JUSINSIGHT: INTELIGÊNCIA ARTIFICIAL E ARQUITETURA RAG NA GESTÃO AUTOMATIZADA DE PUBLICAÇÕES JURÍDICAS

JUSINSIGHT: Artificial Intelligence and RAG Architecture in the Automated Management of Legal Publications

RUAN PEREIRA DO NASCIMENTO
FACULDADE INBEC

WELLINGTON SOUSA AGUIAR
UNIFOR

DANIEL MATOS GOMES
FACULDADE INBEC

FABIO RODRIGUES DE LIMA
UNICESUMAR

Comunicação:
O XIV SINGEP foi realizado em conjunto com a 14th Conferência Internacional do CIK (CYRUS Institute of Knowledge) e com a SPARK (Society for Promoting Academic Research and Knowledge), em formato híbrido, com sede presencial na UNINOVE - Universidade Nove de Julho, no Brasil.

JUSINSIGHT: INTELIGÊNCIA ARTIFICIAL E ARQUITETURA RAG NA GESTÃO AUTOMATIZADA DE PUBLICAÇÕES JURÍDICAS

Objetivo do estudo
Desenvolvimento de um pipeline automatizado de extração semântica de publicações judiciais voltado a escritórios de advocacia Combinando Modelos de Linguagem de Larga Escala (LLM) com a arquitetura de Geração Aumentada por Recuperação (RAG), convertendo textos jurídicos brutos em informações estruturadas.

Relevância/originalidade
Uso das tecnologias digitais e Inteligência Artificial em grandes escritórios de advocacia, gerando inteligência e rapidez confiável nas tarefas rotineiras e previsíveis do dia a dia de um grande escritório, trazendo grande retorno financeiro e redução dos riscos judiciais.

Metodologia/abordagem
Pesquisa aplicada com abordagem qualitativa na análise das estruturas semânticas das publicações judiciais. A dimensão quantitativa também está presente na avaliação de desempenho do modelo, mensurada por meio do tempo de processamento, acurácia de extração e análise de custo-benefício operacional.

Principais resultados
A acurácia média foi de 95,3%, medida sobre 300 publicações anotadas manualmente, com inconsistências residuais causadas majoritariamente por falhas de OCR na fonte e ambiguidades sintáticas O custo de API para processar 600 publicações diárias foi inferior a R$ 1,00.

Contribuições teóricas/metodológicas
Essa pesquisa buscou desenvolver uma pesquisa multidisciplinar e aplicada, onde a computação pode trazer grandes resultados para áreas parceiras, como nesse caso, o uso da IA na redução de custos e riscos para um grande escritório de advocacia.

Contribuições sociais/para a gestão
O modelo apresentou grande retorno financeiro e de qualidade de dados, mitigando os riscos de perder prazos judiciais, usandop recursos de baixo custo. O modelo também apresenta os resultados em dashboards para a gestão estratégica do escritório.

Palavras-chave: Inteligência Artificial Jurídica, Arquitetura RAG, Modelos de Linguagem de Larga Escala, Automação Processual, Processamento de Linguagem Natural

JUSINSIGHT: Artificial Intelligence and RAG Architecture in the Automated Management of Legal Publications

Study purpose
Development of an automated pipeline for the semantic extraction of judicial publications tailored for law firms, combining Large Language Models (LLMs) with Retrieval-Augmented Generation (RAG) architecture to convert raw legal texts into structured information.

Relevance / originality
The use of digital technologies and artificial intelligence in large law firms generates intelligence and reliable speed in routine, predictable daily tasks, resulting in significant financial returns and a reduction in litigation risks.

Methodology / approach
Applied research employing a qualitative approach to analyze the semantic structures of judicial publications. A quantitative dimension is also included in the model's performance evaluation, measured by processing time, extraction accuracy, and an operational cost-benefit analysis.

Main results
The average accuracy was 95.3%, measured across 300 manually annotated publications, with residual inconsistencies caused primarily by source OCR errors and syntactic ambiguities. The API cost to process 600 publications daily was less than R$ 1.00.

Theoretical / methodological contributions
This study aimed to conduct multidisciplinary, applied research demonstrating how computing can yield significant results for partner fields—specifically, the use of AI to reduce costs and risks for a large law firm.

Social / management contributions
The model delivered significant financial returns and high data quality—mitigating the risk of missing court deadlines—while utilizing low-cost resources. It also presents results via dashboards to support the firm's strategic management.

Keywords: Legal Artificial Intelligence, RAG Architecture, Large-Scale Language Models, Procedural Automation, Natural Language Processing

JUSINSIGHT: INTELIGÊNCIA ARTIFICIAL E ARQUITETURA RAG NA GESTÃO AUTOMATIZADA DE PUBLICAÇÕES JURÍDICAS

1 Introdução

A perda de um prazo processual não é uma falha administrativa ordinária. Na advocacia, equivale à supressão de uma oportunidade jurídica concreta do cliente e pode ensejar a responsabilidade civil do profissional com fundamento na teoria da perda de uma chance, consolidada por Savi (2012) e aplicada pelo Superior Tribunal de Justiça no julgamento do Recurso Especial n. 1.637.375/SP. Nesse precedente, a Terceira Turma, sob relatoria do Ministro Ricardo Villas Bôas Cueva, manteve condenação de advogado ao pagamento de indenização por ter perdido o prazo para apresentação de embargos monitórios, reconhecendo que a omissão na observância de prazo processual equivale à privação de uma oportunidade real de obter situação mais favorável na demanda. Quando essa falha tem causa operacional prevenível, a gravidade jurídica e ética do problema se acentua consideravelmente.

