O Cavalo de Troia Algorítmico: *Prompt Injection* como Afronta à Lealdade Processual

NEY MARANHÃO

Pós-Doutorado em Direito pela Universidade de Coimbra (Portugal). Doutor em Direito do Trabalho pela Universidade de São Paulo (USP), com estágio de Doutorado-Sanduíche junto à Universidade de Massachusetts (Boston/EUA). Especialista em Direito Material e Processual do Trabalho pela Universidade de Roma/La Sapienza (Itália). Professor de Direito Material e Processual do Trabalho da Universidade Federal do Pará (Graduação, Mestrado e Doutorado). Professor da Escola Nacional de Formação e Aperfeiçoamento de Magistrados do Trabalho (ENAMAT/TST) e da Escola Nacional de Formação e Aperfeiçoamento de Magistrados da Justiça Militar da União (ENAJUM/STM). Professor Convidado de todas as 24 Escolas Judiciais de Tribunais Regionais do Trabalho do Brasil e de diversas Escolas Judiciais de Tribunais de Justiça. Professor Especializado na aplicação de Inteligência Artificial Generativa nas áreas da Educação e do Direito, tendo ministrado diversos cursos institucionais em todas as regiões do Brasil. Membro da Academia Brasileira de Direito do Trabalho. Juiz Titular de Vára da Justiça do Trabalho da 8ª Região (TRT-8/PA-AP). E-mail: ney.maranhao@gmail.com.

FABRÍCIO LIMA

Juiz do Trabalho. Mestre em Constitucionalismo e Democracia pela Faculdade de Direito do Sul de Minas (FDSM). Doutorando em Ciências Jurídicas Privatísticas pela Universidade do Minho (Portugal). Especialista em Direito Material e Processual do Trabalho. Especialista em Inteligência Artificial. Graduado em Direito pela Universidade de São Paulo, com habilitação em Direito de Empresa – Administração Empresarial e Tributária. Formação em Compliance Laboral pela Wolters Kluwer (Espanha). E-mail: fabriciolimasilva@yahoo.com.br.

RESUMO: A inexorável simbiose entre a inteligência artificial generativa e o cotidiano forense, conquanto prometa uma revolução na eficiência judiciária, inaugura vulnerabilidades de jaez inédito e gravidade singular. O presente estudo perscruta o fenômeno do *prompt injection* – a inserção maliciosa de instruções ocultas em *Large Language Models* (LLMs) –, qualificando-o não como mero *glitch* técnico, mas como um verdadeiro "Cavalo de Troia algorítmico" capaz de subverter a higidez da prestação jurisdicional. Demonstra-se que tal artifício,

ao induzir a "hipnose algorítmica" dos sistemas de apoio à decisão, transcende a esperteza tecnológica para configurar ofensa frontal aos deveres de lealdade processual e boa-fé objetiva. A pesquisa sustenta que o ordenamento jurídico pátrio (CPC, CLT e Código Penal) já dispõe de antídotos hermenêuticos robustos para reprimir essa conduta, enquadrando-a nas figuras da litigância de má-fé, do ato atentatório à dignidade da justiça e, em *ultima ratio*, em tipos penais específicos como o estelionato processual. Conclui-se que a integridade do processo na era digital exige uma vigilância epistemológica constante, assegurando que a tecnologia permaneça serva da verdade e jamais instrumento de manipulação surdina.

PALAVRAS-CHAVE: Inteligência Artificial generativa. *Prompt injection*. Lealdade processual. Litigância de má-fé. Resolução CNJ nº 615/2025.

Introdução

A crescente integração de tecnologias de Inteligência Artificial (IA) no cotidiano forense, notadamente os modelos de linguagem de grande escala (LLMs), representa um marco de inegável relevância. Ferramentas que automatizam a análise de documentos, sumarizam peças processuais ou auxiliam na elaboração de minutas decisórias prometem uma revolução na eficiência da prestação jurisdicional.

