ADVOCACIA 5.0 · PAPER 1

A Falsa Dicotomia

entre Segurança Jurídica e IA Generativa

Como uma arquitetura de auditoria em três camadas responde ao problema das alucinações no Direito

Um ensaio técnico sobre por que confiabilidade em IA jurídica é requisito arquitetural, não feature adicional

LawAgent | Thought Leadership

Abril de 2026 · Draft v1

Sumário Executivo

Em 22 de junho de 2023, um juiz federal da Southern District of New York sancionou dois advogados por submeterem petição contendo seis decisões judiciais inexistentes — fabricadas pelo ChatGPT. O caso Mata v. Avianca tornou-se o marco simbólico de uma preocupação legítima que se espalhou rapidamente pela profissão global: se a IA generativa inventa jurisprudência com tal confiança, ela é incompatível com a prática jurídica.

A pergunta é importante, mas a resposta popular é imprecisa. A incompatibilidade não é entre IA generativa e Direito. É entre um tipo específico de aplicação ingênua da IA e o Direito. Distinguir esses dois cenários não é exercício retórico; é a diferença entre uma profissão paralisada pelo medo e uma profissão que reorganiza suas arquiteturas para absorver a nova tecnologia com segurança.

Este paper sustenta três proposições encadeadas. A primeira é que alucinação em modelos de linguagem é propriedade estatística conhecida — não defeito contornável por mais dados ou por melhores prompts — e que ignorar essa propriedade é o erro de engenharia que gerou Mata v. Avianca e seus mais de seiscentos casos análogos documentados globalmente. A segunda é que arquiteturas de verificação posterior (incluindo RAG aplicado sobre bases jurídicas) reduzem a taxa de alucinação, mas não a eliminam — estudos empíricos rigorosos do Stanford RegLab documentam taxas de erro entre 17% e 33% nas principais ferramentas comerciais de pesquisa jurídica por IA disponíveis no mercado. A terceira é que mitigação eficaz exige auditoria em múltiplos pontos independentes do pipeline, e não em uma única camada de validação ao final.

O paper descreve uma arquitetura de auditoria em três camadas — intervenções independentes antes da redação, durante a estruturação da peça, e após a geração final do texto — e discute suas implicações práticas frente à Recomendação nº 001/2024 do Conselho Federal da OAB e à Resolução nº 615/2025 do Conselho Nacional de Justiça. A tese é que confiabilidade em IA jurídica corporativa não resulta de melhor treinamento ou de prompts mais cuidadosos: resulta de decisões arquiteturais explícitas tomadas no momento em que o sistema é projetado. Um sistema arquitetado assumindo que seu modelo subjacente vai errar é estruturalmente diferente de um sistema arquitetado acreditando que seu modelo subjacente não vai errar — e só o primeiro é adequado à advocacia.

SEÇÃO 1

O Problema Real: Alucinação é Propriedade, não Bug

O que o caso Mata v. Avianca efetivamente estabeleceu

Em fevereiro de 2022, Roberto Mata entrou com ação de indenização contra a companhia aérea Avianca alegando lesão corporal sofrida em voo internacional. Sua defesa jurídica, confiada ao escritório Levidow, Levidow & Oberman, apresentou em março de 2023 uma petição de oposição à moção de extinção do processo. A petição citava seis precedentes favoráveis à tese do autor — Varghese v. China Southern Airlines, Shaboon v. EgyptAir, Petersen v. Iran Air, entre outros. Os advogados da Avianca reportaram ao juízo que não conseguiam localizar essas decisões. O juiz P. Kevin Castel, ao tentar localizá-las ele próprio, chegou à mesma conclusão: os precedentes eram fictícios. [1]

O que aconteceu em seguida é o elemento mais instrutivo do caso. O advogado Steven A. Schwartz, autor material da petição, explicou ao juiz em audiência que havia usado o ChatGPT como ferramenta de pesquisa jurisprudencial. Ao ser questionado pela corte sobre se, em algum momento, havia suspeitado da autenticidade das decisões, Schwartz declarou que, quando consultou o próprio ChatGPT para confirmar se os casos existiam, o sistema respondeu que sim — e inclusive afirmou que poderiam ser localizados no Westlaw e no LexisNexis. Schwartz, em suas palavras registradas em transcrição, estava operando sob a falsa percepção de que o site não poderia possivelmente estar fabricando casos por si próprio.

A sanção final — cinco mil dólares — foi modesta. A sanção reputacional, essa, foi devastadora. E o precedente narrativo foi estabelecido: a partir de Mata, qualquer discussão séria sobre IA na advocacia precisa passar pela questão da alucinação.

Por que LLMs alucinam — e por que isso não vai mudar

Modelos de linguagem de larga escala, os Large Language Models ou LLMs, são sistemas de previsão estatística de texto. Dado um contexto inicial, o modelo calcula, para cada próximo token possível, uma probabilidade condicional baseada nos padrões observados durante seu treinamento. A resposta que o sistema produz é a sequência de tokens mais provável em função desses padrões. Esse mecanismo — extraordinariamente poderoso em sua capacidade de produzir texto fluente — é, por construção, insensível à verdade. O modelo não possui um conceito de realidade contra o qual comparar sua saída. Ele produz o que, estatisticamente, se parece com o que veria em seus dados de treinamento.

---

[1] Mata v. Avianca, Inc., 678 F.Supp.3d 443, U.S. District Court for the Southern District of New York, decisão de 22 de junho de 2023, juiz P. Kevin Castel. Sanção de US$ 5.000 aplicada sob a Regra 11 das Federal Rules of Civil Procedure.

Esse é o fenômeno rotulado como alucinação. Em domínios onde a aparência superficial do texto é o que importa — textos de marketing, resumos genéricos, respostas a perguntas triviais —, alucinações são raras ou irrelevantes. Em domínios onde cada afirmação precisa ser factualmente correta sob pena de dano material — medicina, engenharia, Direito —, alucinações são catastróficas.

