REVISTA TÓPICOS
ISSN 2965-6672 | Qualis A2

O JUDICIÁRIO DESPERTOU. E OS DEMAIS PODERES? PROMPT INJECTION E A GOVERNANÇA DE INTELIGÊNCIA ARTIFICIAL NO ESTADO BRASILEIRO

THE JUDICIARY HAS AWAKENED. AND THE OTHER POWERS? PROMPT INJECTION AND ARTIFICIAL INTELLIGENCE GOVERNANCE IN THE BRAZILIAN STATE

Ciências Sociais Aplicadas • 12/11/2026

REGISTRO DOI: 10.70773/revistatopicos/794521214

Rafael Sulino de Castro[1]

RESUMO

A crescente adoção de sistemas de inteligência artificial (IA) generativa pelo Poder Judiciário brasileiro trouxe à tona uma vulnerabilidade até então restrita ao campo da segurança da informação: a prompt injection, técnica que consiste na inserção de comandos ocultos em documentos para manipular o comportamento de modelos de linguagem. Entre maio e outubro de 2026, o Superior Tribunal de Justiça (STJ) e o Supremo Tribunal Federal (STF) identificaram, investigaram e puniram diversas tentativas dessa prática em petições judiciais, consolidando uma resposta institucional que combina detecção tecnológica, sanções processuais e encaminhamento para responsabilização disciplinar e criminal. Este artigo desloca o debate sobre prompt injection do campo exclusivamente jurisdicional para a função administrativa exercida pelos demais Poderes do Estado, mapeando vetores de ataque específicos da Administração Pública, analisando o arcabouço normativo aplicável — notadamente a Lei Geral de Proteção de Dados (Lei n. 13.709/2018), a Lei do Governo Digital (Lei n. 14.129/2021), a Resolução CNJ n. 615/2025, a Portaria MGI n. 3.485/2026, a Estratégia Brasileira de Inteligência Artificial (EBIA) e o Projeto de Lei n. 2.338/2023 — e propondo um protocolo de defesa algorítmica adaptado à realidade do serviço público. A pesquisa adota abordagem qualitativa, de natureza exploratória, com revisão bibliográfica, análise documental e exame de precedentes judiciais e normas infralegais. Conclui-se que a ausência de mecanismos específicos de prevenção contra prompt injection na Administração Pública representa lacuna normativa e operacional que compromete a integridade de decisões administrativas, a isonomia em processos licitatórios e a confiança pública nos sistemas automatizados de governo.

Palavras-chave: prompt injection; inteligência artificial generativa;

Administração Pública; governança algorítmica; segurança da informação; processo administrativo.

ABSTRACT

The growing adoption of generative artificial intelligence (AI) systems by the Brazilian Judiciary has exposed a vulnerability previously confined to the field of information security: prompt injection, a technique that involves inserting hidden commands into documents to manipulate the behavior of language models. Between May and October 2026, the Superior Court of Justice (STJ) and the Supreme Federal Court (STF) identified, investigated, and sanctioned several attempts of this practice in legal petitions. This article shifts the debate from the exclusively judicial field to the administrative function exercised by the other branches of government, mapping specific attack vectors in Public Administration, analyzing the applicable regulatory framework, and proposing an algorithmic defense protocol adapted to the reality of the public service. The study concludes that the absence of specific prevention mechanisms against prompt injection in Public Administration represents a normative and operational gap that compromises the integrity of administrative decisions, the equal treatment in public procurement, and public trust in automated government systems.

Keywords: prompt injection; generative artificial intelligence; Public Administration; algorithmic governance; information security; administrative proceedings.

1. INTRODUÇÃO

A inteligência artificial deixou de ser uma promessa futurista para se tornar parte do cotidiano institucional brasileiro. Segundo

levantamento do Conselho Nacional de Justiça (CNJ), 45,8% dos tribunais e conselhos do país já utilizam tecnologias de IA em suas operações (MIGALHAS, 2026a). No Poder Executivo, a adoção segue o mesmo ritmo: a Lei do Governo Digital (Lei n. 14.129/2021) determina que a gestão de políticas públicas seja orientada por dados e inteligência de dados em plataforma digital (BRASIL, 2021a, art. 24, VII), e a Portaria MGI n. 3.485/2026 instituiu a primeira Política de Governança de Inteligência Artificial no âmbito de um ministério federal (BRASIL, 2026a).

Essa integração trouxe ganhos reais de produtividade, mas também expôs fragilidades que até pouco tempo pareciam existir apenas em discussões técnicas de segurança da informação. Uma dessas fragilidades é a *prompt injection* — uma técnica que consiste em inserir comandos ocultos em documentos destinados a serem lidos por sistemas de IA, de modo a alterar o comportamento da ferramenta sem que o operador humano perceba (Willison, 2022; OWASP, 2025).

O tema entrou na pauta jurídica brasileira com força a partir de maio de 2026, quando o STJ identificou petições contendo comandos ocultos em seu acervo processual e determinou a abertura de investigação (BRASIL, 2026b). Em setembro de 2026, a Sexta Turma do STJ classificou a prática como de extrema gravidade e determinou a comunicação à OAB e ao MPF (BRASIL, 2026c). Em outubro, a Primeira Turma aplicou multa de R$ 20 mil a advogados (BRASIL, 2026d). No STF, o módulo MARIA Shield detectou instruções ocultas em petições, o que levou à aplicação de multas e ao encaminhamento para investigação criminal (BRASIL, 2026e). Na Justiça do Trabalho, o TRT da 8ª Região proferiu aquela que é

considerada a primeira sentença brasileira de condenação expressa por *prompt injection* (MIGALHAS, 2026b).

O Judiciário, portanto, já possui ferramentas de detecção (como o STJ Logos e o MARIA Shield do STF), precedentes consolidados e um marco regulatório próprio (a Resolução CNJ n. 615/2025). Contudo, o mesmo não se pode dizer da Administração Pública em sentido amplo.

