MANIFESTAÇÃO TÉCNICA CNIAJ 1/2026

INJEÇÃO DE COMANDOS

Considerações técnicas sobre mitigação de riscos de prompt injection em sistemas judiciais de inteligência artificial

Origem	Processo SEI 09953/2026
Requerente	Presidência do CNIAJ
Relatoria	DTI
Órgão	Conselho Nacional de Justiça

Brasília, 27 de maio de 2026.

MANIFESTAÇÃO TÉCNICA CNIAJ 1/2026

Origem: Processo SEI 09953/2026
Requerente: Presidência do CNIAJ
Relatoria: Departamento de Tecnologia da Informação e Comunicação do CNJ
Assunto: Considerações técnicas sobre medidas de mitigação de riscos de *prompt injection* em sistemas judiciais de inteligência artificial.

1. Contextualização

Trata-se de manifestação técnica Comitê Nacional de Inteligência Artificial do Judiciário – CNIAJ, sob relatoria do Departamento de Tecnologia da Informação e Comunicação – DTI/CNJ, em atenção ao encaminhamento de estudo prévio para a instituição do Programa de Segurança Adversarial para Sistemas de Inteligência Artificial do Poder Judiciário Brasileiro e a apresentação de considerações e sugestões acerca da prevenção e mitigação de riscos decorrentes da injeção de comandos ocultos, ou *prompt injection*, em soluções de inteligência artificial desenvolvidas, contratadas ou utilizadas no âmbito do Poder Judiciário.

A presente Manifestação Técnica tem por finalidade contribuir com a análise da matéria sob a perspectiva tecnológica, especialmente quanto à arquitetura segura de sistemas judiciais de IA generativa, aos riscos associados ao processamento de peças processuais, anexos, metadados, documentos externos, bases de conhecimento e textos extraídos por reconhecimento óptico de caracteres (OCR), bem como quanto à viabilidade de estabelecimento de orientação técnica inicial aos tribunais e de estruturação operacional das medidas de maior complexidade.

Reconhece-se a relevância e a oportunidade do Proseg-IA, que apresenta contribuição institucional importante ao organizar, em programa estruturado, tema sensível para a segurança, a confiabilidade e a governança das soluções de inteligência artificial no Poder Judiciário. A proposta tem o mérito de deslocar o debate sobre *prompt injection* de uma perspectiva meramente técnica ou episódica para uma abordagem de gestão de riscos, segurança adversarial, auditoria, contratação, resposta a incidentes e articulação institucional.

A presente manifestação tem por objetivo complementar sua formulação com insumos técnicos dos integrantes do CNIAJ, do DTI/CNJ e do Programa Conecta, incorporar contribuições suscitadas na reunião de apresentação e consolidar, em documento técnico autônomo, as medidas que se mostram adequadas ao enfrentamento inicial e progressivo dos riscos de manipulação adversarial em sistemas judiciais de IA.

2. Pertinência e oportunidade do Proseg-IA

O Proseg-IA é iniciativa tecnicamente adequada, institucionalmente necessária e alinhada ao atual estágio de adoção de inteligência artificial no Poder Judiciário. A vulnerabilidade de *prompt injection* não veicula mais risco meramente hipotético, tratado no âmbito acadêmico, mas ameaça compatível com o próprio modo de funcionamento de modelos de linguagem de grande porte. O risco é inerente aos modelos que processam conteúdos fornecidos por usuários ou terceiros, como peças e documentos processuais, anexos e metadados, recorrendo a componentes da infraestrutura informacional do sistema usados para contextualizar ou orientar as respostas.

Em sistemas judiciais, uma instrução adversarial pode ser inserida no corpo de uma petição, documento anexo, texto oculto, metadado de arquivo, comentário, camada invisível, imagem submetida a OCR ou em base de conhecimento utilizada por mecanismo de recuperação aumentada. A depender da arquitetura da solução o modelo pode considerar esse conteúdo como comando a ser obedecido em vez de interpretá-lo como dado processual a ser analisado.

A manipulação adversarial pode, além de comprometer informações de acesso restrito, induzir o sistema a fornecer saídas anômalas, tais como distorcer resumo, omitir argumento relevante, favorecer determinada tese, alterar prioridade, simular conclusão e produzir resposta desvinculada dos autos, com entrega incompatível com a finalidade institucional da ferramenta. Por essa razão, a matéria deve ser tratada como tema interdisciplinar de governança de inteligência artificial que envolve segurança da informação e proteção de dados pessoais, com reflexos na integridade processual e confiabilidade institucional, o que gera risco reputacional.

