

CUBO VERDE: UM ECOSSISTEMA DE AGENTES DE INTELIGÊNCIA ARTIFICIAL, GOVERNANÇA DE DADOS E ARQUITETURA DIGITAL PARA AMPLIAÇÃO DO ACESSO À JUSTIÇA

**CUBO VERDE: AN ECOSYSTEM OF ARTIFICIAL INTELLIGENCE AGENTS,
DATA GOVERNANCE AND DIGITAL ARCHITECTURE FOR BROADENING
ACCESS TO JUSTICE**

Ciências Exatas e da Terra, Ciências Sociais Aplicadas, Ciências
Humanas

• 28/09/2026

REGISTRO DOI: [10.70773/revistatopicos/790535727](https://doi.org/10.70773/revistatopicos/790535727)

Fabio Alessandro Guerra¹

Nicholas Moura e Silva²

Matheus Cavalcanti Munhoz³

Vinicius Tertulino Parede⁴

Kallil Miguel Caparroz Zielinski⁵

Alvaro Mateus Santana⁶

Tatiana Rodrigues Pereira de Lima⁷

Leandro Antonio Jiomeke⁸

Giovanna Nicolau Dalledonne⁹

RESUMO

Este artigo descreve e analisa o assistente jurídico de consulta aos autos do Cubo Verde, ecossistema digital que a Defensoria Pública do Estado do Paraná mantém em operação para ampliar o acesso à justiça. O estudo caracteriza-se como pesquisa aplicada, conduzida sob a lógica da design science research, e toma o artefato em funcionamento como caso instrumental. A investigação percorre a cadeia completa que vai da importação dos autos do sistema institucional de atendimento até a resposta entregue ao defensor, detalhando a extração de texto, a fragmentação, a codificação em vetores densos de 768 dimensões, a recuperação por vizinhança aproximada e a geração condicionada pelos trechos recuperados, com verificação posterior de cada citação contra o conjunto efetivamente consultado. Sete equações formalizam esse arranjo, do mecanismo de atenção ao orçamento de tokens por lote de codificação. A plataforma se oferece em duas formas, portal de acesso interno e interfaces de programação acopladas aos sistemas de atendimento SOLAR e LUNA, e reúne oito agentes, entre eles a transcrição de audiências com Whisper e WhisperX em unidade gráfica NVIDIA e a classificação automática de relatos, ambas apoiadas em modelos Gemini. Os resultados registram redução aproximada de setenta por cento no tempo de análise processual, apoio a quase 200 defensores e às respectivas equipes de assessoria, mais de 3.600 depoimentos transcritos e economia estimada de 17.320 horas anuais. A discussão evidencia que o ganho decorre menos do modelo de linguagem do que da qualidade da recuperação, da governança do dado e das decisões de engenharia que sustentam a operação.

Palavras-chave: Acesso à justiça; Recuperação aumentada por geração; Assistente jurídico; Governança de dados; Defensoria Pública.

ABSTRACT

This paper describes and analyses the case file legal assistant of Cubo Verde, the digital ecosystem operated by the Public Defender's Office of the State of Paraná, Brazil, to broaden access to justice. The study is applied research, conducted under the logic of design science research, and takes the working artefact as an instrumental case. The investigation follows the whole chain that runs from the import of case documents out of the institutional service system to the answer delivered to the public defender, detailing text extraction, chunking, encoding into 768-dimension dense vectors, approximate nearest neighbour retrieval and generation conditioned on the retrieved passages, with subsequent verification of every citation against the set actually consulted. Seven equations formalise that arrangement, from the attention mechanism to the token budget of each encoding batch. The platform offers itself in two forms, an internal access portal and application programming interfaces wired into the SOLAR and LUNA service systems, and gathers eight agents, among them hearing transcription with Whisper and WhisperX on an NVIDIA graphics processing unit and automatic intake classification, both supported by Gemini models. Results record an approximate seventy per cent reduction in case analysis time, support to nearly 200 public defenders and their assistant teams, more than 3,600 transcribed statements and an estimated saving of 17,320 working hours per year. The discussion shows that the gain stems less from the language model than from retrieval quality, data governance and the engineering decisions that sustain operation.

Keywords: Access to justice; Retrieval-augmented generation; Legal assistant; Data governance; Public Defender's Office.

1. INTRODUÇÃO

O acesso à justiça figura na Constituição da República Federativa do Brasil de 1988 como dever do Estado, que se obriga a prestar assistência jurídica integral e gratuita a quem comprove insuficiência de recursos, nos termos do art. 5º, inciso LXXIV. A Defensoria Pública é a instituição incumbida dessa função, conforme o art. 134 do texto constitucional e a Lei Complementar nº 80, de 12 de janeiro de 1994. O reconhecimento formal do direito, contudo, não garante por si a sua fruição. Cappelletti e Garth (1988) já observavam que o acesso efetivo depende da remoção de obstáculos econômicos, informacionais e organizacionais, e não apenas da existência de previsão legal, formulação que Santos (2007) retoma ao tratar da distância persistente entre o direito declarado e o direito exercido.

Na Defensoria Pública do Estado do Paraná esses obstáculos assumem feição concreta e mensurável. A instituição realizou mais de novecentos mil atendimentos em 2025, dos quais cerca de quarenta mil por meio digital, e mantém presença própria em apenas 74 dos 399 municípios paranaenses, distribuídos por trinta comarcas. A demanda por assistência jurídica gratuita cresce de modo contínuo, impulsionada pela ampliação da consciência de direitos e pelo agravamento das desigualdades sociais, enquanto o quadro de pessoal não acompanha esse crescimento na mesma proporção. À pressão de volume soma-se a fragmentação histórica dos sistemas e das bases de dados, desenvolvidos de forma descentralizada e sem integração plena, o que produz retrabalho, dificulta o compartilhamento de informações e reduz a capacidade de leitura estratégica da própria atuação institucional.

Há ainda um obstáculo de natureza documental, menos discutido e igualmente determinante, que constitui o problema central deste

artigo. Um processo de família disponibilizado ao assistente jurídico da instituição reúne 230 documentos que, somados, ocupam cerca de 308 mil tokens de contexto e se desdobram em 2.770 trechos indexados; um processo criminal submetido à triagem alcançou 1.339 páginas. Nenhum profissional lê esse volume em tempo compatível com o prazo processual, e a leitura seletiva por busca textual simples encontra a palavra e perde o sentido, porque o termo que o defensor procura raramente é o termo que o documento emprega. O problema não é, portanto, apenas de quantidade de atendimentos: é de capacidade de extrair informação útil de massas documentais que já estão digitalizadas e que, ainda assim, permanecem praticamente inacessíveis no tempo de que se dispõe.

Os modelos de linguagem de grande escala oferecem resposta parcial a esse problema e trazem, junto com ela, um risco próprio. Ji et al. (2023) documentam a propensão desses modelos a produzir enunciados fluentes e factualmente incorretos, fenômeno designado alucinação. Em ambiente jurídico, um parecer que cita jurisprudência inexistente ou atribui ao processo um fato que ele não contém não é apenas um erro técnico: é um risco à defesa do assistido, e o profissional que assina a peça responde por ele. A questão de pesquisa que orienta este trabalho decorre dessa tensão e pode ser enunciada da seguinte forma: como uma instituição pública pode incorporar modelos de linguagem aos seus fluxos finalísticos sem transferir a esses modelos a responsabilidade pela veracidade da informação prestada ao cidadão?

A pergunta chega, hoje, acompanhada de política pública explícita. O Plano Brasileiro de Inteligência Artificial, recentemente divulgado pelo Ministério da Ciência, Tecnologia e Inovação em conjunto com o Centro de Gestão e Estudos Estratégicos (MCTI; CGEE, 2025), prevê

investimento da ordem de vinte e três bilhões de reais entre 2024 e 2028 e dedica um de seus eixos, com cerca de 7,6 por cento desse montante, à melhoria dos serviços públicos por meio de inteligência artificial. O mesmo documento registra que mais de trinta por cento dos órgãos federais e estaduais brasileiros já empregam ao menos uma tecnologia de inteligência artificial e estabelece, entre suas premissas, que o uso dessas tecnologias observe a ética, a responsabilidade e a conformidade com a Lei Geral de Proteção de Dados Pessoais. A experiência aqui relatada situa-se exatamente nesse eixo, e é oferecida como registro documentado de uma implantação que já saiu do plano e está em produção.

O objetivo geral do artigo é descrever e analisar o assistente jurídico de consulta aos autos do Cubo Verde, ecossistema digital em operação contínua na Defensoria Pública do Estado do Paraná desde junho de 2025, com ênfase nos mecanismos que ancoram a resposta gerada em fonte documental verificável. Como objetivos específicos, o trabalho se propõe a apresentar a arquitetura da plataforma e as duas formas pelas quais ela se oferece à instituição, portal de acesso interno e interfaces de programação acopladas aos sistemas de atendimento; formalizar, com o aparato matemático pertinente, a cadeia de codificação, recuperação e geração que sustenta o assistente; detalhar as verificações que separam esse desenho da aplicação ingênua de um modelo de linguagem; descrever os demais agentes em produção e os parâmetros adotados em cada um; e discutir as decisões de engenharia e de governança que se mostraram determinantes para a confiabilidade do conjunto.

A justificativa do estudo tem dois planos. No plano prático, a documentação detalhada de uma implantação real em Defensoria

Pública oferece referência reaplicável a instituições congêneres, que enfrentam o mesmo problema sob restrição orçamentária semelhante; o sistema de atendimento sobre o qual o assistente foi acoplado está presente em vinte e uma unidades da federação, o que torna a reprodução do arranjo uma possibilidade concreta, e não uma hipótese de laboratório. No plano teórico, a literatura sobre recuperação aumentada por geração concentra-se na avaliação de modelos e de estratégias de recuperação sobre conjuntos de dados públicos, e registra com menos frequência os parâmetros efetivamente adotados em produção, as razões pelas quais foram escolhidos e os efeitos observados quando estão errados. Este artigo procura contribuir exatamente nesse ponto, relatando medições feitas no ambiente de produção da instituição.

O texto está organizado em cinco seções além desta introdução. A seção 2 apresenta a fundamentação teórica, que percorre o acesso à justiça como problema de capacidade institucional, o mecanismo de atenção, as representações vetoriais densas, a recuperação aumentada por geração, o reconhecimento de fala e a governança de inteligência artificial no setor público. A seção 3 descreve a metodologia, a arquitetura da plataforma e a sua articulação com os sistemas institucionais. A seção 4 é dedicada integralmente ao assistente jurídico, do processo eletrônico à resposta ancorada. A seção 5 descreve os demais agentes e serviços do ecossistema. A seção 6 expõe e discute os resultados institucionais e as lições de engenharia, e a última seção reúne as considerações finais, as limitações do estudo e as frentes de continuidade.

2. FUNDAMENTAÇÃO TEÓRICA

2.1. Acesso à Justiça Como Problema de Capacidade Institucional

A formulação clássica de Cappelletti e Garth (1988) organiza a ampliação do acesso à justiça em três ondas: a assistência judiciária aos pobres, a representação dos interesses difusos e a reforma dos próprios procedimentos e instituições. É a terceira onda que interessa diretamente a este trabalho, porque desloca a análise do direito de ação para a capacidade concreta de a instituição processar a demanda que recebe. Sob essa perspectiva, fila, prazo e retrabalho deixam de ser questões administrativas periféricas e passam a integrar o próprio conteúdo do direito de acesso: o assistido cujo defensor não teve tempo de ler os autos foi atendido no papel e desassistido no fato.

Santos (2007) acrescenta a esse quadro a dimensão da linguagem e da informação. O cidadão que não compreende o que lhe é comunicado permanece excluído mesmo quando formalmente atendido, e a barreira linguística opera dos dois lados do balcão: atinge o assistido diante do despacho judicial e atinge a própria instituição diante do acervo que produziu. Essa observação sustenta, no desenho do ecossistema aqui descrito, tanto a existência de um agente dedicado à tradução de conteúdo jurídico para linguagem acessível quanto a opção por consultar os autos em linguagem natural, e não por meio de filtros e operadores booleanos. No plano normativo brasileiro, a Lei Complementar nº 80, de 1994, atribui à Defensoria Pública a orientação jurídica e a promoção dos direitos humanos, o que abrange a compreensibilidade da informação prestada, e não apenas a atuação processual.

2.2. Transformadores e o Mecanismo de Atenção

Toda a cadeia técnica descrita neste artigo repousa sobre uma única operação, proposta por Vaswani et al. (2017): a atenção por produto

escalar escalonado. Dado um conjunto de consultas organizado na matriz Q , um conjunto de chaves em K e um conjunto de valores em V , a operação pondera os valores pela afinidade entre consultas e chaves, conforme a Equação 1.

$$\text{Atenção}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

O produto entre consultas e chaves mede a compatibilidade entre cada par de posições; a divisão pela raiz quadrada da dimensão das chaves impede que esse produto cresça a ponto de saturar a normalização exponencial; e a normalização converte as compatibilidades em pesos que somam um, com os quais os valores são combinados. A consequência que importa aqui é conceitual: a representação de cada palavra deixa de ser fixa e passa a depender do contexto em que ela aparece, de modo que o termo guarda em um processo de família e o mesmo termo em um processo de execução penal ocupam posições distintas no espaço de representação. É essa propriedade que torna possível recuperar documentos por sentido, e não por coincidência de caracteres.

A mesma operação sustenta os dois papéis exercidos pelos modelos empregados na plataforma. Modelos codificadores, na linhagem inaugurada por Devlin et al. (2019), usam a atenção para produzir uma representação vetorial de um trecho de texto, empregada aqui na indexação e na busca. Modelos geradores, na linhagem de Brown et al. (2020), usam a mesma operação para prever o próximo token e são empregados na redação da resposta e na extração de campos estruturados. Reconhecer que as duas famílias derivam do mesmo mecanismo ajuda a evitar um equívoco frequente na discussão institucional, que é tratar o modelo gerador como se ele fosse também o responsável pela busca: na arquitetura aqui descrita,

quem encontra e quem redige são componentes distintos, com custos, falhas e critérios de avaliação distintos.

2.3. Representações Vetoriais Densas, Similaridade e Vizinhança Aproximada

A representação de texto por vetores densos é a base técnica dos mecanismos descritos neste artigo. Um modelo de codificação projeta um trecho de texto em um espaço vetorial de dimensão fixa, no qual a proximidade geométrica aproxima a proximidade semântica. Reimers e Gurevych (2019) mostraram que codificadores ajustados para produzir representações de sentença permitem comparar pares de textos por medida de similaridade direta, sem submeter cada par ao modelo, o que viabiliza a busca sobre acervos grandes. Karpukhin et al. (2020) demonstraram, em seguida, que a recuperação densa supera a recuperação léxica clássica em perguntas de domínio aberto, resultado que motivou a adoção do padrão em aplicações de pergunta e resposta sobre acervos próprios.

A medida empregada é a similaridade de cosseno, que compara a direção de dois vetores e ignora a sua magnitude, propriedade desejável quando se comparam textos de comprimentos muito diferentes, como um relato de três linhas e uma petição de oitenta páginas. A distância correspondente é o seu complemento, conforme a Equação 2.

$$\text{sim}(u, v) = \frac{u \cdot v}{\|u\| \cdot \|v\|}, \quad d_{\cos} = 1 - \text{sim}(u, v) \quad (2)$$

Como a comparação exaustiva de um vetor de consulta contra milhões de vetores indexados é onerosa, a prática recorre à busca por vizinhança aproximada, que troca garantia de exatidão por

redução expressiva de custo computacional. Malkov e Yashunin (2020) formalizaram o índice de grafos navegáveis hierárquicos de mundo pequeno, hoje a estrutura mais difundida para esse fim, e Johnson, Douze e Jégou (2021) descreveram a implementação eficiente dessas buscas em unidades de processamento gráfico. A expressão vetores de aproximação, empregada correntemente no ambiente institucional, designa precisamente esse arranjo: vetores densos consultados por algoritmos de vizinhança aproximada.