O aspecto crítico, frequentemente ignorado em discussões superficiais sobre o tema, é que alucinação não é um bug a ser corrigido. É uma propriedade estatística do método. Modelos de linguagem mais sofisticados reduzem a frequência de alucinação, mas não a eliminam. Modelos treinados em dados jurídicos específicos reduzem a frequência de alucinação jurídica, mas não a eliminam. Prompts mais cuidadosos reduzem a frequência de alucinação, mas não a eliminam. Essa é a realidade arquitetural da qual qualquer sistema sério precisa partir.

O que dizem as medidas empíricas

A quantificação rigorosa do fenômeno foi conduzida em dois estudos consecutivos pelo Stanford RegLab e pelo Stanford Institute for Human-Centered AI, liderados por Daniel E. Ho. O primeiro estudo, Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models, publicado no Journal of Legal Analysis em 2024, testou modelos generalistas de linguagem — GPT-3.5, GPT-4, Llama 2 e PaLM 2 — em mais de 200.000 consultas jurídicas verificáveis. As taxas de alucinação observadas variaram entre 58% e 88%, dependendo do modelo e do tipo de consulta, com modelos mais antigos exibindo as piores taxas.[2]

O segundo estudo, Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools, publicado no Journal of Empirical Legal Studies em 2025, testou três das principais ferramentas comerciais de pesquisa jurídica por IA vendidas no mercado americano — todas arquitetadas com base em retrieval-augmented generation, ou RAG, a arquitetura que os próprios fornecedores promoviam como solução para o problema das alucinações. Os resultados: o Lexis+ AI apresentou taxa de alucinação superior a 17%; o Westlaw AI-Assisted Research, taxa superior a 33%. Nenhuma das ferramentas eliminou alucinações. Todas as reduziram substancialmente em relação a modelos generalistas, mas a redução não foi suficiente para tornar suas saídas confiáveis sem verificação independente. [3]

A conclusão dos pesquisadores de Stanford é clara: alucinações legais não foram resolvidas. O que foi resolvido foi apenas o caso mais extremo — o de modelos generalistas sem qualquer grounding em bases jurídicas. Entre não ter qualquer grounding e ter grounding em base jurídica robusta, há uma

[2]Dahl, M., Magesh, V., Suzgun, M., e Ho, D.E., "Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models", Journal of Legal Analysis, vol. 16, 2024, pp. 64-93. DOI: 10.1093/jla/laae003.
[3]Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C.D., e Ho, D.E., "Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools", Journal of Empirical Legal Studies, vol. 22, p. 216, 2025. Publicado originalmente como preprint no arXiv em maio de 2024 (2405.20362). Conduzido pelo Stanford RegLab e pelo Stanford Institute for Human-Centered Artificial Intelligence.

melhoria mensurável. Entre ter grounding em base jurídica e produzir saídas confiáveis sem supervisão humana, há ainda um abismo.

O problema no Brasil: três casos que importam

A discussão pública sobre alucinação de IA no Direito, por razões de volume e visibilidade editorial, concentra-se em casos americanos. Mas há casos brasileiros igualmente significativos, ocorridos nos últimos trinta e seis meses, que merecem atenção específica da profissão nacional.

Caso TSE, 2023

Em abril de 2023, o corregedor-geral da Justiça Eleitoral, ministro Benedito Gonçalves, aplicou multa de R$ 2.604 por litigância de má-fé a advogado que havia protocolado petição escrita integralmente com ChatGPT em pedido de ingresso como amicus curiae em investigação eleitoral. A decisão do corregedor registrou que a petição se baseava exclusivamente em um diálogo com uma inteligência artificial, sem contribuição pessoal do requerente, classificando a conduta como temerária e infundada e afirmando que um advogado deve conhecer a inadequação do material apresentado. Esse foi, ao que se sabe, o primeiro caso formalmente documentado de sanção no Brasil envolvendo uso de IA generativa em peça processual.[4]

Caso TJSC, 2025

Em fevereiro de 2025, a 6ª Câmara Civil do Tribunal de Justiça de Santa Catarina aplicou multa de 10% sobre o valor da causa a advogado que havia utilizado ChatGPT para produzir recurso em ação de reintegração de posse. O recurso continha tanto citações jurisprudenciais quanto referências a obras doutrinárias inexistentes — todas fabricadas pelo modelo de linguagem. O advogado alegou uso inadvertido. O tribunal considerou a conduta suficientemente grave para determinar a comunicação formal do caso à OAB/SC para análise disciplinar. O desembargador relator registrou em acórdão que o surgimento de novas tecnologias de Inteligência Artificial exige que os operadores a utilizem com cautela e parcimônia, sob o risco de incorrer em reprodução de informações e fundamentos que não encontram respaldo concreto de existência, e que o exercício da advocacia, verdadeiro múnus público, atrai responsabilidades ímpares.[5]

---

[4]Tribunal Superior Eleitoral, decisão do corregedor-geral ministro Benedito Gonçalves, 2023. Multa de R$ 2.604 por litigância de má-fé aplicada a advogado que protocolou petição redigida com ChatGPT em pedido de amicus curiae.

[5]Tribunal de Justiça de Santa Catarina, 6ª Câmara Civil, decisão de fevereiro de 2025. Multa de 10% sobre o valor da causa aplicada a advogado que utilizou ChatGPT em recurso de reintegração de posse, com comunicação à OAB/SC. Processo em segredo de justiça.

