

IA generativa com arquitetura de RAG na prática do Direito

Autor(a): Vitor Soliano

Publicado em: 10 de agosto de 2025

Publicação: conjur_import

Link JurisTube:

<https://juristube.com.br/colunistas/vitor-soliano/ia-generativa-com-arquitetura-de-rag-na-pratica-do-direito>

Fonte original: <https://conjur.com.br/2025-ago-10/ia-generativa-com-arquitetura-de-rag-na-pratica-do-direito>

Identificador citável: juristube.com.br/colunistas/vitor-soliano/ia-generativa-com-arquitetura-de-rag-na-pratica-do-direito#ep-94a444dd

Resumo

Rumamos para a adoção virtualmente plena de ferramentas de IA, na esteira do que ocorreu previamente com o computador pessoal e a internet

Texto integral

No primeiro semestre de 2025, 82% dos brasileiros já utilizavam inteligência artificial de forma regular ou semi-regular, segundo uma ampla pesquisa da Universidade de Melbourne em parceria com a KPMG [1] . Isso significa que quatro em cada cinco brasileiros já adotaram a IA? Sim. Para confirmar, basta olhar para o colega ao lado. Os inovadores, os “early adopters” (ou primeiros adeptos) e toda a maioria inicial de usuários absorveram e incorporaram a tecnologia, restando parte da maioria tardia e os retardatários [2] . Queira-se ou não, goste-se ou não, rumamos em direção à adoção virtualmente plena de ferramentas de inteligência artificial, na esteira do que ocorreu previamente com o computador pessoal e a internet . Os profissionais das ciências sociais encontraram o equivalente à “calculadora” dos operadores da matemática. E, salvo poucas exceções, já perceberam que uma “calculadora” para a escrita, quando utilizada adequadamente, empodera – e não diminui – a importância de sua ciência. É útil? A taxa de adoção da tecnologia fala por si só. A propósito, comentei em artigo anterior sobre estudos empíricos que atestam o impacto positivo da IA generativa nos produtos de atividades intelectuais, indicando estatísticas sobre a conclusão de tarefas com maior velocidade, com maior qualidade e em maior volume [3] . Para o que nos interessa aqui, nesta coluna, nota-se que a IA generativa transbordou definitivamente do nicho estritamente tecnológico e deu origem a ferramentas profissionais especializadas para o cotidiano jurídico. É a expressão de um fenômeno derivado: a rápida evolução técnica das próprias ferramentas de IA generativa, que ascenderam em capacidades e inauguraram novas funcionalidades e técnicas de uso. Essa evolução ocorreu mediante a incorporação e integração de arquiteturas tecnológicas preexistentes aos sistemas de IA generativa, muitas delas voltadas a mitigar ou, ao menos, a permitir maior controle sobre algumas de suas falhas, a exemplo da alucinação, do viés algorítmico e da caixa-preta (black-box). Em especial, destaco uma arquitetura atualmente já integrada a várias plataformas e que representa o estado da arte dessa tecnologia aplicada ao Direito: a “Retrieval Augmented Generation” (RAG) – em tradução livre, a geração aumentada por recuperação. A arquitetura de RAG: IA sob medida baseada em fontes confiáveis O ano de 2024 foi marcado pelo surgimento – e, ao final, pela proliferação – de aplicações de IA

