

Nexum AI: Plataforma de Inteligência Artificial e Busca Híbrida para Gestão de Litigância em Massa no Setor Público

Rhedson Esashika¹ ,
Carlos M. S. Figueiredo¹ , Tiago de Melo¹ 

¹PPGEEL - Postgraduate Program in Electrical Engineering.
State University of Amazonas (UEA), Manaus, AM, Brazil

{rffe.mee25, cfigueiredo, tmelo}@uea.edu.br

Abstract. *This paper presents an Artificial Intelligence platform to support the management of mass litigation in the public sector. The proposed approach integrates hybrid information retrieval, semantic clustering, and Large Language Models to assist in identifying similar cases, organizing recurring litigation patterns, and supporting the drafting of legal documents. The platform combines lexical and vector search with structured extraction of central elements, such as facts, legal grounds, and claims. In addition, it adopts a human-in-the-loop strategy, preserving the role of judges and court staff in reviewing and validating the generated content. A preliminary evaluation with five analysts indicated a positive perception regarding similar case identification, draft standardization, and workflow organization. Overall, the proposal seeks to contribute to greater efficiency, standardization, and analytical support in repetitive judicial workflows.*

Resumo. *Este artigo apresenta uma plataforma de Inteligência Artificial para apoio à gestão da litigância em massa no setor público. A abordagem proposta integra recuperação híbrida de informação, agrupamento semântico e Modelos de Linguagem de Grande Porte para auxiliar a identificação de processos semelhantes, a organização de padrões recorrentes de litigância e a elaboração assistida de documentos jurídicos. A plataforma combina busca lexical e vetorial com extração estruturada de elementos centrais, como fatos, fundamentos jurídicos e pedidos. Além disso, adota uma estratégia human-in-the-loop, preservando o papel de magistrados e servidores na revisão e validação do conteúdo gerado. Uma avaliação exploratória preliminar com cinco analistas judiciários indicou percepção positiva sobre identificação de casos similares, padronização de minutas e organização do fluxo de trabalho. De modo geral, a proposta busca contribuir para maior eficiência, padronização e suporte analítico em fluxos judiciais repetitivos.*

1. Introdução

A crescente digitalização do Poder Judiciário, com 99,7% de virtualização dos novos processos, ampliou significativamente a disponibilidade de documentos eletrônicos, mas também evidenciou desafios no tratamento de grandes volumes de informação [CNJ Justiça em Números 2025]. Em um cenário com 83,8 milhões de causas em tramitação e forte concentração de demandas repetitivas [CNJ 2025], magistrados e

equipes precisam gerir acervos extensos cujos processos frequentemente compartilham estruturas fáticas e fundamentos jurídicos análogos. Nesse contexto, atividades como localizar precedentes internos e organizar padrões de repetitividade tornam-se operacionalmente custosas e cognitivamente exigentes.

Nos últimos anos, técnicas de Inteligência Artificial têm sido progressivamente incorporadas ao domínio jurídico com o objetivo de apoiar tarefas como recuperação de informação, classificação documental, sumarização, extração estruturada de conteúdo e geração textual. Em particular, os Grandes Modelos de Linguagem (*Large Language Models* – LLMs) têm demonstrado potencial para ampliar a capacidade de análise e síntese de documentos jurídicos [CNJ 2024]. Ainda assim, a literatura aponta limitações importantes relacionadas à confiabilidade, à ocorrência de alucinações e à necessidade de mecanismos de validação humana em aplicações sensíveis. Nesse contexto, abordagens baseadas apenas em geração textual ou em busca lexical tradicional mostram-se insuficientes para atender, de forma integrada, às exigências práticas de fluxos judiciais caracterizados por volume, repetitividade e necessidade de rastreabilidade.

Trabalhos recentes destacam a relevância da recuperação semântica, de mecanismos explicáveis de correspondência entre consultas e documentos jurídicos e do agrupamento semântico de grandes acervos. Sistemas como *LUMI* e abordagens como *JuriDiscovery* indicam que a combinação entre similaridade semântica, organização estrutural e apoio à interpretação humana é promissora para aplicações jurídicas robustas em contexto real [Wehnert et al. 2024, Brandizzi and Rodrigues 2025].

Apesar desses avanços, ainda são relativamente escassos os trabalhos que integram, em um mesmo fluxo, a recuperação híbrida de processos correlatos, o agrupamento de demandas repetitivas, a geração assistida de minutas e uma camada explícita de validação *human-in-the-loop*. Essa lacuna é particularmente relevante no setor público, em que a automação deve ser concebida como instrumento de apoio à decisão, e não como substituição da atuação interpretativa e institucional dos agentes humanos. A necessidade, portanto, não é apenas recuperar ou gerar texto, mas estruturar um processo de apoio que amplie eficiência e uniformização sem comprometer controle, transparência e responsabilidade decisória.