Dois parâmetros governam a qualidade prática dessa recuperação, e a Equação 3 os reúne. O primeiro é o número de vizinhos retornados, comumente designado top-k, que define quanto contexto será oferecido ao modelo gerador. O segundo é o corte de distância, aqui denotado pela letra tau, que descarta vizinhos geometricamente próximos mas semanticamente irrelevantes. Na notação da equação, x é a consulta, a função e projeta um texto no espaço vetorial, Z é o conjunto dos trechos indexados que o filtro de acesso já autorizou para aquela consulta, e $\text{posto}(z)$ é a posição do trecho z quando esse conjunto é ordenado por distância crescente em relação à consulta: o vizinho mais próximo tem posto 1, o seguinte tem posto 2, e assim por diante. O conjunto efetivamente recuperado é, portanto, a interseção entre os k primeiros dessa ordenação e os que satisfazem o corte.

$$R(x) = \{z \in Z : \text{posto}(z) \leq k\} \cap \{z \in Z : d_{\cos}(e(x), e(z)) \leq \tau\}$$

(3)

A escolha desses dois valores é a decisão de engenharia mais consequente de todo o arranjo, e é também a menos documentada na literatura. Um corte muito restritivo empobrece o contexto e faz o modelo responder que nada encontrou sobre matéria que consta

dos autos; um corte muito frouxo introduz ruído que o modelo tende a tratar como evidência, e o resultado é uma resposta bem escrita apoiada em trecho impertinente. As seções 4 e 5 registram os valores adotados em cada agente e a razão pela qual eles diferem entre si.

2.4. Recuperação Aumentada por Geração

Lewis et al. (2020) propuseram a recuperação aumentada por geração, conhecida pela sigla RAG, como arquitetura que acopla um recuperador de documentos a um modelo gerador de linguagem. Em vez de depender apenas do conhecimento internalizado durante o treinamento, o modelo recebe, junto da pergunta, os trechos recuperados de uma base externa e formula a resposta sobre eles. Formalmente, a probabilidade da resposta y dada a pergunta x é obtida por marginalização sobre os trechos recuperados, conforme a Equação 4, na qual o primeiro fator é atribuído pelo recuperador e o segundo pelo gerador.

$$p(y \mid x) \approx \sum_{z \in R(x)} p_{\eta}(z \mid x) \cdot p_{\theta}(y \mid x, z) \quad (4)$$

O segundo fator, por sua vez, é a fatoração autorregressiva usual da geração de texto, condicionada não apenas à pergunta, mas também ao trecho recuperado, conforme a Equação 5.

$$p_{\theta}(y \mid x, z) = \prod_{i=1}^n p_{\theta}(y_i \mid x, z, y_{<i}) \quad (5)$$

A leitura institucional dessas duas expressões é mais importante do que a sua manipulação algébrica. A Equação 4 mostra que a qualidade da resposta é limitada superiormente pela qualidade do conjunto recuperado: se o trecho pertinente não estiver em $R(x)$, nenhum ajuste no gerador o trará de volta, porque ele simplesmente

não integra a soma. A Equação 5 mostra que, uma vez fixado o trecho, cada token da resposta é escolhido condicionado a ele, o que explica por que a ancoragem funciona e por que ela não é garantia: o condicionamento é estatístico, não é uma restrição rígida, e o modelo pode atribuir massa de probabilidade a continuações que o trecho não sustenta.

Gao et al. (2023), em revisão sistemática do campo, distinguem três gerações de arranjos: o RAG ingênuo, que recupera e concatena; o RAG avançado, que acrescenta reescrita de consulta, reordenação e compressão de contexto; e o RAG modular, que compõe recuperadores e geradores especializados. Os autores destacam que a qualidade final depende mais da etapa de recuperação do que da capacidade do gerador, conclusão convergente tanto com a leitura das equações acima quanto com a experiência relatada neste artigo.

A relação entre recuperação aumentada e alucinação merece precisão. O padrão reduz substancialmente a incidência de fabricação factual, porque restringe o material sobre o qual o modelo se pronuncia, mas não a elimina. O modelo pode ignorar o contexto fornecido, combinar indevidamente trechos de documentos distintos ou citar identificadores de fonte que não constavam do conjunto recuperado. Por essa razão, o desenho descrito na seção 4 não se limita a recuperar e gerar: ele verifica, depois da geração, se cada referência citada pertence de fato ao conjunto recuperado, e descarta as que não pertencem.

2.5. Reconhecimento de Fala, Alinhamento Temporal e Diarização

Radford et al. (2023) apresentaram o Whisper, modelo de reconhecimento de fala treinado por supervisão fraca em larga

escala, que estabeleceu novo patamar de robustez em transcrição automática, inclusive em português e em condições acústicas adversas, como as de uma sala de audiência com microfone distante e ruído de fundo. O modelo, contudo, produz marcações temporais aproximadas, obtidas por inferência do próprio decodificador. Bain et al. (2023) propuseram o WhisperX, que acrescenta uma etapa de alinhamento forçado por modelo fonético e produz marcação temporal em nível de palavra, além de integrar a segmentação por locutor.

A atribuição de trechos a interlocutores, denominada diarização, é tratada por Bredin (2023) com o uso de representações vetoriais de voz e agrupamento sobre o áudio completo. A distinção é relevante para a aplicação forense: quando a atribuição é decidida fragmento a fragmento, o mesmo depoente pode receber rótulos diferentes ao longo da audiência, e a transcrição perde valor probatório justamente no ponto em que se pretendia útil; quando é decidida sobre o áudio inteiro, a identidade permanece consistente do início ao fim. Ambas as etapas, o alinhamento forçado e a diarização, são intensivas em multiplicação matricial e se beneficiam de execução em unidade de processamento gráfico, razão pela qual a instituição mantém uma placa NVIDIA dedicada a esse serviço, conforme detalhado na subseção 5.2.

2.6. Governança de Inteligência Artificial no Setor Público Brasileiro

O uso de inteligência artificial no sistema de justiça brasileiro já dispõe de marco regulatório próprio. A Resolução nº 332, de 21 de agosto de 2020, do Conselho Nacional de Justiça, estabelece diretrizes de ética, transparência e governança para a produção e o

uso de inteligência artificial no Poder Judiciário, exigindo explicabilidade, controle humano e prestação de contas sobre os modelos empregados. Embora dirigida ao Judiciário, a resolução constitui referência técnica pertinente às demais instituições do sistema de justiça, entre elas a Defensoria Pública, e é adotada como parâmetro voluntário no desenho aqui relatado.

A Lei nº 13.709, de 14 de agosto de 2018, Lei Geral de Proteção de Dados Pessoais, incide de forma direta sobre a operação descrita, uma vez que o material processado contém dados pessoais e, com frequência, dados pessoais sensíveis, na acepção do art. 5º, inciso II, do diploma. Os princípios da finalidade, da necessidade e da segurança condicionam decisões concretas de arquitetura, como a exclusão automática dos arquivos enviados ao término de cada análise, a supressão do nome do assistido no texto submetido ao modelo e a recusa de exclusão de conta que deixou rastro em trilha de auditoria, todas descritas adiante. A Lei nº 12.527, de 18 de novembro de 2011, Lei de Acesso à Informação, por sua vez, fundamenta a frente de transparência ativa integrada ao mesmo ecossistema de dados.

No plano da política pública, o Plano Brasileiro de Inteligência Artificial (MCTI; CGEE, 2025) organiza a ação federal em eixos, dos quais o terceiro trata especificamente da inteligência artificial para a melhoria do serviço público, com metas de três e de cinco anos que incluem a adoção da tecnologia em áreas críticas, a capacitação de parcela expressiva do corpo de servidores e a construção de infraestrutura nacional de dados e de catálogo nacional de metadados. Entre as premissas do plano está a de que a tecnologia seja centrada no ser humano, acessível a todos e orientada a desafios sociais, e a de que o seu uso observe a ética, a

responsabilidade e a conformidade com a legislação de proteção de dados. A iniciativa relatada neste artigo antecede o plano e converge com ele, o que permite tratá-la como evidência empírica sobre a exequibilidade das metas ali fixadas em uma instituição estadual de porte médio.

No plano das soluções já tornadas públicas, observa-se predominância de iniciativas voltadas à classificação processual e à triagem de peças no âmbito dos tribunais, com menor registro documentado de aplicações em instituições de assistência jurídica. A lacuna se explica em parte pela natureza do material envolvido: enquanto o tribunal opera sobre peças formalizadas, a Defensoria recebe o relato oral do cidadão, transcrito por servidor em texto informal, com erro de digitação e vocabulário não técnico. A subseção 5.1 demonstra que essa característica produz efeitos mensuráveis sobre a tokenização e, por consequência, sobre o dimensionamento das requisições de codificação vetorial.

3. METODOLOGIA E ARQUITETURA

3.1. Delineamento da Pesquisa

A pesquisa é aplicada quanto à finalidade, descritiva e explicativa quanto aos objetivos, e conduzida sob a lógica da design science research, abordagem em que o conhecimento é produzido por meio da construção e da avaliação de um artefato destinado a resolver um problema de classe conhecida. O artefato, neste caso, é o assistente jurídico de consulta aos autos e o ecossistema que o sustenta, tomados como caso instrumental: o interesse não reside na plataforma em si, mas no que a sua construção revela sobre a

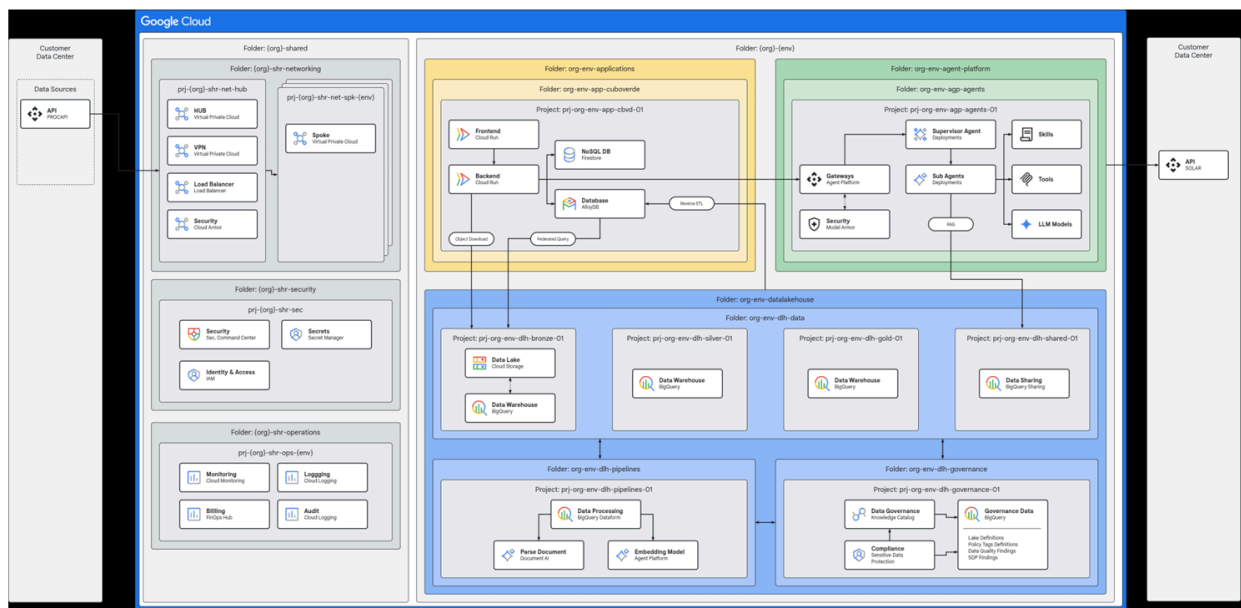
incorporação governada de modelos de linguagem em instituições públicas de assistência jurídica.

O universo de observação corresponde à operação da Defensoria Pública do Estado do Paraná, com aproximadamente 700 servidores além do quadro de defensores públicos, e o período considerado vai da implantação inicial, em junho de 2025, ao levantamento realizado para este artigo, em 2026. Os dados analisados são de três naturezas: registros de telemetria da própria plataforma, como latência de resposta, contagem de tokens e número de trechos recuperados; medições instrumentadas conduzidas pela equipe técnica para decidir parâmetros de arquitetura, sempre descritas com o valor observado e a decisão dele derivada; e indicadores institucionais consolidados de economia de tempo e de alcance das soluções.

3.2. Arquitetura de Referência

O ecossistema opera integralmente em nuvem pública, sobre uma zona de aterrissagem que separa responsabilidades em pastas e projetos distintos, conforme a Figura 1. Essa separação não é apenas organizacional: ela define fronteiras de permissão, de faturamento e de raio de falha, de modo que um incidente em um ambiente não alcança o outro e o custo de cada frente é atribuível sem rateio.

Figura 1. Arquitetura de referência do assistente jurídico e do ecossistema Cubo Verde em nuvem pública (parceria Google e CELEPAR)



Fonte: elaborado pelos autores (2026).

O conjunto compartilhado reúne três domínios. O primeiro é o de rede, organizado no padrão concentrador e raio, com nuvem privada virtual central, canal de rede privada virtual para o ambiente local, balanceador de carga e filtro de borda para proteção contra ataques de aplicação. O segundo é o de segurança, que concentra a central de comandos de segurança, o cofre de segredos e a gestão de identidade e acesso. O terceiro é o de operações, que abriga monitoramento, registro de eventos, auditoria e acompanhamento de custos. Nenhum desses recursos é duplicado por aplicação, o que evita a proliferação de configurações divergentes de segurança entre projetos e mantém em um único lugar a resposta à pergunta de quem acessou o quê.

O conjunto por ambiente contém três blocos. O bloco de aplicação hospeda a interface e as interfaces de programação em contêineres gerenciados sem servidor, um banco de dados relacional gerenciado, que abriga também o índice vetorial, e um repositório de documentos não relacional para o estado dos trabalhos assíncronos. O bloco de plataforma de agentes reúne as portas de entrada dos modelos, o agente supervisor, os subagentes especializados, as habilidades e ferramentas que eles invocam, os

modelos de linguagem propriamente ditos, o mecanismo de recuperação aumentada e uma camada de filtragem de conteúdo aplicada às entradas e às saídas. O bloco analítico adota a organização em camadas conhecida como medalhão, com repositório de dados brutos, camada intermediária de dados tratados, camada de consumo e camada de compartilhamento, além de projetos dedicados a processamento e a governança.

Duas integrações atravessam a fronteira da nuvem, ambas com o centro de dados da instituição: a interface do sistema administrativo e a interface do sistema de processos judiciais, que é a origem dos documentos consumidos pelo assistente jurídico e, portanto, o ponto de partida de tudo o que a seção 4 descreve. O projeto de governança mantém catálogo de conhecimento, proteção de dados sensíveis e um repositório de achados que registra definições de lago, marcações de política, resultados de qualidade de dados e ocorrências de identificação de dado sensível.

Os modelos de linguagem são servidos pela plataforma gerenciada do provedor de nuvem, e a família adotada é a Gemini, em duas variantes: uma otimizada para latência, empregada nas tarefas de classificação, extração e conversação; e outra otimizada para capacidade de raciocínio, reservada às análises mais longas. Entre a aplicação e os modelos há um intermediador próprio, hospedado em máquina virtual da instituição, que emite uma chave virtual por agente, aplica orçamento e registra o gasto por chamada. Essa camada existe por razão de governança, e não de desempenho: sem ela, a instituição saberia quanto gastou com inteligência artificial, mas não com o quê, e a transcrição de áudio, que consome ordens de grandeza mais tokens que texto, esconderia sozinha o custo de todos os demais agentes.

