Utilização de Modelos de Linguagem de Grande Escala (LLMs) para Resumo Automático de Informações em Procedimentos Extrajudiciais

Helyézer Nascimento Freitas Teófilo[1], Fábio Castro Araújo[1]

[1]Departamento de Computação
Universidade Luterana do Brasil – Palmas – TO

helyezerteofilo2@rede.ulbra.com, fabio.araujo@ulbra.br

Resumo: *O crescente volume de informações presentes em procedimentos extrajudiciais representa um desafio para os profissionais dos Ministérios Públicos, que precisam analisar grandes quantidades de documentos em pouco tempo. O processo de leitura e síntese demanda esforço significativo e apresenta variações na qualidade entre analistas. O uso de Large Language Models (LLMs) surge como alternativa promissora para apoiar a triagem e a sumarização automática dessas informações. Este trabalho propõe um pipeline de sumarização baseado em técnicas de processamento de linguagem natural, integrando limpeza textual, sumarização individual de peças e síntese final estruturada. A abordagem visa reduzir o tempo de leitura, aumentar a eficiência e padronizar relatórios no contexto extrajudicial. A avaliação realizada com documentos reais aponta para o potencial dessa tecnologia em apoiar a análise inicial de procedimentos e sua futura integração a sistemas institucionais.*

Abstract: *The increasing volume of information contained in extrajudicial procedures poses a challenge for Public Prosecutor's Offices, which must analyze large amounts of documents within limited timeframes. The reading and synthesis process demands significant effort and exhibits variability in quality among analysts. Large Language Models (LLMs) emerge as a promising alternative to support the triage and automatic summarization of such information. This work proposes a summarization pipeline based on natural language processing techniques, integrating text cleansing, segmented summarization of individual pieces, and structured final synthesis. The approach aims to reduce reading time, increase efficiency, and standardize reports in the extrajudicial context. Evaluation using real documents indicates the potential of this technology to support the initial analysis of procedures and its future integration into institutional systems.*

1. Introdução

Procedimentos extrajudiciais desempenham papel central no âmbito do Ministério Público e de outros órgãos de justiça, constituindo instrumentos essenciais para a defesa de direitos coletivos, a mediação de conflitos e a promoção da cidadania. Tais mecanismos ampliam a capacidade resolutiva institucional sem a necessidade imediata de judicialização, alinhando-se ao movimento de desjudicialização

impulsionado pela Lei nº 11.441/2007 e pela concepção de Justiça Multiportas prevista no Código de Processo Civil de 2015 (HILL; DALLA, 2016).

Apesar dos avanços, a expansão desses instrumentos trouxe novos desafios à gestão da informação jurídica. O grande volume de documentos gerados em procedimentos extrajudiciais exige métodos mais eficientes de leitura e interpretação, sob pena de comprometer a celeridade e a padronização da atuação ministerial.

Dados oficiais divulgados pelo Ministério Público do Estado do Tocantins revelam que, somente em 2023, foram instaurados 17.767 novos procedimentos extrajudiciais (COSTA, 2024). Informações complementares da Assembleia Legislativa do Tocantins indicam que, entre 2020 e 2022, o número anual de autuações variou entre 10.587 e 14.429 procedimentos (ASSEMBLEIA LEGISLATIVA DO TOCANTINS, 2023). Esses dados dimensionam a carga informacional enfrentada pela instituição e reforçam a necessidade de ferramentas que apoiem a análise e a síntese desses documentos.

Nos últimos anos, avanços significativos em Large Language Models (LLMs), como GPT-4, LLaMA, Falcon e Mistral, têm demonstrado elevado potencial na compreensão e geração de linguagem natural em múltiplos domínios. Estudos recentes apontam que resumos produzidos por LLMs são frequentemente preferidos a resumos humanos e aos gerados por modelos ajustados para tarefas específicas, apresentando maior consistência factual e fluência textual (PU; GAO; WAN, 2023). O desempenho robusto desses modelos em cenários *zero-shot* evidencia sua aplicabilidade prática a processos complexos de análise de informação.

No campo jurídico, benchmarks como o *LegalBench* demonstram que esses modelos são capazes de sintetizar decisões e peças legais com qualidade competitiva, ainda que persistam limitações, como risco de alucinações e perda de nuances interpretativas inerentes ao discurso jurídico (GUHA et al., 2023). Nesse contexto, as *LLMs* configuram-se como ferramentas promissoras para apoiar a automação de tarefas de sumarização, com potencial para reduzir a carga de trabalho humano, aumentar a eficiência analítica e padronizar relatórios, desde que acompanhadas de validação e supervisão especializada (BOMMASANI et al., 2021).

Diante da crescente complexidade e do volume documental dos procedimentos extrajudiciais, torna-se relevante investigar soluções baseadas em *LLMs* capazes de otimizar a análise e a síntese de informações, fortalecendo a eficiência institucional e o suporte às atividades de triagem e compreensão inicial dos casos. O objetivo geral deste trabalho consiste em investigar, em caráter exploratório, a aplicação de *Large Language Models (LLMs)* na sumarização automática de informações jurídicas extrajudiciais, analisando sua viabilidade técnica e a qualidade dos resumos gerados. Especificamente, busca-se explorar técnicas de sumarização baseadas em *LLMs* e coletar feedback de promotores, servidores e técnicos, com o intuito de avaliar a utilidade prática da solução proposta em cenários reais de análise preliminar de procedimentos.

2. Fundamentação Teórica

2.1 Processos Extrajudiciais

Os processos extrajudiciais constituem instrumentos administrativos que integram a atuação finalística do Ministério Público Federal (MPF), permitindo a tutela de direitos coletivos, a mediação de conflitos e a promoção da cidadania. A sua ampliação e o fortalecimento inserem-se no contexto mais amplo da desjudicialização e da consolidação do modelo de Justiça Multiportas, que busca oferecer múltiplas vias de acesso à justiça (HILL; DALLA, 2016). A desjudicialização representa um redirecionamento da cultura jurídica brasileira, permitindo que determinadas matérias sejam resolvidas fora do Poder Judiciário, desde que observados os princípios constitucionais do devido processo legal, da publicidade e da imparcialidade (HILL; DALLA, 2016). Esse movimento traduz uma transformação do acesso à justiça, agora compreendido como um sistema plural de resolução de conflitos, no qual instâncias extrajudiciais complementam a atuação jurisdicional, ampliando a efetividade e a celeridade da tutela de direitos.

No plano prático, a adoção dos instrumentos extrajudiciais tem contribuído para aliviar a sobrecarga estrutural do sistema judicial brasileiro. Conforme relatado por MACEDO; SILVA (2021), a Lei nº 11.441/2007, ao permitir a realização de inventários, separações e divórcios pela via extrajudicial, foi um marco nesse processo, retirando mais de 1,3 milhão de ações das varas judiciais. Dados do Conselho Nacional de Justiça indicam que, em 2020, o Judiciário ainda apresentava uma taxa de congestionamento de 68,5% (CNJ, 2020). Esses números evidenciam a importância de soluções administrativas e extrajudiciais para a eficiência da justiça.

O fluxo operacional dos processos extrajudiciais segue etapas padronizadas que asseguram transparência e rastreabilidade. O processo inicia-se com o recebimento da notícia de fato, registrada e avaliada quanto à relevância e competência ministerial (MPF, 2018). Havendo indícios mínimos de irregularidade, instaura-se um procedimento preparatório ou um inquérito civil, conduzido pelo membro do Ministério Público com apoio técnico de servidores e analistas (MPF, 2018). Durante a instrução, podem ser expedidos ofícios, recomendações, termos de ajustamento de conduta (TACs) e outras medidas administrativas destinadas à solução extrajudicial do conflito (MPF, 2018). Concluídas as diligências, o procedimento pode resultar em arquivamento, propositura de ação judicial, celebração de acordo extrajudicial ou encaminhamento a outro órgão competente (MPF, 2018).