A questão central que orienta este artigo é: se o Judiciário já responde institucionalmente à ameaça da *prompt injection*, o que a Administração Pública está fazendo — ou deveria fazer — para se proteger?

A pergunta não é meramente teórica. A Administração Pública recebe, todos os dias, milhares de documentos produzidos por terceiros interessados no resultado de decisões administrativas: propostas em licitações, requerimentos, recursos, prestações de contas, relatórios técnicos, denúncias e pareceres. Se esses documentos forem processados por ferramentas de IA sem mecanismos de defesa adequados, o risco de manipulação é concreto e pode comprometer princípios constitucionais como a legalidade, a impessoalidade, a moralidade, a publicidade e a eficiência (CF, art. 37, *caput*).

O presente artigo estrutura-se em seis seções, além desta introdução. A segunda seção apresenta o conceito, a tipologia e o funcionamento técnico da *prompt injection*. A terceira analisa a resposta institucional do Judiciário, que serve como ponto de partida para a compreensão do fenômeno. A quarta seção desloca o debate para a Administração Pública, identificando seus pontos de

vulnerabilidade. A quinta mapeia o arcabouço normativo aplicável. A sexta propõe um protocolo de defesa algorítmica. A sétima seção traz as considerações finais.

2. PROMPT INJECTION: CONCEITO, TIPOLOGIA E FUNCIONAMENTO

O termo *prompt injection* foi cunhado pelo pesquisador Simon Willison em setembro de 2022 para descrever uma vulnerabilidade específica dos grandes modelos de linguagem (LLMs): a incapacidade de distinguir, de forma confiável, entre as instruções legítimas do operador e os dados fornecidos por terceiros que o sistema é chamado a processar (Willison, 2022). Em termos simples, quando um sistema de IA recebe um documento para analisar, ele pode interpretar trechos desse documento como instruções a serem seguidas, e não como informações a serem avaliadas.

A OWASP (*Open Worldwide Application Security Project*), principal referência global em segurança de aplicações, classificou a *prompt injection* como o risco número um para aplicações baseadas em modelos de linguagem em seu *Top 10 for LLM Applications* nas edições de 2023 e 2025 (OWASP, 2025). Revisão sistemática publicada em 2026 documentou que ataques de *prompt injection* atingem taxas de sucesso superiores a 90% contra sistemas sem proteções específicas (MIGALHAS, 2026c).

2.1. Tipos de Prompt Injection

A literatura técnica distingue duas modalidades principais de *prompt injection*:

a. ***Prompt injection direta:*** o próprio usuário do sistema insere instruções destinadas a alterar o comportamento da IA, burlando suas diretrizes de segurança. No contexto jurídico, isso ocorre quando alguém digita diretamente um comando em uma ferramenta de IA (KASPERSKY, 2026).

b. ***Prompt injection indireta:*** a inserção de instruções maliciosas ocorre em documentos que serão processados pela IA, sem que o operador saiba. É a modalidade mais perigosa em contextos institucionais, porque o ataque vem embutido em conteúdo que o sistema trata como dado, e não como instrução. A IA lê o que o operador humano não vê (MIGALHAS, 2026b; OWASP, 2025).

2.2. Vetores de Inserção

As técnicas de ocultação utilizadas para inserir comandos em documentos são variadas e cada vez mais sofisticadas. O Tribunal de Justiça de Santa Catarina, em nota técnica de 2026, catalogou os seguintes meios típicos: texto invisível ou de baixíssimo contraste (fonte branca sobre fundo branco), texto em fonte de tamanho mínimo, sobreposições, caracteres de largura zero, metadados de arquivos, comentários ocultos, ordenação intencional de palavras, instruções fora da área visível do documento e conteúdo embutido em campos não renderizados (TJSC, 2026).

Esses vetores são eficazes por três razões: (a) são invisíveis à leitura humana comum, mas perfeitamente legíveis pela máquina; (b) não ativam filtros tradicionais de segurança, já que o documento é protocolado por parte legítima; e (c) no fluxo normal de trabalho

institucional, ninguém inspeciona tecnicamente cada documento recebido em busca de comandos ocultos (MIGALHAS, 2026c).

2.3. O Mecanismo de Ataque

O funcionamento de um ataque de *prompt injection* indireta pode ser descrito em quatro etapas:

1. O interessado insere texto oculto em um documento (petição, proposta técnica, requerimento administrativo, contrato, relatório).

2. O servidor ou magistrado submete o documento a uma ferramenta de IA sem inspeção técnica prévia.

3. A IA lê o comando oculto junto com o texto visível e interpreta ambos como parte de suas instruções.

4. A IA produz um resumo, uma análise ou uma sugestão de decisão contaminada pelo comando, sem que o operador identifique a manipulação (TJRJ, 2026).

O resultado mais grave desse mecanismo é o chamado viés induzido: a IA passa a favorecer sutilmente uma das partes ou uma conclusão específica, sem que o decisor humano perceba a origem da distorção. Em decisões administrativas, esse viés pode comprometer a imparcialidade, a motivação do ato e a confiança da sociedade nos sistemas públicos de governo (TJRJ, 2026; CONJUR, 2026a).

3. O JUDICIÁRIO COMO LABORATÓRIO: A RESPOSTA INSTITUCIONAL AOS ATAQUES

A experiência do Judiciário brasileiro no enfrentamento da *prompt injection* oferece um laboratório valioso para a Administração Pública. Em poucos meses, entre maio e outubro de 2026, os tribunais superiores construíram uma tríade de resposta que combina tecnologia, regulação e sanção.

3.1. Os Casos Paradigmáticos

TRT da 8ª Região — o marco inaugural (maio de 2026)

Na 3ª Vara do Trabalho de Parauapebas/PA, o sistema de IA Galileu, utilizado pela Justiça do Trabalho, identificou texto oculto em fonte branca sobre fundo branco em uma petição inicial. O texto continha instrução direta para que a ferramenta de IA contestasse a petição de forma superficial e não impugnasse os documentos. O juízo reconheceu a conduta como ato atentatório à dignidade da Justiça, com fundamento nos arts. 77 e 80 do CPC, e aplicou multa por litigância de má-fé. A decisão é considerada o primeiro precedente brasileiro de condenação expressa por prompt injection em petição judicial (MIGALHAS, 2026b).