Esse risco existia em escala controlável quando os processos tramitavam em papel. A digitalização do Judiciário alterou essa equação. O Conselho Nacional de Justiça registrou, em seu Relatório Justiça em Números, mais de 80 milhões de processos em tramitação no Brasil, cujas movimentações geram diariamente um volume de publicações nos Diários de Justiça Eletrônicos que ultrapassa a capacidade de triagem manual confiável de equipes de porte razoável (CNJ, 2023). Para um escritório com atuação em múltiplos tribunais, acompanhar esse fluxo sem apoio tecnológico é uma das principais fontes de risco operacional da advocacia contemporânea.

A Inteligência Artificial, em particular os Modelos de Linguagem de Larga Escala, abriu uma via concreta para automatizar a leitura e interpretação dessas publicações. Modelos como o GPT da OpenAI, o Claude da Anthropic e os da família Llama demonstraram capacidade de compreender texto jurídico em português com consistência crescente (ZHAO et al., 2023). Nunes, Lucon e Wolkart (2021) observam que a introdução de tecnologias de IA no âmbito processual não elimina a necessidade do jurista, mas redistribui seu trabalho: as atividades mecânicas de baixo valor cognitivo passam a ser mediadas por ferramentas automatizadas, enquanto o profissional concentra suas expertise em tarefas que exigem julgamento e responsabilidade.

Uma aplicação direta de LLMs genéricos a textos jurídicos, porém, esbarra em um problema documentado: a alucinação. Modelos generativos podem produzir números de processo, datas e partes internamente coerentes, porém incorretos. Num domínio em que um prazo errado pode configurar perda de uma chance nos termos de Savi (2012), esse comportamento é inadmissível como base de gestão processual. A arquitetura RAG (Retrieval-Augmented Generation), proposta por Lewis et al. (2020), resolve o problema ao ancorar cada extração em fragmentos verificáveis do documento original, tornando a saída rastreável e auditável.

Este artigo documenta uma pesquisa em finalização. O JusInsight encontra-se em estágio de prova de conceito (MVP) com testes realizados em ambiente controlado com publicações reais. O objetivo desta fase é validar a arquitetura do pipeline de extração semântica, identificar os gargalos técnicos do domínio jurídico e estabelecer métricas que orientem as versões subsequentes. As seções seguintes apresentam a fundamentação teórica (Seção 2), os trabalhos relacionados (Seção 3), a metodologia (Seção 4), a arquitetura técnica (Seção 5), os resultados com análise de viabilidade (Seção 6) e as considerações finais (Seção 7).

2 Referencial Teórico

Da Evolução Tecnológica aos Modelos de Linguagem de Larga Escala

Para situar os LLMs no quadro mais amplo da Inteligência Artificial, é útil recorrer ao modelo de quatro plataformas tecnológicas proposto por Aguiar (2019) em sua tese de doutorado sobre modelos preditivos de mortalidade infantil. Nesse enquadramento, a primeira plataforma (décadas de 1950-60) corresponde aos mainframes de acesso restrito; a segunda (1980-90), à democratização com os PCs e a internet; a terceira (a partir de 2010), à computação em nuvem, smartphones e Big Data, que criou o substrato necessário para o aprendizado de máquina em escala; e a quarta, ainda emergente, com IA generativa avançada. Os LLMs pertencem a essa terceira e quarta plataformas: seu treinamento exige poder computacional da ordem de petaflops e corpora de trilhões de palavras, recursos economicamente acessíveis apenas após 2018.

A definição mais consolidada de Inteligência Artificial na literatura é sistematizada por Aguiar (2019) como a automação do comportamento inteligente, síntese construída a partir de autores clássicos como Barr e Feigenbaum (1981), Charniak e McDermott (1987) e Kurzweil (1990). No interior desse campo, o Machine Learning representa a subárea em que o sistema aprende a partir da experiência em relação a uma classe de tarefas, melhorando seu desempenho com a prática, conforme Mitchell (1997) apud Aguiar (2019). O Deep Learning especializa essa abordagem por meio de redes neurais profundas, extraindo representações hierárquicas dos dados com destaque em visão computacional, reconhecimento de voz e Processamento de Linguagem Natural (AGUIAR, 2019).

O marco técnico para os LLMs modernos foi a publicação de Vaswani et al. (2017), que introduziu a arquitetura Transformer baseada no mecanismo de atenção. Ao capturar relações semânticas entre palavras distantes no texto com eficiência computacional, o Transformer superou as redes recorrentes que dominavam o campo. Sobre essa base surgiram o BERT (Devlin et al., 2019) e a família GPT (Brown et al., 2020), inaugurando o paradigma dos modelos de linguagem de larga escala. Zhao et al. (2023) documentam que esses modelos exibem capacidades não explicitamente programadas: raciocínio em múltiplos passos (chain-of-thought reasoning) e aprendizado a partir de exemplos no contexto (in-context learning).

Uma leitura crítica do trabalho de Aguiar (2019) revela uma distinção tecnológica relevante. Os modelos MLP (Multilayer Perceptron) utilizados por Aguiar (2019) operam sobre dados tabulares estruturados e requerem engenharia manual de características (feature engineering). Essa abordagem é eficaz para variáveis epidemiológicas numéricas, mas encontra limitações diante de textos não estruturados com alta variabilidade de formato. Os LLMs baseados em Transformer processam texto bruto de forma nativa, aprendendo representações semânticas sem pré-processamento manual de campos. A contrapartida é menor interpretabilidade interna, o que reforça a necessidade do RAG para garantir rastreabilidade. O princípio interdisciplinar que orienta os dois trabalhos é o mesmo: Aguiar (2019) demonstrou que disciplinas distintas precisam articular seus conhecimentos para que a IA opere com rigor em domínios especializados, premissa que se aplica integralmente à interseção entre Computação e Direito no JusInsight.