Caso TRT-12, 2025

Em outubro de 2025, a advogada autora de uma petição inicial trabalhista no Tribunal Regional do Trabalho da 12ª Região teve sua cliente multada por litigância de má-fé após o juiz constatar que a peça continha jurisprudência inventada — incluindo o nome de um desembargador fictício supostamente responsável por uma decisão. O magistrado, Alexandre Martins, foi além do contorno técnico do caso individual: em fundamentação da decisão, invocou expressamente a Recomendação nº 001/2024 do Conselho Federal da OAB, afirmando que a norma exige do advogado entendimento adequado das limitações, verificação rigorosa das informações, transparência aos clientes e demais interlocutores, sendo vedada a delegação de atos privativos da profissão sem supervisão qualificada. O juiz caracterizou a petição como ato processual inexistente, consequência sem precedente na jurisprudência brasileira sobre o tema. [6]

Esses três casos — e existem outros menos divulgados, incluindo o episódio de 2023 em que a Corregedoria Regional da Justiça Federal da 1ª Região abriu investigação contra juiz federal que utilizou ferramenta de IA generativa e incluiu em sua decisão jurisprudência inexistente atribuída ao Superior Tribunal de Justiça — desenham um quadro claro. O problema da alucinação é real também no contexto jurídico brasileiro. A reação institucional é consistente, e progressivamente mais rigorosa. E a defesa do uso inadvertido tem se mostrado insuficiente para afastar responsabilidade.

No nível internacional, o quadro é ainda mais grave. A base de dados AI Hallucination Cases mantida publicamente por Damien Charlotin registrava, ao início de 2026, mais de 600 casos judiciais em que jurisprudência alucinada foi apresentada em peça processual. Desses, mais de 128 envolveram advogados formalmente sancionados ou submetidos a procedimento disciplinar. [7] Em Johnson v. Dunn, decisão da Northern District of Alabama em julho de 2025, um escritório de grande porte foi desqualificado do caso, teve suas sanções reportadas às ordens estaduais de todas as jurisdições em que seus advogados atuavam, e foi obrigado a conduzir auditoria independente de todas as suas petições ativas no circuito — precedente cuja gravidade supera em ordem de magnitude a sanção monetária de Mata v. Avianca.[8]

---

[6]Tribunal Regional do Trabalho da 12ª Região, Vara do Trabalho, decisão do juiz Alexandre Martins, outubro de 2025. Petição inicial contendo jurisprudência e nome de desembargador fictícios gerados por IA, com invocação expressa da Recomendação 001/2024 do CFOAB.

[7]Charlotin, D., "AI Hallucination Cases" (base de dados monitorada publicamente). Estado reportado por Corporate Compliance Insights em janeiro de 2026: mais de 600 casos registrados envolvendo mais de 128 advogados sancionados ou disciplinados.

[8]Johnson v. Dunn, No. 2:21-cv-1701, U.S. District Court for the Northern District of Alabama, decisão de 23 de julho de 2025. Firma de advocacia desqualificada do caso; sanções referidas à OAB estadual; determinação de auditoria de todas as petições da firma no circuito.

A inferência correta

A leitura superficial desses casos leva muitos operadores do Direito a uma conclusão defensiva: se a IA gera conteúdo falso, não devemos usar IA. Essa conclusão é compreensível emocionalmente mas é estrategicamente insustentável. Os casos ocorridos não estabelecem que IA generativa seja incompatível com a advocacia. Estabelecem, com precisão, que certas formas de uso da IA — a saber, uso direto de modelos generalistas sem arquitetura de verificação, combinado com ausência de supervisão humana qualificada — são incompatíveis com a advocacia.

A pergunta estratégica, portanto, não é se IA pode ser usada no Direito. É qual arquitetura de IA é compatível com o Direito. É a essa pergunta que as próximas seções deste paper se dedicam.

SEÇÃO 2

Por que Arquiteturas de Verificação Posterior Falham

Quando o problema das alucinações em IA jurídica começou a ser reconhecido como risco operacional sério, a resposta dominante dos fornecedores de tecnologia foi uniforme: grounding. Mais especificamente, grounding por meio de retrieval-augmented generation — o RAG, arquitetura introduzida em artigo acadêmico de 2020 como método geral para ancorar modelos de linguagem em bases factuais externas.

O que RAG faz e o que não faz

A arquitetura RAG, em sua formulação original por Lewis et al. na conferência NeurIPS 2020, funciona da seguinte maneira. Quando uma consulta chega ao sistema, o sistema primeiro recupera — retrieves — documentos relevantes a partir de uma base de conhecimento estruturada. Esses documentos são então fornecidos ao modelo de linguagem como contexto, juntamente com a pergunta original do usuário. O modelo gera sua resposta com base nesse contexto enriquecido. [9] A promessa: o modelo não precisa mais alucinar, porque a informação correta está diante dele. Basta que ele a utilize.

A promessa é parcialmente verdadeira e parcialmente enganosa. É verdadeira no sentido de que RAG reduz de fato a taxa de alucinação em relação a modelos sem grounding — essa redução é consistente e mensurável. É enganosa no sentido de que a redução não chega a zero, e os erros residuais assumem formas particularmente insidiosas.

Os dois tipos de alucinação em sistemas RAG

Os pesquisadores de Stanford que avaliaram Lexis+ AI e Westlaw AI-Assisted Research identificaram que sistemas RAG-based ainda produzem dois tipos distintos de erro. O primeiro tipo é alucinação propriamente dita: o sistema descreve o direito incorretamente ou faz afirmação factual errada, apesar de ter acesso às fontes corretas. O segundo tipo é misgrounding — o sistema descreve o direito corretamente, mas cita como fundamento uma fonte que não sustenta a afirmação.

O segundo tipo de erro é, sob certo aspecto, mais perigoso do que o primeiro. Uma jurisprudência completamente inventada, como em Mata v. Avianca, é teoricamente verificável: a contraparte ou o juízo, ao tentar localizar a decisão, descobre sua inexistência. Uma afirmação jurídica correta com citação misgrounded é muito mais difícil de detectar — a decisão citada existe, é verificável, mas não diz o que o

[9] Lewis, P., Perez, E., Piktus, A. et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks", NeurIPS 2020. Arquitetura originalmente proposta como método de redução de alucinações em modelos generativos.

sistema afirma que ela diz. Esse erro passa pelo filtro de verificação superficial e só é capturado por leitura atenta da fonte original.

