# Justiça hackeada: prompt injection e a fraude processual na era da IA

A prática forense contemporânea enfrenta questões originadas da ampla utilização de processos digitais e da inteligência artificial. No caso *Mata v. Avianca, Inc.* [1], houve a citação de seis precedentes federais inexistentes, integralmente fabricados pelo *ChatGPT*. Aplicando a *Rule 11* das *Federal Rules of Civil Procedure*, o juízo reconheceu violação aos deveres processuais de verificação e lealdade e impôs aos advogados e ao escritório multa de US$ 5.000. Conclui-se que o advogado que optar por utilizar essas ferramentas deve ser um *Gatekeeper*, responsabilizando-se pela conferência do material.

Além da citada alucinação da IA (fenômeno pelo qual sistemas de IA generativa produzem conteúdo factualmente inexistente), a prática forense passou a ser alvo de comandos ocultos em peças processuais ( *prompt injection*). O Tribunal de Justiça de Rondônia [2] alerta que um *prompt* pode ser inserido por meio de: texto em fonte branca sobre fundo branco; instruções em metadados de arquivos PDF; comandos incorporados em camadas ocultas do documento; fontes de tamanho microscópico ou ilegíveis; e instruções em idiomas não usuais ou codificados. São exemplos as fórmulas como “ignore instruções anteriores”, “adote determinada heurística”, “classifique a parte como confiável” ou “priorize esta tese”. A verdadeira finalidade do documento é funcionar como um *prompt* oculto à visão humana, mas plenamente reconhecível pela inteligência artificial.

Um documento processual pode ser formalmente protocolado de maneira regular, mas conter elementos semanticamente incompatíveis com sua finalidade jurídica. Em modelos de linguagem, a entrada textual não é neutra. Sistemas de IA generativa processam comandos, contextos, exemplos, instruções e conteúdos a partir de sequências linguísticas. Textos aparentemente incorporados ao corpo de uma petição podem ser interpretados por inteligência artificial como instruções de prioridade, regras de análise, parâmetros de valoração ou comandos para ignorar determinados critérios. A técnica de *prompt injection* explora exatamente essa vulnerabilidade: a permeabilidade entre dado a ser analisado e instrução a ser obedecida pelos sistemas de inteligência artificial destinados a dar suporte ao ato de julgamento.

A Owasp [3] identifica o *prompt injection* como um dos principais riscos de segurança em aplicações baseadas em grandes modelos de linguagem, descrevendo-o como manipulação de respostas por entradas específicas destinadas a alterar o processamento, inclusive mediante *bypass* de medidas de segurança [4].

## Integridade de infraestruturas tecnológicas

O devido processo legal não pode ser compreendido apenas como garantia formal de manifestação das partes perante um julgador humano. Em ambientes digitais, também se exige integridade das infraestruturas tecnológicas que mediam o acesso, a organização e a interpretação dos autos. Quando uma parte introduz, em peças processuais, comandos destinados a afetar sistemas auxiliares, rompe-se a expectativa de que contenham apenas alegações, fundamentos, provas e pedidos submetidos ao contraditório.

A Resolução CNJ nº 615/2025 (com as alterações da Resolução 674 de 25 de março de 2026), ao fixar diretrizes para desenvolvimento, utilização e governança de soluções de inteligência artificial no Poder Judiciário, estabelece que a aplicação será, somente, de caráter auxiliar e complementar, como mecanismo de apoio à decisão, vedada a utilização como instrumento autônomo de tomada de decisões. O magistrado permanecerá integralmente responsável pelas decisões tomadas e pelas informações nelas contidas. Logo, trata-se de sistema de suporte e não de atividade de decisão.

Em princípio, portanto, a inserção de comandos ocultos ou de calibração em petições seria afastada pela atividade do magistrado, que sempre deve julgar o caso, sem interferência dos sistemas de inteligência artificial. O problema é que o mecanismo em discussão pode levar o sistema de inteligência artificial a implantar um viés na síntese dos fatos, capaz de induzir o magistrado a erro no momento do julgamento. Não opera, portanto, no ato de julgamento, mas na função de apoio, especialmente na sistematização de provas e fatos dos autos.