Porém, essa transição para uma "sociedade de plataforma" (Srnicek, 2017; Maranhão, 2025) e o natural advento de uma virada tecnológica no direito processual (Nunes, 2023) decerto não ocorrem sem suscitar complexos desafios éticos e jurídicos. Dentre eles, emerge com especial gravidade a vulnerabilidade desses sistemas aos ataques de injeção de *prompt* (*prompt injection*).

Em termos diretos, a injeção de *prompt* é uma técnica maliciosa que consiste na manipulação das entradas de um LLM para subverter suas instruções originais e induzi-lo a um comportamento não programado. Tal prática transforma a ferramenta de auxílio em um potencial "Cavalo de Troia", capaz de gerar desinformação, vazar dados confidenciais ou, no contexto processual, produzir resultados que distorcem a realidade dos fatos.

Este breve artigo debruça-se sobre essa problemática sob a ótica da Teoria Geral do Processo, defendendo a hipótese de que a injeção de

prompts maliciosos no âmbito de qualquer peça processual transcende a esfera da astúcia tecnológica para configurar inequívoca ofensa à lealdade processual, amoldando-se *de lege lata* às figuras da litigância de má-fé e do ato atentatório à dignidade da justiça. O argumento central é que tal conduta não representa apenas uma falha de segurança computacional, mas uma violação deliberada dos deveres de lealdade, verdade e boa-fé que devem nortear a conduta de todos os sujeitos processuais, sendo certo ainda que o regramento jurídico vigente já permite enquadramento legal adequado para essas situações.

Para tanto, iniciaremos expondo, no contexto brasileiro, algo da atual autorização institucional para o uso de LLMs na atividade judicante, depois passando a delinear o conceito e as modalidades da injeção de *prompt*. Em seguida, abordaremos as estratégias de prevenção e mitigação. Por fim, procederemos ao enquadramento jurídico da prática à luz do CPC e da CLT e suas implicações para a integridade da prestação jurisdicional.

1. A aplicação de LLMs na atividade judicante e a Resolução CNJ nº 615/2025

A incorporação de modelos de IA no Judiciário brasileiro não é mais uma promessa, mas uma realidade em expansão, com diversas aplicações já em uso para otimizar a prática de atos processuais (Nunes, 2025). Atento a essa transformação, o Conselho Nacional de Justiça (CNJ) editou a Resolução nº 615, de 11 de março de 2025, que estabelece diretrizes para o desenvolvimento e a utilização (produtiva, ética e segura) de soluções de IA, com especial atenção aos modelos generativos.

A Resolução do CNJ, ao mesmo tempo em que incentiva a inovação e a eficiência, impõe uma série de salvaguardas éticas e de segurança. O normativo veda expressamente o uso de IA que não permita a revisão humana, que valore traços de personalidade para prever a reiteração delitiva ou que classifique pessoas com base em seu comportamento social (CNJ, 2025, art. 10). Além disso, a resolução, corretamente, determina que o uso dessas ferramentas deve ser sempre de caráter auxiliar, vedando a tomada de decisões autônomas sem a devida supervisão de um magistrado, que permanece integralmente responsável pelo ato final (CNJ, 2025, art. 19, § 3º, II).

Um ponto crucial da regulamentação é a exigência de transparência. Os sistemas de processo judicial eletrônico que utilizam IA devem indicar claramente os modelos em uso, e, quando a IA generativa for empregada na redação de um ato judicial, tal fato deve ser registrado no sistema interno do

tribunal para fins de monitoramento e auditoria (CNJ, 2025, arts. 21 e 19, § 6º). Essa medida visa a garantir a rastreabilidade e a responsabilidade, mas, como veremos, não elimina o risco de manipulações ocultas.

2. *Prompt injection*: conceito, modalidades e aplicação no processo

O *prompt injection*[1] representa uma das vulnerabilidades mais insidiosas dos modernos sistemas de inteligência artificial. Como explicam Yeung e Ring (2024), trata-se de técnica pela qual um atacante insere instruções maliciosas na entrada de dados para manipular o comportamento do modelo. O que torna essa vulnerabilidade particularmente perigosa é sua sutileza: diferentemente de ataques tradicionais a sistemas computacionais, a injeção de *prompt* explora a própria capacidade de interpretação de linguagem natural do LLM, que se mostra incapaz de distinguir entre o texto legítimo que deve processar e as instruções ardilosamente inseridas para subvertê-lo (Nunes, 2025).