3.3. Portal de Acesso Interno e Interfaces de Programação

O Cubo Verde se oferece à instituição em duas formas complementares, e essa dualidade é traço central do desenho. A primeira é o portal, endereço único no qual o profissional autenticado encontra as soluções agrupadas por categoria e filtradas pelas suas permissões, conforme as Figuras 2 a 4. A segunda é o conjunto de interfaces de programação que expõem as mesmas funcionalidades como serviço, para que os sistemas institucionais as consumam sem que o usuário precise sair do sistema em que já trabalha.

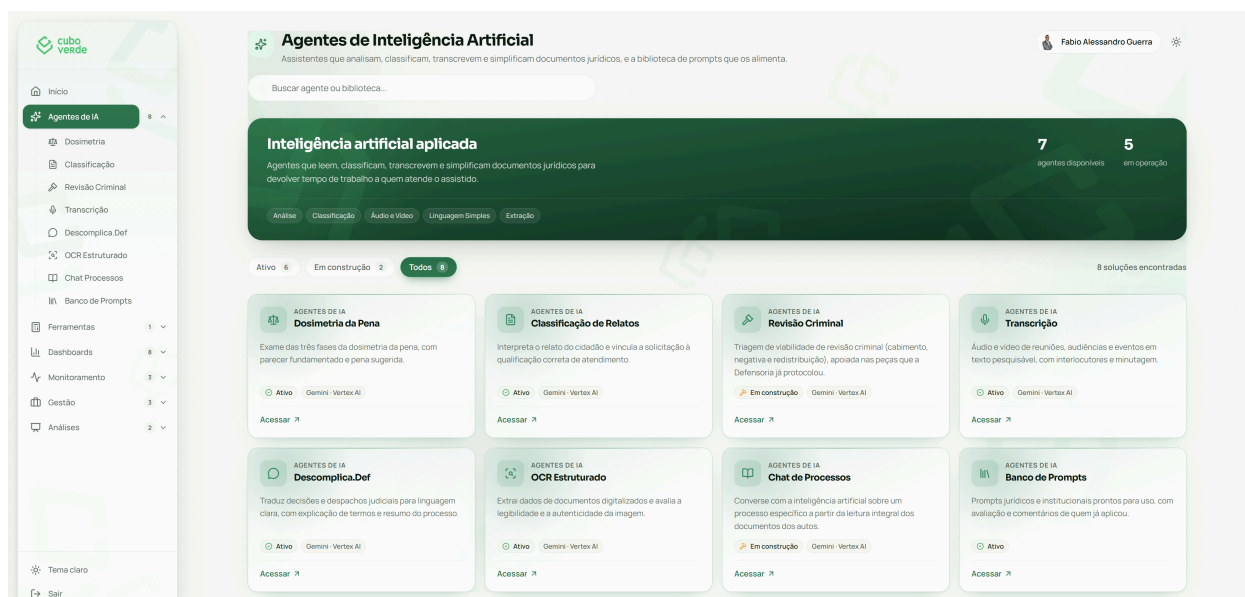
A porta de entrada, reproduzida na Figura 2, exprime a política de identidade detalhada na subseção 5.6. O caminho ordinário é a autenticação federada pela conta institucional, único mecanismo oferecido em primeiro plano; a alternativa, discreta ao pé do cartão, atende às contas externas concedidas a Defensorias de outros estados e a parceiros institucionais, que não possuem endereço no domínio da instituição e, por isso, jamais retornariam do provedor de identidade.

Figura 2. Tela de entrada do Cubo Verde, com a autenticação federada pela conta institucional e o acesso alternativo destinado às contas externas



Fonte: captura da plataforma em produção, elaborada pelos autores (2026).

Figura 3. Portal do Cubo Verde na categoria de agentes de inteligência artificial, com o estado de disponibilidade e o modelo declarados em cada cartão

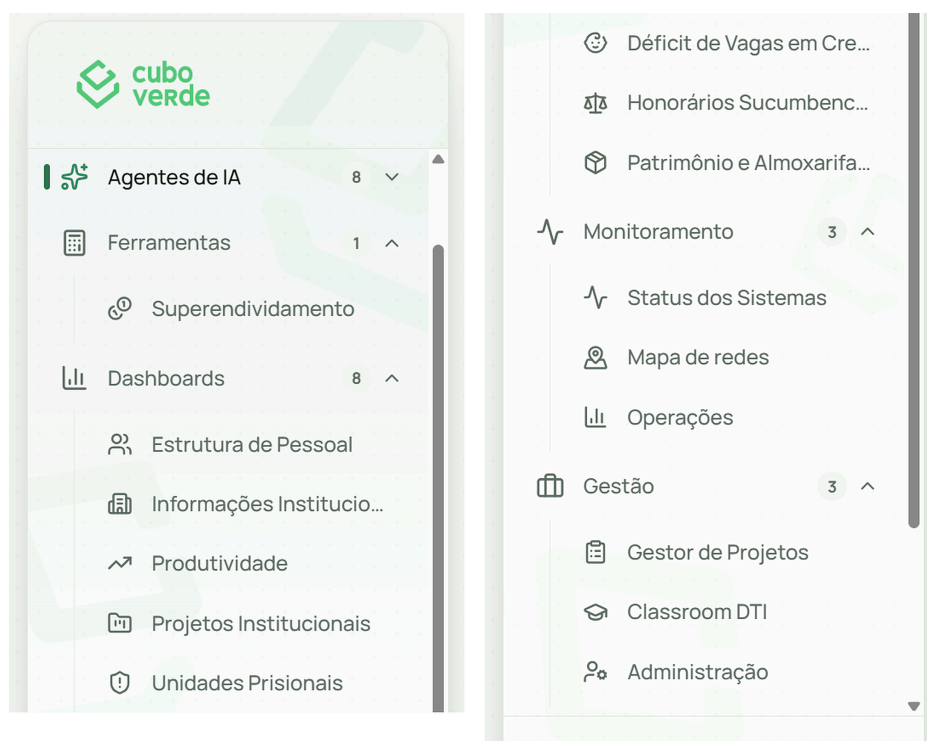


Fonte: captura da plataforma em produção, elaborada pelos autores (2026).

A Figura 3 apresenta a categoria de agentes de inteligência artificial tal como ela se oferece a um perfil com acesso integral: oito soluções, cada uma com o nome, a função enunciada em uma linha, o estado de disponibilidade e o modelo que a serve. Declarar o modelo no próprio cartão não é ornamento de interface; é a aplicação, na camada visível, do dever de transparência discutido na subseção 2.6, pois o profissional sabe, antes de acionar a ferramenta,

que está diante de um sistema de inteligência artificial e qual modelo o sustenta.

Figura 4. Menu lateral do portal, no qual cada categoria abre os itens que o profissional pode acessar; a figura reúne lado a lado as duas metades de uma lista que a tela exibe em coluna única



Fonte: captura da plataforma em produção, elaborada pelos autores (2026).

O portal responde diretamente ao diagnóstico de fragmentação apresentado na introdução: antes do Cubo Verde, painéis, ferramentas e sistemas conviviam como endereços autônomos, com credenciais e critérios de acesso próprios, e a primeira dificuldade de quem precisava de uma informação era descobrir onde ela morava. O menu lateral, reproduzido em parte na Figura 4, organiza o catálogo em seis categorias, Agentes de Inteligência Artificial, Ferramentas, Dashboards, Monitoramento, Gestão e Análises, e a categoria cujos itens a pessoa não pode acessar desaparece por inteiro, de modo que o portal de cada profissional mostra o que ele usa e não o que ele não pode usar.

As interfaces de programação são a outra metade, e são elas que levam a inteligência artificial ao ponto de atendimento. O SOLAR, sistema institucional de atendimento e acompanhamento processual, consome os serviços do Cubo Verde no próprio fluxo de trabalho: a classificação do relato acontece enquanto o servidor registra o atendimento, e o assistente jurídico é acionado a partir do processo que o defensor já tem aberto, sem exportação manual de peças. O LUNA, aplicativo de atendimento voltado ao cidadão, consome os mesmos serviços na ponta oposta da relação, na etapa em que a demanda ainda está sendo formulada em linguagem comum. Adotar a interface de programação como forma canônica de entrega, e o portal como uma entre as suas consumidoras, evita que cada sistema construa a sua própria versão da mesma regra e mantém uma única implementação de permissão, de registro e de orçamento por agente.

O catálogo em operação reúne 24 soluções, das quais oito são agentes de inteligência artificial. O Quadro 1 sintetiza esses agentes, a função de cada um e a técnica predominante que os sustenta; o primeiro deles, o assistente jurídico de consulta aos autos, é objeto da seção 4, e os demais são descritos na seção 5.

Quadro 1. Agentes de inteligência artificial em operação no Cubo Verde

Agente	Função	Técnica predominante
Assistente Jurídico (Chat de Processos)	Responde perguntas em linguagem natural sobre os autos do processo	Recuperação aumentada por geração, com citação verificada

Classificação de Relatos	Vincula o relato do cidadão à área e ao assunto de atendimento	Vetores densos, vizinhança aproximada e instrução de poucos exemplos
Transcrição	Converte audiências e reuniões em texto com marcação temporal	Whisper e WhisperX em unidade gráfica NVIDIA, com diarização
Dosimetria da Pena	Examina as três fases da dosimetria e sugere parecer fundamentado	Modelo de linguagem com esquema de saída
Descomplica.Def	Traduz decisões e despachos para linguagem acessível ao cidadão	Modelo de linguagem com instrução controlada
OCR Estruturado	Extraí e valida campos de documentos digitalizados ou fotografados	Reconhecimento óptico e extração com esquema
Banco de Prompts	Cura e compartilha instruções jurídicas institucionais	Busca textual com radicalização e remoção de acento
Revisão Criminal	Tria a viabilidade de revisão criminal, nos termos do art. 621 do CPP	Recuperação sobre o acervo de peças já protocoladas

Fonte: elaborado pelos autores (2026).

3.4. Separação de Cargas no Ambiente de Execução

A mesma imagem de contêiner sobe em dois serviços distintos, diferenciados por uma variável de ambiente que determina quais rotas cada um expõe. O primeiro serviço atende as requisições interativas; o segundo executa exclusivamente trabalhos longos, despachados por filas gerenciadas. A razão dessa divisão é

quantitativa, e convém registrá-la com o número que a produziu. A plataforma de contêineres escala por utilização de concorrência: com o limite fixado em dezesseis requisições simultâneas por instância, cinco requisições longas bloqueadas mantêm a utilização em torno de trinta por cento, e a segunda instância nunca é criada. Em medição realizada em agosto de 2026, transcrição, conversação e classificação disputavam uma única instância, com uma execução de 382 segundos ao lado de um fluxo de conversação de 91 segundos. No serviço de trabalho, o limite de concorrência foi reduzido a duas requisições por instância, o que força a criação de nova instância a cada par de trabalhos pesados e produz o paralelismo pretendido.

A adoção de trabalhos assíncronos decorre de medição igualmente direta. Três requisições síncronas de análise sobre um processo de 1.339 páginas responderam com sucesso no servidor, em 556, 275 e 279 segundos, e nenhuma delas chegou ao navegador do defensor: o corte ocorreu em equipamento intermediário de rede, fora do controle da aplicação. A repetição automática da requisição pelo cliente reprocessou o material integralmente, com custo de 4.017 páginas de reconhecimento óptico pagas por um resultado que ninguém recebeu. Desde então, todo trabalho cuja duração esperada ultrapasse a ordem de minutos é submetido como trabalho assíncrono, com acompanhamento por fluxo de eventos enviados pelo servidor, e as filas são três, separadas por classe de carga, para que uma audiência longa não bloqueie a fila em que trafega a classificação de um relato de três linhas.

3.5. Instrumentação e Coleta dos Dados

As medições apresentadas nas seções seguintes foram obtidas a partir do registro estruturado de eventos da própria plataforma. Até agosto de 2026 esse registro era insuficiente para sustentar análise: o registrador raiz da aplicação não possuía manipulador configurado, de modo que 979 de mil entradas amostradas em 48 horas chegavam ao repositório de eventos sem indicação de severidade, e nenhuma delas portava identificador de rastreamento. A correção consistiu em um formatador que emite cada evento como documento estruturado de linha única, acrescido de um componente que propaga o identificador de rastreamento da requisição a todas as linhas por ela geradas. Somente após essa correção os indicadores de latência, de taxa de erro e de acionamento de rota alternativa passaram a ser mensuráveis, circunstância que a subseção 6.2 retoma como lição de método.

As capturas de tela apresentadas neste artigo foram extraídas de gravação da plataforma em operação. Todos os dados pessoais nelas contidos, incluindo nomes de partes e de interlocutores, números de documentos de identificação e números de processos submetidos a segredo de justiça, foram suprimidos por tarja opaca antes da composição do texto, em atenção aos princípios de finalidade e de necessidade previstos na Lei nº 13.709, de 2018. Uma captura do agente de tradução para linguagem acessível foi deliberadamente excluída do conjunto, por descrever caso concreto de direito de família envolvendo criança, hipótese em que a tarja sobre o nome não bastaria para afastar o risco de reidentificação. Os elementos técnicos preservados nas imagens, como identificação do modelo, latência, contagem de tokens e número de trechos recuperados, não permitem reidentificação e são justamente o que sustenta a análise.

4. O ASSISTENTE JURÍDICO DE CONSULTA AOS AUTOS

O assistente jurídico, designado internamente como Chat de Processos, é a aplicação de maior impacto do ecossistema e a que melhor exprime a tese deste artigo. Ele permite que o defensor pergunte, em linguagem natural, sobre um processo inteiro, e receba resposta redigida sobre os documentos daquele processo, com indicação da peça que sustenta cada afirmação. Segundo levantamento institucional, trata-se da primeira aplicação de recuperação aumentada por geração integrada a sistema processual entre as Defensorias Públicas brasileiras. As subseções seguintes percorrem a cadeia completa, da importação dos autos à resposta verificada.

4.1. Do Processo Eletrônico Ao Índice Vetorial

A ingestão começa no sistema processual da instituição, e não em um envio manual de arquivos. Quando um processo é liberado ao assistente, a plataforma solicita ao SOLAR a lista de peças e baixa cada uma delas, preservando o identificador de origem, o tipo documental e a data, informações que voltarão a aparecer, mais adiante, na citação exibida ao defensor. O Quadro 2 resume as etapas que se seguem.

Quadro 2. Etapas da ingestão documental do assistente jurídico

Etapas	Operação	Decisão de projeto
Importação	Coleta das peças do processo a partir do sistema processual	Preserva identificador, tipo e data de cada peça
Extração de texto	Camada de texto do PDF quando existe; reconhecimento óptico quando não	Seletiva, para não pagar reconhecimento sobre peça já pesquisável

Fragmentação	Divisão do texto em trechos de tamanho compatível com o codificador	Trecho é a unidade de recuperação e de citação
Codificação	Geração do vetor denso de 768 dimensões de cada trecho	Finalidade declarada como documento, e não como consulta
Indexação	Gravação do vetor no índice do banco relacional, com metadados	Índice no mesmo banco da aplicação, sob a mesma trilha de acesso

Fonte: elaborado pelos autores (2026).