O Proseg-IA compreende marco inicial de organização do tema no âmbito nacional, reunindo diagnóstico e diretrizes programáticas para tratamento coordenado da segurança adversarial em inteligência artificial a partir de eixos de atuação determinados. Sua estruturação permite que o CNJ enfrente essa questão por meio de estratégia estruturada e progressiva, alinhada com a governança de IA definida pela Resolução CNJ n. 615/2025.

O programa propõe abordagem estruturada, com eixos voltados a diagnóstico nacional, protocolo de auditoria adversarial, diretrizes para contratação, requisitos para plataformas nacionais, levantamento junto aos tribunais, protocolo de resposta a incidentes e demais medidas de governança. Essa estrutura, alinhada com as melhores práticas internacionais, viabiliza a atuação coordenada do Conselho Nacional de Justiça na construção progressiva de parâmetros nacionais.

Ao mesmo tempo, verifica-se que alguns pontos já possuem maturidade suficiente para integrar uma nota técnica autônoma a ser submetida ao Plenário. Essas diretrizes se organizam nas seguintes frentes: a) classificação e acompanhamento dos sistemas de IA (identificação de sistemas de maior risco e consolidação das informações na Plataforma Sinapses); b) tratamento de conteúdos não confiáveis (documentos externos, rastreabilidade,

supervisão humana efetiva e filtros de saída); e c) resposta institucional a incidentes ou ocorrências relevantes (certificação nos autos e registro de incidentes).

Assim, recomenda-se que o Proseg-IA seja compreendido em duas dimensões complementares. A primeira, imediata, voltada à aprovação da presente Manifestação Técnica como documento autônomo de orientação e referência técnica. A segunda, estruturante, voltada ao aprofundamento metodológico, à evolução do Sinapses, à definição de protocolos de auditoria adversarial, à padronização de requisitos técnicos e à elaboração de eventuais atos normativos futuros.

3. Insumos técnicos consolidados pelo DTI

A análise técnica indica que a mitigação dos riscos de *prompt injection* exige abordagem em profundidade e deve partir da distinção entre dois grupos principais de risco.

O primeiro grupo envolve **vazamento de dados, credenciais, segredos, trechos de instruções internas, informações protegidas por sigilo ou elementos objetivamente proibidos na saída do sistema**. Nesses casos, a filtragem determinística de saída é útil para bloquear ou sinalizar, antes da entrega ao usuário, respostas que contenham credenciais, *tokens*, dados sensíveis, trechos de *prompt* institucional, identificadores sigilosos ou expressões incompatíveis com a finalidade da ferramenta.

O segundo grupo envolve **sequestro ou manipulação de comportamento**, hipótese em que o objetivo do atacante não é necessariamente obter informação protegida, mas induzir o modelo a agir de forma indevida. Esse risco pode ocorrer quando uma peça processual contém instruções ocultas ou dissimuladas para que o sistema favoreça uma tese que, pela manipulação dos fatos e dos argumentos e pela distorção de resumos, simule uma conclusão.

A filtragem de saída, embora necessária, não é suficiente para todos os cenários. Ela é eficaz quando o conteúdo proibido pode ser previamente definido e objetivamente verificado; entretanto, a manipulação semântica do comportamento do modelo pode ocorrer sem que a resposta final contenha indicativo de manipulação facilmente detectável. Um resumo pode parecer formalmente imparcial e ainda assim ter sido contaminado por instrução oculta em peça processual.

Por essa razão, no contexto judicial, a segurança deve começar na **ingestão segura dos documentos processuais**. Esta etapa de extração, preparação, classificação e apresentação do conteúdo ao modelo é crítica, e o texto extraído de peças, incluindo imagens e metadados, deve ser tratado como dado externo potencialmente não confiável, em linha com recomendações internacionais, acompanhado de informações sobre origem, página, peça, camada, visibilidade, contraste, posição e demais elementos necessários à rastreabilidade.

Recomenda-se que sistemas judiciais de IA preservem, sempre que tecnicamente possível, metadados visuais e estruturais como tamanho da fonte, cor da fonte, contraste com o fundo, posição na página, camada do documento, opacidade, rotação, sobreposição com outros elementos, origem do texto, identificação da peça, página e trecho de origem. Esses elementos viabilizam distinção entre o texto ordinariamente visível e o tecnicamente presente no arquivo, mas potencialmente adversarial.

Também se recomenda a separação entre **texto canônico visível** e **conteúdo suspeito ou anômalo**. O primeiro corresponde ao conteúdo ordinariamente perceptível em condições normais de leitura e pode ser utilizado como base principal para sumarização, classificação, pesquisa ou geração assistida. O segundo compreende elementos como fonte diminuta, texto branco sobre fundo branco, baixo contraste, camadas ocultas, metadados incomuns, comentários, objetos sobrepostos ou trechos fora da área visível da página.