generativa que adotam a arquitetura de RAG. Em essência, a arquitetura de RAG integra uma base de dados externa ao sistema de IA, permitindo que ele “pesquise” informações relevantes antes de gerar a resposta ao usuário. Assim, em vez de confiar apenas no treinamento que recebeu, o modelo (LLM) recupera trechos de textos autorizados pelo sistema – por exemplo, acórdãos, leis, normas infralegais, livros doutrinários – e fundamenta sua resposta nesses conteúdos. Para melhor compreensão, imagine que o sistema de IA seja um jurista contratado. A arquitetura de RAG autoriza que este profissional, em vez de responder apenas com o conhecimento que está em sua memória, consulte ativamente suas anotações, cadernos, jurisprudência ou livros de sua biblioteca antes de responder às perguntas do usuário. Sem RAG, este profissional responderia à questão “de cabeça”, apenas com a sua experiência, memória e intuição formada ao longo de seu treinamento. Sem RAG, seria o equivalente a uma pessoa fazer uma prova técnica sem consulta. Com RAG, é uma prova com consulta. A diferença nos resultados é gritante, por óbvio. O funcionamento é conceitualmente simples, embora sua construção tecnológica seja complexa: o prompt do usuário do sistema é enriquecido com um contexto extraído de uma base de dados em momento anterior ao processamento pela IA generativa. Aplicando-se outra metáfora, é como se, antes da dúvida jurídica (prompt) chegar ao advogado responsável por formular a resposta (LLM), outro profissional (RAG) fizesse uma pesquisa sobre as prováveis fontes jurídicas a serem utilizadas para essa resposta (base de dados), entregando ao advogado (LLM) a dúvida jurídica (prompt) e o conteúdo bruto das fontes de Direito que provavelmente levarão à resposta (contexto que enriquece o prompt , obtido via RAG). Daí porque o índice de alucinações (invenção de conteúdos falsos) da IA generativa com uma boa arquitetura de RAG é ínfimo. Como o sistema obrigatoriamente consulta um banco de dados com conteúdo sobre o tema tratado na pergunta, não há necessidade de inventar conteúdos falsos para responder à questão. Em vez de inventar um artigo de lei ou citar um precedente inexistente, a IA se apoia em textos verificados. Estudos empíricos recentes [4] mostram que uma ferramenta jurídica que se utiliza de IA generativa com RAG é capaz de reduzir as alucinações encontradas nos trabalhos jurídicos a níveis comparáveis àqueles encontrados em trabalhos jurídicos finalizados sem qualquer assistência de inteligência artificial. Com uma boa arquitetura de RAG, portanto, as ferramentas jurídicas de IA generativa podem acoplar bancos de dados com normas, precedentes e doutrina para responder às questões dos usuários. Mitigam-se os riscos de alucinações, de viés algorítmico e de blackbox , pois os outputs da ferramenta passam a referenciar conteúdo de um banco de dados específico. Além disso, o uso de RAG viabiliza – por ser mais simples, mais rápido e mais barato – que seja mantida a atualidade das respostas da IA. Modelos de linguagem tradicionais tinham conhecimento congelado na data de seu treinamento; uma IA com arquitetura de RAG, por outro lado, tem acesso a uma base de dados que pode ser constantemente alimentada e atualizada. Assim, a IA pode responder com dados atualizados em tempo real – seja uma nova lei publicada há poucos dias ou o último entendimento do TCU sobre certo tema. Essa sinergia entre IA e bases de dados resolve a tensão entre confiabilidade e atualidade, fator crítico para a adoção segura de IA no Direito. Costumo brincar que a “simplicidade da receita” de uma ferramenta de IA com arquitetura de RAG esconde a complexidade da operação. É como produzir vinho: as condições do solo, o corte das uvas, o clima, a época da colheita, o método de engarrafamento, a barrica escolhida, a adega... cada etapa de uma receita teoricamente “simples” interfere no resultado. No caso das aplicações de IA com arquitetura de RAG, da mesma maneira: a estruturação dos dados, a curadoria dos documentos (qualidade e quantidade de fontes), a indexação, a recuperação, a integração, a categorização, as redundâncias e as instruções compõem o “segredo do fabricante” e interferem decisivamente no resultado. Essa arquitetura permite tanto pesquisas em bases abertas, como a

