Escrever peças para humanos e máquinas: processo penal

12/06/2026

Quando o órgão julgador já não lê mais sozinho

A incorporação da inteligência artificial à prática forense é irreversível, e a posição pessoal de cada um não altera o estado de coisas. Pressionados pela produtividade, julgadores, assessores e partes recorrem aos grandes modelos de linguagem — LLMs *(large language models)* — para resumos de processos, análise de teses, sugestão de argumentos e minutas, triagem de recursos. Parte admite o uso; outra não. O resultado tende à uniformidade: a peça protocolada pode ser lida primeiro por uma máquina e só depois pelo julgador.

Quem ignora o fenômeno abre mão de uma camada inteira de processamento. O texto passa a ser lido por um intermediário cujos vieses o redator desconhece e que pode distorcer a argumentação sem registro perceptível, antecipando conclusões. A peça tem dois leitores, simultâneos ou sequenciais: o humano (magistrado, assessor, estagiário) e o modelo que filtra, resume e classifica o documento. Desconsiderar o segundo deixa metade do percurso comunicativo sem controle.

O segundo leitor é mensurável: preferências comportam estudo, vieses comportam mapeamento. As estratégias eficazes para a máquina costumam sê-lo também para o humano, porque ambos respondem à clareza estrutural, à especificidade fática e à lógica, ainda que os modelos operem com suporte probabilístico, não determinístico. Propõe-se tratar a redação para além dos planos sintático e semântico, com relevo para o pragmático, em que o uso dos termos dialoga com a engenharia algorítmica. A escrita amadora deixa de contaminar apenas o processamento cognitivo humano (heurísticas e vieses) e passa a afetar o leitor algorítmico, suscetível à má captura da mensagem e à ancoragem de fatos em premissas mal compreendidas.

Como uma LLM lê uma peça jurídica

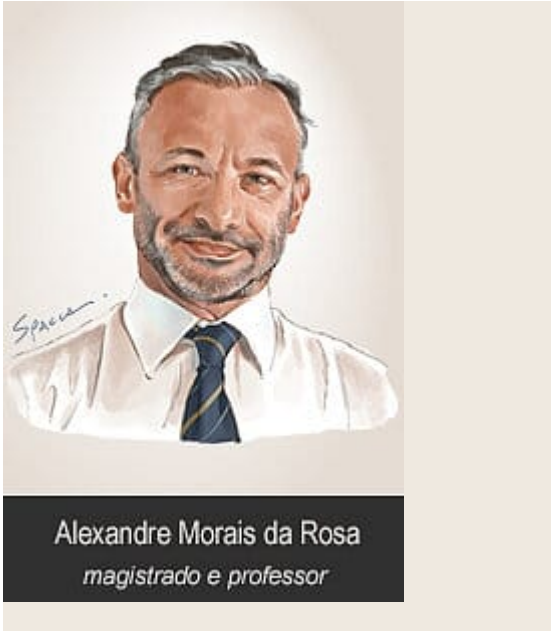
Modelos de linguagem não leem como humanos: processam o texto em camadas. O texto é fracionado em tokens (cerca de 70% de uma palavra), que recebem posição e peso de atenção (arquitetura Transformer), determinando a influência na representação do espaço vetorial semântico, com distribuição variável. Daí o fenômeno *lost in the middle*: a atenção desenha curva em U invertido, com picos no início e no fim e vale no centro. Argumentos no meio de uma peça longa têm peso desproporcionalmente menor na geração de resumos.

Cada palavra, frase e parágrafo converte-se em vetor de alta dimensionalidade, e conceitos semelhantes ocupam regiões próximas. Os termos arguido, réu, acusado e infrator não são sinônimos técnicos, mas pertencem à mesma vizinhança vetorial, com coordenadas que carregam pesos diversos de positividade ou negatividade. A escolha de uma palavra ativa região específica do mapa: “meu constituinte” em vez de “o acusado” desloca a representação para zona em que a humanidade da pessoa se torna mais saliente que sua condição processual.

