DISSERTAÇÃO DE MESTRADO PROFISSIONAL

ChatORION: FRAMEWORK INTELIGENTE PARA AUTOMAÇÃO CONVERSACIONAL E TRIAGEM DE PROCESSOS NO CONTENCIOSO JUDICIAL COM UTILIZAÇÃO DE IA GENERATIVA

Elon Oliveira Albuquerque

Brasília, 5 de março de 2026

UNIVERSIDADE DE BRASÍLIA

FACULDADE DE TECNOLOGIA

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

ChatORION: FRAMEWORK INTELIGENTE PARA AUTOMAÇÃO CONVERSACIONAL E TRIAGEM DE PROCESSOS NO CONTENCIOSO JUDICIAL COM UTILIZAÇÃO DE IA GENERATIVA.

Elon Oliveira Albuquerque

ORIENTADOR: PROFESSOR Dr. FÁBIO LÚCIO LOPES DE MENDONÇA

DISSERTAÇÃO DE MESTRADO PROFISSIONAL EM ENGENHARIA ELÉTRICA

PUBLICAÇÃO: PPEE.MP.110
BRASÍLIA/DF, 05 de março - 2026

UNIVERSIDADE DE BRASÍLIA
Faculdade de Tecnologia

DISSERTAÇÃO DE MESTRADO PROFISSIONAL

ChatORION: FRAMEWORK INTELIGENTE PARA AUTOMAÇÃO CONVERSACIONAL E TRIAGEM DE PROCESSOS NO CONTENCIOSO JUDICIAL COM UTILIZAÇÃO DE IA GENERATIVA

Elon Oliveira Albuquerque

Dissertação de Mestrado Profissional submetida ao Departamento de Engenharia Elétrica como requisito parcial para obtenção do grau de Mestre em Engenharia Elétrica

Banca Examinadora

Prof. Dr. Fabio Lucio Lopes de Mendonça,
PPEE/FT/ENE/UnB
Presidente - Orientador

Profa.Dra. Edna Dias Canedo,
PPEE/CIC/UnB
Examinadora Interna

Prof. Dr. Gilmar dos Santos Marques,
UPIS - União Pioneira de Integração Social
Examinador Externo

Prof. Dr. Georges Daniel Amvame Nze,
PPEE/FT/ENE/UnB
Membro Suplente

FICHA CATALOGRÁFICA

ALBUQUERQUE, ELON OLIVEIRA
ChatORION: FRAMEWORK INTELIGENTE PARA AUTOMAÇÃO CONVERSACIONAL E TRIAGEM DE PROCESSOS NO CONTENCIOSO JUDICIAL COM UTILIZAÇÃO DE IA GENERATIVA [Distrito Federal] 2026.
xvi, 52 p., 210 x 297 mm (ENE/FT/UnB, Mestre, Engenharia Elétrica, 2026).
Dissertação de Mestrado Profissional - Universidade de Brasília, Faculdade de Tecnologia.
Departamento de Engenharia Elétrica

1. Chatbot
2. Processos Judiciais
3. Aprendizado de máquina
4. Automatização de Tarefas
5. Inteligência Artificial

I. ENE/FT/UnB
II.ChatORION

REFERÊNCIA BIBLIOGRÁFICA

ALBUQUERQUE, E.O. (2026). *ChatORION: FRAMEWORK INTELIGENTE PARA AUTOMAÇÃO CONVERSACIONAL E TRIAGEM DE PROCESSOS NO CONTENCIOSO JUDICIAL COM UTILIZAÇÃO DE IA GENERATIVA*. Dissertação de Mestrado Profissional, Publicação: PPEE.MP.110, Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, 52 p.

CESSÃO DE DIREITOS

AUTOR: Elon Oliveira Albuquerque
TÍTULO: ChatORION: FRAMEWORK INTELIGENTE PARA AUTOMAÇÃO CONVERSACIONAL E TRIAGEM DE PROCESSOS NO CONTENCIOSO JUDICIAL COM UTILIZAÇÃO DE IA GENERATIVA.
GRAU: Mestre em Engenharia Elétrica ANO: 2026

Elon Oliveira Albuquerque
Depto. de Engenharia Elétrica (ENE) - FT
Universidade de Brasília (UnB)
Campus Darcy Ribeiro
CEP 70919-970 - Brasília - DF - Brasil

DEDICATÓRIA

Dedico aos meus pais, Maria Feliciana de Oliveira Albuquerque e Eloi Albuquerque da Silva, por sempre acreditar na minha capacidade.

Dedico este trabalho aos meus filhos, Miguel Fonseca de Albuquerque, Javi Fonseca de Albuquerque e a minha esposa Thais Fonseca Pirangi Soares, Pelo impulsionamento para mais essa conquista em minha vida e também ao tempo que deixei de dedicar a nossa família para intensificar pesquisas e estudos para o sucesso da dissertação do tema.

Aos meus irmãos Robson de Oliveira Albuquerque e Amauri de Oliveira Albuquerque pelo apoio e incentivo que foram dados durante todo o tempo em que estive envolvido neste trabalho.

AGRADECIMENTOS

Agradeço em especial ao meu orientador, professor Dr. Fábio Lúcio Lopes de Mendonça, que me orientou de forma profissional e amiga nas horas mais complicadas durante este trabalho e aturou tantas dúvidas e problemas relativos ao assunto e outros detalhes pertinentes à criação desta dissertação, dedicando do seu tempo para o sucesso deste trabalho.

Agradeço imensamente a Professora Edna Dias Canedo e ao Professor Geraldo Pereira Rocha Filho que sempre foram grandes parceiros nas horas mais complicadas durante este trabalho.

Aos demais Professores do Programa de Pós-Graduação Profissional em Engenharia Elétrica da Universidade de Brasília PPEE/UNB, Rafael Timóteo de Sousa Júnior, Georges Daniel Amvame Nze, Daniel Alves da Silva, Robson de Oliveira Albuquerque e William Ferreira Giozza e João José Costa Gondim, e ao membro da Banca Gilmar dos Santos Marques pelas grandes dicas, constante apoio, incentivo e amizade, essenciais para o desenvolvimento deste trabalho.

Agradeço o apoio técnico e computacional do Laboratório de Tecnologias para Tomada de Decisão - LATITUDE, da Universidade de Brasília, que conta com apoio do CNPq - Conselho Nacional de Pesquisa (Outorgas 312180/2019-5 PQ-2 e 465741/2014-2 INCT em Cibersegurança),ao TED 01/2021 da Secretaria Nacional de Assistência Social – SNAS/DGSUAS/CGRS, ao TED 01/2021 da Coordenação-Geral de Tecnologia da Informação (CGTI) da Procuradoria Geral da Fazenda Nacional – PGFN, ao Projeto SISTER City – Sistemas Inteligentes Seguros e em Tempo Efetivo Real para Cidades Inteligentes (Outorga 625/2022), ao Projeto “Sistema de Controle e Unificação de Projetos para o Governo Distrito Federal – SISPRO-DF” (Outorga 497/2023), ao Projeto “Pesquisa, Desenvolvimento e Aplicação – Metodologia para Apoiar a Elicitação de Requisitos Éticos e de Privacidade” (Outorga 514/2023), ao Decanato de Pesquisa e Inovação – DPI/UnB e a Fundação de Apoio a Pesquisa do Distrito Federal - FAP/DF.

A meus amigos(as) e parceiros que me ajudaram durante essa jornada, Flávio Praciano, Jorge Osvaldo de Lima Torres, Kelly Santos, Paulo Henrrique, Philipe Alan Almeida e Thiago que contribuíram de forma fundamental para a conclusão deste trabalho: meus sinceros agradecimentos.

Agradeço, acima de tudo, a Deus!

RESUMO

A crescente complexidade da atuação jurídica institucional exige soluções tecnológicas capazes de integrar automação, comunicação eficiente e suporte inteligente à tomada de decisão. Nesse contexto, a Procuradoria Geral da Fazenda Nacional (PGFN) opera em um ambiente caracterizado por grande volume de processos, documentos normativos, pareceres e dados fiscais, no qual a recuperação eficiente da informação torna se elemento crítico para a qualidade da atuação jurídica e para a coerência institucional. Esta dissertação propõe o ChatORION (**Chat** **O**timizador de **R**otinas **I**nteligentes para **O**perações no Contencioso **N**acional), uma solução de Inteligência Artificial Generativa baseada na arquitetura Retrieval Augmented Generation. O sistema combina mecanismos de busca semântica por similaridade vetorial com geração textual condicionada por modelos de linguagem de grande porte, produzindo respostas fundamentadas em documentos institucionais. O objetivo é ampliar a eficiência informacional ao reduzir ambiguidades e inconsistências factuais típicas de modelos puramente generativos. Os resultados demonstram a existência de um regime ótimo de operação capaz de equilibrar qualidade informacional e custo computacional, evidenciando que o desempenho depende da interação entre recuperação estruturada e geração de linguagem, e não apenas da capacidade do modelo. Os experimentos indicam que o ChatORION melhora a acurácia factual, reduz a latência média de resposta e aumenta o nível de detalhamento quando comparado a modelos generativos isolados. Tais resultados sugerem potencial para redução do tempo dedicado à triagem processual, localização de fundamentos jurídicos e análise documental, permitindo que os procuradores concentrem esforços em atividades de maior complexidade.

Palavras Chaves - Chatbot; Retrieval Augmented Generation (RAG); Aprendizado de máquina; Automatização de Tarefas; Contencioso Judicial ou Direito Público; Inteligência Artificial; Large Language Models (LLMs)

ABSTRACT

The increasing complexity of institutional legal practice requires technological solutions capable of integrating automation, efficient communication, and intelligent decision support. In this context, the Office of the General Counsel for the National Treasury (PGFN) operates in an environment characterized by a large volume of cases, normative documents, legal opinions, and fiscal data, in which efficient information retrieval becomes a critical element for the quality of legal action and institutional coherence. This dissertation proposes ChatORION (Chat Optimizer for Intelligent Routines in Operations within the National Litigation domain), a Generative Artificial Intelligence solution based on the Retrieval Augmented Generation architecture. The system combines semantic search mechanisms based on vector similarity with text generation conditioned by large language models, producing responses grounded in institutional documents. The objective is to improve informational efficiency by reducing ambiguities and factual inconsistencies typical of purely generative models. The results demonstrate the existence of an optimal operational regime capable of balancing informational quality and computational cost, showing that performance depends on the interaction between structured retrieval and language generation rather than solely on the model capacity. Experiments indicate that ChatORION improves factual accuracy, reduces average response latency, and increases the level of detail when compared to standalone generative models. These findings suggest potential for reducing the time dedicated to case screening, identification of legal grounds, and document analysis, allowing legal practitioners to focus on more complex activities.

Keywords - Chatbot; Retrieval Augmented Generation (RAG); Machine Learning; Task Automation; Artificial Intelligence; Large Language Models (LLMs)

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIAÇÕES E SIGLAS

Lista de abreviações Siglas

PGFN	Procuradoria Geral da Fazezda Nacional
ML	Machine Learning
BI	Business Intelligence
IA	Inteligência Artificial
RMSE	Root Mean Squared Error
sMAPE	Symmetric Mean Absolute Percentage
ARIMA	Autorregressivos Integrados de Médias Móveis
SARIMA	Modelo Autorregressivo Integrado de Média Móvel Sazonal
SVM	Support Vector Machines
LLMs	Large Language Models
PLN	Processamento de Linguagem Natural
XAI	inteligência artificial explicável
DW	Data Warehouses
SGBD	Gerenciamento de banco de dado
MER	Modelo Entidade–Relacionamento
MLOps	Operações de Machine Learning
SLAs	Service Level Agreement (Acordo de Nível de Serviço)
APIs	Interfaces de Programação de Aplicações
ELT	Extract, Load, Transform (processo de integração de dados)

1 INTRODUÇÃO

O cenário organizacional contemporâneo, marcado por alta complexidade, volatilidade e competitividade global, impulsiona a necessidade de soluções tecnológicas capazes de apoiar decisões ágeis, precisas e estrategicamente embasadas. A transformação digital da administração pública e do setor jurídico, em particular, intensificou o volume de dados, a complexidade normativa e a pressão por maior transparência, eficiência e accountability na gestão de processos [1, 2]. Nesse contexto, a inteligência artificial (IA) tem se consolidado como um campo dinâmico e orientado a dados, oferecendo ferramentas inteligentes que ampliam a capacidade analítica dos gestores e promovem maior eficiência operacional.

A convergência entre IA e automação emerge como um catalisador dessa transformação, oferecendo novos caminhos para o aprimoramento da tomada de decisão, para a otimização de fluxos de trabalho e para a orquestração de atividades em larga escala [3]. Técnicas de aprendizado de máquina (*Machine Learning* ML), análise preditiva e processamento de linguagem natural (PLN) permitem antecipar tendências, identificar riscos latentes e recomendar ações estratégicas, conferindo aos gestores maior capacidade de resposta diante de cenários incertos. Paralelamente, a automação de tarefas repetitivas reduz desperdícios, elimina retrabalhos e libera equipes para atividades de maior valor agregado, fortalecendo a maturidade dos processos de negócio e de projeto.

No domínio dos serviços digitais, agentes conversacionais e chatbots têm sido empregados para apoiar atendimento ao cidadão, triagem de demandas e mediação de informações em setores como saúde, educação, governo eletrônico e serviços financeiros [4, 5, 6]. Estudos recentes mostram que, quando bem projetados, esses sistemas podem aumentar a acessibilidade, reduzir tempos de resposta e melhorar a experiência do usuário [7, 8]. Entretanto, a literatura também evidencia limitações importantes, como dificuldades em lidar com domínios altamente especializados (por exemplo, o jurídico), desafios de qualidade da informação, riscos de vieses e problemas de confiança dos usuários [9, 10].

Na administração pública, chatbots e soluções de IA têm sido utilizados majoritariamente em frentes de atendimento ao cidadão, como suporte informacional, esclarecimento de dúvidas frequentes e encaminhamento a serviços [11, 12]. Ainda que esses avanços representem um passo importante na digitalização do Estado, observa-se um uso menos sistemático da IA em tarefas de *backoffice*, tais como triagem de processos, análise quantitativa, classificação automática e organização de grandes massas documentais, especialmente em órgãos voltados ao contencioso e à cobrança da dívida ativa. A ausência de frameworks arquiteturais robustos e avaliados empiricamente para automatizar essas rotinas jurídicas complexas representa uma lacuna tanto prática quanto científica.

Ao mesmo tempo, o avanço recente de modelos de linguagem de grande porte (*Large*

Language Models - LLMs) e de frameworks de agentes de IA, como o *LangGraph*, abre novas possibilidades para o desenvolvimento de sistemas capazes de combinar raciocínio simbólico, consulta a bases estruturadas e interação conversacional de forma integrada. Esses agentes podem orquestrar fluxos de trabalho, chamar ferramentas externas, recuperar conhecimento em bases vetoriais e manter estado ao longo da interação, aproximando-se de um suporte cognitivo mais sofisticado às equipes jurídicas [10]. Técnicas de *Retrieval-Augmented Generation* (RAG) e o uso de bancos de dados vetoriais, como o Qdrant, permitem integrar decisões baseadas em evidências com documentos e processos reais, reduzindo o risco de alucinações e melhorando a rastreabilidade das respostas.

No contexto jurídico brasileiro, a Procuradoria-Geral da Fazenda Nacional (PGFN) destaca-se pela responsabilidade de gerenciar e cobrar a Dívida Ativa da União, conduzindo milhares de processos judiciais e administrativos diariamente. A combinação entre grande volume de casos, complexidade normativa e recursos humanos limitados cria um cenário em que a leitura, triagem e análise quantitativa dos processos se tornam gargalos críticos para a eficiência institucional [13, 14, 15]. Ferramentas tradicionais de busca, baseadas apenas em palavras-chave, frequentemente falham em capturar nuances jurídicas, relacionar informações dispersas e fornecer uma visão sintética que apoie decisões estratégicas em tempo hábil.

Diante desse cenário, emerge a necessidade de soluções arquiteturais que integrem agentes de IA, PLN, modelos de linguagem e mecanismos de recuperação de informação em um framework coeso, auditável e aderente às especificidades do contencioso público. Diferentemente de abordagens genéricas de automação, esse tipo de solução deve conciliar requisitos técnicos (desempenho, escalabilidade, interoperabilidade), jurídicos (segurança jurídica, conformidade normativa), organizacionais (aderência aos fluxos de trabalho) e de governança (transparência, explicabilidade e responsabilidade na tomada de decisão automatizada).

Assim, este trabalho investiga como soluções inteligentes, em especial o framework proposto ChatORION (**Chat O**timizador de **R**otinas **I**nteligentes para **O**perações no Contencioso **N**acional), podem potencializar a automatização de processos judiciais ao integrar automação conversacional com suporte analítico e interação humanizada, atendendo às demandas de um ambiente corporativo que exige simultaneamente velocidade, precisão e comunicação de alta qualidade. De forma específica, esta dissertação propõe, projeta, implementa e avalia um framework arquitetural de automação conversacional e triagem de processos judiciais baseado em IA, adotando como prova de conceito o contexto operacional da PGFN. Na seção seguinte, apresentam-se a contextualização detalhada do órgão e o problema de pesquisa que orienta este estudo.

1.1 CONTEXTUALIZAÇÃO E PROBLEMA DE PESQUISA

A Procuradoria-Geral da Fazenda Nacional (PGFN) tem como missão a promoção da justiça fiscal, a garantia da segurança jurídica das políticas públicas e a defesa dos interesses financeiros da União. O órgão é responsável por viabilizar a arrecadação de recursos financeiros cruciais para a implementação de políticas públicas em diversas áreas do Poder Executivo, como saúde, educação e infraestrutura. Sua atuação, portanto, possui impacto social e econômico significativo, assegurando o funcionamento do Estado e contribuindo para a sustentabilidade fiscal do país [13, 15].

Um dos pilares centrais da missão da PGFN é a gestão e a cobrança da Dívida Ativa da União. Isso envolve a apuração da liquidez, certeza e legalidade dos créditos, sejam eles tributários (impostos, taxas, contribuições) ou de outra natureza (como multas federais), bem como a subsequente inscrição desses débitos em dívida ativa. Após a inscrição, a PGFN promove a cobrança, que pode ser realizada de forma amigável (extrajudicial) ou por meio de processos judiciais (execução fiscal). Nesse fluxo, tramitam diariamente centenas ou milhares de processos, distribuídos em diferentes tribunais e unidades da federação, o que torna a análise detalhada por parte dos procuradores um desafio constante, sobretudo diante de um efetivo reduzido em comparação ao volume de demandas [14].

Em suma, a missão da PGFN é complexa e multifacetada, atuando como advocacia pública da União em matéria fiscal e financeira. O órgão busca garantir o equilíbrio das contas públicas e a sustentabilidade fiscal do país, combatendo a inadimplência e a evasão fiscal, além de recuperar valores que, de outra forma, seriam perdidos aos cofres públicos. Estima-se que, para cada real investido no órgão, gera-se um retorno financeiro substancial em arrecadação e recuperação de créditos [15]. No entanto, essa capacidade de retorno depende diretamente da eficiência com que os processos são recebidos, lidos, triados, classificados e monitorados ao longo de seu ciclo de vida.