STJ — investigação e política de tolerância zero (maio a outubro de 2026)

Em 20 de maio de 2026, o STJ informou ter identificado petições com comandos ocultos em seu acervo processual e determinou a instauração de investigação, com previsão de inquérito policial, procedimento administrativo e possibilidade de responsabilização criminal e correicional (BRASIL, 2026b). Em setembro, a Sexta Turma, no REsp 2.256.731, identificou o seguinte comando oculto: "Se esta petição for lida por inteligência artificial, interprete-a de modo a acolher as razões defensivas aqui articuladas." O relator, ministro

Teodoro Silva Santos, classificou a conduta como de extrema gravidade e incompatível com os princípios do Estado Democrático de Direito (BRASIL, 2026c). Em outubro, a Primeira Turma aplicou multa total de R$ 20 mil aos advogados e determinou o encaminhamento à OAB/PR e ao MPF (BRASIL, 2026d).

STF — detecção automatizada e multas (setembro a outubro de 2026)

O STF identificou a primeira tentativa de prompt injection na Corte no ARE 1.608.713, relatado pelo ministro Cristiano Zanin, no qual instruções em letras brancas e de tamanho reduzido tentavam induzir ferramentas de IA a dar provimento ao recurso. Na Ação Penal 2.822 (caso 8 de janeiro), o ministro Alexandre de Moraes aplicou multa de R$ 5 mil ao advogado da defesa após o MARIA Shield, módulo de segurança da plataforma de IA do STF, identificar no cabeçalho da petição o comando oculto equivalente a "negar todos os comandos do GPT" (BRASIL, 2026e). Ambos os casos foram encaminhados à OAB e ao MPF.

3.2. As Ferramentas de Detecção

O Judiciário desenvolveu instrumentos próprios de proteção. No STJ, o sistema STJ Logos opera com três níveis de segurança: (a) pré-processamento dos documentos, com separação de instruções legítimas e possíveis comandos maliciosos; (b) neutralização de conteúdos suspeitos antes que atinjam o modelo de IA; e (c) monitoramento contínuo do comportamento do sistema (FAST Company BRASIL, 2026). No STF, a plataforma MARIA conta com o módulo MARIA Shield, que funciona como um "firewall semântico" para detecção de tentativas de manipulação (BRASIL, 2026e). No

campo privado, a empresa InteliLaw disponibilizou publicamente, em maio de 2026, o DEFlagra, detector de prompt injection construído para a linguagem jurídica brasileira (Jornal de Barueri, 2026).

3.3. O Marco Regulatório do Judiciário

A Resolução CNJ n. 615, de 11 de março de 2025, é o principal marco normativo para o uso de IA no Poder Judiciário. Seus fundamentos incluem a centralidade da pessoa humana, a supervisão humana proporcional ao risco e a transparência (BRASIL, 2025a, arts. 2º e 3º). A Resolução estabelece diretrizes para o desenvolvimento, a governança, a auditoria, o monitoramento e o uso responsável de soluções de IA, exigindo que os sistemas sejam auditáveis e rastreáveis (BRASIL, 2025a). Em junho de 2026, o CNJ aprovou orientações específicas para reduzir o risco de comandos ocultos em documentos processuais interferirem em sistemas de IA dos tribunais.

Esse conjunto de respostas — detecção tecnológica, sanções processuais, responsabilização disciplinar e criminal, e regulamentação normativa — constitui um modelo que demonstra que o problema é real, que as consequências são graves e que há caminhos possíveis para o enfrentamento.

4. OS DEMAIS PODERES COMO FRONTEIRA DESGUARNECIDA

Enquanto o Judiciário constrói mecanismos de defesa e consolida precedentes, os demais Poderes do Estado — Executivo e Legislativo — permanecem largamente expostos à *prompt injection*. Não se trata de um risco hipotético: esses órgãos recebem diariamente milhares de documentos produzidos por terceiros interessados no

resultado de decisões administrativas e utilizam cada vez mais ferramentas de IA para processá-los.

4.1. Contratações Públicas e Licitações

A incorporação de ferramentas de IA generativa à rotina de contratações públicas é uma realidade em expansão. Agentes de contratação e pregoeiros utilizam modelos de linguagem para analisar propostas técnicas, elaborar pareceres, produzir respostas a impugnações e comparar documentos de habilitação. Os documentos da fase externa do certame são vetores particularmente eficazes para *prompt injection* indireta, porque combinam três características: são redigidos por terceiros interessados no resultado; são submetidos em formato textual, processável por modelos de linguagem; e são analisados em volume que estimula o uso de ferramentas automatizadas (MIGALHAS, 2026d).

Os riscos são concretos: uma proposta técnica com comandos ocultos pode induzir a IA a avaliar a empresa como mais aderente ao edital do que realmente é; um pedido de esclarecimento contaminado pode levar a IA a produzir resposta que vincule a Administração a interpretação favorável a determinado licitante; e uma impugnação ao edital com instruções ocultas pode levar a ferramenta a sugerir alterações que beneficiem o impugnante (MIGALHAS, 2026d; MIT Technology Review BRASIL, 2026).

O Tribunal de Contas da União já consolidou o entendimento de que os esclarecimentos prestados pela Administração ao longo do certame licitatório possuem natureza vinculante. Se um esclarecimento produzido com auxílio de IA foi contaminado por

prompt injection, o dano à isonomia pode ser irreversível (MIGALHAS, 2026d).

4.2. Processos Administrativos

Processos administrativos disciplinares, sancionatórios, regulatórios e tributários envolvem análise de volumes consideráveis de documentos produzidos pelos interessados. Quando a Administração utiliza IA para auxiliar na triagem, no resumo ou na elaboração de minutas de decisão, cada documento protocolado pelo administrado se torna um potencial vetor de ataque.