Solicitado o resumo, o modelo não relê o documento: consulta a representação latente já formada e extrai os pontos de maior peso atencional. Frases bem-posicionadas, com alta saliência sintática e alinhadas ao tema global percebido, têm maior probabilidade de constar do resumo. O dado reposiciona o redator, que além de produtor de sentido atua como arquiteto de saliência. Em síntese: LLMs leem com atenção desigual, processam palavras como vetores e geram resumos a partir de representação compacta. Quem escreve para esse leitor controla três variáveis: onde alocar o que importa, quais palavras ativar e qual tema global o documento projeta.

Estratégias semânticas, em ordem de impacto

Spacca



As estratégias seguem ordenadas pela relação entre impacto persuasivo e facilidade de aplicação: as primeiras rendem mais com menor adaptação; as últimas são refinamentos.

Frame inicial (a premissa que o modelo adota)

Os primeiros parágrafos operam como instrução implícita: definem a premissa sob a qual o restante será lido. A tradição forense ensina a abrir a peça reafirmando a denúncia, sob pretexto de situar o julgador. Para o leitor humano, convenção; para o modelo, o procedimento estabelece a narrativa acusatória como ponto de partida da verdade, processando o que segue como tentativa de contestação. A inversão coloca a hipótese a ser examinada já nas primeiras linhas, em vez de reproduzir a tese de qualquer das partes. O custo é nulo; o ganho, estrutural.

Exemplo — abertura de resposta à acusação:

? Versão fraca: *Conforme narra a denúncia, o acusado teria, no dia 15 de março, mediante grave ameaça, subtraído bens da vítima, conduta tipificada no artigo 157 do CP. No entanto, a hipótese acusatória não ocorreu na forma narrada.*

? **Versão otimizada:** *A presente peça demonstrará, com base no acervo probatório, que a imputação não encontra sustentação fática nem jurídica. Três pontos: (a) a ausência de prova da materialidade; (b) a fragilidade da identificação do autor; (c) a contradição entre o depoimento da suposta vítima e as imagens do CFTV.*

Distribuição de argumentos conforme a curva de atenção

Argumento forte no meio perde tração cognitiva. Regra prática: os argumentos determinantes no primeiro e no último terço; o meio carrega o suporte (provas, doutrina, jurisprudência, transcrições). Cada seção é, internamente, um minidocumento sujeito ao mesmo viés, e deve abrir e fechar afirmando a tese, com a técnica no corpo.

***Reframing* temático**

Modelos extraem um tema latente que influencia o resumo mais do que qualquer parágrafo isolado. Peças elaboradas para leitura exclusivamente humana projetam, por inércia, tema vinculado ao crime imputado. A pergunta de controle: se um modelo classificasse este documento em uma frase, qual seria? Resposta que nomeie o crime indica que o frame permanece o da acusação. O objetivo é deslocar o tema percebido para o campo técnico-jurídico.

Sentenças-síntese pré-fabricadas

Ao resumir, modelos extraem frases inteiras do original — proposições autossuficientes que funcionam como sumários internos. O redator pode situá-las ao final de cada seção e na conclusão. Regra prática: de três a cinco por peça, como pontos de extração óbvia; o recurso recomendado é o box destacado, que isola a frase como bloco extraível.

Exemplo — sentença-síntese:

? Versão fraca: *A defesa discordou dos argumentos do Ministério Público, demonstrando que houve diversos problemas na apuração policial e que a prova é frágil.*

? **Versão otimizada:** *Em síntese: a busca pessoal foi realizada sem fundada suspeita, sem testemunhas e sem registro audiovisual, configurando nulidade absoluta nos termos do artigo 157 do CPP — ponto que, isoladamente, impõe a absolvição.*

Escolha dos termos

Cada palavra carrega coordenadas no espaço vetorial. Réu, criminoso, conduta delitiva, vítima, violência e dolo ativam regiões associadas a contextos condenatórios; o uso reiterado desloca o documento para zona em que a culpa é estado padrão. A virada lexical é cirúrgica: em vez de "réu", "o Sr. [Nome]" ou "meu constituinte"; em vez de "conduta criminosa", "conduta imputada"; em vez de "não houve violência" (que ainda ativa o vetor violência), descreve-se positivamente o ocorrido — a interação foi exclusivamente verbal. O nome do constituinte não deve figurar próximo aos verbos da conduta; o sujeito das frases negativas há de ser a prova fraca, a narrativa contraditória, a inconsistência probatória.