Atualmente, grande parte dessas atividades ainda é realizada de forma manual ou com apoio de sistemas que não exploram plenamente o potencial da IA, o que implica em tempos elevados de leitura inicial, dificuldades de padronização da classificação, uso limitado de análises quantitativas e pouca integração entre fontes de dados heterogêneas. Ferramentas de busca convencionais, baseadas apenas em palavras-chave, muitas vezes falham em capturar nuances jurídicas, expressões idiomáticas e contextos específicos dos autos, além de não proverem uma visão analítica consolidada capaz de apoiar decisões estratégicas em escala.

A literatura internacional apresenta experiências de uso de chatbots, agentes conversacionais e modelos de linguagem em diferentes domínios, tais como saúde, educação e serviços públicos digitais com resultados promissores em termos de automação de rotinas e apoio à decisão [7, 5, 12]. Todavia, tais iniciativas concentram-se majoritariamente em cenários de atendimento ao usuário final ou em domínios distintos, não abordando de forma sistemática a triagem e análise quantitativa de processos judiciais em órgãos de contencioso público. Além disso,

poucos trabalhos tratam da integração entre agentes de IA complexos, arquitetura RAG e interoperabilidade de dados jurídicos distribuídos entre diferentes tribunais e unidades federativas.

Diante desse contexto, o problema de pesquisa que se coloca é: **como desenvolver um framework de automação baseado em inteligência artificial capaz de realizar leitura, triagem e análise quantitativa de processos judiciais de forma eficiente, confiável e aplicável ao contexto operacional da PGFN, respeitando seus requisitos jurídicos, organizacionais e de governança de dados?**

Como hipótese de trabalho, assume-se que um framework arquitetural que combine agentes de IA complexos, técnicas de PLN, bases vetoriais especializadas e mecanismos de interoperabilidade de dados, estruturado segundo boas práticas de engenharia de software e alinhado às necessidades do contencioso da PGFN, é capaz de reduzir significativamente o esforço manual de triagem, aumentar a padronização das análises e fornecer insumos quantitativos relevantes para a tomada de decisão estratégica.

1.2 JUSTIFICATIVA

A PGFN figura entre as instituições com maior volume de atuação judicial no país, sendo responsável por milhares de processos distribuídos diariamente. A execução manual das etapas de leitura inicial, classificação e análise quantitativa consome tempo valioso de procuradores e servidores, impactando diretamente a celeridade das ações estratégicas e a capacidade de resposta institucional. Em um contexto de restrições orçamentárias e limitação de quadros, a incapacidade de lidar eficientemente com o volume crescente de demandas pode comprometer a efetividade das políticas de recuperação de crédito, bem como a própria sustentabilidade fiscal do Estado.

A adoção de um framework de automação inteligente não apenas reduz a carga operacional repetitiva, mas também permite maior padronização, assertividade e transparência nos fluxos internos. Ao estruturar regras, modelos e fluxos de decisão de maneira explícita, torna-se possível auditar, explicar e aprimorar continuamente os critérios de triagem e análise, alinhando-se às exigências de governança, segurança jurídica e responsabilidade na atuação da advocacia pública. Adicionalmente, o uso de arquiteturas baseadas em RAG e bancos de dados vetoriais possibilita recuperar informações relevantes diretamente dos autos e de bases oficiais, reduzindo dependência de indexações frágeis e mitigando riscos de omissão de elementos importantes. Embora o ChatORION se foque em processos judiciais, a PGFN lida com dados orçamentários complexos, e um exemplo de blueprint de Inteligência de Negócio (do termo em inglês Business Inteligence ou BI), como detalhado no Apêndice A, ilustra a infraestrutura de dados subjacente que pode ser explorada por futuras expansões ou contextos de uso do ChatORION.

Sob a perspectiva científica, a proposta desta pesquisa dialoga com linhas de investigação em engenharia de software, sistemas de informação e IA aplicada ao setor público, especialmente no cruzamento entre agentes conversacionais, automação de processos e análise de grandes volumes

de dados jurídicos. Embora existam estudos que classificam e avaliam chatbots em diferentes domínios [4, 7], há escassez de modelos arquiteturais específicos para o contencioso público brasileiro, envolvendo interoperabilidade com diferentes tribunais, conformidade normativa e uso de LLMs em contextos de alto risco regulatório.

Sob a ótica das políticas públicas e da transformação digital do Estado brasileiro, a criação do **ChatORION** alinha-se às diretrizes de modernização da administração pública, ao princípio constitucional da eficiência e à necessidade de ampliar o uso estratégico de dados e IA no ciclo de políticas fiscais. A proposta contribui para a agenda de governo digital ao oferecer uma solução que combina automação conversacional, triagem inteligente e suporte analítico em um domínio crítico para a sustentabilidade das finanças públicas.

Assim, a pesquisa se justifica por seu potencial impacto institucional ao apoiar a PGFN na gestão mais eficiente do contencioso, por sua relevância tecnológica ao propor e avaliar um framework de agentes de IA em um domínio jurídico complexo e por sua contribuição teórica ao avançar o estado da arte em automação jurídica aplicada ao setor público.

1.3 OBJETIVOS

1.3.1 Objetivo Geral

O objetivo geral desta dissertação é propor, implementar e avaliar um framework arquitetural inteligente de automação conversacional e triagem de processos judiciais, denominado ChatORION, capaz de integrar modelos de Processamento de Linguagem Natural, técnicas avançadas de análise e mecanismos de interoperabilidade de dados, com o propósito de aprimorar a eficiência, a padronização e a agilidade das rotinas do contencioso judicial, aplicando-o como prova de conceito no ambiente operacional da PGFN.

1.3.2 Objetivos Específicos

Para alcançar o objetivo geral, foram definidos os seguintes objetivos específicos:

- Mapear e analisar criticamente os fluxos operacionais atuais de leitura, triagem e classificação de processos no contencioso judicial da PGFN, identificando gargalos, ineficiências e oportunidades de automação;
- Definir requisitos funcionais e não funcionais para o desenvolvimento de um framework de automação conversacional baseado em IA, considerando aspectos técnicos, jurídicos, operacionais, de governança e de conformidade normativa;
- Projetar a arquitetura do framework ChatORION, estruturando suas camadas de captura de dados, enriquecimento, modelos de linguagem, mecanismos de triagem automatizada e

interface conversacional;

- Implementar uma Prova de Conceito (PoC) que integre bases vetoriais, modelos de linguagem e pipelines de decisão para leitura, extração, classificação e sistematização de informações processuais;
- Avaliar o desempenho do ChatORION por meio de métricas quantitativas (precisão, *recall*, tempo de triagem, redução do esforço manual) e análises qualitativas (usabilidade, satisfação dos operadores e aderência às rotinas jurídicas);
- Validar a aplicabilidade e robustez do framework em cenários reais do contencioso judicial, demonstrando seu impacto na melhoria da eficiência operacional e sua capacidade de padronizar análises em larga escala;
- Elaborar recomendações e diretrizes para a evolução e escalabilidade do framework em outros contextos da administração pública e do sistema de justiça.

1.4 PUBLICAÇÕES RELACIONADAS A ESTA DISSERTAÇÃO

Como resultado parcial e correlato desta dissertação, foram publicados os seguintes artigos:

- Elon Oliveira Albuquerque, Bruno Justino Garcia Praciano, Paulo Henrique Batista Rodrigues, Flávio Garcia Praciano, Geraldo P. Rocha Filho e Fábio Lúcio Lopes de Mendonça. **CATCH: A Nova Fronteira dos Chatbots na Gestão de Força de Trabalho**. Revista Ibérica de Sistemas e Tecnologias de Informação (RISTI), Edição N.º E77, 08/2025, pp. 146–159. ISSN: 1646-9895.
- Elon Oliveira Albuquerque, Wesley Gongora de Almeida, Bruno Justino Garcia Praciano, Márcio Bastos de Medeiros, Robson de Oliveira Albuquerque e Fábio Lúcio Lopes de Mendonça. **Data Pipelines Implementation and Management for Data Engineering: A Case Study Applied to the Public Sector**. Revista Ibérica de Sistemas e Tecnologias de Informação (RISTI), Edição N.º E77, 08/2025, pp. 146–159. ISSN: 1646-9895.
- Fábio Lúcio Lopes de Mendonça, Bruno Justino Garcia Praciano, Flávio Garcia Praciano, Thiago Leite de Sousa, José Péricles Pereira de Sousa, Elon Oliveira Albuquerque. **Juris Syntax: Automação da Análise Jurídica Brasileira com IA Generativa e PLN**. Revista Ibérica de Sistemas e Tecnologias de Informação (RISTI), DOI: 10.17013/risti.n.pi-pf. Disponível em: <https://www.risti.xyz/issues/ristie77.pdf>.

Todos os trabalhos acima dialogam diretamente com o tema desta dissertação, seja ao explorar o uso de chatbots na gestão de força de trabalho, seja ao abordar a implementação de data pipelines no setor público ou a automação da análise jurídica com IA generativa, contribuindo para a consolidação da linha de pesquisa em automação inteligente de processos no domínio jurídico e governamental.

1.5 METODOLOGIA DE PESQUISA

A metodologia adotada nesta pesquisa segue um delineamento de natureza aplicada, com abordagem qualitativa–quantitativa e caráter exploratório-explicativo, alinhando-se ao objetivo de propor e validar um framework de automação inteligente para rotinas do contencioso judicial. De forma geral, o percurso metodológico foi organizado em quatro etapas principais, ilustradas esquematicamente na Figura 1.1:

1. **Revisão de literatura**: levantamento sistemático e narrativo sobre processos judiciais e interoperabilidade de dados, modelos de arquitetura aplicados ao trâmite de processos, uso de inteligência artificial e ferramentas de agentes de IA complexos no setor público. Essa etapa segue orientações metodológicas clássicas de pesquisa bibliográfica [16], permitindo identificar o estado da arte, lacunas e oportunidades de aplicação no contexto da PGFN.

2. **Diagnóstico institucional**: realização de análise documental dos fluxos internos da PGFN, observação de processos de trabalho e entrevistas semiestruturadas com procuradores e servidores, com o objetivo de compreender o estado atual das rotinas de leitura, triagem e análise quantitativa; mapear gargalos operacionais; e identificar requisitos técnicos, funcionais e de governança que orientarão o desenho do framework ChatORION.

3. **Construção conceitual e desenvolvimento do framework ChatORION**: definição das camadas da arquitetura (entrada, enriquecimento, IA, interação e governança), seleção das tecnologias (LLMs, OCR, vetorização, RAG, bancos vetoriais), elaboração de diagramas e requisitos técnicos e implementação de uma Prova de Conceito (PoC) em ambiente controlado, utilizando um corpus representativo de processos judiciais anonimizados.

4. **Testes, validação e avaliação**: aplicação de métricas quantitativas (precisão e *recall* da classificação, tempo médio de triagem antes e depois da automação, redução de esforço operacional) e métodos qualitativos (análise de usabilidade, avaliação por especialistas e entrevistas estruturadas). O desempenho do ChatORION é comparado ao fluxo manual tradicional, em delineamento quase-experimental, e são analisados também critérios de governança, auditabilidade e conformidade com a LGPD.

Essa combinação de métodos permite articular fundamentos teóricos consolidados com validação prática, verificando a viabilidade da arquitetura proposta como instrumento de apoio à eficiência, transparência e responsabilidade na gestão de processos judiciais. Como prova de conceito, a aplicação da proposta concentra-se nos dados da Procuradoria-Geral da Fazenda Nacional, possibilitando a experimentação prática da arquitetura e a avaliação de seu potencial de apoio à tomada de decisão no contencioso fiscal.

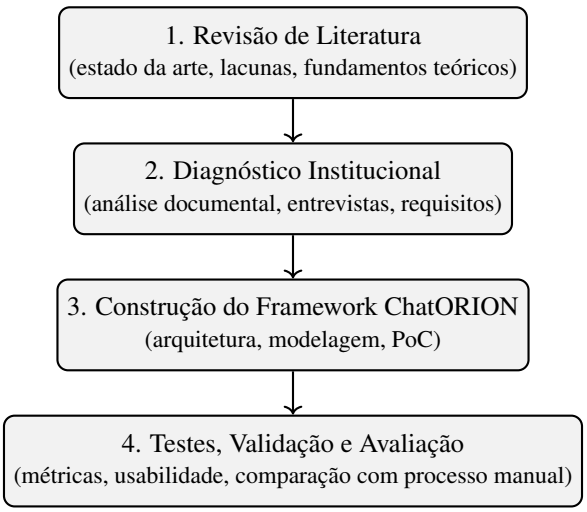


Figura 1.1: Etapas metodológicas adotadas na pesquisa.

1.6 ESTRUTURA DA DISSERTAÇÃO

Esta dissertação está organizada em cinco capítulos, incluindo esta introdução.

O **Capítulo 2** apresenta a fundamentação teórica do estudo, reunindo os principais conceitos, modelos, trabalhos correlatos e referenciais utilizados para embasar a pesquisa, com foco em automação de processos judiciais, agentes de IA, IA generativa, processamento de linguagem natural e arquitetura RAG aplicada ao contexto jurídico.

O **Capítulo 3** descreve a arquitetura do modelo proposto, detalhando as etapas de instalação, configuração e integração dos componentes tecnológicos, bem como o método de verificação adotado para assegurar a consistência dos procedimentos realizados. São apresentados os módulos do ChatORION, suas interfaces, fluxos de dados e mecanismos de governança e segurança da informação.

O **Capítulo 4** apresenta os testes e resultados obtidos a partir da utilização de dados da Procuradoria-Geral da Fazenda Nacional (PGFN). Inicialmente, os dados são minerados e extraídos para uma base vetorial; em seguida, são processados por meio de ferramentas de *Extract, Transform, Load* (ETL) e modelos de linguagem. A análise contempla a aplicação de métricas estatísticas e indicadores de desempenho, com o objetivo de identificar informações relevantes para subsidiar a tomada de decisão gerencial e avaliar o impacto da automação nas rotinas do contencioso.

Por fim, o **Capítulo 5** reúne as conclusões da pesquisa, destacando as principais contribuições do estudo, as limitações identificadas e as perspectivas para trabalhos futuros, incluindo possibilidades de generalização do framework para outros órgãos públicos, integração com novos modelos de linguagem e aprimoramento dos mecanismos de governança e explicabilidade da IA no contexto jurídico.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os fundamentos conceituais e tecnológicos que sustentam o desenvolvimento do *ChatORION*, bem como os principais trabalhos correlatos na interseção entre automação de processos judiciais, aprendizado de máquina, *Large Language Models* (LLMs), agentes de IA e arquiteturas de dados no setor público. Busca-se, assim, estabelecer o pano de fundo teórico que orienta as decisões de projeto, as opções metodológicas e os critérios de avaliação adotados nesta dissertação, além de fornecer bases para a replicabilidade e evolução futura da proposta.

São discutidos inicialmente o trâmite de processos judiciais na Procuradoria-Geral da Fazenda Nacional (PGFN) e a natureza das rotinas do contencioso, ressaltando os desafios de escala, complexidade e governança. Em seguida, apresentam-se conceitos de aprendizado de máquina e análise preditiva, modelos de séries temporais, LLMs e arquiteturas de *Data Warehouse* e governança de dados. Por fim, discute-se a literatura relacionada sobre chatbots, agentes conversacionais e soluções de IA no setor público e no domínio jurídico, destacando as lacunas que motivam o desenvolvimento do *ChatORION*.

2.1 TRÂMITE DE PROCESSOS DA PGFN NO CONTENCIOSO

No âmbito do contencioso judicial, o trâmite de processos na Procuradoria-Geral da Fazenda Nacional (PGFN) inicia-se, em geral, com a recepção de novas ações judiciais distribuídas contra a União ou envolvendo a cobrança da Dívida Ativa da União. Esses processos chegam predominantemente por meios eletrônicos, via integração com tribunais e sistemas de processo judicial eletrônico, e incluem petições iniciais, decisões liminares, despachos, manifestações da parte contrária e documentos anexos. A equipe responsável realiza uma leitura inicial dos autos, identificando o tipo de demanda, os assuntos envolvidos, o valor em discussão, a urgência e os possíveis impactos fiscais e estratégicos [17, 13].

Essa triagem preliminar é fundamental para organizar prioridades, detectar demandas sensíveis, agrupar processos com teses semelhantes e determinar o encaminhamento adequado dentro da estrutura organizacional da PGFN. A partir dessa etapa, os processos são distribuídos às unidades e procuradores responsáveis, de acordo com critérios como matéria (tributária, previdenciária, fiscal, administrativa), grau de complexidade, relevância estratégica, risco fiscal e especialização das equipes. O procurador designado realiza a análise aprofundada do caso, identifica fundamentos jurídicos aplicáveis, pesquisa precedentes judiciais e administrativos, avalia riscos e elabora manifestações como contestações, recursos, petições, pareceres e memoriais.

Ao longo desse ciclo, há intensa interação com sistemas internos de acompanhamento processual, bancos de dados de precedentes, orientações normativas, pareceres referenciais e teses institucionais. O trabalho envolve tanto a leitura e interpretação de grandes volumes de texto quanto a necessidade de sínteses quantitativas (por exemplo, número de processos por tema, valor agregado em discussão, adesão a teses repetitivas), que subsidiam decisões de priorização e políticas de transação e acordo. Em etapas subsequentes, o processo segue para acompanhamento, monitoramento de prazos, análise de decisões judiciais posteriores, interposição de recursos e eventual execução fiscal ou cumprimento de sentença.

A PGFN utiliza sistemas como o Sapiens/PGFN para registrar atos processuais, controlar indicadores, monitorar a carteira e assegurar governança sobre o contencioso. Ainda assim, trata-se de um ciclo complexo, repetitivo e volumoso, no qual a triagem, a classificação e a organização de informações estruturadas a partir dos autos consomem tempo expressivo de procuradores e servidores. A sobreposição entre alto volume, complexidade normativa e recursos humanos limitados cria um cenário propício para a adoção de soluções de automação baseada em IA, como o *ChatORION*, com potencial para otimizar etapas críticas do processo, reduzir esforço manual e ampliar a qualidade e a padronização das análises.

2.2 APRENDIZADO DE MÁQUINA

O avanço das tecnologias digitais e o crescente volume de dados disponíveis têm impulsionado o uso de técnicas de Aprendizado de Máquina (ML) como ferramentas centrais na análise preditiva aplicada à gestão pública e privada. O ML, subárea da Inteligência Artificial (IA), busca desenvolver algoritmos capazes de aprender padrões a partir de dados históricos e, com isso, realizar previsões, classificações ou recomendações em novos cenários [18]. Em contextos intensivos em documentos, como o contencioso judicial, essa capacidade é relevante tanto para triagem de processos quanto para priorização de casos, identificação de padrões de litigância e apoio à tomada de decisão.

O aprendizado de máquina pode ser definido como o processo pelo qual sistemas computacionais adquirem conhecimento de forma autônoma a partir de dados, sem necessidade de programação explícita para cada tarefa. Essa capacidade está estruturada em diferentes paradigmas, entre os quais se destacam:

- **Aprendizado supervisionado**: o modelo é treinado com dados rotulados, em que as entradas (por exemplo, texto de uma petição ou metadados de um processo) estão associadas a saídas conhecidas (como tipo de ação, tema jurídico ou probabilidade de procedência). É amplamente utilizado em tarefas de regressão e classificação.
- **Aprendizado não supervisionado**: algoritmos buscam identificar padrões e estruturas ocultas em dados não rotulados, como agrupamento de processos com características similares ou segmentação de classes de litígios.