A implicação prática é severa. Um advogado que utilize sistema RAG-based para produzir pesquisa jurisprudencial, e que confie na existência de cada citação retornada, ainda assim precisa ler integralmente cada decisão citada para verificar se ela efetivamente sustenta o ponto. Isso reduz — mas não elimina — o ganho de produtividade prometido pela ferramenta. E, crucialmente, o dever de verificação recai inteiramente sobre o advogado.

Por que uma única camada de verificação é insuficiente

A constatação de que sistemas RAG-based ainda alucinam entre 17% e 33% das consultas leva à pergunta seguinte: se uma camada de grounding não resolve, o que resolve?

A resposta emergente — tanto na pesquisa acadêmica quanto na prática industrial — é que verificação confiável exige intervenções independentes em múltiplos pontos do pipeline de geração. Um único ponto de controle, por mais sofisticado que seja, está sujeito a classe de falhas correlacionadas: se o sistema que realiza o grounding inicial também é o sistema que verifica a saída final, a verificação herda os mesmos vieses e limitações do modelo subjacente.

O princípio geral, conhecido em engenharia de sistemas críticos como defense in depth, sustenta que confiabilidade em sistemas complexos resulta da composição de múltiplas camadas de proteção independentes, cada uma desenhada para capturar uma classe específica de falha que as outras camadas podem não capturar. É o mesmo princípio que orienta arquiteturas de segurança em aviação, em medicina intensiva e em energia nuclear. A IA jurídica corporativa contemporânea precisa operar sob a mesma lógica.

O que isso significa em termos arquiteturais

Três pontos de intervenção independentes emergem como inescapáveis em qualquer sistema projetado para produção jurídica séria. Primeiro, uma camada de verificação no momento em que o sistema estrutura a estratégia ou plano de ataque ao problema — antes de produzir texto algum. Segundo, uma camada de verificação no momento em que o sistema executa os passos intermediários do pipeline — pesquisa, estruturação, primeira redação. Terceiro, uma camada de verificação após a produção do texto final, mas antes que esse texto seja entregue ao usuário humano.

Cada uma dessas três camadas tem como missão específica capturar tipos distintos de falha. E cada uma delas opera com mecanismos técnicos independentes, de modo a minimizar a correlação entre suas modalidades de erro.

A próxima seção descreve como essas três camadas se manifestam concretamente em uma arquitetura desenhada para produção jurídica corporativa.

SEÇÃO 3

Três Pontos de Auditoria Independentes

Esta seção descreve, em linguagem técnica sóbria, a arquitetura de auditoria em três camadas que o LawAgent implementa. A descrição é deliberadamente não-proprietária em sua base conceitual: os princípios subjacentes são aplicáveis a qualquer sistema jurídico corporativo projetado para produção séria, e refletem convergência crescente entre pesquisa acadêmica em NLP aplicado ao Direito e prática industrial em sistemas críticos de alta confiabilidade.

Camada 1 — Auditoria de Estratégia

A primeira camada de auditoria opera em um momento conceitualmente contra-intuitivo: antes que qualquer texto jurídico seja produzido. Sua função é validar a estratégia abstrata que o sistema propõe para abordar a questão colocada.

Quando um usuário submete uma consulta jurídica — por exemplo, o pedido de elaboração de uma contestação em determinada ação, ou a produção de parecer sobre viabilidade de tese específica —, um sistema agêntico bem desenhado não inicia imediatamente a redação. Inicia, em vez disso, a construção de um plano: quais teses defensivas ou ofensivas são aplicáveis ao caso, quais artigos legais são invocáveis, quais linhas jurisprudenciais sustentam cada tese, quais precedentes contrários precisam ser preemptivamente endereçados, e qual a sequência argumentativa mais robusta para a peça.

Esse plano é, em si, um artefato sujeito a alucinação. Um modelo mal-governado pode propor tese inaplicável ao caso, invocar artigo revogado, construir linha argumentativa sobre precedente inexistente. Se essa estratégia errada prosseguir para a fase de redação, o erro se propagará por todo o texto produzido, e a correção posterior será substancialmente mais custosa — cognitivamente e computacionalmente — do que a correção no nível da estratégia.

A função do auditor de estratégia é, portanto, validar três coisas antes que a redação comece. Primeiro, a aplicabilidade dos artigos legais invocados ao caso específico — um artigo citado como relevante precisa ser um artigo que efetivamente se aplica à matéria em discussão, e precisa estar em vigor na data da consulta. Segundo, a existência e pertinência dos precedentes jurisprudenciais invocados — cada precedente citado como sustentação precisa ser verificado em base de dados jurisprudencial autorizada, com confirmação de que o precedente existe, está ativo e efetivamente sustenta o ponto alegado. Terceiro, a coerência interna da linha argumentativa — as teses propostas precisam ser compatíveis entre si e cobrir adequadamente os contra-argumentos esperados da parte adversa.

O ganho estrutural dessa camada é que ela captura, no estágio de planejamento, erros que seriam caros de corrigir após produção de texto. O ganho adicional é pedagógico: a estratégia validada torna-se artefato auditável pelo advogado humano antes mesmo de investir tempo em revisar texto, permitindo intervenção precoce.

Camada 2 — Validação Estrutural

A segunda camada opera durante a execução do pipeline de produção, após a estratégia ter sido validada e enquanto o texto da peça está sendo efetivamente construído. Sua função é garantir que o documento em construção respeite as normas processuais aplicáveis e os checklists obrigatórios específicos do tipo de peça em questão.

Tipos distintos de peças jurídicas têm estruturas obrigatórias distintas. Uma petição inicial cível precisa conter qualificação das partes, fundamentos de fato e de direito, pedidos e valor da causa, conforme o Código de Processo Civil. Um recurso de apelação precisa demonstrar tempestividade, preparo, interesse recursal e fundamentos específicos. Um parecer tributário precisa endereçar matriz normativa aplicável, competência tributária envolvida, hipótese de incidência e obrigações acessórias. Cada categoria de peça carrega consigo um conjunto de requisitos formais cuja ausência pode gerar consequências processuais severas — da não-conhecimento do recurso à extinção sem resolução de mérito.

