UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO – CTC
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA
CURSO DE GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

Fernando Carlos Pereira Domenegato

EXTRAÇÃO E DECOMPOSIÇÃO DA SEÇÃO “OBJETO” EM ACORDOS DE NÃO PERSECUÇÃO CIVIL POR MEIO DE LLMs E RAG

Florianópolis, Santa Catarina – Brasil
2026

Fernando Carlos Pereira Domenegato

EXTRAÇÃO E DECOMPOSIÇÃO DA SEÇÃO “OBJETO” EM ACORDOS DE NÃO PERSECUÇÃO CIVIL POR MEIO DE LLMs E RAG

Trabalho de Conclusão de Curso submetido ao curso de Graduação em Sistemas de Informação do Centro Tecnológico da Universidade Federal de Santa Catarina como requisito parcial para a obtenção do título de Bacharel em Sistemas de Informação.
Orientadora: Prof.ª Carina Friedrich Dorneles, Dra.

Florianópolis, Santa Catarina – Brasil
2026

Ficha catalográfica cujo conteúdo foi gerado por meio de sistema automatizado gerenciado pela BU/UFSC, com dados inseridos pelo próprio autor.

Domenegato, Fernando Carlos Pereira

Extração e decomposição da seção “Objeto” em Acordos de Não Persecução Civil por meio de LLMs e RAG / Fernando Carlos Pereira Domenegato; Orientadora, Carina Friedrich Dorneles, 2026.
91 p.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal de Santa Catarina, Centro Tecnológico – CTC, Graduação em Sistemas de Informação, Florianópolis, Santa Catarina – Brasil, 2026.

Inclui referências

1. Sistemas de Informação, 2. Inteligência artificial, 3. Grandes modelos de linguagem, 4. Geração aumentada por recuperação, 5. Extração de informações em documentos jurídicos, I. Dorneles, Carina Friedrich. II. Universidade Federal de Santa Catarina. Graduação em Sistemas de Informação. III. Título.

Fernando Carlos Pereira Domenegato

EXTRAÇÃO E DECOMPOSIÇÃO DA SEÇÃO “OBJETO” EM ACORDOS DE NÃO PERSECUÇÃO CIVIL POR MEIO DE LLMs E RAG

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do título de Bacharel em Sistemas de Informação e aprovado em sua forma final pelo Curso de Graduação em Sistemas de Informação da Universidade Federal de Santa Catarina.

Florianópolis, Santa Catarina – Brasil, 09 de julho de 2026.

Prof. José Eduardo de Lucca, Me.
Coordenador do Curso de Graduação em
Sistemas de Informação

Banca Examinadora:

Prof.ª Carina Friedrich Dorneles, Dra.
Orientadora
Universidade Federal de Santa Catarina – UFSC

Prof. Renato Fileto, Dr.
Membro Titular
Universidade Federal de Santa Catarina – UFSC

Prof.ª Vitória Soares dos Santos, Ma.
Membro Titular
Universidade Federal de Santa Catarina – UFSC

Florianópolis, 2026.

Aos meus pais.

AGRADECIMENTOS

Agradeço ao Senhor Deus, onipresente e de infinita bondade, por me guiar e colocar em meu caminho pessoas que me ajudaram a persistir e a avançar.

À minha companheira de vida, Nakita Verônica Gheller, agradeço pelos laços de amor e admiração que nos unem, pela compreensão, paciência e apoio que sempre me oferece.

À minha orientadora, professora Carina Friedrich Dorneles, agradeço pela oportunidade de desenvolver este trabalho, pela orientação durante sua realização e pelas valiosas contribuições para o seu aprimoramento.

Agradeço à Universidade Federal de Santa Catarina (UFSC), ao corpo docente e aos servidores da Instituição por proporcionarem uma formação acadêmica de excelência.

Por fim, agradeço aos meus amados pais, Domingos e Iracema, por serem minhas eternas referências de bondade e resiliência.

RESUMO

O uso de técnicas de inteligência artificial tem se mostrado relevante para apoiar tarefas de análise documental na área jurídica, especialmente em contextos que envolvem documentos extensos, heterogêneos e com baixa padronização textual. Nesse cenário, este trabalho tem como objetivo desenvolver e avaliar uma abordagem baseada em grandes modelos de linguagem, do inglês *Large Language Models* (LLMs), e geração aumentada por recuperação, do inglês *Retrieval-Augmented Generation* (RAG), para a extração literal da seção denominada “Objeto” em documentos em formato PDF relacionados a Acordos de Não Persecução Civil. Em etapa complementar, a abordagem também é utilizada para decompor o conteúdo extraído em duas variáveis de interesse jurídico: o enquadramento legal do fato e a descrição do fato concreto, quando presentes. Para isso, foi organizada uma base documental oriunda do Ministério Público de Santa Catarina e construído um *dataset* de referência, manualmente curado e normalizado, utilizado para a avaliação dos experimentos. A solução foi implementada em Python e organizada como uma *pipeline* que combina extração do conteúdo textual dos arquivos PDF, recuperação semântica, uso de *prompts* especializados e execução de LLMs. Os resultados da extração da seção “Objeto” e da sua decomposição em variáveis de interesse foram comparados com o *dataset* de referência por meio de métricas de igualdade e de similaridade textual, complementadas por análise qualitativa de erros e casos divergentes. Os resultados indicaram elevada similaridade textual nas melhores configurações avaliadas, demonstrando a viabilidade da abordagem nas condições experimentais e no conjunto documental analisados. Ao todo, foram avaliados 530 documentos, dos quais 470 continham a seção “Objeto” no *dataset* de referência. Considerando apenas os documentos com conteúdo de referência para essa seção, a melhor configuração alcançou 99,11% de similaridade média na extração da seção “Objeto”, com 96,17% dos registros apresentando similaridade igual ou superior a 95%. Em uma avaliação binária complementar da melhor configuração sobre os 530 documentos, foram obtidos 96,23% de acurácia, 95,92% de precisão, 100% de sensibilidade e 66,67% de especificidade na identificação da presença ou ausência da seção. Os resultados devem ser interpretados como evidência exploratória restrita ao conjunto documental analisado, uma vez que não houve conjunto independente de teste.

Palavras-chave: inteligência artificial; grandes modelos de linguagem; geração aumentada por recuperação; extração de informações em documentos jurídicos.

ABSTRACT

The use of artificial intelligence techniques has become increasingly relevant for supporting document analysis tasks in the legal domain, especially in contexts involving lengthy, heterogeneous documents with limited textual standardization. In this context, this study aims to develop and evaluate an approach based on Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) for the literal extraction of the section entitled “Objeto” from PDF documents related to Civil Non-Prosecution Agreements. As a complementary step, the approach is also used to decompose the extracted content into two legally relevant variables: the legal classification of the conduct and the description of the underlying facts, when available. To this end, a document collection from the Public Prosecutor’s Office of the State of Santa Catarina was organized, and a manually curated and normalized reference dataset was constructed for evaluating the experiments. The solution was implemented in Python and structured as a pipeline that combines PDF text extraction, semantic retrieval, specialized prompts, and LLM execution. The results of the extraction of the “Objeto” section and its decomposition into the variables of interest were compared against the reference dataset using textual equality and textual similarity metrics, supplemented by qualitative analysis of errors and divergent cases. The results indicated high textual similarity in the best evaluated configurations, demonstrating the feasibility of the approach under the experimental conditions and within the analyzed document set. A total of 530 documents were evaluated, of which 470 contained the “Objeto” section in the reference dataset. Considering only the documents with reference content for this section, the best configuration achieved an average similarity of 99.11% in extracting the “Objeto” section, with 96.17% of the records presenting similarity equal to or greater than 95%. In a complementary binary evaluation of the best configuration over all 530 documents, the pipeline achieved 96.23% accuracy, 95.92% precision, 100% sensitivity, and 66.67% specificity in identifying the presence or absence of the section. The results should be interpreted as exploratory evidence restricted to the analyzed document set, since no independent test set was used.

Keywords: artificial intelligence; large language models; retrieval-augmented generation; information extraction in legal documents.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANPC	Acordo de Não Persecução Civil
GPU	*Graphics Processing Unit*
IA	Inteligência Artificial
JSON	*JavaScript Object Notation*
LLM	Grande Modelo de Linguagem
MPSC	Ministério Público de Santa Catarina
PDF	*Portable Document Format*
PLN	Processamento de Linguagem Natural
RAG	Geração Aumentada por Recuperação
REN	Reconhecimento de Entidades Nomeadas
UFSC	Universidade Federal de Santa Catarina
XLSX	Formato de planilha baseado em Office Open XML

SUMÁRIO

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO E MOTIVAÇÃO

A análise e o processamento de documentos jurídicos extensos e heterogêneos impõem desafios significativos que têm impulsionado a adoção de soluções baseadas em inteligência artificial (IA). Nesse cenário, destacam-se os Acordos de Não Persecução Civil (ANPCs)[1]. Adotado no âmbito do Ministério Público brasileiro, o Acordo de Não Persecução Civil (ANPC) é um instrumento jurídico consensual voltado à responsabilização por atos de improbidade administrativa, à reparação de danos ao erário e à solução pactuada com agentes públicos ou terceiros. A formalização desses acordos, no entanto, resulta em documentos com significativa variabilidade redacional e estrutural. Tais características tornam a extração e a análise manual de suas informações tarefas trabalhosas e suscetíveis a inconsistências, justificando a necessidade de abordagens automatizadas de extração de conhecimento.

Dentre os elementos que compõem um ANPC, a seção denominada “Objeto”, analisada neste trabalho, pode descrever os fatos concretos das condutas investigadas, os agentes envolvidos, o enquadramento legal do fato e, em alguns casos, sanções ou obrigações pactuadas. Essa seção funciona, portanto, como uma síntese narrativa do acordo, podendo reunir em um mesmo bloco textual informações centrais para a compreensão do caso e de seus efeitos jurídicos. A identificação automatizada das informações contidas nessa seção pode contribuir para facilitar consultas, análises, auditorias e o acompanhamento do cumprimento dos acordos.

A extração da seção “Objeto” é mais complexa do que a identificação de campos estruturados, como datas, números processuais ou valores monetários. Por não ser definida de forma explícita nos normativos que regulamentam o ANPC, essa seção apresenta variação de redação entre diferentes documentos, sobretudo quando elaborados por promotorias de justiça distintas. Assim, a tarefa não consiste apenas em capturar um valor pontual, mas em localizar o início e o fim de uma unidade semântica cuja extensão pode variar e cujo conteúdo pode se confundir com cláusulas de obrigação, pagamento ou fundamentação jurídica. Embora a localização visual do “Objeto” seja simples quando essa denominação aparece expressamente como seção no documento, a ausência de padronização textual e estrutural torna mais complexas a extração e a identificação automáticas de suas informações.

Nesse contexto, tecnologias de IA, como os grandes modelos de linguagem (LLMs) e as técnicas baseadas em geração aumentada por recuperação (RAG),

[1] Neste trabalho, utiliza-se a forma “Acordo de Não Persecução Civil” para se referir ao instituto previsto no artigo 17-B da Lei n. 8.429/1992. A forma “Acordo de Não Persecução Cível” é mantida apenas quando reproduz o rótulo original anotado na curadoria ou o texto dos *prompts* efetivamente utilizados nos experimentos.

apresentam-se como recursos importantes para apoiar a automatização da leitura e da extração de informações em documentos jurídicos. A extração automática do conteúdo da seção “Objeto” em ANPCs constitui o foco principal deste trabalho. Além dessa extração, este estudo também realiza experimentos de decomposição da seção extraída, com o objetivo de identificar duas variáveis de interesse: o enquadramento legal do fato e a descrição do fato concreto. Neste trabalho, a expressão “enquadramento legal do fato” refere-se à capitulação jurídica da conduta descrita na seção “Objeto”, enquanto “descrição do fato concreto” refere-se à narrativa da conduta atribuída ao investigado ou compromissário. Após essa definição, utilizam-se também as formas abreviadas “enquadramento legal” e “descrição do fato” com o mesmo sentido.

Este trabalho integra-se ao Projeto Céos, executado pela Universidade Federal de Santa Catarina (UFSC) em parceria e com financiamento do Ministério Público de Santa Catarina (MPSC). Conforme declarado em seu portal institucional, o objetivo geral do Projeto Céos é “o estudo, desenvolvimento e implementação de fluxos de trabalho (*workflows*) para coleta, integração, organização, processamento e análise de dados voltados à extração de conhecimento de forma automatizada ou semi-automatizada, visando apoiar a tomada de decisão nos processos inerentes às atividades do MPSC” (Projeto Céos, 2026).

1.2 PROBLEMA DE PESQUISA

A tarefa investigada envolve dois desafios encadeados. O primeiro consiste em localizar e extrair literalmente a seção “Objeto” em documentos que apresentam variações de redação, estrutura interna e forma de apresentação. O segundo consiste em decompor o conteúdo extraído em duas variáveis de interesse jurídico: o enquadramento legal e a descrição do fato.

A dificuldade não se limita à localização do termo “Objeto”. Nos documentos analisados, a seção pode ser introduzida por títulos como “DO OBJETO” ou “CLÁUSULA PRIMEIRA – DO OBJETO”, iniciar na mesma linha do título ou ser apresentada diretamente por expressões como “O presente acordo tem por objeto...”. Além disso, sua extensão pode abranger mais de uma página e conter menções a obrigações, pagamentos, ressarcimento ou homologação, sem que essas expressões indiquem necessariamente o início de outra seção. A tarefa exige, portanto, identificar contextualmente o início, a continuidade e o término do bloco textual correspondente.

Na abordagem proposta, a recuperação semântica busca restringir o espaço de busca ao selecionar trechos potencialmente relevantes de cada documento. O LLM, por sua vez, interpreta o contexto recuperado para delimitar a seção, considerar possíveis sobreposições e continuidades entre fragmentos e produzir uma saída estruturada. Essa distribuição de funções não pressupõe que métodos determinísticos sejam in-

capazes de realizar a tarefa, mas delimita o papel de cada componente na *pipeline* avaliada neste trabalho.

Diante desse contexto, a questão que orienta o trabalho é: em que medida uma *pipeline* baseada em RAG e LLMs é capaz de extrair, com fidelidade textual, a seção "Objeto" de documentos relacionados a ANPCs e, nos registros previamente definidos como elegíveis no *dataset* ouro, decompô-la em enquadramento legal e descrição do fato?

1.3 OBJETIVOS (GERAL E ESPECÍFICOS)

O objetivo geral deste trabalho é desenvolver e avaliar, com base em um *dataset* de referência, uma abordagem baseada em LLMs e RAG para extrair, de documentos em formato PDF relacionados a ANPCs, o conteúdo literal da seção "Objeto" e, em etapa complementar, decompor o texto extraído em enquadramento legal e descrição do fato, quando presentes.

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

1. organizar uma base documental composta por documentos em formato PDF relacionados a ANPCs;
2. construir, curar e normalizar um *dataset* de referência para a avaliação dos experimentos;
3. implementar uma *pipeline* que integre extração e pré-processamento textual, segmentação em fragmentos, geração de *embeddings*, recuperação semântica e execução de LLM;
4. desenvolver e refinar *prompts* especializados para a extração literal da seção "Objeto" e para sua decomposição em variáveis de interesse;
5. comparar os textos extraídos automaticamente com o *dataset* de referência por meio de métricas de igualdade e similaridade textual e, na configuração de maior fidelidade textual, avaliar o desempenho da *pipeline* na classificação binária da presença ou ausência da seção "Objeto", por meio de acurácia, precisão, sensibilidade e especificidade;
6. avaliar a decomposição da seção "Objeto" em enquadramento legal e descrição do fato com base nos valores do *dataset* de referência;
7. analisar qualitativamente os principais padrões de divergência observados nas saídas dos modelos.

1.4 CONTRIBUIÇÕES

A principal contribuição deste trabalho está na implementação e na avaliação de uma *pipeline* específica para a extração literal da seção “Objeto” em ANPCs e para a decomposição do conteúdo extraído em variáveis de interesse juridicamente relevantes. A tarefa investigada envolve a delimitação de um bloco textual narrativo, a preservação de sua redação original e a identificação de informações específicas presentes no texto.

Além da aplicação ao acervo documental do MPSC, o trabalho apresenta uma contribuição metodológica ao avaliar um fluxo modular que distingue duas etapas: a localização e preservação da unidade textual relevante e a posterior análise de seu conteúdo para identificação de dados juridicamente significativos. Essa arquitetura oferece um desenho metodológico potencialmente adaptável a outras seções documentais ou a outros tipos de documentos jurídicos, desde que submetido a nova curadoria e validação específica.

Como contribuições complementares, este trabalho organiza e cura um *dataset* de referência para avaliação, compara diferentes modelos, *prompts* e configurações de recuperação semântica, e analisa quantitativa e qualitativamente as saídas produzidas. Esses resultados fornecem evidências sobre as possibilidades e as limitações do uso combinado de LLMs e RAG no domínio documental estudado.

1.5 ESTRUTURA DO TRABALHO

Além desta introdução, o trabalho está organizado em cinco capítulos. O Capítulo 2 apresenta a fundamentação teórica sobre os ANPCs, a extração de informação em documentos jurídicos, os grandes modelos de linguagem, a geração aumentada por recuperação e a engenharia de prompts. O Capítulo 3 discute os trabalhos relacionados. O Capítulo 4 descreve o desenho da pesquisa e a *pipeline* experimental, a construção do *dataset* de referência, a estratégia de extração baseada em RAG e os critérios e métricas de avaliação. O Capítulo 5 apresenta a configuração experimental e discute os resultados da extração e da decomposição da seção “Objeto”. Por fim, o Capítulo 6 reúne as conclusões, as limitações e as possibilidades de trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os fundamentos jurídicos e tecnológicos que sustentam a abordagem desenvolvida neste trabalho. Inicialmente, contextualiza-se o ANPC e a seção denominada “Objeto”. Em seguida, discutem-se a extração de informação em documentos jurídicos e o uso de grandes modelos de linguagem nesse domínio. Por fim, são apresentados os conceitos de geração aumentada por recuperação e engenharia de prompts, que orientam a arquitetura implementada.

2.1 O CONTEXTO DOS ANPCs NO DIREITO BRASILEIRO

Os ANPCs são instrumentos jurídicos voltados à resolução consensual de casos relacionados à improbidade administrativa. Celebrados entre o Ministério Público e os investigados por atos dolosos de improbidade (Brasil, 2021), devem observar o interesse público, a proteção do patrimônio público e a moralidade administrativa (Conselho Nacional do Ministério Público, 2025). Previsto na Lei nº 8.429/1992 (Brasil, 1992), com as alterações introduzidas pela Lei nº 14.230/2021 (Brasil, 2021), e regulamentado pela Resolução nº 306/2025 do Conselho Nacional do Ministério Público (Conselho Nacional do Ministério Público, 2025), o instituto possibilita a reparação do dano ao erário e permite ao compromissário evitar o prosseguimento da persecução civil. Trata-se, assim, de um mecanismo que busca equilibrar eficiência na resolução de conflitos, responsabilização proporcional e segurança jurídica.

Nos termos do artigo 6º da referida Resolução (Conselho Nacional do Ministério Público, 2025), o acordo deve ser formalizado por escrito e conter elementos mínimos obrigatórios, entre os quais se destacam: a identificação das pessoas envolvidas e de seus vínculos; a narrativa circunstanciada da conduta, com indicação de tempo, local e enquadramento jurídico; a quantificação do dano e das vantagens indevidas; a definição das sanções aplicáveis; e os compromissos e garantias relacionados à execução do acordo. O artigo 7º admite, ainda, a inclusão de cláusulas adicionais opcionais, como reparação de dano moral coletivo, medidas de integridade (*compliance*) e negócios jurídicos processuais.

Embora a regulamentação não estabeleça uma seção formal denominada “Objeto”, os primeiros incisos do artigo 6º reúnem os elementos que, na prática, delimitam a matéria do acordo, como a narrativa dos fatos, o enquadramento legal e a extensão do dano. Os documentos analisados frequentemente concentram essas informações sob o título de “Objeto”. A identificação correta dessa parte do documento é essencial para apoiar consultas posteriores, auditorias, estudos estatísticos e demais atividades de análise documental no âmbito do Ministério Público.

Neste trabalho, o termo “Objeto” designa esse bloco textual, cuja extração literal e posterior decomposição em variáveis de interesse constituem o foco empírico da

pesquisa.

2.2 IA NA ÁREA JURÍDICA: LLMs E EXTRAÇÃO DE INFORMAÇÃO

A extração de informação busca identificar e organizar conteúdos relevantes presentes em textos não estruturados. No domínio jurídico, essa tarefa enfrenta dificuldades relacionadas à linguagem técnica, à sintaxe complexa, ao uso recorrente de referências normativas e à alta variabilidade redacional entre documentos (Ariai; Mackenzie; Demartini, 2025). Essas características tornam difícil a aplicação de regras fixas a acervos produzidos por diferentes órgãos, autores e contextos institucionais.

Abordagens tradicionais de extração jurídica empregam correspondência de padrões e expressões regulares (Kowsrihawat; Vateekul, 2015; Cheng et al., 2009). Embora úteis em documentos previsíveis e padronizados, esses métodos apresentam menor flexibilidade diante de ambiguidades e mudanças de estrutura (Aquino et al., 2024). Como consequência, a extração manual ou semiestruturada de contratos e atos oficiais permanece um processo oneroso, suscetível a falhas humanas e fortemente dependente de especialistas no domínio (Aejas et al., 2024; Hassan; Le, 2020). Diante dessas limitações, técnicas de Processamento de Linguagem Natural (PLN) e aprendizado profundo ampliaram as possibilidades de automação, permitindo tarefas como classificação textual e Reconhecimento de Entidades Nomeadas (REN) em domínios jurídicos específicos (Mota, 2023; Aejas et al., 2024).

Nesse cenário, os grandes modelos de linguagem, do inglês *Large Language Models* (LLMs), destacam-se por sua capacidade de processar, compreender e gerar linguagem natural de forma contextual, podendo desempenhar tarefas de compreensão e extração textual mesmo sem ajuste fino adicional específico para cada tarefa (Brown et al., 2020; Wei et al., 2024). Fundamentados na arquitetura Transformer, esses modelos utilizam mecanismos de autoatenção (*self-attention*) para processar o contexto de entrada (Vaswani; Shazeer; Parmar et al., 2017). Na geração autorregressiva, o modelo estima a probabilidade do próximo token a partir da sequência anterior, o que viabiliza a produção de respostas condicionadas às instruções e ao conteúdo fornecido (Jurafsky; Martin, 2026).

A aplicação da tecnologia de LLM à extração jurídica oferece maior flexibilidade para interpretar variações linguísticas e estruturais. Entretanto, o desempenho do modelo pode ser degradado quando a informação relevante está dispersa em contextos longos (Liu et al., 2023), além de haver o risco de alucinações — isto é, a produção de respostas não fundamentadas no documento de origem (Huang et al., 2025). Por isso, seu uso em tarefas que exigem fidelidade textual demanda mecanismos de controle do contexto (como a recuperação semântica), instruções especializadas (*prompts*) e validação da saída.

Métricas de similaridade textual baseadas na forma da redação comparam se-

quências textuais com base em caracteres, tokens ou subsequências coincidentes, sem avaliar diretamente a equivalência semântica entre os textos. Elas são adequadas quando o objetivo é avaliar a fidelidade de reprodução textual, pois permitem estimar o grau de proximidade entre a saída produzida pelo modelo e um trecho de referência, especialmente em tarefas que exigem preservação da redação originalmente registrada (Jurafsky; Martin, 2026).

Neste trabalho, a arquitetura baseada em LLM é utilizada em duas etapas. Na primeira, o modelo recebe trechos recuperados do documento e retorna o intervalo textual correspondente à seção “Objeto”. Na segunda, recebe apenas a seção “Objeto” extraída e realiza sua decomposição nas variáveis de interesse. Para a execução dessas tarefas, foram adotadas as famílias de modelos Gemma 3 (Gemma Team, 2025), Gemma 4 (Google DeepMind, 2026), Qwen 2.5 (Qwen et al., 2025), Qwen 3.6 (Yang et al., 2025) e Llama 3.3 (Dubey et al., 2024). Assim, a solução combina a extração literal de um bloco textual com a identificação posterior de variáveis de interesse específicas.

2.3 RAG E ENGENHARIA DE PROMPTS

A geração aumentada por recuperação, do inglês *Retrieval-Augmented Generation* (RAG), conforme formalizada por Lewis et al. (2020), consiste em uma arquitetura híbrida que integra componentes de memória paramétrica, representada pelos pesos de um modelo de linguagem pré-treinado, e de memória não paramétrica, representada por uma coleção de documentos acessada por um mecanismo de recuperação. Nessa arquitetura, trechos considerados relevantes são selecionados antes da geração e fornecidos ao LLM como contexto adicional.

O uso de RAG busca ancorar a resposta no conteúdo recuperado, reduzindo a dependência exclusiva do conhecimento armazenado nos parâmetros do modelo. Essa estratégia pode mitigar, embora não eliminar, respostas não aderentes ao texto de origem e contribui para lidar com a limitação da janela de contexto, pois prioriza os fragmentos mais relevantes para a tarefa (Aquino et al., 2024). A abordagem também favorece a rastreabilidade, uma vez que permite registrar os trechos enviados ao modelo durante cada execução.

Em sistemas de recuperação semântica, *embeddings* representam textos como vetores numéricos em um espaço vetorial. A proximidade entre a consulta e os fragmentos pode ser estimada por medidas como a similaridade do cosseno, que compara a orientação entre os vetores e permite recuperar os trechos semanticamente mais relevantes para a tarefa (Jurafsky; Martin, 2026; Aquino et al., 2024).

Neste estudo, a base de recuperação é formada pelos fragmentos do próprio PDF em processamento. Após a extração e a segmentação do texto, são gerados embeddings desses fragmentos, e os mais próximos da consulta são selecionados

para compor o contexto da extração. Portanto, o RAG é empregado na localização da seção “Objeto”, e não como fonte externa de conhecimento jurídico.

A interação com os LLMs ocorre por meio de prompts, isto é, instruções textuais que especificam a tarefa, as restrições e o formato esperado da resposta (Jurafsky; Martin, 2026). A engenharia de prompts corresponde ao processo de formular, testar e refinar essas instruções para melhorar o comportamento do modelo em uma tarefa específica (Jurafsky; Martin, 2026).

No trabalho, foram desenvolvidos prompts distintos para a extração literal da seção “Objeto” e para sua decomposição. Essa divisão em duas chamadas ao LLM favorece a controlabilidade da saída, reduz a complexidade de cada etapa e permite avaliar separadamente erros de delimitação textual e de identificação das variáveis. As instruções definem os critérios de elegibilidade do trecho, os limites da seção, as regras de preservação ou normalização dos campos e o formato *JavaScript Object Notation* (JSON) esperado. O refinamento dessas instruções fez parte do desenho experimental, pois mudanças no prompt podem alterar a delimitação do texto, a aderência ao formato de saída e a qualidade dos elementos identificados.