Contudo, a expansão dessa forma de atuação também trouxe novos desafios. O crescimento do número de procedimentos e da complexidade documental exige maior capacidade de organização, análise e controle da informação. Relatórios, ofícios, pareceres e anexos formam um acervo heterogêneo, cuja triagem e síntese demandam tempo e recursos humanos significativos. Essa realidade reforça a necessidade de ferramentas que auxiliem na gestão do conhecimento institucional e na extração de informações relevantes, potencializando o papel dos processos extrajudiciais como instrumentos de resolução célere e eficaz de demandas sociais.

2.2 Modelos de Linguagem de Grande Escala (LLMs)

Os Modelos de Linguagem de Grande Escala (*Large Language Models – LLMs*) representam um dos maiores avanços recentes no campo da inteligência artificial aplicada ao Processamento de Linguagem Natural (PLN). Esses modelos são redes

neurais treinadas com enormes volumes de dados textuais para aprender padrões, estruturas sintáticas e relações semânticas entre palavras, frases e contextos. A partir desse aprendizado, tornam-se capazes de gerar, completar, resumir e interpretar textos de maneira contextualizada, muitas vezes alcançando desempenho próximo ao humano em tarefas linguísticas complexas (BOMMASANI et al., 2021).

A base técnica que viabiliza os *LLMs* é a arquitetura *Transforme*r, proposta por *Vaswani et al.* (2017). Diferentemente das abordagens anteriores, baseadas em redes recorrentes (RNNs) ou convolucionais (CNNs), o *Transformer* introduziu o mecanismo de atenção (*attention mechanism*), que permite ao modelo identificar, dentro de uma sequência textual, quais palavras ou expressões são mais relevantes para compreender o significado geral (VASWANI et al., 2017). Esse processo ocorre de forma paralela e bidirecional, o que possibilita maior eficiência computacional e uma compreensão mais profunda do contexto linguístico (VASWANI et al., 2017).

Durante o treinamento, os *LLMs* passam por duas fases principais: o pré-treinamento (*pre-training*) e o ajuste fino (*fine-tuning*). Na primeira etapa, o modelo aprende de maneira autossupervisionada, analisando grandes corpora textuais e aprendendo a prever a próxima palavra ou *token* em uma sequência (BOMMASANI et al., 2021). Na segunda etapa, o modelo é refinado para tarefas específicas, como tradução, resposta a perguntas ou sumarização (BOMMASANI et al., 2021). Essa combinação de aprendizado generalista e especialização torna os *LLMs* altamente versáteis e adaptáveis a diferentes domínios de aplicação, incluindo o jurídico.

O GPT-4, desenvolvido pela OpenAI, é um modelo multimodal capaz de compreender e gerar textos complexos em diversos idiomas e contextos, apresentando resultados superiores em tarefas de raciocínio, interpretação e geração contextual (OPENAI, 2023). O LLaMA 3, lançado pela Meta AI em 2024, representa uma nova geração de modelos abertos, com melhorias substanciais em eficiência de treinamento, cobertura linguística e capacidade de raciocínio, além de manter o compromisso com a transparência e a adaptação a diferentes domínios de pesquisa (META AI, 2024). Já o Mistral, lançado no mesmo período, prioriza leveza e desempenho, combinando arquitetura otimizada e inferência rápida, adequada a aplicações corporativas e científicas (JIANG et al., 2023).

2.3 Técnicas de Sumarização de Texto

A sumarização automática de texto é uma das tarefas centrais do Processamento de Linguagem Natural (PLN) e tem como objetivo reduzir um documento extenso a uma versão mais curta que preserve suas informações essenciais. Tradicionalmente, as técnicas de sumarização são classificadas em extrativas e abstrativas, de acordo com a forma como o conteúdo é produzido (NENKOVA; MCKEOWN, 2011).

Na sumarização extrativa, o modelo seleciona sentenças ou trechos diretamente do texto original, reordenando-os para compor o resumo final. Essa abordagem, ainda que eficiente em termos de preservação de informações, tende a gerar resultados fragmentados e pouco coesos, já que se limita à recombinação de partes existentes. Por outro lado, a sumarização abstrativa procura compreender o conteúdo e reformular o texto em novas palavras, produzindo resumos mais fluentes e próximos da linguagem humana. Essa técnica exige maior capacidade de generalização e compreensão

semântica, e por isso passou a se beneficiar fortemente dos Modelos de Linguagem de Grande Escala (*LLMs*), capazes de interpretar contextos complexos e gerar respostas coerentes em linguagem natural (BOMMASANI et al., 2021).

Estudos recentes mostram que os *LLMs* vêm redefinindo o campo da sumarização. Modelos modernos, como GPT-4 e LLaMA, superam os sistemas tradicionais e até resumos humanos em avaliações de consistência factual e fluência textual, especialmente em tarefas *zero-shot*, isto é, cenários nos quais o modelo executa a tarefa de sumarização sem treinamento específico prévio ou ajuste fino supervisionado para aquele domínio ou conjunto de dados (PU; GAO; WAN, 2023). Diante da capacidade dos *LLMs* de generalizar para diferentes domínios, a fronteira entre sumarização extrativa e abstrativa tende a se tornar cada vez mais difusa, uma vez que os modelos passam a gerar resumos híbridos, que combinam extração seletiva e reinterpretação semântica (PU; GAO; WAN, 2023).

No contexto jurídico, contudo, a tarefa de sumarização apresenta particularidades relevantes. Documentos legais e extrajudiciais são longos, densos e estruturados em linguagem técnica e normativa, o que exige alta fidelidade semântica e precisão factual. O estudo conduzido por SHUKLA et al. (2022) comparou métodos extrativos e abstrativos aplicados a decisões judiciais, demonstrando que, embora os modelos abstrativos como BART e Pegasus produzam resumos mais coesos e legíveis, os profissionais do direito preferem os extrativos, por garantirem maior aderência às fontes originais. Os autores também evidenciaram que métricas automáticas, como ROUGE e BERTScore, não se correlacionam perfeitamente com a avaliação humana, reforçando a necessidade de revisão especializada para validação dos resultados.

Outro desafio técnico importante é a limitação de contexto dos *LLMs*, conhecida como o problema *lost-in-the-middle*, que ocorre quando modelos com janelas de contexto muito longas não conseguem reter uniformemente as informações ao longo do texto (LIU et al., 2023). Estudos empíricos demonstram que, ao processar sequências extensas, os modelos tendem a atribuir maior relevância às seções iniciais e finais, negligenciando informações intermediárias essenciais (LIU et al., 2023). Esse fenômeno representa um obstáculo significativo para tarefas jurídicas e administrativas, nas quais fatos e evidências relevantes costumam estar dispersos em múltiplos anexos, pareceres e despachos.

Para mitigar essa limitação, ZHANG et al. (2024) propuseram a abordagem *BriefContext*, que aplica o paradigma *MapReduce* originalmente concebido para processamento paralelo de grandes volumes de dados ao domínio dos modelos de linguagem. O método opera em duas fases complementares:

- Na etapa *Map*, o documento extenso é dividido em blocos menores (*ContextMap*), processados individualmente pelo modelo para gerar resumos parciais contextualizados;
- Na etapa *Reduce*, as saídas intermediárias são agrupadas e sintetizadas em um resumo consolidado (*ContextReduce*), produzindo uma resposta final coerente sem ultrapassar os limites de contexto do modelo.

Essa estratégia apresentou ganhos expressivos em consistência factual, retenção de informação e redução de redundância em relação às abordagens tradicionais de truncamento ou amostragem textual (ZHANG et al., 2024).