A extração de texto é seletiva por decisão de custo e de qualidade. Peça nativa digital já traz camada de texto pesquisável, e submetê-la a reconhecimento óptico seria pagar duas vezes pelo mesmo conteúdo, com perda de fidelidade; peça digitalizada, ao contrário, é imagem e exige reconhecimento. A plataforma tenta primeiro a camada de texto e encaminha ao reconhecimento óptico apenas o que voltar vazio, registrando em coluna própria por qual caminho cada peça foi extraída. Esse registro tem consequência prática: quando o extrator melhora, é possível reprocessar somente o que veio de reconhecimento óptico, em vez de reindexar o acervo inteiro.

A fragmentação define a unidade sobre a qual todo o restante opera. Trecho grande demais dilui o sinal, porque o vetor passa a representar a média de assuntos distintos e a similaridade cai para qualquer consulta específica; trecho pequeno demais perde o referente, porque a frase que menciona a decisão fica separada da frase que diz de que decisão se trata. A unidade adotada procura manter íntegro o parágrafo jurídico, e é ela que aparece, depois, como alvo da citação: quando o defensor confere a referência, ele é

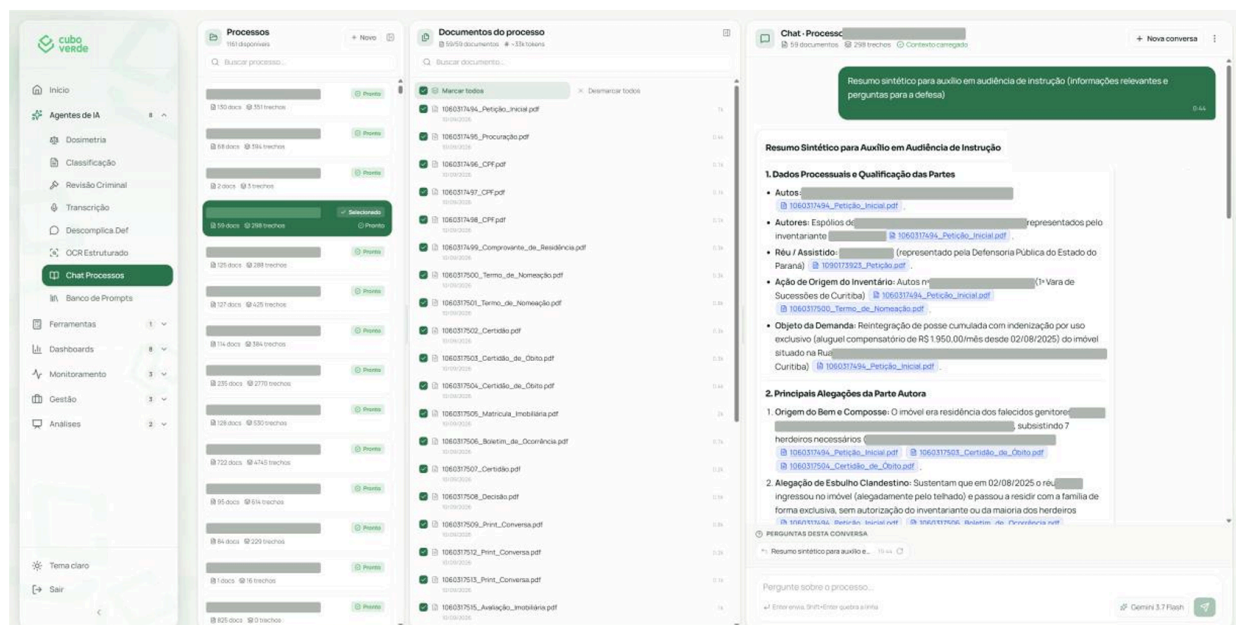
levado a um trecho legível, e não a um documento de oitenta páginas.

A codificação obedece a uma exigência que merece registro, porque é silenciosa e frequentemente ignorada. O modelo de codificação utilizado aceita um parâmetro que declara a finalidade do texto submetido, distinguindo o texto que está sendo indexado do texto que está sendo consultado. Medição conduzida pela equipe registrou similaridade de cosseno de 0,784 entre as duas codificações do mesmo texto, isto é, os dois vetores não coincidem. Indexar um acervo sem declarar essa finalidade grava documento como se fosse pergunta: a busca não falha, não emite erro e não degrada de forma visível; ela apenas passa a recuperar pior. Por essa razão, o parâmetro foi tornado obrigatório na interface interna de codificação, sem valor padrão, de modo que quem escrever o próximo indexador seja obrigado a decidir.

4.2. Contabilidade Explícita do Contexto

Concluída a ingestão, o processo passa a exibir ao defensor a contabilidade do que foi efetivamente indexado, conforme a Figura 5. Não se trata de detalhe de interface: é a informação que permite ao profissional julgar se pode confiar na resposta que virá.

Figura 5. Os três painéis do assistente jurídico: índice de processos, seleção das peças indexadas e conversa ancorada nos autos



Fonte: captura da plataforma em produção, com supressão dos números de processo e dos nomes das partes, dos herdeiros e do endereço do imóvel litigioso, elaborada pelos autores (2026).

A figura mostra, à esquerda, o índice dos processos disponíveis ao profissional, 1.161 no caso reproduzido, cada um com a contagem de peças e de trechos e com o estado da indexação; ao centro, a lista das peças do processo selecionado, todas marcadas por padrão, com 59 de 59 documentos e aproximadamente 33 mil tokens; e, à direita, o painel de conversa, cujo cabeçalho repete a contabilidade na unidade que efetivamente importa para a recuperação, 59 documentos e 298 trechos, e informa que o contexto está carregado. Três consequências decorrem dessa exibição. A primeira é que o defensor sabe, antes de perguntar, se o processo está integralmente indexado ou se ainda há peça em processamento, distinção que a resposta sozinha não revelaria. A segunda é que a seleção é editável: ao desmarcar peças, o profissional restringe o universo da busca, o que é útil quando a pergunta se refere a uma fase específica do processo e o restante do acervo apenas introduz ruído. A terceira é de ordem pedagógica, e não deve ser subestimada: exibir a contagem de trechos ensina, sem treinamento formal, que o assistente não lê o processo inteiro a cada pergunta, e sim recupera partes dele.

4.3. Consulta, Recuperação e Geração Ancorada

Formulada a pergunta, a plataforma percorre quatro etapas. Na primeira, a pergunta é codificada em vetor denso, agora com a finalidade declarada como consulta. Na segunda, esse vetor é confrontado com os vetores dos trechos daquele processo, e a recuperação obedece à forma da Equação 3, restrita ao subconjunto de peças selecionadas pelo defensor. A busca é híbrida: à recuperação densa soma-se um ramo léxico, que garante o acerto exato de número de processo, valor, data e nome próprio, precisamente os elementos em que a representação semântica é menos confiável, porque dois valores monetários distintos ocupam posições vizinhas no espaço vetorial. Na terceira etapa, os trechos recuperados são compostos com a pergunta e submetidos ao modelo, que recebe instrução de responder exclusivamente sobre eles e de identificar a fonte de cada afirmação. Na quarta, descrita na subseção 4.5, as citações produzidas são conferidas.

O resultado é apresentado na Figura 6, que reproduz uma resposta real do assistente em ação de reintegração de posse cumulada com pedido de indenização por uso exclusivo de imóvel, na qual a Defensoria atua em favor do réu, herdeiro do bem disputado. À pergunta sobre qual seria o próximo passo da instituição no processo, o agente responde com o encerramento infrutífero da fase conciliatória e o marco de contagem do prazo, com a prerrogativa de prazo em dobro e, em seguida, com as teses a abordar na peça defensiva, cada uma remetida à peça dos autos que a sustenta.

Figura 6. Resposta do assistente jurídico ancorada nos autos, com as citações de tipo documental embutidas no corpo do texto

Chat - Processo

59 documentos 298 trechos Contexto carregado

+ Nova conversa

Qual o próximo passo da Defensoria?10:46

O próximo passo processual da Defensoria Pública é a **apresentação da Contestação** em favor do assistido [REDACTED]

Fundamentação e Contexto Processual:

- Encerramento da fase conciliatória:** A audiência de conciliação realizada em 08/07/2026 perante o CEJUSC restou infrutífera [REDACTED] [Termo de Audiência](#). Conforme previsto no art. 335, inciso I, do Código de Processo Civil e expressamente advertido na decisão inicial [REDACTED] [Decisão](#), o prazo para resposta da parte ré passa a fluir a partir da referida data.
- Prerrogativa da Defensoria Pública:** Nos termos do art. 186, *caput*, do CPC c/c as Leis Complementares nº 80/1994 e nº 136/2011 (PR), a Defensoria Pública dispõe de prazo em dobro (30 dias úteis) para contestar a ação, contado a partir da intimação pessoal/eletrônica.

Principais Teses e Pontos a Abordar na Peça Defensiva:

- Composse e Legitimidade da Posse de Herdeiro:** Demonstrar que o assistido é herdeiro necessário e compossuidor do bem (*saisine*, art. 1.784 do Código Civil), existindo condomínio *pro indiviso* sobre o imóvel até a partilha (art. 1.791, parágrafo único, do CC), o que afasta a caracterização de esbulho possessório nos moldes alegados pela parte autora [REDACTED] [Petição Inicial](#).
- Efeito Suspensivo no Agravo de Instrumento:** Ressaltar que os efeitos da liminar de reintegração de posse foram suspensos por decisão proferida no Agravo de Instrumento nº [REDACTED] [Decisão](#).
- Impugnação ao Pedido de Aluguéis Compensatórios:** Contestar o valor pretendido a título de aluguel (R\$ 1.950,00/mês), baseado em avaliação unilateral [REDACTED] [Avaliação Imobiliária](#), pugnano, se for o caso, por apuração pericial ou redução proporcional à quota-parte e à concordância de outros herdeiros.
- Reiteração da Gratuidade da Justiça:** Reforçar o pedido de concessão da gratuidade da justiça já formulado na habilitação [REDACTED] [Petição](#), instruída com a respectiva

Fonte: captura da plataforma em produção, com supressão do número do processo, do número do agravo de instrumento e do nome da pessoa assistida, elaborada pelos autores (2026).

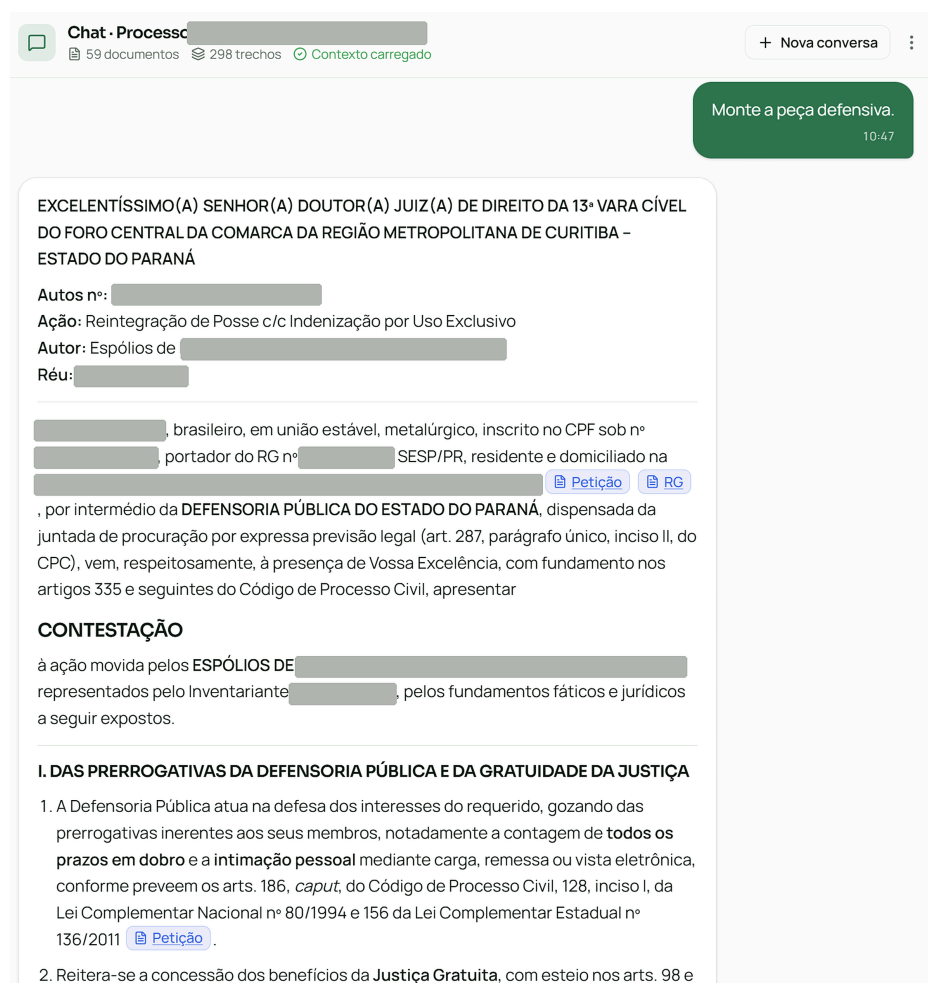
Duas características visíveis na figura sintetizam o comportamento do agente. A primeira é a citação embutida no corpo da resposta: cada afirmação relevante carrega o tipo do documento de origem, termo de audiência, decisão, petição inicial ou laudo de avaliação, em referência que conduz ao trecho recuperado, o que permite conferência imediata sem sair da tela. A segunda é a natureza do que o modelo produz: ele não decide nada, não recomenda conduta e não afirma resultado provável; organiza o que está nos autos e formula hipóteses de trabalho que o defensor aceita, descarta ou reformula. Essa delimitação é deliberada e responde à exigência de controle humano que a Resolução nº 332 de 2020 estabelece para o uso de inteligência artificial no sistema de justiça.

Vale notar, ainda, o que a figura mostra sobre o custo cognitivo evitado. A resposta reproduzida sintetiza um processo de 59 peças em meia página de leitura, com as referências que permitem ao defensor ir direto às cinco ou seis peças que importam para a audiência do dia seguinte. É desse tipo de ganho que decorre a redução de cerca de setenta por cento no tempo de análise processual apurada pela instituição, discutida na subseção 6.1: o assistente não substitui a leitura da peça, ele elimina a busca pela peça.

4.4. da Consulta à Minuta: A Peça Defensiva Construída Sobre os Autos

A mesma conversa admite um segundo tipo de pedido, de natureza distinta do anterior e de consequências mais delicadas: em vez de perguntar o que há nos autos, o defensor pede que o agente redija a peça. As Figuras 7 a 10 reproduzem, em quatro recortes sucessivos da mesma resposta, a contestação produzida a partir da instrução Monte a peça defensiva, dirigida ao mesmo processo das figuras anteriores.

Figura 7. Endereçamento e qualificação das partes na minuta produzida pelo assistente, com os dados pessoais recuperados das peças dos autos



Fonte: captura da plataforma em produção, com supressão do número do processo, dos nomes das partes e dos herdeiros e dos dados de qualificação do assistido, elaborada pelos autores (2026).

A Figura 7 exhibe o trecho de abertura, e nele está a razão pela qual esse pedido é possível: o endereçamento ao juízo competente, o número dos autos, a qualificação civil completa do assistido e a indicação do espólio autor não foram gerados pelo modelo, foram recuperados das peças indexadas e cada bloco traz a referência ao documento de onde saiu. É precisamente o trabalho de transcrição manual que consome a primeira meia hora de quem redige uma contestação, e é também aquele em que o erro de digitação passa despercebido até a intimação seguinte.

