



UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA
CURSO DE GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Felipe Valentin Nascimento

**Extração de Dados a Partir de Publicações do Diário Oficial dos Municípios de Santa
Catarina usando Large Language Models**

Florianópolis
2025

Felipe Valentin Nascimento

Extração de Dados a Partir de Publicações do Diário Oficial dos Municípios de Santa Catarina usando Large Language Models

Trabalho de Conclusão de Curso do Curso de Graduação em Ciência da Computação do Centro Tecnológico da Universidade Federal de Santa Catarina para a obtenção do título de Bacharel em Ciência da Computação.
Orientadora: Profa. Carina Friedrich Dorneles, Dra.

Florianópolis
2025

Ficha catalográfica gerada por meio de sistema automatizado gerenciado pela BU/UFSC.
Dados inseridos pelo próprio autor.

NASCIMENTO, FELIPE VALENTIN

Extração de Dados a Partir de Publicações do Diário Oficial dos Municípios de Santa Catarina usando Large Language Models / FELIPE VALENTIN NASCIMENTO ; orientadora, Carina Friedrich Dorneles, 2025.

71 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Tecnológico, Graduação em Ciências da Computação, Florianópolis, 2025.

Inclui referências.

1. Ciências da Computação. 2. Licitações. 3. Large Language Models. 4. Extração de Dados. 5. Saídas Estruturadas. I. Dorneles, Carina Friedrich. II. Universidade Federal de Santa Catarina. Graduação em Ciências da Computação. III. Título.

Felipe Valentin Nascimento

Extração de Dados a Partir de Publicações do Diário Oficial dos Municípios de Santa Catarina usando Large Language Models

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de “Bacharel em Ciência da Computação” e aprovado em sua forma final pelo Curso de Graduação em Ciência da Computação.

Florianópolis, 23 de 06 de 2025.

Profa. Lúcia Helena Martins Pacheco, Dra.
Coordenadora do Curso

Banca Examinadora:

Profa. Carina Friedrich Dorneles, Dra.
Orientadora
Universidade Federal de Santa Catarina

Prof. Renato Fileto, Dr.
Universidade Federal de Santa Catarina

Vitória Soares dos Santos, Me.
Universidade Federal de Santa Catarina

Florianópolis, 2025.

Este trabalho é dedicado aos meus amigos, minhas
irmãs e aos meus queridos pais.

AGRADECIMENTOS

Agradeço à professora Carina pela orientação, apoio e contribuições fundamentais ao longo deste trabalho, bem como ao projeto CEOS pelos recursos disponibilizados para os experimentos.

Meus pais, Valdir e Iara, mostraram-me o verdadeiro significado do amor incondicional, e minhas irmãs, Cíntia e Cláudia, foram o suporte de que precisei nos momentos mais difíceis.

Também guardo enorme apreço pelos amigos que sempre me arrancam risos sinceros e pela minha namorada, cuja presença suavizou ansiedades e inseguranças.

Por fim, reconheço o cuidado dos doutores Fabrício e Marcelo, que trataram não apenas do corpo, mas acolheram-me com empatia, atenção e respeito.

Sou profundamente grato a cada um de vocês por tornarem possível a conclusão deste trabalho.

RESUMO

A extração de dados estruturados a partir de documentos é fundamental para viabilizar seu uso em sistemas computacionais, análises automatizadas e políticas públicas baseadas em evidências. No caso do Diário Oficial dos Municípios de Santa Catarina (DOM/SC), embora as publicações sejam parcialmente estruturadas em formato JSON, um dos campos contém um texto extenso em linguagem natural, que concentra a maioria das informações relevantes. Diante desse cenário, este trabalho investiga o uso de Large Language Models (LLMs) para extrair automaticamente dados estruturados de documentos de publicações do Diário Oficial dos Municípios de Santa Catarina (DOM/SC) referentes a Licitações Públicas. O foco recai sobre um subconjunto de publicações estratificadas, já que o DOM/SC contempla tanto a divulgação completa quanto a parcial de atos administrativos. Após uma exploração inicial dos dados, que contam com mais de 110 mil publicações filtradas por categoria e tamanho, foi criado um conjunto de dados ouro de 100 documentos que serviu de base para a validação dos experimentos. Foram anotadas 11 variáveis, sendo 6 obrigatórias e 5 opcionais. O fluxo experimental integra pré-processamento, construção de prompts, uso opcional de Retrieval-Augmented Generation (RAG) e validação automática por JSON Schema. O prompt final foi avaliado com quatro modelos abertos de última geração (Llama 3.3-70B, Qwen 2.5-72B, DeepSeek R1-Llama-70B e Gemma 3-27B). Os resultados mostram que a combinação de few-shot com saídas estruturadas, recurso dos provedores de LLM que força a resposta em JSON válido segundo um esquema definido, reduziu a taxa de erros de formatação a zero, alcançando um F1-score médio de 89,2% calculado como a média aritmética dos F1-scores dos 11 campos avaliados, com os valores por campo variando de 81% a 100%. Os diferentes modelos apresentaram métricas semelhantes, com o melhor desempenho obtido pelo llama 3.3-70B. Conclui-se que, com prompts bem elaborados, os LLMs extraem entidades de documentos de licitação de modo confiável, dispensando o uso de RAG em textos curtos, fortalecendo a transparência e a fiscalização das contratações públicas.