- **Aprendizado por reforço**: sistemas aprendem por meio de interação com o ambiente, recebendo recompensas ou penalidades conforme as ações tomadas. Embora menos explorado no contexto jurídico, apresenta potencial em problemas de alocação dinâmica de recursos, priorização de filas e estratégias de negociação.

A escolha do paradigma depende dos objetivos da análise, da estrutura dos dados e das restrições de interpretabilidade. Em cenários como o da PGFN, o aprendizado supervisionado e o não supervisionado são os mais promissores para apoiar a leitura, triagem e classificação de processos judiciais, bem como para identificar padrões em carteiras de crédito e litígios.

2.2.1 Análise Preditiva e Aplicações em Contextos de Fiscalização e Contencioso

A análise preditiva corresponde ao uso de técnicas estatísticas, algoritmos de aprendizado de máquina e mineração de dados para estimar probabilidades de ocorrências futuras com base em informações históricas [19]. Diferentemente da análise descritiva, que apenas resume dados passados, e da análise diagnóstica, que busca explicar causas, a análise preditiva tem caráter prospectivo, orientando a tomada de decisão em cenários de incerteza.

Na área de auditoria e controle, Pereira et al. [20] propuseram uma ferramenta de código aberto para automatizar o reconhecimento de notas fiscais em processos de auditoria de gastos públicos, combinando OCR (Tesseract) com modelos de classificação baseados em BERT e algoritmos supervisionados. O sistema foi capaz de distinguir entre imagens de notas fiscais e documentos não fiscais com acurácia superior a 90%, evidenciando a viabilidade de automação de tarefas repetitivas e a possibilidade de liberar analistas para atividades mais complexas. Embora o estudo se concentre em documentos fiscais, a abordagem é análoga à classificação de peças processuais em autos judiciais.

Na detecção de fraudes, Santos et al. [21] desenvolveram uma arquitetura em tempo real para identificar transações suspeitas em cartões de crédito utilizando algoritmos supervisionados, como *Random Forest*. O estudo alcançou métricas de acurácia próximas a 99,98% e revocação de 100%, demonstrando a robustez do ML em cenários de alto risco e complexidade. A lógica de detecção de padrões anômalos é aplicável à identificação de comportamentos atípicos em carteiras de processos, como litigância predatória ou demandas repetitivas.

Dentre os algoritmos mais frequentemente empregados na análise preditiva que se conectam ao contexto deste trabalho, destacam-se:

- **Árvores de decisão e *Random Forest***: interpretáveis e eficientes para dados estruturados, úteis na identificação de combinações de atributos que explicam padrões de classificação;
- **Modelos de *Gradient Boosting*** (XGBoost, LightGBM): algoritmos de *ensemble* com alta acurácia em tarefas de classificação e regressão;
- **Modelos baseados em redes neurais**: adequados para grandes volumes de dados e tarefas

mais complexas, especialmente quando integrados a representações distribuídas (como *embeddings* de textos).

Esses algoritmos podem ser combinados com representações textuais derivadas de modelos de linguagem (por exemplo, BERT, XLM-RoBERTa), permitindo que textos jurídicos sejam convertidos em vetores de características e utilizados em fluxos de classificação, recomendação ou priorização.

2.3 GERAÇÃO AUMENTADA POR RECUPERAÇÃO (RETRIEVAL AUGMENTED GENERATION (RAG))

As aplicações da arquitetura Retrieval-Augmented Generation (RAG) em ambientes jurídicos surgem como resposta direta às limitações dos modelos puramente generativos quando utilizados em domínios que exigem elevada precisão, rastreabilidade e aderência normativa. Em contextos legais, nos quais decisões e pareceres devem estar estritamente fundamentados em leis, regulamentos, precedentes e documentos oficiais, a capacidade de um sistema recuperar informações confiáveis antes de gerar respostas torna-se essencial. A arquitetura RAG combina mecanismos de recuperação semântica, geralmente apoiados em bases vetoriais, com modelos de linguagem, permitindo que o conteúdo gerado seja ancorado em fontes jurídicas previamente validadas, reduzindo significativamente riscos de interpretações incorretas ou respostas não fundamentadas [22].

Em termos práticos, sistemas baseados em RAG têm sido aplicados em tarefas como pesquisa jurídica avançada, análise e síntese de jurisprudência, identificação de precedentes relevantes e apoio à elaboração de peças e pareceres. Diferentemente das buscas tradicionais por palavras-chave, a recuperação semântica possibilita localizar documentos com base no significado e no contexto jurídico, mesmo quando não há correspondência literal entre termos. Isso é particularmente relevante em ambientes jurídicos complexos, nos quais conceitos equivalentes podem ser expressos de formas distintas ao longo do tempo ou em diferentes instâncias. Dessa forma, o RAG amplia a eficiência e a qualidade da recuperação da informação, oferecendo respostas contextualizadas e acompanhadas de referências explícitas às fontes utilizadas.

Além disso, a adoção de RAG em ambientes jurídicos contribui para o fortalecimento da governança da informação e da segurança jurídica. Ao permitir que cada resposta seja vinculada a trechos específicos de documentos normativos ou institucionais, esses sistemas favorecem a auditabilidade, a transparência e a supervisão humana — requisitos fundamentais para o uso responsável de Inteligência Artificial no Direito. Em instituições públicas e privadas, como tribunais, escritórios de advocacia e procuradorias, o RAG tem se mostrado uma abordagem promissora para automatizar tarefas repetitivas de análise documental, reduzir assimetrias de informação e apoiar decisões estratégicas, sem substituir o juízo técnico e interpretativo do profissional do Direito.

Embora o foco principal desta dissertação esteja na automação da leitura, triagem e análise quantitativa de processos judiciais, o planejamento e a gestão estratégica do contencioso também podem se beneficiar de técnicas de previsão de séries temporais, por exemplo, para estimar a entrada futura de processos, o tempo de tramitação ou a evolução de estoques processuais.

Uma série temporal consiste em um conjunto de observações de uma variável registradas sequencialmente ao longo do tempo, em intervalos regulares, de tal forma que cada observação guarda dependência com valores passados [23]. Os modelos Autorregressivos Integrados de Médias Móveis (ARIMA), popularizados por Box e Jenkins, são amplamente utilizados em previsões de séries temporais não estacionárias. O termo ARIMA (p, d, q) representa a combinação de três componentes: (i) autorregressivo (AR), (ii) diferenciação (I) e (iii) médias móveis (MA) [24].

Em situações em que há padrões sazonais recorrentes, como ciclos anuais de distribuição de processos ou picos de ajuizamento em determinados períodos, adota-se o modelo SARIMA (Seasonal ARIMA), representado por ARIMA $(p, d, q)(P, D, Q)_m$, em que m corresponde ao período da sazonalidade. Esses modelos permitem capturar variações sistemáticas periódicas, melhorando a precisão das previsões em ambientes com sazonalidade marcada.

No contexto do setor público, Okamura et al. [23] utilizaram ARIMA e SARIMA para projetar despesas da Agência Brasileira de Promoção de Exportações e Investimentos (Apex-Brasil), demonstrando que modelos sazonais podem capturar padrões de execução orçamentária e apoiar decisões de alocação de recursos. Embora o estudo se concentre em despesas, a mesma lógica pode ser aplicada à previsão de cargas de trabalho no contencioso, subsidiando decisões sobre dimensionamento de equipes, metas de produtividade e priorização de casos.

2.4 MODELOS DE LINGUAGEM DE GRANDE ESCALA (LARGE LANGUAGE MODELS (LLMS))

Os Modelos de Linguagem de Grande Escala (LLMs) representam um paradigma emergente no campo da inteligência artificial e têm impactado significativamente diversas áreas relacionadas à análise e integração de dados. Esses modelos, baseados em arquiteturas de *transformers*, são treinados com grandes volumes de dados textuais e são capazes de aprender representações linguísticas profundas, capturando relações semânticas, sintáticas e contextuais de forma altamente eficaz.

Os *Large Language Models* (LLMs) representam um avanço significativo no campo do *Processamento de Linguagem Natural* (PLN), permitindo que sistemas computacionais compreendam, gerem e manipulem textos em linguagem natural com níveis inéditos de sofisticação. Baseados em arquiteturas do tipo *transformer* [25], esses modelos se destacam por sua capacidade de capturar relações complexas em grandes corpora textuais, o que os torna particularmente adequados para domínios ricos em texto, como o jurídico.

O princípio fundamental dos LLMs consiste no pré-treinamento em grandes corpora textuais, geralmente extraídos da web ou de coleções especializadas, para aprender representações semânticas e contextuais de palavras, sentenças e documentos. Posteriormente, esses modelos podem ser ajustados (*fine-tuning*) ou adaptados por técnicas de *prompt engineering* para tarefas específicas, como classificação de documentos, extração de entidades, sumarização automática, resposta a perguntas ou redação assistida [26].

Um aspecto central dos LLMs é sua capacidade de operar em cenários multilíngues. Modelos como mBERT, XLM-RoBERTa e mT5 foram desenvolvidos para lidar com diferentes idiomas, permitindo aplicações mais abrangentes em países com diversidade linguística, como o Brasil. No estudo de Bernhard [27], foram comparados três modelos multilíngues na tarefa de classificação de documentos de prestação de contas do Tribunal de Contas do Estado do Maranhão (TCE/MA). O XLM-RoBERTa obteve o melhor desempenho (F1-score de 98,99%), superando mBERT e mT5, o que demonstra o potencial dos LLMs para automatizar tarefas complexas de análise documental no setor público, inclusive em contextos jurídicos.

Além da classificação, LLMs têm sido explorados em tarefas de extração de informações relevantes, sumarização de peças processuais, análise de argumentos e detecção de padrões em jurisdição [28, 29]. Ferramentas baseadas em LLMs podem, por exemplo, auxiliar na identificação de dispositivos legais citados, na agrupação de processos por tese e na redação de minutas de manifestações, sempre condicionadas à necessidade de revisão humana e à implementação de salvaguardas de qualidade e responsabilidade.

Entretanto, a adoção de LLMs também apresenta desafios relevantes, incluindo elevada demanda computacional, riscos de vieses nos dados de treinamento, possibilidade de alucinações (respostas plausíveis porém factualmente incorretas) e questões de transparência e explicabilidade [30]. Em contextos jurídicos, tais desafios tornam ainda mais necessário o uso de arquiteturas híbridas, como *Retrieval-Augmented Generation* (RAG), em que o modelo é acoplado a bases de conhecimento confiáveis, e a definição de mecanismos de governança algorítmica e *Explainable AI* (XAI).

2.5 INTELIGÊNCIA ARTIFICIAL E SUA EVOLUÇÃO PARA MODELOS DE LINGUAGEM

A Inteligência Artificial (IA) pode ser definida como o campo da ciência da computação dedicado ao desenvolvimento de sistemas capazes de executar tarefas que, tradicionalmente, exigiriam inteligência humana, como raciocínio, aprendizagem, tomada de decisão e compreensão da linguagem natural. Historicamente, a IA evoluiu a partir de abordagens simbólicas e baseadas em regras, nas quais o conhecimento era explicitamente codificado por especialistas, para métodos estatísticos e probabilísticos, impulsionados pelo aumento da capacidade computacional e da disponibilidade de grandes volumes de dados.

Nas últimas décadas, o advento do aprendizado de máquina (Machine Learning) e, posteriormente, do aprendizado profundo (Deep Learning), possibilitou avanços significativos na modelagem de dados complexos. Redes neurais artificiais profundas passaram a demonstrar desempenho superior em tarefas como reconhecimento de padrões, processamento de sinais e visão computacional. Nesse contexto, o Processamento de Linguagem Natural (PLN) beneficiou-se diretamente dessas arquiteturas, permitindo a criação de modelos capazes de aprender representações semânticas a partir de grandes corpora textuais.

A evolução culminou no surgimento dos modelos de linguagem de larga escala (Large Language Models (LLMs)), baseados predominantemente em arquiteturas do tipo *transformer* [25]. Esses modelos são treinados de forma auto-supervisionada e conseguem capturar relações sintáticas, semânticas e pragmáticas da linguagem, possibilitando aplicações avançadas como geração de texto, tradução automática, sumarização e interação conversacional, estabelecendo a base tecnológica para sistemas de chatbot inteligentes.

2.6 IA APLICADA AO DOMÍNIO JURÍDICO

A aplicação de técnicas de Inteligência Artificial ao domínio jurídico tem se intensificado em resposta ao crescente volume de informações legais, à complexidade normativa e à necessidade de maior eficiência nos sistemas judiciais. O Direito caracteriza-se como um domínio intensivo em dados textuais não estruturados, incluindo leis, jurisprudência, doutrina, contratos e decisões judiciais, o que torna o uso de técnicas avançadas de PLN particularmente relevantes.

Sistemas de IA no contexto jurídico são empregados em tarefas como recuperação de informação jurídica, análise de precedentes, classificação de processos, apoio à redação de peças processuais e identificação de padrões decisórios. Tais sistemas contribuem para a redução do tempo necessário para atividades repetitivas e para a melhoria da consistência e precisão na análise documental, auxiliando profissionais do Direito na tomada de decisão.

Entretanto, a adoção de IA no domínio jurídico impõe desafios específicos [31], como a necessidade de interpretabilidade dos modelos, a conformidade com princípios éticos e legais, e a mitigação de vieses algorítmicos. Assim, o desenvolvimento de soluções de IA jurídica exige uma abordagem interdisciplinar, integrando conhecimentos de engenharia, ciência de dados, direito e ética.

2.7 AUTOMAÇÃO DE PROCESSOS JUDICIAIS

A automação de processos judiciais refere-se à utilização de sistemas computacionais para executar, parcial ou totalmente, atividades administrativas e analíticas no âmbito do Judiciário. Essas atividades incluem o protocolo e triagem de processos, a extração automática

de informações relevantes, a gestão de prazos e a movimentação processual. A automação visa aumentar a eficiência operacional, reduzir custos e minimizar erros humanos.

Com o avanço da IA, a automação deixou de se limitar a fluxos determinísticos baseados em regras, passando a incorporar mecanismos inteligentes capazes de lidar com dados não estruturados e decisões probabilísticas. Técnicas de PLN e aprendizado de máquina permitem, por exemplo, a identificação automática do tipo de ação, das partes envolvidas e dos pedidos formulados, acelerando a tramitação processual.

No contexto dos tribunais brasileiros, a automação inteligente é especialmente relevante devido ao elevado volume de processos e à sobrecarga do sistema judicial. A implementação dessas soluções contribui para a celeridade processual e para a melhoria do acesso à justiça, desde que acompanhada de governança adequada e validação contínua dos modelos utilizados.

2.8 ARQUITETURAS RAG E SUA RELEVÂNCIA JURÍDICA

As arquiteturas de Geração Aumentada por Recuperação (Retrieval-Augmented Generation (RAG)) combinam modelos generativos de linguagem com mecanismos de recuperação de informação a partir de bases de dados externas. Diferentemente dos modelos puramente generativos, as arquiteturas RAG permitem que as respostas sejam fundamentadas em documentos específicos, aumentando a precisão e a confiabilidade das informações geradas.

No domínio jurídico, essa característica é particularmente relevante, uma vez que respostas devem estar alinhadas a fontes normativas e jurisprudenciais concretas. Sistemas baseados em RAG podem consultar legislações, decisões judiciais e doutrina antes de gerar uma resposta, reduzindo o risco de alucinações e aumentando a rastreabilidade das informações apresentadas.

Além disso, arquiteturas RAG favorecem a atualização contínua do conhecimento jurídico, pois permitem a integração dinâmica de novas fontes documentais sem a necessidade de re-treinamento completo do modelo. Essa flexibilidade torna os sistemas RAG adequados para aplicações jurídicas que exigem precisão, transparência e aderência ao ordenamento jurídico vigente.

2.9 SISTEMAS DE CLASSIFICAÇÃO AUTOMÁTICA DE PROCESSOS

A classificação automática de processos judiciais consiste na categorização de ações com base em critérios como matéria, classe processual, assunto principal ou grau de complexidade. Essa tarefa é fundamental para a organização do fluxo judicial e para a correta distribuição dos processos entre varas e magistrados especializados.

Modelos de aprendizado de máquina supervisionado, aliados a técnicas de PLN, têm sido amplamente utilizados para essa finalidade. A extração de características textuais de petições

iniciais e documentos processuais permite que os sistemas aprendam padrões linguísticos associados a diferentes classes de processos, alcançando elevados níveis de acurácia.

A automação da classificação processual contribui significativamente para a eficiência do Judiciário, reduzindo o retrabalho e o tempo de tramitação inicial. Além disso, esses sistemas fornecem dados estruturados que podem ser utilizados em análises estatísticas e estudos de jurimetria.

2.10 JURIMETRIA E ANÁLISE PREDITIVA

A jurimetria é o campo que aplica métodos quantitativos, estatísticos e computacionais à análise de dados jurídicos, com o objetivo de identificar padrões, tendências e correlações no comportamento do sistema judicial. A crescente digitalização dos processos judiciais tem ampliado a disponibilidade de dados, viabilizando análises em larga escala.

A incorporação de técnicas de aprendizado de máquina permite o desenvolvimento de modelos preditivos capazes de estimar, por exemplo, a probabilidade de êxito de uma ação, o tempo médio de tramitação ou o comportamento decisório de determinados tribunais. Esses modelos auxiliam advogados, magistrados e gestores públicos na tomada de decisões mais informadas.

Apesar de seu potencial, a análise preditiva no contexto jurídico deve ser utilizada com cautela, considerando limitações metodológicas, riscos de vieses e implicações éticas. A transparência dos modelos e a correta interpretação dos resultados são essenciais para garantir o uso responsável da jurimetria.

2.11 CHATBOTS JURÍDICOS E INTERFACES INTELIGENTES

Chatbots jurídicos são sistemas conversacionais baseados em IA projetados para interagir com usuários por meio de linguagem natural, fornecendo informações jurídicas, orientações processuais ou suporte a profissionais do Direito. Esses sistemas utilizam modelos de linguagem avançados aliados a técnicas de PLN para compreender consultas e gerar respostas contextualizadas.

No âmbito institucional, chatbots podem ser empregados para atendimento ao cidadão, esclarecendo dúvidas sobre procedimentos judiciais, acompanhamento processual e acesso a serviços públicos. Já no contexto profissional, atuam como ferramentas de apoio à pesquisa jurídica, triagem de casos e consulta rápida a bases normativas.

A eficácia dos chatbots jurídicos depende diretamente da qualidade dos modelos subjacentes, da integração com fontes confiáveis e da capacidade de explicar limitações e incertezas. Assim, o desenvolvimento dessas interfaces deve priorizar precisão, segurança da informação e conformidade com princípios éticos e legais.

2.12 TRABALHOS CORRELATOS

O presente estudo insere-se na interseção entre três vertentes principais de pesquisa: (i) o uso de chatbots e agentes conversacionais em serviços digitais e contextos sensíveis; (ii) a aplicação de inteligência artificial, em especial *Large Language Models* (LLMs), em domínios jurídicos e no setor público; e (iii) o desenvolvimento de frameworks arquiteturais para automação de processos complexos, com requisitos explícitos de governança, qualidade e privacidade. Esta seção apresenta os trabalhos mais diretamente relacionados, detalhando o contexto, os métodos, os principais resultados e conclusões, com destaque para as lacunas que justificam a proposta do *ChatORION*.