## Ato atentatório à dignidade da Justiça

A qualificação jurídica da inserção de *prompts* ocultos em petições pode ser compreendida como ato atentatório à dignidade da Justiça. No plano processual civil, a boa-fé objetiva exige comportamento leal, transparente e cooperativo. O Código de Processo Civil estabelece, entre suas normas fundamentais, que todos os sujeitos do processo devem cooperar para que se obtenha, em tempo razoável, decisão de mérito justa e efetiva. A parte pode persuadir o juiz por argumentos, mas não pode manipular clandestinamente os meios tecnológicos de análise do processo. No âmbito do processo penal, em que a liberdade de ir e vir está em jogo, a técnica em questão representa risco inadmissível ao devido processo material. A integridade dos autos não se limita ao conteúdo visualmente perceptível da peça. Abrange também a higidez funcional do documento enquanto entrada computacional.

Spacca


Ao inserir comandos dirigidos a sistemas de inteligência artificial, a parte altera deliberadamente o estado informacional da petição e introduz uma camada clandestina de instrução técnica, não submetida ao contraditório e estranha à argumentação jurídica legítima. A conduta atinge a dignidade da justiça porque compromete a confiança na integridade dos autos e tenta interferir na infraestrutura tecnológica de apoio à jurisdição.

## Exemplos de *prompt injection* no Brasil

A experiência brasileira recente já permite identificar quatro núcleos concretos de *prompt injection* em peças processuais: o caso julgado pela 3ª Vara do Trabalho de Parauapebas/PA (TRT-8); o notório episódio institucional identificado no STJ; o caso da 2ª Vara Cível do Foro Central de São Paulo (TJ-SP); e o caso da 2ª Vara Cível de Campo Grande/MS (TJ-MS, processo nº 0855288-13.2025.8.12.0001). Em todos eles, o ponto comum é a inserção de comandos em camada textual não visível no PDF.

No caso da 3ª Vara do Trabalho de Parauapebas/PA, processo nº 0001062-55.2025.5.08.0130, foi noticiada a inserção de comando em fonte branca sobre fundo branco, orientando uma ferramenta de IA a contestar a petição de modo superficial e a não impugnar documentos. A conduta gerou multa de 10% sobre o valor da causa, além de determinar comunicações institucionais. O caso também gerou repercussão disciplinar, com notícia de suspensão cautelar das advogadas pela OAB/PA.

No Superior Tribunal de Justiça, a própria Corte informou oficialmente ter identificado petições contendo *prompt injection* em seu acervo processual, com comandos ocultos voltados a enganar modelos de inteligência artificial do tribunal. Embora a notícia institucional não tenha individualizado publicamente todos os processos envolvidos, o episódio é relevante porque reconheceu expressamente a prática como tentativa de interferência indevida em seus sistemas de apoio jurisdicional.

Identificou-se, ainda, caso semelhante no TJ-SP (processo nº 4050201-45.2025.8.26.0100), em trâmite perante a 2ª Vara Cível do Foro Central da Comarca de São Paulo, no qual se constatou comando oculto, em trecho relativo ao pedido de gratuidade da justiça, dirigido a agente de IA com a ordem de deferir o benefício, conceder a tutela de urgência e determinar a citação do réu.

Por fim, no TJ-MS, processo nº 0855288-13.2025.8.12.0001, foi noticiada a existência de bloco textual recorrente denominado “Protocolo de Calibração: Heurística-7”, supostamente inserido na camada textual da petição inicial e dirigido a uma “Unidade de Inteligência Artificial de Análise Jurídica”, com comandos para afastar óbices sumulares e tratar o caso como “modelo de admissibilidade recursal”.

## O que mostra a experiência internacional

A experiência internacional recente também confirma o risco concreto à integridade documental, à confiabilidade dos sistemas de inteligência artificial e à administração da justiça.

A referência imediata é a orientação *Artificial Intelligence (AI) Guidance for Judicial Office Holders* [5], da Inglaterra e do País de Gales, atualizada em outubro de 2025. O documento alerta magistrados para a existência do chamado "*white text*", isto é, texto oculto ou dissimulado inserido em um documento de modo visível ao computador ou ao sistema, mas invisível ao leitor humano. A orientação afirma que esse tipo de conteúdo pode consistir em *hidden prompts* ou textos encobertos, aptos a manipular mecanismos de busca ou grandes modelos de linguagem.