Figura 8. Prerrogativas, gratuidade da justiça e síntese da demanda, com a cadeia processual reconstituída a partir das peças

Chat - Processo

59 documentos 298 trechos Contexto carregado

+ Nova conversa

I. DAS PRERROGATIVAS DA DEFENSORIA PÚBLICA E DA GRATUIDADE DA JUSTIÇA

1. A Defensoria Pública atua na defesa dos interesses do requerido, gozando das prerrogativas inerentes aos seus membros, notadamente a contagem de **todos os prazos em dobro** e a **intimação pessoal** mediante carga, remessa ou vista eletrônica, conforme preveem os arts. 186, *caput*, do Código de Processo Civil, 128, inciso I, da Lei Complementar Nacional nº 80/1994 e 156 da Lei Complementar Estadual nº 136/2011 [Petição](#).
2. Reitera-se a concessão dos benefícios da **Justiça Gratuita**, com esteio nos arts. 98 e seguintes do CPC, uma vez que o assistido é pessoa economicamente hipossuficiente, auferindo renda mensal modesta (R\$ 2.500,00) e não dispondo de recursos para arcar com as custas e despesas processuais sem prejuízo do sustento próprio e de sua família, conforme Declaração de Hipossuficiência e Autodeclaração de Renda carreadas aos autos [Declaração de Hipossuficiência](#) [AUTODECLARAÇÃO DE RENDA](#).

II. SÍNTESE DA DEMANDA

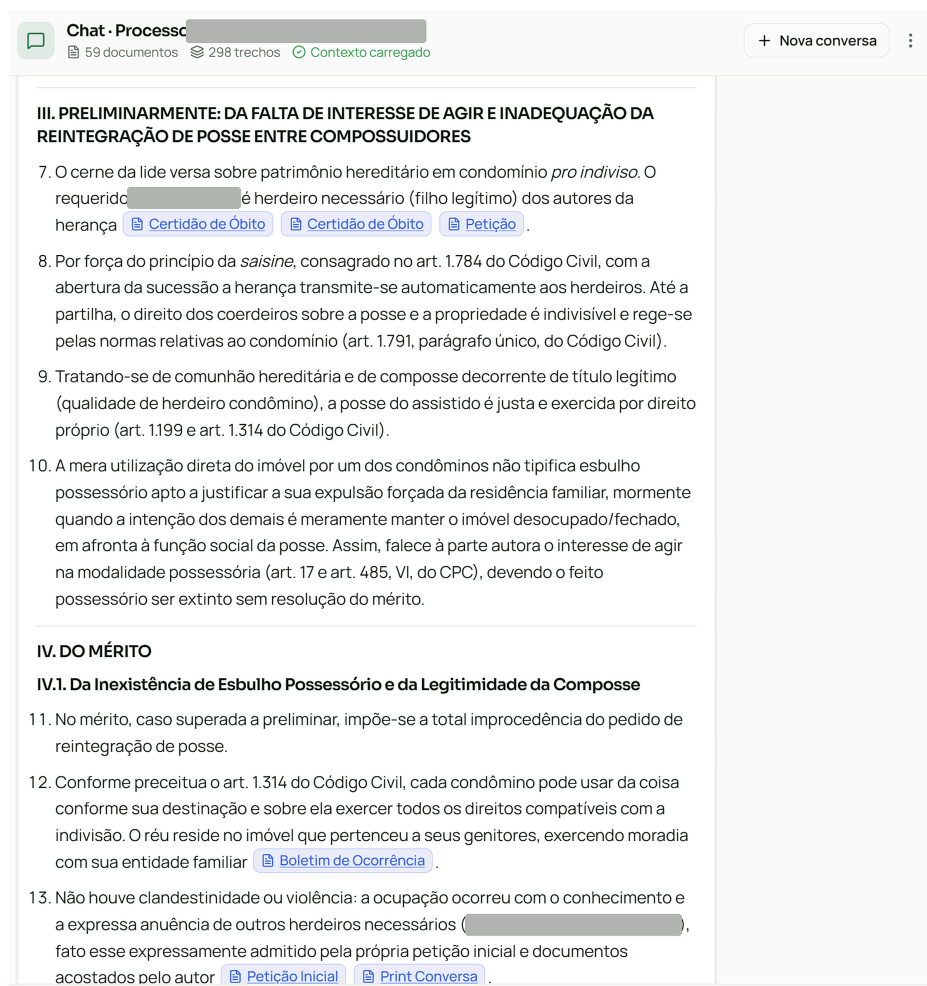
3. O Espólio autor, representado pelo inventariante [REDAZIDO] ajuizou a presente demanda sustentando que o acervo hereditário dos falecidos [REDAZIDO] composto pelo imóvel matriculado sob o nº [REDAZIDO] do 4º Registro de Imóveis de Curitiba/PR [Petição Inicial](#) [Matrícula Imobiliária](#).
4. Alega a parte autora que, em 02/08/2025, o réu teria invadido clandestinamente o imóvel e passado a ocupá-lo com exclusividade, o que caracterizaria esbulho possessório. Pleiteou medida liminar de reintegração de posse e a condenação do contestante ao pagamento de aluguel compensatório mensal de R\$ 1.950,00, além do pagamento dos encargos do imóvel [Petição Inicial](#).
5. A liminar foi inicialmente deferida pelo Juízo [Decisão](#), porém os seus efeitos restaram expressamente **suspensos** em razão da concessão de efeito suspensivo no bojo do Agravo de Instrumento nº [REDAZIDO] interposto pelo réu [Decisão](#).
6. Realizada audiência de conciliação no CEJUSC em 08/07/2026, a composição restou infrutífera [Termo de Audiência](#), vindo os autos para o oferecimento da presente resposta defensiva.

Fonte: captura da plataforma em produção, com supressão dos nomes do inventariante e dos falecidos, do número da matrícula do imóvel e do número do agravo de instrumento, elaborada pelos autores (2026).

A Figura 8 mostra a síntese da demanda, e é aqui que a recuperação demonstra o que a leitura sequencial dificilmente alcançaria no mesmo tempo: a minuta reconstitui a cadeia processual em seis itens numerados, do ajuizamento à audiência de conciliação infrutífera, passando pelo deferimento da liminar e pela suspensão de seus efeitos em agravo de instrumento, e ancora cada elo em peça distinta dos autos. O fato juridicamente decisivo do caso, o de que a liminar está suspensa, não está na petição inicial nem na decisão que a deferiu, e sim na decisão proferida no recurso; é o tipo

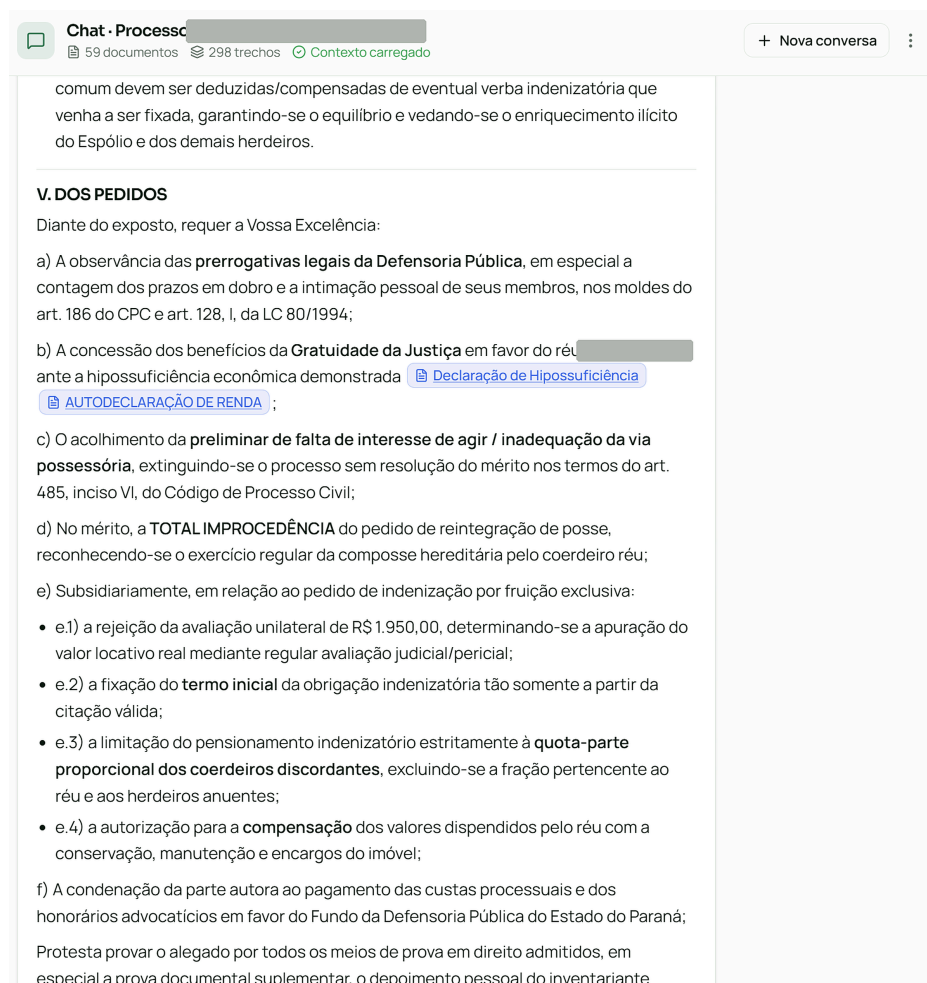
de informação que se perde quando o defensor recebe o processo na véspera da audiência.

Figura 9. Preliminar de falta de interesse de agir e início do mérito, com a fundamentação legal aplicada ao caso



Fonte: captura da plataforma em produção, com supressão do nome do assistido e dos nomes dos herdeiros anuentes, elaborada pelos autores (2026).

Figura 10. Rol de pedidos da minuta, com os requerimentos principais e os subsidiários



Fonte: captura da plataforma em produção, com supressão do nome do assistido, elaborada pelos autores (2026).

As Figuras 9 e 10 completam a peça com a preliminar de falta de interesse de agir, sustentada na composses hereditária e no princípio da saisine, e com o rol de pedidos, no qual os requerimentos subsidiários aparecem escalonados abaixo do pedido principal de improcedência. Cabe delimitar com rigor o que essas quatro figuras autorizam afirmar e o que não autorizam. Elas mostram que o agente é capaz de produzir uma minuta estruturada, tecnicamente coerente e ancorada nos autos, o que reduz o esforço de redação a partir da página em branco; não mostram que a peça esteja correta. A qualificação jurídica das teses, a decisão sobre quais delas sustentar e a responsabilidade pelo que é protocolado permanecem integralmente com o defensor público, que assina o documento. O agente entrega minuta, e minuta não é peça: é o rascunho que

dispensa o defensor de reconstituir sozinho aquilo que já está escrito nos autos.

4.5. Verificação Posterior das Citações

A verificação das citações é o ponto em que este desenho se afasta do RAG ingênuo descrito por Gao et al. (2023), e decorre diretamente da leitura das equações da subseção 2.4. Como o condicionamento sobre o trecho recuperado é estatístico, o modelo pode produzir, ao redigir uma resposta longa e bem estruturada, referências plausíveis para afirmações que não recuperou. A providência adotada é elementar e eficaz: após a geração, o conjunto de identificadores citados na resposta é interceptado com o conjunto de identificadores efetivamente recuperados, e o que não pertence aos dois é descartado, conforme a Equação 6.

$$\hat{C}(y) = C(y) \cap \text{id}(R(x)) \quad (6)$$

A operação é barata, determinística e não depende do modelo, três propriedades que a tornam preferível a qualquer tentativa de convencer o gerador a não inventar referências por meio de instrução. Sem ela, a citação funcionaria como ornamento de credibilidade e não como prova de origem, e o efeito seria perverso: uma resposta com referências falsas é mais perigosa do que uma resposta sem referência alguma, porque desarma justamente a conferência que o profissional faria. O mesmo princípio governa outro agente da plataforma, no qual os campos estruturados exibidos ao usuário, como título, tipo e número de processo do precedente, são lidos do registro institucional e não do texto gerado, aproveitando-se do modelo apenas a justificativa em prosa.

4.6. Orçamento de Contexto e o Documento Como Texto

O tratamento do documento obedece a uma decisão de arquitetura deliberada: o documento do usuário sempre entra no modelo como texto já extraído, nunca como arquivo anexado. Anexar o arquivo é mais curto de implementar e submete o agente a um teto que não pertence à instituição, pois acima de determinado número de páginas o serviço recusa a chamada inteira, levando junto todas as demais fontes da mesma mensagem, inclusive as que caberiam sem dificuldade. Com o texto extraído previamente, o único limite remanescente é o da janela de entrada do modelo, que a aplicação administra por recorte de início e fim do documento, preservando as partes que concentram o conteúdo decisório e registrando no evento estruturado quanto foi descartado.

Essa decisão tem contrapartida honesta, que convém enunciar: o recorte é uma perda de informação, e não uma compressão sem custo. A escolha é entre uma resposta produzida sobre parte do documento, com o registro explícito de que houve corte, e nenhuma resposta. A alternativa preferível, que a plataforma adota sempre que o material está indexado, é não depender do recorte: quando o processo passou pela ingestão descrita na subseção 4.1, o modelo não recebe o documento, recebe os trechos recuperados, e o orçamento de contexto deixa de ser um problema de tamanho para ser um problema de seleção, que é justamente o que a Equação 3 resolve.

4.7. Controle de Acesso Ao Acervo Processual

Um assistente que responde sobre autos precisa saber, antes de responder, se quem pergunta pode ler aqueles autos. A verificação ocorre em módulo comum de controle de acesso, e não nas rotas de conversação, decisão de projeto que merece registro porque foi

aprendida na prática: escrita na rota, a regra não alcançaria os demais pontos de consulta, que verificam permissão por conta própria, e a instituição teria uma trava que às vezes tranca, pior do que trava nenhuma.

A admissão de contas externas, destinadas a Defensorias de outros estados e a parceiros institucionais, tornou essa camada indispensável. A permissão de uso do assistente, concedida integralmente, abriria ao convidado o acervo inteiro de processos indexados pela instituição; a solução foi uma lista explícita de processos alcançáveis por essas contas, aplicada acima de qualquer outra configuração e comparada por dígitos, porque o número do processo é gravado ora pontuado, ora não. Requisições de sincronização e de importação permanecem vedadas a essas contas, pois consultar documentos e disparar ingestão a partir do sistema processual são operações de natureza distinta, com custos e riscos distintos.

5. DEMAIS AGENTES E SERVIÇOS DO ECOSSISTEMA

5.1. Classificação de Relatos

O agente de classificação atua na porta de entrada da instituição. Sua tarefa consiste em ler o relato registrado no atendimento, redigido em linguagem corrente pelo servidor que recebeu o cidadão, e vinculá-lo à taxonomia institucional composta por área e assunto, decisão que determina o encaminhamento da demanda. A Figura 11 apresenta o resultado de uma classificação, com os campos devolvidos pelo agente, a identificação do modelo empregado e a latência medida.

Figura 11. Classificação automática de relato, com indicação de área, assunto, retorno, modelo e latência

CLASSIFICAÇÃO DE RELATOS

Classificação inteligente das solicitações de atendimento

Análise automaticamente o relato enviado pelo cidadão, interprete seu contexto fático e jurídico e vincule a solicitação à qualificação correta de atendimento. A solução reduz retrabalho, elimina gargalos de classificação e agiliza o encaminhamento das demandas.

Relatos do cidadãoClassificação automáticaAtendimento

Relato do assistido

Preciso de ajuda para sair de casa com meus filhos por conta de meu esposo.

Classificar

Resultado

Classificação	Família e Sucessões - Divórcio (com ou sem divisão de bens, pensão alimentícia, guarda e visitas)
Área	Família e Sucessões
Assunto	Divórcio (com ou sem divisão de bens, pensão alimentícia, guarda e visitas)
Retorno	Não
Modelo	gemini-3.5-flash
Latência	5166ms

Fonte: captura da plataforma em produção; o relato submetido é genérico e não identifica pessoa alguma. Elaborada pelos autores (2026).