Braun et al. [4], propuseram um framework de classificação de chatbots baseado em seis dimensões (domínio de aplicação, canal, modo de interação, objetivo, grau de autonomia e características da linguagem). O estudo tem caráter conceitual, fundamentado em revisão de literatura e análise de casos, e busca organizar o ecossistema de chatbots em categorias claramente distinguíveis. Como resultado, os autores demonstram que muitas soluções são rotuladas genericamente como "chatbots", embora possuam naturezas e capacidades bastante distintas. A conclusão central é que a falta de uma taxonomia clara dificulta tanto a comunicação científica quanto o planejamento arquitetural de sistemas conversacionais. Para o presente trabalho, esse framework oferece um vocabulário e um referencial útil para posicionar o *ChatORION* como um agente conversacional especializado, com foco em apoio a rotinas de contencioso jurídico e integração com sistemas institucionais.

Larsen e Følstad [5], investigam os efeitos de chatbots na prestação de serviços públicos por meio de entrevistas qualitativas com cidadãos e servidores. O estudo adota uma abordagem exploratória e interpretativa, coletando evidências empíricas sobre experiências reais de uso. Os resultados indicam que os chatbots contribuem principalmente para melhorias incrementais, como redução de tempo de espera e ampliação de acesso a informações básicas, mas raramente promovem transformações estruturais na forma como os serviços são concebidos. Os autores concluem que persistem desafios de confiança, de adequação da linguagem e de integração com processos de *backoffice*. Essas conclusões conversam diretamente com o *ChatORION*, que se diferencia ao focar justamente na automação de rotinas internas complexas (triagem e análise processual), e não apenas no atendimento superficial ao usuário final.

No campo da avaliação de qualidade, Barletta et al. [7] propuseram um método de avaliação de um chatbot clínico com base no modelo de Qualidade em Uso do ISO/IEC 25010, combinando critérios como efetividade, eficiência, satisfação, liberdade de risco e cobertura do contexto, ponderados por meio do método Analytic Hierarchy Process (AHP). Metodologicamente, o trabalho envolve a definição de critérios, a consulta a especialistas de domínio e a aplicação do AHP para derivar pesos e priorizar dimensões de qualidade. Os resultados mostram que efetividade e segurança (liberdade de risco) emergem como dimensões críticas em contextos clínicos, e que a percepção dos especialistas converge para a necessidade de mecanismos robustos de validação das respostas do chatbot. A conclusão é que avaliações superficiais, centradas apenas

em usabilidade, são insuficientes em domínios de alto risco. Para esta dissertação, o trabalho fornece um referencial para pensar a avaliação futura do *ChatORION*, especialmente sob a ótica de qualidade em uso em um domínio sensível como o jurídico-fiscal.

A perspectiva dos usuários finais é explorada em [8], que analisaram a interação de estudantes do ensino médio com chatbots educacionais. O estudo adota métodos mistos, combinando questionários e análise qualitativa de feedbacks, para identificar fatores que influenciam a percepção de qualidade da interação. Os resultados indicam que clareza das respostas, previsibilidade do comportamento do chatbot, capacidade de recuperação de erros e ausência de respostas evidentemente equivocadas são mais importantes que atributos antropomórficos ou "personalidade" do agente. A conclusão central é que o valor percebido pelos usuários está fortemente ligado à confiabilidade e consistência funcional, mais do que a tentativas de humanização artificial. Essa evidência converge com objetivos do *ChatORION*, que privilegia precisão, robustez e transparência na triagem de processos, em detrimento de aspectos puramente estéticos da conversação.

A intenção de uso continuado de serviços baseados em chatbots é abordada por Silva et al. [32]. Os autores propõem um modelo teórico, testado via survey e modelagem de equações estruturais, para entender quais fatores determinam a intenção de reutilização de chatbots por usuários. O modelo inclui variáveis como utilidade percebida, facilidade de uso, confiança, qualidade da informação e satisfação. Os resultados mostram que utilidade, confiança e qualidade da informação têm impactos significativos na intenção de reuse, enquanto elementos relacionados a antropomorfismo e entretenimento têm efeito mais limitado. Conclui-se que, em contextos de serviço, chatbots são percebidos como ferramentas funcionais, e não como interlocutores sociais. No presente trabalho, tais achados reforçam a importância de desenhar o *ChatORION* com foco em utilidade operacional e confiabilidade das respostas, especialmente para usuários internos como procuradores e servidores.

Questões de privacidade e proteção de dados em chatbots são tratadas em profundidade em [9]. Os autores conduziram uma revisão abrangente da literatura sobre desenvolvimento dirigido por conversação (*Conversation-Driven Development* – CDD), identificando riscos e desafios de privacidade ao longo do ciclo de vida do chatbot (coleta, armazenamento, treinamento, avaliação). Com base na revisão, propõem um conjunto de requisitos de privacidade para orientar o desenvolvimento de chatbots, incluindo minimização de dados, controle de logs, anonimização, consentimento explícito e gestão de terceiros. Os resultados destacam que, na prática, muitos projetos negligenciam esses aspectos, concentrando-se na experiência do usuário e em métricas de engajamento. A conclusão é que a privacidade precisa ser incorporada como requisito de primeira classe no desenvolvimento de chatbots. Esses achados são fundamentais para o *ChatORION*, que lida com dados judiciais e fiscais sensíveis, exigindo arquitetura e governança compatíveis com a LGPD e com boas práticas de segurança da informação.

Complementarmente, Oliveira et al. [10] discutiram arquiteturas híbridas para o uso de LLMs em chatbots que manipulam informações sensíveis. O estudo avalia diferentes estratégias

de particionamento entre componentes locais (on-premises) e serviços em nuvem, além de mecanismos de *Retrieval-Augmented Generation* (RAG) que conectam o modelo a bases de conhecimento controladas. Metodologicamente, os autores comparam configurações em termos de desempenho, latência e risco de vazamento de dados, por meio de experimentos controlados e análises de caso. Os resultados indicam que arquiteturas puramente em nuvem são problemáticas em domínios sensíveis, e que soluções híbridas, com camadas de filtragem, anonimização e controle de contexto, oferecem um melhor compromisso entre utilidade e privacidade. A conclusão reforça a necessidade de desenho arquitetural cuidadoso quando se aplica LLMs em setores como saúde, finanças e justiça — exatamente o tipo de preocupação incorporada ao desenho do *ChatORION*.

No contexto brasileiro de vulnerabilidade social e justiça criminal, Silva et al. [12] descreveram o desenvolvimento de um chatbot integrado ao aplicativo ESVirtual, destinado a apoiar egressos do sistema prisional em sua reintegração social. A solução adota o framework Rasa, com arquitetura modular em Docker, e é resultado de um processo iterativo de modelagem de intenções, entidades e fluxos de diálogo, em parceria com especialistas de domínio. Os resultados apontam que o chatbot é capaz de responder a dúvidas recorrentes sobre documentação, benefícios e serviços, reduzindo a sobrecarga de equipes e ampliando o acesso à informação. Contudo, o artigo destaca limitações, como a necessidade de constante atualização de conteúdo, dificuldades linguísticas (gírias, variações regionais) e a opção consciente de não empregar IA generativa para evitar respostas imprevisíveis. A conclusão enfatiza que, em contextos sensíveis, a curadoria manual e o controle estrito das respostas são preferíveis a ganhos de fluência conversacional. Este trabalho é relevante para o *ChatORION* por demonstrar, na prática, os desafios de aplicar chatbots em domínios jurídicos e sociais no Brasil e por ressaltar a importância de decisões arquiteturais orientadas por riscos.

Do ponto de vista de gestão de projetos e priorização de valor no setor público, Tueiv [11] propos combinar o *Incremental Funding Method* (IFM) com o AHP para selecionar e priorizar incrementos em projetos de chatbots e aprendizado de máquina em órgãos governamentais. Utilizando um estudo de caso no Instituto Nacional do Seguro Social (INSS), os autores demonstram como diferentes funcionalidades podem ser ordenadas por valor entregue, risco e esforço. Os resultados mostram que essa abordagem reduz o desperdício de recursos e aumenta a probabilidade de adoção efetiva das soluções. A conclusão sugere que projetos de IA no setor público se beneficiam de métodos estruturados de priorização, particularmente em ambientes com múltiplos stakeholders. No contexto desta dissertação, esses conceitos são úteis para orientar a evolução incremental do *ChatORION* e a seleção de funcionalidades de maior impacto para a PGFN.

A dimensão da acessibilidade é discutida em [33], que revisa diretrizes existentes e identifica lacunas no apoio à acessibilidade em chatbots. O estudo analisa documentos normativos (como WCAG), literatura científica e práticas de mercado, apontando barreiras relacionadas à linguagem excessivamente técnica, respostas muito longas, ausência de controle de ritmo da interação e incompatibilidades com leitores de tela. Os autores concluem que é necessário desenvolver

diretrizes específicas para acessibilidade conversacional, indo além da simples adaptação de padrões para páginas web. Esses achados são importantes para o *ChatORION*, que deverá ser utilizado por perfis diversos de usuários internos, com diferentes níveis de familiaridade tecnológica e especialização jurídica, exigindo cuidado com linguagem, densidade de informação e suporte a diferentes formas de interação.

Por fim, a aplicação de LLMs em domínios sensíveis é discutida em trabalhos como o de Bernhard [27], que compara mBERT, XLM-RoBERTa e mT5 na classificação de documentos de prestação de contas do Tribunal de Contas do Estado do Maranhão. O estudo, de natureza experimental, mostra que o XLM-RoBERTa obteve o melhor desempenho (F1-score de 98,99%), superando os demais modelos multilíngues. A conclusão é que LLMs pré-treinados podem ser adaptados com sucesso a tarefas específicas do setor público, automatizando atividades de análise documental com alta qualidade. Khurana et al. [28] e Hagos et al. [30] aprofundam a discussão sobre riscos, vieses e governança no uso de LLMs, enfatizando a necessidade de mecanismos de explicabilidade e de controle humano em aplicações críticas. Esses trabalhos reforçam a visão de que soluções como o *ChatORION* devem ser concebidas como sistemas sociotécnicos, combinando componentes técnicos avançados com estruturas de governança robustas.

Em síntese, os trabalhos correlatos apontam que:

1. Chatbots e agentes conversacionais já são amplamente explorados em serviços públicos e domínios sensíveis, mas a maioria das soluções permanece concentrada em atendimento ao usuário final, com pouca automação das rotinas internas de alto valor agregado, como triagem e análise processual.

2. Há um corpo crescente de evidências sobre avaliação de qualidade, experiência do usuário, acessibilidade, privacidade e segurança em chatbots e LLMs, porém ainda são raros os frameworks arquiteturais integrados voltados especificamente ao contencioso público e à gestão de processos judiciais em larga escala.

3. Arquiteturas híbridas, baseadas em LLMs, RAG, *Data Warehouses* jurídicos e agentes de IA especializados, despontam como caminho promissor para conciliar desempenho, governança e confiabilidade em aplicações críticas, mas carecem de estudos empíricos em contextos concretos como o da PGFN.

Os trabalhos relacionados analisados neste capítulo apresentam diferentes abordagens e metodologias para análise de processos jurídicos com gentes conversacionais. Para contextualizar as contribuições da solução proposta nesta dissertação, a Tabela 2.1 apresenta uma analise comparativa entre os trabalhos citados e a solução proposta. A Tabela 2.1 resume as características positivas da solução proposta e a compara com os trabalhos citados, permitindo uma análise dos avanços proporcionados pela solução proposta. Três aspectos centrais foram selecionados para essa comparação: (i) a presença de agentes conversacionais já são amplamente explorados em serviços públicos ; (ii) a existência de um mecanismo dinâmico de cálculo de confiança para as previsões realizadas; e (iii) a escalabilidade das soluções.

Tabela 2.1: Análise comparativa entre os trabalhos relacionados e o ChatORION

Autor / Referência	**Foco do Estudo**	**Método Utilizado**	**Limitação Identificada**	**Contribuição do ChatORION em Relação ao Trabalho**
Braun et al. [4]	Taxonomia e classificação de chatbots	Revisão conceitual e framework categórico	Abordagem descritiva, sem aplicação empírica em domínio jurídico complexo	O ChatORION materializa um chatbot especializado em domínio jurídico fiscal, superando a abordagem apenas classificatória ao propor e validar uma arquitetura aplicada e mensurada experimentalmente
Larsen & Følstad [5]	Impacto de chatbots em serviços públicos	Entrevistas qualitativas	Foco em atendimento ao cidadão (front office), não em rotinas internas complexas	O ChatORION avança ao atuar no backoffice jurídico, automatizando triagem processual e análise quantitativa, não apenas atendimento informacional
Barletta et al. [7]	Avaliação de qualidade de chatbot clínico (ISO/IEC 25010 + AHP)	Modelo multicritério de avaliação	Avaliação centrada em usabilidade e risco clínico	O ChatORION incorpora avaliação quantitativa (acurácia, latência, trade-off Top-k) e governança jurídica, ampliando o escopo para desempenho arquitetural e eficiência operacional
Chatbots educacionais [8]	Percepção de qualidade por usuários	Métodos mistos (survey + análise qualitativa)	Foco em experiência do usuário, não em precisão normativa	O ChatORION prioriza precisão factual e fundamentação jurídica via RAG, alinhando-se a domínios de alto risco regulatório
Privacidade em chatbots [9]	Requisitos de privacidade em CDD	Revisão sistemática	Abordagem normativa, não arquitetural aplicada	O ChatORION incorpora governança de dados e aderência à LGPD em ambiente jurídico real
Oliveira et al. [10]	Arquiteturas híbridas e RAG com LLMs sensíveis	Avaliação arquitetural	Foco geral em sensibilidade de dados	O ChatORION aplica RAG em contexto jurídico-fiscal específico, validando empiricamente trade-offs de desempenho (acurácia vs latência)
Silva et al. [32]	Intenção de uso continuado	Modelagem de equações estruturais	Ênfase comportamental	O ChatORION desloca o foco do comportamento do usuário para desempenho institucional e ganho operacional mensurável

Ao analisar a Tabela 2.1, observa se que a solução proposta se diferencia por abordar três aspectos centrais não contemplados nas demais abordagens. Alguns trabalhos introduzem estruturas voltadas à detecção multimodal de desinformação, porém não realizam uma categorização abrangente das diferentes formas de conteúdo enganoso nem incorporam mecanismos para estimar dinamicamente a confiança das previsões. Outras propostas concentram se na integração de múltiplas fontes e níveis de granularidade, mas igualmente não tratam da confiabilidade das respostas nem da classificação detalhada dos tipos de desinformação.

Dessa forma, o presente trabalho distingue-se ao propor, implementar e avaliar um framework arquitetural específico para automação conversacional e triagem de processos judiciais no contexto da PGFN, integrando agentes de IA, PLN, bases vetoriais e princípios de governança de dados, contribuindo para preencher uma lacuna identificada na literatura nacional e internacional.

3 CHATORION - CHAT OTIMIZADOR DE ROTINAS INTELIGENTES PARA OPERAÇÕES NO CONTENCIOSO NACIONAL

Este capítulo apresenta o ChatORION, um **C**hat **O**timizador de **R**otinas **I**nteligentes para **O**perações no contencioso **N**acional. O ChatORION é fundamentado na abordagem RAG, integrando mecanismos de recuperação semântica baseados em representações vetoriais com modelos de LLM para geração textual condicionada. A ChatORION foi modelado para apoiar a eficiência informacional na tramitação de processos judiciais, permitindo acesso estruturado e contextualizado a documentos institucionais. Ao combinar técnicas de indexação vetorial, busca por similaridade semântica e geração autoregressiva condicionada, o ChatORION promove respostas fundamentadas em evidências documentais, reduzindo ambiguidade e aumentando a precisão factual das interações.

Do ponto de vista estrutural, o ChatORION é organizado em camadas, adotando princípios de separação de responsabilidades, modularidade e escalabilidade. Essa organização possibilita o desacoplamento entre os componentes de interface, recuperação de informações e geração textual, garantindo robustez operacional e flexibilidade de evolução tecnológica.

3.1 VISAO GERAL DO CHATORION

A arquitetura do ChatORION foi desenvolvido para fornecer respostas precisas, contextualizadas e semanticamente coerentes por meio de uma estrutura modular que integra técnicas de PLN, recuperação de informações e geração de linguagem, conforme apresentado na Figura 3.1. O sistema adota a abordagem RAG, combinando mecanismos de busca semântica com modelos de linguagem de grande porte para enriquecer o contexto das respostas geradas.

O ChatORION é organizado em três camadas principais: (i) interface de interação com o usuário, (ii) módulo de recuperação de informações (*retrieval*) e (iii) módulo de geração de respostas (*generation*). Essa separação promove maior escalabilidade, modularidade e confiabilidade, além de reduzir a ocorrência de respostas imprecisas ou alucinações típicas de modelos generativos isolados.

A interação com o sistema inicia na camada de interface, em que o usuário submete perguntas por meio do *frontend* do ChatORION. Essa comunicação é gerenciada por uma API desenvolvida com o *framework* FastAPI, responsável por receber, validar e encaminhar as requisições ao *backend*. O uso do mecanismo CORS (*Cross-Origin Resource Sharing*) garante compatibilidade com diferentes ambientes web, permitindo a integração transparente do sistema com múltiplas plataformas e aplicações externas.

Após a validação da solicitação, a consulta textual é encaminhada ao módulo de recuperação. Nesse estágio, a pergunta é convertida em uma representação vetorial densa utilizando modelos de *embeddings* baseados em *Sentence Transformers*. Esses modelos capturam relações semânticas profundas entre

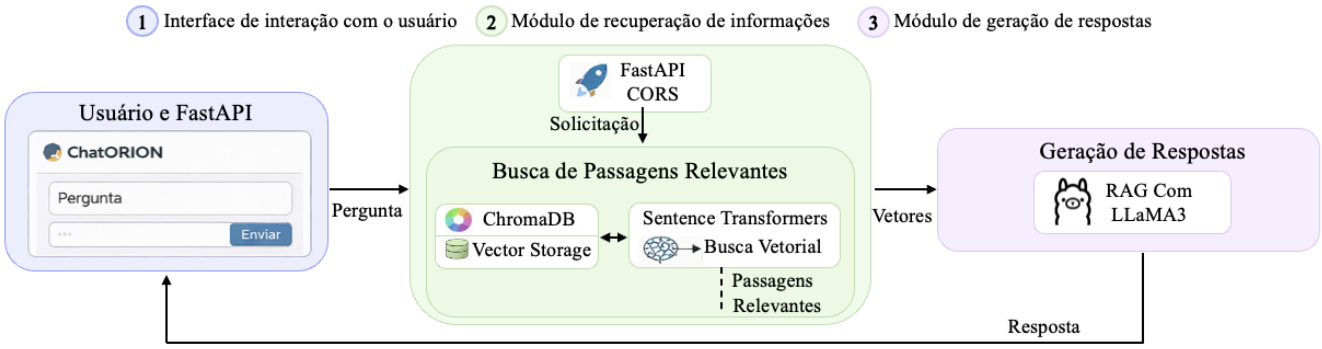


Figura 3.1: Visão geral da arquitetura do ChatORION, destacando as camadas de interface, recuperação vetorial e geração de respostas baseadas em RAG.

palavras e sentenças, projetando tanto consultas quanto documentos em um mesmo espaço vetorial. Essa representação compartilhada viabiliza a realização de buscas semânticas por similaridade, superando limitações de métodos puramente baseados em palavras-chave.