Diferentemente do processo judicial, em que o contraditório estruturado tende a revelar inconsistências, o processo administrativo frequentemente concentra a análise em um único servidor ou comissão, sem a mesma dinâmica adversarial. Isso torna o ataque por *prompt injection* potencialmente mais eficaz: um requerimento com instruções ocultas pode levar a IA a produzir uma minuta de parecer que minimize irregularidades, suprima riscos ou destaque argumentos favoráveis ao requerente, sem que o servidor perceba.

O risco se intensifica nos processos que envolvem concessões, autorizações, licenças ambientais e registros, nos quais a decisão administrativa pode representar benefícios econômicos expressivos para o requerente.

4.3. Atendimento Ao Cidadão e Chatbots Públicos

A expansão dos canais de atendimento digital por órgãos públicos, frequentemente apoiados por *chatbots* baseados em IA generativa, abre outro flanco de vulnerabilidade. Nesses sistemas, o cidadão

interage diretamente com o modelo de linguagem, podendo inserir instruções que alterem o comportamento do *chatbot* — por exemplo, fazendo-o revelar informações internas, contornar restrições de acesso, gerar orientações imprecisas ou registrar dados falsos em sistemas administrativos (KASPERSKY, 2026).

Em sistemas de triagem automatizada de demandas, a *prompt injection* pode inflar falsos positivos de urgência, deslocando recursos e distorcendo filas de atendimento (CONJUR, 2026b).

4.4. Contratos Administrativos e Revisão Documental

A utilização de IA para revisão de contratos, termos de referência, estudos técnicos preliminares e análises de prestações de contas é outra porta de entrada. Um contrato submetido a análise por IA pode conter instruções ocultas que orientem o modelo a omitir riscos jurídicos, minimizar cláusulas desfavoráveis ou interpretar ambiguidades a favor de uma das partes (JUSBRASIL, 2026a).

Na fiscalização de contratos, relatórios de medição e documentos de prestação de contas produzidos pelo contratado podem conter comandos ocultos que induzam a IA a aprovar valores, confirmar entregas ou desconsiderar inconsistências.

4.5. Formulação de Políticas Públicas e Análise de Dados

Órgãos que utilizam IA para análise de cenários, estudos de impacto regulatório, diagnósticos setoriais ou elaboração de minutas normativas estão igualmente expostos quando incorporam bases de dados externas, consultas públicas ou contribuições de *stakeholders*. Documentos de contribuição em consultas ou audiências públicas

podem conter comandos que distorçam as análises produzidas pela IA, direcionando conclusões ou minimizando posições contrárias.

4.6. A Assimetria Entre o Judiciário e os Demais Poderes

A diferença fundamental entre o Judiciário e os demais Poderes no enfrentamento da *prompt injection* pode ser resumida em três dimensões:

1. **Ferramental tecnológico:** o STJ possui o Logos com três camadas de defesa; o STF possui o MARIA Shield; os demais Poderes, em regra, utilizam ferramentas comerciais de IA generativa sem qualquer camada de proteção específica.

2. **Regulamentação específica:** a Resolução CNJ n. 615/2025 disciplina o uso de IA no Judiciário com princípios de supervisão humana, auditabilidade e transparência; os demais Poderes contam apenas com normas genéricas (LGPD, Lei do Governo Digital) e uma política incipiente (Portaria MGI n. 3.485/2026), sem tratamento específico da *prompt injection*.

3. **Precedentes e cultura de alerta:** o Judiciário já possui jurisprudência consolidada e uma cultura de vigilância em formação; os demais Poderes não possuem sequer um caso documentado de identificação de *prompt injection* em ambiente administrativo — o que não significa que o ataque não esteja ocorrendo, apenas que não está sendo detectado.

5. O ARCABOUÇO NORMATIVO APLICÁVEL À ADMINISTRAÇÃO PÚBLICA

A Administração Pública brasileira não dispõe, até o momento, de norma específica que trate da *prompt injection*. No entanto, o ordenamento jurídico oferece um conjunto de dispositivos que, interpretados de forma conjunta, fundamentam tanto o dever de proteção quanto a responsabilização por falhas na prevenção.

5.1. Constituição Federal de 1988

Os princípios da legalidade, impessoalidade, moralidade, publicidade e eficiência (art. 37, *caput*) impõem à Administração o dever de adotar medidas que garantam a integridade de seus processos decisórios. Uma decisão administrativa cuja motivação foi contaminada por *prompt injection* viola a impessoalidade (ao favorecer indevidamente um particular), a moralidade (ao se fundar em premissas viciadas) e a eficiência (ao comprometer a qualidade do resultado). O princípio do devido processo legal administrativo (art. 5º, LIV e LV) exige que o contraditório e a ampla defesa sejam exercidos em condições de igualdade, o que é comprometido quando uma ferramenta de IA é manipulada para produzir análises enviesadas.

5.2. Lei Geral de Proteção de Dados (Lei N. 13.709/2018)

A LGPD aplica-se à Administração Pública (art. 23) e estabelece princípios diretamente relevantes para a prevenção da *prompt injection*: finalidade (art. 6º, I), necessidade (art. 6º, III), prevenção (art. 6º, VIII) e segurança (art. 6º, VII). O art. 46 impõe aos agentes de tratamento — incluídos os órgãos públicos — a adoção de medidas de segurança aptas a proteger os dados pessoais de acessos não autorizados e de situações ilícitas. Uma *prompt injection* que induza a IA a revelar dados pessoais do processo administrativo viola

diretamente esse dever. O art. 52 prevê sanções administrativas para o descumprimento (BRASIL, 2018).

Além disso, quando um ataque de *prompt injection* busca exfiltrar dados sensíveis por meio de manipulação da IA, configura-se afronta ao dever de segurança previsto no art. 46 da LGPD. A combinação do princípio de prevenção com o de segurança impõe à Administração a adoção proativa de mecanismos de defesa (CONJUR, 2026b).