A validação estrutural opera como um sistema de checklists dinâmicos, específicos para cada tipo de peça, que verifica continuamente se os elementos obrigatórios estão sendo incorporados ao texto conforme ele é produzido. Essa camada é, em termos técnicos, mais simples do que as camadas adjacentes — envolve principalmente verificações formais e estruturais, não raciocínio substantivo sobre mérito. Mas é exatamente por essa simplicidade que ela é valiosa: captura classe de falhas cuja raiz não é alucinação, mas omissão. E omissões de elementos formais obrigatórios são, em conjunto com jurisprudência fabricada, a segunda causa mais frequente de peças tecnicamente deficientes produzidas por sistemas de IA mal-arquitetados.

Camada 3 — Auditoria Pós-Redação

A terceira camada opera após a produção do texto final, em revisão integral antes que qualquer saída seja entregue ao advogado humano. Sua função é comparar, afirmação por afirmação, o texto produzido contra fontes de autoridade independentes, identificar qualquer conteúdo não verificado, e removê-lo ou sinalizá-lo explicitamente.

Essa camada é a mais complexa tecnicamente e é onde se concentram as inovações mais recentes em engenharia de sistemas jurídicos de IA. Os mecanismos empregados incluem seis modalidades complementares de verificação, cada uma capturando classe específica de erro residual.

Mecanismos de verificação da Camada 3

Primeiro, verificação de existência de citações — cada caso, decisão ou precedente mencionado no texto é confrontado com base de dados jurisprudencial autorizada, com rejeição de qualquer citação cuja existência não seja confirmada.

Segundo, verificação de grounding — cada afirmação substantiva sobre direito é rastreada até a fonte citada, com validação de que a fonte efetivamente sustenta o que está sendo afirmado no texto. É aqui que se captura o erro de misgrounding identificado no estudo de Stanford.

Terceiro, verificação de vigência — cada norma invocada é confrontada contra bases de dados de legislação atualizada, rejeitando-se citação a artigos revogados, súmulas superadas ou entendimentos jurisprudenciais não mais vinculantes.

Quarto, verificação de consistência interna — o texto é analisado contra si mesmo para detectar contradições lógicas entre afirmações feitas em seções diferentes da peça.

Quinto, verificação de conformidade com o plano estratégico — o texto final é comparado com a estratégia validada na Camada 1, garantindo que o que foi efetivamente escrito corresponde ao que foi planejado e auditado.

Sexto, sinalização explícita de afirmações não ancoradas — qualquer conteúdo substantivo que não possa ser rastreado a uma fonte verificável é sinalizado ou removido, com transparência para o advogado revisor sobre quais partes do texto passaram por cada nível de verificação.

Por que as três camadas precisam ser independentes

A independência entre as três camadas é condição necessária de sua eficácia conjunta. Se as três auditorias fossem conduzidas pelo mesmo mecanismo ou pelo mesmo modelo subjacente, suas falhas seriam correlacionadas: um erro que escapasse à primeira camada teria probabilidade elevada de também escapar às duas seguintes, porque as três estariam cometendo o mesmo tipo de erro pelo mesmo motivo estrutural.

A engenharia de sistemas críticos resolve esse problema com o princípio da diversidade técnica: cada camada deve ser implementada com mecanismos distintos, operando sobre dados distintos, avaliando critérios distintos. Em aviação, os sistemas redundantes que sustentam voo automático usam sensores e

algoritmos diferentes entre si — não cópias do mesmo mecanismo. Em energia nuclear, os sistemas de desligamento de emergência usam tecnologias físicas diferentes das usadas pelos sistemas primários. A lógica é idêntica em IA jurídica corporativa: as três camadas de auditoria empregam técnicas distintas, acessam bases distintas e avaliam aspectos distintos do mesmo artefato.

O que essa arquitetura não promete

É fundamental caracterizar com precisão o que essa arquitetura entrega — e, sobretudo, o que ela não entrega. Três observações importam para uma leitura honesta.

Primeiro, nenhuma arquitetura elimina completamente a possibilidade de erro. Sistemas de três camadas reduzem drasticamente a probabilidade de erro em relação a sistemas de uma única camada, mas a redução é de ordem de magnitude, não é absoluta. A promessa correta é redução severa de risco, não eliminação.

Segundo, a arquitetura só funciona se cada camada for efetivamente desenhada com rigor. Uma arquitetura de três camadas mal-implementadas não é superior a uma arquitetura de uma camada bem-implementada. O que importa não é o número formal de camadas, mas a qualidade substantiva de cada uma e sua independência técnica.

Terceiro, e mais importante, nenhuma arquitetura substitui o advogado humano no papel de revisor final. A próxima seção aborda essa questão diretamente.

SEÇÃO 4

Human-in-the-Loop: O Que a Arquitetura Não Substitui

A melhor arquitetura de auditoria automatizada possível ainda é insuficiente para substituir o juízo humano qualificado em um ponto crítico do processo: o momento em que o documento é assinado e protocolado. Esta seção explicita por que isso é verdadeiro — e por que, longe de ser limitação da tecnologia, é princípio arquitetural deliberado.

O que o caso Mata realmente ensina

Retornando ao caso que abriu este paper com nova lente: o juiz P. Kevin Castel, em sua decisão sancionatória, foi explícito em uma distinção que frequentemente se perde em discussões públicas sobre o caso. Castel não sancionou os advogados pelo uso do ChatGPT como ferramenta. Citando literalmente a decisão: avanços tecnológicos são comuns, e não há nada inerentemente impróprio no uso de uma ferramenta de inteligência artificial confiável para assistência. Os advogados foram sancionados por uma coisa diferente: por terem abandonado o seu papel de guardiões — em inglês, gatekeeping role — da exatidão de suas petições.