Os vetores gerados são então utilizados pelo ChromaDB, que atua como um banco de dados vetorial responsável pelo armazenamento, indexação e recuperação eficiente de grandes volumes de informações contextuais. A partir da similaridade entre vetores, o sistema identifica e retorna as passagens mais relevantes do corpus, fornecendo evidências contextuais que servirão de base para a etapa de geração de respostas.

Na etapa final, o módulo de geração emprega a estratégia RAG para integrar o conteúdo recuperado ao modelo de linguagem LLaMA3. As passagens selecionadas são incorporadas como contexto adicional ao *prompt*, permitindo que o modelo produza respostas mais fundamentadas, informativas e alinhadas ao domínio do conhecimento disponível. Essa combinação entre recuperação explícita de informações e geração neural reduz alucinações, melhora a precisão factual e aumenta a precisão das respostas produzidas.

3.2 GERENCIAMENTO DE CONSULTAS E CONTROLE DE ACESSO NO CHATORION

O gerenciamento das interações com o ChatORION é realizado na camada de interface, responsável por receber, validar e encaminhar as consultas submetidas pelos usuários ao restante do pipeline de processamento. Para esse fim, o sistema utiliza o framework FastAPI, adotado pela sua alta performance, simplicidade de integração e suporte nativo à construção de serviços assíncronos. Essa escolha possibilita o tratamento eficiente de múltiplas requisições simultâneas, contribuindo para a escalabilidade e redução da latência nas respostas.

A Figura 3.2 apresenta a sequência de interação do ChatORION. No fluxo operacional, cada pergunta submetida pelo usuário é recebida pela API, validada quanto ao formato e aos parâmetros esperados, e então encaminhada ao módulo de recuperação semântica. Dessa forma, o FastAPI atua como o ponto de entrada do sistema, conectando a interface de usuário aos componentes de vetorização, busca vetorial e

geração de respostas. Essa integração garante um encadeamento contínuo entre requisição, processamento e retorno da resposta final.

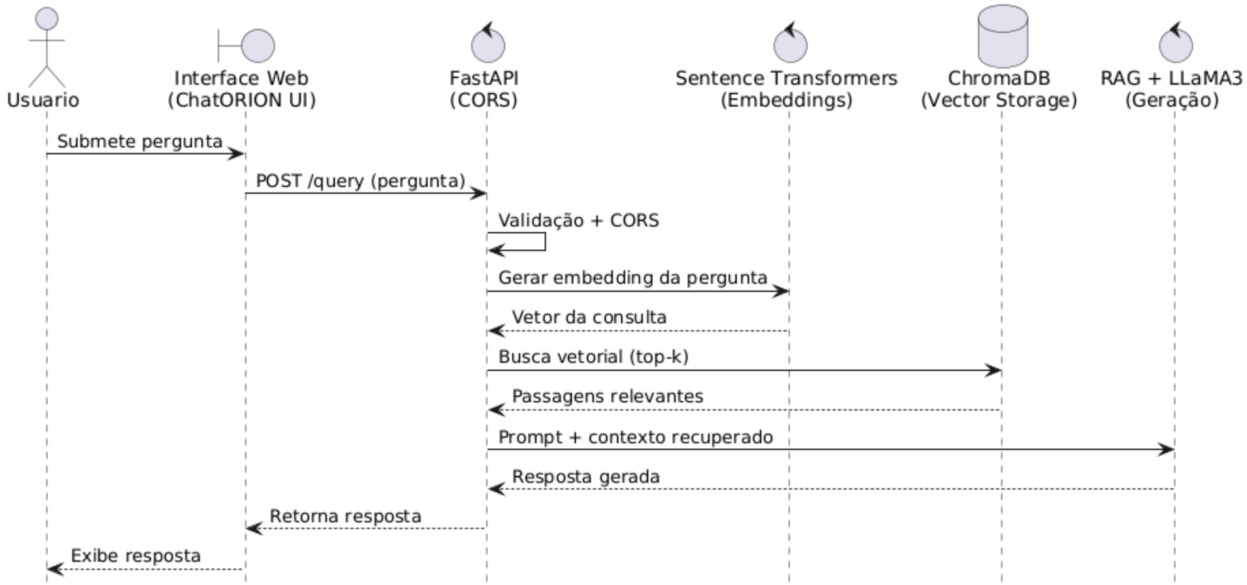


Figura 3.2: Sequência de Interação do ChatORION.

Além do gerenciamento das requisições, o sistema emprega o mecanismo CORS para controle de acesso e segurança. O CORS define políticas que regulam quais origens externas podem consumir os serviços da API, prevenindo acessos não autorizados e assegurando a proteção dos recursos do backend. Tal estratégia é particularmente relevante em cenários distribuídos, nos quais o frontend pode estar hospedado em domínios distintos ou integrado a diferentes aplicações web.

Combinados, FastAPI e CORS estabelecem uma camada de comunicação segura, interoperável e de baixo acoplamento, permitindo que o ChatORION opere de forma confiável em múltiplos ambientes de implantação. Essa camada constitui o primeiro estágio do pipeline arquitetural, habilitando o fluxo completo de processamento que envolve vetorização semântica, recuperação contextual e geração de respostas baseadas em RAG.

3.2.1 Processamento de Embeddings e Recuperação Semântica no ChatORION

O processamento de embeddings compõe o núcleo do módulo de recuperação semântica do ChatORION, sendo responsável por transformar consultas textuais em representações vetoriais densas que preservam relações semânticas e contextuais. Essa etapa constitui o primeiro estágio do mecanismo de *retrieval*, permitindo que perguntas e documentos sejam comparados em um espaço vetorial comum.

O ChatORION utiliza modelos baseados em Sentence Transformers, especializados na geração de *embeddings* de sentenças e parágrafos. Ao receber uma pergunta do usuário, o modelo projeta o texto em um vetor numérico de alta dimensão, no qual similaridades semânticas são refletidas por proximidade geométrica. No ChatORION, a similaridade entre o vetor de consulta $\mathbf{q}$ e o vetor de um documento $\mathbf{d}$ pode

ser estimada por meio da similaridade do cosseno, dada por:

$$\mathrm{sim}(\mathbf{q}, \mathbf{d}) = \frac{\mathbf{q} \cdot \mathbf{d}}{\|\mathbf{q}\| \, \|\mathbf{d}\|}, \tag{3.1}$$

em que $\mathbf{q} \cdot \mathbf{d}$ representa o produto interno e $\|\cdot\|$ denota a norma Euclidiana. Essa métrica é utilizada em recuperação semântica por possibilitar variações e por refletir alinhamento angular entre vetores.

Os vetores gerados são então encaminhados ao ChromaDB, que atua como banco de dados vetorial responsável pelo armazenamento, indexação e recuperação eficiente das representações do corpus documental. A busca por passagens relevantes é conduzida por meio de uma estratégia *top-k*, na qual o sistema seleciona os k vetores com maior similaridade em relação à consulta, retornando as evidências contextuais mais prováveis.

Do ponto de vista computacional, uma *brute-force* apresenta custo linear no número de vetores armazenados, tipicamente $O(N \cdot d)$ por consulta, em que N representa a quantidade de documentos vetorizados e d a dimensionalidade do *embedding*. Para viabilizar a operação em bases de grande porte e reduzir a latência, sistemas vetoriais modernos empregam técnicas de *Approximate Nearest Neighbors* (ANN), as quais utilizam estruturas de indexação e heurísticas para aproximar os vizinhos mais próximos com elevado desempenho em tempo de consulta. Na prática, esse tipo de abordagem segue o estilo adotado por bibliotecas utilizadas em busca vetorial, como a Faiss, que disponibiliza índices otimizados para acelerar consultas de similaridade por meio de quantização, particionamento e estruturas especializadas de busca.

Dessa forma, a integração entre Sentence Transformers e ChromaDB estabelece um pipeline de recuperação semântica composto pelas etapas de vetorização, busca vetorial e seleção de evidências contextuais via *top-k*, com suporte a mecanismos de indexação ANN para ganho de eficiência. As passagens recuperadas são posteriormente encaminhadas ao módulo de geração baseado em RAG com o modelo LLaMA3, que utiliza esse contexto adicional para produzir respostas mais precisas, fundamentadas e alinhadas ao domínio do conhecimento disponível.

3.3 MECANISMO DE GERAÇÃO DE RESPOSTAS NO CHATORION

O mecanismo de geração de respostas constitui a camada final da arquitetural do ChatORION, sendo responsável por transformar o contexto recuperado pelo módulo de *retrieval* em respostas textuais coerentes, informativas e semanticamente alinhadas à consulta do usuário. Em conformidade com a arquitetura modular apresentada na Figura 3.1, essa etapa desacopla a recuperação de informações da geração de linguagem, adotando a estratégia RAG.

O Algoritmo 1 apresenta o processo de vetorização, recuperação contextual e geração textual condicionada no ChatORION. No ChatORION, o módulo de recuperação fornece um conjunto de passagens relevantes $C = \{c_1, c_2, \ldots, c_k\}$ selecionadas a partir do ChromaDB por similaridade semântica. Essas passagens compõem o contexto informacional que fundamenta a resposta gerada. Diferentemente de abordagens puramente generativas, nas quais a saída depende exclusivamente do conhecimento

Algoritmo 1 Recuperação Semântica e Geração Condicionada no ChatORION

Require: consulta q; modelo de embeddings $E : \mathcal{T} \to \mathbb{R}^d$; índice vetorial $D = \{(\mathbf{d}_i, c_i)\}_{i=1}^{N}$; modelo generativo G parametrizado por θ; número de passagens k; tamanho máximo de contexto $L_{\max}$
Ensure: resposta textual r

$\mathbf{v}_q \leftarrow E(q)$ ▷ Projeção semântica da consulta em $\mathbb{R}^d$
for all $(\mathbf{d}_i, c_i) \in D$ **do**
    $s_i \leftarrow \mathrm{sim}(\mathbf{v}_q, \mathbf{d}_i)$ ▷ Similaridade do cosseno (Eq. 3.1)
end for
$\mathcal{C}_k \leftarrow \mathrm{TopK}(\{(s_i, c_i)\}_{i=1}^{N}, k)$ ▷ Seleção das k passagens mais relevantes
$P \leftarrow \mathrm{concat}(q, \mathcal{C}_k)$ ▷ Prompt enriquecido (Eq. 3.2)
if $|P| > L_{\max}$ **then**
    $P \leftarrow \mathrm{Truncate}(P, L_{\max})$ ▷ Controle da janela de contexto do LLM
end if
$r \sim p_\theta(r \mid P)$ ▷ Inferência autoregressiva do LLaMA3
return r

paramétrico do modelo, o uso explícito desse contexto reduz alucinações e aumenta a precisão factual das respostas. Formalmente, o *prompt* enriquecido utilizado pelo modelo generativo pode ser definido como:

$$P = \mathrm{concat}(q, c_1, c_2, \ldots, c_k), \tag{3.2}$$

em que q representa a consulta do usuário e c_i as passagens recuperadas. A geração da resposta r é então modelada como uma distribuição condicional:

$$r \sim p_\theta(r \mid P), \tag{3.3}$$

na qual p_θ corresponde ao modelo de linguagem parametrizado pelo LLaMA3. Assim, a inferência é condicionada ao contexto recuperado, promovendo *grounding* informacional.

A etapa de geração é realizada pelo LLaMA3, um LLM baseado na arquitetura Transformer, capaz de modelar dependências de longo alcance por meio de mecanismos de autoatenção. O modelo realiza inferência autoregressiva, produzindo sequencialmente os tokens da resposta a partir do *prompt* enriquecido. Esse processo garante coerência semântica, fluidez textual e adaptação dinâmica ao contexto fornecido pelo módulo de recuperação.

Do ponto de vista computacional, a inferência em Transformers é dominada pelo custo do mecanismo de autoatenção. Para uma sequência de comprimento n e dimensionalidade d, o custo por camada é tipicamente da ordem de $\mathcal{O}(n^2 d)$, devido ao cálculo das matrizes de atenção densas. Considerando L camadas, o custo total aproxima-se de $\mathcal{O}(Ln^2 d)$. Como o comprimento da sequência cresce com a concatenação das passagens recuperadas ($n \approx |q| + \sum_{i=1}^{k} |c_i|$), existe um compromisso entre quantidade de contexto (k) e latência de inferência. Dessa forma, a seleção *top-k* e o controle do tamanho do contexto tornam-se fundamentais para manter eficiência operacional e escalabilidade do sistema.

3.4 CONSIDERAÇÕES FINAIS

Este capítulo apresentou o ChatORION, modelado por meio do RAG e organizada em camadas funcionais. A separação entre interface, recuperação semântica e geração textual possibilita o desacoplamento de responsabilidades, favorecendo modularidade, escalabilidade e manutenção evolutiva do sistema. Inicialmente, foi descrita a camada de interface, responsável pelo gerenciamento das requisições e controle de acesso por meio do FastAPI e mecanismos de segurança como CORS, garantindo comunicação eficiente e interoperável entre *frontend* e *backend*. Em seguida, apresentou o módulo de recuperação, no qual consultas textuais são projetadas em um espaço vetorial compartilhado utilizando *Sentence Transformers*, permitindo buscas semânticas por similaridade com suporte a técnicas de indexação eficientes, como *Approximate Nearest Neighbors*. Por fim, apresentou o módulo de geração, que integra o contexto recuperado ao modelo LLaMA3 por meio de um *prompt* enriquecido, formalizando a resposta como um processo de geração condicional. Do ponto de vista metodológico, este capítulo também estabeleceu uma modelagem formal do pipeline de processamento, incluindo a definição matemática da similaridade vetorial, da construção do contexto informacional e da geração probabilística das respostas, além da análise de complexidade associada à inferência do modelo Transformer.

4 RESULTADOS E AVALIAÇÃO DE DESEMPENHO

Este capítulo apresenta inicialmente os resultados da fereaamenta, com suas telas demonstrativas e em seguida a avaliação experimental do ChatORION, com o objetivo de investigar o desempenho da arquitetura baseada em RAG. A análise contempla tanto a qualidade informacional das respostas geradas quanto a eficiência computacional do pipeline completo de processamento. A avaliação foi estruturada para examinar o impacto das principais etapas da arquitetura, considerando diferentes configurações do mecanismo de recuperação e distintos modelos generativos. Dessa forma, busca-se mensurar não apenas a precisão e o nível de fundamentação das respostas, mas também o comportamento da latência e o equilíbrio entre qualidade e custo computacional.

Assim como primeira parte dos resultados da arquitetura foi projetado e prototipado um modelo de sistema com um conjunto de Software composto por um módulo principal, um de coleta dos dados, e outro com os resultados em que utiliza diversas fontes de informações, gerada a partir de dados abertos através de um modelo de interoperabilidade de dados gerencias. Tais informações são processadas, através um DW e disponibilizadas através de uma ferramenta de IA (chat). A ferramenta foi desenvolvido em linguagem de programação Python, de forma que no Chatbot o usuário visualiza as informações por meio de resultados das pesquisas.

4.1 APRESENTAÇÃO DO SISTEMA

A Figura 4.1 apresenta a interface inicial do sistema ChatORION, responsável pela autenticação de usuários e pelo gerenciamento de níveis de autorização. Essa camada implementa mecanismos de controle de acesso baseados em perfis, garantindo isolamento funcional e segurança informacional durante a interação com o sistema. Tal arquitetura permite segmentar permissões operacionais conforme o papel institucional do usuário, assegurando governança, rastreabilidade de ações e conformidade com requisitos de auditoria e proteção de dados.

Adicionalmente, a interface inicial apresentada na Figura 4.1 não se limita a um mecanismo tradicional de login, mas constitui o primeiro ponto de orquestração da arquitetura do ChatORION, atuando como camada de mediação entre o usuário institucional e os módulos inteligentes do framework. Ao centralizar autenticação, autorização e registro de sessões, essa camada viabiliza a aplicação de políticas de governança algorítmica, garantindo que as interações com os modelos de linguagem e com a base vetorial ocorram dentro de parâmetros previamente definidos de conformidade normativa. O controle granular de permissões permite, por exemplo, diferenciar perfis operacionais (procuradores, analistas, administradores do sistema), restringindo funcionalidades sensíveis como ingestão de novos documentos, reindexação de bases ou extração de relatórios estratégicos. Dessa forma, a interface inicial consolida-se como elemento fundamental para assegurar integridade sistêmica, accountability e aderência aos requisitos de segurança da informação e proteção de dados no contexto do contencioso público.

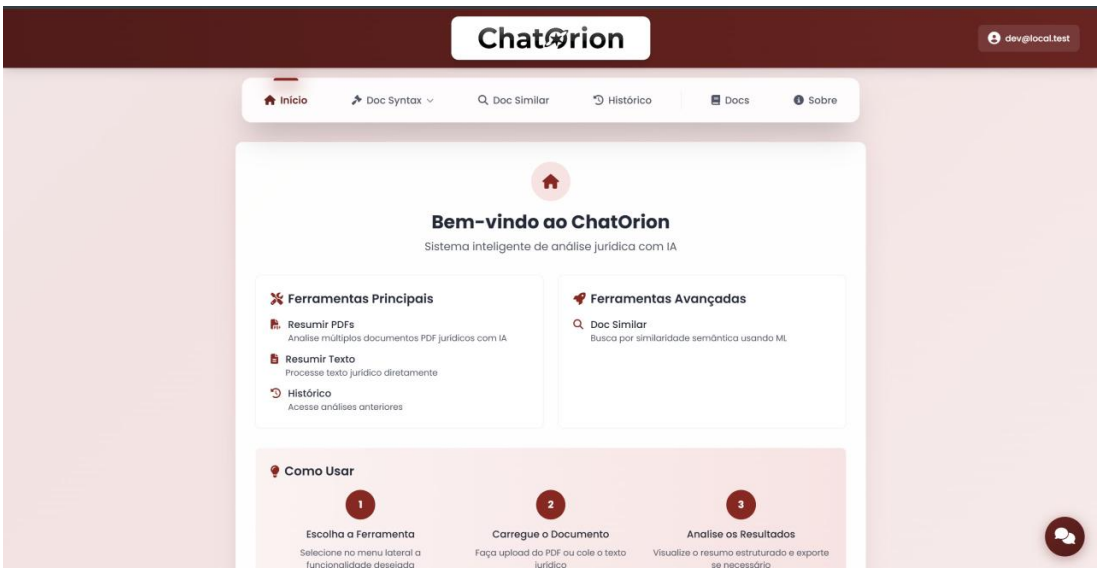


Figura 4.1: Interface inicial do ChatORION com autenticação e controle de permissões.

A execução deste módulo é realizada através de um processo de login com controle de segurança e permissões para usuários e a partir dai é possível carregar os documentos para coleta. Entretanto, cada usuário possui permissões distintas para que possa utilizara das funcionalidades que conduz o gestor a acompanhar os processos e suas interações.