Esse conteúdo suspeito não deve ser descartado silenciosamente sem o devido registro, pois sua existência pode ter relevância para auditoria, averiguação administrativa, análise processual, segurança da informação ou eventual identificação de tentativa de manipulação, com vistas inclusive à apuração de responsabilidades. O procedimento recomendável é segregá-lo do contexto principal enviado ao modelo, registrar o motivo técnico da segregação, preservar o trecho, indicar peça e página de origem, gerar trilha de auditoria e permitir revisão humana quando necessário.

Preferencialmente, o tratamento prévio do dado deve ocorrer em camada ou sistema autônomo de preparação documental, separado da aplicação de IA responsável pela geração da resposta. Essa camada deve executar a extração, normalização, classificação, validação, segregação de conteúdo suspeito, preservação de metadados, geração de trilhas de auditoria e disponibilização do conteúdo canônico em formato próprio para consumo seguro por soluções de IA.

O *Datalake* do Poder Judiciário contribui para esta finalidade na medida em que pode, sem substituir os sistemas processuais, apoiar a consolidação de dados processuais previamente tratados, rastreáveis, auditáveis e governados, observadas as regras de segurança, sigilo, proteção de dados e controle de acesso aplicáveis. Sua utilização, alinhada às regras de governança de IA, pode reduzir a exposição direta das ferramentas generativas a documentos brutos e conteúdos externos não confiáveis, além de metadados anônimos.

Além da ingestão segura, o DTI entende que o Proseg-IA deve contemplar diretrizes específicas sobre o **tratamento da saída gerada por sistemas de IA**, pois parte relevante dos riscos adversariais somente se materializa quando a resposta é apresentada ao usuário ou utilizada como apoio a uma atividade judicial ou administrativa.

A filtragem determinística de saída é entendida como camada técnica relevante, externa ao modelo, destinada a bloquear e sinalizar respostas que contenham elementos

objetivamente incompatíveis com a finalidade do sistema, submetendo-a à revisão humana. Essa filtragem pode incidir, entre outros pontos, sobre credenciais, *tokens*, chaves de integração, dados pessoais sensíveis, informações protegidas por segredo de justiça, trechos de *prompt* institucional, instruções internas, identificadores sigilosos ou padrões textuais que indiquem tentativa de exposição indevida de regras operacionais do sistema.

No contexto judicial, recomenda-se que essa camada também contemple verificações relacionadas à natureza da resposta esperada. Sistemas destinados apenas à sumarização, classificação, pesquisa assistida, triagem ou apoio informacional não devem produzir saídas com aparência de decisão judicial. Assim, quando incompatíveis com a finalidade declarada da ferramenta, expressões como “julgo procedente”, “condeno”, “defiro a tutela”, “indefiro o pedido” ou comandos equivalentes podem ser objeto de bloqueio ou de alerta com encaminhamento obrigatório à revisão.

Sempre que tecnicamente viável, sem prejuízo da busca de palavras e expressões, recomenda-se igualmente a adoção de controles combinados, compostos por regras determinísticas, validação de formato, checagem de campos obrigatórios, verificação de referências às fontes utilizadas, identificação de ausência de evidências e, em sistemas de maior risco, mecanismos auxiliares de classificação ou pontuação de risco da resposta. Esses mecanismos auxiliares podem contribuir para indicar respostas potencialmente inadequadas ou desvinculadas dos documentos de origem, sem que com isso substituam controles determinísticos nem supervisão humana.

Também se recomenda que as saídas sejam submetidas a **contratos de resposta**, isto é, formatos previamente definidos conforme a finalidade do sistema. Em uma ferramenta de resumo, por exemplo, a saída deve limitar-se a campos como fatos principais, pedidos, fundamentos invocados, documentos relevantes, pontos controvertidos e incertezas, preferencialmente com indicação da peça, página ou trecho de suporte. Já nas ferramenta de classificação, a resposta deve explicitar a classe sugerida, o grau de confiança, os elementos utilizados e as hipóteses alternativas, evitando conclusões categóricas sem lastro documental.

Esse tratamento estruturado da saída reduz a margem para respostas livres e dificulta que uma instrução adversarial produza efeitos sem deixar sinais verificáveis. A exigência de vinculação a evidências, embora não elimine o risco de manipulação, permite que as equipes técnica e finalística e eventuais mecanismos automatizados verifiquem se a resposta efetivamente decorre do conteúdo canônico dos autos ou se apresenta algum tipo de anomalia (deslocamento semântico, omissão relevante, ênfase indevida, conclusão não suportada pelo material processual).