Diante desse contexto, este trabalho apresenta a plataforma *Nexum AI*, concebida para apoiar a gestão da litigância em massa por meio da integração entre processamento documental, busca híbrida, agrupamento semântico e geração assistida de peças jurídicas. A solução extrai informações estruturadas, recupera processos semelhantes, organiza grupos tematicamente coesos e oferece mecanismos de auditoria e comparação, mantendo a revisão humana como etapa central do fluxo.

Como contribuições, este artigo: (i) apresenta uma arquitetura aplicada de apoio à litigância repetitiva no setor público; (ii) integra busca híbrida, *clustering* textual não supervisionado e geração assistida por LLMs em um fluxo unificado; (iii) explicita o papel de mecanismos de auditoria e validação humana na utilização prática dessas técnicas; e (iv) discute o potencial de uso da plataforma em cenários judiciais marcados por elevado volume processual e recorrência temática.

O restante deste artigo está organizado da seguinte forma. A Seção 2 apresenta os trabalhos relacionados. Em seguida, a Seção 3 descreve a arquitetura da plataforma

e suas principais etapas de funcionamento. A Seção 4 discute o cenário de aplicação, os potenciais ganhos e as limitações da abordagem. Por fim, a Seção 5 apresenta as conclusões e os trabalhos futuros.

2. Trabalhos Relacionados

A aplicação de técnicas de Inteligência Artificial ao domínio jurídico tem evoluído rapidamente, sobretudo com a convergência entre recuperação semântica de informação, grandes modelos de linguagem (LLMs), organização não supervisionada de documentos e mecanismos de validação humana. Uma revisão recente da literatura mostra que LLMs vêm sendo empregados de forma crescente em tarefas como sumarização, classificação e recuperação de informação jurídica, com destaque para modelos como GPT-4 [OpenAI et al. 2024], BERT [Devlin et al. 2019] e Llama 2 [Touvron et al. 2023]. Ao mesmo tempo, a literatura aponta limitações relevantes, como vieses, alucinações e variações de confiabilidade, indicando que o uso dessas técnicas em fluxos jurídicos sensíveis exige mecanismos adicionais de controle, auditoria e validação [Belarmino et al. 2025].

No eixo da recuperação jurídica, estudos buscam superar métodos puramente lexicais por meio de representações semânticas e mecanismos explicáveis de correspondência entre consulta e documento. O sistema *LUMI*, por exemplo, destaca trechos relevantes em documentos recuperados, favorecendo a interpretação da similaridade [Wehnert et al. 2024]. De modo complementar, abordagens de raciocínio por casos (*case-based reasoning*) seguem relevantes por apoiar decisões a partir de precedentes semelhantes, preservando transparência e inteligibilidade para o operador do direito [Nouwens et al. 2024].

Mais recentemente, a literatura tem convergido para arquiteturas de *hybrid search*, nas quais sinais lexicais, semânticos e estruturais são combinados em uma única *pipeline* de recuperação. O levantamento realizado indica que trabalhos recentes no contexto do COLIEE [COLIEE 2026] defendem explicitamente a combinação entre recuperação da informação (*Information Retrieval* – IR) tradicional, busca vetorial, reformulação semântica da consulta e uso de metadados jurídicos para re-ranqueamento, reduzindo a fragmentação entre pesquisa textual e contexto institucional. Na mesma direção, a pipeline *TraceRetriever* combina BM25, busca vetorial por bi-encoders e etapas de *re-ranking* por *cross-encoders*. Além disso, explora papéis retóricos para ponderar diferentes partes dos documentos recuperados. Em conjunto, esses estudos sugerem que, em aplicações jurídicas reais, a recuperação mais robusta tende a emergir da integração entre correspondência lexical, similaridade semântica e estrutura argumentativa [Wehnert et al. 2025, Nigam et al. 2025a].

Outro eixo relevante é o agrupamento não supervisionado de documentos, especialmente em contextos de litigância repetitiva, nos quais é necessário identificar grupos recorrentes de processos, temas e padrões argumentativos. O *JuriDiscovery* demonstra essa possibilidade ao combinar *topic modeling*, BERTopic [Grootendorst 2022] e refinamento de rótulos por LLMs para organizar milhares de decisões judiciais em categorias interpretáveis [Brandizzi and Rodrigues 2025]. Estudos com petições iniciais brasileiras também indicam que a combinação entre extração factual, embeddings jurídicos, UMAP e HDBSCAN pode reduzir ruídos e aumentar a coesão dos agrupa-

mentos [Esashika et al. 2026]. Essa perspectiva é aderente à triagem, gestão de acervo e elaboração padronizada de minutas em demandas repetitivas.