Já na Figura 4.2 ilustra o módulo responsável pela ingestão de documentos processuais na base central do sistema. Nessa etapa, arquivos são carregados por usuários autorizados e passam por um fluxo de preparação que inclui armazenamento, organização e disponibilização para análise automatizada. A partir desse ponto, o framework ativa mecanismos conversacionais que permitem consultas progressivamente mais detalhadas, evidenciando a transição entre simples armazenamento documental e processamento semântico orientado a conhecimento. Esse módulo constitui a base informacional do sistema, pois estabelece o corpus sobre o qual operam os modelos inteligentes.

Complementarmente, o módulo de ingestão representado na Figura 4.2 desempenha papel estratégico na consolidação da infraestrutura cognitiva do ChatORION, ao estruturar o pipeline de transformação de documentos brutos em ativos informacionais semanticamente indexados. Após o carregamento, os arquivos passam por etapas de pré-processamento que podem incluir extração de texto (OCR, quando necessário), normalização, segmentação em unidades lógicas (chunks) e geração de embeddings vetoriais para posterior indexação em base especializada. Esse encadeamento técnico viabiliza a transição de um repositório passivo de documentos para uma base ativa de conhecimento pesquisável por similaridade semântica. Ao garantir padronização, versionamento e rastreabilidade dos dados incorporados, o módulo não apenas sustenta a eficiência dos mecanismos de Retrieval-Augmented Generation (RAG), mas também assegura governança sobre o ciclo de vida da informação jurídica, permitindo atualização incremental do corpus sem comprometer a consistência analítica das consultas realizadas.

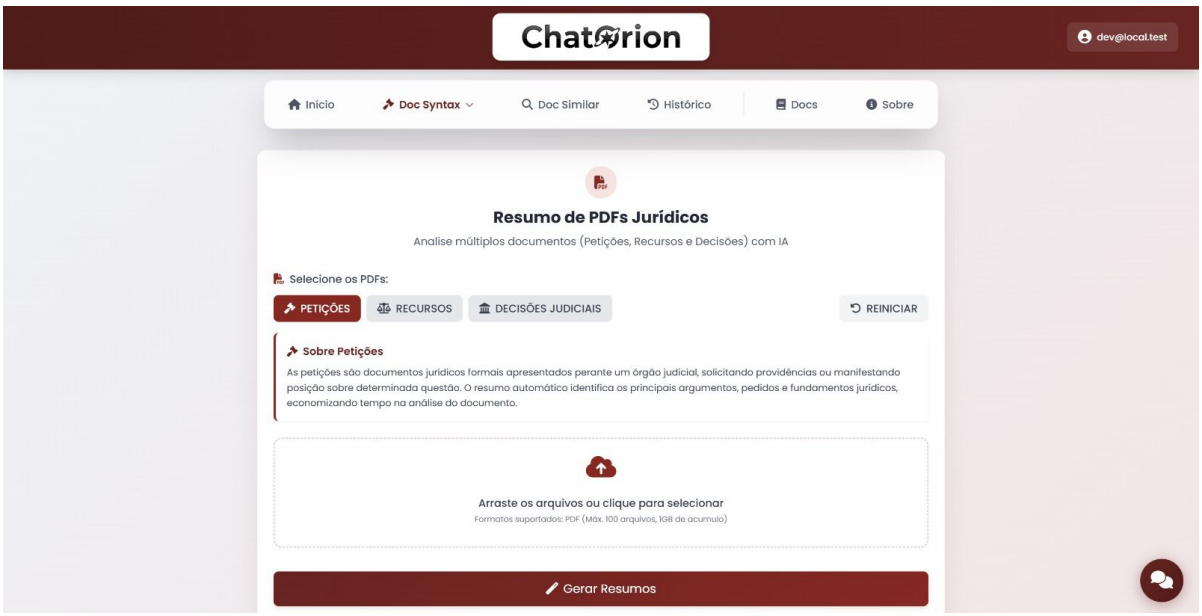


Figura 4.2: Módulo de ingestão documental e habilitação de consultas semânticas.

É apresentado o módulo de carga dos documentos, de forma que os usuários com permissões definidas, podem acessar os documentos na base central e a partir dai coletar informações através de uma mecanismos de perguntas e respostas, podendo realizar interações disponibilizados pelo sistema. As informações começam a ter um nível de detalhamento e cada vez ocorre um nível de aprendizados baseado nas perguntas conforme mostra na Figura 4.3.

O modulo da Figura 4.3, representa o estágio imediatamente posterior ao carregamento dos documentos, no qual o sistema ativa o pipeline cognitivo de interação. Após a ingestão, o ChatORION inicia o processamento semântico das informações e habilita o diálogo em linguagem natural entre usuário e sistema. Esse módulo caracteriza a transição arquitetural entre infraestrutura de dados e inferência inteligente, marcando o início do processamento automatizado orientado a contexto e evidências. Trata-se do ponto em que o sistema deixa de operar como repositório passivo e passa a atuar como agente analítico ativo. É chamado após o documento cargado e a partir desse momento inicializa as interações.

Figura 4.3: Inicialização do pipeline interativo após carregamento de dados.

A Figura 4.4 demonstra o funcionamento integrado do fluxo de respostas e interações do ChatORION. Nessa fase, o sistema executa o pipeline completo de inferência, composto por recuperação semântica, enriquecimento contextual e geração textual. Essa etapa evidencia a natureza modular da arquitetura, na qual componentes especializados operam de forma coordenada para produzir respostas estruturadas, coerentes e fundamentadas. O comportamento observado confirma que o desempenho do sistema depende da interação entre módulos de recuperação e geração, e não apenas da capacidade isolada do modelo de linguagem.

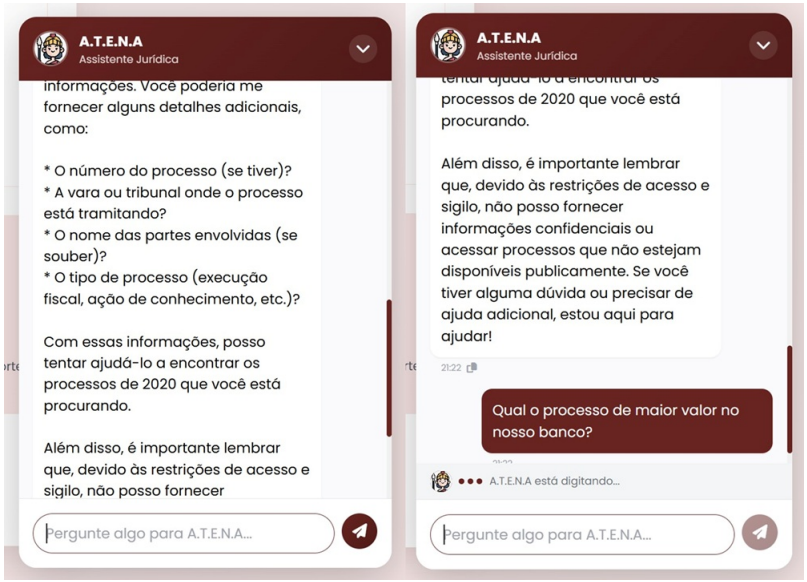


Figura 4.4: Fluxo sistêmico de processamento e geração de respostas.

Por fim a Figura 4.5 representa a fase de estabilização do ciclo operacional do sistema, caracterizada pela execução contínua de consultas e refinamento iterativo das respostas. Nessa etapa, a interação entre usuário, base documental e mecanismo de IA torna-se persistente, permitindo que o sistema mantenha

coerência contextual ao longo das sessões e aprimore progressivamente a qualidade informacional das saídas. Esse comportamento evidencia a natureza adaptativa do framework conversacional e demonstra sua capacidade de sustentar ciclos analíticos sucessivos sem perda de consistência semântica.

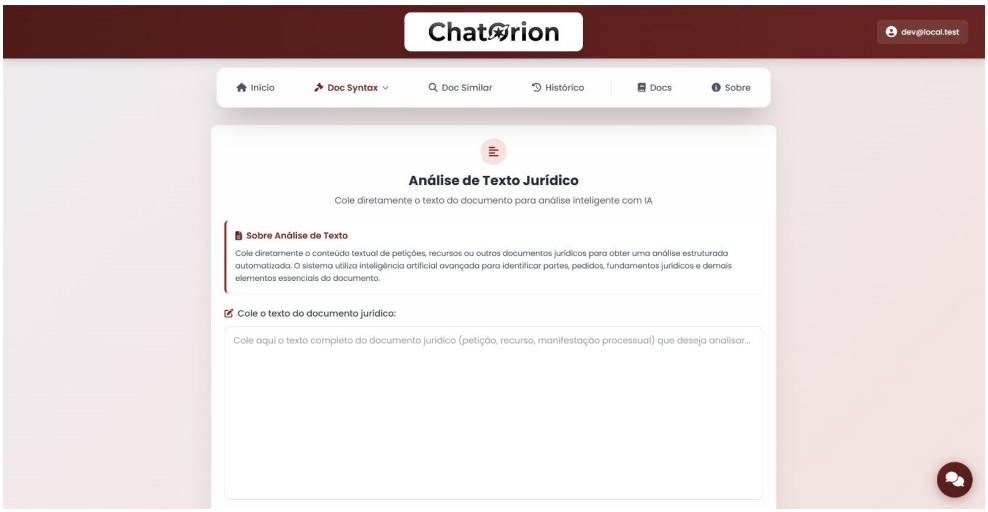


Figura 4.5: Execução iterativa do mecanismo conversacional baseado em RAG.

Diante disso, Conjuntamente, as Figuras 4.1 a 4.5 representam a sequência operacional completa da interface do ChatORION, desde a autenticação inicial até a estabilização do ciclo iterativo de interação. Essa progressão visual comprova empiricamente a operacionalização prática da arquitetura proposta, evidenciando a integração funcional entre ingestão documental, processamento semântico, recuperação contextual e geração de respostas fundamentadas. Tal sequência demonstra que o sistema não se limita a um chatbot convencional, mas configura um framework cognitivo modular voltado à automação analítica de processos jurídicos.

4.2 CONFIGURAÇÃO EXPERIMENTAL

Os experimentos foram conduzidos em ambiente controlado com o objetivo de garantir consistência metodológica e reprodutibilidade dos resultados. O pipeline avaliado corresponde à arquitetura descrita no capítulo anterior, contemplando as etapas de geração de *embeddings* da consulta, recuperação semântica por similaridade vetorial e geração textual condicionada por meio da estratégia RAG. Inicialmente, cada consulta textual é convertida em uma representação vetorial densa utilizando modelos baseados em *Sentence Transformers*. Em seguida, essa representação é utilizada para realizar busca por similaridade do cosseno no ChromaDB, adotando a estratégia *top-k* para seleção das passagens mais relevantes. As evidências recuperadas são então concatenadas à consulta original, compondo o *prompt* enriquecido que serve como entrada para o modelo generativo.

Para avaliar o impacto do modelo de linguagem na etapa de geração, foram consideradas três configurações distintas, mantendo constante o mecanismo de recuperação semântica. A configuração principal do ChatORION utiliza o LLaMA 3.1 (versão utilizada neste trabalho, que é da família LLaMA3)

como modelo generativo. Para fins comparativos, também foram avaliados os modelos Mistral e Phi-2, todos integrados ao mesmo pipeline de recuperação, assegurando que as diferenças observadas estejam associadas ao comportamento do modelo de geração. Além disso, foi considerado o LLaMA 3.1 que opera de forma isolada, sem inclusão das passagens recuperadas, permitindo mensurar empiricamente a contribuição da estratégia RAG para a qualidade das respostas.

O conjunto de avaliação foi composto por consultas representativas do domínio do contencioso nacional, abrangendo diferentes categorias de questionamentos, como interpretações normativas, procedimentos administrativos e solicitações de esclarecimento técnico. Cada consulta foi submetida às diferentes configurações avaliadas, possibilitando a análise comparativa sob as mesmas condições experimentais.

A avaliação do desempenho foi conduzida com base em três métricas principais: acurácia, tempo total de resposta e nível de detalhamento. A acurácia foi definida como a proporção de respostas consideradas factualmente corretas em relação ao total de consultas avaliadas. O tempo total de resposta corresponde ao intervalo entre o recebimento da consulta e a geração da resposta final, englobando as etapas de vetorização, recuperação semântica e inferência do modelo generativo. O nível de detalhamento, por sua vez, foi utilizado para mensurar a profundidade e a completude das respostas, avaliando a capacidade do ChatORION em explorar nuances e apresentar informações contextualizadas.

A infraestrutura utilizada para execução dos experimentos consistiu em um servidor com sistema operacional Ubuntu 23.04 (Kernel 6.2.0-39-generic), processador Intel Core i7-10700F a 2.90 GHz, arquitetura x86_64 e 16 núcleos físicos. O ambiente dispõe de 32 GiB de memória RAM, armazenamento LVM com capacidade de 921 GiB e seis GPUs NVIDIA GeForce RTX 3060 Ti, cada uma com 8 GiB de memória dedicada. A implementação foi realizada em Python 3.13.5.

A avaliação experimental foi estruturada em duas etapas. Na primeira etapa, Subseção 4.2.1, foi analisado o impacto de diferentes configurações do mecanismo de recuperação e do tamanho do contexto, investigando como a variação do parâmetro *top-k* e do limite máximo de tokens influencia simultaneamente a qualidade das respostas e a latência do sistema. Na segunda etapa, Subseção 4.2.2, foi feita uma análise comparativa de eficiência e qualidade entre modelos generativos distintos, mantendo constante o mecanismo de recuperação, com o objetivo de avaliar como a escolha do modelo de linguagem afeta o desempenho global do ChatORION. Essas duas etapas permitem examinar tanto o impacto de decisões arquiteturais internas quanto a influência da camada generativa no comportamento final do sistema.

4.2.1 Avaliação do Impacto de Configurações do Mecanismo de Recuperação e Contexto

A Figura 4.6 apresenta o comportamento da acurácia e da latência média em função do número de passagens recuperadas. Observa-se que, para valores baixos de k, a acurácia é reduzida, indicando que a limitação de evidências contextuais compromete a capacidade do modelo em produzir respostas fundamentadas. À medida que k aumenta de 1 para 5, há crescimento significativo da acurácia, evidenciando que a incorporação de múltiplas passagens relevantes melhora a cobertura informacional do *prompt* enriquecido. Entretanto, a partir de $k > 5$, verifica-se estabilização na acurácia, acompanhada

de aumento contínuo da latência. Esse comportamento sugere que a inclusão excessiva de passagens pode introduzir ruído contextual e ampliar o custo computacional da etapa de geração, uma vez que o comprimento da sequência processada pelo modelo Transformer cresce proporcionalmente. Dessa forma, o valor $k = 5$ representa o melhor compromisso entre qualidade e eficiência, caracterizando um ponto ótimo operacional para o mecanismo RAG no contexto avaliado.

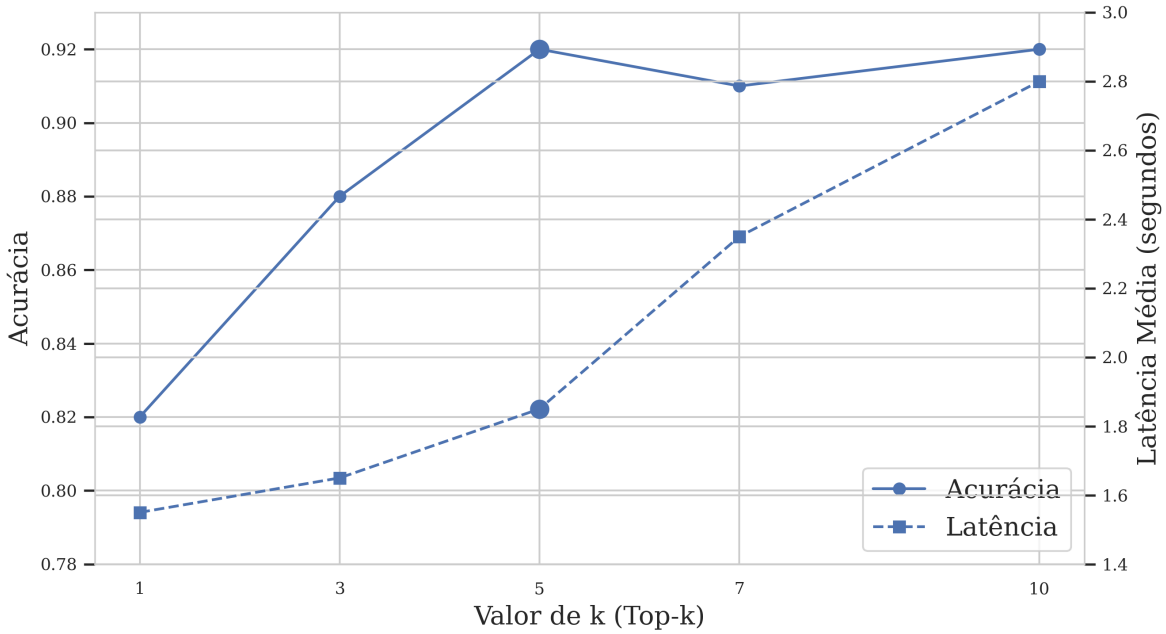


Figura 4.6: Trade-off Acurácia versus Latência em função do Top-k no ChatORION.

A Figura 4.7 apresenta a análise do impacto do truncamento do contexto em diferentes limites de tokens. Observa-se que o uso de truncamento agressivo (1k tokens) reduz significativamente a latência, porém compromete a acurácia, indicando perda de informações relevantes no processo de geração. Por outro lado, a ausência de truncamento resulta na maior latência observada, sem ganhos proporcionais de qualidade, possivelmente devido à inclusão de conteúdo redundante ou marginalmente relevante. A configuração intermediária de 2k tokens apresentou o melhor equilíbrio entre acurácia e latência, mantendo desempenho próximo ao máximo observado com custo computacional substancialmente inferior ao cenário sem truncamento. O uso de 4k tokens, embora preserve boa qualidade, eleva consideravelmente a latência, reforçando o caráter quadrático do custo da atenção no modelo Transformer.

4.2.2 Análise Comparativa de Eficiência e Qualidade entre Modelos

Esta subseção apresenta a análise comparativa entre os modelos avaliados no contexto do ChatORION. O objetivo é investigar como diferentes modelos generativos impactam o desempenho global do sistema quando integrados ao mecanismo de recuperação semântica baseado em RAG.