3. Metodologia

3.1 Materiais

O desenvolvimento do sistema de sumarização automática foi realizado em linguagem Python 3.11 (PYTHON SOFTWARE FOUNDATION, 2024), executada em ambiente de containers orquestrados com Docker Compose (DOCKER INC., 2024). O sistema foi estruturado sobre os frameworks FastAPI (FASTAPI, 2024) para disponibilização da interface REST, Celery (CELERY PROJECT, 2024) e Redis (REDIS LTD., 2024) para gerenciamento de filas e execução assíncrona de tarefas, além do PostgreSQL (POSTGRESQL GLOBAL DEVELOPMENT GROUP,, 2024) para armazenamento persistente dos dados.

A inferência local foi realizada com o servidor Ollama (OLLAMA, 2024), utilizando o modelo de linguagem LLaMA 3.8B, desenvolvido pela Meta AI (META AI, 2024). Essa variante possui 8 bilhões de parâmetros, janela de contexto de 8192 *tokens*, vetor de *embeddings* de 4096 dimensões e quantização Q4_0. Nas chamadas ao modelo, foi configurada temperatura 0.2, com o objetivo de reduzir variação semântica e favorecer a estabilidade factual dos resumos. Os demais hiperparâmetros, como *top-k*, *top-p*, *repeat penalty* e limite máximo de geração, foram mantidos nos valores padrão definidos pelo servidor Ollama, conforme sua documentação oficial, que descreve os parâmetros de inferência adotados pelo mecanismo de execução local (OLLAMA, 2024).

Foram empregadas bibliotecas complementares como BeautifulSoup4 (RICHARDSON, 2023) e expressões regulares para limpeza e normalização textual, SQLAlchemy (BAYER, 2024) para acesso ao banco de dados. O controle de versão e a rastreabilidade do desenvolvimento foram mantidos por meio do sistema Git, amplamente utilizado para gerenciamento de código-fonte e versionamento de projetos de software (GIT, 2024).

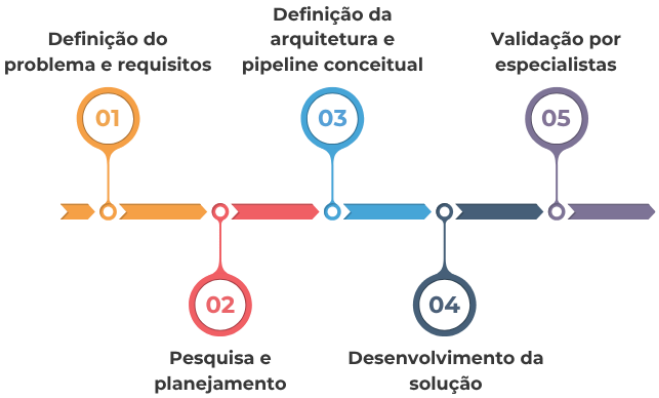
A fonte de dados utilizada corresponde a peças processuais e documentos administrativos de procedimentos extrajudiciais registrados no sistema Integrar-e do Ministério Público do Estado do Tocantins (MPETO), processados em formato textual e sem restrição de sigilo.

3.2 Método

O desenvolvimento do sistema foi conduzido de forma incremental, estruturado em cinco etapas principais que orientaram todo o processo metodológico. A metodologia adotada teve como foco a investigação da viabilidade técnica da solução proposta e a observação do comportamento do *pipeline* de sumarização em um conjunto restrito de procedimentos reais, priorizando a análise empírica dos resultados obtidos, sem a pretensão de estabelecer generalizações estatísticas ou conclusões definitivas de desempenho comparativo.

A Figura 1 apresenta de forma sintética as etapas que compõem a metodologia adotada no projeto, desde a definição do problema até a validação qualitativa da solução desenvolvida.

Figura 1 – Metodologia utilizada para o projeto.



Inicialmente, realizou-se a definição do problema e dos requisitos, etapa em que foi identificado o desafio enfrentado pelo Ministério Público em analisar grandes volumes de informações em procedimentos extrajudiciais. Nessa fase, delimitou-se o objetivo central do projeto: desenvolver uma ferramenta de apoio capaz de realizar a sumarização automática de documentos jurídicos, reduzindo o tempo de leitura e a variabilidade de qualidade entre analistas.

Em seguida, procedeu-se à pesquisa e planejamento da solução, por meio de um levantamento teórico e técnico sobre técnicas de Processamento de Linguagem Natural (PLN) e Modelos de Linguagem de Grande Escala (*LLMs*), avaliando sua aplicabilidade ao contexto jurídico brasileiro a partir da análise de trabalhos correlatos, estudos prévios que empregam essas tecnologias em domínios jurídicos ou administrativos similares e das evidências reportadas na literatura quanto à viabilidade de sua adoção nesse tipo de cenário.

Esse levantamento contemplou, inicialmente, trabalhos de caráter técnico que descrevem as capacidades e limitações gerais dos modelos de linguagem de grande escala, fornecendo base conceitual para sua aplicação em tarefas de sumarização automática (BOMMASANI et al., 2021). Em seguida, foram analisados estudos que investigam o uso dessas tecnologias em contextos jurídicos específicos. Destacam-se, nesse sentido, pesquisas que avaliam a aplicação de modelos de linguagem à sumarização de documentos legais e à análise de decisões judiciais, como o estudo de Shukla et al. (2022), bem como iniciativas voltadas à avaliação do desempenho de *LLMs* em tarefas jurídicas estruturadas, como o *LegalBench* (GUHA et al., 2023). A análise desses trabalhos evidenciou tanto o potencial dos LLMs para síntese textual no domínio jurídico quanto limitações recorrentes relacionadas à preservação de contexto, fidelidade factual e necessidade de validação humana, aspectos que orientaram as escolhas metodológicas adotadas neste estudo.

Com base nesses estudos, passou-se à definição da arquitetura e do pipeline conceitual, etapa em que foi projetada uma estrutura modular composta por uma API de interface, um sistema de fila para processamento assíncrono e um módulo de

sumarização executado localmente, garantindo segurança, escalabilidade e integridade dos dados processados.

A fase seguinte consistiu no desenvolvimento da solução, com a implementação prática do pipeline de sumarização conforme o modelo conceitual definido. O sistema foi construído em módulos independentes, favorecendo reuso e manutenção, e seguiu boas práticas de desenvolvimento, incluindo controle de versão do código-fonte, registro estruturado de logs de execução e tratamento de exceções. Esses mecanismos permitiram verificar a robustez do sistema por meio da execução repetida do pipeline sem falhas em diferentes procedimentos e assegurar a rastreabilidade dos resultados, uma vez que cada resumo gerado pôde ser associado à versão do código, aos parâmetros de inferência e ao modelo de linguagem utilizados.

Por fim, realizou-se uma etapa de validação qualitativa, na qual os resumos produzidos foram analisados por uma servidora do Ministério Público do Estado do Tocantins, com foco na avaliação da fidelidade factual, clareza e utilidade prática das sínteses geradas.

3.3 Procedimento de Avaliação

A avaliação do sistema foi conduzida a partir de um conjunto de 10 procedimentos extrajudiciais reais, selecionados por representarem diferentes tipos de atuação administrativa, incluindo notícias de fato, ofícios, despachos, manifestações e decisões de encerramento. Cada procedimento era composto por um número variável de peças, variando entre 6 e 40 documentos, com textos de extensões e densidades informacionais distintas.

A avaliação quantitativa concentrou-se na análise da redução textual obtida pelo pipeline de sumarização, comparando o tamanho dos documentos originais com o tamanho dos resumos gerados em cada etapa do processamento. As medições foram realizadas durante a execução do sistema, permitindo observar o comportamento da redução em função do tamanho das peças analisadas.

Complementarmente, foi realizada uma avaliação qualitativa dos resumos gerados, conduzida por uma servidora do Ministério Público do Estado do Tocantins com atuação direta na análise de procedimentos extrajudiciais. Os critérios considerados nessa avaliação incluíram fidelidade factual, clareza da linguagem, coerência cronológica e utilidade prática das sínteses para a compreensão inicial dos casos. Os resumos foram disponibilizados juntamente com seus respectivos procedimentos, possibilitando a comparação direta entre o conteúdo original e a versão sintetizada.