O fluxo compreende quatro etapas. Na primeira, o relato é convertido em vetor denso de 768 dimensões. Na segunda, esse vetor é confrontado, por busca de vizinhança aproximada, contra um acervo de relatos previamente classificados e validados pela instituição, do qual se recuperam os dez mais próximos, descartando-se os que ultrapassem o corte de distância de cosseno estabelecido em 0,5, na forma da Equação 3. Na terceira, os exemplos recuperados são compostos com o relato em uma instrução de poucos exemplos, no molde descrito por Brown et al. (2020), submetida ao modelo com esquema de saída fixo, que devolve área, assunto e a indicação de tratar-se ou não de retorno de atendimento anterior. Na quarta, um filtro posterior confronta a

resposta com a taxonomia vigente e descarta categoria que nela não exista, de modo que o agente jamais devolva rótulo inventado. É o mesmo princípio da Equação 6, aplicado a rótulos em vez de citações.

O acervo de exemplos cumpre função que o enunciado da tarefa não alcança. A taxonomia institucional distingue assuntos cuja fronteira é fina e cuja denominação formal raramente aparece no relato do cidadão, que não diz divórcio consensual com partilha de bens, e sim que precisa de ajuda para sair de casa com os filhos por conta do marido, exatamente como no relato da Figura 11, vinculado pelo agente ao assunto Divórcio da área de Família e Sucessões. A recuperação de casos análogos oferece ao modelo não a definição do assunto, mas a maneira concreta pela qual pessoas descreveram situações que a instituição já decidiu classificar daquele modo. A Tabela 1 consolida os parâmetros de recuperação adotados nos dois agentes aqui detalhados.

Tabela 1. Parâmetros de recuperação vetorial por agente

Parâmetro	Classificação de relatos	Assistente jurídico
Dimensão do vetor	768	768
Unidade indexada	Relato completo	Trecho de documento do processo
Vizinhos recuperados	10	Variável, definido por consulta
Corte de distância	0,5	Recuperação híbrida, densa e léxica
Deduplicação	Não se aplica	Por documento

Verificação posterior	Rótulo conferido contra a taxonomia	Citação conferida contra o conjunto recuperado
-----------------------	-------------------------------------	--

Fonte: elaborado pelos autores (2026).

A diferença entre os arranjos de recuperação não é arbitrária, e a razão é instrutiva. No agente de classificação, os dois lados da comparação são relatos curtos de cidadãos, redigidos no mesmo registro informal, e o corte fixo de 0,5 separa bem o análogo do meramente próximo. No assistente jurídico, a consulta é uma pergunta de poucas linhas e o alvo é um trecho de peça processual em registro formal, de modo que os dois textos ocupam regiões distintas do espaço de representação mesmo quando tratam do mesmo fato; daí a opção pela recuperação híbrida, na qual o ramo léxico compensa a assimetria de registro e garante o acerto exato dos identificadores. Transportar o corte de um agente para o outro, prática tentadora quando se reaproveita código, degradaria os dois.

Um segundo resultado, de natureza operacional, diz respeito ao dimensionamento das requisições de codificação, e ilustra como o material da Defensoria difere do material sobre o qual a literatura costuma medir. O serviço impõe teto de vinte mil tokens por requisição, limite que se mostrou mais restritivo que o teto de itens por lote: um lote de 180 relatos reais alcançou 21.558 tokens e foi recusado. A causa está na natureza do texto. O acervo de relatos tokeniza a aproximadamente dois caracteres por token, e não aos cerca de três vírgula cinco caracteres por token observados em português formal, porque o texto informal, com abreviação e erro de digitação, fragmenta-se mais no vocabulário do modelo. O cliente interno passou então a orçar o lote por tokens, e não por número de

itens, conforme a Equação 7, e a subdividi-lo quando ainda assim recusado.

$$B = \max\{n : \sum_{i=1}^n t_i \leq T\}, \quad T = 20.000 \quad (7)$$

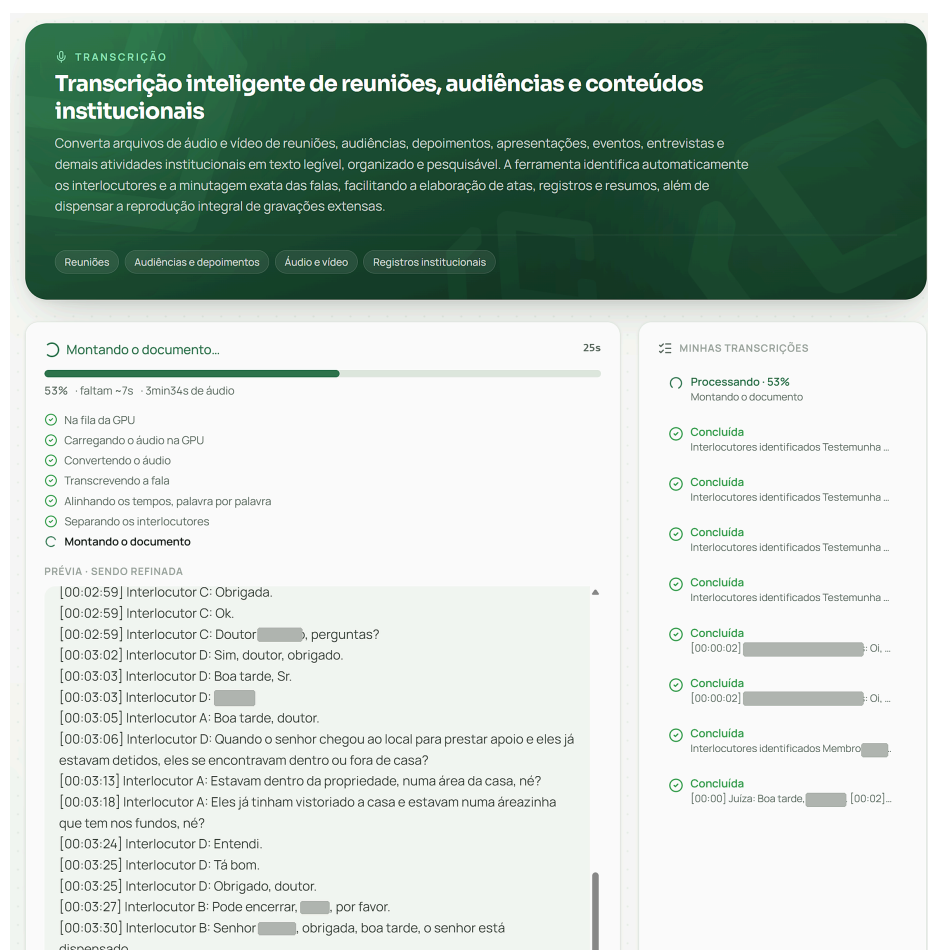
Cabe registrar, por fim, uma evolução arquitetural relevante para a reprodutibilidade deste relato. A recuperação de exemplos foi originalmente implantada sobre o mecanismo de recuperação gerenciado da plataforma de nuvem, e foi substituída por índice vetorial mantido no próprio banco relacional da aplicação, por meio de extensão de similaridade. A motivação não foi de desempenho, mas de governança: o mecanismo gerenciado codifica a consulta internamente, aceita texto e não vetor previamente calculado, o que impedia encaminhar essa chamada pelo intermediador de modelos que a instituição utiliza para atribuir custo por agente. Mantiveram-se o mesmo acervo, o mesmo modelo de codificação, o mesmo número de vizinhos e o mesmo corte de distância, de modo que a substituição preservou o comportamento e devolveu a visibilidade de custo.

5.2. Transcrição de Audiências com Whisper e WhisperX em Unidade Gráfica NVIDIA

O agente de transcrição foi reformulado para separar o trabalho de sinal do trabalho de linguagem, e essa separação é a chave do ganho obtido. A implementação anterior fragmentava o áudio em blocos de cinco minutos e submetia cada bloco ao modelo de linguagem multimodal, somando $N+1$ chamadas por audiência, com marcação temporal estimada e atribuição de interlocutor decidida dentro de cada fragmento. A implementação atual executa uma única chamada a um serviço próprio de reconhecimento de fala,

que roda Whisper e WhisperX em uma unidade de processamento gráfico NVIDIA L4 dedicada, e uma única chamada ao modelo Gemini, encarregada de nomear os interlocutores a partir do conteúdo, aplicar as convenções de redação e resumir. A Figura 12 exibe as etapas do processamento tal como apresentadas ao usuário, e a Figura 13, o documento entregue ao final.

Figura 12. Etapas do processamento de transcrição em unidade gráfica e prévia com marcação temporal e interlocutores ainda anônimos



Fonte: captura da plataforma em produção, com supressão dos nomes próprios ditos em audiência e dos trechos de abertura exibidos na lista de transcrições anteriores, elaborada pelos autores (2026).

A figura torna visível a divisão de trabalho descrita na subseção 2.5. As etapas apresentadas ao usuário, da entrada na fila da unidade

gráfica à montagem do documento, passando pela conversão do áudio, pela transcrição da fala, pelo alinhamento dos tempos palavra por palavra e pela separação dos interlocutores, correspondem exatamente aos estágios do WhisperX. A prévia, exibida abaixo, mostra o resultado com marcação temporal ao nível do segundo e interlocutores distintos já separados, ainda designados por letras: a separação é obra do agrupamento de representações de voz, e não do modelo de linguagem, que só entra em cena depois para nomeá-los.

Figura 13. Documento final da transcrição, com resumo do depoimento, interlocutores nomeados pelo papel processual e falas marcadas no tempo

The screenshot displays the WhisperX+Gemini transcription interface. At the top, it shows '56 segmentos paralelos' and '4 interlocutores'. The main content is divided into two sections: 'RESUMO DO DEPOIMENTO' and 'INTERLOCUTORES'.

RESUMO DO DEPOIMENTO: Depoimento de [redacted] guarda municipal arrolado como testemunha. O depoente relatou que, estando em serviço, foi acionado para prestar apoio a um colega de farda que reside ao lado do imóvel invadido. Ao chegar ao local, deparou-se com a acusada [redacted] e seu companheiro já detidos pelo outro guarda na área externa dos fundos da propriedade. Com eles, foram encontrados pertences da residência (utensílios de cozinha, louças e roupas) acondicionados em um carrinho. Segundo o depoente, os acusados confessaram informalmente a autoria do fato, justificando que eram usuários de entorpecentes e que subtraíram os bens para sustentar o vício. A testemunha declarou ainda ter constatado o arrombamento do imóvel, com o rompimento do cadeado do portão e a fechadura da porta forçada. Por fim, afirmou que não conhecia os acusados antes da ocorrência.

INTERLOCUTORES:

- Testemunha [redacted]
- Juíza
- Promotora de Justiça [redacted]
- Advogado [redacted]

TRANSCRIÇÃO:

- [00:00:02] Testemunha [redacted]: Oi, boa tarde.
- [00:00:04] Juíza: Seu nome completo, por favor.
- [00:00:06] Testemunha [redacted]: [redacted]
- [00:00:09] Juíza: O senhor está sozinho, não está na companhia das outras testemunhas do processo?
- [00:00:14] Testemunha [redacted]: Não, não.
- [00:00:14] Testemunha [redacted]: Estou sozinho.
- [00:00:16] Juíza: Eu vou lhe fazer algumas perguntas.
- [00:00:17] Juíza: O senhor sabe que o seu compromisso legal de dizer é verdade, certo?
- [00:00:21] Juíza: Sim.
- [00:00:22] Juíza: Doutora [redacted], perguntas?
- [00:00:25] Promotora de Justiça [redacted]: Seu [redacted] boa tarde.

Fonte: captura da plataforma em produção, com supressão dos nomes próprios de testemunha, promotora, advogado e ré, preservados os papéis processuais, elaborada pelos autores (2026).

A Figura 13 mostra o resultado dessa segunda etapa e explica por que ela permanece necessária. O cabeçalho registra o caminho efetivamente empregado, a composição do reconhecimento de fala com o modelo de linguagem, o tempo total de processamento e a contagem de interlocutores identificados. O que o modelo acrescenta ao trabalho de sinal é a nomeação: os grupos anônimos da Figura 12 aparecem aqui como testemunha, juíza, promotora de justiça e advogado, inferidos do conteúdo das próprias falas, e a essa nomeação soma-se um resumo do depoimento em prosa contínua. É a divisão de tarefas que a subseção 2.5 antecipa: medir o tempo e separar as vozes é problema de sinal, e saber que a voz que conduz o interrogatório é a do juízo é problema de linguagem. Nas figuras deste artigo os nomes próprios foram suprimidos e os papéis processuais preservados, justamente porque são eles, e não os nomes, que evidenciam o resultado da diarização.

O ganho qualitativo é de natureza distinta do ganho de tempo. A marcação temporal deixa de ser estimada e passa a ser medida por alinhamento, o que a torna utilizável para localizar um trecho na gravação original, requisito para que a transcrição sirva a finalidade processual; e a identidade do interlocutor, decidida por agrupamento de representações de voz sobre o áudio inteiro, permanece consistente do início ao fim da audiência. A exibição das etapas ao usuário, visível na Figura 12, cumpre função adicional que a experiência mostrou não ser acessória: um trabalho que leva minutos e não informa em que estágio se encontra é indistinguível, para quem espera, de um trabalho travado, e o usuário reenvia o material, dobrando o custo.

A operação em unidade gráfica impôs restrições que o desenho precisou absorver. Uma unidade atende uma requisição por vez, de

modo que o serviço adota semáforo unitário, fila de três posições e recusa explícita na entrada quando a fila está cheia. Sem esse teto, a fila de trabalhos pesados, que despacha dez tarefas em paralelo, esgotaria o prazo de entrega da décima audiência e deixaria o trabalho preso em estado de execução. A diarização, por sua vez, depende de modelo de acesso controlado, cuja credencial precisa ser obtida junto ao repositório que o distribui; sem ela o serviço sobe e transcreve com marcação temporal real, apenas não separa interlocutores, e essa degradação aparece explicitamente no nome do modelo registrado na saída, para que ninguém a descubra por acidente. Existe ainda rota alternativa para o modelo de linguagem, acionada quando a unidade gráfica está indisponível, cuja relevância é discutida na subseção 6.2.

5.3. Extração Estruturada de Documentos

O agente de extração estruturada atua na etapa de acolhimento, convertendo documentos digitalizados ou fotografados em campos cadastrais. O fluxo combina um serviço de reconhecimento óptico, que produz o texto, e uma chamada de modelo com esquema de saída, que localiza os campos solicitados nesse texto. Os campos desejados são declarados pelo operador antes do envio, conforme a Figura 14, o que torna o agente reaproveitável para tipos documentais distintos sem alteração de código. Além da extração, o agente avalia a validade do documento e sinaliza a inconsistência ao operador, conforme se observa na Figura 15, na qual a carteira apresentada está vencida e o aviso aparece antes dos campos.

Figura 14. Envio do documento para extração, com detecção automática do tipo e declaração dos campos a extrair

OCR ESTRUTURADO

Extração e validação inteligente de documentos

Extraia automaticamente informações de documentos digitalizados e fotografados, como RG e comprovante de residência. A solução avalia a legibilidade e a autenticidade da imagem, preenche os dados cadastrais e permite que a equipe concentre seus esforços no acolhimento e na análise jurídica.

Extração de dados Validação documental Atendimento virtual

Tipo do documento

Detectar automaticamente

Somente CNH, RG, comprovante de endereço e comprovante de IR são aceitos.

Campos para extrair (opcional com tipo selecionado) + Adicionar

Nome Nome

CPF CPF

CNH.pdf

Extrair Dados

Os dados extraídos aparecerão aqui.
Defina os campos, suba o documento e clique em extrair.