Deve-se reconhecer, contudo, que a filtragem de saída é mais efetiva contra vazamentos e saídas objetivamente proibidas do que contra manipulações semânticas sutis, sendo necessária mas não suficiente. Por isso, o tratamento da saída deve ser articulado com a

ingestão segura, a segregação de conteúdo suspeito, o encapsulamento dos autos como dados não confiáveis, a rastreabilidade das fontes, os logs de execução e a supervisão humana.

Por fim, reitera-se o papel central exercício pela supervisão humana, ainda que tampouco seja suficiente para o enfrentamento dos riscos. A revisão humana somente se torna efetiva quando apoiada por rastreabilidade, logs, explicitação das fontes utilizadas, indicação de trechos suspeitos, marcação de incertezas, controles de saída e documentação do comportamento esperado do sistema. Sem tais elementos, há risco de que a supervisão humana se torne meramente formal, especialmente em ambientes de alto volume processual.

4. Propostas de complementação ao Proseg-IA

4.1 Diagnóstico nacional de vulnerabilidades e uso da Plataforma Sinapses

Recomenda-se que o diagnóstico nacional previsto no Proseg-IA não se limite ao levantamento declaratório da existência de sistemas de inteligência artificial nos tribunais, mas avance para a identificação do grau de exposição adversarial de cada solução, especialmente daquelas que utilizem modelos de linguagem, IA generativa, mecanismos de recuperação aumentada de conhecimento ou processamento de documentos externos.

Para esse fim, sugere-se que a **Plataforma Sinapses seja indicada como solução para a realização, consolidação e manutenção do inventário nacional de sistemas judiciais de IA e de sua exposição a riscos adversariais**, em razão de sua vocação institucional como ambiente nacional de registro, governança, compartilhamento e acompanhamento de soluções de inteligência artificial no Poder Judiciário.

A utilização do Sinapses como base de consolidação do inventário permite evitar a criação de instrumento paralelo de coleta, reduzir a dispersão informacional, aproveitar estrutura tecnológica já vinculada à governança nacional de IA e estabelecer repositório permanente para atualização do diagnóstico. Em vez de levantamento pontual e estático, o inventário de riscos adversariais pode ser incorporado como dimensão específica do cadastro, homologação, documentação ou acompanhamento de soluções de IA na plataforma.

Recomenda-se, portanto, que o Proseg-IA preveja a evolução do Sinapses para contemplar campos específicos de segurança adversarial, incluindo, no mínimo: a) finalidade do sistema; b) tipo de modelo utilizado; c) existência de IA generativa; d) processamento de documentos externos; e) uso de OCR; f) integração com sistemas processuais; g) uso de RAG ou bases de conhecimento; h) tratamento de dados sigilosos; i) existência de filtros de saída; j) segregação de conteúdo suspeito; k) trilhas de auditoria; l) supervisão humana; m) mecanismos de resposta a incidentes; n) histórico de testes adversariais; e o) estágio de homologação ou produção.

Também se sugere que o Sinapses permita o registro de atributos técnicos de exposição adversarial, a serem considerados na avaliação e categorização de risco prevista na Resolução CNJ nº 615/2025, sem criação de classificação paralela. Esses atributos devem contemplar, especialmente, soluções que processem documentos externos e produzam respostas em linguagem natural, resumos, classificações, triagens, recomendações, minutas ou análises de conteúdo processual, podendo subsidiar a definição de prioridades para auditoria, adequação contratual, resposta a incidentes e eventual exigência de requisitos adicionais para integração com plataformas nacionais.

4.2 Protocolo de auditoria adversarial

Recomenda-se que o protocolo de auditoria adversarial contemple testes sobre o modelo e sobre o *pipeline* documental. Devem ser previstos testes com petições, PDFs, imagens e bases documentais contendo texto oculto, fonte diminuta, baixo contraste, metadados adversariais, comentários, objetos sobrepostos, camadas invisíveis, texto fora da área visível, links ativos, OCR induzido a erro e conteúdo malicioso inserido em bases de recuperação.

A auditoria deve verificar se o sistema preserva metadados visuais e estruturais, separa texto canônico de conteúdo suspeito, registra a origem dos trechos utilizados, encapsula documentos como dados não confiáveis e reduz o risco de que instruções contidas nos autos sejam interpretadas como comandos operacionais.

Recomenda-se, ainda, que as avaliações sejam proporcionais ao risco do sistema. Soluções que processem documentos externos, utilizem IA generativa, produzam resumos, minutas, classificações ou recomendações devem receber maior nível de atenção. Testes excessivamente breves podem gerar falsa percepção de segurança, razão pela qual a metodologia nacional deve considerar avaliações adaptativas, cenários variados e ciclos periódicos de verificação.