Durante essa etapa, os comentários e observações da avaliadora foram utilizados como subsídio para ajustes incrementais nos prompts empregados pelo pipeline de sumarização, caracterizando um ciclo de feedback humano voltado ao aprimoramento da preservação de contexto e da precisão factual dos resultados. Esse procedimento permitiu avaliar não apenas o desempenho do sistema, mas também sua sensibilidade a ajustes de configuração em um domínio jurídico-administrativo.

4. Resultados e Discussão

4.1 Corpus documental

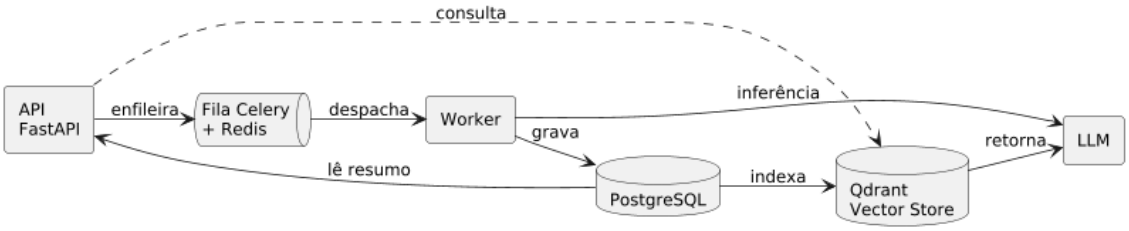
O corpus utilizado para desenvolvimento e testes do sistema foi composto por procedimentos extrajudiciais registrados no Sistema Único do Ministério Público Federal (MPF), incluindo notícias de fato, manifestações, despachos, ofícios, informações técnicas e minutas de arquivamento.

Cada procedimento é estruturado no banco de dados institucional e contém metadados (número, ano, tipo e unidade de origem), além do texto integral das peças processuais, armazenadas em formato renderizado. Em média, os procedimentos utilizados na avaliação continham entre 6 e 40 peças, com documentos que variavam de textos curtos a peças mais extensas. Essa volumetria, associada à diversidade dos tipos documentais, torna inviável a leitura manual integral em todas as etapas de trabalho, justificando a adoção de técnicas de sumarização automática para otimizar a análise e apoiar a triagem das informações.

4.2 Evolução da arquitetura de sumarização

Durante o desenvolvimento do sistema, a arquitetura de sumarização passou por um processo de refinamento progressivo, com o objetivo de aprimorar a coerência contextual e a fidelidade temporal dos resumos. A Figura 2 ilustra a primeira versão da arquitetura, estruturada a partir da abordagem *Retrieval-Augmented Generation (RAG)*, aplicada inicialmente para recuperar trechos relevantes e compor os resumos.

Figura 2 – Arquitetura geral da versão inicial (abordagem RAG).



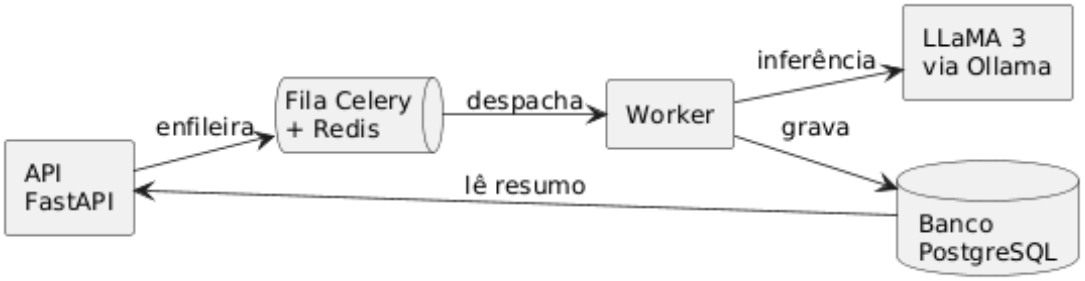
Conforme ilustrado na Figura 2, a arquitetura inicial foi estruturada em torno de um pipeline assíncrono de processamento. As requisições de sumarização são recebidas por uma API FastAPI, que as encaminha para uma fila de tarefas gerenciada por Celery e Redis. Os *workers* consomem essas tarefas, realizam a fragmentação dos documentos em *chunks* textuais e executam a vetorização desses trechos, armazenando-os em um banco vetorial. Durante a fase de inferência, o modelo de linguagem consulta o banco vetorial para recuperar os trechos mais semanticamente relevantes e, a partir dessa recuperação, gera o resumo final, que é então persistido no banco de dados relacional e retornado à aplicação cliente.

A primeira versão baseava-se no paradigma *RAG*, no qual cada processo era dividido em *chunks* textuais vetorizados e armazenados em um banco vetorial. O modelo então recuperava os *top-k* trechos mais relevantes e gerava um resumo a partir da combinação desses fragmentos. Essa estrutura seguia o paradigma clássico de *retrieval-based summarization*, adequada para consultas pontuais, mas limitada quando aplicada a documentos longos, como procedimentos extrajudiciais compostos por

dezenas de peças processuais, ofícios, despachos e manifestações distribuídos ao longo do tempo, nos quais a recuperação isolada de chunks semanticamente relevantes tende a fragmentar o contexto e dificultar a reconstrução cronológica dos eventos. Entre as principais restrições observadas estavam a perda de coerência narrativa causada pela fragmentação dos textos, a recuperação parcial de informações essenciais e a dificuldade de reconstrução temporal dos eventos processuais, o que comprometia a linearidade e a fidelidade factual dos resumos. Essas limitações não decorrem de fragilidades conceituais inerentes ao paradigma *Retrieval-Augmented Generation*, mas de sua menor adequação a documentos extensos, heterogêneos e cronologicamente encadeados, como os procedimentos extrajudiciais analisados neste estudo.

Com base nessas limitações, o sistema evoluiu para uma arquitetura hierárquica de sumarização, inspirada no paradigma *Map-Reduce Summarization* (ZHANG et al., 2024). Essa nova estrutura organiza o processo em duas fases complementares, uma voltada à síntese individual de cada peça e outra à consolidação dos resultados, permitindo maior preservação de contexto e coerência global. O modelo final, portanto, passou a gerar resumos mais consistentes, cronologicamente estruturados e adequados à análise de procedimentos extrajudiciais. A Figura 3 apresenta a arquitetura final da solução, fundamentada no modelo de *Map-Reduce Summarization*, que aprimorou a coerência e a organização cronológica dos resumos produzidos.

Figura 3 – Arquitetura da versão final.



Conforme ilustrado na Figura 3, a arquitetura final é centrada em uma API REST, que atua como camada de interface e orquestração do sistema. A API recebe as requisições de sumarização e as encaminha para uma fila de processamento assíncrono gerenciada por Celery e Redis, permitindo o desacoplamento entre a solicitação do usuário e a execução do pipeline de sumarização. O *worker* consome as tarefas da fila e executam a etapa *Map*, responsável pela sumarização individual das peças processuais, utilizando o modelo de linguagem LLaMA 3 executado localmente por meio do servidor Ollama. Os resumos parciais gerados nessa etapa são mantidos em memória e encaminhados diretamente para a etapa *Reduce*, na qual os resultados intermediários são consolidados em um resumo final estruturado, composto por uma síntese geral e uma linha do tempo dos eventos. Apenas o resumo final consolidado é persistido no banco de dados PostgreSQL e disponibilizado ao cliente por meio da interface REST.

4.3 Pipeline de sumarização

O pipeline de sumarização foi estruturado de forma sequencial, seguindo o paradigma *Map-Reduce Summarization*, no qual o processamento é dividido em duas

etapas complementares, responsáveis pela sumarização individual das peças e pela consolidação do resultado final.