3 TRABALHOS RELACIONADOS

A aplicação de técnicas de inteligência artificial no domínio jurídico tem crescido de forma significativa em diferentes atividades, como a análise de contratos públicos, a extração de informações de documentos e a análise automática de peças processuais. Nesse contexto, estudos voltados à análise automatizada de contratos, à identificação de padrões textuais e à extração de informações jurídicas oferecem referências relevantes para a presente pesquisa. Este capítulo está estruturado em quatro subseções: abordagens clássicas e estatísticas, reconhecimento de entidades nomeadas e supervisão fraca, soluções baseadas em LLMs e RAG, e uma análise comparativa dos estudos em relação ao escopo deste TCC.

3.1 ABORDAGENS CLÁSSICAS E ESTATÍSTICAS EM DOCUMENTOS JURÍDICOS

As abordagens clássicas de PLN aplicadas a documentos jurídicos incluem métodos baseados em regras, expressões regulares, vetorização textual e classificadores estatísticos tradicionais. Esses métodos constituíram uma base importante para tarefas de classificação e extração de informação, especialmente em cenários nos quais padrões textuais são relativamente estáveis. No entanto, sua aplicação tende a depender de regras definidas por especialistas, de etapas de pré-processamento e de ajustes específicos ao domínio, o que pode limitar sua flexibilidade diante de documentos com grande variabilidade redacional e estrutural.

Uma abordagem representativa é apresentada por Hassan e Le (2020), que propuseram um método baseado em técnicas de PLN e aprendizado de máquina para classificar sentenças de contratos de construção civil em duas categorias: requisitos e não requisitos. O estudo compara estratégias baseadas em regras, como a detecção de termos obrigacionais, com algoritmos tradicionais de aprendizado de máquina, incluindo *Naïve Bayes*, Regressão Logística e Máquinas de Vetores de Suporte (SVM). Os resultados demonstraram que os modelos SVM atingiram precisão de até 98% na classificação das sentenças, especialmente quando combinados com vetorização *Bag-of-Words* e *Word2Vec*. Embora relevantes, tais resultados limitam-se à classificação de sentenças curtas e pouco variáveis.

No contexto brasileiro, destaca-se a dissertação de Lima (2021), que aborda a detecção automática de riscos de conluio em licitações publicadas no Diário Oficial da União. A pesquisa apresenta uma *pipeline* de pré-processamento em grande escala, empregando técnicas de vetorização, como TF-IDF e *Word2Vec*, para a representação textual. O autor utilizou classificadores como *Passive-Aggressive*, redes neurais recorrentes (Bi-LSTM) e outros modelos supervisionados para identificar padrões associados a possíveis indícios de fraude, atingindo F1-Score de 95,1%. Para este trabalho, a contribuição de Lima (2021) está sobretudo na demonstração da viabilidade de PLN

em língua portuguesa e em grandes volumes de documentos públicos, ainda que em uma tarefa voltada à classificação probabilística de risco, e não à extração literal de blocos textuais.

Esses estudos evidenciam que regras, vetorização e classificadores tradicionais podem alcançar bons resultados em tarefas delimitadas. Contudo, também indicam limitações relevantes para o problema tratado neste TCC: a necessidade de adaptação manual ao domínio, a dificuldade de lidar com redações heterogêneas e a concentração em sentenças, classes ou indicadores específicos, em vez da delimitação de unidades narrativas extensas.

3.2 RECONHECIMENTO DE ENTIDADES NOMEADAS E SUPERVISÃO FRACA

Outra linha relevante de trabalhos jurídicos concentra-se no reconhecimento de entidades nomeadas (REN), voltado à identificação de elementos como pessoas, organizações, datas, valores, referências contratuais e outros termos de interesse. Em comparação com métodos clássicos, modelos de aprendizado profundo e arquiteturas baseadas em transformadores ampliaram a capacidade de representação contextual dos textos, mas mantiveram como desafio a necessidade de corpora anotados de forma consistente.

Avanços nessa direção são apresentados por Aejas et al. (2024), que investigaram o uso de técnicas de REN para automatizar a análise de contratos legais norte-americanos e identificar entidades relevantes, como partes, datas, valores e referências contratuais. O estudo avaliou diferentes arquiteturas, incluindo BERT, RoBERTa e uma variante especializada denominada Contracts-BERT, além de combinações com camadas BiLSTM e *Conditional Random Fields* (CRF). Os resultados mostram que modelos pré-treinados em domínios jurídicos superam abordagens tradicionais na identificação de elementos contratuais relevantes, representando um salto qualitativo em relação a técnicas menos sensíveis ao contexto.

Ainda no cenário nacional, o trabalho de Mota (2023) buscou mitigar a dificuldade e o alto custo da anotação manual de documentos jurídicos governamentais aplicando técnicas de supervisão fraca. Utilizando publicações de licitações e contratos do Distrito Federal, a autora demonstrou que a combinação de funções de rótulo, baseadas em conhecimento heurístico e correspondência de palavras, com arquiteturas Bi-LSTM-CNN e CRF obteve resultados altamente competitivos na extração de entidades nomeadas, alcançando um F1-Score de até 96% na extração de dados de contratos. O estudo comprova a viabilidade do PLN em língua portuguesa e propõe uma alternativa ao custo da anotação manual.

Apesar de sua relevância, os trabalhos de REN concentram-se principalmente na rotulação de entidades curtas e delimitáveis, como datas, valores, partes e referências. O problema investigado neste TCC possui granularidade distinta: a seção “Objeto”

de um ANPC não corresponde a um termo isolado, mas a um bloco narrativo que pode reunir fatos, enquadramento jurídico, agentes envolvidos, sanções e obrigações pactuadas. Assim, a tarefa vai além da extração de entidades, pois exige delimitar e preservar uma unidade textual longa antes de decompô-la em variáveis jurídicas específicas.

3.3 LLMs E RAG NA EXTRAÇÃO DOCUMENTAL

Mais recentemente, a extração de informações em documentos jurídicos passou a incorporar modelos generativos de linguagem e mecanismos de recuperação contextual. Nesse paradigma, LLMs podem ser instruídos por *prompts* a localizar, resumir ou estruturar informações, enquanto arquiteturas de RAG recuperam trechos semanticamente relevantes para compor o contexto da resposta. Essa combinação busca mitigar limitações de janela de contexto, reduzir respostas sem apoio no documento analisado e aumentar a flexibilidade diante de variações redacionais.

Nesse grupo, Aquino et al. (2024) propuseram um *workflow* para extração de informações de documentos jurídicos brasileiros relacionados a fraudes em processos de contratação pública, a partir de documentos fornecidos pelo Ministério Público de Santa Catarina. A abordagem utilizou LLMs integrados a uma arquitetura RAG e avaliou diferentes configurações experimentais, incluindo modelos de linguagem, sobreposição de *chunks* e variação do número de fragmentos recuperados, correspondente ao parâmetro *top-k*. O estudo reportou desempenho superior ao de uma estratégia baseada em expressões regulares e indicou a relevância do ajuste de parâmetros em pipelines RAG aplicadas a documentos jurídicos longos.

A pesquisa de Aquino et al. (2024) estabelece a base metodológica mais próxima ao escopo aqui proposto, ao apresentar resultados sobre o uso combinado de RAG e LLM na extração de informações em acervo de documentos jurídicos. Em alinhamento com essa abordagem, o presente trabalho explora a aplicação dessa arquitetura em um novo nível de granularidade textual, dedicando-se à delimitação e à extração literal de blocos longos, especificamente a seção completa do “Objeto” do ANPC. De forma complementar, este estudo insere uma segunda etapa analítica na *pipeline*, instruindo o LLM a decompor o texto extraído para a identificação de variáveis de interesse jurídico.

3.4 ANÁLISE COMPARATIVA

A Tabela 1 sintetiza os trabalhos analisados, destacando o domínio de aplicação, a tarefa investigada, os métodos utilizados, as métricas ou resultados reportados no recorte analisado, bem como a granularidade da extração e a lacuna observada em relação ao presente TCC.

Tabela 1 – Comparação entre os trabalhos relacionados e o presente TCC.

Trabalho	**Corpus/domínio e tarefa**	**Método**	**Métrica ou resultado citado**	**Granularidade da extração e lacuna em relação a este TCC**
Hassan e Le (2020)	Contratos de construção civil; classificação de sentenças em requisitos e não requisitos.	Regras, *Naïve Bayes*, Regressão Logística, SVM, *Bag-of-Words* e *Word2Vec*.	Precisão de até 98%.	Atua sobre sentenças curtas. Não trata da extração literal de blocos narrativos longos nem da decomposição posterior de informações jurídicas.
Lima (2021)	Licitações publicadas no Diário Oficial da União; detecção de risco de conluio.	Pré-processamento em larga escala, TF-IDF, *Word2Vec*, *Passive-Aggressive*, Bi-LSTM e modelos supervisionados.	F1-Score de 95,1%.	Classifica documentos ou publicações por risco. Não realiza a delimitação e a preservação de uma seção textual específica.
Aejas et al. (2024)	Contratos legais norte-americanos; reconhecimento de entidades contratuais.	BERT, RoBERTa, Contracts-BERT, BiLSTM e CRF.	Sem valor numérico citado.	Extrai entidades curtas, como partes, datas, valores e referências. Não se concentra na extração de uma unidade narrativa extensa.
Mota (2023)	Licitações e contratos do Distrito Federal; reconhecimento de entidades nomeadas com menor dependência de anotação manual.	Supervisão fraca, funções de rótulo, Bi-LSTM-CNN e CRF.	F1-Score de até 96%.	Atua sobre entidades nomeadas em documentos governamentais. Mitiga o custo de anotação, mas permanece voltado à rotulação de entidades curtas.
Aquino et al. (2024)	Documentos jurídicos brasileiros relacionados a irregularidades em contratações públicas; extração do identificador do processo licitatório e do município da irregularidade.	LLMs integrados a RAG, com variação de *chunks*, sobreposição e *top-k*; comparação com expressões regulares.	Acurácia média aproximada de 90%.	Extrai variáveis curtas a partir de trechos recuperados. Aproxima-se da arquitetura deste TCC, mas não se concentra na delimitação literal de uma seção narrativa completa seguida de decomposição.
Este trabalho	530 documentos do acervo analisado do MPSC; extração literal da seção "Objeto" e decomposição em enquadramento legal e descrição do fato.	RAG intradocumental, LLMs, prompts especializados, persistência e avaliação textual e binária.	99,11% de similaridade média nos 470 documentos com referência; 96,23% de acurácia na avaliação binária complementar sobre 530 documentos, na melhor configuração.	Extrai um bloco narrativo longo antes de decompô-lo em variáveis jurídicas; os resultados são exploratórios e restritos ao conjunto avaliado.

Nota: os resultados apresentados não são diretamente comparáveis entre si, pois foram obtidos em conjunto documental, tarefas, protocolos e métricas diferentes. Esta tabela tem finalidade descritiva e busca situar a granularidade e o método de cada trabalho.

Fonte: Elaborada pelo autor.

A comparação indica uma evolução das abordagens aplicadas a documentos jurídicos. Os métodos clássicos e estatísticos oferecem uma base importante, mas tendem a depender de padrões textuais mais estáveis e de maior intervenção especializada. Os trabalhos com REN e aprendizado profundo ampliam a capacidade de identificação contextual, especialmente para entidades curtas, enquanto a supervisão fraca busca reduzir o custo da anotação manual. Por sua vez, arquiteturas com LLMs e RAG introduzem maior flexibilidade para lidar com documentos longos e heterogêneos.

Diante desse panorama, o presente trabalho diferencia-se por investigar uma tarefa de granularidade intermediária e mais complexa do que a extração de entidades isoladas: a delimitação e extração literal de um bloco narrativo longo, a seção "Objeto" de ANPCs, seguida da decomposição desse conteúdo em variáveis jurídicas. Essa combinação permite preservar a redação original da seção e, ao mesmo tempo, estruturar informações relevantes para análise posterior.

4 METODOLOGIA

4.1 DESENHO DA PESQUISA E PIPELINE EXPERIMENTAL

O estudo foi conduzido sobre 530 documentos PDF, cada um tratado como uma unidade de análise. O objetivo metodológico foi avaliar uma pipeline automatizada para duas tarefas encadeadas: a extração literal da seção denominada “Objeto” em ANPCs e a decomposição do conteúdo extraído em duas variáveis de interesse jurídico, o enquadramento legal e a descrição do fato. As saídas produzidas foram comparadas com um dataset ouro por meio de métricas de fidelidade textual, avaliação binária da presença ou ausência da seção, indicadores de execução e inspeção qualitativa dos casos divergentes.

A execução metodológica foi organizada em quatro etapas:

1. seleção, curadoria e normalização do dataset de referência, denominado dataset ouro;
2. pré-processamento textual e recuperação semântica dos trechos relevantes;
3. extração da seção “Objeto” e decomposição do conteúdo extraído;
4. persistência e avaliação dos resultados.

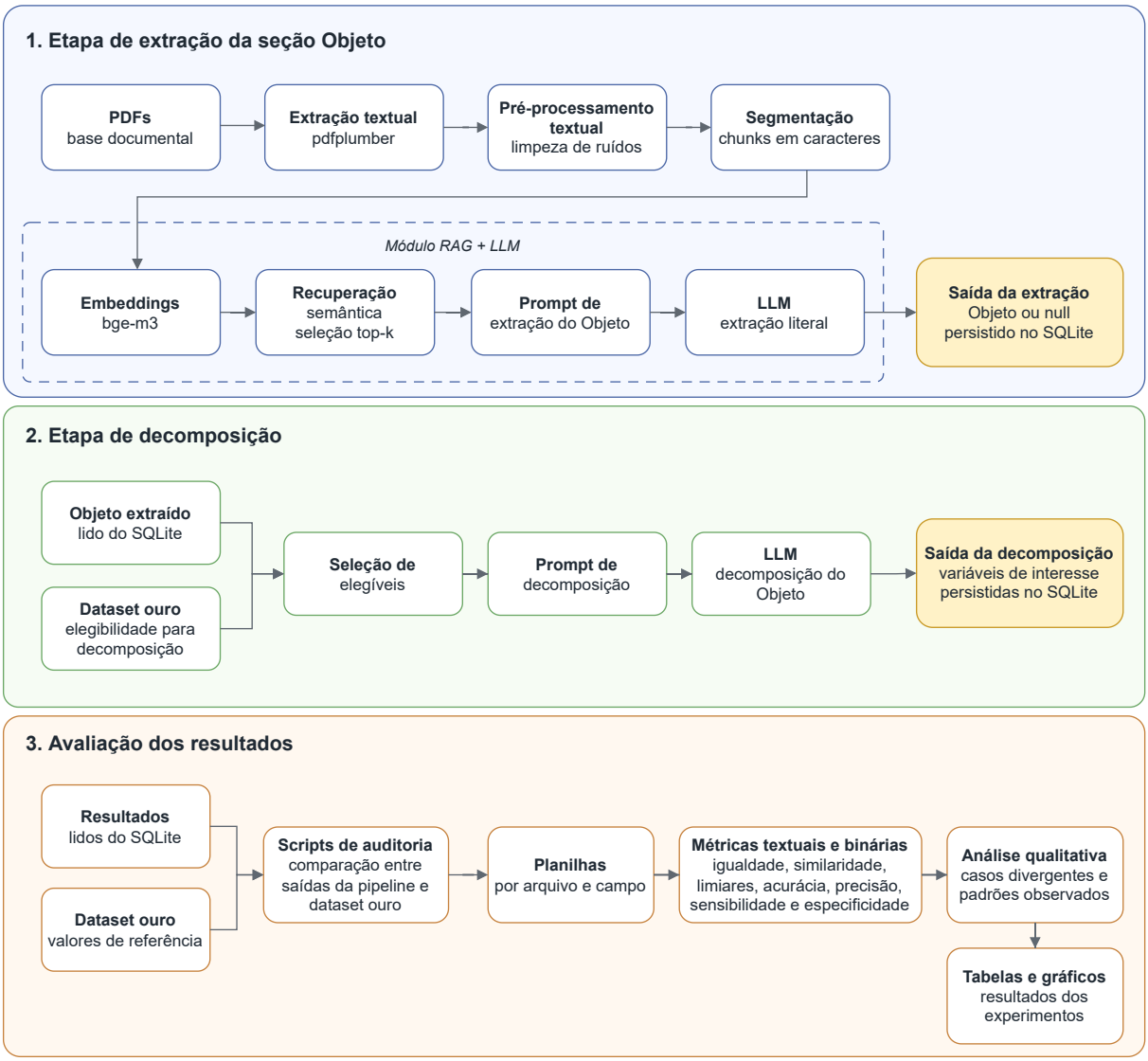
As configurações experimentais combinaram o modelo de linguagem, a versão do prompt, os parâmetros de recuperação semântica e os parâmetros de geração. Nas execuções mais recentes, também foram avaliadas configurações base e ampliada de recuperação semântica. Os valores de referência individuais do dataset ouro não foram fornecidos aos modelos durante a execução da pipeline; apenas a indicação de elegibilidade foi utilizada para selecionar os documentos submetidos à etapa de decomposição.

A Figura 1 apresenta a visão geral da pipeline. Após o pré-processamento dos documentos e a recuperação dos trechos semanticamente relevantes, o LLM foi acionado para extrair a seção “Objeto”. Em seguida, o texto extraído foi submetido a uma nova chamada do modelo, restrita ao conteúdo da própria seção, para identificação do enquadramento legal e da descrição do fato, conforme os critérios definidos na curadoria do dataset ouro. Os dados intermediários, os parâmetros e as respostas dos modelos foram registrados em banco SQLite, permitindo vincular cada saída ao documento de origem, ao prompt, ao modelo e aos parâmetros empregados. Desse modo, o SQLite funcionou não apenas como mecanismo de persistência, mas também como apoio à rastreabilidade, à auditoria e à governança dos experimentos.

A avaliação da decomposição foi condicionada aos registros previamente marcados como elegíveis no dataset ouro. Portanto, os resultados medem a qualidade da

decomposição quando a aplicabilidade da etapa já é conhecida, e não a capacidade da pipeline de decidir autonomamente quais documentos devem ser decompostos.

Figura 1 – Pipeline para extração da seção “Objeto”, decomposição do conteúdo e avaliação dos resultados.



Fonte: Elaborada pelo autor.

4.2 CONSTRUÇÃO E CURADORIA DO DATASET

Esta seção descreve a seleção do conjunto documental e a construção, curadoria e normalização do dataset ouro utilizado como referência nos experimentos. O dataset ouro corresponde ao conjunto de registros manualmente curados e normalizados empregado na avaliação das saídas da pipeline.

A curadoria foi realizada manualmente pelo autor. Não houve dupla anotação independente nem cálculo de concordância entre anotadores, o que constitui uma limitação do trabalho. Para reduzir inconsistências, foram aplicados critérios uniformes

de anotação, revisão dos casos duvidosos e normalização dos campos usados na avaliação.

4.2.1 Coleta, Normalização e Caracterização

A construção do dataset ouro compreendeu quatro atividades principais:

1. **Seleção e composição do conjunto documental:** A base disponível era composta por 890 documentos do MPSC, cedidos ao Projeto Céos e disponibilizados ao autor mediante a assinatura de termo de confidencialidade. Desse acervo, foram selecionados 530 documentos, correspondentes a 59,55% do total, por amostragem não probabilística. O acervo completo não foi utilizado devido ao custo da curadoria manual e, posteriormente, do processamento dos experimentos. A seleção foi estruturada em três recortes: os 450 documentos iniciais da listagem disponibilizada; 50 documentos de uma região intermediária da mesma listagem, com exclusão de cópias identificadas por sufixos nos nomes dos arquivos; e 30 documentos localizados ao final da listagem, identificados após inspeção preliminar por estarem relacionados a ANPCs e conterem a seção “Objeto”. O terceiro recorte foi incluído intencionalmente com a finalidade de ampliar o número de casos úteis para a avaliação da tarefa principal deste trabalho. Por se tratar de amostragem não probabilística, a composição do dataset não permite inferência estatística sobre todo o acervo. Ainda assim, a seleção buscou evitar concentração exclusiva em um único segmento da listagem, combinar documentos de diferentes posições da base disponível e ampliar a quantidade de casos positivos para a tarefa de extração da seção “Objeto”. Essa estratégia pode ter ampliado a variedade de padrões redacionais submetidos à avaliação no conjunto selecionado, mas não elimina possíveis vieses de seleção, especialmente em razão da inclusão intencional dos 30 documentos do terceiro recorte. Como não foram utilizados metadados específicos para verificar variação temporal, regional ou institucional dos documentos, a amostragem não é apresentada como representativa dessas dimensões. Assim, os resultados devem ser interpretados como evidências de desempenho da pipeline no conjunto documental analisado, de forma compatível com o objetivo experimental do estudo, e não como inferência estatística sobre a totalidade do acervo. Em razão do termo de confidencialidade, os documentos originais e o dataset ouro não são disponibilizados publicamente; o trabalho apresenta apenas resultados agregados e exemplos sintéticos, sem reprodução de informações identificáveis do acervo.

2. **Construção e curadoria:** As informações foram anotadas manualmente em planilha eletrônica no formato XLSX. A planilha registrou metadados documentais, como nome do arquivo, órgão de origem, identificador, tipo documental e títulos

das seções. Para a avaliação da extração e da decomposição, foram utilizados principalmente os campos `conteudo-secao-objeto`, referente ao texto anotado na seção "Objeto", com `null` quando a seção não foi encontrada; `Enquadramento Legal do Fato`, correspondente à capitulação jurídica expressa no Objeto, ou ao texto padrão quando não identificada; e `Descrição do Fato`, correspondente à narrativa da conduta concreta, ou ao texto padrão quando não identificada. A Tabela 2 apresenta exemplos sintéticos que ilustram a decomposição da seção "Objeto" nas variáveis de interesse adotadas neste estudo, incluindo situações em que determinado elemento informacional não está presente e deve ser registrado por texto-padrão.

3. **Normalização:** A normalização dos campos foi necessária para reduzir variações de grafia, formatação e redação que poderiam afetar as comparações. Os campos de órgão de origem e tipo documental foram padronizados para apoiar a caracterização do conjunto. Para os experimentos, foram especialmente relevantes os campos `elegível_categorização_anpc`, que indica a elegibilidade para a decomposição da seção "Objeto"; `Enquadramento Legal do Fato - normalizado`, correspondente à versão padronizada da capitulação jurídica; e `Descrição do Fato - normalizado`, correspondente à versão revisada para conter apenas a narrativa factual concreta. Foram marcados como elegíveis os documentos do tipo ANPC, ou variações diretamente relacionadas, que possuíam seção "Objeto". Foram excluídos os documentos sem essa seção e três documentos da área penal que estavam fora do escopo do trabalho. A normalização do enquadramento legal incluiu a expansão de abreviações, como `art.` e `inc.`, a uniformização da referência à Lei n. 8.429/92 e a remoção de rótulos redundantes. O literal `null` foi mantido quando não havia seção "Objeto", e o texto padrão foi preservado quando a seção existia, mas não continha enquadramento identificável. A descrição do fato foi revisada para conter apenas a narrativa da conduta, excluindo trechos relativos a obrigações, sanções, ressarcimento, multa civil, referências processuais e fórmulas genéricas que não descreviam o fato concreto.

4. **Caracterização:** Ao final da curadoria, o dataset ouro continha 530 documentos, distribuídos em 19 tipos documentais e vinculados a 90 órgãos de origem distintos. A Tabela 3 apresenta os números centrais do conjunto. A Tabela 4 detalha a ocorrência do enquadramento legal e da descrição do fato nos 467 documentos elegíveis para decomposição.

Tabela 2 – Exemplos sintéticos de decomposição da seção “Objeto” nas variáveis de interesse.

Ex.	Padrão representado	Objeto sintético/adaptado	Enquadramento legal do fato	Descrição do fato concreto
1	Com enquadramento e fato concreto	O presente acordo tem por objeto o fato subsumido à hipótese prevista no artigo 10, caput e inciso VIII, da Lei n. 8.429/1992, em razão de o compromissário ter contratado serviços, no exercício de função pública, em valor superior ao limite legal, dispensando indevidamente a licitação pública.	artigo 10, caput e inciso VIII, da Lei n. 8.429/1992	o compromissário ter contratado serviços, no exercício de função pública, em valor superior ao limite legal, dispensando indevidamente a licitação pública.
2	Com enquadramento, sem descrição do fato	O presente acordo tem por objeto fatos subsumidos às hipóteses típicas previstas nos artigos 9º, caput e inciso XI, 10, caput, e 11, caput e inciso I, da Lei n. 8.429/1992.	artigos 9º, caput e inciso XI; artigo 10, caput; e artigo 11, caput e inciso I, da Lei n. 8.429/1992	Não foi identificada descrição do fato concreto na seção Objeto.
3	Sem enquadramento, com descrição do fato	O presente acordo tem por objeto a conduta consistente em autorizar o uso de bem público para finalidade particular, sem autorização legal e sem contraprestação ao ente público.	Não foi identificado enquadramento legal do fato na seção Objeto.	autorizar o uso de bem público para finalidade particular, sem autorização legal e sem contraprestação ao ente público.
4	Sem enquadramento e sem fato concreto	O presente acordo tem por objeto compelir a parte compromissária ao pagamento de multa civil e ao cumprimento de obrigação de não fazer, evitando-se a judicialização da controvérsia.	Não foi identificado enquadramento legal do fato na seção Objeto.	Não foi identificada descrição do fato concreto na seção Objeto.

Nota: os exemplos são sintéticos e foram elaborados a partir de padrões recorrentes observados no conjunto documental, sem correspondência direta com caso específico do acervo. Foram omitidos ou generalizados elementos potencialmente identificadores.
Fonte: Elaborada pelo autor.

Tabela 3 – Visão geral do dataset ouro.

Métrica	Total	Percentual
Total de documentos	530	100,0%
Órgãos de origem distintos	90	–
Tipos documentais distintos	19	–
Documentos do tipo ANPC ou variações correlatas	497	93,77%
Documentos de outros tipos	33	6,23%
Documentos com seção “Objeto” identificada	470	88,68%
Documentos sem seção “Objeto” identificada (`null`)	60	11,32%

Nota: Entre os documentos com seção “Objeto”, 467 são do tipo ANPC e 3 são da área penal. Entre os documentos sem a seção, 30 são do tipo ANPC e 30 são de outros tipos documentais.
Fonte: Elaborada pelo autor.

Tabela 4 – Caracterização do dataset ouro quanto ao enquadramento legal e à descrição do fato.