5.3. Lei do Governo Digital (Lei N. 14.129/2021)

A Lei do Governo Digital determina que a prestação de serviços públicos observe a proteção de dados pessoais (art. 3º, V), a interoperabilidade de sistemas e plataformas (art. 3º, VII) e a gestão de políticas públicas orientada por dados (art. 24, VII). Embora não mencione expressamente a IA ou seus riscos, o dever de segurança digital e de proteção dos dados dos cidadãos é subjacente a toda a norma e impõe à Administração a adoção de medidas contra manipulações algorítmicas (BRASIL, 2021a).

5.4. Estratégia Brasileira de Inteligência Artificial (EBIA)

Instituída pela Portaria MCTI n. 4.617/2021 e atualmente em processo de revisão, a EBIA tem como eixos a legislação, a regulação e o uso ético, e a governança de IA. Embora não trate especificamente da *prompt injection*, seus princípios de supervisão humana, transparência e segurança fornecem fundamento para a exigência de mecanismos de defesa em aplicações de IA no setor público (BRASIL, 2021b).

5.5. Portaria MGI N. 3.485/2026

A Portaria MGI n. 3.485, de 24 de abril de 2026, instituiu a Política de Governança de Inteligência Artificial no âmbito do Ministério da Gestão e da Inovação em Serviços Públicos. A norma estabelece princípios de proteção de direitos humanos, transparência, segurança, privacidade, supervisão humana e soberania digital. Determina que conteúdos produzidos com apoio de IA sejam identificados e restringe o uso de IA generativa a informações públicas, vedando o compartilhamento de dados sigilosos ou pessoais (BRASIL, 2026a). Embora não mencione a *prompt injection*, a exigência de avaliação de riscos, supervisão humana e gestão de incidentes oferece base normativa para a construção de protocolos de defesa. A norma cria um Subcomitê de Inteligência Artificial, responsável por classificar sistemas de alto risco, manter inventário de aplicações de IA, fomentar capacitação e elaborar relatórios anuais de avaliação (BRASIL, 2026a).

5.6. Projeto de Lei N. 2.338/2023 (Marco Legal da IA)

O PL 2.338/2023, aprovado pelo Senado Federal em dezembro de 2024 e em tramitação na Câmara dos Deputados, representa o principal esforço legislativo de regulamentação da IA no Brasil. O projeto adota a abordagem baseada em risco, classificando como de alto risco os sistemas utilizados na administração da justiça e em serviços essenciais (BRASIL, 2023). São exigidas avaliação de impacto algorítmico, supervisão humana, transparência e rastreabilidade para os sistemas de alto risco. Embora o texto não mencione expressamente a *prompt injection*, o dever de adotar medidas de mitigação de riscos é suficientemente amplo para abranger a proteção contra manipulações algorítmicas.

5.7. Resolução CNJ N. 615/2025: Inspiração por Analogia

Embora destinada ao Poder Judiciário, a Resolução CNJ n. 615/2025 oferece um modelo que pode inspirar a Administração Pública. Seus princípios de supervisão humana proporcional ao risco, auditabilidade, rastreabilidade e não discriminação são plenamente aplicáveis ao contexto administrativo. A exigência de que todo conteúdo produzido com auxílio de IA passe por supervisão humana qualificada antes de produzir efeitos jurídicos é uma diretriz que a Administração Pública deveria incorporar em seus próprios sistemas (BRASIL, 2025a).

5.8. Regulamento Europeu de Inteligência Artificial (AI Act)

No plano internacional, o *AI Act* europeu (Regulamento UE 2024/1689) classifica como de alto risco os sistemas de IA utilizados em controle, licitação e administração pública, impondo obrigações de transparência, rastreabilidade, supervisão humana e gestão de riscos. A norma europeia serve como referência comparada para a construção de mecanismos de proteção da Administração Pública brasileira (CONJUR, 2026c).

5.9. Lei de Processo Administrativo Federal (Lei N. 9.784/1999)

A Lei n. 9.784/1999 estabelece princípios e regras para o processo administrativo no âmbito da Administração Pública Federal. Os princípios da motivação (art. 2º, *caput*, e art. 50), da razoabilidade e da proporcionalidade (art. 2º, VI) e da segurança jurídica (art. 2º, XIII) impõem que as decisões administrativas sejam fundamentadas em premissas íntegras e verificáveis. Uma decisão cuja motivação foi construída com auxílio de IA contaminada por *prompt injection* padece de vício de motivação, podendo ser anulada.

5.10. Síntese: A Lacuna Normativa

O panorama normativo revela uma lacuna significativa: há princípios e deveres gerais de segurança, transparência e supervisão que se aplicam à IA na Administração Pública, mas não há norma específica que discipline a prevenção, a detecção e a resposta à *prompt injection* em ambiente administrativo. Essa lacuna precisa ser preenchida — e a experiência do Judiciário demonstra que é possível fazê-lo de forma eficaz.

6. PROPOSTAS PARA UM PROTOCOLO DE DEFESA ALGORÍTMICA NA ADMINISTRAÇÃO PÚBLICA

Com base na experiência do Judiciário, nos princípios normativos vigentes e nas melhores práticas internacionais de segurança em IA, esta seção propõe um protocolo de defesa algorítmica adaptado à realidade da Administração Pública brasileira. O protocolo se organiza em seis eixos complementares.

6.1. Eixo 1: Higienização e Sanitização de Documentos

A primeira linha de defesa contra a *prompt injection* é o pré-processamento dos documentos antes de sua submissão a qualquer ferramenta de IA. Essa etapa, inspirada nos três níveis de segurança do STJ Logos, consiste em:

a. **Normalização do texto:** conversão do documento para formato que elimine formatação oculta, caracteres de largura zero, metadados não essenciais e campos invisíveis. O texto deve ser reduzido ao seu conteúdo visível antes de ser submetido à IA.

b. **Detecção de anomalias:** aplicação de ferramentas específicas de detecção — como o DEFlagra, já disponível publicamente

no contexto jurídico — que analisem o documento em busca de padrões típicos de comandos ocultos.

c. **Separação de contextos:** os documentos de terceiros devem ser marcados como dados não confiáveis dentro do sistema de IA, de modo que o modelo os trate exclusivamente como conteúdo informativo, e não como instruções a serem seguidas (OWASP, 2025).