No que se refere à estrutura interna dos documentos jurídicos, a literatura também tem avançado no reconhecimento automático de componentes argumentativos e papéis retóricos. O levantamento complementar indica que o modelo MARRO explora *Multi-Head Attention* para rotular papéis retóricos em documentos legais, permitindo distinguir segmentos como fatos, fundamentos, análise e decisão [Bambroo et al. 2025]. Esse tipo de modelagem é especialmente importante para aplicações que, como a aqui tratada, operam sobre campos jurídicos diferenciados (fatos, fundamentos e pedidos, por exemplo) e dependem de uma representação mais fina da topologia argumentativa do processo. Em complemento, estudos recentes no contexto brasileiro têm explorado LLMs para extrair informações estruturadas de petições iniciais, organizando automaticamente elementos como partes, fatos, fundamentos jurídicos e pedidos em formatos formais de saída, o que reforça o potencial desses modelos para transformar documentos processuais não estruturados em dados jurídicos mais tratáveis analiticamente [Esashika et al. 2025]. De forma semelhante, trabalhos de extração de conteúdo normativo com LLMs mostram que esses modelos podem automatizar a identificação de enunciados normativos em textos jurídicos, embora ainda apresentem limitações, como erros de *parsing* e alucinação, o que reforça a necessidade de estratégias de checagem posterior [Zin et al. 2024].

No eixo da geração automática, a literatura recente tem buscado sair do paradigma de geração genérica de texto e avançar para a produção jurídica estruturada e controlada. Evidencia-se uma abordagem *wrapper* agnóstica ao modelo base para a geração estruturada de documentos jurídicos, denominado *VidhikDastaavej*, que enfatiza a organização formal da peça e o controle da saída textual [Nigam et al. 2025b]. Essa direção é coerente com aplicações de apoio à elaboração de decisões, nas quais a qualidade da geração não depende apenas da fluência do texto, mas também da aderência à estrutura esperada do documento judicial. Em paralelo, estudos sobre geração argumentativa multiagente têm proposto mecanismos reflexivos para mitigar manipulação e reforçar persuasão e consistência do texto gerado, aproximando-se de uma lógica de auditoria interna do próprio sistema [Zhang and Ashley 2025].

Apesar desses avanços, a literatura ainda revela uma lacuna importante: poucos trabalhos integram, em um fluxo unificado, a recuperação híbrida de jurisprudência, o agrupamento de processos, a geração estruturada de minutas e a validação humana no laço. Em geral, os estudos concentram-se em camadas isoladas, tais como recuperação, classificação, extração, sumarização ou geração, sem tratar de forma sistêmica o ciclo completo de apoio à decisão judicial. Assim, o presente trabalho posiciona-se nessa interseção, propondo uma arquitetura que combina busca híbrida sobre documentos processuais, agrupamento de casos repetitivos, geração assistida de peças e um mecanismo de revisão *human-in-the-loop*, buscando maior utilidade prática, rastreabilidade e confiabilidade no contexto do Judiciário.

3. Arquitetura do Sistema

3.1. Visão geral da plataforma

A solução proposta neste trabalho consiste na plataforma *Nexum AI*, desenvolvida para apoiar a análise de documentos processuais e a elaboração de decisões judiciais em con-

textos caracterizados por elevada repetitividade de demandas. A ferramenta foi concebida para auxiliar magistrados e equipes de apoio na organização, comparação e tratamento de processos com características semelhantes, contribuindo para a redução de tarefas operacionais e para a racionalização do fluxo de trabalho. De modo geral, a plataforma processa documentos jurídicos, extrai informações estruturadas do caso e emprega mecanismos de indexação lexical e vetorial para recuperar processos correlatos e organizar padrões de repetitividade processual.

Do ponto de vista funcional, a arquitetura integra técnicas de inteligência artificial ao domínio jurídico por meio da combinação entre extração estruturada de conteúdo, representações vetoriais de texto, métodos de agrupamento não supervisionado e LLMs voltados à geração e auditoria de minutas. Além disso, a plataforma oferece suporte ao processamento em lote e à análise comparativa de documentos, favorecendo a uniformização de procedimentos em unidades com grande volume processual. A proposta adota uma abordagem na qual toda sugestão produzida pelo sistema permanece sujeita à supervisão, validação e eventual ajuste por parte do usuário humano, preservando o controle decisório e o papel interpretativo do julgador.

Em termos de implementação, a prova de conceito foi instanciada com uma API RESTful em FastAPI, responsável pela orquestração do fluxo e pela comunicação entre os módulos. O motor de busca híbrida foi implementado sobre o Elasticsearch, enquanto a interface do usuário, desenvolvida em React, oferece um ambiente de consulta, comparação e validação. Nesta versão, a solução materializa os módulos centrais da arquitetura como demonstração funcional, e não como avaliação experimental conclusiva de desempenho.

