prático de RAG em escritórios.

**9. Conclusão**

RAG representa uma alternativa viável para combinar a flexibilidade dos LLMs com a robustez de fontes documentais na prática jurídica. Quando adotado com governança clara, mecanismos de provenance e revisão humana, contribui tanto para eficiência quanto para maior transparência nas argumentações. A tecnologia não elimina a necessidade do trabalho jurídico qualificado, mas amplia o repertório de ferramentas à disposição do profissional.

**Referências**

LEWIS, P.; PIKTUS, A.; PETRONI, F.; et al. Retrieval‑Augmented Generation for Knowledge‑Intensive NLP Tasks. NeurIPS, 2020. (arXiv).  
  
CHALKIDIS, I.; FURFARO, F.; MALIK, M.; et al. LEGAL‑BERT: The Muppets straight out of Law School. Findings of EMNLP, 2020.  
  
PIPITONE, N.; HOUIR ALAMI, G. LegalBench‑RAG: A Benchmark for Retrieval‑Augmented Generation in the Legal Domain. arXiv, 2024.  
  
THOMSON REUTERS. Intro to retrieval‑augmented generation (RAG) in legal tech. Thomson Reuters Legal Blog, 2024.