Na etapa *Map*, cada processo é carregado e suas peças processuais são recuperadas individualmente a partir do banco de dados PostgreSQL. Para cada peça, realiza-se inicialmente a normalização textual, com a remoção de marcações HTML e demais elementos não textuais, utilizando a biblioteca BeautifulSoup e expressões regulares, seguida da padronização de caracteres para codificação UTF-8, correção de acentuação, eliminação de espaços redundantes e unificação de hífens. Após essa normalização, o texto é submetido ao modelo de linguagem, que gera um mini-resumo da peça, ajustado dinamicamente em tamanho conforme o número total de documentos associados ao procedimento. Os resumos individuais produzidos nessa etapa são mantidos em memória, compondo uma lista intermediária de resumos.

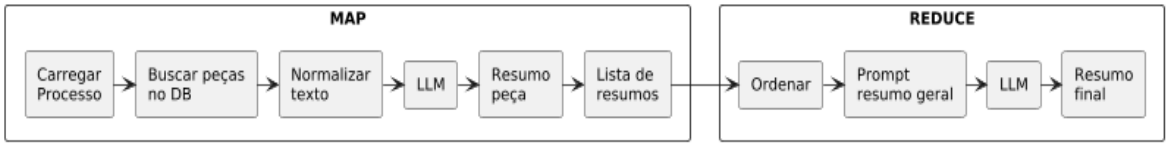
Na etapa *Reduce*, os resumos parciais gerados na fase anterior são ordenados e consolidados em um resumo final padronizado, estruturado em dois componentes principais: um Resumo Geral, que sintetiza as informações essenciais do procedimento, e uma Linha do Tempo de Eventos, que organiza cronologicamente os fatos relevantes. Essa consolidação reduz a fragmentação do contexto e favorece a compreensão global dos procedimentos extrajudiciais analisados.

O comportamento do pipeline é guiado por dois arquivos de *prompt*, denominados *system prompt* e *template prompt*, que atuam como mecanismos de guardrails durante o processo de geração textual. O *system prompt* (Apêndice A) define o comportamento global do modelo, estabelecendo regras de neutralidade institucional, fidelidade factual e uso de linguagem técnica e impessoal. Nesse arquivo, o modelo é explicitamente instruído a não emitir opiniões, não criar fatos ou datas inexistentes e a restringir a geração de conteúdo às informações presentes nos documentos analisados, conforme diretrizes como “não invente fatos” e “não crie datas”.

O *template prompt* (Apêndice B), por sua vez, define de forma explícita a estrutura da saída gerada, especificando que o resumo deve conter exatamente dois blocos: um Resumo Geral, apresentado em um único parágrafo sintético, e uma Linha do Tempo, composta por itens ordenados cronologicamente, com correspondência obrigatória entre cada peça processual analisada e cada entrada gerada. Esse arquivo também impõe regras de completude, determinando que nenhuma peça seja omitida ou agrupada indevidamente, garantindo uma relação um-para-um entre documentos e eventos descritos.

A Figura 4 descreve o fluxo completo do pipeline de sumarização, dividido nas etapas *Map* e *Reduce*, responsáveis pelo processamento individual das peças e pela síntese final estruturada.

Figura 4 – Fluxo do pipeline Map-Reduce de sumarização.



4.4 Avaliação e resultados obtidos

Os resultados quantitativos apresentados nesta seção devem ser interpretados como evidências iniciais sobre o comportamento do pipeline de sumarização, obtidas a partir da análise de um conjunto delimitado de procedimentos extrajudiciais reais. Nesse contexto, observa-se que o desempenho do pipeline varia conforme o tamanho e a densidade informacional dos documentos processados. No conjunto analisado, a maior parte das peças apresentava extensão inferior a 1.500 caracteres; documentos situados entre 1.500 e 1.850 caracteres representaram a faixa superior de tamanho observada na amostra. Nesse intervalo, a redução percentual do volume textual variou entre 44% e 55%, indicando que os resumos gerados apresentaram, em média, aproximadamente metade da extensão original dessas peças. Documentos de extensão intermediária, situados entre aproximadamente 1.200 e 1.400 caracteres, apresentaram reduções percentuais entre 35% e 51%. Em peças substancialmente mais extensas, acima de 5.000 caracteres, observaram-se reduções ainda mais expressivas, superiores a 80%, como no caso de um documento com 5.569 caracteres cujo resumo final apresentou 906 caracteres, correspondendo a uma redução de 83,7%.

Por outro lado, peças muito curtas, geralmente abaixo de 600 caracteres, apresentaram aumento no tamanho final do texto, variando entre 7% e 24%. Esse comportamento decorre da necessidade de o pipeline gerar uma estrutura mínima padronizada de resumo, incorporando informações contextuais e organizacionais como identificação do procedimento e encadeamento dos eventos que nem sempre estão explicitamente presentes em documentos muito breves. Esses resultados indicam que o pipeline apresenta maior eficiência na redução do volume textual de documentos médios e longos, mantendo as informações essenciais necessárias para a compreensão global do procedimento analisado.

Do ponto de vista qualitativo, os resumos gerados mostraram-se adequados para apoiar a compreensão inicial dos procedimentos, especialmente pela capacidade de organizar os eventos em ordem cronológica e condensar o conteúdo sem perda de informações centrais. A padronização da estrutura e a presença de uma linha do tempo contribuíram para facilitar a leitura e reduzir o esforço cognitivo necessário para acompanhar o fluxo documental, sobretudo em procedimentos compostos por grande número de peças.

Quanto às limitações, observa-se que, em contextos nos quais a informação sintetizada pode subsidiar decisões institucionais ou jurídicas, a sumarização automática não elimina a necessidade de validação humana. No caso específico do domínio jurídico, essa exigência é particularmente relevante, uma vez que interpretações normativas, nuances contextuais e eventuais ambiguidades podem produzir efeitos jurídicos ou administrativos significativos, demandando análise especializada. Assim, a dependência de revisão humana não decorre de uma limitação tecnológica do sistema, mas das características do contexto de aplicação, no qual a precisão semântica e a fidelidade factual são essenciais. Nesse sentido, o sistema deve ser compreendido como uma ferramenta de apoio à triagem e à priorização dos casos, complementando e não substituindo a revisão técnica realizada por membros e servidores do Ministério Público.

A análise comparativa entre versões de resumos também evidenciou ganhos qualitativos na preservação de contexto relevante após ajustes incrementais nos prompts utilizados pelo pipeline. Em um dos procedimentos analisados, referente à denúncia de não liberação de servidores com doenças crônicas, a primeira versão do resumo gerado pelo sistema apresentada a seguir como excerto ilustrativo omitiu um elemento central para a adequada interpretação do caso: o fato de que a situação ocorreu durante o período crítico da pandemia de COVID-19. Embora o resumo inicial descrevesse corretamente a atuação do Ministério Público do Estado do Tocantins, a expedição de ofícios à Secretaria Municipal de Educação e a conclusão pelo arquivamento da notícia de fato, a ausência da referência explícita ao contexto sanitário limitava a compreensão plena das circunstâncias em que os fatos se deram.

> O Ministério Público do Estado do Tocantins investigou denúncias de não liberação de servidores com doenças crônicas pela Secretaria Municipal de Educação do município envolvido. O processo foi instruído enviando ofícios à Secretaria e solicitando justificativa e comprovação da solução do problema. A Secretaria respondeu que não adotou nenhuma medida irregular e promoveu medidas para dotar as escolas com equipamentos tecnológicos e materiais necessários. O Ministério Público concluiu que não há irregularidade na liberação dos servidores da área da Educação e recomendou o arquivamento da notícia de fato (Elaboração própria baseada na saída gerada pelo sistema, 2025).