Especificidade factual e ancoragem quantitativa

Modelos calibram confiança por especificidade: afirmações vagas viram opinião; específicas, fato verificável. Converter adjetivos em dados quebra inferências causais implícitas. Ao registrar "às 22h15 a vítima foi encontrada ferida, e às 22h47 o investigado ingressou no local — 37 minutos após o ocorrido (ev. 3)", insere-se entre dois fatos um dado quantitativo que torna estatisticamente improvável a inferência causal automática.

Hierarquia estrutural e títulos como teses

LLMs, treinadas em documentos estruturados (Markdown, HTML, LaTeX), têm forte capacidade de extração hierárquica: títulos, subtítulos e primeiras frases recebem peso atencional ampliado, e o modelo com frequência resume um documento longo lendo apenas títulos e frases iniciais. Títulos genéricos como "Dos Fatos" ou "Do Direito" desperdiçam saliência. Cada título pode ser proposição completa e autossuficiente. Regra estética: no máximo três níveis; negrito com parcimônia, três a cinco destaques por página.

Exemplo — títulos como teses:

? Versão fraca: *II – Dos Fatos. III – Do Direito. IV – Do Pedido.*

? **Versão otimizada:** *II – A Acusação Sustenta-se em Prova Oral Isolada e Contraditória. III – Inexiste Dolo: a Conduta Foi Praticada em Erro de Tipo Essencial. IV – A Absolvição é a Única Conclusão Compatível com o Acervo Probatório.*

Pré-refutação de objeções esperadas

Ao gerar resumos ou pareceres, o modelo simula objeções no plano latente; as não respondidas permanecem abertas no resumo, identificadas como fragilidades. A solução inclui seção explícita que antecipa as principais objeções e as responde, condicionando o tratamento dos contra-argumentos. É a versão jurídica da prolepse: antecipa-se a objeção, nomeia-se, responde-se. O efeito sobre o leitor humano sempre foi conhecido; sobre o algorítmico, é agora igualmente mensurável.

Exemplo — pré-refutação:

? Versão fraca: *A peça apenas afirma suas teses, sem dialogar com a possível resposta da parte contrária. O modelo, ao resumir, pode preencher a lacuna com a versão esperada.*

? **Versão otimizada:** *Poder-se-ia objetar que a confissão informal narrada pelo policial supriria a ausência de outras provas. Tal argumento esbarra em três obstáculos: (a) a confissão extrajudicial sem corroboração não sustenta condenação (STF, HC 174.099); (b) o policial admitiu em juízo (fl. 198) que não houve registro audiovisual; (c) o investigado negou a confissão na primeira oportunidade processual (fl. 76).*

Containers visuais: quadros, tabelas e timelines

Sem violar o formalismo processual, é possível incorporar três containers de alta densidade informacional: o quadro-síntese de teses (no início), a tabela comparativa de provas (no meio) e a linha do tempo (em disputa cronológica). Julgadores conservadores aderem a eles quando bem executados, e LLMs os processam com fidelidade quase perfeita. Cuidado prático: cores não devem carregar informação; tabelas complexas devem ter a informação crítica replicada em prosa; o conteúdo argumentativo deve ser texto, não imagem, sob risco de não sobreviver à ausência de OCR.