Arquitetura Retrieval-Augmented Generation (RAG)

O principal obstáculo ao uso de LLMs em contextos de alta responsabilidade é a alucinação: a tendência do modelo de produzir afirmações coerentes, porém factualmente incorretas. Para o direito processual, onde um prazo ou número de processos equivocado

invalida completamente a informação gerada, esse comportamento representa uma barreira técnica objetiva que precisa ser tratada arquiteturalmente.

A arquitetura RAG, proposta por Lewis et al. (2020), resolve o problema ao separar o conhecimento factual do modelo de linguagem. Em vez de depender exclusivamente do que foi armazenado nos pesos da rede durante o treinamento, o sistema RAG incorpora um componente de recuperação que, antes de cada extração, busca os fragmentos mais relevantes de uma base de conhecimento externa e os entrega ao modelo como contexto verificável. A resposta gerada baseia-se em documentos concretos, não em memória paramétrica potencialmente imprecisa.

No protótipo JusInsight, o RAG opera em quatro etapas encadeadas. Na ingestão, cada texto de publicação é transformado em vetor numérico por um modelo de embeddings e armazenado no banco vetorial com metadados identificadores. Na recuperação, o sistema realiza uma busca de similaridade por cosseno no banco vetorial para localizar os fragmentos semanticamente mais próximos da publicação em processamento. Na geração aumentada, esses fragmentos são concatenados ao prompt enviado ao LLM, que recebe instrução explícita de basear sua extração exclusivamente nos documentos fornecidos. Por fim, o sistema registra quais trechos do documento original fundamentaram cada campo extraído, criando uma cadeia de rastreabilidade completa e auditável.

Inteligência Artificial e o Direito no Brasil

O debate sobre a aplicação de IA no Judiciário brasileiro ganhou contornos normativos com a Resolução CNJ n. 332/2020, que estabelece diretrizes éticas para o uso de IA em sistemas judiciais, com ênfase em transparência, não discriminação e supervisão humana. Nunes, Lucon e Wolkart (2021) documentam que a introdução de tecnologias de IA no âmbito processual não substitui a função do jurista, mas redistribui seu trabalho: as atividades repetitivas passam a ser mediadas por ferramentas automatizadas, enquanto o profissional concentra suas expertise nas atividades de maior densidade intelectual. Esse modelo de colaboração humano-máquina é precisamente o que o JusInsight visa implementar na camada operacional da triagem de publicações, mantendo a revisão profissional como etapa inegociável do fluxo neste estágio de desenvolvimento.

A conformidade com a Lei n. 13.709/2018 (LGPD) é requisito não negociável nessa equação. Ferramentas que processam dados de partes processuais, mesmo quando provenientes de publicações de acesso público, devem observar os princípios da finalidade, adequação, necessidade e segurança. O JusInsight incorpora controles de privacidade desde as primeiras decisões de projeto, abordagem detalhada na Seção 5.4.

3 Trabalho relacionados

A aplicação de Inteligência Artificial ao sistema judiciário brasileiro possui antecedentes institucionais relevantes. O projeto VICTOR, desenvolvido pelo Supremo Tribunal Federal em parceria com a Universidade de Brasília a partir de 2018, é o caso mais documentado: a ferramenta utiliza redes neurais para classificar peças processuais e identificar automaticamente a existência de repercussão geral nos recursos extraordinários submetidos à corte, reduzindo significativamente o tempo de triagem pelos gabinetes dos ministros (ANDRADE et al., 2020). O SÓCRATES, do Tribunal de Justiça do Distrito Federal e Territórios, aplica técnicas de PLN (Processamento de Linguagem Natural) à classificação e priorização de petições. A Resolução CNJ n. 332/2020, ao estabelecer diretrizes para o uso de IA no Judiciário, reconhece expressamente a existência de múltiplos sistemas de IA em operação nos tribunais brasileiros, evidenciando a maturidade do ecossistema de tecnologia jurídica no país (CNJ, 2020).

O JusInsight diferencia-se dessas iniciativas em dois aspectos centrais. Do ponto de vista do público-alvo, as ferramentas institucionais como VICTOR e SÓCRATES foram desenvolvidas para uso interno dos tribunais, com foco em apoio à atividade judicante. O protótipo aqui descrito volta-se especificamente aos escritórios de advocacia privados, que são simultaneamente destinatários das publicações dos Diários de Justiça e responsáveis pela gestão dos prazos delas decorrentes. Do ponto de vista técnico, as iniciativas existentes de classificação judicial utilizam predominantemente modelos de aprendizado supervisionado treinados sobre categorias fechadas. O JusInsight adota a arquitetura RAG combinada com LLMs de uso geral, o que permite a extração aberta de campos estruturados a partir de texto livre, sem treinamento supervisionado sobre categorias pré-definidas e sem restrição ao conjunto de tipos documentais antecipados.

No plano acadêmico, a literatura sobre Processamento de Linguagem Natural aplicado ao português jurídico tem crescido desde 2018, com trabalhos em conferências como PROPOR e STIL abordando classificação de sentenças, extração de entidades nomeadas e sumarização de acórdãos. Esses trabalhos operam predominantemente sobre textos de decisões já proferidas para fins de análise jurisprudencial. A extração estruturada de publicações de Diários de Justiça Eletrônicos com cálculo de prazos processuais, que é o objeto do JusInsight, representa uma tarefa técnica distinta: opera sobre textos de natureza notificatória, com heterogeneidade de formatação entre tribunais e com requisito explícito de rastreabilidade para fins de responsabilidade profissional do advogado.