3.2. Pipeline de processamento dos documentos processuais

O *pipeline* de processamento foi concebido para transformar documentos processuais brutos em uma base estruturada apta a alimentar as etapas de recuperação, agrupamento e geração assistida. Inicialmente, os arquivos são carregados pela plataforma e submetidos a uma etapa de extração de informação orientada à identificação de campos relevantes do processo, com ênfase na narrativa fática, nos fundamentos jurídicos e nos pedidos formulados. Essa etapa pode combinar instruções estruturadas e procedimentos de processamento textual voltados à identificação e organização de campos juridicamente relevantes do documento.

A Figura 1 apresenta uma visão simplificada e linear das etapas principais de processamento, recuperação, geração e validação.

Após a extração, os conteúdos estruturados são convertidos em representações vetoriais (*embeddings*), preservando informações semânticas relevantes para a comparação entre casos. Na prova de conceito, a representação vetorial foi priorizada sobre os campos juridicamente mais informativos — especialmente fatos, fundamentos e pedidos —, pois esses elementos concentram a narrativa do conflito e reduzem ruídos decorrentes de cabeçalhos, citações processuais e argumentação repetitiva. Como referência metodológica, foram considerados modelos de *embeddings* jurídicos e gerais previamente avaliados em experimentos correlatos, incluindo Legal-BERT, text-embedding-3-large e gemini-embedding-001, com normalização vetorial e comparação por similaridade cosseno. Essa escolha favorece a recuperação de processos semanticamente próximos

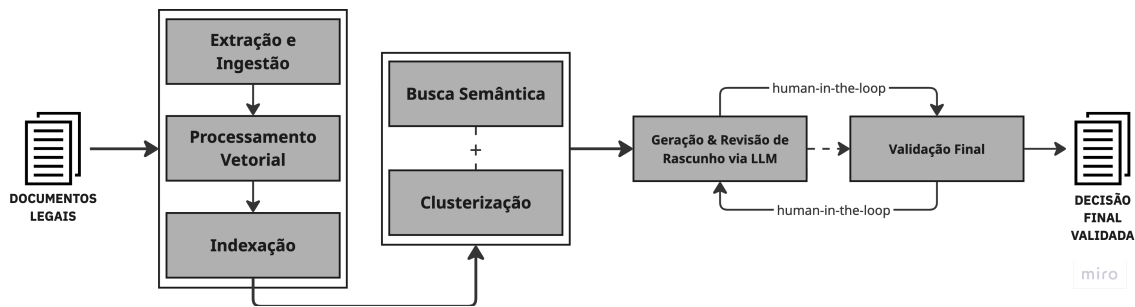


Figura 1. Fluxograma simplificado da Arquitetura do Sistema Nexum AI.

mesmo quando redigidos com vocabulário distinto.

3.3. Busca híbrida e recuperação de processos similares

Com a representação textual e vetorial estabelecida na etapa anterior, a plataforma executa a recuperação de processos semelhantes por meio de uma estratégia de busca híbrida, combinando sinais lexicais e semânticos na localização de documentos correlatos. Diferentemente de abordagens tradicionais baseadas apenas em palavras-chave, esse método busca capturar o contexto e o significado global da peça jurídica, permitindo identificar documentos relacionados mesmo quando redigidos com variações terminológicas ou estruturais profundas.

Em termos operacionais, os campos textuais estruturados permanecem indexados para consulta lexical, enquanto os *embeddings* associados aos documentos são utilizados para estimar a similaridade semântica no espaço vetorial. Desse modo, a recuperação não depende exclusivamente de coincidências literais de vocabulário, mas também da proximidade de conteúdo entre os casos analisados. A recuperação híbrida combina dois sinais complementares. O primeiro é lexical, baseado na correspondência textual entre termos da consulta e campos indexados do processo, favorecendo a recuperação de documentos com aderência terminológica explícita. O segundo é vetorial, baseado na similaridade semântica entre o caso analisado e os documentos previamente indexados. Na arquitetura proposta, esses sinais são utilizados de forma complementar: a busca lexical contribui para preservar precisão terminológica e rastreabilidade, enquanto a busca vetorial amplia a recuperação de casos semanticamente semelhantes mesmo quando há variação vocabular. A lista final de processos correlatos é então apresentada ao usuário como apoio à comparação e à validação humana, e não como conclusão automática de identidade entre casos.

Essa estratégia é especialmente relevante no contexto judicial, no qual demandas materialmente semelhantes podem apresentar diferenças de estilo argumentativo, estrutura textual ou terminologia empregada pelas partes. Ao recuperar processos correlatos com base em similaridade semântica e correspondência lexical, a plataforma fornece subsídios relevantes para a identificação de precedentes internos da vara ou tribunal, facilitando a comparação entre fundamentos já adotados e favorecendo maior uniformidade no tratamento de casos repetitivos.