Métrica	**Total**	**Percentual**
Total de documentos avaliados	467	100,0%
Documentos com enquadramento legal identificado	413	88,44%
Documentos sem enquadramento legal identificado	54	11,56%
Documentos com descrição do fato identificada	342	73,23%
Documentos sem descrição do fato identificada	125	26,77%

Nota: Os percentuais foram calculados sobre os 467 documentos elegíveis, a partir dos campos normalizados do dataset ouro.
Fonte: Elaborada pelo autor.

4.2.2 Delimitação do Objeto de Extração

A tarefa principal foi delimitada à extração textual da seção “Objeto”. A decomposição posterior também foi restrita a essa seção, sem busca de enquadramento legal ou descrição do fato em outras partes do documento.

Essa delimitação manteve a avaliação alinhada ao dataset ouro. Como a seção nem sempre contém todas as variáveis de interesse, a ausência de enquadramento ou descrição não representa necessariamente falha do modelo. A busca no documento integral exigiria nova curadoria e foi reservada para trabalhos futuros.

4.3 ESTRATÉGIA DE EXTRAÇÃO BASEADA EM RAG

A estratégia computacional foi organizada em duas etapas. Na primeira, a pipeline selecionou, em cada documento, os trechos mais relevantes para a localização da seção “Objeto” e utilizou o LLM para realizar a extração textual dessa seção. Na segunda, uma nova chamada ao LLM recebeu somente o Objeto extraído e o decompôs nas variáveis de interesse: o enquadramento legal do fato e a descrição do fato concreto. O RAG foi utilizado apenas na etapa de extração; a decomposição não realizou nova recuperação no documento.

4.3.1 Arquitetura e Recuperação Semântica

O texto dos PDFs foi extraído com a biblioteca `pdfplumber` e submetido a pré-processamento para reduzir espaços excessivos, quebras irregulares, cabeçalhos, rodapés e marcas de paginação. As rotinas de limpeza textual baseadas em expressões regulares foram implementadas com o módulo `re` da biblioteca padrão do Python, enquanto a normalização Unicode complementar foi realizada com o módulo `unicodedata`. Em seguida, o conteúdo foi segmentado por uma função própria baseada em janela deslizante. Os parâmetros `rag-chunk` e `rag-overlap`, medidos em caracteres, controlaram, respectivamente, o tamanho dos fragmentos e a sobreposição entre trechos consecutivos. Essa sobreposição buscou preservar a continuidade de seções que

atravessavam a fronteira entre dois fragmentos.

Operacionalmente, o deslocamento entre dois fragmentos consecutivos foi definido como a diferença entre `rag-chunk` e `rag-overlap`. Assim, a parte final de um fragmento era repetida no início do fragmento seguinte, buscando preservar sentenças ou seções que atravessassem a fronteira entre as janelas.

A segmentação foi definida em caracteres por simplicidade de implementação e reprodutibilidade, evitando dependência do tokenizador específico do modelo de *embedding* ou dos LLMs avaliados. Embora modelos de *embedding*, como o `bge-m3:latest`, e modelos de linguagem operem internamente com limites de *tokens*, a quantidade de caracteres foi tratada como uma aproximação operacional, e não como controle exato da janela de *tokens*. Essa escolha facilitou a aplicação uniforme dos mesmos parâmetros aos diferentes experimentos, mas constitui uma limitação da implementação; métodos baseados em *tokens* ou na estrutura documental podem ser avaliados em trabalhos futuros.

Os *embeddings* foram gerados pelo modelo `bge-m3:latest` (Chen et al., 2025). Em cada execução, a base de recuperação foi composta exclusivamente pelos fragmentos do próprio PDF processado, e não por uma base externa de conhecimento jurídico. Assim, o RAG atuou como mecanismo de seleção de trechos relevantes dentro do próprio documento em análise.

A recuperação utilizou um conjunto fixo de consultas de busca relacionadas ao título e às formas recorrentes de introdução da seção “Objeto”. Para cada fragmento, calculou-se a maior similaridade de cosseno entre seu *embedding* e os *embeddings* dessas consultas. Essa medida foi combinada com uma pontuação adicional, de menor peso, baseada na presença de termos e expressões característicos da seção. Os fragmentos foram ordenados pela pontuação resultante, e os *top-k* mais relevantes, conforme o parâmetro `rag-topk`, foram reorganizados de acordo com sua posição original no documento antes da construção do contexto enviado ao LLM.

Antes do envio ao LLM, a pipeline aplicou ainda uma verificação estrutural simples sobre os próprios trechos recuperados, buscando indícios textuais da seção “Objeto”. Quando identificado, esse candidato era apenas acrescentado ao contexto enviado ao modelo, sem substituir a recuperação semântica e sem utilizar informações do dataset ouro. Nos experimentos oficiais, não foi utilizado o mecanismo alternativo de recuperação baseado apenas em palavras-chave.

4.3.2 Desenho dos Prompts e Validação

Foram avaliadas três versões do prompt de extração. O Prompt Objeto 2.1 adotou critérios mais restritivos; o Prompt Objeto 2.2 ajustou esses critérios para preservar conteúdos legítimos associados à seção “Objeto”; e o Prompt Objeto 2.3 acrescentou regras para evitar que quebras visuais de linha fossem reproduzidas como quebras de

parágrafo. Os textos integrais dessas versões são detalhados nas seções A.1.2, A.1.3 e A.1.4 do Apêndice A.

A saída da extração foi exigida em formato JSON, contendo o campo `objeto` com o texto extraído ou o valor `null` quando a seção não pudesse ser identificada com segurança. Respostas inválidas ou sem o campo esperado foram registradas como erro técnico. O retorno `null` podia representar um acerto, quando não havia seção no dataset ouro, ou uma falha, quando existia conteúdo de referência.

Nos experimentos principais de decomposição, uma segunda chamada ao LLM recebeu apenas o Objeto extraído e retornou, também em JSON, o enquadramento legal e a descrição do fato. O Prompt de Decomposição 2.0 orientou o modelo a preservar a redação encontrada e a utilizar textos padrão quando determinado elemento não estivesse presente. Essa etapa foi aplicada aos 467 documentos marcados como elegíveis no dataset ouro.

Os prompts também modelaram explicitamente saídas negativas controladas, em consonância com os casos de ausência representados nos exemplos sintéticos da Tabela 2. Na extração, quando a seção “Objeto” não pudesse ser identificada com segurança, a saída esperada era o retorno JSON `{"objeto": null}`. Na decomposição, quando o enquadramento legal ou a descrição do fato não estivessem presentes no texto extraído, a saída esperada era o texto padrão de ausência correspondente. Essa escolha buscou reduzir o risco de respostas inferidas, preenchimentos não apoiados no texto ou alucinações, tratando a ausência de informação como uma saída válida da pipeline, e não como lacuna a ser completada pelo modelo.

Após a consolidação dos experimentos principais, foi realizado um experimento complementar com o Prompt de Decomposição 2.1. Essa versão manteve as regras de extração da descrição do fato e acrescentou normalização formal somente ao enquadramento legal, incluindo padronização de abreviações, artigos, incisos e referências à Lei n. 8.429/92. As regras não autorizavam inferir capitulações ausentes do texto. O novo prompt foi avaliado com Gemma 4 e Qwen 3.6, nas configurações base e ampliada, reutilizando os Objetos gerados nas quatro execuções com o Prompt Objeto 2.3 e mantendo os demais parâmetros. Não houve nova extração dos PDFs nem nova recuperação semântica; a alteração intencional concentrou-se no prompt empregado na segunda etapa. Como a saída esperada passou a incluir transformação formal do enquadramento, esse experimento foi tratado como complementar e não substituiu os resultados principais.

As regras do Prompt de Decomposição 2.1 foram derivadas do padrão de normalização adotado no próprio dataset ouro e avaliadas sobre os mesmos registros elegíveis. Assim, o experimento mede a aderência dos modelos ao esquema de padronização definido neste trabalho, mas não constitui validação independente de generalização para outros acervos ou critérios de normalização. Os prompts de decomposição são apresentados nas seções A.2.1 e A.2.2 do Apêndice A.

4.4 CRITÉRIOS E MÉTRICAS DE AVALIAÇÃO

4.4.1 Métricas e procedimentos de comparação

A avaliação foi organizada em quatro frentes: fidelidade textual das saídas, classificação binária da presença da seção “Objeto”, indicadores de execução e análise qualitativa dos casos divergentes. Na extração da seção “Objeto”, a pipeline processou os 530 documentos, sem filtro prévio por tipo documental. As métricas textuais foram calculadas sobre os 470 registros que possuíam conteúdo de referência no dataset ouro, incluindo 467 ANPCs e três documentos da área penal. Na etapa de decomposição, foram considerados os 467 registros previamente marcados como elegíveis no dataset ouro, e o enquadramento legal e a descrição do fato foram avaliados separadamente.

A **taxa de igualdade textual após limpeza técnica** indica a proporção de casos em que a saída da pipeline e o texto de referência se tornaram idênticos após um tratamento destinado a reduzir artefatos de representação, como caracteres de controle, representações inconsistentes de caracteres e espaços excedentes. Esse procedimento preservou aspectos relevantes da forma textual, como caixa, acentuação, pontuação e separação entre parágrafos. Assim, qualquer acréscimo, omissão, reordenação, alteração ou substituição de termos que permanecesse após o tratamento fazia o indicador individual de igualdade assumir valor zero.

A **similaridade textual** foi calculada com o método `ratio()` da classe `SequenceMatcher`, pertencente à biblioteca padrão `difflib` do Python. Como as entradas foram fornecidas como cadeias de caracteres, os elementos comparados pelo método foram caracteres individuais. O `SequenceMatcher` identifica blocos de caracteres coincidentes e contíguos; o conjunto desses blocos não apresenta sobreposição e preserva sua ordem relativa nas duas sequências. Consequentemente, a posição e a ordem dos conteúdos influenciam o resultado, não se tratando de uma simples contagem de caracteres comuns independentemente de sua localização. A razão é calculada por $2M/T$, em que M corresponde à soma dos comprimentos dos blocos coincidentes e T à soma dos comprimentos das duas sequências comparadas (Python Software Foundation, 2025).

Para cada registro i, definiu-se P_i como o texto produzido pela pipeline, O_i como o texto correspondente no dataset ouro e $\mathrm{comp}(\cdot)$ como a transformação aplicada antes da comparação. A similaridade percentual foi calculada por:

$$S_i = 100 \times \mathrm{ratio}\left(\mathrm{comp}(P_i), \mathrm{comp}(O_i)\right),$$

em que $S_i \in [0, 100]$. Em todas as comparações, manteve-se a saída da pipeline como primeira sequência e o texto do dataset ouro como segunda sequência.

Seja Σ^* o conjunto das cadeias de caracteres consideradas na avaliação. Definiu-se $\mathrm{comp} : \Sigma^* \to \Sigma^*$ como uma transformação determinística empregada exclusivamen-

te no cálculo da similaridade textual. Conforme a sequência implementada no código, essa função converte o texto para letras minúsculas, elimina acentos gráficos e sinais de pontuação, substitui quebras de linha por espaços, reduz sequências de espaços a um único espaço e remove espaços no início e no fim do texto. A mesma transformação foi aplicada à saída da pipeline e ao texto de referência. Esse procedimento é distinto da limpeza mais conservadora utilizada na igualdade textual, que preservou caixa, acentuação, pontuação e separação entre parágrafos.

Na implementação, empregou-se a chamada `SequenceMatcher(None, a, b).ratio()`, em que a representa a saída normalizada da pipeline e b o texto normalizado do dataset ouro. O parâmetro `autojunk` não foi explicitamente informado e, portanto, permaneceu com seu valor padrão `True`. Os valores percentuais foram obtidos pela multiplicação da razão por 100 e pelo arredondamento para duas casas decimais.

Optou-se por uma métrica de similaridade baseada na forma textual como critério principal porque a tarefa central deste estudo envolve a extração literal e a preservação da redação original da seção “Objeto”. Nesse contexto, acréscimos, omissões, reordenações e paráfrases devem ser penalizados, pois podem modificar a forma do relato jurídico ou do enquadramento legal registrado no documento. Métricas semânticas, como *BERTScore* ou medidas baseadas em proximidade vetorial, poderiam atribuir pontuação elevada a textos semanticamente próximos; contudo, uma paráfrase semanticamente adequada não é suficiente quando o objetivo é reproduzir o trecho de referência. Métricas de precisão, revocação e *F1-score* baseadas em *tokens* também poderiam complementar a avaliação, mas não foram adotadas como critério principal porque não representam diretamente a ordem, a continuidade e a delimitação integral do bloco textual extraído.

Os registros sem conteúdo de referência não integraram o denominador das métricas textuais. Quando havia conteúdo de referência no dataset ouro, mas a pipeline não retornava a saída correspondente, os indicadores de igualdade e similaridade assumiam valor zero. Na decomposição, os textos-padrão definidos para indicar ausência de enquadramento legal ou de descrição do fato também integraram a avaliação. Desse modo, as métricas agregadas dessa etapa refletem tanto a fidelidade dos conteúdos identificados quanto o reconhecimento correto da ausência de determinado elemento.

A partir das pontuações individuais, foram calculadas a similaridade média e as proporções de registros com similaridade igual ou superior a 80%, 90% e 95%. Esses limiares foram adotados como critérios para comparar configurações completas e, quando aplicável, combinações diretamente comparáveis de modelos, *prompts* e parâmetros. Portanto, não constituem pontos de corte estatisticamente validados nem certificam, isoladamente, a correção jurídica das saídas.

Os tratamentos aplicados para o cálculo das métricas não modificaram os valores armazenados no banco de dados ou no dataset ouro. Tais procedimentos também

se distinguem da normalização manual, realizada durante a curadoria do dataset, e da normalização formal, solicitada ao LLM via Prompt de Decomposição 2.1. No experimento complementar, as mesmas métricas foram utilizadas para comparar o enquadramento legal produzido pelos modelos com a respectiva coluna normalizada do dataset ouro, mantendo-se os 467 registros elegíveis.

Além das métricas textuais, foram registrados os textos extraídos da seção "Objeto", os retornos `null`, as respostas fora do formato esperado, a presença simultânea de referência e saída, e os tempos de processamento. Os casos de maior divergência também foram examinados qualitativamente para identificar padrões como: retorno indevido de texto em documentos sem conteúdo de referência, duplicações, truncamentos, respostas inválidas e retornos `null` quando havia conteúdo de referência.

Como análise complementar, foi realizada uma avaliação sob a ótica de **classificação binária** para medir a capacidade da pipeline de distinguir documentos com e sem a seção "Objeto". Essa análise considerou apenas a configuração de maior fidelidade textual — Gemma 4 (31B), Prompt Objeto 2.3 e configuração ampliada — e utilizou os mesmos 530 documentos, o dataset ouro e as saídas já persistidas, sem nova execução do modelo.

Considerou-se positiva a classe correspondente à presença de texto de referência no dataset ouro. A predição foi classificada como positiva quando a pipeline retornou um Objeto textual e negativa quando retornou `null`. Assim, um verdadeiro positivo indica apenas que a presença da seção foi identificada, enquanto sua fidelidade textual permanece avaliada separadamente pelas métricas de igualdade e similaridade.

Foram calculados verdadeiros positivos (VP), verdadeiros negativos (VN), falsos positivos (FP) e falsos negativos (FN), além das seguintes métricas:

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN}, \qquad \text{Precisão} = \frac{VP}{VP + FP},$$

$$\text{Sensibilidade} = \frac{VP}{VP + FN}, \qquad \text{Especificidade} = \frac{VN}{VN + FP}.$$

Essa avaliação binária responde se a pipeline reconheceu corretamente a presença ou a ausência da seção. Ela não mede a fidelidade do conteúdo textual retornado, que permanece avaliada separadamente pelas métricas de igualdade e similaridade.

4.4.2 Validade experimental e ausência de *hold-out*

Em avaliações experimentais de modelos, é usual separar os dados empregados no desenvolvimento, ajuste ou seleção de configurações daqueles reservados à avaliação final. Neste trabalho, entende-se por *hold-out test set* um subconjunto de documentos previamente apartado, não utilizado na inspeção intermediária dos resultados, no refinamento dos prompts ou na seleção de parâmetros, e empregado apenas

para estimar o desempenho final da pipeline. Essa separação busca reduzir estimativas excessivamente otimistas decorrentes de decisões tomadas com base no mesmo conjunto utilizado para avaliação (Kohavi, 1995; Hastie; Tibshirani; Friedman, 2009).

No caso deste estudo, não foi reservado um conjunto independente desse tipo; além disso, os prompts foram refinados a partir da inspeção de resultados obtidos no mesmo dataset utilizado na avaliação final. Esse procedimento pode introduzir risco de sobreajuste de prompt ao conjunto analisado, isto é, de adaptação das instruções aos padrões sintáticos e redacionais observados nos documentos avaliados. Em tarefas com modelos de linguagem, a seleção de prompts e hiperparâmetros também pode depender de exemplos ou conjuntos de desenvolvimento; por isso, quando essa etapa utiliza os mesmos documentos empregados na avaliação final, a estimativa de desempenho fica condicionada ao conjunto analisado (Perez; Kiela; Cho, 2021). Por essa razão, o estudo deve ser interpretado como exploratório e descritivo, com caráter de prova de conceito da pipeline proposta, e as métricas apresentadas não devem ser tratadas como estimativa de generalização para outros acervos.

Outras limitações de robustez decorrem da ausência de um protocolo de repetições sistemáticas de cada configuração experimental, da curadoria realizada por um único anotador, sem cálculo de concordância inter-anotadores, e da amostragem não probabilística do acervo. Em tarefas que envolvem anotação manual de conjuntos documentais, medidas de concordância entre anotadores são usualmente empregadas para avaliar a consistência das decisões de anotação (Artstein; Poesio, 2008). De modo complementar, como a seleção combinou recortes da listagem com inclusão intencional de casos positivos, a composição do conjunto documental não foi desenhada para estimar parâmetros da totalidade dos documentos do MPSC; por isso, os resultados não são apresentados como inferência estatística sobre o acervo completo. Essas escolhas foram compatíveis com o escopo e os recursos disponíveis para o trabalho, mas impedem isolar com segurança a variabilidade associada ao modelo, ao prompt ou aos parâmetros. Para uso em produção ou implantação institucional, seria necessária validação em conjunto independente, preferencialmente por meio de um *hold-out test set* definido antes do refinamento dos prompts, além de nova curadoria e validação específica quando a pipeline for aplicada a outros acervos, tipos documentais ou critérios de normalização.

5 AVALIAÇÃO EXPERIMENTAL

Este capítulo apresenta a avaliação experimental da pipeline, contemplando as configurações executadas, os resultados da extração da seção “Objeto”, a decomposição do conteúdo extraído e a discussão dos achados. A extração foi avaliada sob duas perspectivas complementares: a fidelidade textual nos 470 documentos com conteúdo de referência e, para a configuração de melhor desempenho textual, a identificação binária da presença ou ausência da seção nos 530 documentos.

Os critérios e métricas de avaliação são definidos na Seção 4.4. Os Prompts Objeto e os Prompts de Decomposição utilizados são descritos na Seção 4.3.2 e apresentados no Apêndice A. Quando houve reexecução de um mesmo modelo e configuração, considerou-se a execução mais recente. O capítulo está organizado em quatro partes: configuração experimental, desempenho da extração, análise da decomposição e discussão abrangente dos resultados.

5.1 CONFIGURAÇÃO EXPERIMENTAL

Esta seção apresenta os modelos, os parâmetros e o ambiente computacional empregados nos experimentos. Os critérios e as métricas de avaliação foram definidos na Seção 4.4.

O desenho experimental não constituiu um experimento fatorial completo, pois nem todos os modelos foram avaliados com todas as versões de prompt e com todos os parâmetros de recuperação e geração. Assim, os resultados devem ser interpretados como desempenho de configurações completas da pipeline — isto é, combinações entre modelo, prompt, parâmetros de recuperação semântica e parâmetros de geração —, e não como estimativa isolada do mérito de cada componente. Os resultados, portanto, não permitem isolar causalmente o efeito individual do modelo, do prompt ou dos parâmetros. As comparações entre modelos, prompts ou parâmetros têm caráter descritivo e devem ser consideradas apenas nos casos em que as combinações avaliadas forem diretamente comparáveis.

5.1.1 Modelos, parâmetros e ferramentas

Neste trabalho, o LLM atua como o componente central de interpretação textual em duas etapas. Na primeira, processa o contexto fornecido pela recuperação semântica para delimitar e retornar a seção denominada “Objeto” dos ANPCs. Na segunda, realiza a decomposição do trecho extraído nas variáveis de interesse. Para a execução dessas tarefas, foram adotadas as famílias de modelos Gemma 3 (Gemma Team, 2025), Gemma 4 (Google DeepMind, 2026), Qwen 2.5 (Qwen et al., 2025), Qwen 3.6 (Yang et al., 2025) e Llama 3.3 (Dubey et al., 2024), disponibilizadas para execução

local no servidor Ollama.

Os modelos avaliados são apresentados na Tabela 5, com seus nomes formais, número aproximado de parâmetros e identificadores utilizados no Ollama. Gemma 3, Qwen 2.5 e Llama 3.3 foram avaliados com os Prompts Objeto 2.1 e 2.2. Posteriormente, Gemma 4 e Qwen 3.6 foram avaliados com o Prompt Objeto 2.3, em configurações base e ampliada. O experimento complementar de normalização reutilizou os Objetos obtidos nas quatro execuções de Gemma 4 e Qwen 3.6, sem repetir a extração da seção “Objeto”.

Tabela 5 – Modelos de linguagem avaliados.

Desenvolvedor	**Modelo**	**Parâmetros aprox. (B)**	**Identificador no Ollama**
Meta	Llama 3.3	70	`llama3.3:70b`
Google DeepMind	Gemma 3	27	`gemma3:27b`
	Gemma 4	31	`gemma4:31b`
Qwen Team Alibaba Cloud	Qwen 2.5	32	`qwen2.5:32b-instruct`
	Qwen 2.5	72	`qwen2.5:72b-instruct`
	Qwen 3.6	36	`qwen3.6:latest`

Nota: A coluna “Parâmetros aprox. (B)” indica o número aproximado de parâmetros do modelo em bilhões. No servidor Ollama utilizado nos experimentos, o artefato `qwen3.6:latest` apresentou o metadado `parameter_size=36.0B`. As tags `qwen3.6:latest` e `qwen3.6:35b` apontavam para o mesmo artefato, identificado pelo *digest* `07d35212591fc27746f0a317c975a6d68754fb38e9053d82e25f06057af28522`, com quantização `Q4_K_M`.
Fonte: Elaborada pelo autor.

O desenho experimental concentrou-se nas famílias de modelos apresentadas na Tabela 5 e não incluiu modelos de linguagem de menor porte. Portanto, os resultados não permitem avaliar o *trade-off* entre modelos menores e os LLMs empregados quanto à qualidade, ao tempo e ao custo computacional.

A Tabela 6 apresenta os parâmetros definidos nos testes preliminares. Nas execuções com o Prompt Objeto 2.3, a configuração ampliada aumentou a sobreposição, o número de trechos recuperados e a janela de contexto, mantendo os demais valores. Optou-se por evitar valores muito altos de `rag-topk`, pois o envio de muitos fragmentos aumenta o contexto, o tempo de processamento e o risco de introdução de ruído na resposta. O experimento complementar de decomposição manteve os parâmetros da execução de origem e alterou somente o prompt empregado na segunda chamada ao LLM.

A pipeline foi implementada em Python 3.11 e executada em ambiente Linux por meio do JupyterLab. O SQLite foi utilizado para persistir documentos, extrações, respostas, parâmetros e resultados. A extração dos PDFs empregou `pdfplumber`, enquanto os LLMs e o modelo de *embeddings* foram acessados em servidor Ollama remoto por API HTTP. A configuração física da *Graphics Processing Unit* (GPU), os drivers e as bibliotecas internas do servidor não estavam sob controle do autor.

Tabela 6 – Parâmetros utilizados nas execuções.

Parâmetro	**Valor**	**Função no experimento**
`embed-model`	`bge-m3:latest`	Modelo de geração dos *embeddings*.
`rag-chunk`	2000	Tamanho máximo do fragmento, em caracteres.
`rag-overlap`	100 (base) ou 150 (ampliada)	Sobreposição entre fragmentos consecutivos, em caracteres.
`rag-topk`	8 (base) ou 10 (ampliada)	Número de trechos recuperados para compor o contexto.
`temperature`	0	Controle da aleatoriedade da geração.
`timeout`	1800 segundos	Tempo máximo de espera por requisição.
`num-predict`	1536	Número máximo de *tokens* gerados na resposta.
`num-ctx`	8192 (base) ou 12288 (ampliada)	Janela máxima de contexto, em *tokens*.
`think`	`false`	Desativação do modo de raciocínio, quando suportado.

Fonte: Elaborada pelo autor.

5.2 DESEMPENHO DA EXTRAÇÃO DA SEÇÃO “OBJETO”

Antes da análise da qualidade textual, a Tabela 7 apresenta o panorama operacional das 12 execuções principais. Os dados distinguem Objetos textuais extraídos, retornos nulos e respostas fora do formato JSON, além do tempo total de cada rodada.

Tabela 7 – Panorama operacional das execuções experimentais consideradas.

Prompt	**Modelo**	**Tempo total**	**Objeto extraído**	**Objeto nulo**	**Erro JSON**	**Total**	**Config.**
2.1	Gemma 3 (27B)	02:13:24	522	7	1	530	base
2.1	Qwen 2.5 (32B)	02:55:14	475	53	2	530	base
2.1	Qwen 2.5 (72B)	03:56:13	483	46	1	530	base
2.1	Llama 3.3 (70B)	03:08:04	500	29	1	530	base
2.2	Gemma 3 (27B)	01:49:49	523	6	1	530	base
2.2	Qwen 2.5 (32B)	02:11:32	474	54	2	530	base
2.2	Qwen 2.5 (72B)	03:11:13	480	48	2	530	base
2.2	Llama 3.3 (70B)	03:49:21	527	3	0	530	base
2.3	Gemma 4 (31B)	10:58:23	484	43	3	530	base
2.3	Gemma 4 (31B)	06:40:08	490	40	0	530	ampliada
2.3	Qwen 3.6 (36B)	01:13:46	510	19	1	530	base
2.3	Qwen 3.6 (36B)	01:04:52	514	16	0	530	ampliada

Nota: “Objeto extraído” indica que a pipeline retornou e persistiu uma seção “Objeto” textual; “Objeto nulo” indica retorno `{"objeto": null}` ou equivalente; e “Erro JSON” indica resposta fora do formato esperado. Nas linhas do Prompt Objeto 2.3, “base” corresponde a `rag-overlap=100`, `rag-topk=8` e `num-ctx=8192`, enquanto “ampliada” corresponde a `rag-overlap=150`, `rag-topk=10` e `num-ctx=12288`.