4.3 Diretrizes para contratação, desenvolvimento e homologação

Recomenda-se que futuras contratações e desenvolvimentos de soluções de IA generativa ou de sistemas que processem documentos judiciais contenham requisitos mínimos de segurança adversarial, incluindo tratamento de documentos externos como dados não confiáveis, extração segura de PDFs e imagens, preservação de metadados visuais e estruturais, segregação de conteúdo oculto ou anômalo, filtragem determinística de saída, validação de formato da resposta, registro de *logs*, rastreabilidade entre resposta e fonte documental, documentação de arquitetura, plano de resposta a incidentes e mecanismos de revisão humana.

Para contratos em execução, a adequação deve ocorrer de forma progressiva e proporcional ao risco, evitando paralisação de serviços essenciais ou imposição técnica sem

avaliação prévia de viabilidade jurídica, contratual e operacional. Para soluções experimentais, de pesquisa ou de baixa criticidade, os requisitos podem ser graduados conforme o nível de risco, sem prejuízo da adoção de salvaguardas mínimas quando houver processamento de dados reais, em especial sigilosos, ou oriundos de autos judiciais.

4.4 Requisitos para PDPJ-Br, Datalake e plataformas nacionais

Recomenda-se que sistemas de IA generativa integrados à PDPJ-Br e a demais plataformas nacionais observem requisitos mínimos de tratamento dos autos como dados não confiáveis, encapsulamento de conteúdo processual, identificação da origem de cada trecho, filtragem de saída, validação de formato, logs auditáveis, segregação de conteúdo suspeito e possibilidade de auditoria técnica.

Para sistemas que utilizem recuperação aumentada de conhecimento, recomenda-se atenção específica à qualidade, procedência, atualização e integridade das bases utilizadas, inclusive com controle contra envenenamento de base, inserção de documentos adversariais e recuperação de conteúdo não validado.

No que se refere à preparação e à disponibilização de dados, recomenda-se avaliar a utilização de camada ou sistema autônomo de tratamento prévio, preferencialmente desacoplado da ferramenta generativa, que permita sanear, classificar, enriquecer, auditar e versionar os dados antes de sua utilização por modelos de IA. O *Datalake* do Poder Judiciário pode constituir caminho institucional para essa finalidade, especialmente quando associado a controles de origem, rastreabilidade, governança de acesso, preservação de metadados e segregação de conteúdo suspeito ou anômalo.

A integração com plataformas nacionais deve observar, sempre que possível, critérios de rastreabilidade, documentação técnica e transparência operacional, de modo a permitir que o CNJ e os tribunais compreendam quais dados foram utilizados, quais mecanismos de controle foram acionados e quais salvaguardas estão presentes na geração da resposta.

4.5 Protocolo de resposta a incidentes e Resolução CNJ n.º 396/2021

Recomenda-se que o protocolo de resposta a incidentes contemple eventos de manipulação adversarial, suspeita de documento processual contaminado, vazamento de dados, geração de saída indevida, contaminação de base de conhecimento e falha de segregação documental.

O protocolo deve prever preservação de evidências, registro do documento de origem, hash dos trechos relevantes, versão do modelo, versão do *prompt* institucional, parâmetros de execução, logs de entrada e saída, usuário solicitante, contexto da consulta, filtros acionados,

alertas emitidos, medidas de contenção, comunicação interna e eventual encaminhamento para análise pelas áreas técnicas, de segurança da informação, de governança de IA ou por outras unidades competentes, conforme a natureza do evento.

Na ocorrência de incidente de segurança cibernética ou de evento que indique comprometimento de confidencialidade, integridade, disponibilidade, autenticidade ou funcionamento seguro de sistemas judiciais de IA, recomenda-se observar os dispositivos de tratamento, resposta, comunicação e preservação de evidências previstos na Resolução CNJ n. 396/2021, que institui a Estratégia Nacional de Segurança Cibernética do Poder Judiciário.

Em especial, devem ser considerados os processos de resposta e tratamento a incidentes, a atuação das Equipes de Tratamento e Resposta a Incidentes de Segurança Cibernética – ETIR, a comunicação interna e externa aplicável, o reporte ao Centro de Prevenção, Tratamento e Resposta a Incidentes Cibernéticos do Poder Judiciário – CPTRIC-PJ, quando cabível, e a observância dos protocolos de prevenção, gerenciamento de crises e investigação de ilícitos cibernéticos instituídos no âmbito da ENSEC-PJ.