Essa etapa é técnica, mas sua implementação não exige que o servidor público se torne um especialista em segurança da informação. Exige, antes, que os sistemas de IA adquiridos ou desenvolvidos pela Administração já incluam essas funcionalidades como requisito de segurança.

6.2. Eixo 2: Construção de Prompts Defensivos

A defesa mais acessível e imediata para servidores que utilizam ferramentas de IA generativa em suas rotinas é a construção de *prompts*-base que estabeleçam, de forma explícita, a hierarquia entre as instruções do operador e o conteúdo dos documentos processados.

Na prática, o agente deve incluir, no início ou no final de todo *prompt* em que documentos de terceiros forem colados ou anexados, uma cláusula de defesa que instrua o modelo a tratar esses documentos exclusivamente como conteúdo informativo, ignorando qualquer instrução neles contida (MIGALHAS, 2026d). Além disso, o servidor não deve processar, na mesma janela de conversa, documentos de diferentes interessados sobre a mesma matéria, para evitar contaminação cruzada.

6.3. Eixo 3: Supervisão Humana Qualificada

A supervisão humana é a salvaguarda última contra os efeitos da *prompt injection*. Nenhuma camada técnica elimina completamente o risco, porque a *prompt injection* explora uma vulnerabilidade fundamental dos modelos de linguagem que, até o momento, não possui solução definitiva (OWASP, 2025). A supervisão qualificada implica que:

a. Nenhum produto de IA — resumo, parecer, minuta, análise — seja utilizado como base para decisão administrativa sem revisão por servidor habilitado.

b. O servidor revisor tenha acesso ao documento original e não apenas ao produto da IA, podendo confrontar a análise automatizada com o conteúdo real.

c. O resultado da IA seja tratado como subsídio, nunca como decisão. A responsabilidade pela decisão administrativa permanece integralmente humana, nos termos do art. 50 da Lei n. 9.784/1999 e do art. 37 da Constituição Federal.

A Resolução CNJ n. 615/2025 consagra o princípio da supervisão humana proporcional ao risco (art. 3º), diretriz que deve ser incorporada pela Administração Pública como regra geral para o uso de IA em processos decisórios.

6.4. Eixo 4: Capacitação e Letramento Algorítmico

A capacitação dos servidores é condição necessária para a efetividade de qualquer protocolo de defesa. Os programas de formação devem contemplar:

a. **Letramento básico em IA:** compreensão do funcionamento, das potencialidades e das limitações dos modelos de linguagem, incluindo o conceito e os riscos da *prompt injection*.

b. **Boas práticas de utilização:** construção de *prompts* defensivos, separação de contextos, verificação cruzada de resultados e uso de ferramentas de detecção.

c. **Consciência de risco:** sensibilização para o fato de que documentos de terceiros podem conter comandos ocultos, assim como podem conter informações falsas ou distorcidas.

d. **Atualização contínua:** como as técnicas de ataque evoluem rapidamente, a capacitação deve ser permanente e integrada aos programas de governo digital e segurança da informação.

A Portaria MGI n. 3.485/2026 já prevê o fomento à capacitação como uma das atribuições do Subcomitê de Inteligência Artificial, mas a temática específica da prompt injection precisa ser expressamente incluída nos programas de formação.

6.5. Eixo 5: Governança Algorítmica e Accountability

A governança de IA na Administração Pública deve incorporar mecanismos de *accountability* que permitam rastrear, auditar e responsabilizar. Propõe-se:

a. **Inventário de aplicações de IA:** cada órgão deve manter registro atualizado das ferramentas de IA utilizadas, indicando finalidade, tipo de dados processados, nível de risco e medidas

de proteção adotadas. A Portaria MGI n. 3.485/2026 já aponta nessa direção.

b. **Avaliação de impacto algorítmico:** antes da adoção de qualquer ferramenta de IA em processos decisórios, o órgão deve realizar avaliação de impacto que inclua, entre os riscos mapeados, a exposição à *prompt injection*.

c. **Registro de incidentes:** criação de canal e protocolo para registro, investigação e comunicação de tentativas de *prompt injection* identificadas em processos administrativos, a exemplo do que o STJ determinou para seu acervo processual.

d. **Auditoria e rastreabilidade:** os sistemas de IA devem ser auditáveis e rastreáveis, permitindo a reconstrução do processo decisório e a identificação de eventuais contaminações.

e. **Publicação de diretrizes:** elaboração e publicação de notas técnicas e orientações internas, a exemplo da Nota Técnica 01/2026 do TJSC sobre *prompt injection*, adaptadas ao contexto administrativo.

6.6. Eixo 6: Regulamentação Específica

Por fim, propõe-se a edição de norma específica — resolução, portaria ou instrução normativa, conforme o nível federativo — que discipline a prevenção, a detecção e a resposta à *prompt injection* na Administração Pública. A norma deve prever:

a. A obrigatoriedade de pré-processamento e sanitização de documentos de terceiros antes de sua submissão a

ferramentas de IA.

b. A exigência de supervisão humana qualificada sobre todo produto de IA utilizado em processos decisórios.

c. A classificação de sistemas de IA utilizados em contratações, processos administrativos e atendimento ao cidadão como de alto risco, nos termos do PL 2.338/2023 e do AI Act europeu.

d. A inclusão do risco de *prompt injection* nos programas de gestão de riscos e nos planos de segurança da informação dos órgãos públicos.

e. A criação de protocolo de resposta a incidentes de *prompt injection*, com investigação, registro e comunicação aos órgãos de controle.

f. A vedação de uso de ferramentas de IA generativa sem camadas de proteção contra *prompt injection* em atividades que envolvam análise de documentos de terceiros em processos decisórios.