3.4. Agrupamento de processos repetitivos

Como desdobramento natural da etapa de recuperação, a arquitetura emprega agrupamento não supervisionado sobre as representações vetoriais dos documentos, tomando a similaridade semântica entre *embeddings* como critério de proximidade para formar conjuntos tematicamente coerentes. Na prova de conceito, o agrupamento foi concebido como uma etapa não supervisionada baseada na proximidade semântica entre *embeddings* dos documentos. Para esse fim, foi adotada uma estratégia compatível com redução de dimensionalidade por UMAP e agrupamento por densidade com HDBSCAN, abordagem adequada a corpora jurídicos por não exigir a definição prévia do número de grupos e por permitir isolar documentos atípicos como ruído. Essa configuração é especialmente útil em litigância repetitiva, pois os grupos podem emergir a partir da concentração semântica dos fatos narrados, em vez de depender de categorias previamente rotuladas.

Em termos operacionais, o agrupamento é orientado pela proximidade entre as representações vetoriais dos documentos, possibilitando reunir processos em *clusters* coerentes sob a perspectiva fática e jurídica. Essa lógica permite organizar lotes de demandas com alto grau de semelhança, o que é particularmente útil em contextos de litigância de massa, nos quais a identificação de recorrências constitui etapa relevante para triagem, comparação e tratamento coordenado do acervo.

Essa funcionalidade pode oferecer utilidade prática imediata em unidades submetidas a litígios repetitivos. Ao agrupar automaticamente documentos correlatos, o sistema pode favorecer uma triagem processual mais célere e preparar o terreno para fluxos de trabalho orientados ao processamento em lote. O *clustering* atua, assim, como um mecanismo de apoio à análise coordenada, sempre preservando a possibilidade de validação, ajuste ou reconfiguração dos grupos sugeridos por parte do usuário humano.

3.5. Geração e auditoria de minutas jurídicas com LLMs

Consolidados os agrupamentos e recuperados os precedentes estruturais, a plataforma utiliza LLMs para apoiar a elaboração de minutas jurídicas. Na prova de conceito, a camada generativa foi estruturada sobre a API do Google Gemini, empregada em tarefas de extração, geração assistida e auditoria textual. As chamadas ao modelo utilizam *prompts* estruturados e baixa aleatoriedade, buscando respostas mais aderentes aos dados do processo. A saída, contudo, é tratada apenas como minuta preliminar, sujeita à validação, correção e aprovação final por magistrados ou servidores autorizados.

Além da geração textual proativa, a solução incorpora uma etapa de auditoria automatizada e análise comparativa no próprio módulo de *LLM Draft Generation & Review*, representado na Figura 1. Nesse estágio, instruções específicas de geração e revisão orientam tanto a produção inicial da minuta quanto uma verificação subsequente, na qual são confrontados os elementos estruturados extraídos do caso, o conteúdo da minuta produzida e os precedentes associados ao *cluster*. Com isso, a plataforma busca examinar a aderência fática e argumentativa da saída gerada, com o objetivo de mitigar inconsistências e reduzir o risco de alucinações.

Na interface da aplicação, conforme ilustra a Figura 2, essa auditoria se materializa em painéis comparativos, *sidebars* e modais de inspeção, que destacam visualmente pontos de convergência e divergência entre o caso atual, os documentos correlatos e o texto sugerido pelo modelo. Essa transparência algorítmica busca favorecer, na prática,

a revisão crítica e a validação final pelo usuário humano, reforçando a lógica de apoio à decisão que orienta toda a arquitetura proposta.

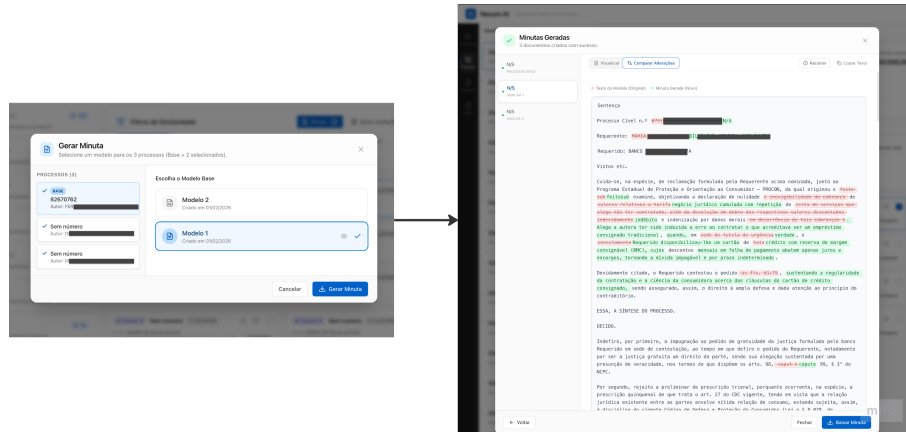


Figura 2. Módulo de geração assistida de minuta jurídica e comparação entre texto de referência e versão produzida pelo modelo.