internet, ou fechadas, como um banco de dados proprietário, sob curadoria de especialistas. Dado o caráter generalista das principais aplicações de IA (ChatGPT, Gemini, Perplexity), um movimento natural desses grandes atores foi simplesmente conectar os seus sistemas à internet (base de dados aberta) para buscar informações. Em teoria, isso expande o conhecimento da IA em tempo real. Na prática, porém, os LLMs, ao “lerem” resultados de buscadores na internet, podem facilmente se deparar com informações não verificadas ou apresentar conteúdo fora de contexto. Um estudo recente trouxe um alerta ao concluir que ferramentas de busca na internet em sistemas de IA generativa frequentemente fabricam links e citam versões alteradas de artigos e notícias [5]. No contexto jurídico, a arquitetura de RAG com base de dados fechada permite que a IA seja alimentada com um repertório personalizado e controlado: decisões de tribunais locais, documentos de referência ou peças processuais bem-sucedidas, doutrina especializada, normas infralegais, casos semelhantes e até manuais e pareceres internos. O sistema passa então a responder embasado apenas nessas fontes específicas, reduzindo drasticamente o risco de alucinações e aumentando a assertividade. O resultado são respostas alinhadas com a realidade jurídica e personalizadas ao contexto do usuário. É uma espécie de assessor jurídico disponível em tempo real, experiente e incansável, conhecedor das normas e procedimentos. É a base dos assistentes jurídicos virtuais. Foi, por exemplo, o que construímos na startup Schief.ai e na ferramenta licito.guru IA, especializada em licitações públicas: usuários podem criar minutas complexas de documentos usados na instrução de uma licitação pública e receber respostas personalizadas de acordo com o seu próprio contexto jurídico, pois a IA se alimenta de um banco de dados com jurisprudência selecionada, normas locais e documentos de referência próprios. A arquitetura de RAG é a memória expandida da IA generativa. É o texto escrito que auxilia o profissional. Isso nos remete a um diálogo socrático. O mito renovado de Thoth e Tâmus: IA versus memória e escrita Em junho de 2025, um estudo divulgado por pesquisadores do MIT [6] sobre possíveis custos cognitivos em usuários de IA generativa gerou manchetes sensacionalistas, a despeito do esforço dos autores em indicar que o estudo ainda carece de revisão por pares, que foi feito com uma amostra limitada e que não conclui sobre dano cerebral, atrofia ou emburrecimento. A hipótese é intuitiva: o esforço cerebral para a produção de um texto técnico diminui com o apoio da IA. Surge então a reflexão: estaríamos revivendo, em pleno século 21, o antigo mito de Thoth e Tâmus? No diálogo socrático relatado por Platão, em Fedro, conta-se que o deus egípcio Thoth presenteou a humanidade com a escrita, prometendo tornar as pessoas mais sábias e com memórias mais eficientes. O rei Tâmus, porém, advertiu-o como ingênuo, pois que, em vez de aprimorar a memória, a escrita iria enfraquecê-la. Com a escrita, os homens passariam a depender de registros externos em vez de exercitar a própria lembrança. A escrita geraria uma falsa sabedoria, externa ao indivíduo, apenas aparente. Thoth, como inventor daquela tecnologia, estaria inebriado por um amor parental, como um pai que exagera ou enxerga qualidades inexistentes em seus filhos. Substitua-se a invenção da escrita pela IA generativa. O paralelo salta aos olhos. E a similitude fica ainda mais nítida quando considerado que o estado da arte da IA generativa opera apoiada em RAG. Ou seja, o “registro externo” que preocupava Tâmus não é mais um pergaminho repousando na biblioteca, mas um vetor semanticamente indexado que a IA consulta em microssegundos. Quanto mais confiamos nesse registro, menor é a tentação (e a necessidade) de manter tudo na memória biológica. Quanto melhores as respostas geradas pela IA generativa, potencializadas agora com a arquitetura de RAG, menos precisamos tê-las previamente em nossa memória. Tal como no mito, há quem celebre a IA generativa como uma libertação: com a ajuda dessa tecnologia, será possível evoluir o conhecimento e produzir conteúdo de forma tão dinâmica e eficiente que a mente humana será

liberada para tarefas mais elevadas, estratégicas e criativas. Afinal, por que saber detalhes de normas, jurisprudência ou doutrina, quando a IA pode fazê-lo em segundos? Seria a promessa de Thoth reimaginada: uma ferramenta que amplifica nossa capacidade cognitiva. Por outro lado, ecoando o ceticismo de Tâmus, alguns alertam para o risco de “atrofia intelectual”. Se delegamos tarefas intelectuais essenciais – analisar casos, argumentar, escrever petições –, corremos o perigo de perder nossas habilidades. As habilidades jurídicas poderiam se enfraquecer se nos acostumarmos a sempre perguntar à máquina. Da mesma forma, a arte da boa escrita, construída na prática e refinada pelo estudo, poderia se degradar se o jurista apenas editar textos gerados pela IA, em vez de redigi-los do zero. Teme-se, em suma, que prospere uma sabedoria de fachada, em que profissionais apresentam trabalhos excelentes, gerados por IA, mas não dominam a matéria por trás deles. É realmente um receio análogo aos “sábios ignorantes” do mito de Tâmus e Thoth, apoiados em “informações exteriores”, sem conhecimento real. E aí, caro leitor, você se inclina para a opinião do rei Tâmus ou do deus Thoth? A história da escrita nos ensinou algo valioso: apesar dos receios do rei Tâmus, a escrita não destruiu a memória humana. Ao contrário, transformou-a e ampliou o alcance do conhecimento. Agora, novamente, as habilidades intelectuais humanas estão sendo redefinidas. É um cenário em que algumas competências ganham maior relevância: saber formular as perguntas certas, declarar premissas, testar hipóteses, auditar e interpretar criticamente as respostas da IA, combinar intuição jurídica com análise de dados fornecidos pela máquina. O objetivo central de quem vive esta nova era do Direito, certamente, não deve ser saber “apertar os botões corretos”. Pelo contrário, o profissional contemporâneo deve se empoderar da ferramenta para extrair delas a capacidade de desenvolver maiores competências analíticas, criativas e estratégicas. Como num mito renovado de Thoth e Tâmus, cabe a nós, operadores do Direito, o discernimento para equilibrar o uso da IA de forma a potencializar a atuação humana e a performance do sistema da justiça. Com senso crítico, a inteligência artificial poderá repetir a jornada da escrita: de ameaça a aliada indispensável. E assim evolui o Direito, sem perder a essência humana que lhe justifica. [1] GILLESPIE, N.; LOCKEY, S.; WARD, T.; MACDADE, A.; HASSED, G. Trust, attitudes and use of artificial intelligence: A global study 2025. The University of Melbourne and KPMG, 2025, p.21. Disponível em: <https://conjur.com.br/wp-content/uploads/2025/08/trust-attitudes-and-use-of-ai-global-report.pdf> [2] São essas as categorias de usuários da teoria da difusão da inovação de Everett Rogers. Cf. ROGERS, Everett M. Diffusion of Innovations. 3rd ed. New York: The Free Press, 1983. [3] Cf. DELL'ACQUA, Fabrizio et al. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. Harvard Business School Working Paper, No. 24-013, September 2023. [4] SCHWARCZ, Daniel et al. AI-Powered Lawyering: AI Reasoning Models, Retrieval Augmented Generation, and the Future of Legal Practice. Minnesota Legal Studies Research Paper No. 25-16, U of Michigan Public Law Research Paper No. 24-058, 2025. Disponível em: <https://ssrn.com/abstract=5162111> or <http://dx.doi.org/10.2139/ssrn.5162111> [5] SANTOS, Gilmar. IA falha ao citar conteúdo jornalístico com referências erradas, 2025. Disponível em: <https://intelligence.garden/ia-falha-ao-citar-conteudo-jornalistico-com-referencias-erradas/> Acesso em 25 jul. 2025 [6] KOSMYNA Nataliya et. al. Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task, 2025. Disponível em: <https://arxiv.org/abs/2506.08872> Acesso em 25 jul. 2025. <https://www.media.mit.edu/projects/your-brain-on-chatgpt/overview>