Pedido como cadeia de raciocínio explícita

O final do documento é lido com altíssima atenção. O pedido, tratado pela tradição como mero protocolo, é, para a LLM, o output esperado da cadeia de raciocínio. Estruturá-lo como conclusão inevitável dos argumentos fecha o ciclo persuasivo. Cada item deve referenciar a seção que o sustenta, recapitulando a estrutura nos parágrafos finais, onde o viés de recência amplifica a atenção.

Síntese operacional: as dez estratégias

Estratégia	Efeito principal
1. Frame inicial	Define a premissa de leitura do documento em vez de reproduzir a tese de pa acusatória.
2. Distribuição de teses	Aloca o que importa onde a atenção do modelo é maior.
3. Reframing temático	Faz o documento ser classificado como discussão jurídica, não como narrativ
4. Sentenças-síntese	Planta no texto as frases que aparecerão no resumo do modelo.
5. Substituição lexical	Desloca o constituinte de regiões criminógenas do espaço vetorial.
6. Especificidade factual	Eleva a confiança do modelo nas afirmações apresentadas.
7. Títulos como teses	Transforma cada heading em ponto de extração argumentativa.
8. Pré-refutação	Condiciona o modelo sobre como tratar objeções esperadas.
9. Containers visuais	Cria blocos de alta densidade informacional parseáveis com fidelidade.
10. Pedido encadeado	Fecha a peça como conclusão lógica dos argumentos.

Por que já não se trata de estratégia meramente opcional

Argumenta-se que o juiz lê a peça pessoalmente, que o assessor é qualificado, que a IA é mero adjunto. Cada afirmação comporta resposta. Primeiro: o tempo médio de leitura por magistrado de vara criminal de grande movimento é menor do que se supõe, e em muitos gabinetes IAs são empregadas como acessório de produção — mapas mentais, resumos e insights que situam o agente humano, à maneira de quem lê o sumário antes dos capítulos.

O material derivado fornece informação filtrada que antecede e condiciona o conhecimento do julgador, ativando gatilhos, heurísticas e vieses ao longo da leitura e da votação. Segundo: ainda que o julgador leia diretamente, o uso de modelos para pesquisa, sumarização e redação está disseminado; se o modelo distorce as teses ao resumir, a decisão se constrói sobre fundamentos distorcidos. Terceiro: a tendência institucional é de aprofundamento.

O CNJ regula o uso de IA (Res. 615), tribunais desenvolvem ferramentas próprias e legaltechs ofertam triagem; a regulamentação não altera o modo como os modelos processam dados, e mesmo com modelos treinados, prompts específicos, baixa temperatura e janela ampliada, a estrutura basilar de leitura permanece. Quarto: a peça mal estruturada para o leitor algorítmico costuma sê-lo também para o humano, de modo que as estratégias coincidem com boas práticas de redação clara — sem prejuízo dos riscos das LLMs (vieses, opacidade decisória, alucinações e reforço de padrões injustos).

Conclusão

Há agentes processuais com interesses próprios, regras formais e informais, movimentos estratégicos, blefes, ganhos esperados e custos calculados. A peça é um movimento; o processo, uma sequência de atos; o julgamento, um resultado. O que se alterou é a entrada de um novo elemento no tabuleiro: o modelo de linguagem que influencia a leitura dos sentidos atribuídos no contexto processual. A teoria dos jogos aplicada ao processo penal sempre foi teoria sobre informação assimétrica: quem domina mais regras, conhece mais atalhos e antecipa mais movimentos ganha com maior frequência.

A fluência algorítmica configura assimetria informacional, exigindo adaptação ao modelo mental do julgador. Não se pretende converter o redator em programador, mas reconhecer que as regras do jogo mudaram e que o tabuleiro ganhou camada nova e mensurável. A escrita processual sempre foi exercício de transmissão controlada de informação a quem decide; hoje, parte dessa decisão se apoia em informação intermediada por filtro algorítmico. Além de saber o que dizer, importa saber como transmitir e onde posicionar o argumento.

Fonte: https://conjur.jumps.com.br/2026-jun-12/escrever-pecas-para-humanos-e-maquinas-processo-penal/