4. Cenário de Aplicação e Discussão

4.1. Contexto de aplicação

Como forma de demonstrar a viabilidade da arquitetura proposta, esta seção discute o cenário prático de uso da plataforma *Nexum AI* e seus impactos qualitativos esperados, antecedendo futuras avaliações empíricas de desempenho. A aplicação situa-se no contexto crítico de unidades judiciais marcadas pela presença de demandas repetitivas, nas quais processos distintos compartilham estruturas fáticas semelhantes, fundamentos jurídicos recorrentes e padrões decisórios frequentemente reiterados. Em cenários dessa natureza, a atividade jurisdicional passa a conviver com desafios relacionados não apenas ao volume processual, mas também à necessidade de organizar, comparar e tratar de maneira consistente grandes conjuntos documentais. A repetitividade, embora possa favorecer a identificação de regularidades, também amplia o esforço operacional exigido de magistrados e equipes de apoio, sobretudo em tarefas como a leitura inicial dos autos, a localização de casos correlatos, a verificação de fundamentos anteriormente adotados e a elaboração de minutas compatíveis com o histórico decisório da unidade. Nesse ambiente, torna-se essencial o emprego de instrumentos computacionais capazes de apoiar a recuperação de informações, a organização do acervo e a condução de fluxos de trabalho mais estruturados, sem afastar a centralidade da supervisão humana no processo decisório.

4.2. Cenário de uso da plataforma

No cenário de uso proposto, a *Nexum AI* integra-se às rotinas de análise processual como ferramenta de apoio à triagem documental, à recuperação de processos semelhantes e à elaboração assistida de textos jurídicos. Sua utilização é particularmente pertinente em contextos nos quais a identificação de similaridades entre casos representa uma etapa relevante para o tratamento coordenado de demandas, seja para fins de comparação entre autos, seja para subsidiar a redação de minutas em processos de alta recorrência temática. A plataforma é empregada, portanto, como meio de suporte à organização das informações

processuais e à construção de uma base de consulta operacionalmente útil à atividade jurisdicional.

A dinâmica de uso da solução pressupõe o recebimento e o processamento dos documentos das ações, sua transformação em representações adequadas para a indexação semântica e a posterior recuperação de casos correlatos, que podem, então, ser agrupados de acordo com seu grau de proximidade. A partir dessa base, o sistema oferece subsídios para a comparação entre demandas, para a inspeção de casos semelhantes e para a geração inicial de textos juridicamente contextualizados, como ilustrado na Figura 3.

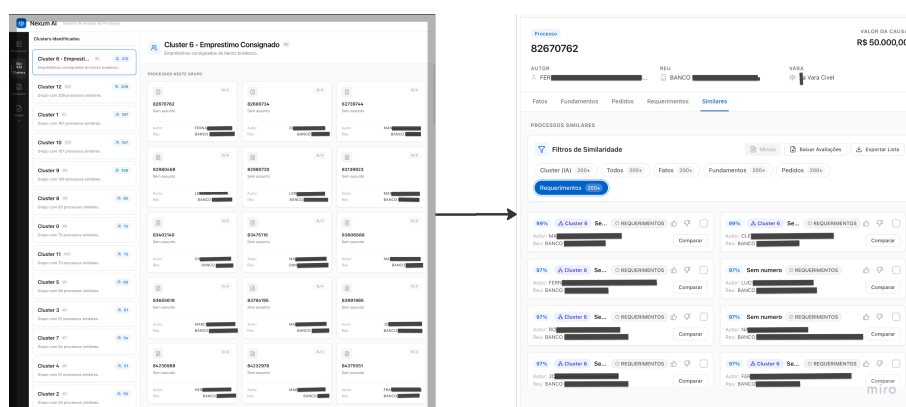


Figura 3. Interface da plataforma para agrupamento de processos e recuperação de casos similares com filtros de similaridade.