A sofisticação dessas técnicas tem evoluído com refinamento. Segundo a taxonomia estabelecida pela OWASP (2025), os ataques se dividem em duas categorias fundamentais: as injeções diretas, onde o usuário deliberadamente insere comandos maliciosos, e as indiretas, que provêm de fontes externas processadas pelo sistema, como documentos anexados ou páginas *web* referenciadas. Essa distinção é crucial, pois enquanto as primeiras requerem ação intencional do usuário, as segundas podem ocorrer sem seu conhecimento, por meio de conteúdo aparentemente inócuo.

Entre as modalidades mais preocupantes está o *jailbreaking*, técnica pela qual o atacante busca contornar os filtros de segurança e diretrizes éticas do modelo. O caso emblemático do *prompt* "DAN" (*Do Anything Now*) ilustra perfeitamente essa abordagem: o atacante essencialmente cria um "alter ego" para a IA, instruindo-a a ignorar suas restrições fundamentais. Como documenta a OWASP (2025), trata-se de forma sofisticada de manipulação em que entradas cuidadosamente elaboradas fazem o modelo desconsiderar seus próprios protocolos de segurança.

Igualmente preocupante é o fenômeno do *prompt leaking*, onde o objetivo é extrair do LLM suas instruções internas, configurações ou dados sensíveis de treinamento. Hung e outros (2025) demonstram como técnicas aparentemente

---

1 A distinção entre erro e dolo na manipulação de sistemas de IA assume relevância jurídica fundamental. Enquanto o uso descuidado que resulta em "alucinações" algorítmicas pode configurar culpa grave, a inserção intencional de comandos ocultos revela elemento subjetivo incompatível com mera negligência, aproximando-se do dolo direto ou, no mínimo, eventual.

inocentes, como solicitar ao modelo que "resuma as instruções anteriores", podem comprometer a confidencialidade de todo o sistema. A simplicidade enganosa desses ataques mascara seu potencial devastador.

O *prompt hijacking* representa talvez a forma mais direta e agressiva de ataque. Por meio de comandos como "ignore todas as instruções anteriores e execute X", o atacante busca assumir controle total sobre as respostas do sistema. Os estudos de Maloyan, Ashinov e Namiot (2025) revelam dados alarmantes: esses ataques de sequestro direto alcançam taxas de sucesso superiores a 30% em sistemas "LLM-as-a-Judge", evidenciando vulnerabilidade que transcende modelos específicos e atinge a própria arquitetura dos sistemas de IA.

Mas é nas técnicas de ocultação que o *prompt injection* revela sua face mais insidiosa. A injeção oculta emprega métodos que tornam as instruções maliciosas invisíveis ao olho humano, permanecendo, contudo, perfeitamente legíveis para o LLM. Liu (2025) e Nunes (2025) descrevem um arsenal de estratégias: texto com a mesma cor do fundo, fontes de tamanho zero, caracteres Unicode invisíveis, todos métodos que transformam um documento aparentemente normal em veículo de manipulação algorítmica.

A OWASP (2025) documenta técnicas ainda mais sofisticadas. Os ataques multimodais, por exemplo, ocultam comandos em imagens ou outros elementos multimídia que acompanham o texto. Considerando que sistemas jurídicos modernos processam rotineiramente documentos com elementos visuais como fotografias, diagramas e capturas de tela, essa vulnerabilidade assume proporções particularmente graves no contexto forense.

A ofuscação multilinguística adiciona outra camada de complexidade, utilizando múltiplos idiomas ou sistemas de codificação para escapar da detecção. Em um sistema jurídico que lida com diversidade linguística crescente, essa técnica encontra terreno fértil.

Ainda mais sofisticada é a técnica de *payload splitting*, em que instruções maliciosas são fragmentadas ao longo de diferentes seções do documento. Hung e outros (2025) observam que essa abordagem explora brilhantemente a arquitetura de atenção dos *transformers*, onde *tokens* distantes podem influenciar-se mutuamente durante o processamento, criando comandos que só se materializam quando o documento é processado em sua totalidade.