Fonte: Elaborada pelo autor.

Nas execuções principais, foram registrados 14 casos de resposta fora do formato JSON. Não foram identificados casos de timeout ou falha de comunicação com o servidor nos logs considerados. O retorno `null`, por sua vez, não representa necessariamente erro, pois pode ser adequado quando o documento não contém a seção "Objeto".

As métricas de igualdade e similaridade textual da Tabela 8 foram calculadas sobre os 470 documentos com Objeto no dataset ouro. A coluna "LLM" indica a quantidade de documentos, entre os 530 avaliados, para os quais a pipeline retornou e persistiu um Objeto textual; "Ambos" indica os casos com Objeto no dataset ouro e na saída da pipeline; e "Aus." indica os documentos com referência no dataset ouro, mas sem Objeto textual retornado.

Os 60 documentos sem Objeto no dataset ouro não compõem essas métricas, pois não há texto de referência para comparação. Quando havia Objeto no dataset ouro, mas a pipeline não retornava conteúdo, igualdade e similaridade eram consideradas iguais a zero. Uma avaliação binária complementar da presença ou ausência da seção, aplicada à configuração de melhor fidelidade textual, é apresentada na Seção 5.2.1.

Tabela 8 – Resultados da extração da seção "Objeto" por modelo, prompt e configuração.

Prompt	Modelo	Config.	LLM	Ambos	Aus.	Igual.	Sim. média	≥80%	≥90%	≥95%
2.1	Gemma 3 (27B)	base	522	468	2	65,74	93,62	87,02	85,74	81,91
2.1	Qwen 2.5 (32B)	base	475	464	6	17,02	93,04	89,79	86,60	83,40
2.1	Qwen 2.5 (72B)	base	483	467	3	81,49	97,68	96,81	95,74	93,83
2.1	Llama 3.3 (70B)	base	500	469	1	65,53	89,35	81,91	80,00	77,66
2.2	Gemma 3 (27B)	base	523	466	4	77,87	95,76	92,77	91,70	88,51
2.2	Qwen 2.5 (32B)	base	474	461	9	45,32	95,70	93,40	92,34	90,21
2.2	Qwen 2.5 (72B)	base	480	462	8	80,43	96,63	95,53	94,26	92,13
2.2	Llama 3.3 (70B)	base	527	469	1	74,89	92,87	90,85	88,72	86,38
2.3	Gemma 4 (31B)	base	484	466	4	87,23	97,89	97,66	96,38	94,68
2.3	Gemma 4 (31B)	ampl.	490	470	0	87,45	99,11	98,94	97,87	96,17
2.3	Qwen 3.6 (36B)	base	510	465	5	17,23	89,98	87,23	84,04	82,34
2.3	Qwen 3.6 (36B)	ampl.	514	470	0	17,66	90,37	86,81	83,40	81,91

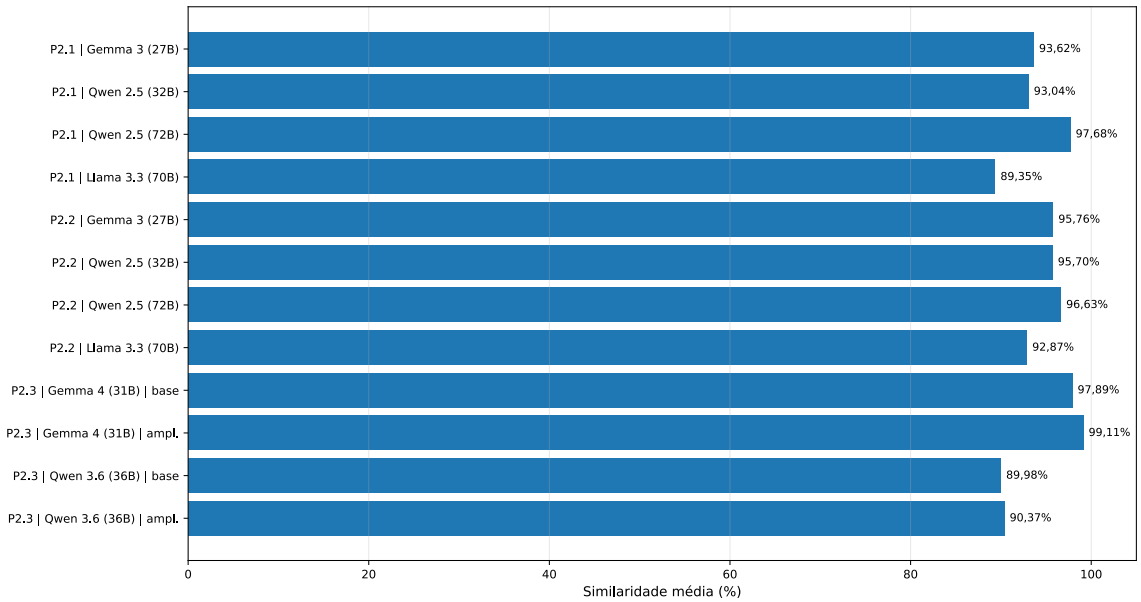
Nota: as métricas de igualdade e similaridade foram calculadas sobre os 470 documentos com Objeto no dataset ouro. "LLM" indica documentos com Objeto textual retornado pela pipeline; "Ambos" indica documentos com Objeto tanto no dataset ouro quanto na saída da pipeline; "Aus." indica documentos com Objeto no dataset ouro, mas sem Objeto textual retornado pela pipeline. "Sim. média" corresponde à similaridade textual média calculada pelo método `SequenceMatcher`, após a transformação comp descrita na Seção 4.4.1. Documentos sem Objeto no dataset ouro não compõem o denominador dessas métricas. Nas linhas do Prompt 2.3, "base" corresponde a `overlap=100`, `topk=8` e `num-ctx=8192`; "ampl." corresponde a `overlap=150`, `topk=10` e `num-ctx=12288`.
Fonte: Elaborada pelo autor.

Na melhor configuração, correspondente ao Gemma 4 (31B) com o Prompt Objeto 2.3 em configuração ampliada, todos os 470 documentos que possuíam seção "Objeto" no dataset ouro receberam uma saída textual da pipeline, sem casos de ausência entre os positivos de referência. Entretanto, a mesma execução retornou conteúdo

textual classificado como “Objeto” em 20 dos 60 documentos sem seção “Objeto” de referência, uma vez que foram registrados 490 retornos textuais no total, dos quais 470 correspondiam a documentos que também possuíam Objeto no dataset ouro. Assim, a similaridade média de 99,11% deve ser interpretada como medida de fidelidade textual nos documentos com referência, e não como medida global de detecção da presença da seção.

A Figura 2 sintetiza a similaridade média das execuções. O melhor resultado foi obtido pelo Gemma 4 (31B), com o Prompt Objeto 2.3 e a configuração ampliada, alcançando 99,11% de similaridade média. Entre os modelos avaliados com os Prompts Objeto 2.1 e 2.2, o Qwen 2.5 (72B) também apresentou desempenho elevado.

Figura 2 – Similaridade média na extração da seção “Objeto” nos 470 documentos com conteúdo de referência, por modelo, prompt e configuração.



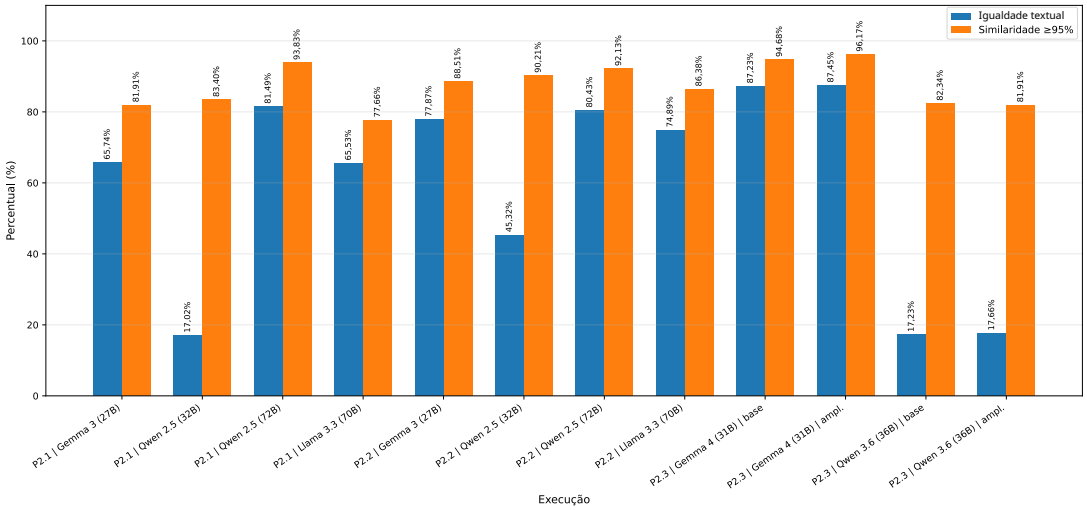
Fonte: Elaborada pelo autor.

A comparação entre os Prompts Objeto 2.1 e 2.2 mostra comportamentos distintos conforme o modelo. No Gemma 3 (27B), a igualdade aumentou de 65,74% para 77,87%, e a similaridade média passou de 93,62% para 95,76%. No Qwen 2.5 (72B), o Prompt Objeto 2.1 apresentou resultado ligeiramente superior ao 2.2, embora ambas as execuções tenham mantido alta similaridade.

Em relação às versões anteriores, o Prompt Objeto 2.3 manteve as regras de elegibilidade e de delimitação final consolidadas no Prompt Objeto 2.2, mas acrescentou orientações específicas para não interpretar quebras visuais de linha dos PDFs como quebras reais do texto e para unir linhas pertencentes ao mesmo parágrafo. Essa alteração pode ter reduzido divergências meramente formatacionais e favorecido a aderência textual das saídas ao trecho de referência. Como a versão 2.3 foi avaliada em rodada posterior, com combinações distintas de modelos e parâmetros, esse efeito não pode ser isolado experimentalmente.

A Figura 3 evidencia que igualdade e alta similaridade textual medem aspectos diferentes. O Qwen 3.6 (36B), por exemplo, apresentou baixa igualdade, mas manteve parte das saídas acima do limiar de 95%. O Gemma 4 (31B), especialmente na configuração ampliada, combinou os maiores valores nas duas medidas.

Figura 3 – Igualdade textual após limpeza técnica e similaridade textual igual ou superior a 95% na extração da seção “Objeto” nos 470 documentos com conteúdo de referência.



Fonte: Elaborada pelo autor.

5.2.1 Avaliação binária da presença da seção “Objeto”

Além da fidelidade textual nos documentos com referência, foi realizada uma análise binária complementar sobre os 530 documentos para a configuração Gemma 4 (31B), Prompt Objeto 2.3 e configuração ampliada. A análise foi produzida na etapa de revisão final a partir dos mesmos logs, banco de dados e dataset ouro, sem nova execução do modelo.

A Tabela 9 apresenta a matriz de confusão. Nessa avaliação, retornar algum texto foi interpretado como indicação de presença da seção, enquanto retornar `null` foi interpretado como indicação de ausência. Essa classificação não avalia se o texto extraído está correto; a fidelidade da transcrição permanece medida separadamente pelas métricas textuais.

A configuração alcançou acurácia de 96,23%, precisão de 95,92%, sensibilidade de 100% e especificidade de 66,67%. A sensibilidade de 100% decorre da ausência de falsos negativos: todos os 470 documentos que possuíam Objeto no dataset ouro receberam uma saída textual. A menor especificidade decorre dos 20 falsos positivos entre os 60 documentos sem Objeto de referência.

A ocorrência de um falso positivo indica que a pipeline retornou algum trecho textual em um documento registrado como negativo no dataset ouro. Esse resultado, por si só, não permite concluir se houve geração de conteúdo inexistente ou seleção

Tabela 9 – Matriz de confusão da avaliação binária da presença da seção “Objeto”.

Referência no dataset ouro	**Predição da pipeline**	
	Objeto textual	**Sem Objeto (`null`)**
Objeto textual ($n = 470$)	**470** (VP)	0 (FN)
Sem Objeto ($n = 60$)	20 (FP)	**40** (VN)

Nota: a classe positiva corresponde à presença de Objeto textual. VP = verdadeiro positivo; VN = verdadeiro negativo; FP = falso positivo; e FN = falso negativo. A predição de ausência corresponde ao retorno `null`. A configuração avaliada não apresentou respostas fora do formato esperado.
Fonte: Elaborada pelo autor.

indevida de um trecho real de outra seção do documento; essa distinção exige inspeção qualitativa das saídas.

Como a avaliação binária complementar foi calculada apenas para a configuração selecionada pela maior fidelidade textual, os resultados não permitem concluir que essa configuração também seja a melhor entre todas as execuções para discriminar a presença e a ausência da seção “Objeto”. A comparação binária de todas as configurações foi reservada para trabalhos futuros.

5.3 ANÁLISE DA DECOMPOSIÇÃO EM ENQUADRAMENTO LEGAL E DESCRIÇÃO DO FATO

A decomposição foi realizada a partir do Objeto previamente extraído, sem nova leitura do PDF ou nova recuperação semântica. Nos experimentos principais, o Prompt de Decomposição 2.0 recebeu esse texto e retornou o enquadramento legal do fato e a descrição do fato concreto. Nas tabelas desta seção, a coluna “Prompt Objeto” identifica a versão do prompt que gerou o texto usado como entrada da decomposição. As comparações principais reúnem as execuções disponíveis com os Prompts Objeto 2.2 e 2.3.

A primeira avaliação, apresentada na Tabela 10, compara as saídas com as colunas não normalizadas do dataset ouro. Nessa comparação, o Gemma 4 (31B) apresentou os maiores valores de similaridade para o enquadramento legal não normalizado. Na configuração ampliada, alcançou 98,18% de similaridade média e 93,79% dos registros com similaridade igual ou superior a 95%. Para a descrição do fato, a mesma configuração obteve 89,17% de similaridade média e 77,73% dos registros no limiar de 95%. A maior igualdade na descrição foi observada no Qwen 2.5 (32B), com 47,75%.

A segunda avaliação, apresentada na Tabela 11, utiliza como referência as colunas normalizadas do dataset ouro. A diferença entre os percentuais de igualdade textual e de similaridade média decorre do contraste entre uma métrica binária, que exige

Tabela 10 – Decomposição da seção “Objeto” em comparação com as colunas não normalizadas do dataset ouro.

Prompt Objeto	Modelo	Campo	Igual.	Sim. média	≥80%	≥90%	≥95%
2.2	Gemma 3 (27B)	Enquadramento	50,32	84,98	58,67	54,39	54,18
2.2	Gemma 3 (27B)	Fato	37,04	81,87	74,09	65,95	58,89
2.2	Qwen 2.5 (32B)	Enquadramento	53,75	85,29	61,88	58,03	57,39
2.2	Qwen 2.5 (32B)	Fato	47,75	84,43	79,87	73,23	67,88
2.2	Qwen 2.5 (72B)	Enquadramento	51,39	85,12	59,96	55,89	55,25
2.2	Qwen 2.5 (72B)	Fato	23,77	85,25	79,87	73,88	70,45
2.2	Llama 3.3 (70B)	Enquadramento	51,39	85,57	59,31	55,25	54,60
2.2	Llama 3.3 (70B)	Fato	10,71	75,05	63,38	55,89	52,03
2.3	Gemma 4 (31B) base	Enquadramento	87,58	97,36	95,93	94,22	93,15
2.3	Gemma 4 (31B) base	Fato	40,69	88,16	85,01	80,09	77,30
2.3	Gemma 4 (31B) ampl.	Enquadramento	87,79	98,18	96,57	94,86	93,79
2.3	Gemma 4 (31B) ampl.	Fato	39,83	89,17	85,65	80,51	77,73
2.3	Qwen 3.6 (36B) base	Enquadramento	68,95	91,28	78,16	74,52	73,02
2.3	Qwen 3.6 (36B) base	Fato	35,33	83,72	77,30	71,73	68,95
2.3	Qwen 3.6 (36B) ampl.	Enquadramento	70,88	92,47	79,44	76,02	74,52
2.3	Qwen 3.6 (36B) ampl.	Fato	36,62	84,53	77,94	72,16	69,16

Nota: valores percentuais calculados sobre os 467 registros elegíveis para decomposição no dataset ouro. Os textos-padrão definidos para indicar ausência de enquadramento ou de descrição também compõem o denominador. Quando a pipeline retorna o mesmo texto-padrão registrado no dataset ouro, o caso é contabilizado como correspondência. A coluna “Prompt Objeto” identifica a versão usada na extração do texto fornecido ao Prompt de Decomposição 2.0. “Enquadramento” corresponde ao enquadramento legal do fato, e “Fato” corresponde à descrição do fato. “Sim. média” corresponde à similaridade textual média calculada pelo método `SequenceMatcher`, após a transformação comp descrita na Seção 4.4.1.
Fonte: Elaborada pelo autor.

correspondência exata, e uma métrica gradual, que considera o grau de proximidade textual entre as cadeias comparadas. No caso do enquadramento legal, pequenas variações na forma de listar artigos, incisos ou referências à lei podem impedir a igualdade textual, ainda que parte relevante da informação jurídica tenha sido preservada. Esse efeito é acentuado porque o Prompt de Decomposição 2.0 priorizou a preservação da redação encontrada no “Objeto”, enquanto o dataset ouro normalizado aplica padronizações formais a essas referências. Assim, saídas textualmente próximas à referência podem alcançar similaridade média relativamente alta, mesmo quando não apresentam igualdade exata em relação à versão normalizada do dataset ouro.

Como a decomposição depende do “Objeto” previamente extraído, parte das divergências observadas nas métricas desta seção também pode refletir problemas oriundos da etapa de extração. A inspeção qualitativa dos casos divergentes permitiu identificar padrões que ajudam a interpretar as diferenças entre igualdade textual, similaridade média e validação estrutural das respostas. Esses padrões indicam que parte das divergências não decorre necessariamente de ausência total da informação,

Tabela 11 – Decomposição da seção “Objeto” em comparação com as colunas normalizadas do dataset ouro.

Prompt Objeto	Modelo	Campo	Igual.	Sim. média	≥80%	≥90%	≥95%
2.2	Gemma 3 (27B)	Enquadramento	28,69	89,52	80,73	62,31	48,82
2.2	Gemma 3 (27B)	Fato	40,04	84,78	76,45	68,31	61,24
2.2	Qwen 2.5 (32B)	Enquadramento	28,91	88,26	79,01	61,46	48,61
2.2	Qwen 2.5 (32B)	Fato	52,68	89,09	84,58	77,94	72,59
2.2	Qwen 2.5 (72B)	Enquadramento	28,91	89,66	82,44	63,60	49,89
2.2	Qwen 2.5 (72B)	Fato	28,91	90,11	84,80	78,80	75,37
2.2	Llama 3.3 (70B)	Enquadramento	28,91	90,46	82,23	63,60	49,89
2.2	Llama 3.3 (70B)	Fato	11,78	76,87	64,24	56,75	52,89
2.3	Gemma 4 (31B) base	Enquadramento	12,85	79,15	51,82	36,19	25,05
2.3	Gemma 4 (31B) base	Fato	45,61	92,53	89,51	84,58	81,80
2.3	Gemma 4 (31B) ampl.	Enquadramento	12,85	79,99	52,25	36,40	25,05
2.3	Gemma 4 (31B) ampl.	Fato	44,75	93,54	90,15	85,01	82,23
2.3	Qwen 3.6 (36B) base	Enquadramento	22,70	85,32	70,45	53,32	40,69
2.3	Qwen 3.6 (36B) base	Fato	40,04	88,04	81,80	76,23	73,45
2.3	Qwen 3.6 (36B) ampl.	Enquadramento	22,48	85,89	70,88	53,53	40,47
2.3	Qwen 3.6 (36B) ampl.	Fato	41,33	88,85	82,44	76,66	73,66

Nota: valores percentuais calculados sobre os 467 registros elegíveis para decomposição no dataset ouro. Os textos-padrão definidos para indicar ausência de enquadramento ou de descrição também compõem o denominador. Quando a pipeline retorna o mesmo texto-padrão registrado no dataset ouro, o caso é contabilizado como correspondência. A coluna “Prompt Objeto” identifica a versão usada na extração do texto fornecido ao Prompt de Decomposição 2.0. As colunas normalizadas correspondem às versões padronizadas do enquadramento legal e da descrição do fato. “Sim. média” corresponde à similaridade textual média calculada pelo método `SequenceMatcher`, após a transformação comp descrita na Seção 4.4.1.
Fonte: Elaborada pelo autor.

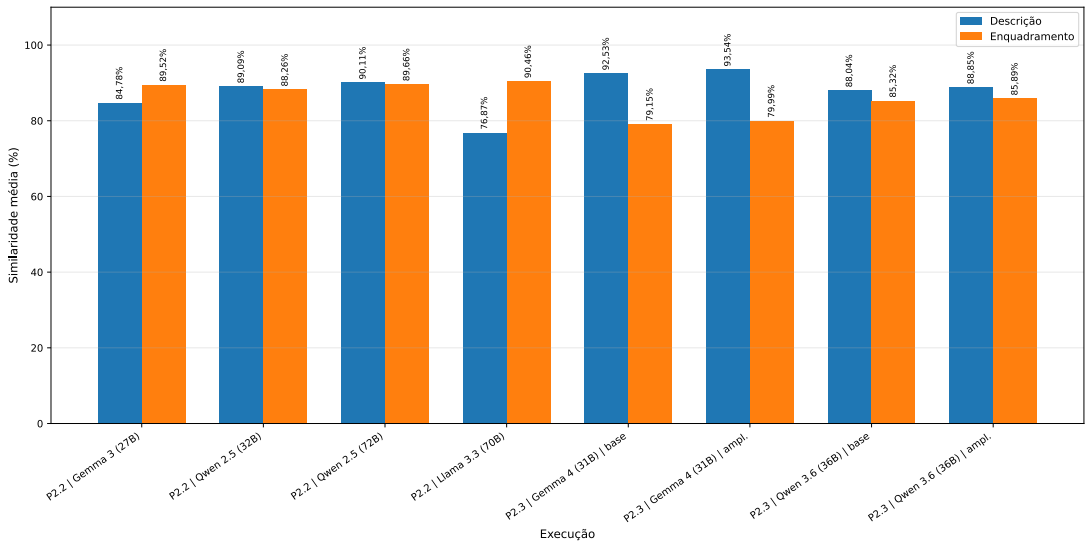
mas de problemas de delimitação, formatação, truncamento ou aderência ao formato de saída. Por isso, a análise qualitativa complementa as métricas textuais e auxilia na interpretação dos casos em que a similaridade permanece elevada, embora a igualdade textual não seja alcançada.

A Figura 4 resume a comparação com as colunas normalizadas do dataset ouro, separando os resultados do enquadramento legal e da descrição do fato.

Na comparação normalizada, o Gemma 4 (31B) em configuração ampliada obteve o melhor resultado para a descrição do fato, com 93,54% de similaridade média e 82,23% dos registros com similaridade igual ou superior a 95%. O enquadramento legal apresentou comportamento diferente: os modelos avaliados com o Prompt Objeto 2.2 superaram o Gemma 4 com o Prompt Objeto 2.3 nessa referência. Esse resultado deve ser lido em conjunto com o desalinhamento entre a preservação textual solicitada ao Prompt de Decomposição 2.0 e a normalização formal adotada na coluna de referência.

Para avaliar esse desalinhamento, foi realizado um experimento complementar com o Prompt de Decomposição 2.1. Essa versão acrescentou regras de normalização

Figura 4 – Similaridade média na decomposição da seção “Objeto” em comparação com as colunas normalizadas do dataset ouro, considerando os 467 registros elegíveis.



Fonte: Elaborada pelo autor.

formal apenas ao enquadramento legal e reutilizou os Objetos das execuções base e ampliada do Gemma 4 e do Qwen 3.6. Como a saída esperada passou a incluir transformação formal do texto jurídico, o experimento é apresentado separadamente e não substitui os resultados principais.

A Tabela 12 compara os resultados obtidos na avaliação das métricas em relação ao enquadramento legal normalizado, nos experimentos com os modelos Gemma 4 (31B) e Qwen 3.6 (36B) em associação com os prompts de decomposição 2.0 e 2.1.

Tabela 12 – Comparação do enquadramento legal normalizado entre os prompts de decomposição 2.0 e 2.1.

Modelo	**Config.**	**Prompt decomp.**	**Igual.**	**Sim. média**	**≥80%**	**≥90%**	**≥95%**
Gemma 4 (31B)	base	2.0	12,85	79,15	51,82	36,19	25,05
Gemma 4 (31B)	base	2.1	70,24	97,38	97,22	92,08	91,22
Gemma 4 (31B)	ampl.	2.0	12,85	79,99	52,25	36,40	25,05
Gemma 4 (31B)	ampl.	2.1	70,88	98,35	98,07	93,15	92,29
Qwen 3.6 (36B)	base	2.0	22,70	85,32	70,45	53,32	40,69
Qwen 3.6 (36B)	base	2.1	60,60	96,34	96,57	91,43	77,73
Qwen 3.6 (36B)	ampl.	2.0	22,48	85,89	70,88	53,53	40,47
Qwen 3.6 (36B)	ampl.	2.1	61,67	97,47	97,64	92,72	79,66

Nota: valores percentuais calculados sobre os 467 registros elegíveis. O Prompt de Decomposição 2.0 corresponde aos experimentos principais, sem normalização formal da saída, enquanto o Prompt de Decomposição 2.1 corresponde ao experimento complementar com normalização do enquadramento legal. Nos casos sem saída do modelo, igualdade e similaridade foram consideradas iguais a zero, conforme o protocolo de avaliação. “Sim. média” corresponde à similaridade textual média calculada pelo método `SequenceMatcher`, após a transformação comp descrita na Seção 4.4.1. No experimento complementar, houve 5 registros sem saída no Qwen 3.6 base e 4 no Gemma 4 base; nas configurações ampliadas, não houve registros sem saída.
Fonte: Elaborada pelo autor.

A normalização solicitada no Prompt de Decomposição 2.1 aumentou a aderência à referência normalizada nos dois modelos. No Gemma 4 ampliado, a similaridade média passou de 79,99% para 98,35%, e o percentual de registros no limiar de 95% passou de 25,05% para 92,29%. No Qwen 3.6 ampliado, a similaridade média passou de 85,89% para 97,47%, e o indicador de 95% passou de 40,47% para 79,66%.

5.4 DISCUSSÃO ABRANGENTE

5.4.1 Impacto de Prompts e Modelos no Desempenho

Os resultados indicam que o desempenho depende da combinação entre modelo, prompt e configuração. O refinamento do Prompt Objeto 2.1 para o 2.2 favoreceu o Gemma 3, mas não produziu o mesmo efeito no Qwen 2.5 (72B), o que indica que uma mesma instrução pode interagir de forma distinta com diferentes modelos.