Como citar este documento

As citações abaixo seguem os padrões ABNT NBR 6023:2018, APA 7ª edição, Chicago Author-Date (17ª ed.) e BibTeX (LaTeX). Copie o formato exigido pelo periódico ou gerenciador bibliográfico (Zotero, Mendeley, EndNote).

ABNT (NBR 6023:2018)

SOLIANO, Vitor. IA generativa com arquitetura de RAG na prática do Direito. *conjur_import*, 10 ago. 2025. Disponível em: <https://conjur.com.br/2025-ago-10/ia-generativa-com-arquitetura-de-rag-na-pratica-do-direito>. Acesso via: JurisTube — Acervo Digital de Direito. Disponível em: <https://juristube.com.br/colunistas/vitor-soliano/ia-generativa-com-arquitetura-de-rag-na-pratica-do-direito>. Acesso em: 30 set. 2026.

APA 7ª edição

Soliano, V. (2025, August 10). IA generativa com arquitetura de RAG na prática do Direito. **conjur_import**. <https://conjur.com.br/2025-ago-10/ia-generativa-com-arquitetura-de-rag-na-pratica-do-direito>

Chicago Author-Date (17ª ed.)

SOLIANO, Vitor. 2025. "IA generativa com arquitetura de RAG na prática do Direito." **conjur_import**, August 10. <https://conjur.com.br/2025-ago-10/ia-generativa-com-arquitetura-de-rag-na-pratica-do-direito>

BibTeX

```
@article{vitor-soliano-ia-generativa-com-arquitetura-de-rag-na--2025,  
  author = {Soliano, Vitor},  
  title = {IA generativa com arquitetura de RAG na prática do Direito},  
  journal = {conjur_import},  
  year = {2025},  
  url = {https://conjur.com.br/2025-ago-10/ia-generativa-com-arquitetura-de-rag-na-pratica-do-direito},  
  urldate = {2025-08-10}  
}
```



Licença de Uso — CC BY-NC-SA 4.0

Este conteúdo está licenciado sob a licença Creative Commons Atribuição-NãoComercial-CompartilhaIgual 4.0 Internacional (CC BY-NC-SA 4.0). O uso para fins acadêmicos e educacionais é permitido, desde que citada a fonte original (JurisTube) conforme as normas técnicas indicadas neste documento.

Termos completos:

<https://creativecommons.org/licenses/by-nc-sa/4.0/deed.pt-br>

JurisTube — Acervo Digital de Direito

<https://juristube.com.br/colunistas/vitor-soliano/ia-generativa-com-arquitetura-de-rag-na-pratica-do-direito>

PDF gerado em 2026-09-30 19:21 UTC