4 Metodologia

A pesquisa classifica-se como aplicada, pois tem por objetivo gerar conhecimento orientado à solução de um problema concreto: o risco operacional da triagem manual de publicações judiciais em escritórios de advocacia (GIL, 2008). Por se tratar de uma investigação em andamento, este artigo documenta os resultados parciais obtidos até a fase de prova de conceito. As etapas subsequentes, detalhadas na Seção 7, preveem validação em ambiente produtivo, expansão da cobertura de tribunais e o desenvolvimento de versões com maior grau de autonomia operacional.

A abordagem é mista. A dimensão qualitativa manifesta-se na análise da estrutura semântica das publicações judiciais, no mapeamento do vocabulário técnico forense necessário para a instrução do pipeline e na avaliação das implicações jurídicas e éticas do uso de IA no Direito, seguindo os princípios de Minayo (2012). A dimensão quantitativa está presente na avaliação de desempenho do protótipo, mensurada por meio do tempo de processamento, acurácia de extração e análise de custo-benefício operacional.

Os meios de investigação são bibliográficos e documentais. O referencial teórico foi construído a partir de revisão da literatura em bibliotecas digitais, repositórios e indexadores acadêmicos (arXiv, ACM Digital Library, IEEE Xplore e Google Scholar), com recorte prioritário entre 2017 e 2024. Foram igualmente analisados textos normativos como a Resolução CNJ n. 332/2020 e a LGPD. O caráter multidisciplinar é estrutural, articulando Ciência da Computação, Direito e Engenharia de Software. Como observa Aguiar (2019) para a interseção entre Computação e Saúde Pública, diferentes disciplinas precisam compartilhar conhecimentos para que a IA opere com rigor em domínios especializados, premissa que se aplica diretamente à interseção entre Computação e Direito que fundamenta este trabalho.

O desenvolvimento do protótipo seguiu uma abordagem iterativa e incremental inspirada no arcabouço Scrum, com ciclos de desenvolvimento de duas semanas e refinamento contínuo da engenharia de prompts do módulo de extração semântica. A validação do artefato ocorreu por meio de três rodadas de teste com publicações reais de Diários de Justiça Eletrônicos, cujos resultados empíricos são discutidos na Seção 6.

5 DESENVOLVIMENTO DA ARQUITETURA TÉCNICA

5.1 Arquitetura de Microsserviços

O protótipo foi estruturado segundo o paradigma de microsserviços, no qual o sistema é decomposto em unidades autônomas com responsabilidades bem definidas, comunicação via APIs e capacidade de escalonamento independente. Esse modelo oferece disponibilidade contínua, tolerância a falhas isoladas e escalabilidade horizontal em momentos de pico, como a ingestão das edições diárias dos Diários de Justiça.

A arquitetura organiza-se em três camadas funcionais. A camada de ingestão recebe e préprocessa os dados brutos: arquivos CSV ou XLSX passam por validação de schema, normalização de encoding e extração do texto. A camada de processamento semântico constitui o núcleo do protótipo, integrada pelo módulo de embeddings para vetorização, o módulo de extração estruturada que instrui o LLM a identificar os campos-alvo, e o módulo de classificação de urgência. A camada de recuperação e interface oferece busca semântica, painéis de prazos e exportação de dados para sistemas de gestão.

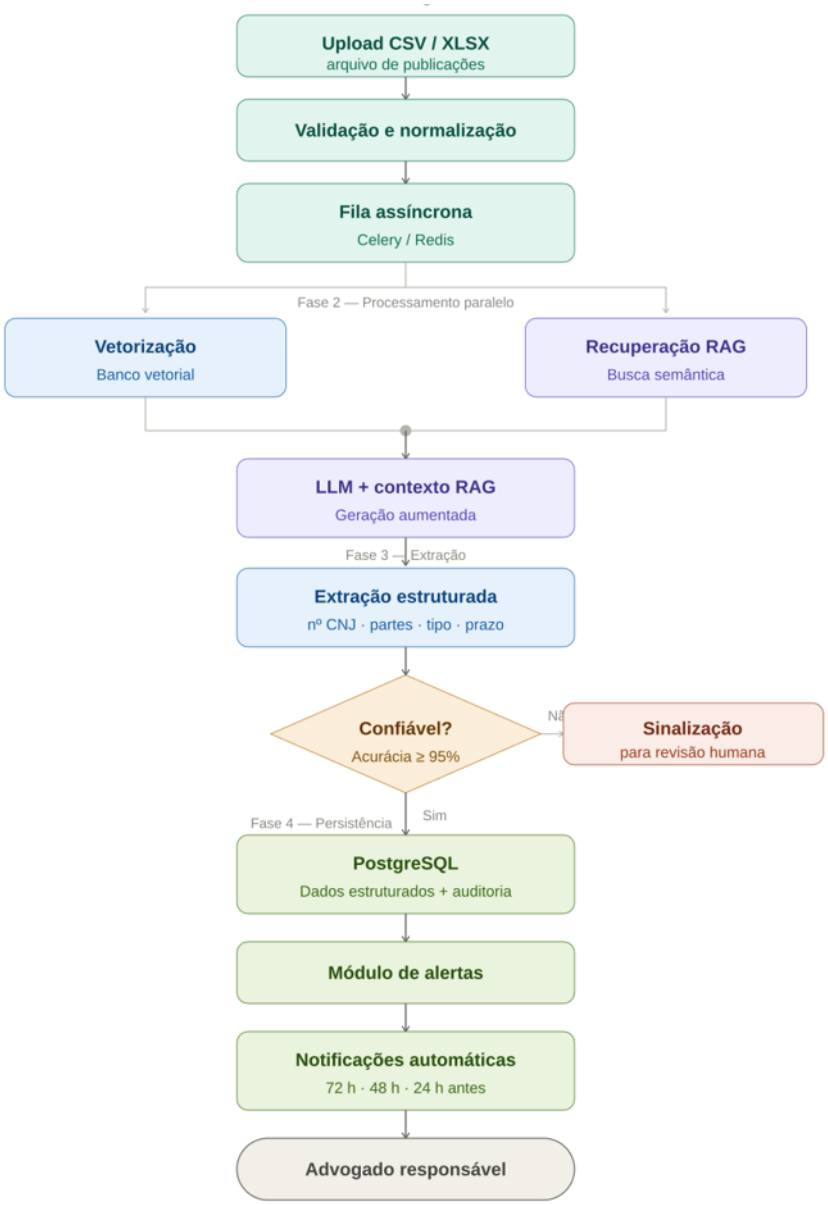
5.2 Fluxo do Pipeline de Extração Semântica