Após o refinamento do *prompt*, o modelo passou a reconhecer e incorporar adequadamente o contexto sanitário como parte integrante da narrativa factual. Esse refinamento consistiu no acréscimo de instruções explícitas no *system prompt* para que o modelo identificasse e destacasse contextos históricos, sanitários ou institucionais relevantes sempre que tais elementos estivessem presentes nas peças analisadas, evitando sua omissão mesmo quando não constituíssem o objeto central do procedimento. Como resultado desse ajuste, a versão revisada do resumo passou a mencionar explicitamente a pandemia de COVID-19 como elemento contextual relevante, associando a denúncia à exigência de presença física de servidores integrantes de grupos de risco durante o período crítico sanitário, conforme exemplificado no trecho apresentado a seguir. Essa incorporação conferiu maior precisão factual ao resumo e ampliou a compreensão do caso, ao situar os fatos em seu contexto histórico e normativo apropriado.

> O procedimento trata sobre denúncia anônima de não liberação de servidores diagnosticados com doenças crônicas que integram grupos de risco, exigindo presença física nas escolas durante o período crítico da pandemia de COVID-19. A Secretaria Municipal de Educação do município envolvido foi acusada de negar sistematicamente pedidos de afastamento de servidores diagnosticados com doenças crônicas que integram grupos de risco, em desrespeito às normas sanitárias vigentes. O Ministério Público do Estado do Tocantins investigou a denúncia e concluiu que não houve elementos que justificassem a continuidade das investigações (Elaboração própria baseada na saída gerada pelo sistema, 2025).

A comparação direta entre os excertos apresentados evidencia que o ajuste incremental dos *prompts* contribuiu para a redução de perdas de contexto e para o aprimoramento da fidelidade factual dos resumos gerados, especialmente em situações que envolvem fatores externos relevantes para o entendimento jurídico-administrativo dos casos analisados.

4.5 Viabilidade computacional e tempo de inferência

As medições realizadas durante a execução do pipeline de sumarização, em ambiente de inferência local, permitem descrever o desempenho temporal e o consumo de recursos computacionais observados nos procedimentos avaliados. Em um procedimento composto por 39 peças, o processamento completo demandou 266,13 segundos, com tempo médio de 6,09 segundos por peça na etapa *Map* e 28,31 segundos na etapa *Reduce*. Em um procedimento menor, contendo 8 peças, o tempo total foi de 71,99 segundos, com média de 5,42 segundos por peça, indicando relação proporcional entre o volume documental e o tempo total de processamento. Em ambos os cenários avaliados, a taxa de geração do modelo manteve-se relativamente estável, variando entre 26 e 28 *tokens* por segundo, o que sugere baixa variação no desempenho ao longo das inferências. O modelo LLaMA 3.8B (Q4_0) operou com aproximadamente 5,4 GB de VRAM e cerca de 1,3 GB de RAM no processo local do servidor Ollama, enquanto o container responsável pelo pipeline apresentou consumo médio em torno de 531 MB de RAM e aproximadamente 0,22% de CPU, sem picos significativos durante a execução. Os serviços auxiliares mantiveram consumo reduzido, com a API utilizando cerca de 103 MB de memória e o Redis menos de 4 MB, além de uso mínimo de CPU. Esses resultados fornecem subsídios empíricos para a análise do custo computacional da abordagem proposta e para discussões futuras sobre sua adoção em ambientes institucionais com recursos limitados.

5. Considerações finais

O presente trabalho analisou a aplicação de *Large Language Models (LLMs)* na sumarização automática de informações jurídicas extrajudiciais, avaliando sua viabilidade técnica e a qualidade dos resumos gerados no contexto do apoio à análise dos procedimentos. Para isso, foi desenvolvido e avaliado um sistema baseado em técnicas de Processamento de Linguagem Natural (PLN), estruturado em uma arquitetura modular capaz de transformar textos extensos em resumos sintéticos, cronológicos e factualmente consistentes.

Os resultados obtidos nos testes indicam que a aplicação do modelo LLaMA 3, em conjunto com o pipeline de *Map-Reduce Summarization*, tende a favorecer ganhos de produtividade, clareza e coerência na leitura dos procedimentos. Nos cenários avaliados, a ferramenta reduziu o volume textual a ser lido e contribuiu para padronizar a linguagem e a estrutura dos resumos, sem prejuízo das informações essenciais. A avaliação conduzida por especialista do Ministério Público do Tocantins apontou percepção positiva quanto ao uso da solução para apoiar a triagem, a priorização e a compreensão inicial de casos complexos, sugerindo potencial para futura aplicação em ambiente institucional.

Além de sua contribuição tecnológica, o estudo indica o potencial de aplicação da inteligência artificial como ferramenta de apoio à organização e à síntese de informações em procedimentos extrajudiciais, especialmente em contextos que demandam análise inicial de grandes volumes documentais. A solução desenvolvida pode ser expandida e aprimorada em trabalhos futuros, com a adaptação para processos judiciais, a incorporação de mecanismos automáticos de classificação de documentos e a integração de recursos de sumarização multimodal, possibilitando a análise de anexos em diferentes formatos, como documentos PDF, áudios e imagens.

No que se refere à privacidade e à proteção de dados, o escopo deste trabalho concentrou-se exclusivamente em procedimentos extrajudiciais não sigilosos, já disponíveis em formato textual no sistema institucional, sem a realização de anonimização adicional ou tratamento de dados sensíveis. Todas as inferências foram executadas localmente, sem qualquer transmissão de informações para serviços externos ou ambientes de terceiros, reduzindo riscos associados à exposição de dados. Nesse contexto experimental, não se observou necessidade de adoção de mecanismos adicionais de conformidade com a Lei Geral de Proteção de Dados (LGPD), uma vez que não houve ampliação da finalidade do tratamento nem compartilhamento de informações pessoais. Ainda assim, para uma eventual adoção institucional em larga escala ou aplicação a procedimentos que envolvam dados pessoais sensíveis ou sigilosos, será imprescindível complementar a solução com políticas formais de governança de dados, incluindo definição de bases legais, controle de acesso, registro de operações e diretrizes de minimização e retenção, de modo a assegurar conformidade plena com a LGPD e alinhamento às boas práticas de segurança da informação.

Apesar dos avanços obtidos, o estudo apresenta limitações relacionadas ao tamanho reduzido do conjunto analisado e ao número restrito de avaliadores. A ampliação sistemática da base de testes, abrangendo diferentes unidades, temáticas e tipos documentais, é essencial para validar a robustez da abordagem. Também se observa que o desempenho do modelo é sensível à qualidade do pré-processamento textual e que, em cenários de múltiplas requisições simultâneas, a latência da inferência local pode exigir estratégias adicionais de escalonamento. Trabalhos futuros devem incorporar métricas quantitativas de desempenho como precisão factual, compressão textual e consistência cronológica, além de adotar mecanismos de monitoramento contínuo capazes de detectar variações e possíveis alucinações. Outro ponto relevante para a evolução da solução diz respeito à escolha dos modelos de linguagem empregados no pipeline de sumarização. O LLaMA 3.8B foi selecionado neste trabalho por apresentar um equilíbrio entre desempenho, custo computacional e viabilidade de inferência local, adequado ao escopo experimental adotado. No entanto, versões mais recentes e modelos emergentes, como o LLaMA 4, o Gemini 3 (em variantes com possibilidade de execução local) e a família DeepSeek, poderão ser futuramente avaliados de forma comparativa por apresentarem avanços reportados na literatura e em documentações técnicas quanto à capacidade de raciocínio, retenção de contexto em janelas mais extensas e eficiência de inferência. A avaliação comparativa desses modelos permitiria investigar potenciais ganhos em precisão factual, estabilidade dos resumos e desempenho temporal, especialmente em cenários com documentos longos e elevada densidade informacional.