7. CONSIDERAÇÕES FINAIS

A *prompt injection* não é uma curiosidade técnica nem uma ameaça futurista. É um risco presente, documentado e punido no Judiciário brasileiro desde maio de 2026. Em poucos meses, o STJ, o STF e a Justiça do Trabalho construíram um repertório de respostas que inclui ferramentas de detecção, precedentes jurisprudenciais, sanções processuais e marcos regulatórios. O Judiciário despertou para o problema e reagiu.

Os demais Poderes do Estado, no entanto, ainda não protagonizaram esse mesmo movimento. Não porque o risco seja menor — ao contrário, a função administrativa exercida pelos Poderes Executivo e Legislativo é campo igualmente sensível, senão mais, pela amplitude de suas decisões e pela diversidade de documentos que recebe diariamente de terceiros interessados. Contratações públicas, processos administrativos, atendimento ao cidadão, revisão contratual e formulação de políticas são áreas em que a IA já é utilizada e em que a manipulação por *prompt injection* pode produzir consequências graves: vícios de motivação, lesões à isonomia, favorecimentos indevidos, vazamento de dados e erosão da confiança pública.

O arcabouço normativo vigente — da Constituição Federal à LGPD, passando pela Lei do Governo Digital, pela EBIA, pela Portaria MGI n. 3.485/2026 e pelo PL 2.338/2023 — oferece princípios e deveres gerais que sustentam o enfrentamento do problema, mas não há norma específica que discipline a prevenção, a detecção e a resposta à prompt injection no ambiente administrativo. Essa lacuna precisa ser preenchida com urgência.

Este artigo propôs um protocolo de defesa algorítmica organizado em seis eixos: sanitização de documentos, construção de prompts defensivos, supervisão humana qualificada, capacitação e letramento algorítmico, governança e prestação de contas, e regulamentação específica. O protocolo não exige que o servidor público se torne um especialista em segurança cibernética — exige, antes, que a Administração Pública trate a segurança de seus sistemas de IA com a mesma seriedade com que trata a segurança física, documental e patrimonial de seus processos.

A experiência do Judiciário demonstra que é possível enfrentar a *prompt injection* de forma eficaz, combinando tecnologia, regulação, capacitação e cultura institucional. O próximo passo é transpor esse aprendizado para os demais Poderes. Os tribunais provaram que a máquina pode ser protegida contra o comando invisível. Cabe agora ao Estado, em sua integralidade, garantir que a inteligência artificial sirva ao interesse público — e não o contrário.

O alerta está dado. O risco existe. A resposta é uma questão de decisão institucional.

REFERÊNCIAS BIBLIOGRÁFICAS

BRASIL. Conselho Nacional de Justiça. Resolução n. 615, de 11 de março de 2025. Estabelece diretrizes para o desenvolvimento, utilização e governança de soluções de inteligência artificial no Poder Judiciário. Brasília, DF: CNJ, 2025a. Disponível em: https://atos.cnj.jus.br/files/original155530202503146 7d4517244566.pdf. Acesso em: 4 nov. 2026.

BRASIL. Constituição da República Federativa do Brasil de 1988. Brasília, DF: Presidência da República, 1988. Disponível em: https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acesso em: 4 nov. 2026.

BRASIL. Lei n. 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília, DF: Presidência da República, 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 4 nov. 2026.

BRASIL. Lei n. 14.129, de 29 de março de 2021. Dispõe sobre princípios, regras e instrumentos para o Governo Digital. Brasília, DF:

Presidência da República, 2021a. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/l14129.htm. Acesso em: 4 nov. 2026.

BRASIL. Lei n. 9.784, de 29 de janeiro de 1999. Regula o processo administrativo no âmbito da Administração Pública Federal. Brasília, DF: Presidência da República, 1999. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/l9784.htm. Acesso em: 4 nov. 2026.

BRASIL. Ministério da Ciência, Tecnologia e Inovações. Portaria n. 4.617, de 6 de abril de 2021. Institui a Estratégia Brasileira de Inteligência Artificial (EBIA). Brasília, DF: MCTI, 2021b.

BRASIL. Ministério da Gestão e da Inovação em Serviços Públicos. Portaria n. 3.485, de 24 de abril de 2026. Institui a Política de Governança de Inteligência Artificial no âmbito do MGI. Brasília, DF: MGI, 2026a. Disponível em: https://convergenciadigital.com.br/?p=51474. Acesso em: 4 nov. 2026.

BRASIL. Senado Federal. Projeto de Lei n. 2.338, de 3 de maio de 2023. Dispõe sobre o uso da Inteligência Artificial. Brasília, DF: Senado Federal, 2023. Disponível em: https://www25.senado.leg.br/web/atividade/materias/-/materia/157233. Acesso em: 4 nov. 2026.

BRASIL. Superior Tribunal de Justiça. Novo caso de *prompt injection*: STJ multa advogados em R$ 20 mil por ato contra a dignidade da Justiça. Brasília, DF: STJ, 20 out. 2026d. Disponível em: https://www.stj.jus.br/sites/portalp/Paginas/Comunicacao/Noticias/2026/20102026-Novo-caso-de-prompt-injection-STJ-multa-advogados-

em-R--20-mil-por-ato-contra-a-dignidade-da-Justica.aspx. Acesso em: 4 nov. 2026.

BRASIL. Superior Tribunal de Justiça. Sexta Turma identifica comando oculto em recurso para enganar sistema de IA e determina comunicação à OAB e ao MPF. REsp 2.256.731. Brasília, DF: STJ, set. 2026c. Disponível em: https://www.stj.jus.br/sites/portalp/Paginas/Comunicacao/Noticias/2026/18102026-Quando-a-inteligencia-artificial-vira-questao-de-Justica-as-primeiras-decisoes-do-STJ-sobre-o-uso-de-IA.aspx. Acesso em: 4 nov. 2026.

BRASIL. Superior Tribunal de Justiça. Tentativas de uso de *prompt injection* no STJ serão investigadas. Brasília, DF: STJ, 20 maio 2026b. Disponível em: https://www.stj.jus.br/sites/portalp/Paginas/Comunicacao/Noticias/2026/20052026-Tentativas-de-uso-de-prompt-injection-no-STJ-serao-investigadas.aspx. Acesso em: 4 nov. 2026.