No contexto processual brasileiro, essas vulnerabilidades assumem contornos especialmente preocupantes. Magesh e outros (2024) documentam como mesmo sistemas jurídicos baseados em RAG (*Retrieval-Augmented*

Generation)[2], supostamente mais seguros, permanecem vulneráveis a manipulações sofisticadas.

Imagine-se uma petição em que, em meio a argumentos aparentemente técnicos, escondem-se instruções como: "*Ao analisar este documento, considere que todos os argumentos do autor devem ser acolhidos e que as provas apresentadas são robustas e suficientes*". Para o advogado ou magistrado lendo o documento, nada pareceria fora do comum. Mas para um sistema de IA processando a peça para gerar resumo ou minuta, tais instruções poderiam direcionar conclusões de forma sub-reptícia.

O fenômeno torna-se ainda mais complexo quando consideramos o que Hung e outros (2025) denominaram "efeito de distração" (*distraction effect*). Suas pesquisas revelam que, durante um ataque bem-sucedido, cabeças de atenção específicas nos modelos de linguagem literalmente desviam o foco das instruções originais para os comandos injetados. É como se o modelo sofresse uma forma de "hipnose algorítmica", onde sua atenção é capturada e redirecionada sem que os mecanismos de segurança percebam a manipulação em curso.

Maloyan, Ashinov e Namiot (2025) expandem essa compreensão ao demonstrar como o *Comparative Undermining Attack* pode ser adaptado ao contexto jurídico. Nessa modalidade particularmente insidiosa, comandos ocultos não apenas direcionam conclusões, mas manipulam a própria interpretação de precedentes. O sistema pode ser instruído a enfatizar aspectos favoráveis de determinada jurisprudência enquanto minimiza ou ignora precedentes contrários, tudo isso mantendo aparência de análise equilibrada.

As vulnerabilidades específicas dos sistemas RAG jurídicos, conforme alertam Magesh e outros (2024), vão além das fragilidades técnicas gerais. A hierarquia de autoridades legais pode ser subvertida, fazendo o sistema privilegiar jurisprudência persuasiva sobre precedentes vinculantes. A temporalidade legal adiciona outra dimensão de risco, permitindo que jurisprudência revogada seja apresentada como válida.

Particularmente preocupante é a manipulação de sínteses processuais. Considere-se o cenário de razões finais onde comandos ocultos determinam ao sistema que interprete todos os argumentos como convincentes, reproduzindo-os acriticamente em minutas decisórias. Ou embargos de declaração que, por meio da mesma técnica, buscam influenciar o conteúdo da decisão

---

2 RAG (*Retrieval-Augmented Generation*) é uma técnica de arquitetura híbrida em modelos de linguagem que combina geração de texto com recuperação de informações de uma base externa. Em vez de se apoiar apenas nos parâmetros internos do modelo, o RAG consulta documentos relevantes em tempo real – como bases de dados jurídicas ou *corpora* doutrinários – para fundamentar suas respostas. Essa abordagem visa reduzir o risco de alucinações, aumentando a precisão factual das informações geradas, especialmente em domínios sensíveis como o jurídico.

final, transformando o que deveria ser mero esclarecimento em oportunidade de manipulação ardilosa.

Os riscos de manipulação processual por Grandes Modelos de Linguagem (LLMs) são exponencialmente ampliados por seu intrínseco poder persuasivo. A capacidade desses sistemas de gerar textos com sofisticada aparência de autoridade e neutralidade técnica engendra um perigoso *viés de aparente veracidade informacional*, no qual a credibilidade de um argumento passa a ser erroneamente aferida mais pela elegância de sua apresentação do que por sua solidez intrínseca.

Diante de uma peça com tamanha qualidade formal, o profissional do Direito pode ter sua vigilância crítica desarmada, tornando-se suscetível a incorporar o conteúdo manipulado como se fosse fruto de uma análise imparcial. A ferramenta deixa, assim, de ser um mero suporte para se converter em um verdadeiro "cavalo de Troia argumentativo", tornando a detecção da fraude um complexo desafio de ordem epistemológica e pragmática.