4.3. Avaliação preliminar do uso da plataforma

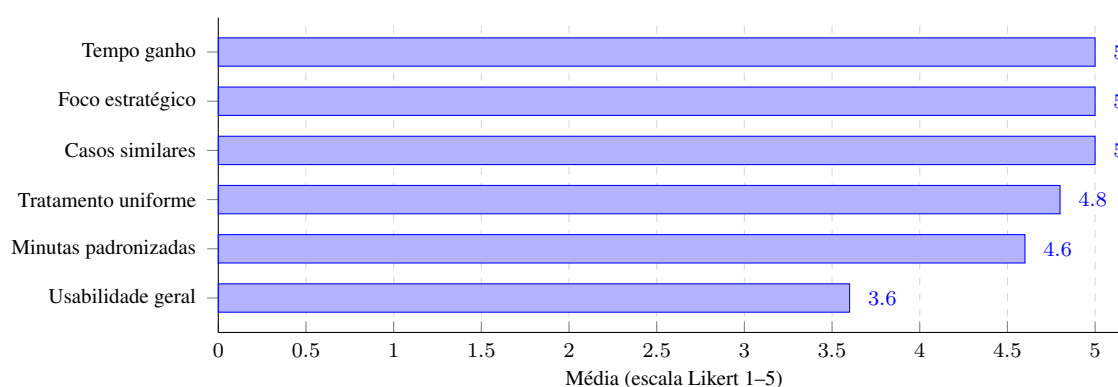


Figura 4. Percepção dos usuários sobre efeitos da Nexum AI no fluxo de trabalho.

Como complemento ao cenário de aplicação, foi realizado um teste exploratório da *Nexum AI* com cinco analistas judiciários atuantes em varas do Tribunal de Justiça do Amazonas. A avaliação utilizou formulário com escala Likert de 1 a 5 e campos abertos, contemplando ganho de tempo, foco estratégico, identificação de casos similares, uniformidade de tratamento, padronização de minutas e usabilidade geral. Embora a amostra seja reduzida, os resultados oferecem evidências iniciais sobre os efeitos percebidos da plataforma.

Conforme a Figura 4, que sintetiza as médias das respostas em escala Likert de 1 a 5, os maiores retornos percebidos concentraram-se na redução do tempo em tarefas repetitivas, no redirecionamento do esforço para atividades mais estratégicas e no agrupamento de casos semelhantes, todos com média 5,0. Também foram bem avaliadas a uniformidade de tratamento (4,8) e a padronização das minutas (4,6). O menor resultado foi o de usabilidade geral (3,6), indicando espaço para aprimoramentos de interface.

Nos relatos abertos, os principais ganhos foram associados à organização do acervo, à formação de grupos de processos e à geração de minutas em lote. Em conjunto, os achados sugerem que o principal retorno percebido da plataforma está menos em substituir a decisão humana e mais em apoiar a organização do trabalho, a padronização de rotinas e o tratamento uniforme de demandas repetitivas.

4.4. Limitações e desafios da abordagem

Apesar dos ganhos observados, a abordagem ainda apresenta limitações técnicas, normativas e institucionais que precisam ser reconhecidas. A qualidade dos resultados depende da consistência do processamento documental e da capacidade dos modelos de representar adequadamente o conteúdo semântico dos textos jurídicos. Em matéria judicial, pequenas distinções fáticas, procedimentais ou argumentativas podem ser decisivas, de modo que mecanismos de similaridade e agrupamento, embora úteis para organizar o acervo, não eliminam o risco de aproximações indevidas entre processos apenas aparentemente análogos.

Há, ainda, desafios específicos no uso de modelos de linguagem para geração e auditoria de minutas, especialmente diante das exigências de governança, transparência, auditabilidade e supervisão humana reforçadas pela Resolução CNJ nº 615/2025. Embora esses sistemas possam apoiar a redação inicial e a verificação de coerência textual, seu emprego exige cautela quanto à precisão das informações, à aderência aos autos e à compatibilidade entre o texto sugerido e os fundamentos jurídicos aplicáveis. Por isso, a plataforma deve ser compreendida como instrumento de apoio, e não como substituta da atividade interpretativa e decisória, mantendo toda saída sujeita à revisão crítica, correção e aprovação final pelo usuário humano.

Além da revisão humana, a adoção institucional de uma arquitetura dessa natureza requer mecanismos de governança, como registro das interações, versionamento de *prompts* e parâmetros, rastreabilidade das fontes utilizadas e documentação das intervenções sobre as minutas sugeridas. Também exige cuidados no tratamento de dados pessoais e sensíveis, com observância da finalidade de uso, minimização da exposição informacional, controles de acesso e respeito ao segredo de justiça. Esses elementos ampliam a transparência, favorecem a auditabilidade e reforçam a aderência da ferramenta às exigências de responsabilidade próprias do domínio jurídico.

5. Conclusão e Trabalhos Futuros

Este trabalho apresentou a plataforma *Nexum AI*, concebida para apoiar a análise de documentos processuais e a elaboração assistida de minutas em contextos judiciais caracterizados pela recorrência de demandas. A solução proposta articula processamento documental, busca híbrida, agrupamento de processos correlatos e uso de modelos de linguagem de grande porte para oferecer suporte à identificação de padrões, à recuperação

de casos semelhantes e à produção inicial de textos jurídicos. Ao reunir essas funcionalidades em um fluxo integrado, a plataforma é proposta como um instrumento de apoio à racionalização de tarefas operacionais e à organização do tratamento de demandas repetitivas, preservando a centralidade da supervisão humana na validação dos resultados e na condução da atividade jurisdicional.