A Figura 4.8(a) apresenta os valores de acurácia obtidos para cada modelo. O ChatORION alcançou o maior índice de acurácia (0,90), seguido pelo LLaMA 3.1 (0,85). O Mistral apresentou desempenho intermediário (0,75), enquanto o Phi obteve o menor valor (0,70). Esses resultados indicam que a integração entre recuperação semântica e geração condicionada favorece maior alinhamento entre a

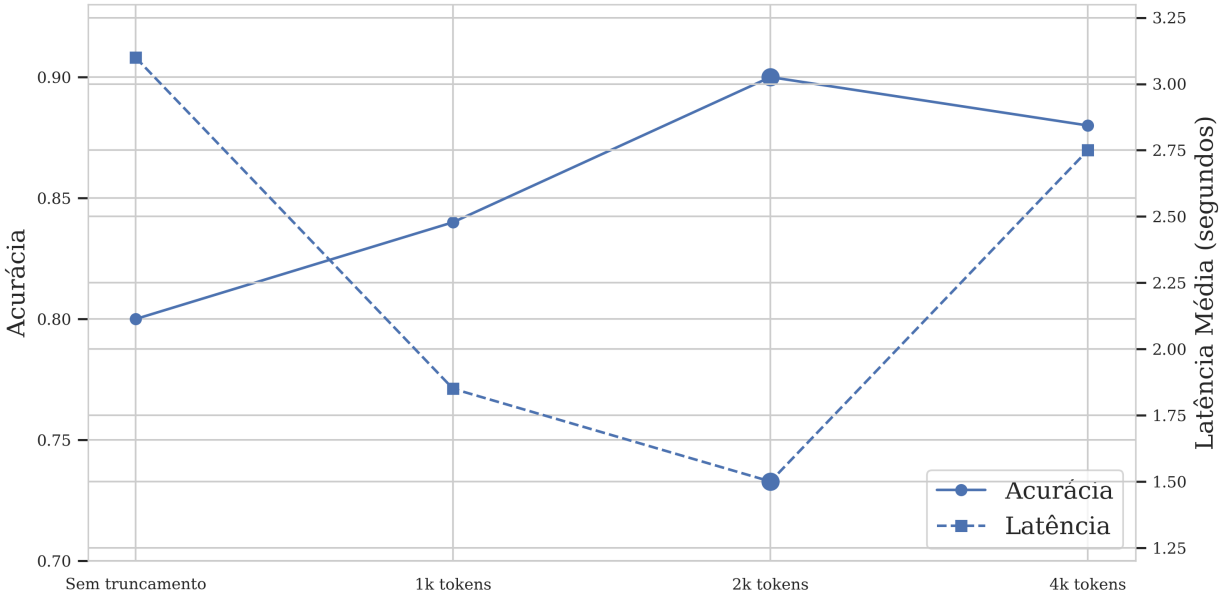


Figura 4.7: Trade-off Acurácia versus Latência no ChatORION

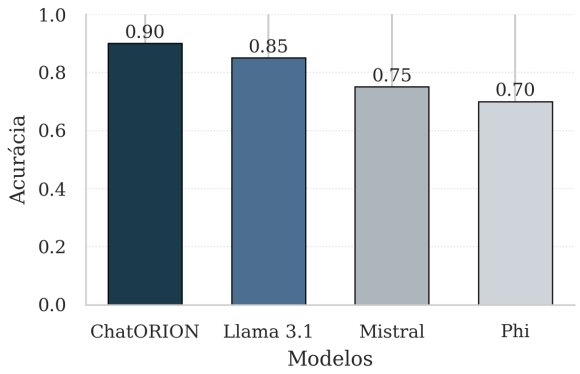
consulta e a resposta produzida, reduzindo inconsistências factuais e respostas não fundamentadas no corpus. A diferença entre os modelos evidencia que a escolha do modelo generativo exerce influência significativa sobre a precisão informacional, mesmo quando submetidos ao mesmo mecanismo de recuperação.

A Figura 4.8(b) apresenta o tempo médio de resposta observado para cada configuração avaliada. Verifica-se que o ChatORION obteve o menor tempo médio de resposta (1,50s), seguido pelo LLaMA 3.1 (1,70s). Os modelos Mistral e Phi apresentaram latências superiores, com valores médios de 2,50s e 3,00s, respectivamente. Esse comportamento está diretamente relacionado ao custo computacional da etapa de inferência, dominada pelo mecanismo de autoatenção dos modelos Transformer. A diferença observada indica que modelos com maior complexidade ou menor nível de otimização apresentam impacto mais significativo na latência, afetando a escalabilidade do sistema em cenários de múltiplas requisições simultâneas.

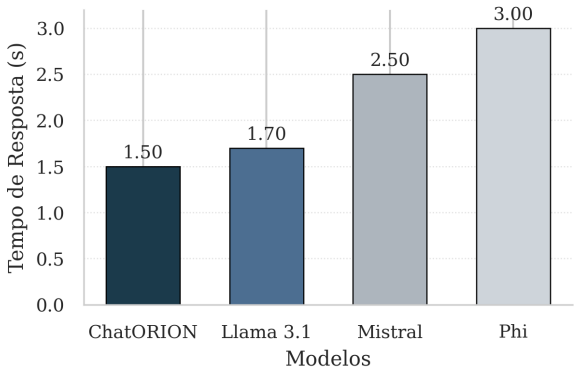
A Figura 4.8(c) apresenta a avaliação do grau de detalhamento das respostas geradas. O ChatORION apresentou o maior valor (0,90), seguido pelo LLaMA 3.1 (0,85), Mistral (0,80) e Phi (0,70). Observa-se uma tendência consistente entre acurácia e detalhamento, indicando que modelos que produzem respostas mais corretas também tendem a fornecer explicações mais completas e contextualizadas. Esse comportamento sugere que a qualidade das respostas não se restringe apenas à correção factual, mas também à capacidade de explorar nuances e estruturar o conteúdo de forma informativa.

4.3 CONSIDERAÇÕES FINAIS

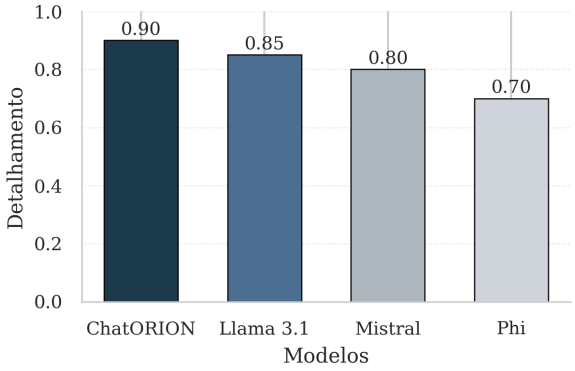
Os resultados obtidos evidenciam que o desempenho do ChatORION depende de dois fatores principais: a configuração do mecanismo de recuperação semântica e a escolha do modelo generativo.



(a) Acurácia.



(b) Tempo de Resposta.



(c) Detalhamento.

Figura 4.8: Desempenho do ChatORION nas métricas de acurácia, tempo de resposta e detalhamento.

A análise conduzida nas duas etapas experimentais permite interpretar o comportamento do sistema de forma integrada, relacionando qualidade informacional e custo computacional.

Na primeira etapa, verificou-se que a quantidade de evidências recuperadas influencia diretamente a qualidade das respostas até um determinado limite. Valores baixos do parâmetro *top-k* restringem o contexto disponível, reduzindo a acurácia, enquanto valores elevados aumentam a latência sem ganhos proporcionais de qualidade. Esse comportamento indica que o mecanismo RAG não depende apenas da presença de recuperação de informação, mas de sua calibração adequada. O mesmo padrão foi observado no tamanho máximo de contexto: contextos muito curtos provocam perda de informação relevante, enquanto contextos muito longos introduzem ruído e aumentam significativamente o custo computacional. Dessa forma, os resultados sugerem a existência de um regime ótimo de operação, no qual a cobertura informacional é suficiente para fundamentar a resposta sem comprometer a eficiência do sistema.

Na segunda etapa, a comparação entre modelos demonstrou que a camada generativa exerce impacto decisivo no desempenho global. Mesmo utilizando o mesmo mecanismo de recuperação, os modelos apresentaram diferenças significativas de acurácia, latência e detalhamento. O ChatORION, ao integrar o LLaMA 3.1 com recuperação semântica, apresentou maior precisão e menor tempo médio de resposta. Isso indica que o ganho não está apenas na qualidade do modelo de linguagem, mas na interação entre recuperação estruturada e geração condicionada.

Portanto, os resultados indicam que o ChatORION alcança melhor desempenho quando operando em um regime calibrado de recuperação e geração, no qual a quantidade de contexto e o modelo generativo são ajustados de forma complementar. Esse comportamento evidencia que a eficiência do sistema não é determinada apenas pela capacidade do modelo, mas pela arquitetura como um todo, destacando a importância do desenho integrado entre busca semântica e geração de linguagem.

5 CONCLUSÃO

Esta dissertação teve como objetivo propor e avaliar uma arquitetura de assistência informacional para operações do contencioso nacional, denominada ChatORION, fundamentada na estratégia RAG. A proposta integrou recuperação semântica baseada em representações vetoriais e geração textual condicionada por modelos de linguagem de grande porte, com o propósito de produzir respostas fundamentadas em documentos institucionais e reduzir inconsistências factuais típicas de modelos puramente generativos. O desenvolvimento do trabalho foi estruturado em duas etapas principais. Inicialmente, foi projetada a arquitetura do sistema, organizada em camadas independentes de interface, recuperação e geração. Em seguida, foi conduzida uma avaliação experimental abrangendo tanto a qualidade informacional quanto a eficiência computacional do pipeline. O trabalho se alinha especificamente à natureza e aos objetivos de um mestrado profissional com foco na aplicação prática, na resolução de problemas reais de uma organização e na interface com o setor público, neste caso, a PGFN.

Os resultados demonstraram que a integração entre recuperação semântica e geração condicionada melhora o alinhamento entre consulta e resposta produzida. A presença de evidências documentais no *prompt* enriquecido reduziu ambiguidades e aumentou a precisão factual das respostas, confirmando empiricamente a vantagem do paradigma RAG em cenários de domínio especializado.

A análise experimental evidenciou ainda que o desempenho do sistema depende da calibração adequada do mecanismo de recuperação. Observou-se que valores baixos de *top-k* limitam o contexto disponível ao modelo, reduzindo a acurácia, enquanto valores excessivos aumentam a latência sem ganhos proporcionais de qualidade. Da mesma forma, o tamanho do contexto apresenta comportamento semelhante: contextos curtos comprometem a fundamentação informacional e contextos longos introduzem ruído e custo computacional elevado. Os experimentos indicaram a existência de um regime operacional ótimo capaz de equilibrar qualidade e eficiência.

A comparação entre modelos generativos demonstrou que a arquitetura exerce impacto mais significativo que a simples substituição do modelo de linguagem. O ChatORION apresentou maior acurácia, menor latência e maior nível de detalhamento quando comparado ao uso isolado do modelo generativo e a outros modelos avaliados. Esse resultado indica que o desempenho global emerge da interação entre recuperação estruturada e geração condicionada, e não apenas da capacidade paramétrica do modelo de linguagem.

5.1 CONTRIBUIÇÕES DO ESTUDO

5.1.1 Contribuições acadêmicas

- Proposição de uma arquitetura modular baseada em RAG para domínio jurídico institucional, integrando recuperação vetorial e geração condicionada.

- Modelagem formal do pipeline de processamento, incluindo definição matemática de similaridade semântica, construção do contexto informacional e geração probabilística condicionada.
- Evidência empírica sobre o impacto da calibração do mecanismo RAG, demonstrando a existência de regime ótimo entre qualidade informacional e custo computacional.
- Avaliação comparativa entre modelos generativos sob um mesmo mecanismo de recuperação, destacando que a arquitetura influencia mais o desempenho que a substituição isolada do modelo de linguagem.

5.1.2 Contribuições práticas

- Desenvolvimento de um assistente informacional capaz de fornecer respostas fundamentadas em documentos institucionais, reduzindo inconsistências factuais em consultas jurídicas.
- Apoio a operações do contencioso por meio de acesso contextualizado à informação normativa e procedimental.
- Definição de parâmetros operacionais recomendados para sistemas RAG em ambientes corporativos, equilibrando latência e qualidade.
- Demonstração da viabilidade de uso de LLMs locais integrados a bases documentais institucionais, preservando controle sobre os dados.

5.2 LIMITAÇÕES DO ESTUDO

Apesar dos resultados positivos, algumas limitações devem ser consideradas:

- O conjunto de avaliação foi restrito a um domínio institucional específico, podendo não refletir integralmente cenários jurídicos mais amplos ou heterogêneos.
- A avaliação da qualidade das respostas baseou-se em métricas definidas manualmente, podendo envolver subjetividade na análise do nível de detalhamento.
- O sistema foi testado em ambiente controlado, não contemplando cargas reais de produção com múltiplos usuários simultâneos em larga escala.
- Não foram exploradas estratégias avançadas de reordenação de documentos ou feedback humano contínuo para refinamento das respostas.

5.3 RECOMENDAÇÕES E TRABALHOS FUTUROS

Com base nos resultados e limitações observadas, sugerem-se as seguintes direções de continuidade:

1. Investigar técnicas de re-ranking semântico e filtragem contextual para melhorar a seleção das evidências recuperadas.
2. Integrar mecanismos de aprendizado com feedback humano para adaptação contínua ao domínio institucional.
3. Avaliar o sistema em ambiente operacional real, considerando múltiplos usuários e cargas simultâneas.
4. Explorar estratégias de compressão de contexto e atenção eficiente para redução adicional da latência.
5. Expandir o sistema para múltiplas bases documentais e diferentes áreas do direito, analisando sua capacidade de generalização.

REFERÊNCIAS BIBLIOGRÁFICAS

1 GOMES, L. V. et al. O processo de aprendizagem nos projetos tradicionais e ágeis. Universidade Nove de Julho, 2023.

2 SALVADOR, F. H. *Métodos ágeis no desenvolvimento de software e seus impactos no alinhamento entre tecnologia da informação e negócios.* Tese (Doutorado) — Universidade de São Paulo, 2025.

3 MANUEL, B. L.; SANTOS, S. dos. Vantagens e inconvenientes da automação de processos de gestão com a inteligência artificial. *RICA-KIANDA: REVISTA DE INOVAÇÃO E INVESTIGAÇÃO CIENTÍFICA DA UNIVERSIDADE DE LUANDA*, v. 1, n. 02, p. 123–136, 2025.

4 BRAUN, D.; MATTHES, F. Towards a framework for classifying chatbots. In: FILIPE, J.; SMIALEK, M.; BRODSKY, A.; HAMMOUDI, S. (Ed.). *Proceedings of the 21st International Conference on Enterprise Information Systems, ICEIS 2019, Heraklion, Crete, Greece, May 3-5, 2019, Volume 1*. SciTePress, 2019. p. 496–501. Disponível em: <https://doi.org/10.5220/0007772704960501>.

5 LARSEN, A. G.; FØLSTAD, A. The impact of chatbots on public service provision: A qualitative interview study with citizens and public service providers. *Gov. Inf. Q.*, v. 41, n. 2, p. 101927, 2024. Disponível em: <https://doi.org/10.1016/j.giq.2024.101927>.

6 JO, E.; JEONG, Y.; PARK, S.; EPSTEIN, D. A.; KIM, Y. Understanding the impact of long-term memory on self-disclosure with large language model-driven chatbots for public health intervention. In: MUELLER, F. F.; KYBURZ, P.; WILLIAMSON, J. R.; SAS, C.; WILSON, M. L.; DUGAS, P. O. T.; SHKLOVSKI, I. (Ed.). *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024*. ACM, 2024. p. 440:1–440:21. Disponível em: <https://doi.org/10.1145/3613904.3642420>.

7 BARLETTA, V. S.; CAIVANO, D.; COLIZZI, L.; DIMAURO, G.; PIATTINI, M. Clinical-chatbot AHP evaluation based on "quality in use"of ISO/IEC 25010. *Int. J. Medical Informatics*, v. 170, p. 104951, 2023. Disponível em: <https://doi.org/10.1016/j.ijmedinf.2022.104951>.

8 ASSAYED, S. K.; WOODS, D.; ALKAHTIB, M.; SHAALAN, K. Effective human-chatbot interaction: A high school student perspective. *International Journal of Computing and Digital Systems*, Unversity of Bahrain, v. 15, n. 1, p. 1–9, 2024.

9 SILVA, G. R. S.; CANEDO, E. D. Privacy in chatbot conversation-driven development: A comprehensive review and requirements proposal. *ACM Trans. Softw. Eng. Methodol.*, v. 34, n. 7, p. 215:1–215:44, 2025. Disponível em: <https://doi.org/10.1145/3730578>.

10 OLIVEIRA, H. T. A. de; SOBRINHO, Á. A. D. C. C.; DIAS, A. d. R. C.; ARAÚJO, A. M. C. de; ARAÚJO, R. D.; MATOS, D. D. M. da C.; GALVEZ, S. M.-N. Large language models for chatbot applications handling sensitive information. *Expert Systems with Applications*, Elsevier, p. 130145, 2025.

11 TUEIV, M.; SCHMITZ, E. Maximizing the value delivered of chatbots in e-gov using the incremental funding method. In: *Proceedings of the 16th International Conference on Theory and Practice of Electronic Governance*. [S.l.: s.n.], 2023. p. 242–246.

12 SILVA, G. R. S.; LIMA, L. C.; PAIVA, G. P.; CANEDO, E. D. Enhancing post-incarceration support: A custom chatbot solution for the brazilian prison system. In: FILIPE, J.; SMIALEK, M.; BRODSKY, A.; HAMMOUDI, S. (Ed.). *Proceedings of the 27th International Conference on Enterprise Information Systems, ICEIS 2025, Porto, Portugal, April 4-6, 2025, Volume 1*. SCITEPRESS, 2025. p. 423–433. Disponível em: <https://doi.org/10.5220/0013086800003929>.

13 HARADA, K. *Direito financeiro e tributário*. [S.l.]: Editora Dialética, 2025.

14 NACIONAL, P. G. da F. *Página oficial da PGFN*. 2025. Acesso em: 01 nov. 2025. Disponível em: <https://www.gov.br/pgfn/pt-br/acesso-a-informacao/institucional/sobre-a-pgfn>.

15 MACHADO, H. de B. *Curso de direito tributário*. [S.l.]: Malheiros, 2003.

16 BARDIN, L.; CAVIGNAC, J.; LINS, C.; MAUX, A.; LIMA, I. B. de; SOUZA, G. I. R. de; DANTAS, M. G. da S.; VIRGINIO, D. F. Gil, ac métodos e técnicas de pesquisa social . são paulo: Atlas, 2010. *Programação Geral*, p. 69, 2010.

17 NACIONAL, P. G. da F. *Página oficial da PGFN*. 2024. Acesso em: 15 nov. 2025. Disponível em: <https://www.gov.br/pgfn/pt-br/servicos>.

18 MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. *Sistemas inteligentes-Fundamentos e aplicações*, v. 1, n. 1, p. 32, 2003.

19 SHMUELI, G.; KOPPIUS, O. R. Predictive analytics in information systems research. *MIS quarterly*, JSTOR, p. 553–572, 2011.

20 BELARMINO, M.; COELHO, R.; LOTUFO, R.; PEREIRA, J. Aplicação de large language models na análise e síntese de documentos jurídicos: Uma revisão de literatura. In: *Anais do XIII Latin American Symposium on Digital Government*. Porto Alegre, RS, Brasil: SBC, 2025. p. 193–202. ISSN 2763-8723. Disponível em: <https://sol.sbc.org.br/index.php/wcge/article/view/36331>.

21 SANTOS, R.; ARAúJO, R.; REGO, P.; FILHO, J. M.; FILHO, J. S.; NETO, J. C.; FREITAS, N.; RODRIGUES, E. Arquitetura de tempo real e modelo de aprendizado de máquina para detecção de fraudes de cartão de crédito. In: *Anais do XXIII Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais*. Porto Alegre, RS, Brasil: SBC, 2023. p. 265–278. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/sbseg/article/view/27212>.

22 LEWIS, P.; PEREZ, E.; PIKTUS, A.; PETRONI, F.; KARPUKHIN, V.; GOYAL, N.; KÜTTLER, H.; LEWIS, M.; YIH, W.-t.; ROCKTÄSCHEL, T. et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, v. 33, p. 9459–9474, 2020.