A existência de procedimento mínimo de resposta evita perda de evidências ao permitir reprocessamento controlado e preservar a rastreabilidade da ocorrência, o que serve como apoio para eventual apuração pelas áreas competentes.

5. Orientações complementares: certificação nos autos e capacitação

5.1 Certificação nos autos da detecção de conteúdo adversarial

A segregação de conteúdo suspeito ou anômalo, recomendada no âmbito da ingestão segura de documentos, produz efeito útil sobretudo quando o fato é tornado conhecido das partes e do juízo. Por essa razão, sugere-se orientar os tribunais a que, identificada pelo sistema de IA a presença de instrução adversarial, como texto oculto, comando dissimulado, metadado anômalo ou equivalente, tal ocorrência seja certificada nos autos, de forma clara e objetiva, preservando-se elementos que permitam rastreabilidade .

A presente orientação limita-se à certificação do fato detectado. As providências eventualmente cabíveis, sejam elas de natureza sancionatória, disciplinar, comunicacional ou de qualquer outra, inserem-se na esfera de competência e de independência funcional do magistrado e na autonomia administrativa de cada tribunal, não constituindo objeto desta Manifestação Técnica.

A título ilustrativo, e sem caráter vinculante ou exaustivo, a certificação poderá contemplar elementos mínimos de rastreabilidade que dialogam com as trilhas de auditoria já recomendadas, tais como a peça ou documento em que o conteúdo foi detectado, com indicação do respectivo identificador no sistema processual; a natureza do conteúdo detectado, como texto

invisível, fonte diminuta, caractere de largura zero ou metadado anômalo; a ação adotada pelo sistema de IA diante do conteúdo, indicando se houve segregação, descarte técnico ou processamento controlado; e a data e hora da detecção, bem como a versão do modelo e das instruções institucionais aplicáveis.

5.2 Capacitação de magistrados, servidores e equipes técnicas

A supervisão humana efetiva pressupõe supervisores aptos a reconhecer saídas anômalas e indícios de manipulação adversarial. Uma supervisão desprovida de preparo técnico tende a tornar-se meramente formal, sobretudo em ambientes de elevado volume processual.

Por essa razão, sugere-se reforçar a orientação para que magistrados e servidores, tanto da área-meio quanto da área-fim, e equipes técnicas que venham a utilizar ferramentas de inteligência artificial recebam capacitação prévia e continuada, proporcional à criticidade dos sistemas e às funções exercidas, em especial quando encarregados de gestão ou de auditoria de modelos.

A capacitação deve abranger, entre outros aspectos, noções básicas sobre *prompt injection* e manipulação adversarial, identificação de sinais de saídas potencialmente manipuladas, boas práticas de uso seguro das ferramentas, delimitação do escopo dos comandos, preservação do controle humano sobre as saídas geradas e procedimentos internos para registro e encaminhamento de ocorrências às áreas técnicas e de segurança da informação.

Recomenda-se que a capacitação seja desenvolvida em articulação com as escolas judiciais e com as diretrizes nacionais aplicáveis, preferencialmente sob diretrizes estabelecidas pelo CNJ e pelas Escolas Nacionais (Enaju, Enfam e Enamat), em formato que combine educação a distância, materiais práticos e atualização periódica, de modo a acompanhar a evolução das técnicas de ataque e de defesa.

6. Aprovação, condições de viabilidade e caminho de operacionalização

A partir das contribuições suscitadas na reunião de apresentação, recomenda-se que a presente Manifestação seja submetida ao Plenário do CNJ como documento autônomo, de natureza técnica, apto a consolidar os encaminhamentos iniciais sobre segurança adversarial em sistemas judiciais de IA, como marco inicial das ações subsequentes que serão estruturadas no Proseg-IA.

A execução do Programa é tecnicamente viável e estrategicamente recomendada, desde que apoiada em metodologia progressivamente consolidada, com iniciativas faseadas adotadas considerando os riscos presentes e futuros. A implementação imediata de todas as recomendações pode gerar dificuldades técnicas e orçamentárias, além de externalidades

negativas como a baixa aderência dos tribunais, tendo como produto indesejado a produção de respostas meramente formais.

Por essa razão, recomenda-se que a aprovação plenária de eventual Nota Técnica elaborada a partir deste insumo tenha caráter de linha de orientação técnica inicial, enquanto o CNIAJ, com apoio do DTI e das demais unidades competentes, aprofunda a metodologia de diagnóstico, auditoria, homologação, monitoramento e resposta a incidentes. Com isso, espera-se que o documento cumpra sua função de indução inicial de boas práticas, sem pretender substituir os instrumentos técnicos e operacionais necessários à implementação progressiva das salvaguardas.