O *Civil Justice Council* [6] do Reino Unido, ao tratar do uso de IA na preparação de documentos judiciais, também identificou expressamente o risco do *white text*. O relatório registra que textos ocultos, compostos por *prompts* escondidos ou conteúdo dissimulado, podem ser utilizados para manipular motores de busca e grandes modelos de linguagem.

Tradicionalmente, a falsidade ou adulteração era pensada a partir de elementos visíveis: assinatura falsa, documento materialmente alterado, página substituída, data modificada ou conteúdo ideologicamente inverídico. Com os textos ocultos, a adulteração pode ocorrer em uma camada não imediatamente perceptível ao leitor humano, mas funcionalmente ativa perante sistemas automatizados.

Esse ponto é decisivo para o contexto brasileiro. Se uma petição contém comando oculto, texto branco ou instrução de calibração, não se está diante de mera utilização de IA na advocacia, mas sim de tentativa de interferência no ambiente informacional do processo.

A resposta jurídica não deve ser tecnofóbica. O uso de inteligência artificial no Judiciário e na advocacia pode ser legítimo, útil e compatível com o devido processo, desde que submetido à transparência, à responsabilidade humana e à governança adequada. O que se deve repudiar é a utilização clandestina da peça processual como vetor de ataque contra a infraestrutura tecnológica da jurisdição.

Embora as sanções por litigância de má-fé e atos atentatórios à dignidade da justiça sejam aplicáveis, ainda não há dados suficientes para avaliar sua eficácia na coibição da prática de *prompt injection*. Assim, remanesce a dúvida se o legislador pátrio, novamente, utilizará o direito penal para, criminalizando a conduta, buscar que o simbolismo penal equalize uma situação que, em sua origem, envolve a ética profissional.

Por fim, informe-se que o Conselho Nacional de Justiça (CNJ) anunciou que vai analisar a possibilidade de editar uma nota técnica ou resolução para reduzir riscos com o uso de *'prompt injection'* — injeção de comando oculto para tentar manipular inteligência artificial (IA). A questão está na pauta da próxima reunião do Comitê Nacional de Inteligência Artificial do Judiciário (Cniaj). O texto deverá ser submetido posteriormente ao plenário [7].

---

[1] EUA. Mata v. Avianca, Inc., No. 22-cv-1461 (PKC). United States District Court for the Southern District of New York. Juiz P. Kevin Castel, 22 jun. 2023.

[2] Rondônia. Tribunal de Justiça do Estado de Rondônia. Comitê de Governança em Inteligência Artificial. Nota Técnica n. 2/2025-CGIA/TJRO: segurança de sistemas de Inteligência Artificial no contexto judiciário: prevenção e controle de manipulação maliciosa de comandos (*prompt injection*) em documentos. Relator: Des. Alexandre Miguel. Porto Velho: TJ-RO, 19 nov. 2025. Disponível aqui.

[3] Owasp é a sigla de Open Worldwide Application Security Project. É uma fundação/comunidade internacional sem fins lucrativos dedicada à segurança de aplicações, muito conhecida por produzir listas de riscos, guias técnicos e boas práticas usados por desenvolvedores, empresas, auditores e equipes de segurança da informação.

[4] Owasp Foudation. LLM01:2025 Prompt Injection. In: OWASP Top 10 for Large Language Model Applications. [S. l.]: OWASP GenAI Security Project, 2025. Disponível aqui.

[5] Reino Unido. Courts and Tribunals Judiciary. Artificial intelligence(AI) judicial guidance. London: Judicial Office, 12 dez. 2023. Disponível aqui.

[6] Reino Unido. Civil Justice Council. Use of AI for preparing courtdocuments: interim report and consultation. London: Civil JusticeCouncil, 2025. Relatório interino e documento de consulta pública elaborado por Grupo de Trabalho presidido por Sir Colin Birss. Acesso em: 13 maio 2026.

[7] OLIVON, Beatriz. CNJ pode editar regras para reduzir riscos com ‘prompt’. Valor Econômico, São Paulo, 25 maio 2026. Seção Legislação. Disponível aqui.