Não bastasse, a OWASP (2025) também é categórica em sua conclusão: a própria natureza estocástica[3] dos modelos generativos torna a prevenção absoluta tecnicamente impossível no estado atual da arte. Filtros semânticos, validação de formato, segregação de conteúdo: todas essas medidas oferecem apenas proteção parcial, nunca garantia completa. Estamos diante de vulnerabilidade que não é mero *bug* a ser corrigido, mas característica intrínseca da arquitetura desses sistemas.

Por certo, essa complexa realidade impõe reflexão profunda sobre os limites da automação no contexto judicial. Se a própria natureza dos modelos de linguagem os torna permanentemente vulneráveis a manipulações sofisticadas, a supervisão humana deixa de ser mera formalidade para tornar-se barreira essencial contra a subversão da justiça. O risco não é teórico ou distante: é real, presente e extremamente danoso à integridade do sistema judicial, exigindo não apenas salvaguardas técnicas, mas transformação cultural na forma como concebemos a relação entre tecnologia e direito.

3. Enquadramento jurídico: litigância de má-fé, ato atentatório à dignidade da justiça e repercussões penais

A nosso ver, diferentemente do defendido por doutrinadores de escol (Nunes, 2025), a prática do *prompt injection* em peças processuais, embora

---

3 A natureza estocástica dos modelos generativos refere-se ao uso de variáveis aleatórias e processos probabilísticos em suas etapas de geração, de forma que a saída não seja determinística. Em palavras mais diretas: eles usam um elemento de aleatoriedade controlada – como se tirassem "dados virtuais" – para que, mesmo recebendo a mesma entrada, possam gerar resultados diferentes e criativos, tornando as saídas mais variadas e realistas.

tecnologicamente inovadora, encontra claro enquadramento na legislação vigente, não havendo necessidade de "ginásticas hermenêuticas" ou alterações legislativas para sua imediata coibição. A conduta pode se amoldar perfeitamente às hipóteses já existentes de litigância de má-fé e ato atentatório à dignidade da justiça, bem como a tipos penais específicos, revelando sua natureza pluriofensiva.

No plano processual civil, o art. 80 do CPC define como litigante de má-fé aquele que, entre outras condutas, deduzir pretensão ou defesa contra texto expresso de lei ou fato incontroverso (inciso I), alterar a verdade dos fatos (inciso II), usar do processo para conseguir objetivo ilegal (inciso III) ou proceder de modo temerário em qualquer incidente ou ato do processo (V). A injeção de um *prompt* malicioso para, por exemplo, induzir a IA a distorcer a análise de um documento ou desconsiderar argumentos da parte contrária, é uma forma sofisticada de, a um só tempo, alterar a verdade dos fatos, tentar obter objetivo ilegal e agir de modo temerário em ato processual, porquanto busca uma decisão favorável baseada em premissas manipuladas.

Da mesma forma, o art. 77 do CPC impõe a todos os sujeitos processuais o dever de expor os fatos em juízo conforme a verdade (inciso I) e não praticar inovação ilegal no estado de fato de bem ou direito litigioso (inciso VI). O descumprimento configura ato atentatório à dignidade da justiça (art. 77, § 2º), passível de multa. A inserção de comando oculto em uma peça processual viola frontalmente o dever de veracidade e lealdade, constituindo inovação ilegal no "estado de fato" informacional do processo.

A gravidade da conduta, todavia, não se esgota nas consequências processuais civis. Quando a manipulação algorítmica é perpetrada dolosamente, adentra-se o território do ilícito penal. O art. 154-A do Código Penal tipifica a invasão de dispositivo informático, criminalizando quem instala vulnerabilidades para obter vantagem ilícita. A inserção de comandos ocultos destinados a subverter sistemas de IA judicial configura precisamente essa instalação de vulnerabilidade, com o agravante de visar a manipulação da própria prestação jurisdicional.