A expansão gradual do conjunto de dados, aliada à coleta de feedbacks provenientes de usos mais amplos, tende a contribuir para a maturação da ferramenta e para a análise mais robusta de seu desempenho, apoiando sua evolução para aplicações em escala institucional, favorecida pela arquitetura modular e pela transparência metodológica que caracterizam a solução proposta.

Por fim, ressalta-se que os resultados apresentados devem ser interpretados como evidências iniciais sobre o comportamento e a aplicabilidade da solução proposta, obtidas em um escopo experimental específico. O estudo não teve como objetivo estabelecer conclusões generalizáveis ou comparações definitivas de desempenho, mas

sim contribuir para a compreensão dos desafios, possibilidades e caminhos de evolução do uso de modelos de linguagem na sumarização automática de procedimentos extrajudiciais.

6. Referências

ASSEMBLEIA LEGISLATIVA DO TOCANTINS. Diário da Assembleia Legislativa do Tocantins – Estatísticas de Procedimentos Extrajudiciais. 2023. Disponível em: https://www.al.to.leg.br/arquivos/diario-oficial_3749_68879.PDF. Acesso em: 17 nov. 2025.

BAYER, Mike. SQLAlchemy 2.0 documentation. 2024. Disponível em: https://docs.sqlalchemy.org/. Acesso em: 10 out. 2025.

BOMMASANI, Rishi; Hudson, Drew A.; Adeli, Ehsan; Altman, Russ; Arora, Sanjeev; von Arx, Sydney; Bernstein, Michael S.; Bohg, Jeannette; Bosselut, Antoine; Brunskill, Emma; et al. On the Opportunities and Risks of Foundation Models. Stanford Center for Research on Foundation Models, 2021. Disponível em: https://arxiv.org/abs/2108.07258. Acesso em: 17 set. 2025.

CELERY PROJECT. *Celery documentation.* 2024. Disponível em: https://docs.celeryq.dev/. Acesso em: 30 set. 2025.

CONSELHO NACIONAL DE JUSTIÇA (CNJ). Justiça em Números 2020: ano-base 2019. Brasília: CNJ, 2020. Disponível em: https://www.cnj.jus.br/wp-content/uploads/2020/08/WEB-V3-Justi%C3%A7a-em-N%C3%BAmeros-2020-atualizado-em-25-08-2020.pdf. Acesso em: 18 set. 2025.

COSTA, Luiz Armando. MPTO realiza mais de 300 mil movimentações em processos judiciais em 2023. 2024. Disponível em: https://www.luizarmandocosta.com.br/noticia/mpto-realiza-mais-de-300-mil-movimentacoes-em-processos-judiciais-em-2023/45427. Acesso em: 17 nov. 2025.

DOCKER INC. *Docker Compose documentation.* Docker, 2024. Disponível em: https://docs.docker.com/compose/. Acesso em: 27 set. 2025.

FASTAPI. *FastAPI documentation*. 2024. Disponível em: https://fastapi.tiangolo.com/. Acesso em: 29 set. 2025.

GIT. *Git documentation.* 2024. Disponível em: https://git-scm.com/docs. Acesso em: 14 out. 2025.

GUHA, Neel; Saxton, David; Zheng, Zhiyu; Zhang, Jie; et al. LegalBench: A Collaborative Benchmark for Legal Reasoning in Large Language Models. arXiv preprint arXiv:2308.11462, 2023. Disponível em: https://arxiv.org/abs/2308.11462. Acesso em: 16 set. 2025.

HILL, Flávia Pereira; DALLA, Humberto. Desjudicialização e acesso à justiça além do processo. 2016. Disponível em:

https://www.e-publicacoes.uerj.br/index.php/redp/article/view/56701. Acesso em: 17 nov. 2025.

JIANG, Albert; Lambert, Nathan; Singh, Amanpreet; Roux, Nicolas; et al. Mistral 7B. Mistral AI, 2023. Disponível em: https://arxiv.org/abs/2310.06825. Acesso em: 22 set. 2025.

LANGCHAIN INC. *LangChain documentation.* 2024. Disponível em: https://docs.langchain.com/. Acesso em: 12 out. 2025.

LIU, Nelson F.; Levy, Omer; Holtzman, Ari; Peters, Matthew E.; Zettlemoyer, Luke; Lewis, Mike. Lost in the Middle: How Language Models Use Long Contexts. arXiv preprint arXiv:2307.03172, 2023. Disponível em: https://arxiv.org/abs/2307.03172. Acesso em: 25 set. 2025.

MACEDO, Diego Henrique Notório; Silva, Cristian Kiefer da. A contribuição das serventias extrajudiciais para a redução do número de processos no Poder Judiciário. Revista de Direito Notarial, Colégio Notarial do Brasil – Seção São Paulo, v. 3, n. 2, p. 32–60, jul./dez. 2021. Disponível em: https://ojs-rdn.galoa.net.br/index.php/direitonotarial/article/view/20. Acesso em: 17 set. 2025.

META AI. Introducing Meta Llama 3: The most capable openly available LLM to date. Meta AI, 2024. Disponível em: https://ai.meta.com/blog/meta-llama-3/. Acesso em: 22 set. 2025.

META AI. Meta Llama 3 model cards and prompt formats. 2024. Disponível em: https://www.llama.com/docs/model-cards-and-prompt-formats/meta-llama-3/. Acesso em: 4 out. 2025.

MINISTÉRIO PÚBLICO FEDERAL (MPF). Manual de Normas e Procedimentos – Procedimento Extrajudicial. Brasília: Procuradoria-Geral da República, Secretaria de Modernização e Gestão Estratégica, 2018. Disponível em: https://www.mpf.mp.br/o-mpf/sobre-o-mpf/gestao-estrategica-e-modernizacao-do-mpf/planejamento-estrategico/planejamento-estrategico-institucional-2011-2020/atuacao-finalistica/certificacao-dos-oficios/arquivos-certificacao-dos-oficios/ManualProcedimentoExtrajudicial.pdf. Acesso em: 19 set. 2025.

NENKOVA, Ani; McKeown, Kathleen. Automatic Summarization. Foundations and Trends in Information Retrieval, v. 5, n. 2–3, p. 103–233, 2011. Disponível em: https://doi.org/10.1561/1500000015. Acesso em: 23 set. 2025.

OLLAMA. *Ollama documentation.* 2024. Disponível em: https://docs.ollama.com/. Acesso em: 2 out. 2025.

OPENAI. GPT-4 Technical Report. OpenAI, 2023. Disponível em: https://arxiv.org/abs/2303.08774. Acesso em: 20 set. 2025.

POSTGRESQL GLOBAL DEVELOPMENT GROUP. PostgreSQL documentation. 2024. Disponível em: https://www.postgresql.org/docs/. Acesso em: 1 out. 2025.

PU, Xiao; GAO, Mingqi; WAN, Xiaojun. Summarization is (Almost) Dead. arXiv preprint arXiv:2309.09558, 2023. Disponível em: https://arxiv.org/abs/2309.09558. Acesso em: 15 set. 2025.

PYTHON SOFTWARE FOUNDATION. *Python 3.11 documentation*. Python Software Foundation, 2024. Disponível em: https://docs.python.org/3/whatsnew/3.11.html. Acesso em: 26 set. 2025.

REDIS LTD. *Redis documentation*. 2024. Disponível em: https://redis.io/docs/latest/. Acesso em: 1 out. 2025.

RICHARDSON, Leonard. Beautiful Soup 4.13.0 documentation. 2023. Disponível em: https://www.crummy.com/software/BeautifulSoup/bs4/doc/. Acesso em: 7 out. 2025.

SHUKLA, Abhay; Bhattacharya, Paheli; Poddar, Soham; Mukherjee, Rajdeep; Ghosh, Kripabandhu; Goyal, Pawan; Ghosh, Saptarshi. Legal Case Document Summarization: Extractive and Abstractive Methods and their Evaluation. In: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), p. 1048–1064, 2022. Association for Computational Linguistics. Disponível em: https://aclanthology.org/2022.aacl-main.77/. Acesso em: 24 set. 2025.