O processamento de cada publicação percorre as etapas sequenciais mapeadas no fluxo operacional da Figura 1. O arquivo é recebido pelo backend FastAPI e enfileirado no Celery via Redis. Um worker consome a tarefa, divide o lote em unidades individuais e inicia o processamento paralelo. Para cada publicação, o texto é vetorizado e persistido no banco vetorial com seus respectivos metadados.

O módulo de extração semântica envia o texto ao LLM com um prompt configurado para o domínio jurídico, solicitando os campos em formato JSON validado: número do processo no padrão CNJ (NNNNNNN-DD.AAAA.J.TT.OOOO), partes envolvidas, tipo de ato judicial e prazo processual aplicável. Os dados estruturados são persistidos no PostgreSQL e vinculados ao cliente e ao usuário responsável. O módulo de alertas verifica os prazos extraídos e dispara notificações escalonadas (72h, 48h e 24h antes do vencimento) via interface web ou e-mail.

Cada campo extraído carrega uma referência direta ao trecho do documento original que o fundamentou. Essa rastreabilidade é diretamente relevante para a segurança jurídica: o profissional que confere o resultado pode verificar a origem de cada prazo calculado diretamente no texto oficial, sem depender da memória paramétrica do modelo.

Figura 1 – Pipeline de processamento do JusInsight.



Fonte: Elaborado pelos autores (2026).

5.3 Stack Tecnológica

A Tabela 1 apresenta a stack tecnológica do protótipo, organizada por camada funcional.

Tabela 1 - *Stack tecnológica do JusInsight por camada funcional.*

Camada	Descrição Técnica
Frontend	React 18 com TypeScript, Tailwind CSS e Redux Toolkit. Interface modelada segundo princípios de UX jurídica, com fluxos otimizados para triagem e conferência de publicações.
Backend	Python 3.11+ com FastAPI. Orquestração do processamento semântico, integração com APIs de LLMs (Groq, OpenAI, Anthropic Claude) e endpoints RESTful documentados via OpenAPI/Swagger.
Processamento Assíncrono	Celery com broker Redis. Filas de tarefas distribuídas com controle de taxa adaptativo para respeitar os rate limits das APIs de LLM e processamento paralelo de lotes.
Bancos de Dados	PostgreSQL para dados relacionais e armazenamento cifrado (AES-256) do texto bruto com controle RBAC. Pinecone ou ChromaDB como banco vetorial para busca semântica do RAG, sem persistência de texto original.
Segurança e DevOps	Docker/Docker Compose, autenticação JWT, TLS 1.3 em trânsito, AES-256 em repouso, log de auditoria imutável. Conformidade com a LGPD (Lei n. 13.709/2018).

Fonte: Elaborada pelos autores (2026).

5.4 Segurança, Privacidade e Conformidade com a LGPD

A arquitetura de segurança foi concebida com base no princípio de Privacy by Design (ISO/IEC 29101), integrando controles de privacidade desde as decisões iniciais de projeto. O controle de acesso adota o modelo RBAC (Role-Based Access Control), com perfis para administradores, advogados seniores, advogados associados e estagiários, cada qual com permissões granulares sobre leitura, edição e exportação.

A proteção dos dados processuais no banco vetorial merece atenção técnica específica. Embeddings são representações matemáticas de alta dimensionalidade (tipicamente 1.536 dimensões) que codificam o significado semântico de um texto em espaço vetorial contínuo. Morris et al. (2023) demonstraram experimentalmente que sistemas com acesso iterativo ao modelo gerador conseguem reconstruir textos a partir de seus vetores com fidelidade significativa, evidenciando que embeddings isolados não constituem, por si sós, garantia absoluta de privacidade.

O JusInsight trata essa vulnerabilidade por meio de controles complementares que atuam diretamente sobre as condições necessárias para os ataques documentados: o texto bruto das publicações não é armazenado no banco vetorial, permanecendo cifrado em AES-256 no PostgreSQL com acesso restrito por autenticação JWT; o banco vetorial contém exclusivamente os vetores numéricos e metadados identificadores; e o modelo gerador não é exposto a consultas externas, eliminando a possibilidade de acesso iterativo que os ataques de Morris et al. (2023) pressupõem.