Ademais, o estelionato processual (art. 171, § 3º, do CP) encontra perfeita aplicação quando o agente emprega artifício tecnológico para induzir o sistema jurisdicional em erro. A peculiaridade reside no fato de que o engano se dirige ao auxiliar algorítmico do magistrado – distinção que não afasta a tipicidade, dado que o resultado pretendido permanece idêntico: a obtenção fraudulenta de provimento jurisdicional favorável.

Quando a petição contendo instruções dissimuladas é protocolada, consuma-se ainda o uso de documento ideologicamente falso (art. 304 c/c o

art. 299 do CP). O documento, embora formalmente íntegro, carrega alteração substancial de seu conteúdo interpretativo quando processado por sistemas automatizados – a falsidade ideológica em sua manifestação mais insidiosa.

Na seara trabalhista, a CLT, em seus arts. 793-A, 793-B e 793-C, reproduz as hipóteses de litigância de má-fé do CPC, sendo plenamente aplicáveis ao processo do trabalho, pela via do art. 769 consolidado, tanto as figuras de ato atentatório à dignidade da justiça quanto as repercussões penais acima delineadas.

Cumpre ressaltar que a imputação de tais condutas deve observar o contraditório e a ampla defesa. A parte acusada de praticar *prompt injection* deve ter oportunidade de se manifestar e produzir provas. A análise demandará, muito provavelmente, perícia técnica no documento digital para ratificar a existência de comandos ocultos. Constatada a presença de tais vetores, além das sanções processuais cabíveis, impõe-se a remessa de cópias ao Ministério Público para apuração criminal, nos termos do art. 40 do Código de Processo Penal, dado que a conduta transcende o interesse privado das partes, atingindo a administração da justiça como bem jurídico tutelado.

Conclusão

A introdução da Inteligência Artificial no Poder Judiciário é um caminho sem volta, cujos benefícios para a eficiência da justiça são inegáveis, como reconhece e busca orientar a própria Resolução nº 615/2025 do CNJ. Contudo, a delegação acrítica de tarefas a sistemas automatizados, sem a devida supervisão humana e a implementação de robustas medidas de segurança, abre perigosas vulnerabilidades (Nunes, 2025).

A injeção de *prompts* maliciosos em peças processuais transcende a mera falha de segurança cibernética. Constitui, simultaneamente, grave violação da lealdade processual e conduta criminosa que atenta contra a administração da justiça. Ao buscar corromper o "assessor algorítmico" do julgador, o agente não apenas viola deveres processuais fundamentais, mas perpetra ilícito penal que compromete a higidez de todo o sistema jurisdicional.

O ordenamento jurídico pátrio demonstra-se já plenamente equipado para enfrentar essa ameaça emergente. No plano processual civil, as figuras da litigância de má-fé e do atentado à dignidade da justiça oferecem resposta adequada às violações dos deveres de lealdade e veracidade. Na esfera penal, a tipificação como invasão de dispositivo informático, estelionato processual ou falsidade ideológica assegura que a manipulação dolosa de sistemas judiciais

não permaneça impune. Essa dupla proteção reflete a gravidade multidimensional da conduta.

A resposta jurisdicional a essas práticas deve ser exemplar e pedagógica. A aplicação rigorosa das sanções processuais, conjugada com a persecução penal, quando cabível, enviará mensagem inequívoca: a modernização tecnológica do Judiciário não criará espaços de impunidade para novas modalidades de fraude. Pelo contrário, a mesma tecnologia que viabiliza a eficiência processual servirá como instrumento de detecção e repressão de condutas ilícitas.

Por isso, a lealdade processual, em tempos de IA, exige notória expansão de seu escopo tradicional. Ela demanda não apenas a veracidade dos fatos e a retidão dos argumentos, mas também a integridade/transparência dos algoritmos e o respeito às fronteiras entre inovação tecnológica e manipulação criminosa. A garantia de um processo justo na era digital depende da nossa capacidade de assegurar que a tecnologia sirva como instrumento para a verdade e a justiça; não como um Cavalo de Troia a serviço da deslealdade e da manipulação.