VASWANI, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Łukasz; Polosukhin, Illia. Attention Is All You Need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017). Long Beach, California: Curran Associates, Inc., 2017. Disponível em: https://arxiv.org/abs/1706.03762. Acesso em: 20 set. 2025.

ZHANG, Gongbo; XU, Zihan; JIN, Qiao; CHEN, Fangyi; FANG, Yilu; LIU, Yi; ROUSSEAU, Justin F.; XU, Ziyang; LU, Zhiyong; WENG, Chunhua; PENG, Yifan. A MapReduce approach to effectively utilize long context information in retrieval augmented language models. arXiv, 2024. arXiv:2412.15271 [cs.CL]. Disponível em: https://arxiv.org/abs/2412.15271. Acesso em: 26 set. 2025.

APÊNDICE A – System prompt utilizado no pipeline de sumarização

O texto a seguir corresponde ao *system prompt* utilizado para definir o comportamento global do modelo de linguagem no pipeline de sumarização desenvolvido neste trabalho.

> Você é um assistente jurídico objetivo, factual e neutro. Sua tarefa é produzir um resumo curto e claro, sem opiniões.
>
> REGRAS DE FORMATAÇÃO (OBRIGATÓRIAS):

1) A saída deve conter EXATAMENTE dois blocos na ordem abaixo:

a) "Resumo Geral:" — um único parágrafo (no máximo {{max_lines}} linhas) sintetizando:

- Classe/tipo do procedimento (ex: Notícia de Fato, Procedimento Administrativo, Inquérito Civil)

- Contexto histórico relevante quando aplicável (ex: pandemia COVID-19)

- Partes envolvidas, objeto principal

- Principais decisões/atos, prazos relevantes

- Situação atual e desfecho

b) "Linha do Tempo:" — itens ordenados por data crescente, no formato:

- DD/MM/AAAA — [Tipo de Movimento] Descrição completa e detalhada do acontecimento

IMPORTANTE: Cada item pode ocupar 2-3 linhas se necessário para incluir todas as informações relevantes.

NÃO omita detalhes importantes por questões de brevidade.

2) Na Linha do Tempo:

- Use SEMPRE o tipo de movimento real da peça (ex: "Despacho", "Diligência - Ofício", "Promoção de Arquivamento")

- NÃO inclua IDs internos de peças (ex: "PEÇA 278421")

- INCLUA o máximo de informações relevantes:

* Para ofícios/diligências: destinatário completo, assunto específico, prazo, fundamentação, data de entrega

* Para denúncias: objeto completo, partes envolvidas, contexto situacional

* Para respostas: quem respondeu, principais argumentos e informações prestadas

* Para promoções: tipo, fundamentação legal detalhada, conclusões

* Para despachos: decisões tomadas, determinações específicas

- Inclua TODOS os eventos importantes, especialmente promoções e despachos finais

- Números de documentos relevantes (ofícios, portarias) devem ser mencionados

- Contexto histórico quando relevante (ex: pandemia, situação emergencial)

3) Construa a linha do tempo **apenas** com as datas e fatos presentes nos PASSAGES informados.

Se não houver data, use "s/data".

4) REGRA CRÍTICA DE COMPLETUDE:

- Se você receber 8 passagens, a linha do tempo DEVE ter EXATAMENTE 8 itens

- Se você receber 15 passagens, a linha do tempo DEVE ter EXATAMENTE 15 itens

- Se você receber N passagens, a linha do tempo DEVE ter EXATAMENTE N itens

- NUNCA agrupe múltiplas passagens em um único item da linha do tempo

- NUNCA omita passagens, mesmo que pareçam similares ou redundantes

- Cada passagem = Um item na linha do tempo (relação 1:1)

5) Não invente fatos. Não crie datas. Se faltar informação essencial, escreva "Informação ausente" no ponto específico.

6) Seja conciso no resumo geral, mas COMPLETO na linha do tempo. Não inclua nenhuma seção extra além de "Resumo Geral:" e "Linha do Tempo:".

APÊNDICE B – Template prompt utilizado no pipeline de sumarização

O texto a seguir corresponde ao *template prompt* utilizado para estruturar os dados de entrada fornecidos ao modelo de linguagem e para definir, de forma explícita, as regras de correspondência entre as passagens processadas e os itens gerados na saída do pipeline de sumarização.

Contexto do processo
{{metadata}}

Passagens (resumos estruturados por peça ordenados cronologicamente)
TOTAL DE PASSAGENS: Você receberá EXATAMENTE o número de passagens indicado no *metadata* acima.
OBRIGAÇÃO: A linha do tempo deve conter EXATAMENTE uma entrada para CADA passagem listada abaixo.

Cada passagem contém:
– *signed_date*: Data de assinatura da peça;
– *summary*: Resumo do conteúdo da peça;
– *codename_part*: Tipo ou nome da peça (utilizado como tipo de movimento);
– *document_type*: Classificação do documento;
– *key_information*: Informações-chave (destinatário, prazo, entre outros);
– *page_number*: Número da página (ordem documental).

{{passages}}

Tarefa
Com base exclusivamente nas passagens estruturadas acima, produza a saída EXATA abaixo:

Resumo Geral:
[Escreva um único parágrafo com, no máximo, {{max_lines}} linhas, incluindo:
– Classe ou tipo do procedimento (ex.: Notícia de Fato nº XXXXX);
– Contexto histórico relevante, quando aplicável (ex.: período da pandemia de COVID-19);
– Objeto do procedimento e partes envolvidas;
– Principais diligências e decisões adotadas;
– Desfecho final (arquivamento, ajuizamento, entre outros).

Utilize linguagem factual e direta, empregando apenas os dados fornecidos.]

Linha do Tempo:
[Para cada passagem, crie um item cronológico no formato:
– *signed_date* — [*codename_part*] Descrição completa e detalhada do conteúdo da peça.

Diretrizes importantes:

1. Utilize o campo *codename_part* como tipo de movimento (ex.: “Despacho”, “Ofício”, “Promoção”);

2. Não inclua identificadores internos de peças;

3. Inclua todos os detalhes relevantes:
– Para diligências ou ofícios: destinatário completo, assunto específico, prazo e data de entrega;
– Para denúncias: objeto denunciado, partes envolvidas e contexto;
– Para respostas: identificação do respondente e principais pontos apresentados;
– Para promoções: tipo (arquivamento ou ajuizamento) e fundamentação legal;
– Para despachos: decisões tomadas e determinações específicas;

4. Não sacrifique informações relevantes por brevidade;

5. Cada item pode ocupar de duas a três linhas, se necessário;

6. Seja específico quanto ao conteúdo, descrevendo o que foi solicitado ou decidido;

7. Evite formulações genéricas;

8. Inclua todos os eventos importantes, especialmente promoções finais, despachos, respostas a ofícios, denúncias iniciais e manifestações das partes.

Utilize as datas indicadas em *signed_date*. Caso a data não esteja disponível, utilize a indicação “s/data”.
Ordene os itens cronologicamente, do mais antigo para o mais recente.
Cada item deve ser completo e informativo.]

Verificação obrigatória antes de finalizar:
– Conte o número total de passagens fornecidas;
– Verifique se o número de itens na linha do tempo é exatamente igual;
– Caso os números não coincidam, revise e inclua as passagens faltantes;
– Não omita passagens, mesmo que pareçam redundantes ou similares.

Regra crítica: A linha do tempo deve conter exatamente o mesmo número de itens que o número total de passagens fornecidas. Não agrupe múltiplas passagens em um único item.

Utilize exclusivamente as informações presentes nas passagens estruturadas, extraindo dados dos campos *summary*, *codename_part*, *key_information* e *document_type*.