A transmissão entre todos os componentes é protegida por TLS 1.3, e um log de auditoria imutável registra cada operação de leitura, consulta e exportação, atendendo ao princípio de responsabilização (accountability) do art. 6º, X, da LGPD. A minimização de dados prevista no art. 6º, III, é aplicada pelo processamento exclusivo dos campos estritamente necessários à gestão de prazos, sem retenção de dados além do período operacionalmente justificado.

6 PROVA DE CONCEITO, DESAFIOS TÉCNICOS E RESULTADOS

6.1 Metodologia da Prova de Conceito

Para verificar a viabilidade do protótipo em condições representativas, foram conduzidos três testes com publicações autênticas de Diários de Justiça Eletrônicos. O conjunto total foi de 2.000 publicações, distribuídas ao longo das três rodadas, com média de 667 registros por teste. Esse volume reflete o fluxo diário típico de um escritório de médio porte, estimado entre 600 e 1.000 publicações por dia em atuação multitribunal. A configuração de infraestrutura foi mantida uniforme: três workers Celery, acesso à API de LLM com autenticação padrão e banco vetorial em instância de desenvolvimento.

A acurácia foi calculada como a proporção de campos corretamente extraídos em relação ao total esperado, tomando como referência a anotação manual dos próprios autores sobre 100 publicações por teste (total: 300 publicações anotadas). Considerou-se correto um campo quando o valor extraído correspondia exatamente ao valor de referência. O tempo médio de três minutos por publicação para a triagem manual foi estimado a partir de cronometragem realizada pelos autores durante a preparação do conjunto de referência, na qual profissionais com formação jurídica realizaram a extração manual dos mesmos campos avaliados pelo sistema.

6.2 Desafios Técnicos e o RAG como Memória Jurídica

O maior obstáculo não foi computacional. Foi de domínio. Antes de configurar qualquer prompt de instrução, foi necessário compreender o que é prazo em dobro para a Fazenda Pública, como se distingue uma intimação pessoal de uma via DJe e porque os ritos da Justiça do Trabalho diferem dos da Justiça Federal. Esse aprendizado, conduzido em colaboração com profissional de formação jurídica, confirmou na prática o princípio que Aguiar (2019) articula para a saúde pública: a interdisciplinaridade não é um atributo desejável do projeto — é condição para que funcione.

O desafio mais recorrente foi a determinação de prazos implícitos: publicações em que o magistrado remete à norma que estabelece o prazo sem declará-lo numericamente. Um exemplo concreto é a redação "Intime-se o embargado nos termos do art. 1.023 do CPC". Esse artigo, em seu parágrafo segundo, determina que, quando o julgamento dos embargos de declaração puder implicar a extinção do processo ou o julgamento do mérito, a parte contrária seja intimada para manifestar-se no prazo de cinco dias. Nas versões iniciais do protótipo, essa estrutura resultava no campo de prazo vazio ou preenchido com a referência legislativa literal, sem cálculo.

A solução revela o funcionamento concreto do RAG e desfaz o efeito de caixa-preta frequentemente associado à IA generativa. O banco vetorial foi alimentado com registros curados que associam referências legislativas ao seu valor em dias: o art. 1.023, §2°, do CPC ao prazo de 5 dias para embargos de declaração; o art. 335 ao prazo de 15 dias para contestação; o art. 496 ao prazo de 15 dias para apelação; entre outros. Quando o protótipo encontra uma publicação com referência ao art. 1.023, a etapa de recuperação do RAG localiza o fragmento curado correspondente e o injeta no contexto enviado ao LLM. O modelo lê o prazo no contexto verificável fornecido, sem necessidade de inferência por memorização de treinamento. O resultado é rastreável, corrigível e escalável: cada nova regra de prazo implícito inserida no banco vetorial expande a cobertura sem retreinamento do modelo, criando uma base de conhecimento jurídico viva que evolui com o uso.

Para fins de reprodutibilidade, apresenta-se abaixo uma versão esquemática do template de prompt utilizado nos testes, em que {contexto} representa os fragmentos recuperados do banco vetorial e {publicação} o texto bruto da publicação:

```
Sistema: Você é um analisador de publicações do Diário de Justiça
Eletrônico brasileiro.
Extraia os campos solicitados exclusivamente com base no texto e no
contexto fornecidos.
Retorne null para campos ausentes ou incertos. Não infira
informações.
Contexto jurídico de referência:
{contexto}
Publicação:
{publicacao}
Retorne apenas JSON no seguinte formato:
{"numero_cnj": string|null, "autor": string|null, "reu":string|null,
 "tipo_ato": "despacho"|"decisao"|"sentenca"|"acordao"|null,
 "prazo_dias": integer|null, "confianca": "alta"|"media"|"baixa"}
```

Fonte: Elaborada pelos autores (2026).

No plano da infraestrutura, identificou-se um gargalo de concorrência entre os workers Celery e a API do provedor de LLM. Com múltiplos workers disparando requisições simultâneas, o sistema atingia os limites de taxa da API (rate limits), gerando falhas e reprocessamentos. A solução foi um mecanismo de controle de taxa com fila de prioridade no nível do worker: cada instância verifica o limite disponível antes de disparar a próxima requisição e aguarda o janelamento necessário sem bloquear a fila global. A heterogeneidade de formatação entre tribunais também exigiu atenção: um módulo de normalização identifica o tribunal de origem a partir dos metadados do arquivo e aplica o parser correspondente, sinalizando para revisão humana as publicações que escapam a qualquer mapeamento disponível.