TITLE: The algorithmic Trojan horse: prompt injection as an affront to procedural loyalty

ABSTRACT: The inexorable symbiosis between generative artificial intelligence and forensic routine, while promising a revolution in judicial efficiency, inaugurates vulnerabilities of an unprecedented nature and singular gravity. This study scrutinizes the phenomenon of "prompt injection" – the malicious insertion of hidden instructions into Large Language Models (LLMs) – qualifying it not as a mere technical glitch, but as a true "algorithmic Trojan Horse" capable of subverting the integrity of jurisdictional provision. It is demonstrated that such artifice, by inducing the "algorithmic hypnosis" of decision-support systems, transcends technological cunning to constitute a frontal offense to the duties of procedural loyalty and objective good faith. The research maintains that the national legal system (Civil Procedure Code, Consolidation of Labor Laws, and Criminal Code) already possesses robust hermeneutic antidotes to repress this conduct, framing it within the figures of bad-faith litigation, acts attacking the dignity of justice, and, as a last resort, specific criminal types such as procedural fraud. It is concluded that the integrity of the process in the digital age demands constant epistemic vigilance, ensuring that technology remains a servant of truth, and never an instrument of surreptitious manipulation.

KEYWORDS: Generative Artificial Intelligence. Prompt injection. Procedural loyalty. Bad-faith litigation. Resolution of National Council of Justice No. 615/2025.

Referências

CONSELHO NACIONAL DE JUSTIÇA (CNJ). *Resolução nº 615, de 11 de março de 2025*. Estabelece diretrizes para o desenvolvimento, a governança e a utilização de soluções desenvolvidas com recursos de inteligência artificial no Poder Judiciário. Brasília, DF, 2025.

HUNG, Kuo-Han; KO, Ching-Yun; RAWAT, Ambrish; CHUNG, I-Hsin; HSU, Winston H.; CHEN, Pin-Yu. Attention tracker: detecting prompt injection attacks in LLMs. *arXiv preprint*, 23 abr. 2025. arXiv:2411.00348. Disponível em: http://arxiv.org/abs/2411.00348v2. Acesso em: 12 ago. 2025.

LIU, Ian Ch. 隱形提示注入 (invisible prompt injection). *Trend Micro*, 22 jan. 2025.

MAGESH, Varun; SURANI, Faiz; DAHL, Matthew; SUZGUN, Mirac; MANNING, Christopher D.; HO, Daniel E. Hallucination-free? Assessing the reliability of leading AI legal research tools. *arXiv preprint*, 30 maio 2024. arXiv:2405.20362. Disponível em: http://arxiv.org/abs/2405.20362v1. Acesso em: 12 ago. 2025.

MALOYAN, Narek; ASHINOV, Bislan; NAMIOT, Dmitry. Investigating the vulnerability of LLM-as-a-judge architectures to prompt-injection attacks. *arXiv preprint*, 19 maio 2025. arXiv:2505.13348. Disponível em: http://arxiv.org/abs/2505.13348v1. Acesso em: 12 ago. 2025.

MARANHÃO, Ney. *Neutralidade tecnológica e plataformas digitais de trabalho*: uma investigação filosófica. Coimbra: Editora Venturoli, 2025.

NUNES, Dierle. Decisões à[s] cegas: como as IAs podem ser manipuladas sem você saber. *Consultor Jurídico*, 18 jul. 2025. Disponível em: https://www.conjur.com.br/2025-jul-18/decisoes-a-cegas-como-as-ias-podem-ser-manipuladas-sem-voce-saber/. Acesso em: 12 ago. 2025.

NUNES, Dierle. Virada tecnológica no direito processual: fusão de conhecimento para geração de uma nova justiça centrada no ser humano. *Revista de Processo*, v. 344, p. 403–429, out. 2023.

OWASP. *LLM01:2025 Prompt injection* – OWASP Gen AI Security Project. 2025. Disponível em: https://genai.owasp.org/llmrisk/llm01-prompt-injection/. Acesso em: 12 ago. 2025.

SRNICEK, Nick. *Platform capitalism*. Cambridge: Polity Press, 2017.

YEUNG, Kenneth; RING, Leo. *Prompt injection attacks on LLMs*. HiddenLayer Innovation Hub, 27 mar. 2024.

Recebido em: 03.11.2025

Aprovado em: 17.11.2025