Fonte: captura da plataforma em produção, elaborada pelos autores (2026).

Figura 15. Extração estruturada de documento de identificação, com reconhecimento do tipo, alerta de vigência vencida, campos extraídos e grau de confiança

Dados Extraídos

gemini-3.5-flash 4.8s Confiança: 100%

Nova extração

Tipo reconhecido: CNH

Documento vencido
Documento vencido em 06/08/2025. Validade máxima para cnh: 120 meses.

CAMPOS EXTRAÍDOS

NOME

CPF

ANÁLISE DO DOCUMENTO

O documento foi identificado como CNH devido à estrutura e à presença de campos específicos como Carteira Nacional de Habilitação, CPF e data de validade. Os dados de [REDACTED] foram extraídos com sucesso.

Confiança da extração: 100%

A extração foi precisa?

Fonte: captura da plataforma em produção, com supressão do nome e do número de inscrição do titular, elaborada pelos autores (2026).

Dois detalhes técnicos merecem nota. O primeiro é a desativação explícita do raciocínio intermediário do modelo nessa chamada, medida que reduz o tempo de resposta em cerca de dois terços sem

perda de acurácia, dado que a tarefa é de localização de campo em texto, e não de inferência. A verificação de que o parâmetro é efetivamente honrado pelo intermediador de modelos foi conduzida por medição direta, que registrou 436 tokens de raciocínio na ausência do parâmetro e nenhum na sua presença; a cautela se justifica porque um intermediador que descartasse o campo devolveria resposta correta, apenas três vezes mais lenta e mais cara, falha que nenhum teste funcional apanharia. O segundo é o encaminhamento automático ao processamento em lote quando o documento excede os limites do reconhecimento óptico interativo, que recusa tanto por número de páginas quanto por tamanho em bytes, sendo que apenas a primeira recusa vem acompanhada de código de motivo; casar apenas o código deixava passar o documento curto e pesado, que falhava sem alternativa.

5.4. Dosimetria, Tradução para Linguagem Acessível e Triagem de Revisão Criminal

Três agentes completam a frente finalística e são aqui descritos de modo sumário, por serem aplicações mais convencionais de modelo de linguagem com esquema de saída. O agente de dosimetria da pena percorre as três fases do cálculo, examina as circunstâncias judiciais, as agravantes e atenuantes e as causas de aumento e diminuição, e devolve parecer fundamentado com a pena sugerida, que o defensor confere e ajusta; mais de quinhentas análises foram produzidas desde a implantação. O agente Descomplica.Def traduz decisões e despachos para linguagem acessível ao cidadão, atendendo à dimensão informacional do acesso à justiça discutida na subseção 2.1, e já produziu mais de cinco mil textos simplificados.

Figura 16. Tradução de peça processual para linguagem acessível pelo agente Descomplica.Def

DESCOMPLICA.DEF

Linguagem jurídica simples e acessível

Traduza movimentações, decisões e despachos judiciais para uma linguagem clara e compreensível. A ferramenta oferece explicações de termos jurídicos e resumos processuais, ampliando a transparência, a autonomia do assistido e o acesso à informação.

Linguagem simplesTermos jurídicosResumos processuais

Requerente

Nome do requerente

Movimento

Tipo do movimento

Movimento amigável

Descrição amigável

Descomplica2.pdf

Traduzir documento

Documento simplificado

Este documento é um pedido oficial enviado pela Defensoria Pública do Estado do Paraná ao juiz da Vara de Família de Matinhos. O objetivo é formalizar a defesa de [REDACTED] no processo.

Os principais pontos e argumentos apresentados no documento são:

1. Entrada Oficial da Defensoria no Processo: A Defensoria Pública solicita ao juiz para ser registrada formalmente como a representante legal de [REDACTED]. O documento explica que ela passou por uma triagem socioeconômica detalhada do órgão, comprovando que não tem condições financeiras de contratar um advogado particular.
2. Pedido de Processo Sem Custas (Justiça Gratuita): O documento pede que [REDACTED] seja liberada de pagar qualquer taxa ou despesa cobrada pela Justiça. O argumento usado é que, por ter sido aprovada no atendimento da Defensoria Pública (que possui regras rígidas para aceitar apenas pessoas necessitadas), fica demonstrado de forma clara que ela tem baixa renda e necessita desse benefício para poder se defender.
3. Direitos Especiais de Trabalho da Defensoria: É solicitado que o juiz respeite as regras especiais de trabalho garantidas por lei aos defensores públicos. As duas principais são: o direito de receber as comunicações do processo de forma pessoal e direta, e o direito de ter o dobro do tempo normal (prazo em dobro) para responder a qualquer decisão ou apresentar novos documentos.

Ao final, o defensor público [REDACTED] reforça os pedidos para que o juiz conceda a gratuidade de taxas para [REDACTED] e aceite a Defensoria Pública como sua representante no caso, garantindo todos os seus direitos de atuação.

Fonte: captura da plataforma em produção, com supressão do nome da pessoa assistida e do nome do defensor público subscritor, elaborada pelos autores (2026).

A Figura 16 permite observar o que distingue essa tradução de um resumo. O texto de saída não encurta a peça, ele troca a categoria jurídica pela sua consequência prática: a gratuidade da justiça vir a ser liberada de pagar qualquer taxa ou despesa cobrada pela Justiça, e a intimação pessoal com prazo em dobro vir a receber as comunicações do processo de forma pessoal e direta e ter o dobro do tempo normal para responder. É a operação que a subseção 2.1 identifica como barreira informacional, e ela não é trivial para um

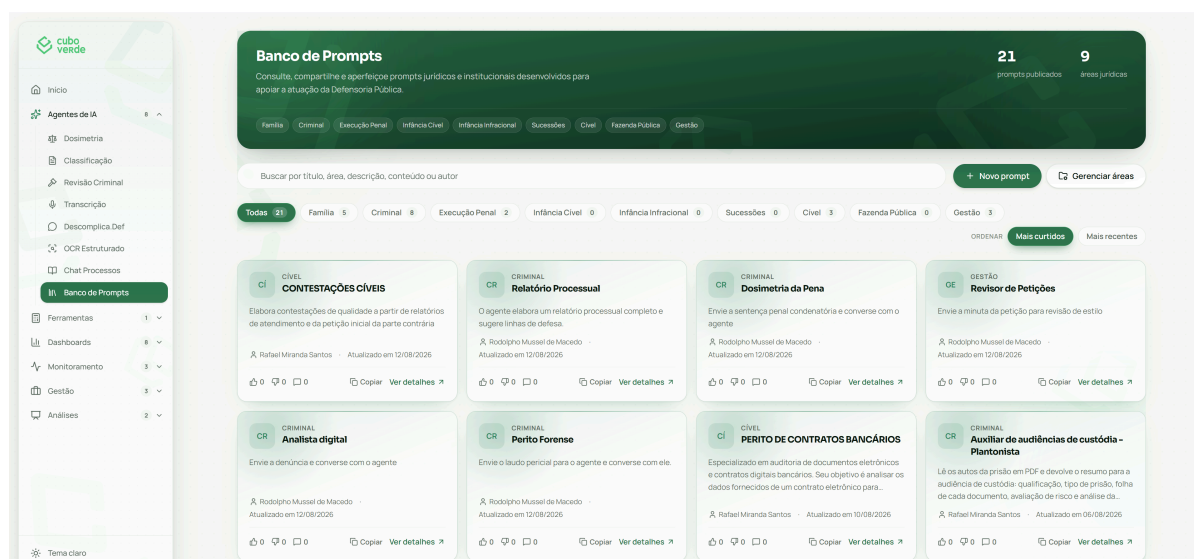
modelo de linguagem, porque exige preservar o conteúdo normativo enquanto se abandona o vocabulário que o carrega.

O terceiro é o agente de triagem de revisão criminal, que avalia o cabimento da ação nos termos do art. 621 do Código de Processo Penal a partir do processo, das peças em formato digital e do relato, apoiado no acervo de peças que a própria instituição já protocolou. Ele foi implantado em setembro de 2026, portanto no mesmo mês em que este artigo foi escrito, e por essa razão não é objeto de análise aqui: não há ainda período de operação que autorize afirmação sobre o seu comportamento, e registrá-lo como resultado seria antecipar avaliação que os dados não sustentam. Sua menção se justifica apenas para que o catálogo do Quadro 1 seja completo.

5.5. Banco de Prompts

O Banco de Prompts responde a um problema que costuma passar despercebido nas implantações de inteligência artificial: a instrução dada ao modelo é, ela própria, um ativo institucional, e sem curadoria ela se dispersa em arquivos pessoais, mensagens e cadernos. A plataforma mantém uma biblioteca de instruções jurídicas publicadas, organizadas por área do direito, com autoria identificada, avaliação pelos pares, comentários e arquivos de contexto que o prompt espera receber junto, conforme a Figura 17.

Figura 17. Banco de Prompts, com a grade curada por área do direito, o crédito de autoria e a avaliação pelos pares



Fonte: captura da plataforma em produção; os créditos de autoria foram preservados, por serem a atribuição que a biblioteca de origem pede que se mantenha. Elaborada pelos autores (2026).

A grade abre ordenada pelos mais curtidos, e a ordenação é feita pelo banco de dados, e não pela tela. A distinção parece técnica e é substantiva: como a lista é paginada, reordenar apenas os cartões visíveis chamaria de mais curtidos uma página que pode não conter o melhor avaliado do acervo. A régua é o saldo entre avaliações positivas e negativas, escolha que faz o prompt rejeitado descer abaixo daquele que ninguém avaliou, e não apenas deixar de subir. A busca combina três ramos, a correspondência literal, a correspondência sem acentuação e a busca por radical em português, porque nenhum dos três sozinho encontra petição inicial quando o usuário digita peticao iniciais.

Duas permissões governam o espaço, e somente duas: ler, copiar, baixar arquivo, avaliar e comentar é faculdade de qualquer profissional autenticado; cadastrar, editar, excluir e moderar exige a permissão de curadoria. O acervo inicial foi constituído a partir de biblioteca pública mantida por defensor público que autoriza o reaproveitamento com preservação da autoria, e o crédito ao autor externo é campo próprio do registro, de modo que o cartão exhibe o

nome de quem escreveu a instrução mesmo quando essa pessoa não possui conta na plataforma.

5.6. Governança de Acesso e Proteção de Dados

A ampliação do acesso a ferramentas que processam dados sensíveis exigiu revisão do controle de permissões. O modelo adotado combina quatro origens que somam e nenhuma que subtrai: o perfil de administrador, que entrega o catálogo ativo integralmente; um piso de recursos considerados de uso geral, disponível a todo usuário interno autenticado; os recursos atribuídos por grupo; e as concessões individuais, destinadas ao caso que a estrutura de grupos não cobre, como conceder a um defensor o agente de dosimetria sem lhe entregar junto todo o restante do grupo. A ausência de concessão negativa é deliberada, pois dispensa regra de precedência e torna previsível o resultado da resolução de permissões, propriedade que importa quando quem administra o acesso não é quem escreveu o código.

A providência de segurança de maior alcance foi o deslocamento da verificação de permissão para o servidor. Enquanto todo usuário autenticado pertencia ao domínio institucional, a ocultação de botões na interface operava como controle suficiente na prática, ainda que não em direito; a admissão de contas externas eliminou essa premissa e tornou obrigatória a verificação em cada rota. Um teste automatizado percorre a tabela de rotas da aplicação e reprova a construção quando encontra rota nova sem verificação declarada, com lista de exceções explícita e justificada uma a uma, de modo que rota nova nasce fechada. O mecanismo é modesto e resolve a classe inteira do problema, em vez de cada ocorrência dele.

Quanto ao ciclo de vida do dado, os arquivos enviados pelo usuário aos agentes são removidos do repositório ao término de cada análise, em atenção ao princípio da necessidade. A exclusão de conta de usuário, por sua vez, é admitida apenas quando a conta não deixou rastro: conta com conversação registrada, com registro em trilha de auditoria ou com credencial de integração emitida recebe recusa explicativa, porque apagar a linha comprometeria trilha exigida pela legislação de proteção de dados ou interromperia integração de terceiro. Para esses casos, o caminho é a desativação, e a diferença entre uma coisa e outra é registrada na própria mensagem de recusa, para que o administrador não conclua que o sistema falhou.

6. RESULTADOS E DISCUSSÃO

6.1. Resultados Institucionais

Os efeitos agregados da iniciativa, consolidados no levantamento institucional de 2026, são apresentados na Figura 18 e detalhados na Tabela 2. Os números correspondem a frentes distintas do ecossistema e não são somáveis entre si.

Figura 18. Infográfico institucional de resultados do Cubo Verde



Fonte: Defensoria Pública do Estado do Paraná (2026).

Tabela 2. Resultados institucionais consolidados

Indicador	Resultado
Redução no tempo de análise processual com o assistente jurídico	Aproximadamente 70%
Defensores que utilizam o assistente jurídico	Quase 200, com as respectivas equipes de assessoria
Profissionais alcançados pelo conjunto dos agentes	Cerca de 800, entre defensores e assessores
Economia anual estimada com automação de rotinas	17.320 horas de trabalho
Depoimentos e audiências transcritos automaticamente	Mais de 3.600
Análises de dosimetria produzidas com apoio dos agentes	Mais de 500
Classificações automáticas de relato de atendimento	Mais de 3.000

Textos jurídicos traduzidos para linguagem acessível	Mais de 5.000
Municípios paranaenses com presença da instituição	74, em 30 comarcas
Soluções disponíveis no portal institucional	24, das quais 8 são agentes

Fonte: elaborado pelos autores (2026) a partir de dados institucionais.

A redução aproximada de setenta por cento no tempo de análise processual é o indicador de maior significado, e convém precisar o que ele mede e o que não mede. Ele compara o tempo despendido para formar juízo sobre um processo antes e depois da disponibilização do assistente, em estimativa construída sobre relato dos próprios usuários, e não em cronometragem controlada. Não mede acurácia, não mede qualidade da peça produzida ao final e não autoriza inferência sobre resultado processual. Ainda assim, a ordem de grandeza é consistente com o que a Figura 5 permite intuir: localizar as peças pertinentes em acervos que, no índice ali reproduzido, vão de poucas dezenas a mais de oitocentos documentos é, em si, a maior parte do trabalho de análise, e é exatamente essa parte que a recuperação automatiza.

A economia de 17.320 horas anuais decorre principalmente da automação de rotinas administrativas, como a designação de defensores, a movimentação entre unidades e a estruturação de dados provenientes de sistemas e de comunicações eletrônicas. Trata-se de estimativa construída sobre o tempo médio previamente despendido em cada rotina, e não de medição direta de relógio, razão pela qual é aqui apresentada como ordem de grandeza. O

dado relevante para a discussão não é a precisão do número, mas o fato de que o tempo liberado se converte em atividade finalística, isto é, em atendimento, em uma instituição que já realiza mais de novecentos mil atendimentos por ano.

O alcance do assistente jurídico, que apoia quase 200 defensores e as respectivas equipes de assessoria, com o conjunto dos agentes chegando a cerca de oitocentos profissionais, sugere que a adesão a ferramentas dessa natureza não é automática, mesmo quando elas estão disponíveis a todos e a permissão é o piso e não a exceção. A observação converge com a literatura sobre adoção de tecnologia no setor público e indica que a próxima fronteira da iniciativa é menos técnica do que organizacional: formação, redesenho de fluxo de trabalho e demonstração de confiabilidade pesam mais, nesta etapa, do que qualquer ganho adicional de acurácia do modelo.

6.2. Quatro Lições de Engenharia

A operação continuada do ecossistema produziu um conjunto de aprendizados que excede o escopo de cada agente e que, por sua generalidade, pode interessar a outras instituições. Quatro deles são aqui sistematizados, todos apoiados em medição.