6.3 Resultados Operacionais

Os três testes produziram resultados consistentes. O processamento de cada lote de aproximadamente 667 publicações foi concluído em 3 a 4 minutos. Dois revisores conduziram a conferência dos campos extraídos em cerca de 20 minutos, sem necessidade de cadastro manual no sistema, pois o protótipo alimenta diretamente a base de dados de gestão. O tempo total do fluxo mediado pelo JusInsight ficou em aproximadamente 23 minutos, ante os 1.000 minutos que dois analistas demandariam para a análise manual do mesmo volume, uma redução de 98% no tempo operacional e nas horas-pessoa empregadas.

A acurácia média na extração dos quatro campos principais (número CNJ, partes, tipo de ato e prazo) foi de 95,3%. Nenhuma falha silenciosa foi registrada: todas as publicações com extração incerta foram corretamente sinalizadas para revisão.

A análise qualitativa dos 4,7% de extrações incorretas revela quatro categorias distintas de falha, o que traz mais valor científico do que o número isolado. A primeira, responsável por aproximadamente 2%, envolveu publicações cujo template de formatação não estava coberto pelos parsers disponíveis: o módulo de normalização rejeitou corretamente o registro, e a sinalização funcionou como esperado. A segunda categoria, em torno de 1,5%, compreendeu referências legislativas de prazo ainda não catalogadas no banco vetorial — artigos de legislações esparsas, como a Lei n. 9.099/1995 (Juizados Especiais) e a Lei n. 6.830/1980 (Execução Fiscal), cujos prazos específicos não foram inseridos durante a fase de testes. A terceira, estimada em 0,9%, correspondeu a publicações com ambiguidade sintática genuína: intimações coletivas em que o texto não especificava inequivocamente a qual parte o prazo se aplicava — um desafio que permanece aberto mesmo para leitores humanos experientes. Por fim, cerca de 0,3% originaram-se de erros tipográficos nos próprios Diários: números de

processo com formatação inválida ou nomes de partes truncados por artefatos de OCR no arquivo-fonte. As duas últimas categorias indicam um limite estrutural que o aprimoramento do prompt ou da base vetorial não resolve sozinho; demandam melhoria na qualidade dos documentos de origem.

Do ponto de vista do risco jurídico, o resultado mais relevante é a eliminação da dependência de atenção humana ininterrupta sobre um volume que supera a capacidade de triagem confiável de uma equipe reduzida. O REsp 1.637.375/SP (STJ, 2020) traduz em obrigação de indenizar exatamente a falha que esse tipo de sobrecarga produz: a perda de prazo por omissão operacional. Automatizar a triagem não é uma melhoria incremental de processo; é a remoção da causa estrutural de um risco juridicamente caracterizado.

6.4 Análise de Viabilidade Financeira

Para avaliar a viabilidade econômica do protótipo, foram calculados e comparados os custos operacionais dos dois fluxos para o volume médio diário de 600 publicações. A Tabela 2 abaixo, sintetiza os indicadores principais.

Tabela 2 – Comparativo operacional e financeiro entre o fluxo manual e o protótipo.

Indicador	Fluxo Manual	JusInsight (PoC)
Volume diário típico	600 publicações	600 publicações
Tempo de processamento (667 pub.)	~1.000 min (2 analistas)	~23 min (2 revisores)
Custo diário de pessoal	R$ 318,18 (2 analistas)	< R$ 1,00 (tokens de API)
Custo mensal estimado	R$ 7.000,00+ (salários brutos)	< R$ 200,00 (infraestrutura)
Cadastro no sistema	Manual por publicação	Automatizado pelo protótipo
Rastreabilidade	Dependente do profissional	Por campo extraído, auditável

Fonte: Elaborada pelos autores (2026).

O profissional encarregado da triagem manual recebe pelo piso salarial médio da categoria aproximadamente R$ 3.500,00 mensais. Considerando 22 dias úteis por mês, o custo diário é de R$ 159,09, representando R$ 318,18 por dia para uma equipe mínima de dois analistas, ou R$ 7.000,00 mensais em salários brutos. Com os encargos previdenciários e trabalhistas (estimados entre 65% e 80% sobre o salário bruto, conforme o regime de contratação), o custo mensal total ultrapassa R$ 11.500,00.

No fluxo automatizado, o processamento do lote diário de 600 publicações consumiu, nos testes realizados, menos de R$ 1,00 em tokens de API do modelo GPT-3.5. Somados os custos de infraestrutura em nuvem, o custo operacional mensal estimado para o protótipo permanece abaixo de R$ 200,00. A relação é de mais de 35 vezes no custo direto de pessoal e superior a 57 vezes quando incorporados os encargos trabalhistas. Para escritórios com volumes superiores a 2.000 publicações diárias, a vantagem se acentua: o custo de API escala linearmente, enquanto o custo de pessoal para triagem manual cresceria de forma proporcional ao número de contratações necessárias

7 Conclusões e Trabalhos futuros

A perda de um prazo processual, enquadrada juridicamente pela teoria da perda de uma chance (SAVI, 2012; STJ, REsp 1.637.375/SP, 2020), tem na sobrecarga informacional da triagem manual sua principal causa estrutural nos escritórios de advocacia contemporâneos. Os resultados desta prova de conceito tornam essa relação quantificável: o lote diário de um escritório de médio porte, que demandaria mais de 16 horas de dois analistas na triagem manual,

foi processado pelo protótipo em menos de 25 minutos, com acurácia de 95,3% e sinalização correta de todas as extrações incertas. A diferença não é apenas de velocidade: é uma mudança estrutural na viabilidade da supervisão completa do fluxo de publicações, que é o que a diligência profissional exige.