Esse conceito de gatekeeping role é o ponto doutrinário que importa, e é ele que a jurisprudência americana subsequente vem consolidando. Em Johnson v. Dunn, a corte da Northern District of Alabama foi ainda mais específica: o advogado cuja assinatura está em uma peça é responsável por cada afirmação feita como verdadeira naquela peça, independentemente de quem a tenha redigido originalmente. O fato de o erro ter sido cometido por IA, por um subordinado, ou por uma fonte anterior, não afasta a responsabilidade do signatário. [10]

A arquitetura técnica de auditoria em três camadas que este paper descreve não tem como objetivo remover essa responsabilidade do advogado. Tem como objetivo, pelo contrário, proteger o advogado de classes de erro que — sem a arquitetura — dependeriam de sua atenção manual e exaustiva para serem capturadas. A arquitetura dá ao advogado um ponto de partida mais limpo, mais rastreável e mais auditável. Ela não o libera da obrigação de revisão final.

O princípio Human-in-the-Loop

Em engenharia de sistemas de IA, o termo Human-in-the-Loop, frequentemente abreviado como HITL, designa arquiteturas em que decisões de consequência material exigem intervenção humana explícita antes de serem executadas ou entregues. O oposto de HITL é automação integral, em que o sistema age por conta própria sem validação humana de cada ato.

Em domínios de baixa consequência, automação integral é aceitável e frequentemente desejável. Em domínios de alta consequência — e a advocacia é, por todas as métricas relevantes, um domínio de alta consequência —, HITL é requisito inegociável. Um sistema jurídico corporativo que opere sem HITL não está desenhando tecnologia: está desenhando passivo reputacional e regulatório.

O que HITL significa na prática

HITL não significa que o advogado precisa redigir tudo manualmente, o que anularia o valor da tecnologia. Significa que, entre a produção do texto pela máquina e o ato processual de protocolar a peça, há pelo menos um ponto em que um advogado habilitado lê, avalia, aprova e assume responsabilidade sobre o conteúdo. Esse ponto pode ser rápido — minutos, não horas, se o sistema entregou um produto de alta qualidade — mas não pode ser ausente.

O desenho correto de HITL em IA jurídica corporativa envolve três elementos distintos. Primeiro, transparência sobre o que o sistema fez e o que o sistema não verificou: o advogado revisor precisa saber quais partes do texto foram validadas pelas três camadas de auditoria, quais foram sinalizadas como incertas, e quais foram geradas apenas por inferência do modelo. Segundo, rastreabilidade até a fonte: cada afirmação relevante no texto precisa ter ligação direta a fonte verificável que o advogado pode consultar se quiser aprofundar-se. Terceiro, facilidade de intervenção: a interface que apresenta o texto produzido deve facilitar — não dificultar — que o advogado revise, edite, reescreva ou rejeite porções do conteúdo.

Alinhamento com a Recomendação CFOAB nº 001/2024

O princípio HITL não é apenas boa prática de engenharia. É, no contexto brasileiro, exigência normativa explícita. A Recomendação nº 001/2024 do Conselho Federal da OAB, em seu item 3.3, estabelece textualmente que a dependência excessiva de ferramentas de IA é inconsistente com a prática da advocacia e não pode substituir a análise realizada pelo advogado. O item 3.5 acrescenta que, ao optar pelo uso de IA generativa, o advogado deve se envolver em contínua aprendizagem sobre os conteúdos gerados por IA e suas implicações para a prática jurídica, mediante capacitações constantes. [11]

A Recomendação estabelece ainda, em seu item 2.2, dever de diligência na escolha do fornecedor de IA — exigindo que o advogado verifique se o fornecedor garante proteção das informações, adoção de medidas de segurança, e vedação ao uso dos dados fornecidos para treinamento de sistemas. Essa

[11]Recomendação CFOAB nº 001/2024, item 3.3: "A dependência excessiva de ferramentas de IA é inconsistente com a prática da advocacia e não pode substituir a análise realizada pelo advogado." Item 2.2: diligência na escolha do fornecedor de IA quanto a garantias de confidencialidade e vedação de uso dos dados para treinamento de sistemas.

exigência é particularmente relevante em contexto empresarial: um sistema jurídico corporativo operando sobre informações sensíveis de clientes precisa contratualmente garantir tenant isolation — isolamento lógico entre dados de clientes distintos — e vedação explícita ao uso dos dados processados para treinamento de modelos compartilhados. [12]

O princípio do sigilo profissional e a arquitetura do sistema

O Art. 7º, inciso XIX, do Estatuto da Advocacia (Lei nº 8.906/1994) estabelece o sigilo profissional como prerrogativa — e dever — do advogado. A Lei Geral de Proteção de Dados (Lei nº 13.709/2018) adiciona camada complementar de obrigações quanto ao tratamento de dados pessoais. A intersecção dessas duas normas tem implicações arquiteturais específicas para qualquer sistema de IA aplicado à advocacia. [13]

A primeira implicação é que dados processados pelo sistema precisam ser logicamente isolados entre clientes. Nenhuma informação da banca A pode ser acessível ou inferível a partir de interações com a banca B, mesmo que ambas sejam clientes do mesmo fornecedor. Isso se materializa em tenant isolation arquitetural estrito — separação lógica no nível do banco de dados, dos modelos de inferência, dos logs de auditoria, e dos contextos de conversação.

A segunda implicação é que dados processados não podem ser utilizados para treinamento de modelos que serão compartilhados com outros clientes. A Recomendação CFOAB nº 001/2024 é explícita nesse ponto: o advogado deve verificar contratualmente a não-utilização dos dados fornecidos para treinamento de sistemas. Sistemas que usem os dados de um cliente para melhorar suas respostas a outros clientes violam simultaneamente a Recomendação da OAB e os princípios de confidencialidade subjacentes ao Estatuto.