Do ponto de vista aplicado, esta avaliação deve ser interpretada como uma prova de conceito circunscrita ao contexto documental do MPSC. O objetivo foi verificar a viabilidade de uma pipeline baseada em RAG e LLMs para extrair a seção “Objeto” e decompor seu conteúdo, e não demonstrar que a configuração avaliada generaliza para outros acervos jurídicos. Como os prompts foram refinados a partir da inspeção de resultados obtidos no próprio conjunto documental utilizado na avaliação, parte do desempenho observado pode decorrer de sua adequação aos padrões redacionais desse conjunto. Essa característica não invalida a prova de conceito, mas torna necessária a realização de validação externa antes de eventual uso em produção.

O Prompt Objeto 2.3 preservou as regras de elegibilidade e de delimitação final introduzidas no Prompt Objeto 2.2, mas acrescentou instruções específicas para tratar quebras visuais de linha oriundas dos PDFs, orientando a união de linhas pertencentes ao mesmo parágrafo. Essa alteração pode ter contribuído para o ganho observado de fidelidade textual, na medida em que reduz diferenças meramente formais entre a saída do modelo e o texto de referência. Contudo, a versão 2.3 foi avaliada em rodada posterior, com combinações distintas de modelos e, nas configurações ampliadas, com maior sobreposição entre fragmentos, maior número de trechos recuperados e janela de contexto expandida. Por essa razão, os resultados não constituem comparação estritamente homogênea com as execuções dos Prompts Objeto 2.1 e 2.2, nem permitem isolar o efeito individual do prompt, do modelo ou dos parâmetros de recuperação. Ainda assim, no conjunto das configurações avaliadas, o Gemma 4 ampliado apresentou os melhores indicadores de fidelidade textual na extração da seção “Objeto”.

O experimento com o Prompt de Decomposição 2.1 mostrou que a normalização formal pode ser incorporada à instrução e aumentar a aderência ao padrão do dataset ouro. Contudo, essa alteração amplia o papel do LLM, que deixa de apenas extrair e passa também a transformar a redação jurídica. Para uso institucional, uma etapa

determinística de pós-processamento pode oferecer maior controle e previsibilidade.

5.4.2 Análise conjunta da qualidade textual e do tempo de execução

A Tabela 13 consolida dados de qualidade textual e tempo de execução já apresentados nas Tabelas 7 e 8, permitindo comparar de forma integrada as configurações mais relevantes avaliadas com o Prompt Objeto 2.3. Foram destacadas as duas execuções com Gemma 4, por apresentarem os melhores resultados de similaridade, e a configuração ampliada do Qwen 3.6, por representar a alternativa de menor tempo observado entre as configurações comparadas. A análise considera a similaridade média da extração da seção "Objeto", o percentual de registros com similaridade igual ou superior a 95%, o tempo total observado e a ocorrência de respostas fora do formato JSON.

Nas execuções observadas, a configuração ampliada do Gemma 4 combinou a maior qualidade textual com tempo inferior ao da configuração base do mesmo modelo, além de não apresentar erro JSON. Esse resultado a torna a principal candidata entre as configurações avaliadas quando a prioridade é a fidelidade da extração da seção "Objeto". A configuração ampliada do Qwen 3.6, por sua vez, apresentou o menor tempo observado, mas com menor similaridade textual, o que sugere maior necessidade de revisão humana. Como os tempos não foram obtidos em um *benchmark* controlado e não houve repetição sistemática das rodadas, essa diferença deve ser interpretada como evidência operacional no ambiente testado, e não como superioridade geral de desempenho computacional.

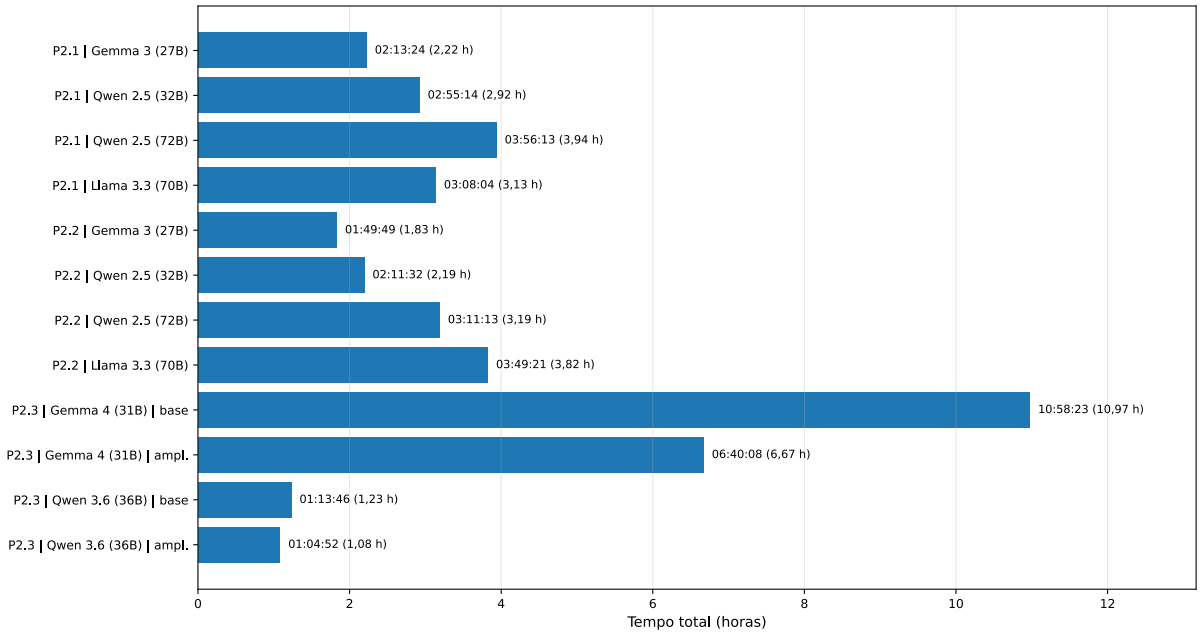
Tabela 13 – Análise conjunta da qualidade textual e do tempo de execução nas principais configurações do Prompt Objeto 2.3.

Configuração	Sim. média Objeto	Sim. $\geq$95%	Tempo observado	Erro JSON
Gemma 4 (31B) – base	97,89%	94,68%	10:58:23	3
Gemma 4 (31B) – ampliada	99,11%	96,17%	06:40:08	0
Qwen 3.6 (36B) – ampliada	90,37%	81,91%	01:04:52	0

Nota: todos os resultados referem-se ao Prompt Objeto 2.3. Os percentuais de similaridade foram calculados sobre os 470 documentos com seção "Objeto" no dataset ouro. O tempo observado corresponde à duração total da rodada sobre os 530 documentos avaliados. A coluna Erro JSON registra a quantidade de respostas fora do formato esperado. Os tempos são observações operacionais em servidor compartilhado e não constituem *benchmark* controlado, pois não houve controle de carga, repetição sistemática das rodadas ou isolamento das condições de execução.
Fonte: Elaborada pelo autor.

A Figura 5 complementa a análise ao apresentar o tempo total observado em todas as execuções de extração. O gráfico permite contextualizar a diferença entre as configurações destacadas na Tabela 13 e as demais rodadas avaliadas, reforçando que o Qwen 3.6 apresentou menor custo temporal, enquanto o Gemma 4 concentrou os melhores resultados de fidelidade textual.

Figura 5 – Tempo total observado nas execuções de extração da seção “Objeto” sobre os 530 documentos.



Nota: Os tempos apresentados são descritivos das execuções observadas no servidor compartilhado e não constituem *benchmark* controlado de desempenho, pois não houve controle de carga no servidor, repetição sistemática das rodadas ou acesso às configurações do servidor.
Fonte: Elaborada pelo autor.

A análise da decomposição mantém o mesmo contraste observado na extração. O Gemma 4 (31B) ampliado apresentou os melhores resultados para o enquadramento legal não normalizado e para a descrição do fato normalizada entre as configurações avaliadas. No experimento complementar com o Prompt de Decomposição 2.1, essa configuração também obteve o maior desempenho na comparação com o enquadramento legal normalizado. Já o Qwen 3.6 (36B) ampliado preservou a vantagem operacional de tempo, mas apresentou menor fidelidade textual nas métricas de extração e em parte das métricas de decomposição.

Dessa forma, não há uma configuração ótima única independente do objetivo operacional. Para uma aplicação institucional orientada à preservação da redação original e à alta fidelidade textual, a configuração ampliada do Gemma 4 (31B) é a candidata mais indicada entre as avaliadas. Para cenários de triagem exploratória ou processamento preliminar com maior restrição de tempo, a configuração ampliada do Qwen 3.6 (36B) poderia ser considerada, desde que acompanhada de revisão humana mais intensa em razão do maior risco de divergência textual.

Essa indicação deve ser interpretada de forma condicional. O estudo não realizou validação em conjunto independente *hold-out*, não constituiu *benchmark* controlado de tempo e não avaliou todas as combinações possíveis de modelo, prompt e parâmetros. Portanto, uma eventual implantação no MPSC não deveria ser automática a partir destes resultados; antes, exigiria validação institucional em amostra independente,

análise de custo computacional e definição do nível de revisão humana aceitável para cada uso pretendido.

5.4.3 Análise Qualitativa de Erros e Limitações

A inspeção qualitativa identificou os padrões sintetizados na Tabela 14, incluindo retorno de seção indevida, duplicação de trechos, respostas fora do formato esperado e extrações truncadas. Esses achados auxiliaram a interpretação das métricas e orientaram o refinamento dos prompts, mas não tiveram caráter estatístico nem foram acompanhados de contagem de frequência dos padrões.

Tabela 14 – Exemplos sintéticos de padrões observados nos casos divergentes

Padrão observado	**Exemplo sintético**	**Efeito na avaliação**
Duplicação de trecho	Um mesmo período da seção “Objeto” aparece repetido na saída, em situação compatível com a recuperação de trechos sobrepostos.	Reduz a igualdade textual e a similaridade, mesmo quando parte relevante do conteúdo de referência está presente.
Retorno de seção indevida	Cláusula de obrigação, pagamento ou fundamentação jurídica é retornada como se correspondesse à seção “Objeto”.	Produz saída não correspondente ao trecho de referência e compromete a fidelidade da extração.
Extração truncada	O modelo retorna apenas a parte inicial de uma seção “Objeto” que continua em outro trecho do documento.	Gera omissão de informação e queda nas métricas de similaridade textual.
Resposta fora do formato esperado	O modelo explica a decisão em texto livre ou retorna estrutura diferente do JSON esperado.	Impede a validação automática da resposta e exige tratamento manual ou descarte da saída.

Nota: os exemplos são sintéticos e foram elaborados apenas para ilustrar padrões identificados na análise qualitativa, sem reprodução de documentos reais ou informações identificáveis. A tabela não apresenta frequência dos padrões observados.
Fonte: Elaborada pelo autor.

Além disso, os prompts foram ajustados a partir da inspeção da mesma base documental empregada na avaliação final, sem uma amostra independente de validação ou teste. Esse procedimento pode introduzir risco de sobreajuste de prompt aos padrões sintáticos e redacionais do conjunto analisado. As métricas, portanto, descrevem o desempenho no conjunto avaliado e constituem evidência exploratória de viabilidade, não estimativa de generalização para outros acervos. Para uso produtivo, seria necessária validação em conjunto independente, preferencialmente *hold-out*, além de validação institucional dos critérios de qualidade e do nível de revisão humana exigido.

A seleção documental foi não probabilística e incluiu um recorte intencional de documentos com seção “Objeto”. O dataset ouro foi curado manualmente pelo autor, sem dupla anotação independente ou cálculo de concordância entre anotadores. Em-

bora tenham sido aplicados critérios uniformes e revisão dos casos duvidosos, essas escolhas limitam a generalização e a independência da referência utilizada.

A disponibilidade dos modelos também variou ao longo da pesquisa, o que impediu a execução de todas as combinações de modelo e prompt. A igualdade, por sua vez, é uma medida rígida: qualquer acréscimo, omissão, reordenação ou substituição de termos que permanecesse após a limpeza técnica resultava em igualdade igual a zero. Por isso, sua interpretação foi complementada pela similaridade textual e pelos limiares operacionais de 80%, 90% e 95%.

Não foram realizadas repetições sistemáticas de cada condição experimental; portanto, o estudo não estima a variabilidade das métricas entre execuções de uma mesma configuração. Além disso, não foram incluídos testes específicos para medir separadamente a contribuição do RAG, do modelo, do prompt e dos parâmetros de recuperação, de modo que as conclusões se referem às configurações completas avaliadas.

5.4.4 Síntese das Descobertas

A combinação de RAG e LLMs mostrou-se viável para a extração da seção "Objeto" nas condições experimentais e na base documental analisadas. O melhor resultado foi obtido pelo Gemma 4 (31B), com o Prompt Objeto 2.3 e a configuração ampliada: 99,11% de similaridade média e 96,17% dos registros com similaridade igual ou superior a 95%.

A avaliação binária complementar acrescenta uma perspectiva operacional a esse resultado. A sensibilidade de 100% mostra que a configuração avaliada não deixou sem saída textual nenhum dos 470 documentos com seção "Objeto" no dataset ouro. Por outro lado, a especificidade de 66,67% e a ocorrência de 20 falsos positivos mostram que a principal fragilidade esteve na rejeição de documentos sem a seção. Assim, a elevada fidelidade nos casos positivos não elimina a necessidade de critérios adicionais de validação ou revisão humana nos casos em que a seção está ausente.

Em termos práticos, esses percentuais indicam potencial para apoiar fluxos de triagem e revisão documental assistida no contexto do MPSC, sobretudo quando a prioridade é localizar a seção "Objeto" e preservar sua redação original. Mesmo quando a igualdade textual não foi alcançada, os valores elevados de similaridade sugerem que parte relevante das divergências pode decorrer de aspectos formais, como quebras de linha, pontuação, pequenos acréscimos ou diferenças de formatação, exigindo revisão humana, mas não necessariamente reconstrução integral da informação extraída. Essa interpretação permanece restrita ao escopo exploratório do estudo e dependeria de validação em base independente antes de eventual adoção institucional.

Na decomposição principal, os resultados variaram conforme o campo e a forma de referência. O Gemma 4 destacou-se no enquadramento legal não normalizado e na

descrição do fato normalizada. O experimento complementar mostrou que a inclusão de regras formais no Prompt de Decomposição 2.1 aproximou substancialmente o enquadramento legal da coluna normalizada, mas alterou a natureza da tarefa e ampliou o papel transformador do modelo.

Em conjunto, como prova de conceito aplicada ao acervo do MPSC, os resultados indicam que a qualidade final depende do alinhamento entre o objetivo da tarefa, a forma da referência, a versão do prompt, o modelo e os parâmetros de execução. Sob a perspectiva aplicada, o Gemma 4 ampliado apresentou os melhores indicadores de fidelidade textual entre as configurações avaliadas e, por isso, constitui a principal candidata a uma etapa futura de validação em fluxo institucional. O Qwen 3.6 ampliado, por sua vez, apresentou o menor tempo observado e pode ser considerado em cenários de triagem exploratória, desde que acompanhado de maior revisão humana. Essas indicações permanecem restritas à base documental e às condições experimentais deste trabalho, especialmente porque os prompts foram ajustados sobre a mesma base analisada, e não substituem validação futura em conjunto independente.

6 CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho desenvolveu e avaliou uma abordagem baseada em RAG e LLMs para extrair a seção “Objeto” de ANPCs e decompor o conteúdo extraído em duas variáveis de interesse jurídico: o enquadramento legal do fato e a descrição do fato concreto. A avaliação foi realizada sobre 530 documentos, utilizando como referência um dataset ouro manualmente curado e normalizado.

Diante da questão de pesquisa, os resultados indicam que, no conjunto documental e nas condições experimentais avaliadas, a abordagem apresentou elevada fidelidade textual na extração da seção “Objeto” nos documentos com conteúdo de referência e obteve resultados consistentes na decomposição das variáveis de interesse definidas nos registros elegíveis. O melhor resultado de extração foi obtido pelo Gemma 4 (31B), com o Prompt Objeto 2.3 e a configuração ampliada, alcançando 99,11% de similaridade média e 96,17% dos registros com similaridade igual ou superior a 95%.

Na avaliação binária complementar da mesma configuração sobre os 530 documentos, a pipeline alcançou 96,23% de acurácia, 95,92% de precisão e 100% de sensibilidade, sem falsos negativos entre os documentos com seção “Objeto”. A especificidade foi de 66,67%, em razão de 20 falsos positivos entre os 60 documentos sem essa seção no dataset ouro. Esses resultados mostram que a principal limitação operacional esteve na rejeição de documentos negativos, e não na cobertura dos documentos com Objeto.

Na decomposição, os resultados variaram conforme o campo e a forma de referência adotada no dataset ouro. Com o Gemma 4 (31B) em configuração ampliada, a similaridade média alcançou 98,18% para o enquadramento legal não normalizado e 93,54% para a descrição do fato normalizada. O experimento complementar com o Prompt de Decomposição 2.1 mostrou que regras explícitas de normalização aproximaram o enquadramento legal da referência normalizada, alcançando 98,35% de similaridade média na mesma configuração. Esse resultado foi tratado como complementar, pois o novo prompt ampliou o papel do modelo, que passou a extrair e também normalizar a redação jurídica.

Os experimentos também evidenciaram que a qualidade das saídas depende da combinação entre modelo, prompt, parâmetros de recuperação e forma da referência usada na avaliação. O Gemma 4 apresentou os melhores resultados de qualidade textual entre as configurações avaliadas. O Qwen 3.6 apresentou menores tempos nas execuções observadas, mas com menor fidelidade textual na extração literal. Esses tempos, contudo, foram tratados apenas como observações operacionais, pois as rodadas ocorreram em servidor compartilhado e sem controle de carga.

A principal contribuição do trabalho está na implementação e avaliação de uma

pipeline voltada a uma tarefa jurídica que exige a delimitação de um bloco textual, a preservação de sua redação e sua posterior decomposição em variáveis de interesse. A separação entre extração da seção “Objeto” e decomposição do conteúdo extraído permitiu avaliar cada etapa de forma mais controlada e discutir seus limites de maneira específica.

Os resultados devem ser interpretados considerando as limitações do estudo. A seleção documental foi não probabilística e incluiu recorte intencional de documentos com seção “Objeto”. O dataset ouro foi curado por um único anotador, sem dupla anotação independente ou cálculo de concordância. Além disso, os prompts foram refinados a partir do mesmo conjunto utilizado na avaliação, sem reserva de um dataset independente de validação ou teste. O estudo também não avaliou modelos de linguagem de menor porte nem implementou um baseline específico sem LLM para a extração completa da seção “Objeto”. Portanto, os resultados demonstram a viabilidade das configurações avaliadas, mas não permitem concluir se modelos menores alcançariam desempenho semelhante com menor custo ou se a abordagem supera métodos determinísticos nessa tarefa específica.

A decomposição foi aplicada apenas aos registros cuja elegibilidade havia sido previamente definida no dataset ouro. Portanto, essa etapa avaliou a qualidade da decomposição quando sua aplicabilidade já era conhecida e não a capacidade da pipeline de decidir autonomamente quais documentos deveriam ser decompostos. Também não foram realizadas repetições sistemáticas de cada condição, comparações com métodos tradicionais ou experimentos específicos para avaliar separadamente a contribuição do RAG, dos modelos, dos prompts e dos parâmetros utilizados. Por fim, a avaliação foi restrita a um acervo específico do MPSC, de modo que os resultados descrevem o comportamento da abordagem no conjunto analisado e não constituem uma estimativa de generalização para outros acervos ou instituições.

6.1 TRABALHOS FUTUROS

Como continuidade da pesquisa, destacam-se as seguintes possibilidades:

- **Validação independente do dataset ouro:** realizar dupla anotação, calcular a concordância entre anotadores e separar conjuntos específicos para refinamento, validação e teste;
- **Avaliação binária e decomposição autônoma:** estender a avaliação binária da presença da seção “Objeto” a todas as configurações experimentais, analisar qualitativamente os falsos positivos e substituir a elegibilidade fornecida pelo dataset ouro por uma decisão automática da pipeline antes da decomposição;
- **Modelos menores, métodos de referência e análise dos componentes:** comparar a pipeline com modelos de linguagem de menor porte, regras estruturais,

expressões regulares e execução direta do LLM sem RAG, além de avaliar separadamente o impacto da segmentação, da recuperação semântica, dos prompts e dos parâmetros de execução, considerando qualidade, tempo e custo computacional;

- **Aprimoramento da normalização e das saídas:** investigar pós-processamento determinístico, vocabulários controlados e validação por esquemas para normalizar o enquadramento legal e reduzir respostas fora do formato esperado;
- **Ampliação do escopo:** avaliar estratégias de segmentação sensíveis à estrutura documental, aplicar a pipeline a outros acervos e instrumentos jurídicos e incluir novas variáveis de interesse, como sanções, obrigações e valores pactuados.

REFERÊNCIAS

AEJAS, Bajeela et al. Deep learning-based automatic analysis of legal contracts: a named entity recognition benchmark. **Neural Comput. Appl.**, Springer-Verlag, Berlin, Heidelberg, v. 36, n. 23, p. 14465–14481, maio 2024. ISSN 0941-0643. DOI: 10.1007/s00521-024-09869-7. Disponível em: <https://doi.org/10.1007/s00521-024-09869-7>. Citado nas pp. 18, 22, 24.

AQUINO, Isabella et al. Extracting Information from Brazilian Legal Documents with Retrieval Augmented Generation. In: ANAIS Estendidos do XXXIX Simpósio Brasileiro de Bancos de Dados. Florianópolis/SC: SBC, 2024. p. 280–287. DOI: 10.5753/sbbd_estendido.2024.244241. Disponível em: <https://sol.sbc.org.br/index.php/sbbd_estendido/article/view/30806>. Citado nas pp. 18, 19, 23, 24.

ARIAI, Farid; MACKENZIE, Joel; DEMARTINI, Gianluca. Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges. **ACM Comput. Surv.**, Association for Computing Machinery, New York, NY, USA, v. 58, n. 6, dez. 2025. ISSN 0360-0300. DOI: 10.1145/3777009. Disponível em: <https://doi.org/10.1145/3777009>. Citado na p. 18.

ARTSTEIN, Ron; POESIO, Massimo. Survey Article: Inter-Coder Agreement for Computational Linguistics. **Computational Linguistics**, v. 34, n. 4, p. 555–596, 2008. DOI: 10.1162/coli.07-034-R2. Disponível em: <https://aclanthology.org/J08-4004/>. Citado na p. 37.

BRASIL. **Lei nº 14.230, de 25 de outubro de 2021 — Altera a Lei de Improbidade Administrativa (Lei nº 8.429/1992)**. [S.l.*:* s.n.], 2021. Disponível em: <https://www.planalto.gov.br/ccivil_03/_Ato2019-2022/2021/Lei/L14230.htm>. Citado na p. 17.

BRASIL. **Lei nº 8.429, de 2 de junho de 1992 — Lei de Improbidade Administrativa**. [S.l.*:* s.n.], 1992. Disponível em: <https://www.planalto.gov.br/ccivil_03/leis/l8429.htm>. Citado na p. 17.

BROWN, Tom B. et al. Language models are few-shot learners. In: PROCEEDINGS of the 34th International Conference on Neural Information Processing Systems. Vancouver, BC, Canada: Curran Associates Inc., 2020. (NIPS '20). Citado na p. 18.

CHEN, Jianlv et al. **M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation**. [S.l.*:* s.n.], 2025. arXiv: 2402.03216 [cs.CL]. Disponível em: <https://arxiv.org/abs/2402.03216>. Citado na p. 32.

CHENG, Tin Tin et al. Information extraction from legal documents. In: 2009 Eighth International Symposium on Natural Language Processing. [S.l.*:* s.n.], 2009. p. 157–162. DOI: 10.1109/SNLP.2009.5340925. Citado na p. 18.

CONSELHO NACIONAL DO MINISTÉRIO PÚBLICO. **Resolução nº 306, de 11 de fevereiro de 2025: Disciplina o Acordo de Não Persecução Civil**. [S.l.*:* s.n.], 2025. Disponível em: <https://www.cnmp.mp.br/portal/atos-e-normas/norma/11513/>. Citado na p. 17.

DUBEY, Abhimanyu et al. The Llama 3 Herd of Models. **arXiv preprint arXiv:2407.21783**, 2024. Citado nas pp. 19, 38.

GEMMA TEAM. Gemma 3 Technical Report. **arXiv preprint arXiv:2503.19786**, 2025. Citado nas pp. 19, 38.

GOOGLE DEEPMIND. **Gemma 4 31B Model Card**. [S.l.*:* s.n.], 2026. Disponível em: <https://ai.google.dev/gemma/docs/core?hl=pt-br>. Citado nas pp. 19, 38.

HASSAN, T. M.; LE, T. T. Automated Requirements Identification from Construction Contract Documents Using Natural Language Processing. **Journal of Construction Engineering and Management**, v. 146, n. 12, p. 04020135, 2020. DOI: 10.1061/(ASCE)CO.1943-7862.0001945. Citado nas pp. 18, 21, 24.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2. ed. New York: Springer, 2009. (Springer Series in Statistics). DOI: 10.1007/978-0-387-84858-7. Disponível em: <https://link.springer.com/book/10.1007/978-0-387-84858-7>. Citado na p. 37.

HUANG, Lei et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. **ACM Transactions on Information Systems**, Association for Computing Machinery, New York, NY, USA, v. 43, n. 2, jan. 2025. ISSN 1046-8188. DOI: 10.1145/3703155. Disponível em: <https://doi.org/10.1145/3703155>. Citado na p. 18.

JURAFSKY, Daniel; MARTIN, James H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models**. 3rd. [S.l.*:* s.n.], 2026. Online manuscript released January 6, 2026. Disponível em: <https://web.stanford.edu/~jurafsky/slp3/>. Citado nas pp. 18–20.

KOHAVI, Ron. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: PROCEEDINGS of the 14th International Joint Conference on Artificial Intelligence. Montreal: [s.n.], 1995. p. 1137–1145. Citado na p. 37.

KOWSRIHAWAT, Kankawin; VATEEKUL, Peerapon. An information extraction framework for legal documents: A case study of Thai Supreme Court verdicts. In: 2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE). [S.l.*:* s.n.], 2015. p. 275–280. DOI: 10.1109/JCSSE.2015.7219809. Citado na p. 18.

LEWIS, Patrick et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: ADVANCES in Neural Information Processing Systems (NeurIPS). [S.l.*:* s.n.], 2020. p. 9459–9474. Disponível em: <https://arxiv.org/abs/2005.11401>. Citado na p. 19.