A discussão ao longo do artigo indica que abordagens dessa natureza podem aumentar a eficiência, a comparabilidade e a padronização de rotinas em unidades com alto volume processual. Ao mesmo tempo, o uso de mecanismos de similaridade, agrupamento e geração textual no domínio jurídico exige cautela, dada a sensibilidade das distinções fáticas entre casos e a necessidade de aderência estrita aos autos. Assim, a plataforma deve ser vista como apoio à decisão, e não como substituta da atuação interpretativa de magistrados e servidores.

Como trabalhos futuros, pretende-se aprofundar a avaliação da plataforma em cenários reais de uso, incorporando métricas mais objetivas de desempenho relacionadas à recuperação de processos semelhantes, à qualidade das minutas geradas e aos ganhos operacionais obtidos em fluxos de trabalho com processamento em lote. Também se mostram relevantes o refinamento dos mecanismos de auditoria jurídica automatizada, a ampliação das estratégias de explicabilidade dos resultados e a investigação de formas de adaptação da solução a diferentes classes processuais e contextos institucionais. Tais desdobramentos podem contribuir para consolidar o uso de arquiteturas híbridas de inteligência artificial como suporte qualificado à atividade jurisdicional, com maior segurança, transparência e aderência às exigências do domínio jurídico.

Referências

- Bambroo, P., Adhikary, S., Bhattacharya, P., Chakraborty, A., Ghosh, S., and Ghosh, K. (2025). Marro: multi-headed attention for rhetorical role labeling in legal documents. *Artificial Intelligence and Law*, pages 1–30.
- Belarmino, M., Coelho, R., Lotufo, R., and Pereira, J. (2025). Aplicação de large language models na análise e síntese de documentos jurídicos: Uma revisão de literatura.
- Brandizzi, L. and Rodrigues, G. (2025). Juriscovery: Automated analysis of legal dispute triggered by public policies using artificial intelligence. In *Legal Knowledge and Information Systems*. IOS Press.
- CNJ (2024). O uso da inteligência artificial generativa no poder judiciário brasileiro: relatório de pesquisa. Technical report, CNJ, Brasília. Acesso em: 15 mar. 2026.
- CNJ (2025). Painel de grandes litigantes. Acesso em: 15 mar. 2026.
- CNJ Justiça em Números (2025). Justiça em números 2025: ano-base 2024. Technical report, CNJ, Brasília. Acesso em: 15 mar. 2026.
- COLIEE (2026). Overview – competition on legal information extraction/entailment 2026.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

- Esashika, R., Figueiredo, C. M., and de Melo, T. (2025). Structuring information from initial petitions using llms: A study in brazilian courts of justice. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 546–557. SBC.
- Esashika, R., Figueiredo, C. M., and de Melo, T. (2026). Discovery of legal patterns in civil petitions via llm-based fact extraction and density clustering. In *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)-Vol. 1*, pages 416–424.
- Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure.
- Nigam, S. K. et al. (2025a). Traceretriever: A hybrid pipeline integrating bm25, vector search, and cross-encoder re-ranking.
- Nigam, S. K., Patnaik, B. D., Thomas, A. V., Shallum, N., Ghosh, K., and Bhattacharya, A. (2025b). Structured legal document generation in india: A model-agnostic wrapper approach with vidhikdastaavej.
- Nouwens, J., Borg, A., and Prakken, H. (2024). An application of case-based reasoning to decision-making in dutch administrative law. *Legal Knowledge and Information Systems*, page 131.
- OpenAI, Achiam, J., Adler, S., and Agarwal, S. (2024). Gpt-4 technical report.
- Touvron, H., Martin, L., Stone, K., and Albert, P. (2023). Llama 2: Open foundation and fine-tuned chat models.
- Wehnert, S., Padmanabhan, V., and Luca, E. W. D. (2024). Lumi: Legal understanding and matching through interactive highlighting. In *Legal Knowledge and Information Systems*. IOS Press.
- Wehnert, S., Valappil, B. B. C., and De Luca, E. W. (2025). Llms, knowledge graphs, and hybrid search: Task-specific approaches to legal ai in coliee. In *Workshop of the Tenth Competition on Legal Information Extraction/Entailment (COLIEE'2025) in the 21th International Conference on Artificial Intelligence and Law (ICAIL)*.
- Zhang, L. and Ashley, K. D. (2025). Mitigating manipulation and enhancing persuasion: A reflective multi-agent approach for legal argument generation. *arXiv preprint arXiv:2506.02992*.
- Zin, M. M. et al. (2024). Leveraging llm for identification and extraction of normative statements. In *Legal Knowledge and Information Systems*. IOS Press.