A terceira implicação é que logs de auditoria precisam ser preservados e acessíveis ao próprio cliente — tanto para fins de compliance quanto para eventual reconstrução forense de como determinada peça foi produzida. Cada interação com o sistema deve ser rastreável por identificador único de fluxo, com registro de qual informação foi consultada, qual foi a saída produzida, e quais validações foram executadas pelas três camadas de auditoria.

Human-in-the-Loop não é disclaimer — é arquitetura

---

[12]Conselho Federal da Ordem dos Advogados do Brasil, Recomendação nº 001/2024, aprovada em 11 de novembro de 2024 por meio da Ementa nº 045/2024/COP, Proposição nº 49.0000.2024.007325-9. Publicada no Diário Eletrônico da OAB em 14 de novembro de 2024.

[13]Lei nº 8.906/1994 (Estatuto da Advocacia e da OAB), Art. 7º, inciso XIX, que estabelece o dever de sigilo profissional; Lei nº 13.709/2018 (Lei Geral de Proteção de Dados Pessoais).

É comum que fornecedores de IA incluam, em termos de serviço, cláusulas isentando-se de responsabilidade por saídas incorretas e exigindo que o usuário final verifique tudo. Essa prática, embora juridicamente defensável sob certa ótica contratual, não cumpre o que se entende por HITL na engenharia moderna. HITL não é um disclaimer legal. É uma propriedade técnica do sistema.

Um sistema que simplesmente imprime um aviso dizendo verifique tudo não é HITL. Um sistema que estrutura sua interface, seus metadados e seus fluxos de trabalho de modo a facilitar a revisão humana qualificada, sim, é HITL. A distinção importa tanto para a eficácia prática quanto para a legitimidade regulatória.

SEÇÃO 5

O Quadro Regulatório Brasileiro e suas Implicações

As seções anteriores descreveram o problema das alucinações e a arquitetura de mitigação em três camadas combinada com supervisão humana. Esta seção coloca ambos dentro do marco regulatório brasileiro específico — que, ao contrário do que muitas discussões superficiais sugerem, já existe com contornos razoavelmente definidos e com consequências práticas imediatas.

A Recomendação CFOAB nº 001/2024 em detalhe

Aprovada em 11 de novembro de 2024 pelo Conselho Pleno da OAB por meio da Ementa nº 045/2024/COP, e publicada no Diário Eletrônico da OAB em 14 de novembro de 2024, a Recomendação nº 001/2024 é o principal documento normativo brasileiro específico sobre uso de IA generativa pela advocacia. Estruturada em quatro pilares — Legislação Aplicável, Confidencialidade e Privacidade, Prática Jurídica Ética, e Comunicação sobre o Uso de IA Generativa —, a Recomendação foi elaborada pelo Observatório Nacional de Cibersegurança, Inteligência Artificial e Proteção de Dados do CFOAB, sob relatoria do conselheiro federal Francisco Queiroz Caputo Neto. [14]

Embora não seja norma sancionatória em sentido estrito — como o próprio relator observou em sua apresentação, sanções exigem reserva legal —, a Recomendação estabelece parâmetros de diligência cujo descumprimento é juridicamente invocável tanto em procedimentos disciplinares na OAB quanto em decisões judiciais que avaliem responsabilidade de advogados por peças processuais defeituosas. O caso TRT-12 discutido anteriormente é o exemplo canônico dessa eficácia indireta: um juiz federal invocou a Recomendação 001/2024 como fundamento para caracterizar uma petição como ato processual inexistente.

Os itens mais operacionalmente relevantes

Três itens específicos da Recomendação têm implicações diretas para arquitetura de sistemas e para rotinas de uso cotidiano.

O item 2.1 estabelece que ao incluir informações em sistemas de IA, o advogado deve zelar pela confidencialidade e sigilo profissional dos dados apresentados, com especial atenção a dados sensíveis. O item 2.2 impõe diligência na escolha do sistema de IA para garantir que o fornecedor protege as informações, adota medidas de segurança, e não utiliza os dados fornecidos para treinamento de sistemas. O item 3.3 é o mais categórico: A dependência excessiva de ferramentas de IA é inconsistente com a prática da advocacia e não pode substituir a análise realizada pelo advogado. [15]

O item 4.1.1 acrescenta uma camada menos discutida mas igualmente importante: o advogado que optar por utilizar ferramentas ou sistemas de IA na prestação de serviços advocatícios deve, previamente ao início de sua utilização, formalizar tal intenção ao cliente. Essa exigência de comunicação prévia — por contrato, aviso expresso ou outro meio adequado — tem implicações práticas ainda pouco absorvidas pela advocacia brasileira: escritórios que adotam IA precisam revisar seus instrumentos contratuais com clientes para garantir transparência sobre o uso da tecnologia.

A Resolução CNJ nº 615/2025 como ancoragem regulatória adjacente

Embora dirigida formalmente ao Poder Judiciário — e não à advocacia privada —, a Resolução nº 615, de 11 de março de 2025, publicada pelo Conselho Nacional de Justiça e em vigor desde 14 de julho de 2025, tem relevância regulatória indireta significativa para fornecedores de tecnologia que operam no ecossistema jurídico brasileiro. A Resolução estabelece princípios obrigatórios para soluções de IA utilizadas no Judiciário, incluindo supervisão humana, explicabilidade, classificação de risco, cadastro no Sinapses (plataforma de catalogação de sistemas de IA no Judiciário), e avaliação de impacto algorítmico para sistemas de alto risco. [16]

Fornecedores de IA que operam simultaneamente no Judiciário e na advocacia privada tendem a alinhar arquitetural e operacionalmente suas soluções aos padrões estabelecidos pela Resolução 615/2025 — tanto por eficiência de engenharia quanto porque esses padrões se tornam, de fato, referência informal de boas práticas setoriais. Um escritório de advocacia que contrata fornecedor cuja arquitetura não satisfaz os critérios da Resolução 615/2025 adota tecnologia de categoria inferior ao estado da arte regulatório brasileiro, mesmo que isso não seja tecnicamente exigido por norma dirigida à advocacia privada.