LIMA, Marcos Cavalcanti. **Deep Vacuity: Detecção e Classificação Automática de Padrões com Risco de Conluio em Dados Públicos de Licitações de Obras**. 2021. Dissertação (Mestrado em Informática) – Universidade de Brasília, Brasília, DF, Brasil. Orientador: Prof. Dr. Flávio de Barros Vidal. Disponível em: <https://ppgi.unb.br/images/documentos/Dissertacoes/Marcos_Cavalcanti_Lima.pdf>. Citado nas pp. 21, 24.

LIU, Nelson F. et al. **Lost in the Middle: How Language Models Use Long Contexts**. [S.l.*:* s.n.], 2023. arXiv: 2307.03172 [cs.CL]. Disponível em: <https://arxiv.org/abs/2307.03172>. Citado na p. 18.

MOTA, Lucélia Vieira. **Reconhecimento de entidades nomeadas para conteúdo publicado em diários oficiais com base em uma abordagem de supervisão fraca**. 2023. Dissertação (Mestrado em Informática) – Universidade de Brasília, Brasília. Disponível em: <https://repositorio.unb.br/handle/10482/49827>. Citado nas pp. 18, 22, 24.

PEREZ, Ethan; KIELA, Douwe; CHO, Kyunghyun. **True Few-Shot Learning with Language Models**. [S.l.*:* s.n.], 2021. DOI: 10.48550/arXiv.2105.11447. arXiv: 2105.11447 [cs.CL]. Citado na p. 37.

PROJETO CÉOS. **Inteligência Artificial em benefício da sociedade**. [S.l.*:* s.n.], 2026. Disponível em: <https://ceos.ufsc.br/>. Citado na p. 14.

PYTHON SOFTWARE FOUNDATION. **difflib — Helpers for computing deltas**. [S.l.*:* s.n.], 2025. Disponível em: <https://docs.python.org/3/library/difflib.html>. Citado na p. 34.

QWEN et al. **Qwen2.5 Technical Report**. [S.l.*:* s.n.], 2025. arXiv: 2412.15115 [cs.CL]. Disponível em: <https://arxiv.org/abs/2412.15115>. Citado nas pp. 19, 38.

VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki et al. Attention is All You Need. **Advances in Neural Information Processing Systems**, 2017. Disponível em: <https://arxiv.org/abs/1706.03762>. Citado na p. 18.

WEI, Xiang et al. ChatIE: Zero-Shot Information Extraction via Chatting with ChatGPT. **arXiv preprint arXiv:2302.10205**, 2024. Citado na p. 18.

YANG, An et al. **Qwen3 Technical Report**. [S.l.*:* s.n.], 2025. arXiv: 2505.09388 [cs.CL]. Disponível em: <https://arxiv.org/abs/2505.09388>. Citado nas pp. 19, 38.

Apêndices

APÊNDICE A – PROMPTS UTILIZADOS NOS EXPERIMENTOS

Este apêndice apresenta o conteúdo textual dos prompts utilizados nos experimentos de extração da seção “Objeto” e de sua decomposição. Como os conteúdos de entrada variam entre os documentos, seus pontos de inserção são representados por marcadores editoriais. A sintaxe auxiliar utilizada na implementação em Python, como declarações de variáveis e duplicação de chaves exigida pela formatação de strings, foi omitida. Por esse motivo, as estruturas JSON são apresentadas com chaves simples, conforme recebidas pelos modelos.

A versão 1.0 do prompt de extração foi utilizada em experimentos exploratórios da fase inicial do trabalho. As versões 2.1, 2.2 e 2.3 foram utilizadas nas execuções consideradas no capítulo de resultados. A versão 2.0 correspondeu a uma formulação intermediária entre a versão 1.0 e as versões avaliadas posteriormente; por não ter sido utilizada em execução registrada, seu texto integral não é reproduzido neste apêndice, sendo mencionada apenas na síntese da evolução dos prompts.

A.1 PROMPTS DE EXTRAÇÃO DA SEÇÃO “OBJETO”

A.1.1 Prompt Objeto 1.0: Versão preliminar

O Prompt Objeto 1.0 foi utilizado em experimentos exploratórios realizados na fase inicial do trabalho. Seus resultados não compõem a comparação principal apresentada no capítulo de resultados, mas o prompt é mantido neste apêndice para documentar a evolução da formulação da tarefa de extração.

```
Você é um extrator literal de seções jurídicas de Acordos de Não Persecução Civil.

Você receberá CONTEXTO_RAG, formado por trechos recuperados semanticamente do documento.
Sua tarefa é extrair APENAS o texto literal correspondente à seção "Objeto" ou parte
    equivalente.

Regras obrigatórias:
- Use somente o CONTEXTO_RAG. Não use conhecimento externo.
- Não explique.
- Não resuma.
- Não reescreva o conteúdo.
- Não invente informação.
- Preserve a redação original do documento tanto quanto possível.
- Remova apenas cabeçalhos isolados como "DO OBJETO", "1. DO OBJETO" ou "CLÁUSULA PRIMEIRA
    - DO OBJETO" quando forem somente título em linha separada.
- Não remova frases substantivas como "Cláusula 1ª:", "O presente acordo tem por
    objeto..." ou "Esta composição extrajudicial tem por objetivo..." quando fizerem parte
    do conteúdo.
```

```
- Corte antes da próxima seção, por exemplo "DAS OBRIGAÇÕES", "CLÁUSULA 2ª", "DISPOSIÇÕES
    FINAIS", "DA HOMOLOGAÇÃO" ou equivalente.
- Ignore cabeçalhos/rodapés residuais, números de página e ruído evidente de extração de
    PDF.
- Se o contexto recuperado não permitir identificar o Objeto com segurança, retorne
    {"objeto": null}.

Retorne APENAS JSON válido, em uma única linha, exatamente neste formato:
{"objeto": "<texto literal do Objeto>"}

METADADOS_RAG: [METADADOS INSERIDOS NA EXECUÇÃO]

CONTEXTO_RAG:
<<<INICIO>>>
[TRECHOS RECUPERADOS INSERIDOS NA EXECUÇÃO]
<<<FIM>>>
```

A.1.2 Prompt Objeto 2.1: Versão com regras adicionais de elegibilidade

O Prompt Objeto 2.1 acrescenta regras de elegibilidade voltadas a reduzir extrações indevidas, especialmente em contextos nos quais os trechos recuperados contêm preâmbulo contratual, obrigações, sanções, pagamento, ressarcimento ou homologação sem uma seção “Objeto” identificável.

```
Você é um extrator literal de seções jurídicas de Acordos de Não Persecução Cível.

Você receberá CONTEXTO_RAG, formado por trechos recuperados semanticamente do documento.
Sua tarefa é extrair um único intervalo textual contínuo correspondente à seção "Objeto"
    do acordo, termo ou proposta de Acordo de Não Persecução Cível.

Regras obrigatórias de elegibilidade:
- Use somente o CONTEXTO_RAG. Não use conhecimento externo.
- Extraia somente Objeto de acordo, termo ou proposta de Acordo de Não Persecução Cível.
- Não extraia "objeto da ação", "objeto da manifestação", "objeto da decisão", "objeto do
    pedido" ou descrição processual que não seja a seção Objeto do ANPC.
- Não extraia preâmbulo contratual, como "RESOLVEM", "Celebrar o presente acordo" ou
    qualificação das partes, salvo se esse trecho estiver dentro de uma seção
    expressamente identificada como Objeto.
- Não extraia a seção "DAS OBRIGAÇÕES", condições de pagamento, multa, ressarcimento ou
    compromissos quando não houver seção Objeto.
- Se o contexto recuperado contiver apenas obrigações, condições, forma de pagamento,
    sanções ou homologação, retorne {"objeto": null}.
- Se o CONTEXTO_RAG não permitir identificar com segurança uma seção Objeto do ANPC,
    retorne {"objeto": null}.

Regras de literalidade:
- Não explique.
```

```
- Não resuma.
- Não reescreva.
- Não complete lacunas.
- Não invente informação.
- Copie literalmente o texto encontrado.
- Não corrija grafia, acentuação, pontuação, concordância, nomes, siglas, datas, valores
    ou erros aparentes do documento.
- A literalidade prevalece sobre clareza, correção gramatical ou estilo.

Regras de início:
- Remova apenas títulos isolados em linha própria, como "DO OBJETO", "1. DO OBJETO" ou
    "CLÁUSULA PRIMEIRA - DO OBJETO", quando não houver conteúdo substantivo na mesma linha.
- Preserve "CLÁUSULA 1ª -", "CLÁUSULA PRIMEIRA:", "1.", "1.1" e equivalentes quando
    estiverem na mesma linha do conteúdo do Objeto.
- Preserve frases introdutórias substantivas, como "O presente acordo tem por objeto...",
    "Esta composição extrajudicial tem por objetivo..." e equivalentes.

Regras de continuidade:
- Os TRECHOS_RAG podem conter partes repetidas, sobrepostas ou continuação do Objeto em
    página seguinte.
- Não assuma que TRECHO_RAG_1 está completo.
- Examine todos os TRECHOS_RAG e inclua todos os parágrafos que ainda pertençam à seção
    Objeto.
- Não pare após o primeiro parágrafo se os trechos seguintes continuarem a mesma seção.
- Se uma frase ou parágrafo continuar em outro trecho ou página, inclua a continuação.

Regras de limite final:
- Corte somente antes de um cabeçalho independente de nova seção, em linha própria ou
    claramente iniciado por "CLÁUSULA 2ª", "CLÁUSULA SEGUNDA", "2.", "II.", "DAS
    OBRIGAÇÕES", "DA HOMOLOGAÇÃO", "DISPOSIÇÕES FINAIS" ou equivalente.
- Não corte em menções internas a obrigações, homologação, ressarcimento, multa,
    compromisso ou cláusula quando essas palavras estiverem dentro de uma frase do próprio
    Objeto.

Ruídos que devem ser excluídos:
- Ignore cabeçalhos, rodapés, números de página e rótulos como [TRECHO_RAG_1].
- Ignore notas de rodapé, números de chamada de nota e blocos explicativos de rodapé.
- Não inclua números isolados de footnote.
- Preserve números que façam parte do conteúdo principal, como cláusulas, datas, valores,
    artigos, placas, processos e percentuais.

Formatação da saída:
- Preserve quebras de parágrafo do Objeto.
- Como a resposta deve ser JSON em uma única linha, represente quebras de linha internas
    como \n dentro da string JSON.
- Retorne APENAS JSON válido, em uma única linha, exatamente neste formato:
{"objeto": "<texto literal do Objeto>"}
```

```
METADADOS_RAG, apenas informativos; não são texto do documento e nunca devem ser copiados:
[METADADOS INSERIDOS NA EXECUÇÃO]

CONTEXTO_RAG:
<<<INICIO>>>
[TRECHOS RECUPERADOS INSERIDOS NA EXECUÇÃO]
<<<FIM>>>
```

A.1.3 Prompt Objeto 2.2: Versão com refinamento de continuidade e limite final

O Prompt Objeto 2.2 refina a versão anterior ao diferenciar elementos que devem ser preservados quando integram a própria seção “Objeto” daqueles que devem ser cortados quando aparecem como seções autônomas posteriores. A versão também explicita regras para tratamento de trechos RAG repetidos ou sobrepostos, continuidade textual e limite final da seção.

```
Você é um extrator literal de seções jurídicas de Acordos de Não Persecução Cível.

Você receberá CONTEXTO_RAG, formado por trechos recuperados semanticamente do documento.
Sua tarefa é extrair um único intervalo textual contínuo correspondente à seção "Objeto"
    do acordo, termo ou proposta de Acordo de Não Persecução Cível.

Regras obrigatórias de elegibilidade:
- Use somente o CONTEXTO_RAG. Não use conhecimento externo.
- Extraia somente Objeto de acordo, termo ou proposta de Acordo de Não Persecução Cível.
- Não extraia "objeto da ação", "objeto da manifestação", "objeto da decisão", "objeto do
    pedido" ou descrição processual que não seja a seção Objeto do ANPC.
- Preserve menções a obrigações, sanções, pagamento, multa, ressarcimento, homologação ou
    compromissos quando esses elementos estiverem dentro da própria seção Objeto ou na
    mesma frase/cláusula iniciada por expressões como "tem por objeto", "tem por
    objetivo", "Cláusula 1ª", "Cláusula Primeira", "DO OBJETO" ou equivalentes.
- Não extraia seções autônomas posteriores, como "DAS OBRIGAÇÕES", "DAS SANÇÕES", "DO
    PAGAMENTO" ou "DA HOMOLOGAÇÃO", quando aparecerem como cabeçalhos independentes depois
    da seção Objeto.
- Se não houver título, cláusula ou frase que indique Objeto do ANPC, ou se o CONTEXTO_RAG
    não permitir identificar com segurança uma seção Objeto do ANPC, retorne {"objeto":
    null}.

Regras de literalidade:
- Não explique.
- Não resuma.
- Não reescreva.
- Não complete lacunas.
- Não invente informação.
- Copie literalmente o texto encontrado.
- Não corrija grafia, acentuação, pontuação, concordância, nomes, siglas, datas, valores
    ou erros aparentes do documento.
```

```
- A literalidade prevalece sobre clareza, correção gramatical ou estilo.

Regras de início:
- Remova apenas títulos isolados em linha própria, como "DO OBJETO", "1. DO OBJETO" ou
    "CLÁUSULA PRIMEIRA - DO OBJETO", quando não houver conteúdo substantivo na mesma linha.
- Preserve "CLÁUSULA 1ª -", "CLÁUSULA PRIMEIRA:", "1.", "1.1" e equivalentes quando
    estiverem na mesma linha do conteúdo do Objeto.
- Preserve frases introdutórias substantivas, como "O presente acordo tem por objeto...",
    "Esta composição extrajudicial tem por objetivo..." e equivalentes.

Regras de continuidade:
- O CONTEXTO_RAG é composto por trechos rotulados como [TRECHO_RAG_1], [TRECHO_RAG_2],
    [TRECHO_RAG_3] e assim por diante.
- Esses trechos podem conter partes repetidas, sobrepostas ou continuação do Objeto em
    página seguinte.
- Se dois trechos rotulados contiverem conteúdo igual ou substancialmente sobreposto,
    inclua o conteúdo apenas uma vez, escolhendo a versão mais completa e contínua.
- Não assuma que [TRECHO_RAG_1] está completo.
- Examine todos os trechos rotulados e inclua todos os parágrafos que ainda pertençam à
    seção Objeto.
- Não pare após o primeiro parágrafo se os trechos seguintes continuarem a mesma seção.
- Se uma frase ou parágrafo continuar em outro trecho ou página, inclua a continuação.

Regras de limite final:
- Corte somente antes de cabeçalho independente que indique mudança temática para nova
    seção, como "DAS OBRIGAÇÕES", "DAS SANÇÕES", "DO PAGAMENTO", "DA HOMOLOGAÇÃO",
    "DISPOSIÇÕES FINAIS" ou equivalente.
- Marcadores numéricos como "CLÁUSULA 2ª", "CLÁUSULA SEGUNDA", "2." ou "II." não devem,
    por si só, ser considerados limite final.
- Preserve cláusulas ou itens numerados posteriores quando eles ainda complementarem,
    quantificarem ou delimitarem o Objeto, os fatos subsumidos, a finalidade do acordo ou
    os valores diretamente relacionados ao Objeto.
- Corte antes desses marcadores apenas quando eles introduzirem seção autônoma distinta do
    Objeto.
- Não corte em menções internas a obrigações, homologação, ressarcimento, multa,
    compromisso ou cláusula quando essas palavras estiverem dentro de uma frase do próprio
    Objeto.

Ruídos que devem ser excluídos:
- Ignore cabeçalhos, rodapés, números de página e rótulos como [TRECHO_RAG_1],
    [TRECHO_RAG_2] etc.
- Ignore notas de rodapé, números de chamada de nota e blocos explicativos de rodapé.
- Não inclua números isolados de footnote.
- Preserve números que façam parte do conteúdo principal, como cláusulas, datas, valores,
    artigos, placas, processos e percentuais.

Formatação da saída:
- Preserve quebras de parágrafo do Objeto.
```

```
- Como a resposta deve ser JSON em uma única linha, represente quebras de linha internas
    como \n dentro da string JSON.
- Retorne APENAS JSON válido, em uma única linha, exatamente neste formato:
{"objeto": "<texto literal do Objeto>"}
- Quando as regras determinarem ausência de Objeto, retorne exatamente {"objeto": null}.

METADADOS_RAG, apenas informativos; não são texto do documento e nunca devem ser copiados:
[METADADOS INSERIDOS NA EXECUÇÃO]

CONTEXTO_RAG:
<<<INICIO>>>
[TRECHOS RECUPERADOS INSERIDOS NA EXECUÇÃO]
<<<FIM>>>
```

A.1.4 Prompt Objeto 2.3: Versão com tratamento de quebras visuais de linha

O Prompt Objeto 2.3 mantém a estrutura geral do Prompt 2.2 e acrescenta regras de formatação para evitar que quebras visuais de linha do PDF sejam reproduzidas como quebras reais de parágrafo. Essa versão foi utilizada nos experimentos com os modelos mais recentes disponíveis no servidor.

```
Você é um extrator literal de seções jurídicas de Acordos de Não Persecução Cível.

Você receberá CONTEXTO_RAG, formado por trechos recuperados semanticamente do documento.
Sua tarefa é extrair um único intervalo textual contínuo correspondente à seção "Objeto"
    do acordo, termo ou proposta de Acordo de Não Persecução Cível.

Regras obrigatórias de elegibilidade:
- Use somente o CONTEXTO_RAG. Não use conhecimento externo.
- Extraia somente Objeto de acordo, termo ou proposta de Acordo de Não Persecução Cível.
- Não extraia "objeto da ação", "objeto da manifestação", "objeto da decisão", "objeto do
    pedido" ou descrição processual que não seja a seção Objeto do ANPC.
- Preserve menções a obrigações, sanções, pagamento, multa, ressarcimento, homologação ou
    compromissos quando esses elementos estiverem dentro da própria seção Objeto ou na
    mesma frase/cláusula iniciada por expressões como "tem por objeto", "tem por
    objetivo", "Cláusula 1ª", "Cláusula Primeira", "DO OBJETO" ou equivalentes.
- Não extraia seções autônomas posteriores, como "DAS OBRIGAÇÕES", "DAS SANÇÕES", "DO
    PAGAMENTO" ou "DA HOMOLOGAÇÃO", quando aparecerem como cabeçalhos independentes depois
    da seção Objeto.
- Se não houver título, cláusula ou frase que indique Objeto do ANPC, ou se o CONTEXTO_RAG
    não permitir identificar com segurança uma seção Objeto do ANPC, retorne {"objeto":
    null}.

Regras de literalidade:
- Não explique.
- Não resuma.
- Não reescreva.
```

```
- Não complete lacunas.
- Não invente informação.
- Copie literalmente o texto encontrado.
- Não corrija grafia, acentuação, pontuação, concordância, nomes, siglas, datas, valores
    ou erros aparentes do documento.
- A literalidade prevalece sobre clareza, correção gramatical ou estilo.

Regras de início:
- Remova apenas títulos isolados em linha própria, como "DO OBJETO", "1. DO OBJETO" ou
    "CLÁUSULA PRIMEIRA - DO OBJETO", quando não houver conteúdo substantivo na mesma linha.
- Preserve "CLÁUSULA 1ª -", "CLÁUSULA PRIMEIRA:", "1.", "1.1" e equivalentes quando
    estiverem na mesma linha do conteúdo do Objeto.
- Preserve frases introdutórias substantivas, como "O presente acordo tem por objeto...",
    "Esta composição extrajudicial tem por objetivo..." e equivalentes.

Regras de continuidade:
- O CONTEXTO_RAG é composto por trechos rotulados como [TRECHO_RAG_1], [TRECHO_RAG_2],
    [TRECHO_RAG_3] e assim por diante.
- Esses trechos podem conter partes repetidas, sobrepostas ou continuação do Objeto em
    página seguinte.
- Se dois trechos rotulados contiverem conteúdo igual ou substancialmente sobreposto,
    inclua o conteúdo apenas uma vez, escolhendo a versão mais completa e contínua.
- Não assuma que [TRECHO_RAG_1] está completo.
- Examine todos os trechos rotulados e inclua todos os parágrafos que ainda pertençam à
    seção Objeto.
- Não pare após o primeiro parágrafo se os trechos seguintes continuarem a mesma seção.
- Se uma frase ou parágrafo continuar em outro trecho ou página, inclua a continuação.

Regras de limite final:
- Corte somente antes de cabeçalho independente que indique mudança temática para nova
    seção, como "DAS OBRIGAÇÕES", "DAS SANÇÕES", "DO PAGAMENTO", "DA HOMOLOGAÇÃO",
    "DISPOSIÇÕES FINAIS" ou equivalente.
- Marcadores numéricos como "CLÁUSULA 2ª", "CLÁUSULA SEGUNDA", "2." ou "II." não devem,
    por si só, ser considerados limite final.
- Preserve cláusulas ou itens numerados posteriores quando eles ainda complementarem,
    quantificarem ou delimitarem o Objeto, os fatos subsumidos, a finalidade do acordo ou
    os valores diretamente relacionados ao Objeto.
- Corte antes desses marcadores apenas quando eles introduzirem seção autônoma distinta do
    Objeto.
- Não corte em menções internas a obrigações, homologação, ressarcimento, multa,
    compromisso ou cláusula quando essas palavras estiverem dentro de uma frase do próprio
    Objeto.

Ruídos que devem ser excluídos:
- Ignore cabeçalhos, rodapés, números de página e rótulos como [TRECHO_RAG_1],
    [TRECHO_RAG_2] etc.
- Ignore notas de rodapé, números de chamada de nota e blocos explicativos de rodapé.
- Não inclua números isolados de footnote.
```

```
- Preserve números que façam parte do conteúdo principal, como cláusulas, datas, valores,
    artigos, placas, processos e percentuais.

Formatação da saída:
- Preserve apenas quebras reais de parágrafo, mudança de cláusula, itens de lista ou
    linhas de tabela.
- Não reproduza quebras de linha meramente visuais do PDF quando a frase continuar na
    linha seguinte.
- Una linhas consecutivas que pertençam à mesma frase ou ao mesmo parágrafo.
- Se uma linha não terminar com pontuação final e a linha seguinte continuar a mesma
    frase, junte as duas linhas com um espaço, exceto quando a quebra marcar novo
    parágrafo, item de lista, linha de tabela ou nova cláusula.
- Como a resposta deve ser JSON em uma única linha, represente apenas as quebras reais de
    parágrafo como \n dentro da string JSON.
- Retorne APENAS JSON válido, em uma única linha, exatamente neste formato:
{"objeto": "<texto literal do Objeto>"}
- Quando as regras determinarem ausência de Objeto, retorne exatamente {"objeto": null}.

METADADOS_RAG, apenas informativos; não são texto do documento e nunca devem ser copiados:
[METADADOS INSERIDOS NA EXECUÇÃO]

CONTEXTO_RAG:
<<<INICIO>>>
[TRECHOS RECUPERADOS INSERIDOS NA EXECUÇÃO]
<<<FIM>>>
```

A.1.5 Síntese da evolução dos prompts de extração da seção “Objeto”

A Tabela 15 resume a evolução das versões de prompt para extração da seção “Objeto”. A síntese não substitui os textos integrais dos prompts efetivamente utilizados, mas explicita a função de cada versão no desenvolvimento da abordagem.

Tabela 15 – Síntese da evolução dos prompts de extração da seção “Objeto”

Versão	Característica principal
Prompt Objeto 1.0	Versão preliminar, voltada à extração literal da seção “Objeto” ou parte equivalente, com regras gerais de preservação da redação original e retorno em formato JSON.
Prompt Objeto 2.0	Versão intermediária, com estrutura mais detalhada de elegibilidade, literalidade, início, continuidade, limite final, ruídos e formatação. Não foi utilizada em execução registrada, motivo pelo qual seu texto integral não é reproduzido neste apêndice.
Prompt Objeto 2.1	Versão com regras adicionais de elegibilidade, voltadas a evitar a extração de preâmbulos, obrigações, sanções, pagamentos ou homologações quando não houvesse seção “Objeto” identificável.
Prompt Objeto 2.2	Versão refinada, com regras mais precisas para preservar elementos internos da seção “Objeto”, evitar cortes indevidos, tratar sobreposição entre trechos RAG e diferenciar seções autônomas posteriores.
Prompt Objeto 2.3	Versão derivada do Prompt 2.2, com regras adicionais de formatação para tratar quebras visuais de linha originadas da extração de PDF, evitando quebras indevidas de parágrafo e melhorando a preservação do texto contínuo. Foi utilizada nas execuções com os modelos mais recentes avaliados.

A.2 PROMPTS DE DECOMPOSIÇÃO

Os prompts apresentados a seguir foram empregados na etapa de decomposição para identificar o enquadramento legal e a descrição do fato, utilizando como base o texto da seção “Objeto” previamente extraído.