O Programa Justiça [+Segura] pode operacionalizar ações executivas de maior complexidade decorrentes do Proseg-IA, no que for compatível com seu escopo. O projeto BRA/25/025, fruto de cooperação técnica entre o CNJ e o Programa das Nações Unidas para o Desenvolvimento (Pnud), tem por objeto o fortalecimento da segurança cibernética do Poder Judiciário, promovendo inovações com previsão de desenvolvimento e implementação de tecnologias, integração de sistemas e plataformas, monitoramento de ferramentas digitais e produção de metodologias de apoio à transformação digital segura do Judiciário.

Assim, recomenda-se que medidas como auditorias adversariais, testes técnicos, validação de bases de recuperação aumentada, adaptação de pipelines documentais, estruturação de logs e trilhas de auditoria, evolução da Plataforma Sinapses, definição de modelos de resposta a incidentes e monitoramento de soluções de IA em produção sejam avaliadas, no que couber, para estruturação como trilha executiva no âmbito do Justiça [+Segura]. Dado o suporte técnico e orçamentário oferecido pelo projeto, os desdobramentos técnicos da iniciativa serão desenvolvidos e monitorados em ambiente institucional com estrutura específica e vocacionado ao fortalecimento da segurança cibernética e à evolução segura dos serviços digitais do Poder Judiciário.

Também se recomenda que o primeiro ciclo de implementação priorize sistemas de maior exposição adversarial, assim entendidos aqueles que utilizem IA generativa ou modelos de linguagem para processar documentos externos, desempenhando atividades tais como a realização de sumarização, classificação, triagem e pesquisa assistida, a geração de minutas e recomendações ou, ainda, a recuperação aumentada de conhecimento.

Sem prejuízo da classificação de risco prevista na Resolução CNJ n. 615/2025, os requisitos técnicos de mitigação podem ser organizados por níveis de maturidade, como forma de apoiar implementação progressiva e proporcional:

a) **medidas iniciais**: documentação da finalidade do sistema, identificação das fontes de dados, supervisão humana, logs, política de tratamento de dados sigilosos, controles básicos de saída e declaração de processamento de documentos externos;

b) **medidas recomendadas**: extração segura, segregação de conteúdo suspeito, encapsulamento de dados não confiáveis, respostas estruturadas e vinculação a evidências; e
c) **medidas avançadas**: *red team*, testes adaptativos, detecção semântica auxiliar, validação de bases RAG, métricas de robustez e auditoria adversarial periódica.

Destaca-se, por fim, que níveis de maturidade não veiculam nova classificação de risco em substituição ao disposto na Resolução CNJ n. 615/2025, constituindo-se em parâmetros operacionais para apoiar a evolução técnica das salvaguardas.

7. Síntese das propostas do DTI

Diante da análise realizada, o DTI sugere que o CNIAJ e, se assim deliberado, o Plenário do CNJ, considerem as seguintes propostas:

a) prestigiar o Proseg-IA como programa estruturante de segurança adversarial para sistemas de inteligência artificial do Poder Judiciário;

b) submeter ao Plenário do CNJ a presente Manifestação Técnica como documento autônomo de referência técnica, sem minuta de recomendação anexa neste momento, preservando a possibilidade de elaboração de ato orientativo próprio em etapa posterior;

c) indicar a Plataforma Sinapses como solução nacional para realização, consolidação e manutenção do inventário nacional de sistemas judiciais de IA e de sua exposição a riscos adversariais;

d) incorporar ao Proseg-IA requisitos de ingestão segura de documentos processuais, com preservação de metadados visuais e estruturais;

e) prever, preferencialmente, camada ou sistema autônomo de tratamento prévio dos dados, anterior ao consumo por modelos generativos, com possibilidade de utilização do Datalake do Poder Judiciário como caminho institucional para consolidação de bases tratadas, rastreáveis e auditáveis;

f) prever separação entre texto canônico visível e conteúdo suspeito ou anômalo, com segregação, preservação e registro auditável;

g) orientar que documentos dos autos, anexos, metadados, OCR e bases externas sejam tratados como dados não confiáveis, devidamente encapsulados e identificados antes de apresentados ao modelo;

h) recomendar filtragem determinística de saída, ou controles equivalentes externos ao modelo, como camada relevante para mitigação de vazamento de dados,

credenciais, trechos de prompt institucional, segredo de justiça, informações sensíveis e saídas objetivamente incompatíveis com a finalidade do sistema;

i) prever controles específicos para evitar que sistemas de sumarização, triagem, classificação, pesquisa assistida ou apoio informacional gerem respostas com aparência de decisão judicial, comando dispositivo ou conclusão jurisdicional autônoma;