O que um sistema jurídico corporativo brasileiro precisa satisfazer

A combinação da Recomendação CFOAB nº 001/2024 com a Resolução CNJ nº 615/2025 — e com os princípios subjacentes do Estatuto da Advocacia e da LGPD — permite derivar uma lista operacional de requisitos que sistemas jurídicos corporativos sérios precisam satisfazer no Brasil.

- Arquitetura de supervisão humana obrigatória, com Human-in-the-Loop como propriedade técnica do sistema e não como disclaimer contratual.
- Tenant isolation lógico e verificável, com dados de clientes distintos estritamente segregados nos níveis de armazenamento, inferência e logging.

[16] Conselho Nacional de Justiça, Resolução nº 615, de 11 de março de 2025. Vigência a partir de 14 de julho de 2025. Estabelece diretrizes para desenvolvimento, utilização e governança de soluções de IA no Poder Judiciário.

- Vedação contratual e técnica ao uso de dados de clientes para treinamento de modelos compartilhados com outros clientes.
- Arquitetura de auditoria em múltiplas camadas independentes, capaz de reduzir taxas de alucinação a níveis substancialmente inferiores aos observados em sistemas de camada única.
- Rastreabilidade completa por identificador único de fluxo, com logs preserváveis e acessíveis ao cliente final para fins de compliance e auditoria forense.
- Mecanismos de sinalização explícita de incertezas, permitindo ao advogado revisor distinguir entre afirmações verificadas contra fontes autorizadas e afirmações geradas por inferência do modelo.
- Transparência arquitetural para o cliente-escritório, incluindo documentação clara sobre quais verificações são executadas, quais bases são consultadas, e quais limitações permanecem.
- Conformidade comprovada com LGPD, incluindo base legal clara para tratamento de dados, medidas técnicas e administrativas de segurança, e canais de exercício de direitos do titular quando aplicável.

Essa lista não é exaustiva nem eterna. O quadro regulatório brasileiro está em evolução, e novos critérios provavelmente emergirão nos próximos vinte e quatro meses. Mas ela captura os requisitos que, ao início de 2026, separam sistemas adequados à advocacia brasileira corporativa daqueles que carregam risco regulatório relevante para seus usuários.

CONCLUSÃO

A Segurança não é Feature — é Arquitetura

O argumento central deste paper pode ser resumido em uma proposição simples: confiabilidade em IA jurídica corporativa não é uma feature que se adiciona ao produto no final do desenvolvimento. É uma decisão arquitetural tomada no primeiro dia e refletida em cada escolha técnica subsequente.

Essa distinção importa porque ela separa duas filosofias de engenharia. A primeira assume que o modelo subjacente é confiável e trata erros como casos excepcionais que podem ser tratados por camadas de verificação superficial. A segunda assume que o modelo subjacente vai errar — não porque seja mal-projetado, mas porque modelos de linguagem são, por construção, sistemas estatísticos insensíveis à verdade — e desenha todo o produto para capturar esses erros antes que cheguem ao advogado revisor.

A primeira filosofia produziu os cenários descritos na Seção 1: Mata v. Avianca, os casos brasileiros no TSE, TJSC e TRT-12, os mais de seiscentos casos documentados internacionalmente. A segunda filosofia produz sistemas que, embora imperfeitos — e nenhum sistema jamais será perfeito —, são adequados ao padrão de diligência que a profissão jurídica exige e que a Recomendação CFOAB nº 001/2024 codifica.

> *Um sistema de IA jurídica corporativa não é confiável porque seu modelo subjacente é bom. É confiável porque sua arquitetura assume, desde o primeiro dia, que seu modelo subjacente vai errar — e captura esses erros antes que cheguem ao advogado.*

Essa é a tese que orienta o desenho do LawAgent: segurança é requisito arquitetural primário, não feature complementar. Auditoria em três camadas independentes, Human-in-the-Loop como propriedade técnica, tenant isolation estrito, rastreabilidade por flow_id, sinalização explícita de conteúdo não verificado — todos esses elementos compõem uma decisão única e coerente sobre como um sistema compatível com a advocacia brasileira precisa ser construído.

A falsa dicotomia entre segurança jurídica e IA generativa é falsa porque assume que a tecnologia é o que ela era em 2023, quando o caso Mata foi julgado. A tecnologia não é a mesma. A arquitetura não é a mesma. Os padrões regulatórios não são os mesmos. E as expectativas da profissão não são as mesmas. O que permanece é a responsabilidade do advogado — e essa, nunca foi objeto de terceirização.

Entre uma profissão paralisada pelo medo de uma tecnologia que ainda não entende, e uma profissão que absorve com discernimento os avanços técnicos que maturam em cada ciclo de pesquisa, existe

apenas a disposição de compreender, em detalhe, como as arquiteturas contemporâneas efetivamente funcionam. Este paper buscou contribuir para essa compreensão. Futuros papers desta série aprofundarão aspectos complementares — preservação de identidade estilística, gestão de conhecimento institucional, disciplina financeira por fluxo, entre outros. A espinha dorsal comum, porém, já está traçada: sistemas sérios para uma profissão séria.

Sobre o LawAgent

O LawAgent é uma infraestrutura de inteligência jurídica corporativa desenvolvida para atuar como orquestrador agêntico ao lado de advogados em bancas brasileiras. Sua arquitetura integra auditoria em três camadas independentes para mitigação de risco de alucinação, tenant isolation estrito para preservação de sigilo profissional, rastreabilidade completa por identificador de fluxo, e Human-in-the-Loop como propriedade técnica inegociável. O LawAgent opera em conformidade com a Recomendação nº 001/2024 do Conselho Federal da OAB, com a Resolução nº 615/2025 do Conselho Nacional de Justiça, e com a Lei Geral de Proteção de Dados.

Contato institucional: insights@lawagent.com.br