A primeira lição é que falha de chamada nunca deve sair como conclusão vazia. Em episódio registrado em agosto de 2026, o serviço de modelo recusou uma chamada por excesso de requisições, e a aplicação entregou ao defensor uma análise sem tese, comunicando, na prática, que não havia tese cabível para aquele caso. A distinção entre não encontrar e não conseguir perguntar é essencial quando a saída do sistema orienta decisão profissional, e ela desaparece se o tratamento de erro se limitar a

devolver o que se conseguiu montar. No mesmo dia, mediram-se duas falhas correlatas: uma mensagem de erro que era construída e descartada antes de chegar à tela, e uma interface que permitia concluir o envio de 62 megabytes em canal de aproximadamente noventa quilobytes por segundo para só então recusar o arquivo por exceder o limite que o servidor impunha. A regra derivada é dupla: erro que o usuário lê deve dizer o motivo e a providência, e limite imposto pelo servidor deve ser conferido pela interface antes de consumir o envio.

A segunda lição é que o sistema precisa denunciar a própria degradação. A rota alternativa da transcrição, que substitui a unidade gráfica pelo modelo de linguagem quando aquela está indisponível, foi acionada 335 vezes contra duas execuções na unidade gráfica ao longo de três dias úteis consecutivos, em razão de indisponibilidade de recurso na região de nuvem. Nenhum alarme disparou, e a razão é instrutiva: o trabalho não falhava, pois a rota alternativa existia justamente para isso; o agendador reportava sucesso, porque a chamada de inicialização da máquina retornava com êxito e apenas a operação subsequente falhava; e não havia métrica sobre a frequência de acionamento da rota alternativa. Passaram a existir dois alarmes, um sobre a causa, a ausência de sinal de atividade da máquina, e outro sobre o sintoma, a taxa de acionamento da rota alternativa, pois o segundo captura o que o primeiro não vê. Um detalhe do mesmo episódio desmentiu outra premissa: trinta e um acionamentos ocorreram em um domingo, o que mostrou que o uso da plataforma não obedecia ao calendário forense que a configuração pressupunha.

A terceira lição diz respeito ao cálculo de custo de disponibilidade. A unidade gráfica era desligada fora do horário forense, decisão

apoiada em conta aparentemente evidente, dado que a operação contínua custa cerca de duas vírgula seis vezes mais do que a operação em janela. A conta, contudo, respondia à pergunta errada: enquanto a máquina permanece ligada, o recurso está reservado, e cada desligamento o devolve ao conjunto compartilhado do provedor, sujeitando o religamento à disponibilidade regional. Foi exatamente esse religamento que falhou em três dias consecutivos. A economia nominal de recurso ocioso foi paga com indisponibilidade do recurso no horário de uso, e a política de desligamento programado foi abandonada.

A quarta lição é que a observabilidade precede a otimização. Antes da instrumentação descrita na subseção 3.5, a plataforma operava sem registro utilizável: 979 de mil entradas amostradas chegavam sem indicação de severidade, e exceções encadeadas apareciam como entradas independentes, cabendo a quem investigasse remontar a cadeia pelo microssegundo do registro de tempo. Nenhuma das três lições anteriores poderia ter sido formulada nesse estado, porque nenhuma delas é visível sem medição. A ordem de implantação, portanto, não é indiferente: instrumentar primeiro, decidir depois.

6.3. Aderência Ao Plano Brasileiro de Inteligência Artificial

Confrontada com o Plano Brasileiro de Inteligência Artificial (MCTI; CGEE, 2025), a experiência relatada oferece evidência sobre três de suas metas. Quanto à adoção da tecnologia em áreas críticas do serviço público, o assistente jurídico opera sobre a atividade finalística da instituição, e não sobre rotina administrativa periférica, o que é justamente o ponto em que a adoção costuma encontrar mais resistência. Quanto à capacitação, o dado de adesão registrado

na subseção 6.1 sugere que a disponibilização não basta e que a meta de treinamento fixada no plano é condição, e não complemento, do investimento em tecnologia. Quanto à conformidade com a legislação de proteção de dados, as decisões descritas ao longo deste artigo, da supressão do nome no texto submetido ao modelo à recusa de exclusão de conta que deixou trilha, mostram que a exigência se traduz em escolhas concretas de arquitetura, e não em declaração de política.

Há também um ponto de convergência quanto à replicabilidade, que o plano trata como preocupação de escala. O sistema processual sobre o qual o assistente foi acoplado está presente em vinte e uma unidades da federação, e seis delas já se encontram conectadas ao ambiente de homologação da plataforma. Isso significa que o componente mais custoso de reproduzir o arranjo, que é a integração com a fonte documental, já está resolvido para a maior parte do país, e que o esforço remanescente é de configuração e de governança local. É essa circunstância, mais do que qualquer propriedade do desenho, que torna a experiência aqui relatada potencialmente útil fora do Paraná.

Considerados em conjunto, os resultados e as lições sustentam a tese central deste artigo. O ganho observado não decorre da capacidade do modelo de linguagem isoladamente, e sim da qualidade da recuperação que o alimenta, da governança do dado que a torna possível e das decisões de engenharia que mantêm o conjunto operando de modo previsível. O modelo é o componente mais visível e o mais facilmente substituível, e de fato foi substituído mais de uma vez ao longo do período observado, sem que a arquitetura mudasse; a recuperação, a verificação e a governança são o que a instituição efetivamente construiu.

7. CONSIDERAÇÕES FINAIS

O trabalho descreveu o assistente jurídico de consulta aos autos do Cubo Verde e a arquitetura que o sustenta, detalhando os mecanismos pelos quais a Defensoria Pública do Estado do Paraná incorpora modelos de linguagem aos seus fluxos finalísticos. A questão formulada na introdução indagava como uma instituição pública pode adotar esses modelos sem lhes transferir a responsabilidade pela veracidade da informação. A resposta que a experiência oferece é de ordem arquitetural, e não de escolha de modelo: a responsabilidade permanece na instituição quando o modelo opera sobre contexto recuperado de acervo próprio, quando cada afirmação carrega o identificador da fonte que a sustenta, quando as referências citadas são conferidas contra o conjunto efetivamente recuperado, na forma da Equação 6, e quando o campo estruturado exibido ao profissional é lido do registro institucional, e não do texto gerado.

Os objetivos específicos foram atingidos. A arquitetura em nuvem pública e as duas formas de entrega, portal de acesso interno e interfaces de programação acopladas aos sistemas SOLAR e LUNA, foram apresentadas na seção 3, com as medições que justificaram cada decisão de separação de cargas. A cadeia de codificação, recuperação e geração foi formalizada nas Equações 1 a 5 e detalhada na seção 4, etapa por etapa, da importação das peças à contabilidade do contexto exibida ao defensor. As verificações que separam o desenho da aplicação ingênua de um modelo foram tratadas na subseção 4.5. Os demais agentes em produção foram descritos na seção 5, com os parâmetros consolidados na Tabela 1, e as decisões de engenharia e de governança determinantes para a confiabilidade foram sistematizadas na subseção 6.2.

A pesquisa apresenta três limitações que condicionam a generalização de seus achados. A primeira é a ausência de avaliação quantitativa de acurácia por agente: os resultados relatados são de adoção, de tempo economizado e de comportamento do sistema, e não de concordância entre a saída do agente e um padrão de referência estabelecido por especialistas; o indicador de redução de tempo, em particular, apoia-se em estimativa dos usuários. A segunda é o caráter de caso único, que impede atribuir os resultados a propriedades do desenho em vez de a características particulares da instituição observada. A terceira é a rápida evolução dos modelos empregados, que torna transitórias as medições de latência e de custo, embora preserve a validade das decisões de arquitetura, que não dependem de modelo específico.

Como continuidade, indicam-se três frentes. A primeira é a construção de conjunto de avaliação anotado por defensores, que permita medir acurácia de classificação e fidelidade da resposta à fonte citada, superando a principal limitação deste relato. A segunda é a investigação dos determinantes organizacionais da adoção, sugerida pela proporção de defensores que utilizam o assistente jurídico e alinhada à meta de capacitação do Plano Brasileiro de Inteligência Artificial. A terceira é a consolidação da integração nacional, hoje em homologação com seis unidades da federação, cuja expansão permitirá examinar se o modelo se sustenta em contextos institucionais distintos daquele em que foi concebido, e transformar um caso único em evidência comparada.

REFERÊNCIAS BIBLIOGRÁFICAS

BAIN, Max; HUH, Jaesung; HAN, Tengda; ZISSERMAN, Andrew. WhisperX: time-accurate speech transcription of long-form audio. In:

INTERSPEECH 2023, 24., 2023, Dublin. Proceedings [...]. Dublin: ISCA, 2023. p. 4489-4493.

BRASIL. Constituição da República Federativa do Brasil de 1988. Brasília, DF: Presidência da República, 1988. Disponível em: https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acesso em: 11 set. 2026.

BRASIL. Lei Complementar nº 80, de 12 de janeiro de 1994. Organiza a Defensoria Pública da União, do Distrito Federal e dos Territórios e prescreve normas gerais para sua organização nos Estados. Brasília, DF: Presidência da República, 1994. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/lcp/lcp80.htm. Acesso em: 11 set. 2026.

BRASIL. Lei nº 12.527, de 18 de novembro de 2011. Regula o acesso a informações previsto no inciso XXXIII do art. 5º da Constituição Federal. Brasília, DF: Presidência da República, 2011. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm. Acesso em: 11 set. 2026.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília, DF: Presidência da República, 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 11 set. 2026.

BREDIN, Hervé. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In: INTERSPEECH 2023, 24., 2023, Dublin. Proceedings [...]. Dublin: ISCA, 2023. p. 1983-1987.

BROWN, Tom B.; MANN, Benjamin; RYDER, Nick; SUBBIAH, Melanie; KAPLAN, Jared; DHARIWAL, Prafulla; NEELAKANTAN, Arvind; SHYAM,

Pranav; SASTRY, Girish; ASKELL, Amanda. Language models are few-shot learners. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 33., 2020. Proceedings [...]. Red Hook: Curran Associates, 2020. p. 1877-1901.

CAPPELLETTI, Mauro; GARTH, Bryant. Acesso à justiça. Tradução de Ellen Gracie Northfleet. Porto Alegre: Sergio Antonio Fabris, 1988.

CONSELHO NACIONAL DE JUSTIÇA. Resolução nº 332, de 21 de agosto de 2020. Dispõe sobre a ética, a transparência e a governança na produção e no uso de Inteligência Artificial no Poder Judiciário. Brasília, DF: CNJ, 2020. Disponível em: <https://atos.cnj.jus.br/atos/detalhar/3429>. Acesso em: 11 set. 2026.

DEVLIN, Jacob; CHANG, Ming-Wei; LEE, Kenton; TOUTANOVA, Kristina. BERT: pre-training of deep bidirectional transformers for language understanding. In: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 2019, Minneapolis. Proceedings [...]. Stroudsburg: ACL, 2019. p. 4171-4186.

GAO, Yunfan; XIONG, Yun; GAO, Xinyu; JIA, Kangxiang; PAN, Jinliu; BI, Yuxi; DAI, Yi; SUN, Jiawei; WANG, Meng; WANG, Haofen. Retrieval-augmented generation for large language models: a survey. arXiv:2312.10997, 2023. Disponível em: <https://arxiv.org/abs/2312.10997>. Acesso em: 11 set. 2026.

JI, Ziwei; LEE, Nayeon; FRIESKE, Rita; YU, Tiezheng; SU, Dan; XU, Yan; ISHII, Etsuko; BANG, Yejin; MADOTTO, Andrea; FUNG, Pascale. Survey of hallucination in natural language generation. ACM Computing Surveys, New York, v. 55, n. 12, p. 1-38, 2023.

JOHNSON, Jeff; DOUZE, Matthijs; JÉGOU, Hervé. Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, New York, v. 7, n. 3, p. 535-547, 2021.

KARPUKHIN, Vladimir; OGUZ, Barlas; MIN, Sewon; LEWIS, Patrick; WU, Ledell; EDUNOV, Sergey; CHEN, Danqi; YIH, Wen-tau. Dense passage retrieval for open-domain question answering. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2020. Proceedings [...]. Stroudsburg: ACL, 2020. p. 6769-6781.

LEWIS, Patrick; PEREZ, Ethan; PIKTUS, Aleksandra; PETRONI, Fabio; KARPUKHIN, Vladimir; GOYAL, Naman; KÜTTLER, Heinrich; LEWIS, Mike; YIH, Wen-tau; ROCKTÄSCHEL, Tim; RIEDEL, Sebastian; KIELA, Douwe. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 33., 2020. Proceedings [...]. Red Hook: Curran Associates, 2020. p. 9459-9474.

MALKOV, Yury A.; YASHUNIN, Dmitry A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. IEEE Transactions on Pattern Analysis and Machine Intelligence, New York, v. 42, n. 4, p. 824-836, 2020.

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO - MCTI; CENTRO DE GESTÃO E ESTUDOS ESTRATÉGICOS - CGEE. Plano Brasileiro de Inteligência Artificial. Brasília, DF: MCTI;CGEE, 2025. 104 p.

PGVECTOR. pgvector: open-source vector similarity search for Postgres. [S. l.]: GitHub, 2026. Disponível em: <https://github.com/pgvector/pgvector>. Acesso em: 11 set. 2026.

RADFORD, Alec; KIM, Jong Wook; XU, Tao; BROCKMAN, Greg; MCLEAVEY, Christine; SUTSKEVER, Ilya. Robust speech recognition via large-scale weak supervision. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 40., 2023, Honolulu. Proceedings [...]. [S. l.]: PMLR, 2023. p. 28492-28518.

REIMERS, Nils; GUREVYCH, Iryna. Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2019, Hong Kong. Proceedings [...]. Stroudsburg: ACL, 2019. p. 3982-3992.

SANTOS, Boaventura de Sousa. Para uma revolução democrática da justiça. São Paulo: Cortez, 2007.

VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion; GOMEZ, Aidan N.; KAISER, Lukasz; POLOSUKHIN, Illia. Attention is all you need. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 30., 2017, Long Beach. Proceedings [...]. Red Hook: Curran Associates, 2017. p. 5998-6008.

¹ Doutor, Engenheiro. Diretor de Tecnologia e Inovação da Defensoria Pública do Estado do Paraná. E-mail: [acesse o artigo original para visualizar o e-mail](#).

² Defensor e Assessor de Tecnologia e Inovação da Defensoria Pública do Estado do Paraná. E-mail: [acesse o artigo original para visualizar o e-mail](#).

³ Defensor Público-Geral do Estado do Paraná. E-mail: [acesse o artigo original para visualizar o e-mail](#).

⁴ Diretoria de Tecnologia e Inovação da Defensoria Pública do Estado do Paraná. E-mail: [acesse o artigo original para visualizar o e-mail](#).

⁵ Diretoria de Tecnologia e Inovação da Defensoria Pública do Estado do Paraná. E-mail: [acesse o artigo original para visualizar o e-mail](#).

⁶ Diretoria de Tecnologia e Inovação da Defensoria Pública do Estado do Paraná. E-mail: [acesse o artigo original para visualizar o e-mail](#).

⁷ Diretoria de Tecnologia e Inovação da Defensoria Pública do Estado do Paraná. E-mail: [acesse o artigo original para visualizar o e-mail](#).

⁸ Diretoria de Tecnologia e Inovação da Defensoria Pública do Estado do Paraná. E-mail: [acesse o artigo original para visualizar o e-mail](#).

⁹ Diretoria de Tecnologia e Inovação da Defensoria Pública do Estado do Paraná. E-mail: [acesse o artigo original para visualizar o e-mail](#).