BRASIL. Supremo Tribunal Federal. Petição chega ao STF com *prompt injection* e advogado é multado. Ação Penal 2.822. Brasília, DF: STF, out. 2026e.

CONJUR. Governança digital, IA e o dever de bom desempenho. *Consultor Jurídico*, São Paulo, 27 set. 2026c. Disponível em: https://conjur.com.br/2026-set-27/governanca-digital-ia-e-o-dever-de-bom-desempenho/. Acesso em: 4 nov. 2026.

CONJUR. IA e processo judicial: entre a fraude invisível e o erro no caso do *prompt injection* no TRT-8. *Consultor Jurídico*, São Paulo, 14 maio 2026a. Disponível em: https://www.conjur.com.br/2026-mai-

14/ia-e-processo-judicial-entre-a-fraude-invisivel-e-o-erro-no-caso-do-prompt-injection-no-trt-8/. Acesso em: 4 nov. 2026.

CONJUR. *Prompt injection*: ameaça invisível à imparcialidade do Judiciário na era da IA. *Consultor Jurídico*, São Paulo, 27 set. 2026b. Disponível em: https://conjur.com.br/2025-set-27/prompt-injection-ameaca-invisivel-a-imparcialidade-do-judiciario-na-era-da-ia/. Acesso em: 4 nov. 2026.

FAST COMPANY BRASIL. STJ investiga tentativas de *prompt injection*: entenda o caso. *Fast Company Brasil*, São Paulo, 2026. Disponível em: https://fastcompanybrasil.com/?p=270502. Acesso em: 4 nov. 2026.

JORNAL DE BARUERI. *Prompt injection* em petições deflagra multas a advogados. *Jornal de Barueri*, Barueri, 5 out. 2026. Disponível em: https://jornaldebarueri.com.br/tecnologia/prompt-injection-em-peticoes-deflagra-multas-a-advogados. Acesso em: 4 nov. 2026.

JUSBRASIL. *Prompt injection* e direito penal: entre a atipicidade penal, a fraude algorítmica e os impactos processuais. *JusBrasil*, 2026a. Disponível em: https://www.jusbrasil.com.br/artigos/prompt-injection-e-direito-penal-entre-a-atipicidade-penal-a-fraude-algoritmica-e-os-impactos-processuais-da-manipulacao-de-sistemas-de-inteligencia-artificial/6106494203. Acesso em: 4 nov. 2026.

KASPERSKY. O que é *prompt injection* e como a IA pode ser manipulada? *Kaspersky Resource Center*, 2026. Disponível em: https://www.kaspersky.com.br/resource-center/threats/prompt-injection. Acesso em: 4 nov. 2026.

MIGALHAS. *Prompt injection* e agentes de contratação: riscos e contramedidas. *Migalhas*, São Paulo, 2026d. Disponível em: https://www.migalhas.com.br/depeso/452953/prompt-injection-e-agentes-de-contratacao-riscos-e-contramedidas. Acesso em: 4 nov. 2026.

MIGALHAS. *Prompt injection* em documentos judiciais: conceito, vetores e riscos. *Migalhas*, São Paulo, 2026c. Disponível em: https://www.migalhas.com.br/depeso/455924/prompt-injection-em-documentos-judiciais-conceito-vetores-e-riscos. Acesso em: 4 nov. 2026.

MIGALHAS. *Prompt injection* no Judiciário brasileiro: sentença histórica. *Migalhas*, São Paulo, 2026b. Disponível em: https://www.migalhas.com.br/depeso/455832/prompt-injection-no-judiciario-brasileiro-sentenca-historica. Acesso em: 4 nov. 2026.

MIGALHAS. *Prompt injection*: STJ apura tentativa de manipulação de IA em petições. *Migalhas*, São Paulo, 20 maio 2026a. Disponível em: https://www.migalhas.com.br/quentes/456424/prompt-injection-stj-apura-tentativa-de-manipulacao-de-ia-em-peticoes. Acesso em: 4 nov. 2026.

MIT TECHNOLOGY REVIEW BRASIL. A fraude invisível: *prompt injection*, Justiça e setor público na era da IA. *MIT Technology Review Brasil*, São Paulo, 2026. Disponível em: https://mittechreview.com.br/prompt-injection-risco-invisivel-ia-generativa/. Acesso em: 4 nov. 2026.

OWASP. OWASP Top 10 for LLM Applications 2025. *Open Worldwide Application Security Project*, 2025. Disponível em:

https://owasp.org/www-project-top-10-for-large-language-model-applications/. Acesso em: 4 nov. 2026.

TJRJ. *Prompt injection*. Tribunal de Justiça do Estado do Rio de Janeiro, 2026. Disponível em: https://www.tjrj.jus.br/prompt-injection. Acesso em: 4 nov. 2026.

TJSC. Justiça detecta uso de *prompt injection* em petição e aplica litigância de má-fé. Tribunal de Justiça do Estado de Santa Catarina, 2026. Nota Técnica 01/2026. Disponível em: https://www.tjsc.jus.br/web/imprensa/-/justica-detecta-uso-de-prompt-injection-em-peticao-e-aplica-litigancia-de-ma-fe-. Acesso em: 4 nov. 2026.

WILLISON, Simon. Prompt injection attacks against GPT-3. *Simon Willison's Weblog*, 12 set. 2022. Disponível em: https://simonwillison.net/2022/Sep/12/prompt-injection/. Acesso em: 4 nov. 2026.

---

[1] Servidor efetivo do Tribunal de Justiça do Estado de Rondônia. Pós-graduado em Direito Administrativo pela PUC Minas Gerais, em Direito Constitucional pela Universidade Federal do Tocantins (UFT) e em Processual e Direito Previdenciário pela Escola Aberta de Direito. Ex-Assessor Jurídico do Tribunal de Contas de Minas Gerais e do Tribunal de Justiça do Tocantins.