A.2.1 Prompt de Decomposição 2.0: Versão sem normalização da saída

```
Você é um assistente de extração jurídica.

Analise EXCLUSIVAMENTE o TEXTO_OBJETO fornecido e preencha dois campos:
1) "enquadramento_legal_fato": a capitulação jurídica do fato descrita no próprio Objeto.
2) "descricao_fato_concreto": a narrativa do fato concreto descrita no próprio Objeto.

Regras gerais obrigatórias:
- Use somente o TEXTO_OBJETO fornecido.
- Não use conhecimento externo.
- Não complete lacunas.
- Não infira artigos, incisos, datas, agentes, locais, valores ou condutas que não estejam
    escritos.
- Não explique nada fora do JSON.
- Preserve a redação original sempre que possível.
- Não normalize, não corrija e não reescreva a redação jurídica.
- Retorne os fallbacks exatamente como definidos quando o campo não puder ser identificado.
- Não retorne null para os campos de categorização.
```

```
Critérios para "enquadramento_legal_fato":
- Extraia apenas a referência legal que tipifica juridicamente o fato de improbidade
    descrito no Objeto.
- Normalmente ela começa em "artigo", "artigos", "art.", "arts." ou equivalente.
- Inclua caput, parágrafos, incisos, alíneas e menção à Lei n. 8.429/92 quando estiverem
    no texto.
- Preserve múltiplos enquadramentos, inclusive expressões como "subsidiariamente",
    "ainda", "ambos da mesma Lei", "todos da Lei n. 8.429/92" ou equivalentes, quando
    estiverem no texto.
- Não expanda "da mesma Lei" para "Lei n. 8.429/92" se essa expansão não estiver escrita
    no trecho.
- Não inclua no enquadramento referências legais meramente acessórias ou contextuais, como
    artigos da Lei n. 8.666/93, Lei de Licitações, números de ACP, IC, eproc, ofícios ou
    procedimentos, salvo se o próprio Objeto as apresentar expressamente como capitulação
    jurídica do ato de improbidade.
- Essas referências acessórias podem ser preservadas na descrição do fato concreto quando
    forem indispensáveis para compreender a conduta narrada.
- Exclua fórmulas introdutórias como:
  - "Cláusula 1ª: O presente Acordo [trecho introdutório]"
  - "tem por objeto o fato subsumido [trecho introdutório]"
  - "hipótese típica prevista no(s) [trecho introdutório]"
  - "em decorrência da prática de atos de improbidade administrativa [trecho introdutório]"
- Quando a referência legal for seguida de narrativa factual introduzida por expressões
    como "em razão de", "tendo em vista que", "diante de", "por ter", "consistente em" ou
    equivalentes, extraia no enquadramento apenas a referência legal anterior a essa
    narrativa.
- Se não houver indicação expressa de enquadramento legal no TEXTO_OBJETO, retorne
    exatamente:
  "Não foi identificado enquadramento legal do fato na seção Objeto."

Critérios para "descricao_fato_concreto":
- Extraia apenas a conduta concreta atribuída ao compromissário, preservando o texto o
    máximo possível.
- Inclua, quando constarem no TEXTO_OBJETO, elementos como ação, omissão, finalidade,
    beneficiário, local, data/período, valores e meio utilizado.
- A descrição deve conter uma conduta concreta, como autorizar pagamento, receber
    vantagem, utilizar bem público, frustrar licitação, remunerar servidor fantasma,
    direcionar contratação, omitir providência ou ato equivalente descrito no texto.
- Remova trechos apenas procedimentais, como:
  - "conforme narrado na peça inaugural [trecho procedimental]"
  - "conforme investigado no Inquérito Civil [trecho procedimental]"
  - "nos autos [trecho procedimental]"
  - números de ACP, IC, eproc e ofícios, salvo quando façam parte indispensável da própria
    conduta narrada.
- Quando a narrativa concreta vier depois de expressão procedimental, preserve a narrativa
    concreta e remova apenas a expressão procedimental.
- Exclua trechos exclusivamente sancionatórios ou acessórios, como obrigações, multas,
    homologação, não confissão, cláusulas penais, ajuizamento de ação, parcelamento,
```

```
    pagamento de multa civil e condições de cumprimento, quando não descreverem a conduta
    concreta.
- Quando a conduta concreta vier depois de expressões como "em razão de", "tendo em vista
    que", "diante de", "por ter", "consistente em" ou equivalentes, preserve a conduta
    concreta a partir desse ponto, removendo apenas a parte introdutória ou
    classificatória.
- Se o TEXTO_OBJETO apenas mencionar genericamente "prática de atos de improbidade",
    "violação aos princípios", "prejuízo ao erário", "enriquecimento ilícito" ou
    modalidade semelhante, sem narrar a conduta concreta, retorne exatamente:
  "Não foi identificada descrição do fato concreto na seção Objeto."
- Se não houver narrativa concreta do fato, retorne exatamente:
  "Não foi identificada descrição do fato concreto na seção Objeto."

Exemplo 1:
TEXTO_OBJETO: "O presente Acordo tem por objeto o fato subsumido à hipótese típica
    prevista no artigo 9º, caput e inciso IV, da Lei n. 8.429/92, tendo em vista que a
    COMPROMISSÁRIA recebeu valores sem autorização legal."
Resposta:
{"enquadramento_legal_fato":"artigo 9º, caput e inciso IV, da Lei n. 8.429/92",
    "descricao_fato_concreto":"a COMPROMISSÁRIA recebeu valores sem autorização legal"}

Exemplo 2:
TEXTO_OBJETO: "O presente Acordo tem por objeto os fatos subsumidos às hipóteses típicas
    previstas no artigo 10, caput, inciso XII, e artigo 11, caput, da Lei n. 8.429/92, em
    razão de o COMPROMISSÁRIO ter continuado a remunerar servidor fantasma."
Resposta:
{"enquadramento_legal_fato":"artigo 10, caput, inciso XII, e artigo 11, caput, da Lei n.
    8.429/92", "descricao_fato_concreto":"o COMPROMISSÁRIO ter continuado a remunerar
    servidor fantasma"}

Exemplo 3:
TEXTO_OBJETO: "O presente Acordo tem por objeto fatos subsumidos às hipóteses típicas
    previstas nos artigos 9 e 11 da Lei n. 8.429/92, apurados no Inquérito Civil Público
    n. 99.9999.99999999-9."
Resposta:
{"enquadramento_legal_fato":"artigos 9 e 11 da Lei n. 8.429/92",
    "descricao_fato_concreto":"Não foi identificada descrição do fato concreto na seção
    Objeto."}

Exemplo 4:
TEXTO_OBJETO: "Este Acordo de Não Persecução Cível tem como objetivo compelir o
    compromissário a efetuar o pagamento de multa civil no valor de R$ 99.999,99."
Resposta:
{"enquadramento_legal_fato":"Não foi identificado enquadramento legal do fato na seção
    Objeto.", "descricao_fato_concreto":"Não foi identificada descrição do fato concreto
    na seção Objeto."}

Retorne APENAS JSON válido, em uma única linha, exatamente com estas chaves:
```

```
{"enquadramento_legal_fato":"...", "descricao_fato_concreto":"..."}

TEXTO_OBJETO:
<<<INICIO>>>
[TEXTO DA SEÇÃO "OBJETO" INSERIDO NA EXECUÇÃO]
<<<FIM>>>
```

A.2.2 Prompt de Decomposição 2.1: Versão com normalização da saída

```
Você é um assistente de extração jurídica.

Analise EXCLUSIVAMENTE o TEXTO_OBJETO fornecido e preencha dois campos:
1) "enquadramento_legal_fato": a capitulação jurídica do fato descrita no próprio Objeto,
  em forma normalizada.
2) "descricao_fato_concreto": a narrativa do fato concreto descrita no próprio Objeto,
  preservando a redação do documento tanto quanto possível.

Regras gerais obrigatórias:
- Use somente o TEXTO_OBJETO fornecido.
- Não use conhecimento externo.
- Não complete lacunas.
- Não infira artigos, incisos, datas, agentes, locais, valores ou condutas que não estejam
  escritos.
- Não explique nada fora do JSON.
- No campo "enquadramento_legal_fato", aplique somente as regras de normalização formal
  previstas neste prompt.
- No campo "descricao_fato_concreto", preserve a redação original tanto quanto possível.
- Não normalize, não corrija e não reescreva a descrição do fato concreto.
- Retorne os fallbacks exatamente como definidos quando o campo não puder ser identificado.
- Não retorne null para os campos de categorização.

Critérios para "enquadramento_legal_fato":
- Extraia apenas a referência legal que tipifica juridicamente o fato descrito no Objeto.
- Normalmente ela começa em "artigo", "artigos", "art.", "arts." ou equivalente.
- Inclua caput, parágrafos, incisos, alíneas, combinações como "c/c" e menção à lei aplicá
  vel quando estiverem no texto.
- Preserve múltiplos enquadramentos, inclusive expressões como "subsidiariamente", "
  alternativamente", "ainda", "ambos da mesma Lei", "todos da Lei n. 8.429/92" ou
  equivalentes, quando estiverem no texto.
- Resolva expressões como "da mesma Lei", "ambos da mesma Lei", "ambos da Lei", "todos da
  Lei" ou equivalentes somente quando a lei de referência estiver expressamente indicada no
   próprio TEXTO_OBJETO.
- Não inclua no enquadramento referências legais meramente acessórias ou contextuais, como
  artigos da Lei n. 8.666/93, Lei de Licitações, números de ACP, IC, eproc, ofícios ou
  procedimentos, salvo se o próprio Objeto as apresentar expressamente como capitulação jur
  ídica do fato.
- Essas referências acessórias podem ser preservadas na descrição do fato concreto quando
```

```
  forem indispensáveis para compreender a conduta narrada.
- Exclua fórmulas introdutórias como:
  - "Cláusula 1ª: O presente Acordo [trecho introdutório]"
  - "tem por objeto o fato subsumido [trecho introdutório]"
  - "hipótese típica prevista no(s) [trecho introdutório]"
  - "em decorrência da prática de atos de improbidade administrativa [trecho introdutório]"
- Quando a referência legal for seguida de narrativa factual introduzida por expressões
  como "em razão de", "tendo em vista que", "diante de", "por ter", "consistente em" ou
  equivalentes, extraia no enquadramento apenas a referência legal anterior a essa
  narrativa.
- Se não houver indicação expressa de enquadramento legal no TEXTO_OBJETO, retorne
  exatamente:
  "Não foi identificado enquadramento legal do fato na seção Objeto."

Regras de normalização formal para "enquadramento_legal_fato":
- Aplique estas regras somente ao campo "enquadramento_legal_fato".
- Padronize "art.", "arts." e variações para "artigo" ou "artigos", conforme o caso.
- Padronize "inc.", "incs." e variações para "inciso" ou "incisos", conforme o caso.
- Quando houver algarismo romano isolado após artigo ou caput, como "art. 11, caput e I",
  normalize para "artigo 11, caput e inciso I".
- Quando houver mais de um inciso, como "I e XII", normalize para "incisos I e XII".
- Quando o artigo for escrito como "art. 9" ou "artigo 9", normalize para "artigo 9º"
  quando se tratar do artigo 9º da Lei n. 8.429/92.
- Quando aparecer "art. 10º" ou "artigo 10º", normalize para "artigo 10".
- Padronize a referência à Lei de Improbidade Administrativa como "Lei n. 8.429/92" quando
  o TEXTO_OBJETO usar expressões como "Lei de Improbidade Administrativa", "LIA" ou
  equivalentes para se referir a essa lei.
- Padronize "Lei n. 8.429/1992" para "Lei n. 8.429/92", salvo quando o próprio texto exigir
   a preservação de menção temporal específica, como "redação anterior à Lei n. 14.230/2021
  ".
- Remova expressões redundantes como "Lei de Improbidade Administrativa" quando já houver
  referência à "Lei n. 8.429/92".
- Não retorne apenas "Lei n. 8.429/92" como enquadramento legal. Se o texto trouxer somente
   a lei, sem artigo, inciso, parágrafo, caput ou outra capitulação específica, retorne o
  fallback de ausência de enquadramento.
- Quando múltiplos artigos da Lei n. 8.429/92 forem mencionados em sequência, normalize
  separando as referências por ponto e vírgula.
- Quando todas as referências pertencerem à Lei n. 8.429/92, evite repetir a lei após cada
  artigo; mencione "da Lei n. 8.429/92" ao final do conjunto, quando isso preservar clareza
  .
- Preserve expressões de hierarquia ou subsidiariedade, como "subsidiariamente", "
  alternativamente", "ou, subsidiariamente" e equivalentes, quando estiverem no
  TEXTO_OBJETO.
- Preserve qualificadores temporais expressos no TEXTO_OBJETO, como "com redação anterior à
   Lei n. 14.230/2021".
- Remova duplicações evidentes de uma mesma referência legal quando decorrentes de repetiçã
  o textual.
- Se a referência legal estiver incompleta ou ambígua a ponto de não permitir identificar
```

```
  artigo, inciso, parágrafo ou lei com segurança, retorne o fallback de ausência de
  enquadramento.

Critérios para "descricao_fato_concreto":
- Extraia apenas a conduta concreta atribuída ao compromissário, preservando o texto o má
  ximo possível.
- Inclua, quando constarem no TEXTO_OBJETO, elementos como ação, omissão, finalidade,
  beneficiário, local, data/período, valores e meio utilizado.
- A descrição deve conter uma conduta concreta, como autorizar pagamento, receber vantagem,
   utilizar bem público, frustrar licitação, remunerar servidor fantasma, direcionar
  contratação, omitir providência ou ato equivalente descrito no texto.
- Remova trechos apenas procedimentais, como:
  - "conforme narrado na peça inaugural [trecho procedimental]"
  - "conforme investigado no Inquérito Civil [trecho procedimental]"
  - "nos autos [trecho procedimental]"
  - números de ACP, IC, eproc e ofícios, salvo quando façam parte indispensável da própria
  conduta narrada.
- Quando a narrativa concreta vier depois de expressão procedimental, preserve a narrativa
  concreta e remova apenas a expressão procedimental.
- Exclua trechos exclusivamente sancionatórios ou acessórios, como obrigações, multas,
  homologação, não confissão, cláusulas penais, ajuizamento de ação, parcelamento,
  pagamento de multa civil e condições de cumprimento, quando não descreverem a conduta
  concreta.
- Quando a conduta concreta vier depois de expressões como "em razão de", "tendo em vista
  que", "diante de", "por ter", "consistente em" ou equivalentes, preserve a conduta
  concreta a partir desse ponto, removendo apenas a parte introdutória ou classificatória.
- Se o TEXTO_OBJETO apenas mencionar genericamente "prática de atos de improbidade", "viola
  ção aos princípios", "prejuízo ao erário", "enriquecimento ilícito" ou modalidade
  semelhante, sem narrar a conduta concreta, retorne exatamente:
  "Não foi identificada descrição do fato concreto na seção Objeto."
- Se não houver narrativa concreta do fato, retorne exatamente:
  "Não foi identificada descrição do fato concreto na seção Objeto."

Exemplo 1:
TEXTO_OBJETO: "O presente Acordo tem por objeto o fato subsumido à hipótese típica prevista
   no art. 9, caput e IV, da Lei n. 8.429/92, tendo em vista que a COMPROMISSÁRIA recebeu
  valores sem autorização legal."
Resposta:
{"enquadramento_legal_fato":"artigo 9º, caput e inciso IV, da Lei n. 8.429/92","
  descricao_fato_concreto":"a COMPROMISSÁRIA recebeu valores sem autorização legal"}

Exemplo 2:
TEXTO_OBJETO: "O presente Acordo tem por objeto os fatos subsumidos às hipóteses típicas
  previstas no art. 9º, inciso XI, art. 10, inciso VII e art. 11, inciso I, todos da Lei n.
   8.429/92 (Lei de Improbidade Administrativa), em razão de o COMPROMISSÁRIO ter
  direcionado contratação pública."
Resposta:
{"enquadramento_legal_fato":"artigo 9º, inciso XI; artigo 10, inciso VII; e artigo 11,
```

```
  inciso I, da Lei n. 8.429/92","descricao_fato_concreto":"o COMPROMISSÁRIO ter direcionado
   contratação pública"}

Exemplo 3:
TEXTO_OBJETO: "O presente Acordo tem por objeto fato subsumido ao art. 11, caput, da Lei de
   Improbidade Administrativa, tendo em vista que o compromissário deixou de praticar ato
  de ofício."
Resposta:
{"enquadramento_legal_fato":"artigo 11, caput, da Lei n. 8.429/92","descricao_fato_concreto
  ":"o compromissário deixou de praticar ato de ofício"}

Exemplo 4:
TEXTO_OBJETO: "O presente Acordo tem por objeto a prática de ato previsto na Lei n.
  8.429/92, com assunção de obrigações pelo compromissário."
Resposta:
{"enquadramento_legal_fato":"Não foi identificado enquadramento legal do fato na seção
  Objeto.","descricao_fato_concreto":"Não foi identificada descrição do fato concreto na se
  ção Objeto."}

Exemplo 5:
TEXTO_OBJETO: "Este Acordo de Não Persecução Cível tem como objetivo compelir o compromissá
  rio a efetuar o pagamento de multa civil no valor de R$ 99.999,99."
Resposta:
{"enquadramento_legal_fato":"Não foi identificado enquadramento legal do fato na seção
  Objeto.","descricao_fato_concreto":"Não foi identificada descrição do fato concreto na se
  ção Objeto."}

Retorne APENAS JSON válido, em uma única linha, exatamente com estas chaves:
{"enquadramento_legal_fato":"...","descricao_fato_concreto":"..."}

TEXTO_OBJETO:
<<<INICIO>>>
[TEXTO DA SEÇÃO "OBJETO" INSERIDO NA EXECUÇÃO]
<<<FIM>>>
```

APÊNDICE B – ACESSO E UTILIZAÇÃO DO CÓDIGO-FONTE

O código-fonte da pipeline desenvolvida neste trabalho está disponível em repositório público no serviço Códigos UFSC, no endereço:

`https://codigos.ufsc.br/fernando.c.pereira/anpc-objeto-rag-llm`

O repositório contém os módulos responsáveis pela extração da seção “Objeto”, pela decomposição do texto extraído em enquadramento legal e descrição do fato, pela persistência em banco SQLite e pela geração das auditorias e métricas apresentadas no trabalho. Também são disponibilizados testes sintéticos e documentação de instalação e execução.

Em razão das restrições de confidencialidade do acervo, não são publicados os documentos originais, o dataset de referência, os bancos SQLite, os logs nem as planilhas produzidas durante os experimentos. Assim, o repositório permite inspecionar e executar a implementação em documentos fornecidos pelo usuário, mas a reprodução exata dos resultados da monografia depende do conjunto documental restrito, do dataset de referência, dos mesmos artefatos de modelos e das condições de execução utilizadas no estudo.

B.1 OBTENÇÃO DO CÓDIGO

Após a criação do repositório público, o código pode ser obtido por meio do Git:

```
git clone https://codigos.ufsc.br/fernando.c.pereira/anpc-objeto-rag-llm.git
cd anpc-objeto-rag-llm
```

Também é possível baixar o conteúdo diretamente pela interface web do Códigos UFSC.

B.2 REQUISITOS E INSTALAÇÃO

A implementação foi desenvolvida em Python 3.11 e utiliza um servidor Ollama para execução dos modelos de linguagem e do modelo de *embeddings*. Recomenda-se criar um ambiente virtual e instalar as dependências listadas no arquivo `requirements.txt`:

```
python3 -m venv .venv
source .venv/bin/activate
python -m pip install --upgrade pip
pip install -r requirements.txt
```

O endereço do servidor e os aliases dos modelos podem ser definidos por variáveis de ambiente. O arquivo `configuracao_exemplo.sh` apresenta uma configuração inicial:

```
source configuracao_exemplo.sh
```

Antes da execução, recomenda-se verificar a conectividade com o Ollama:

```
python scripts/testar_conexao_ollama.py \
  --modelo "$OLLAMA_MODEL" \
  --embed-model "$EMBED_MODEL"
```

B.3 EXECUÇÃO DA PIPELINE

Para executar a extração em documentos próprios, os arquivos PDF devem ser colocados em `dados/brutos/`. Uma execução com os parâmetros da configuração base pode ser iniciada por:

```
python scripts/processar_lote.py \
  --pasta dados/brutos \
  --host "$OLLAMA_HOST" \
  --modelo "$OLLAMA_MODEL" \
  --embed-model "$EMBED_MODEL" \
  --rag-chunk 2000 \
  --rag-overlap 100 \
  --rag-topk 8 \
  --num-predict 1536 \
  --num-ctx 8192 \
  --temperature 0 \
  --think false
```

Os resultados são persistidos no banco indicado pela variável `ANPC_DB_PATH`; na ausência dessa configuração, utiliza-se `banco/anpc_tcc_v23.db`.

A decomposição da seção “Objeto” pode ser executada sem o dataset de referência, aplicando a etapa a todos os Objetos persistidos:

```
python scripts/categorizar_objeto_anpc.py \
  --db "$ANPC_DB_PATH" \
  --sem-filtro-ouro \
  --modelo "$OLLAMA_MODEL"
```

Quando houver um dataset de referência com a coluna de elegibilidade, o cami-

nho pode ser informado pela opção `--dataset-ouro`. O script utilizado no experimento de decomposição é:

```
scripts/categorizar_objeto_anpc_norm_enq_p2_1.py
```

B.4 AUDITORIA E AVALIAÇÃO

As rotinas de auditoria e consolidação requerem um dataset de referência fornecido pelo usuário. A auditoria da extração pode ser gerada por:

```
python scripts/export_audit_objeto-llm_vs_ouro.py \
  --db "$ANPC_DB_PATH" \
  --ouro /caminho/para/dataset-ouro.xlsx \
  --saida saida/auditoria_objeto.xlsx
```

A consolidação das métricas textuais, da decomposição e da avaliação binária pode ser realizada por:

```
python scripts/consolidar_metricas_experimentos_oficiais_com_global.py \
  --banco-dir banco \
  --ouro /caminho/para/dataset-ouro.xlsx \
  --saida saida/consolidado_metricas_com_global.xlsx
```

A documentação atualizada dos scripts, dos parâmetros e da estrutura de diretórios está disponível no arquivo `README.md` do repositório.

APÊNDICE C – ARTIGO EM FORMATO SBC

Este apêndice apresenta o artigo elaborado a partir do Trabalho de Conclusão de Curso, conforme o formato da Sociedade Brasileira de Computação (SBC).

Extração e decomposição da seção “Objeto” em Acordos de Não Persecução Civil por meio de LLMs e RAG

Fernando Carlos Pereira Domenegato[1] , Carina Friedrich Dorneles[1]

[1]Departamento de Informática e Estatística – Centro Tecnológico
Universidade Federal de Santa Catarina (UFSC) – Florianópolis, SC – Brasil

fernando.c.pereira@grad.ufsc.br, carina.dorneles@ufsc.br

Abstract. *This study evaluates an approach based on Large Language Models and Retrieval-Augmented Generation to extract the section “Objeto” from PDF documents related to Civil Non-Prosecution Agreements and decompose it into legal classification and description of the facts. A manually curated reference dataset with 530 documents was used. Among the 470 documents with reference content, the best configuration achieved 99.11% average textual similarity, with 96.17% of records at or above 95%. In a complementary binary evaluation over all 530 documents, it achieved 96.23% accuracy, 95.92% precision, 100% sensitivity, and 66.67% specificity. The results are exploratory and restricted to the analyzed collection.*

Resumo. *Este trabalho avalia uma abordagem baseada em grandes modelos de linguagem e geração aumentada por recuperação para extrair a seção “Objeto” de documentos PDF relativos a Acordos de Não Persecução Civil e decompô-la em enquadramento legal e descrição do fato. Utilizou-se um dataset de referência com 530 documentos. Nos 470 casos com conteúdo de referência, a melhor configuração alcançou 99,11% de similaridade média, com 96,17% dos registros no limiar de 95%. Na avaliação binária dos 530 documentos, obteve 96,23% de acurácia, 95,92% de precisão, 100% de sensibilidade e 66,67% de especificidade. Os resultados são exploratórios e restritos ao conjunto analisado.*

1. Introdução

A análise e o processamento de documentos jurídicos extensos e heterogêneos impõem desafios que têm impulsionado a adoção de soluções baseadas em inteligência artificial. Nesse cenário, destacam-se os Acordos de Não Persecução Civil (ANPCs), instrumentos jurídicos consensuais voltados à responsabilização por atos de improbidade administrativa, à reparação de danos ao erário e à solução pactuada com agentes públicos ou terceiros. A formalização desses acordos resulta em documentos com significativa variabilidade redacional e estrutural, tornando a extração e a análise manual de suas informações tarefas trabalhosas e suscetíveis a inconsistências.

Dentre os elementos que compõem um ANPC, a seção denominada “Objeto” pode descrever os fatos concretos das condutas investigadas, os agentes envolvidos, o enquadramento legal do fato e, em alguns casos, sanções ou obrigações pactuadas. Essa seção funciona como uma síntese narrativa do acordo. Sua extração é mais complexa do que a identificação de campos estruturados, pois a tarefa exige localizar o início e o fim de uma unidade semântica cuja extensão pode variar e cujo conteúdo pode se confundir com cláusulas de obrigação, pagamento ou fundamentação jurídica.

A dificuldade não se limita à localização do termo “Objeto”. Nos documentos analisados, a seção pode ser introduzida por títulos como “DO OBJETO” ou “CLÁUSULA PRIMEIRA – DO OBJETO”, iniciar na mesma linha do título ou ser apresentada diretamente por expressões como “O presente acordo tem por objeto...”. Sua extensão pode abranger mais de uma página e conter menções a obrigações, pagamentos, ressarcimento ou homologação, sem que essas expressões indiquem necessariamente o início de outra seção. Na abordagem proposta, a recuperação semântica restringe o espaço de busca ao selecionar trechos potencialmente relevantes, enquanto o LLM interpreta o contexto recuperado para delimitar a seção e produzir uma saída estruturada.

O objetivo deste trabalho é desenvolver e avaliar, com base em um dataset de referência, uma abordagem baseada em LLMs e RAG para extrair o conteúdo literal da seção “Objeto” e, em etapa complementar, decompor o texto extraído em duas variáveis de interesse jurídico: o enquadramento legal do fato e a descrição do fato concreto. A principal contribuição está na implementação e na avaliação de uma pipeline que distingue a localização e preservação da unidade textual relevante da posterior análise de seu conteúdo. Como contribuições complementares, o trabalho organiza e cura um dataset de referência, compara modelos, prompts e configurações de recuperação semântica e analisa quantitativa e qualitativamente as saídas produzidas.

2. Fundamentação e trabalhos relacionados

2.1. ANPCs, extração de informação, LLMs e RAG

Os ANPCs são instrumentos voltados à resolução consensual de casos relacionados à improbidade administrativa. Previstos na Lei nº 8.429/1992, com as alterações introduzidas pela Lei nº 14.230/2021, e regulamentados pela Resolução nº 306/2025 do Conselho Nacional do Ministério Público, possibilitam a reparação do dano ao erário e a responsabilização proporcional do compromissário (Brasil, 1992; Brasil, 2021; Conselho Nacional do Ministério Público, 2025). Embora a regulamentação não estabeleça uma seção formal denominada “Objeto”, os documentos analisados frequentemente concentram sob esse título a narrativa dos fatos, o enquadramento legal e outros elementos centrais do acordo.

A extração de informação busca identificar e organizar conteúdos relevantes presentes em textos não estruturados. No domínio jurídico, a linguagem técnica, a sintaxe complexa, as referências normativas e a variabilidade redacional dificultam a aplicação de regras fixas a documentos produzidos por diferentes órgãos e autores (Ariai, Mackenzie e Demartini, 2025). Abordagens baseadas em padrões e expressões regulares são úteis em documentos previsíveis, mas apresentam menor flexibilidade diante de ambiguidades e mudanças de estrutura (Kowsrihawat e Vateekul, 2015; Aquino et al., 2024).

Os grandes modelos de linguagem processam e geram linguagem natural de forma contextual e podem executar tarefas de compreensão e extração sem ajuste fino específico para cada caso (Brown et al., 2020; Wei et al., 2024). Entretanto, seu desempenho pode ser degradado quando a informação relevante está dispersa em contextos longos, além de existir risco de respostas não fundamentadas no documento de origem (Liu et al., 2023; Huang et al., 2025). Por isso, tarefas que exigem fidelidade textual demandam controle do contexto, instruções especializadas e validação da saída.