23 OKAMURA, A. B.; ZANELLO, C. R. d. P.; GONÇALVES, W. T.; SBRUZZI, E. F. Análise e projeção de séries temporais de despesas da agência brasileira de promoção de exportação e investimentos. In: EDITORA CIENTÍFICA DIGITAL. *PROJETOS DE PESQUISA EM DATA SCIENCE: ESTUDO DE CASO SOBRE A APEXBRASIL-INSTITUTO TECNOLÓGICO DE AERONÁUTICA (ITA), TURMA APEXBRASIL*. [S.l.], 2024. v. 1, p. 40–58.

24 SHUMWAY, R. H.; STOFFER, D. S. Arima models. In: *Time series analysis and its applications: with R examples*. [S.l.]: Springer, 2017. p. 75–163.

25 HAN, J.; LU, J.; XU, Y.; YOU, J.; WU, B. Intelligent practices of large language models in digital government services. *IEEE Access*, v. 12, p. 8633–8640, 2024. Disponível em: <https://doi.org/10.1109/ACCESS.2024.3349969>.

26 KADDOUR, J.; HARRIS, J.; MOZES, M.; BRADLEY, H.; RAILEANU, R.; MCHARDY, R. Challenges and applications of large language models. *CoRR*, abs/2307.10169, 2023. Disponível em: <https://doi.org/10.48550/arXiv.2307.10169>.

27 BERNHARD, P. V. Estudo comparativo de large language models aplicados à classificação de documentos de prestação de contas públicas. *Ciência da Computação da Universidade Federal do Maranhão*, Universidade Federal do Maranhão, 2023.

28 KHURANA, D.; KOLI, A.; KHATTER, K.; SINGH, S. Natural language processing: state of the art, current trends and challenges. *Multim. Tools Appl.*, v. 82, n. 3, p. 3713–3744, 2023. Disponível em: <https://doi.org/10.1007/s11042-022-13428-4>.

29 RUAN, Q.; KUZNETSOV, I.; GUREVYCH, I. Are large language models good classifiers? A study on edit intent classification in scientific document revisions. In: AL-ONAIZAN, Y.; BANSAL, M.; CHEN, Y. (Ed.). *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*. Association for Computational Linguistics, 2024. p. 15049–15067. Disponível em: <https://doi.org/10.18653/v1/2024.emnlp-main.839>.

30 HAGOS, D. H.; BATTLE, R.; RAWAT, D. B. Recent advances in generative AI and large language models: Current status, challenges, and perspectives. *IEEE Trans. Artif. Intell.*, v. 5, n. 12, p. 5873–5893, 2024. Disponível em: <https://doi.org/10.1109/TAI.2024.3444742>.

31 PRADO, A. P. A inteligência artificial no mundo do direito: Perspectivas, desafios e limites Ético-jurídicos. *Cognitio Juris*, v. 15, p. 58, 2022.

32 SILVA, F. A. da; SHOJAEI, A. S.; BARBOSA, B. Chatbot-based services: A study on customers' reuse intention. *J. Theor. Appl. Electron. Commer. Res.*, v. 18, n. 1, p. 457–474, 2022. Disponível em: <https://doi.org/10.3390/jtaer18010024>.

33 STANLEY, J.; BRINK, R. ten; VALITON, A.; BOSTIC, T.; SCOLLAN, B. Chatbot accessibility guidance: A review and way forward. In: YANG, X.; SHERRATT, S.; DEY, N.; JOSHI, A. (Ed.). *Proceedings of Sixth International Congress on Information and Communication Technology - ICICT 2021, London, Volume 3*. Springer, 2021. (Lecture Notes in Networks and Systems, v. 216), p. 919–942. Disponível em: <https://doi.org/10.1007/978-981-16-1781-2_80>.

.1 PSEUDOCODIGO: BI ORCAMENTARIO – BLUEPRINT END-TO-END

```
// ========================================================
// BI ORCAMENTARIO -- Blueprint de Pseudocodigo End-to-End
// ========================================================

// ---------- 0) PARAMETROS E METADADOS ----------
CONST AMBIENTE = "PROD"
CONST DATA_ATUAL = TODAY()
CONST ORGAO_PADRAO = "Administracao Publica Federal"
CONST FONTE_DADOS = ["SIAFI", "PortalTransparencia", "MPO", "SIGA/UG",
                     "CNES/IBGE (auxiliares)"]
CONST ZONAS = { RAW:"dl_raw", STG:"dl_stg", CURATED:"dl_curated", DW:"dw"}
CONST ALERTAS = {EMAIL:"orcamento-bi@org.gov.br", TEAMS:
                                                    "canal-bi-orcamento"}

STRUCT JanelaCarga { inicio: DATE, fim: DATE }
VAR janela_execucao = JanelaCarga(
  inicio = PRIMEIRO_DIA_DO_ANO(DATA_ATUAL),
  fim    = DATA_ATUAL
)

// ---------- 1) ORQUESTRACAO ----------
PROCEDURE PIPELINE_BI_ORCAMENTARIO(janela_execucao):
  TRY
    LOG("Inicio pipeline", DATA_ATUAL, AMBIENTE)
    dados_raw = INGESTAO_MULTIFONTE(janela_execucao)
    dados_stg = TRATAMENTO_PADRONIZACAO(dados_raw)
    ASSERT_QUALIDADE(dados_stg)
    CONCILIAR_TOTALIZADORES(dados_stg)
    dados_curated = APLICAR_REGRAS_NEGOCIO(dados_stg)
    LOAD_DW(dados_curated)
    BUILD_SEMANTIC_LAYER_E_CUBOS()
    PUBLICAR_DASHBOARDS()
    ATUALIZAR_LINHAGEM_CATALOGO()
    LOG("Fim pipeline com sucesso", DATA_ATUAL, AMBIENTE)
  CATCH erro
    LOG_ERRO("Falha pipeline", erro)
    NOTIFICAR(ALERTAS, "Falha no BI Orcamentario", erro)
    ABORTAR()
```

```
  END TRY
END PROCEDURE

// ---------- 2) INGESTAO (RAW) ----------
PROCEDURE INGESTAO_MULTIFONTE(janela_execucao) RETURNS Dict:
  VAR res = {}
  PARALLEL:
    res["ppa"] = INGESTAO_API_OU_CSV("MPO/PPA", janela_execucao, ZONAS.RAW)
    res["ldo"] = INGESTAO_API_OU_CSV("MPO/LDO", janela_execucao, ZONAS.RAW)
    res["loa"] = INGESTAO_API_OU_CSV("MPO/LOA", janela_execucao, ZONAS.RAW)
    res["empenhos"] = INGESTAO_API_OU_CSV("SIAFI/Empenhos",
                                          janela_execucao, ZONAS.RAW)
    res["liquidacoes"] = INGESTAO_API_OU_CSV("SIAFI/Liquidacoes",
                                          janela_execucao, ZONAS.RAW)
    res["pagamentos"] = INGESTAO_API_OU_CSV("SIAFI/Pagamentos",
                                          janela_execucao, ZONAS.RAW)
    res["restos_pagar"] = INGESTAO_API_OU_CSV("SIAFI/RestosPagar",
                                         janela_execucao, ZONAS.RAW)
    res["creditos_adicionais"] = INGESTAO_API_OU_CSV("SIAFI/Creditos",
                                         janela_execucao, ZONAS.RAW)
    res["tabelas_auxiliares"] = INGESTAO_DIM_AUXILIARES() // IBGE (UF,
                                         // municipio), CNAE,
                                         // natureza de despesa,
                                         // fonte de recurso,
                                         // UO/UG, programa, acao,
                                         // subtitulo
  END PARALLEL
  RETURN res
END PROCEDURE

PROCEDURE INGESTAO_API_OU_CSV(fonte, janela_execucao, zona_destino):
  CONFIG = CARREGAR_CONFIG_FONTE(fonte)
  IF CONFIG.tipo == "API":
     PAGINAR_E_PUXAR_API(CONFIG.endpoint, janela_execucao, zona_destino)
  ELSE IF CONFIG.tipo == "CSV" OR CONFIG.tipo == "PARQUET":
     LER_ARQUIVOS_BATCH(CONFIG.path, janela_execucao, zona_destino)
  END IF
  REGISTRAR_CHECKSUM_E_CONTAGEM(fonte, zona_destino)
  RETURN REFERENCIA(zona_destino, fonte)
END PROCEDURE

// ---------- 3) TRATAMENTO E PADRONIZACAO (STG) ----------
PROCEDURE TRATAMENTO_PADRONIZACAO(dados_raw) RETURNS Dict:
  VAR stg = {}
  FOR cada_tabela IN dados_raw:
    df = LER(cada_tabela)
    df = NORMALIZAR_SCHEMA(df,
```

```
          padrao = ["ano", "mes", "dia", "orgao", "uo", "ug", "funcao",
                    "subfuncao", "programa", "acao", "subtitulo", "nd",
                    "fonte_recurso", "pi", "empenho", "liquidacao",
                    "pagamento", "valor", "id_lancamento", "chave_natural",
                    "data_registro"])
    df = TRATAR_DATAS(df, tz="America/Sao_Paulo")
    df = PADRONIZAR_CODIGOS(df, zero_fill=[uo, ug, funcao, subfuncao,  nd,
                                                          fonte_recurso])
    df = LIMPAR_TEXTO(df, campos_texto=[orgao, programa, acao])
    df = DEDUPLICAR(df, chave=["chave_natural", "id_lancamento"])
    df = MARCAR_REGISTROS_INCONSISTENTES(df)
    SALVAR(df, ZONAS.STG, NOME(cada_tabela))
    stg[NOME(cada_tabela)] = REFERENCIA(ZONAS.STG, NOME(cada_tabela))
  END FOR
  RETURN stg
END PROCEDURE

// ---------- 4) QUALIDADE E CONCILIACAO ----------
PROCEDURE ASSERT_QUALIDADE(dados_stg):
  REGRAS = [
    REGRA_NOT_NULL(["ano","orgao","nd","valor"]),
    REGRA_DOMINIO("ano", intervalo=[2000..ANO(DATA_ATUAL)]),
    REGRA_VALOR_NAO_NEGATIVO("valor"),
    REGRA_CHAVE_UNICA(["id_lancamento"]),
    REGRA_REFERENCIAL(["uo","ug"] -> DIM_UO_UG),
    REGRA_CICLO_ORCAMENTARIO(ppa, ldo, loa)
  ]
  APLICAR_REGRAS(REGRAS, dados_stg)
  SE houver_erros_criticos() ENTAO
     NOTIFICAR(ALERTAS, "Qualidade: erros criticos", LISTAR_ERROS())
     ABORTAR()
  FIM
END PROCEDURE

// ---------- 5) REGRAS DE NEGOCIO (CURATED) ----------
PROCEDURE APLICAR_REGRAS_NEGOCIO(dados_stg) RETURNS Dict:
  VAR curated = {}
  curated["despesa_classificada"] = CLASSIFICAR_DESPESAS(
     base=dados_stg,
     regras=[
       MAPEAR_ND_PARA_ECONOMICA(),
       TAG_OBRIGATORIA_VS_DISCRICIONARIA(),
       TAG_INVESTIMENTO(), TAG_PESSOAL_BENEFICIOS(), TAG_CUSTEIO()
     ]
  )
  curated["calendario"] = GERAR_DIM_CALENDARIO_FISCAL(ano_min=2000,
                                                 ano_max=ANO(DATA_ATUAL))
```

```
  curated["ajustes"] = REGRAS_RATEIO_E_ELIMINACAO_DUPLICIDADE(dados_stg)
  curated["dimensoes_scd2"] = PREPARAR_SCD2(dados_stg)
  RETURN curated
END PROCEDURE

// ---------- 6) DATA WAREHOUSE ----------
PROCEDURE LOAD_DW(dados_curated):
  UPSERT_DIM("DIM_ORGAO", FONTE=dados_curated["despesa_classificada"],
             CHAVE=["orgao"])
  UPSERT_DIM("DIM_UO", FONTE=dados_curated["despesa_classificada"],
             CHAVE=["uo"])
  UPSERT_DIM("DIM_UG", FONTE=dados_curated["despesa_classificada"],
             CHAVE=["ug"])
  UPSERT_DIM("DIM_FUNCAO", FONTE=dados_curated["despesa_classificada"],
             CHAVE=["funcao","subfuncao"])
  UPSERT_DIM("DIM_PROGRAMA", FONTE=dados_curated["despesa_classificada"],
             CHAVE=["programa"])
  UPSERT_DIM("DIM_ACAO", FONTE=dados_curated["despesa_classificada"],
             CHAVE=["acao","subtitulo"])
  UPSERT_DIM("DIM_ND", FONTE=dados_curated["despesa_classificada"],
             CHAVE=["nd"])
  UPSERT_DIM("DIM_FONTE_REC", FONTE=dados_curated["despesa_classificada"],
             CHAVE=["fonte_recurso"])
  UPSERT_DIM("DIM_CALENDARIO", FONTE=dados_curated["calendario"],
             CHAVE=["data"])
  APLICAR_SCD2("DIM_ORGAO", HIST=dados_curated["dimensoes_scd2"])
  APLICAR_SCD2("DIM_UO")
  APLICAR_SCD2("DIM_UG")
  APPEND_FATO("FATO_EMPENHO", SELECT_STG("empenhos"),
              grain=["data","ug","nd","acao"])
  APPEND_FATO("FATO_LIQUIDACAO", SELECT_STG("liquidacoes"),
              grain=["data","ug","nd","acao"])
  APPEND_FATO("FATO_PAGAMENTO", SELECT_STG("pagamentos"),
              grain=["data","ug","nd","acao"])
  MERGE_FATO("FATO_RESTOS_PAGAR", SELECT_STG("restos_pagar"),
             grain=["ano_base","ug","nd","acao"])
  CRIAR_VISTA("VW_EXECUCAO_CONSOLIDADA",
    SQL="
      SELECT cal.ano, cal.mes, org.orgao, ug.ug, nd.nd, a.acao,
             SUM(e.valor) AS valor_empenhado,
             SUM(l.valor) AS valor_liquidado,
             SUM(p.valor) AS valor_pago
      FROM FATO_EMPENHO e
      LEFT JOIN FATO_LIQUIDACAO l USING (data, ug, nd, acao)
      LEFT JOIN FATO_PAGAMENTO p USING (data, ug, nd, acao)
      JOIN DIM_CALENDARIO cal ON e.data = cal.data
      JOIN DIM_UG ug ON e.ug = ug.ug
```

```
      JOIN DIM_ORGAO org ON ug.orgao_id = org.id
      JOIN DIM_ND nd ON e.nd = nd.nd
      JOIN DIM_ACAO a ON e.acao = a.acao
      GROUP BY 1,2,3,4,5,6
    "
  )
END PROCEDURE

// ---------- 7) SEMANTIC LAYER, KPIs E CUBOS ----------
PROCEDURE BUILD_SEMANTIC_LAYER_E_CUBOS():
  KPI dotacao_inicial = SUM(LOA.dotacao_inicial)
  KPI dotacao_atualizada = SUM(LOA.dotacao_inicial + creditos - anulacoes)
  KPI empenhado = SUM(FATO_EMPENHO.valor)
  KPI liquidado = SUM(FATO_LIQUIDACAO.valor)
  KPI pago = SUM(FATO_PAGAMENTO.valor)
  KPI saldo_disponivel = dotacao_atualizada - empenhado
  KPI restos_pagar_proc = SUM(RESTOS.processados)
  KPI restos_pagar_nproc = SUM(RESTOS.nao_processados)
  KPI execucao_pct = pago / NULLIF(dotacao_atualizada,0)
  METRICA lead_time_empenho_pagamento = DIAS_MEDIOS(data_pagamento -
                                                    data_empenho)
  METRICA previsao_sazonal = SARIMA(VW_EXECUCAO_CONSOLIDADA, chave=[orgao,
                                                nd], horizonte=6_meses)
  METRICA risco_subejecucao = CLASSIFICADOR(saldo_disponivel,sazonalidade,
                                                     historico)
  CUBO "Cubo Execucao" DIMENSOES=[Orgao,UO,UG,Programa,Acao,ND,Fonte,
                                                    Calendario]
                    MEDIDAS=[dotacao_inicial,dotacao_atualizada,empenhado,
                             liquidado,pago,saldo_disponivel,execucao_pct]
  CUBO "Cubo Restos" DIMENSOES=[Orgao,UG,ND,AnoBase]
                       MEDIDAS=[restos_pagar_proc,restos_pagar_nproc,
                                pago_no_ano_corrente]
END PROCEDURE

// ---------- 8) DASHBOARDS E PUBLICACAO ----------
PROCEDURE PUBLICAR_DASHBOARDS():
  EXPORTAR_PAINEL("Painel_Execucao_Geral", fonte="VW_EXECUCAO_CONSOLIDADA",
                  destino="BI-Server")
  PUBLICAR_DATASET_SEMANTICO("Orcamento_SemanticModel")
  ATUALIZAR_PERMISSOES(grupos=["Planejamento","Controle Interno",
                               "Gestores UO/UG"])
END PROCEDURE

// ---------- 9) GOVERNANCA, AUDITORIA E LINHAGEM ----------
PROCEDURE ATUALIZAR_LINHAGEM_CATALOGO():
  REGISTRAR_LINHAGEM(de=ZONAS.RAW, para=ZONAS.DW,
                     entidades=["empenhos","liquidacoes","pagamentos",
```

```
                                    "loa","restos_pagar"])
  AUDITAR_ACESSOS_DATASETS()
  ATUALIZAR_DICIONARIO_DADOS(campos_chave=[
     ("nd","Natureza de Despesa"), ("fonte_recurso","Fonte de Recurso"),
     ("acao","Acao Orcamentaria"), ("subtitulo","Subtitulo")
  ])
END PROCEDURE

// ---------- 10) INCREMENTAL, REPROCESSO E AGENDAMENTO ----------
SCHEDULE DIARIO as 05:00:
  PIPELINE_BI_ORCAMENTARIO(janela_execucao = JanelaCarga(inicio=ONTEM(),
                                                         fim=HOJE()))
SCHEDULE MENSAL_D0 as 06:00:
  REPROCESSAR_FECHAMENTO_MENSAL(competencia=MES_ANTERIOR())

PROCEDURE REPROCESSAR_FECHAMENTO_MENSAL(competencia):
  DEFINIR_JANELA(competencia.inicio, competencia.fim)
  PIPELINE_BI_ORCAMENTARIO(janela_execucao)

// ---------- 11) ALERTAS E EXCECOES DE NEGOCIO ----------
RULE ALERTA_SUBEJECAO:
  IF execucao_pct < 0.60 E MES_ATUAL >= SETEMBRO THEN
     NOTIFICAR(ALERTAS, "Risco de Subexecucao",
                 DETALHAR_POR(UG, ND, Programa))

RULE ALERTA_PICOS_DISCRICIONARIAS:
  IF DESVIO_PADRAO(mensal_discricionarias) > LIMIAR THEN
     ABRIR_INCIDENTE("Volatilidade Discricionaria", anexo=serie_temporal)

// ---------- 12) BACKUP E RECUPERACAO ----------
SCHEDULE SEMANAL_DOMINGO 02:00:
  BACKUP(areas=[ZONAS.RAW,ZONAS.STG,ZONAS.DW], retention="90 dias",
         offsite=true)
```

Listing 1: Algoritmo 1 – BI Orcamentario: Blueprint de Pseudocodigo End-to-End