A fundamentação teórica posicionou o JusInsight numa linhagem científica que parte de Aguiar (2019), cujos modelos preditivos com Machine Learning clássico estabeleceram a viabilidade da IA em domínios especializados do setor público brasileiro, e avança pela arquitetura Transformer (VASWANI et al., 2017), pelos LLMs contemporâneos (ZHAO et al., 2023) e pela técnica RAG (LEWIS et al., 2020). A comparação crítica entre os dois paradigmas, que contrapõe a IA preditiva sobre dados estruturados à extração semântica sobre linguagem natural não estruturada, constitui uma das contribuições teóricas deste trabalho.

As contribuições técnicas incluem a descrição do pipeline LLM-RAG para o domínio jurídico, o mapeamento dos desafios de prompt engineering em textos forenses brasileiros com solução concreta pelo mecanismo de injeção de conhecimento curado, a solução para o gargalo de concorrência em ambientes com APIs de LLM sob rate limit e a análise do risco de inversão de embeddings com fundamento em Morris et al. (2023), acompanhada da descrição dos controles arquiteturais que o mitigam.

Como limitação, os testes foram conduzidos com publicações de um subconjunto de tribunais. A cobertura nacional, dada a diversidade de templates dos 92 tribunais do sistema judiciário brasileiro, demandará a expansão contínua do módulo de normalização e da base de conhecimento jurídico curado no banco vetorial. A acurácia de 95,3% deve ser tratada como resultado de fase inicial, sujeito a variação conforme a expansão do escopo de análise.

A visão de longo prazo do projeto é a evolução do pipeline atual para um sistema com grau crescente de autonomia, capaz de incorporar o cadastro automatizado no sistema de gestão do escritório e reduzir a intervenção humana à supervisão de exceções. Essa evolução pressupõe a acumulação de dados operacionais reais e o amadurecimento da base de conhecimento jurídico curado.

Os trabalhos futuros imediatos incluem: * A validação do protótipo em parceria com um escritório de advocacia em ambiente produtivo; * A integração direta com as APIs dos Diários de Justiça Eletrônicos, eliminando a necessidade de carregamento manual de arquivos; * A investigação do fine-tuning de um modelo de linguagem especializado em publicações judiciais em português para reduzir a dependência de um prompt engineering extenso.

REFERÊNCIAS

AGUIAR, Wellington Sousa. Desenvolvimento de modelos preditivos da mortalidade infantil com base em inteligência artificial no estado do Ceará. 2019. 163 f. Tese (Doutorado em Saúde Coletiva) – Programa de Pós-Graduação em Saúde Coletiva, Departamento de Saúde Comunitária, Universidade Federal do Ceará, Fortaleza, 2019.

BRASIL. Superior Tribunal de Justiça (3. Turma). Recurso Especial nº 1.637.375/SP. Relator: Ministro Ricardo Villas Bôas Cueva. Diário da Justiça Eletrônico, Brasília, 25 nov. 2020.

ANDRADE, Mariana Dionísio de; PINTO, Eduardo Régis Girão de Castro; LIMA, Isabela Braga de; GALVÃO, Alex Renan de Sousa. Inteligência artificial para o rastreamento de ações com repercussão geral: o projeto Victor e a realização do princípio da razoável duração do processo. Revista Eletrônica de Direito Processual, Rio de Janeiro, v. 21, n. 1, 2020. DOI: 10.12957/redp.2020.42717.

BROWN, Tom B. et al. Language models are few-shot learners. In: Advances in Neural Information Processing Systems (NeurIPS), v. 33, p. 1877-1901, 2020. DOI: 10.48550/arXiv.2005.14165.

CONSELHO NACIONAL DE JUSTIÇA (CNJ). Justiça em Números 2023: ano-base 2022. Brasília: CNJ, 2023. Disponível em: https://www.cnj.jus.br/pesquisas-judiciarias/justica-em-numeros/. Acesso em: jan. 2025.

CONSELHO NACIONAL DE JUSTIÇA (CNJ). Resolução n. 332, de 21 de agosto de 2020. Dispõe sobre ética, transparência e governança na produção e no uso de Inteligência Artificial no Poder Judiciário. Brasília: CNJ, 2020.

DEVLIN, Jacob et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), p. 4171-4186, 2019. DOI: 10.18653/v1/N19-1423.

GIL, Antonio Carlos. Métodos e técnicas de pesquisa social. 6. ed. São Paulo: Atlas, 2008.

LEWIS, Patrick et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems (NeurIPS), v. 33, p. 9459-9474, 2020. DOI: 10.48550/arXiv.2005.11401.

MINAYO, Maria Cecília de Souza. Pesquisa social: teoria, método e criatividade. 31. ed. Petrópolis: Vozes, 2012.

MORRIS, John X. et al. Text embeddings reveal (almost) as much as text. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), p. 12448-12460, 2023. DOI: 10.48550/arXiv.2310.06816.

NUNES, Dierle; LUCON, Paulo Henrique dos Santos; WOLKART, Erik Navarro (Coord.). Inteligência artificial e direito processual: os impactos da virada tecnológica no direito processual. 2. ed. Salvador: Editora JusPodivm, 2021.

SAVI, Sérgio. Responsabilidade civil por perda de uma chance. 3. ed. São Paulo: Atlas, 2012.

VASWANI, Ashish et al. Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS), v. 30, 2017. DOI: 10.48550/arXiv.1706.03762.

ZHAO, Wayne Xin et al. A survey of large language models. arXiv preprint arXiv:2303.18223, 2023. DOI: 10.48550/arXiv.2303.18223.