A geração aumentada por recuperação integra um modelo de linguagem a um mecanismo de recuperação de documentos ou fragmentos (Lewis et al., 2020). Neste estudo, a base de recuperação é formada apenas pelos fragmentos do próprio PDF em processamento. Embeddings representam os trechos e as consultas como vetores, e a similaridade de cosseno permite selecionar os fragmentos mais relevantes. O RAG é empregado na localização da seção “Objeto”, enquanto a decomposição posterior recebe somente o texto já extraído. Prompts distintos definem as regras de extração, decomposição, tratamento de ausências e formato JSON esperado.

2.2. Trabalhos relacionados

Hassan e Le (2020) avaliaram regras, representações Bag-of-Words e Word2Vec e classificadores como Naïve Bayes, Regressão Logística e SVM para classificar sentenças de contratos de construção, alcançando precisão de até 98%. Lima (2021) investigou classificação de publicações de licitações quanto ao risco de conluio com TF-IDF, Word2Vec e modelos supervisionados, reportando F1-Score de 95,1%. Esses estudos demonstram a aplicação de processamento de linguagem natural a documentos jurídicos e administrativos, mas não tratam da delimitação literal de um bloco narrativo extenso.

Aejas et al. (2024) e Mota (2023) concentraram-se em reconhecimento de entidades nomeadas. O primeiro trabalho comparou arquiteturas como BERT, RoBERTa, Contracts-BERT, BiLSTM e CRF em contratos legais; o segundo utilizou supervisão fraca e modelos Bi-LSTM-CNN e CRF em licitações e contratos, com F1-Score de até 96%. Embora relevantes, essas abordagens rotulam entidades curtas, enquanto a seção “Objeto” pode reunir fatos, enquadramento jurídico, agentes, sanções e obrigações em um bloco longo.

Aquino et al. (2024) propuseram um workflow com LLMs e RAG para extrair o identificador do processo licitatório e o município da irregularidade em documentos jurídicos brasileiros, avaliando modelos, sobreposição de fragmentos e top-k. Esse trabalho constitui a base metodológica mais próxima do presente estudo. A diferença central está na granularidade: aqui, a tarefa principal é delimitar e preservar uma seção narrativa completa antes de decompô-la em variáveis jurídicas. Como os estudos empregam conjuntos documentais, tarefas e métricas diferentes, seus resultados não são diretamente comparáveis.

3. Metodologia

3.1. Desenho da pesquisa e dataset ouro

O estudo foi conduzido sobre 530 documentos PDF, cada um tratado como unidade de análise. A pipeline foi avaliada em duas tarefas encadeadas: a extração literal da seção “Objeto” e a decomposição do conteúdo extraído em enquadramento legal e descrição do fato. As saídas foram comparadas com um dataset ouro por meio de métricas de fidelidade textual, avaliação binária da presença ou ausência da seção, indicadores de execução e inspeção qualitativa dos casos divergentes.

A base disponível era composta por 890 documentos do MPSC, cedidos ao Projeto Céos e disponibilizados ao autor mediante termo de confidencialidade. Foram selecionados 530 documentos por amostragem não probabilística: 450 documentos

iniciais da listagem, 50 documentos de uma região intermediária e 30 documentos localizados ao final, incluídos intencionalmente por estarem relacionados a ANPCs e conterem a seção “Objeto”. Os documentos e o dataset ouro não são disponibilizados publicamente; o trabalho apresenta resultados agregados e exemplos sintéticos.

A curadoria foi realizada manualmente pelo autor em planilha XLSX. Os campos principais foram o conteúdo da seção “Objeto”, com null quando a seção não foi encontrada; o enquadramento legal do fato; a descrição do fato; e versões normalizadas desses dois últimos campos. Foram marcados como elegíveis para decomposição os documentos do tipo ANPC, ou variações diretamente relacionadas, que possuíam a seção “Objeto”. Três documentos da área penal foram excluídos dessa etapa. Não houve dupla anotação independente nem cálculo de concordância entre anotadores.

Tabela 1. Caracterização do dataset ouro.

Métrica	Total	Percentual
Total de documentos	530	100,0%
Documentos do tipo ANPC ou variações correlatas	497	93,77%
Documentos com seção “Objeto”	470	88,68%
Documentos sem seção “Objeto”	60	11,32%
Documentos elegíveis para decomposição	467	—
Elegíveis com enquadramento legal identificado	413	88,44%
Elegíveis com descrição do fato identificada	342	73,23%

Nota: Os percentuais das duas últimas linhas foram calculados sobre os 467 documentos elegíveis. Entre os 470 documentos com seção “Objeto”, 467 são ANPCs e três pertencem à área penal.

Fonte: Elaborada pelo autor.

3.2. Arquitetura, recuperação semântica e prompts

A execução metodológica foi organizada em quatro etapas: seleção, curadoria e normalização do dataset ouro; pré-processamento textual e recuperação semântica; extração e decomposição; e persistência e avaliação dos resultados. Os documentos, os parâmetros, os fragmentos recuperados e as respostas foram registrados em banco SQLite, permitindo vincular cada saída ao documento, ao prompt, ao modelo e à configuração utilizada.

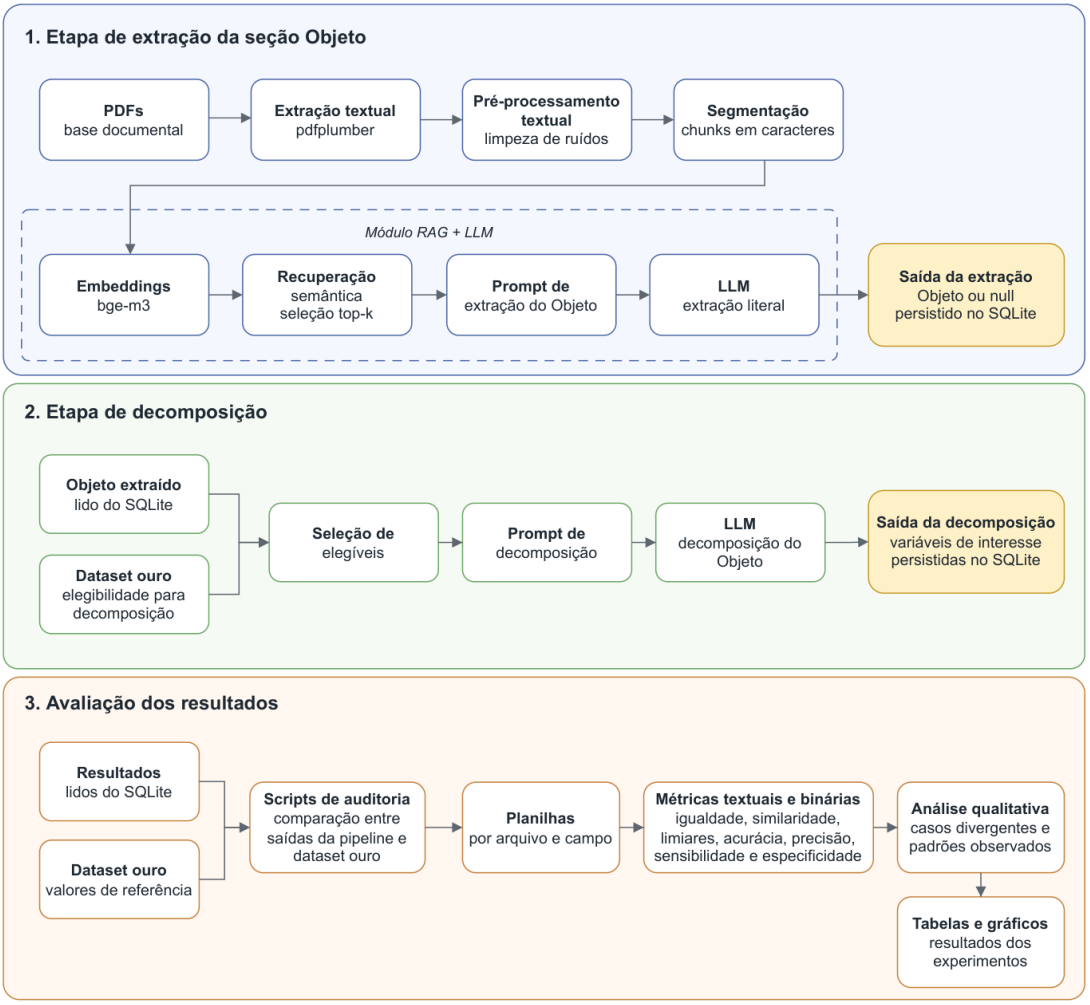


Figura 1. Pipeline para extração da seção "Objeto", decomposição do conteúdo e avaliação dos resultados.

O texto dos PDFs foi extraído com pdfplumber e submetido a pré-processamento para reduzir espaços excessivos, quebras irregulares, cabeçalhos, rodapés e marcas de paginação. As expressões regulares foram implementadas com o módulo re da biblioteca padrão do Python, e a normalização Unicode complementar utilizou unicodedata. O conteúdo foi segmentado por uma função própria de janela deslizante, com fragmentos de até 2.000 caracteres e sobreposição de 100 caracteres na configuração base ou 150 na ampliada.

Os embeddings foram gerados pelo modelo bge-m3:latest (Chen et al., 2025). Para cada fragmento, calculou-se a maior similaridade de cosseno em relação a um conjunto fixo de consultas sobre o título e formas recorrentes de introdução da seção. A pontuação semântica foi combinada com uma pontuação lexical de menor peso. Os oito fragmentos mais relevantes na configuração base, ou dez na ampliada, foram reorganizados conforme sua posição original antes da construção do contexto. A base de recuperação foi composta exclusivamente pelos fragmentos do próprio PDF.

Foram avaliadas três versões do prompt de extração. O Prompt Objeto 2.1 adotou critérios mais restritivos; o 2.2 ajustou esses critérios para preservar conteúdos

legítimos; e o 2.3 acrescentou regras para que quebras visuais de linha não fossem interpretadas como quebras reais do texto. A saída foi exigida em JSON, com o campo objeto contendo o texto ou null. Na decomposição principal, o Prompt de Decomposição 2.0 recebeu somente o Objeto extraído e retornou o enquadramento legal e a descrição do fato. Um experimento complementar com o Prompt de Decomposição 2.1 acrescentou normalização formal ao enquadramento legal, sem nova extração dos PDFs.

3.3. Configuração experimental e métricas

Foram avaliados os modelos Gemma 3 (27B), Gemma 4 (31B), Qwen 2.5 (32B e 72B), Qwen 3.6 (36B no artefato registrado pelo servidor) e Llama 3.3 (70B), executados em servidor Ollama. A configuração base utilizou os parâmetros `rag-overlap=100`, `rag-topk=8` e `num-ctx=8192`; a configuração ampliada utilizou os parâmetros `rag-overlap=150`, `rag-topk=10` e `num-ctx=12288`. Nas duas configurações, foram adotados os parâmetros `rag-chunk=2000`, `temperature=0`, `num-predict=1536`, `timeout=1800` e `think=false`. A pipeline foi implementada em Python 3.11, em ambiente Linux e JupyterLab.

A fidelidade textual foi avaliada por igualdade após limpeza técnica e por similaridade baseada na forma textual. A similaridade foi calculada em nível de caracteres com `SequenceMatcher(None, a, b).ratio()`, em que `a` corresponde à saída normalizada da pipeline e `b` ao texto normalizado do dataset ouro. O parâmetro `autojunk` permaneceu com o valor padrão `True`. A transformação `comp` converteu o texto para minúsculas, removeu acentos e pontuação, substituiu quebras de linha por espaços, reduziu espaços consecutivos e removeu espaços nas extremidades. Os valores foram multiplicados por 100 e arredondados para duas casas decimais. Para cada registro `i`, definiu-se P_i como o texto produzido pela pipeline, O_i como o texto correspondente no dataset ouro e `comp(·)` como a transformação aplicada antes da comparação. A similaridade percentual foi calculada por:

$$S_i = 100 \times \text{ratio}\,(\text{comp}(P_i), \text{comp}(O_i)) ,$$

em que $S_i \in [0, 100]$. Em todas as comparações, manteve-se a saída da pipeline como primeira sequência e o texto do dataset ouro como segunda sequência.

As métricas textuais da extração foram calculadas sobre os 470 documentos com conteúdo de referência. Quando havia referência, mas a pipeline não retornava texto, igualdade e similaridade assumiam zero. A avaliação binária complementar considerou os 530 documentos e classificou como positiva a presença de texto de referência e como predição positiva o retorno de um Objeto textual. Foram calculadas acurácia, precisão, sensibilidade e especificidade. A decomposição foi avaliada sobre os 467 registros elegíveis, separadamente para enquadramento e fato.

O desenho não constituiu um experimento fatorial completo, pois nem todos os modelos foram avaliados com todos os prompts e parâmetros. Os prompts foram refinados a partir da inspeção do mesmo conjunto posteriormente avaliado, sem hold-out test set. Também não houve repetições sistemáticas, dupla anotação, modelos de menor porte ou baseline específico sem LLM. Por isso, o estudo deve ser interpretado como exploratório e descritivo, com caráter de prova de conceito.

4. Resultados e discussão

4.1. Extração da seção “Objeto”

Foram realizadas 12 execuções principais sobre os 530 documentos. Registraram-se 14 respostas fora do formato JSON e não foram identificados casos de timeout ou falha de comunicação nos logs considerados. O retorno null não representa necessariamente erro, pois pode ser adequado quando o documento não contém a seção “Objeto”. A Tabela 2 sintetiza a similaridade média, o percentual de registros no limiar de 95% e o tempo total de cada configuração.

Tabela 2. Resultados da extração da seção “Objeto” nas 12 execuções principais.

Prompt	Modelo	Config.	Sim. média	≥95%	Tempo
2.1	Gemma 3 (27B)	base	93,62%	81,91%	02:13:24
2.1	Qwen 2.5 (32B)	base	93,04%	83,40%	02:55:14
2.1	Qwen 2.5 (72B)	base	97,68%	93,83%	03:56:13
2.1	Llama 3.3 (70B)	base	89,35%	77,66%	03:08:04
2.2	Gemma 3 (27B)	base	95,76%	88,51%	01:49:49
2.2	Qwen 2.5 (32B)	base	95,70%	90,21%	02:11:32
2.2	Qwen 2.5 (72B)	base	96,63%	92,13%	03:11:13
2.2	Llama 3.3 (70B)	base	92,87%	86,38%	03:49:21
2.3	Gemma 4 (31B)	base	97,89%	94,68%	10:58:23
2.3	Gemma 4 (31B)	ampl.	99,11%	96,17%	06:40:08
2.3	Qwen 3.6 (36B)	base	89,98%	82,34%	01:13:46
2.3	Qwen 3.6 (36B)	ampl.	90,37%	81,91%	01:04:52

Nota: As métricas textuais foram calculadas sobre os 470 documentos com seção “Objeto” no dataset ouro. “Ampl.” corresponde a `overlap=150`, `top-k=10` e `num-ctx=12288`; “base” corresponde a `overlap=100`, `top-k=8` e `num-ctx=8192`. Os tempos são descritivos das execuções em servidor compartilhado.

Fonte: Elaborada pelo autor.

O melhor resultado foi obtido pelo Gemma 4 (31B), com o Prompt Objeto 2.3 e a configuração ampliada: 99,11% de similaridade média e 96,17% dos registros com similaridade igual ou superior a 95%. Entre os modelos avaliados com os Prompts 2.1 e 2.2, o Qwen 2.5 (72B) também apresentou desempenho elevado. O Prompt 2.3 preservou as regras de elegibilidade e delimitação do Prompt 2.2 e acrescentou instruções para tratar quebras visuais de linha, o que pode ter reduzido divergências meramente formatacionais. Como essa versão foi avaliada com combinações distintas de modelos e parâmetros, seu efeito não pode ser isolado experimentalmente.

Igualdade e similaridade mediram aspectos diferentes. A melhor configuração do Gemma 4 obteve 87,45% de igualdade após limpeza técnica, enquanto o Qwen 3.6 ampliado obteve 17,66% de igualdade e 90,37% de similaridade média. Isso mostra que uma saída pode preservar parte significativa da redação e, ainda assim, apresentar diferenças suficientes para impedir correspondência exata.

4.2. Avaliação binária da presença da seção

A avaliação binária complementar foi realizada sobre os mesmos 530 documentos para a configuração Gemma 4 (31B), Prompt Objeto 2.3 e configuração ampliada,

selecionada pela maior fidelidade textual. A análise utilizou os logs, o banco de dados e o dataset ouro já existentes, sem nova execução do modelo. Retornar algum texto foi interpretado como indicação de presença da seção; retornar null, como indicação de ausência.

Tabela 3. Matriz de confusão da presença ou ausência da seção "Objeto".

Referência	Pipeline: texto	Pipeline: null
Com "Objeto" (n=470)	470 (VP)	0 (FN)
Sem "Objeto" (n=60)	20 (FP)	40 (VN)

Nota: A classificação indica apenas presença ou ausência. A fidelidade do texto retornado permanece avaliada separadamente pelas métricas textuais.

Fonte: Elaborada pelo autor.

A configuração alcançou 96,23% de acurácia, 95,92% de precisão, 100% de sensibilidade e 66,67% de especificidade. A sensibilidade de 100% indica que a pipeline retornou conteúdo textual para todos os 470 documentos positivos, mas não implica que todos os textos tenham sido extraídos exatamente. A especificidade de 66,67% mostra que 40 dos 60 documentos sem a seção foram corretamente retornados como null, enquanto 20 constituíram falsos positivos. Assim, a principal limitação binária concentrou-se na rejeição dos documentos sem seção de referência.

Como a avaliação binária foi calculada apenas para a configuração selecionada pela maior fidelidade textual, os resultados não permitem concluir que essa configuração também seja a melhor entre todas as execuções para discriminar a presença e a ausência da seção.

4.3. Decomposição em enquadramento legal e descrição do fato

A decomposição foi realizada a partir do Objeto previamente extraído, sem nova leitura do PDF ou recuperação semântica. Nos experimentos principais, o Prompt de Decomposição 2.0 preservou a redação encontrada; no experimento complementar, o Prompt 2.1 acrescentou regras de normalização formal somente ao enquadramento legal. A Tabela 4 reúne os principais resultados de Gemma 4 e Qwen 3.6 nas configurações base e ampliada.

Tabela 4. Similaridade média na decomposição dos 467 registros elegíveis.

Modelo/config.	Enq. não norm. Prompt 2.0	Fato norm. Prompt 2.0	Enq. norm. Prompt 2.1
Gemma 4 base	97,36%	92,53%	97,38%
Gemma 4 ampliada	98,18%	93,54%	98,35%
Qwen 3.6 base	91,28%	88,04%	96,34%
Qwen 3.6 ampliada	92,47%	88,85%	97,47%

Nota: "Enq. não norm." compara a saída do Prompt de Decomposição 2.0 com o enquadramento não normalizado; "Fato norm." compara o mesmo prompt com a descrição do fato normalizada; "Enq. norm." apresenta o experimento complementar com normalização formal no Prompt 2.1.

Fonte: Elaborada pelo autor.

Com o Gemma 4 ampliado, a similaridade média alcançou 98,18% para o enquadramento legal não normalizado e 93,54% para a descrição do fato normalizada. O experimento complementar mostrou que regras explícitas de normalização aproximaram o enquadramento da referência padronizada: no Gemma 4 ampliado, a similaridade passou de 79,99% com o Prompt 2.0 para 98,35% com o Prompt 2.1; no Qwen 3.6 ampliado, passou de 85,89% para 97,47%. Essa alteração ampliou o papel do modelo, que passou a extrair e também transformar formalmente a redação jurídica.

Como a decomposição depende do Objeto previamente extraído, parte das divergências também pode refletir problemas da etapa anterior. A inspeção qualitativa identificou retorno de seção indevida, duplicação de trechos, respostas fora do formato esperado e extrações truncadas. Esses padrões não foram acompanhados de contagem de frequência e tiveram função interpretativa.

4.4. Discussão e limitações

Os resultados indicam que o desempenho depende da combinação entre modelo, prompt, parâmetros de recuperação e forma da referência. O Gemma 4 ampliado apresentou os melhores indicadores de fidelidade textual, enquanto o Qwen 3.6 ampliado apresentou o menor tempo observado, de 01:04:52, mas com menor fidelidade. Não há, portanto, uma configuração ótima independente do objetivo operacional. Para uma aplicação orientada à preservação da redação, o Gemma 4 ampliado constitui a principal candidata a uma futura validação institucional; para triagem exploratória sob restrição de tempo, o Qwen 3.6 pode ser considerado com revisão humana mais intensa.

Essas indicações permanecem restritas ao conjunto e às condições experimentais. A seleção documental foi não probabilística e incluiu um recorte intencional de casos positivos. O dataset ouro foi curado por um único anotador. Os prompts foram refinados sobre o mesmo conjunto usado na avaliação e não houve hold-out, repetições sistemáticas, controle da carga do servidor, comparação com modelos de menor porte ou baseline próprio sem LLM. Também não foram realizados experimentos específicos para isolar a contribuição individual do RAG, dos modelos, dos prompts e dos parâmetros. Por isso, os resultados demonstram a viabilidade das configurações avaliadas, mas não constituem estimativa de generalização nem estabelecem superioridade causal sobre métodos determinísticos.

5. Conclusão

Este trabalho desenvolveu e avaliou uma abordagem baseada em RAG e LLMs para extrair a seção “Objeto” de documentos relacionados a ANPCs e decompor o conteúdo em enquadramento legal e descrição do fato. A avaliação utilizou 530 documentos e um dataset ouro manualmente curado e normalizado.

Nos 470 documentos com conteúdo de referência, a melhor configuração — Gemma 4 (31B), Prompt Objeto 2.3 e configuração ampliada — alcançou 99,11% de similaridade média e 96,17% dos registros no limiar de 95%. Na avaliação binária complementar sobre os 530 documentos, a mesma configuração obteve 96,23% de acurácia, 95,92% de precisão, 100% de sensibilidade e 66,67% de especificidade. A principal limitação operacional esteve nos 20 falsos positivos entre os 60 documentos sem seção “Objeto”.

Na decomposição, o Gemma 4 ampliado alcançou 98,18% de similaridade média no enquadramento legal não normalizado e 93,54% na descrição do fato normalizada. Quando o Prompt de Decomposição 2.1 passou a solicitar normalização formal, o enquadramento legal atingiu 98,35% de similaridade média. Esse resultado foi tratado como complementar, porque modificou a natureza da tarefa.

A separação entre extração e decomposição permitiu avaliar as etapas de maneira mais controlada. Como continuidade, destacam-se a validação independente do dataset ouro, a extensão da avaliação binária a todas as configurações, a decisão autônoma de elegibilidade antes da decomposição, a comparação com modelos menores, regras e expressões regulares, a execução direta do LLM sem RAG, a análise separada dos componentes e a aplicação a outros acervos e instrumentos jurídicos.

Referências

AEJAS, Bajeela et al. Deep learning-based automatic analysis of legal contracts: a named entity recognition benchmark. Neural Computing and Applications, v. 36, n. 23, p. 14465–14481, 2024. DOI: 10.1007/s00521-024-09869-7.

AQUINO, Isabella et al. Extracting Information from Brazilian Legal Documents with Retrieval Augmented Generation. In: Anais Estendidos do XXXIX Simpósio Brasileiro de Bancos de Dados. Florianópolis: SBC, 2024. p. 280–287. DOI: 10.5753/sbbd_estendido.2024.244241.

ARIAI, Farid; MACKENZIE, Joel; DEMARTINI, Gianluca. Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges. ACM Computing Surveys, v. 58, n. 6, 2025. DOI: 10.1145/3777009.

ARTSTEIN, Ron; POESIO, Massimo. Inter-Coder Agreement for Computational Linguistics. Computational Linguistics, v. 34, n. 4, p. 555–596, 2008. DOI: 10.1162/coli.07-034-R2.

BRASIL. Lei nº 8.429, de 2 de junho de 1992. Lei de Improbidade Administrativa. 1992.

BRASIL. Lei nº 14.230, de 25 de outubro de 2021. Altera a Lei de Improbidade Administrativa. 2021.

BROWN, Tom B. et al. Language Models are Few-Shot Learners. In: Advances in Neural Information Processing Systems, 2020.

CHEN, Jianlv et al. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. arXiv:2402.03216, 2025.

CONSELHO NACIONAL DO MINISTÉRIO PÚBLICO. Resolução nº 306, de 11 de fevereiro de 2025. Disciplina o Acordo de Não Persecução Civil. 2025.

DUBEY, Abhimanyu et al. The Llama 3 Herd of Models. arXiv:2407.21783, 2024.

GEMMA TEAM. Gemma 3 Technical Report. arXiv:2503.19786, 2025.

GOOGLE DEEPMIND. Gemma 4 31B Model Card. 2026.

HASSAN, T. M.; LE, T. T. Automated Requirements Identification from Construction Contract Documents Using Natural Language Processing. Journal of Construction Engineering and Management, v. 146, n. 12, 2020. DOI: 10.1061/(ASCE)CO.1943-7862.0001945.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2. ed. New York: Springer, 2009.

HUANG, Lei et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Transactions on Information Systems, v. 43, n. 2, 2025. DOI: 10.1145/3703155.

JURAFSKY, Daniel; MARTIN, James H. Speech and Language Processing. 3. ed. Online manuscript, 2026.

KOHAVI, Ron. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence. Montreal, 1995. p. 1137–1145.

KOWSRIHAWAT, Kankawin; VATEEKUL, Peerapon. An Information Extraction Framework for Legal Documents: A Case Study of Thai Supreme Court Verdicts. In: 12th International Joint Conference on Computer Science and Software Engineering. 2015. p. 275–280.

LEWIS, Patrick et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: Advances in Neural Information Processing Systems. 2020. p. 9459–9474.

LIMA, Marcos Cavalcanti. Deep Vacuity: Detecção e Classificação Automática de Padrões com Risco de Conluio em Dados Públicos de Licitações de Obras. 2021. Dissertação (Mestrado em Informática) – Universidade de Brasília, Brasília.

LIU, Nelson F. et al. Lost in the Middle: How Language Models Use Long Contexts. arXiv:2307.03172, 2023.

MOTA, Lucélia Vieira. Reconhecimento de entidades nomeadas para conteúdo publicado em diários oficiais com base em uma abordagem de supervisão fraca. 2023. Dissertação (Mestrado em Informática) – Universidade de Brasília, Brasília.

PEREZ, Ethan; KIELA, Douwe; CHO, Kyunghyun. True Few-Shot Learning with Language Models. arXiv:2105.11447, 2021.

PYTHON SOFTWARE FOUNDATION. difflib — Helpers for Computing Deltas. 2025.

QWEN et al. Qwen2.5 Technical Report. arXiv:2412.15115, 2025.

VASWANI, Ashish et al. Attention Is All You Need. In: Advances in Neural Information Processing Systems. 2017.

WEI, Jason et al. Emergent Abilities of Large Language Models. Transactions on Machine Learning Research, 2024.

YANG, An et al. Qwen3 Technical Report. arXiv:2505.09388, 2025.