j) orientar a adoção de respostas estruturadas, com contratos de resposta adequados à finalidade do sistema, validação de formato e vinculação das afirmações relevantes a evidências extraídas dos autos;

k) deixar claro que a filtragem de saída é necessária, mas não suficiente para mitigar manipulação semântica ou sequestro de comportamento, devendo ser combinada com ingestão segura, segregação de conteúdo suspeito, encapsulamento de dados não confiáveis, rastreabilidade, logs e supervisão humana;

l) recomendar logs auditáveis e rastreabilidade entre entrada, trecho utilizado, resposta gerada, versão do modelo, versão do prompt, filtros acionados, alertas emitidos e usuário solicitante;

m) prever a certificação nos autos da detecção de conteúdo adversarial ou anômalo, quando identificada por sistema de IA, preservada a autonomia decisória do magistrado quanto às providências cabíveis;

n) reforçar a necessidade de capacitação prévia e continuada de magistrados, servidores e equipes técnicas que utilizem, gerenciem ou auditem ferramentas de IA, proporcionalmente ao risco e à criticidade da solução;

o) orientar que, na ocorrência de incidentes, sejam observados os dispositivos de tratamento, resposta, comunicação e preservação de evidências previstos na Resolução CNJ n.º 396/2021 e nos protocolos dela decorrentes;

p) orientar implementação faseada, priorizando sistemas de maior risco e maior exposição adversarial;

q) calibrar inventário nacional, exigências de avaliação adversarial, requisitos de homologação e controles técnicos segundo capacidade operacional, maturidade metodológica, criticidade dos sistemas e classificação de risco prevista na Resolução CNJ n.º 615/2025, considerando os atributos técnicos de exposição a prompt injection e manipulação adversarial como elementos complementares da avaliação de risco;

r) recomendar que as medidas de implementação operacional mais complexas decorrentes do Proseg-IA sejam estruturadas, no que couber, como trilha executiva no âmbito do Programa Justiça [+Segura], especialmente aquelas relacionadas a auditorias adversariais, testes técnicos, evolução de sistemas em produção, adaptação de pipelines documentais, estruturação de instrumentos de monitoramento, resposta a incidentes e desenvolvimento de metodologias de apoio aos tribunais.

8. Conclusão

O Comitê manifesta-se favoravelmente ao prosseguimento do Proseg-IA e reconhece a relevância técnica e institucional da proposta. A matéria é atual, e sua sensibilidade demanda atuação coordenada do Conselho Nacional de Justiça, especialmente diante da expansão de soluções judiciais baseadas em modelos de linguagem e IA generativa.

Entende-se, ainda, que há elementos suficientes para submissão da presente Manifestação Técnica ao Plenário do CNJ como documento autônomo de orientação técnica inicial para sistemas judiciais de IA sujeitos a riscos de prompt injection e manipulação adversarial, sem prejuízo do posterior desenvolvimento de protocolos, requisitos, recomendações ou atos normativos próprios pelo CNIAJ e pelo CNJ.

Recomenda-se, portanto, que o Proseg-IA seja prestigiado como programa estruturante, ao mesmo tempo em que a presente Manifestação, submetida ao Plenário para avaliação da oportunidade e conveniência de adoção de Nota Técnica, seja aprovada como referência técnica inicial sobre inventário, classificação de risco, ingestão segura de documentos, tratamento de autos como dados não confiáveis, filtragem de saída, estruturação das respostas, certificação nos autos de ocorrências relevantes, capacitação, rastreabilidade, supervisão humana, registro das soluções na Plataforma Sinapses e tratamento de incidentes conforme a Resolução CNJ n. 396/2021.

No ambiente judicial, a segurança começa na forma como os autos são lidos, extraídos, preparados, classificados, segregados e apresentados ao modelo, preferencialmente por camada autônoma de tratamento prévio de dados, e prossegue na forma como a resposta é estruturada, filtrada, rastreada, vinculada a evidências e submetida à revisão humana proporcional ao risco. Essa premissa deve orientar a evolução do Proseg-IA e a futura regulamentação nacional sobre o tema.

Por fim, considerando que parte das providências sugeridas demandará execução técnica progressiva, apoio especializado, testes, auditorias, evolução de sistemas e eventual adaptação de serviços já disponíveis em produção, recomenda-se que as ações operacionais de maior complexidade sejam desde logo avaliadas para estruturação como trilha executiva no âmbito do Programa Justiça [+Segura], sem prejuízo da governança própria do Proseg-IA, da Plataforma Sinapses, do CNIAJ e das unidades técnicas competentes do CNJ.

É a Manifestação Técnica.