Palavras-chave: Licitações. Large Language Models. Extração de Dados. Saídas Estruturadas. Retrieval-Augmented Generation.

ABSTRACT

Extracting structured data from documents is essential for their integration into computational systems, automated analyses, and evidence-based public policy. In the Diário Oficial dos Municípios de Santa Catarina (DOM/SC), although notices are partly provided in JSON, one field still contains a lengthy free-text description that holds most of the relevant information. This study investigates the use of large language models (LLMs) to automatically extract structured data from DOM/SC procurement notices. We focus on a stratified subset of publications, because the DOM/SC includes both full and partial disclosures of administrative acts. After exploring more than 110 000 publications filtered by category and size, we compiled a 100-document gold corpus to validate the experiments. Eleven variables were annotated six mandatory and five optional. The experimental pipeline comprises preprocessing, prompt construction, optional Retrieval-Augmented Generation (RAG), and automatic validation through a JSON Schema. The final prompt was tested on four state-of-the-art open models (Llama-3.3-70B, Qwen-2.5-72B, DeepSeek R1-Llama-70B, and Gemma-3-27B). Results show that combining few-shot prompting with structured outputs, a feature that forces models to return valid JSON according to a predefined schema, eliminated formatting errors and achieved an average F1-score of 89.2%, calculated as the arithmetic mean across the eleven fields (individual scores ranged from 81% to 100%). All models produced similar metrics, with Llama-3.3-70B performing best. We conclude that well-crafted prompts enable LLMs to extract entities from short procurement documents reliably, obviating RAG in this context and strengthening transparency and oversight of public contracting.

Keywords: Public procurement. Large Language Models. Information Extraction. Structured outputs. Retrieval-Augmented Generation.

LISTA DE FIGURAS

Figura 1 – Exemplo do JSON de uma publicação do DOM/SC	14
Figura 2 – Exemplo de NER	18
Figura 3 – Exemplo de JSON schema	21
Figura 4 – Exemplo simplificado de RAG	22
Figura 5 – Visão geral do desenvolvimento	26
Figura 6 – Exemplo completo de objeto JSON do DOM	28
Figura 7 – Exemplo do conjunto de dados ouro anotado	29
Figura 8 – Exemplo de Documento de Licitação	30
Figura 9 – Pipeline de extração com a definição das etapas por configuração	33
Figura 10 – Exemplo de arquivo YAML de configuração de experimento	35
Figura 11 – F1-score dos experimentos 0 e 9	40
Figura 12 – Linha de base e melhor resultado com exemplos de saída	42
Figura 13 – Contexto bruto abreviado	42
Figura 14 – Comparação entre linha de base e melhor resultado RAG	43
Figura 15 – Contexto abreviado construído com Retrieval-Augmented Generation (RAG) do documento	44
Figura 16 – Resultado do quarto experimento com diferentes LLMs	45

LISTA DE TABELAS

Tabela 1 – Modalidades de Licitação: comparação entre a Lei nº 8.666/1993 e a Lei nº 14.133/2021	17
Tabela 2 – Modelos de linguagem selecionados: comparação de parâmetros e capacidade de contexto	19
Tabela 3 – Comparação dos trabalhos relacionados	25
Tabela 4 – Campos do arquivo JSON de publicações do DOM	27
Tabela 5 – Tipo de Documento e frequência	31
Tabela 6 – Município e frequência	31
Tabela 7 – Modalidade e frequência	32
Tabela 8 – Formato da Modalidade e frequência	32
Tabela 9 – Completude dos outros campos	32
Tabela 10 – Campos de configuração de cada experimento	34
Tabela 11 – Configuração 0 do primeiro experimento	38
Tabela 12 – Configuração 9 do primeiro experimento (melhor resultado)	38
Tabela 13 – Comparação dos resultados do primeiro experimento	39
Tabela 14 – Resultados da configuração 0 do primeiro experimento	39
Tabela 15 – Resultado da configuração 9 do primeiro experimento	39
Tabela 16 – Comparação dos resultados do segundo experimento	41
Tabela 17 – Resultados do Few-shot (saída apenas)	41
Tabela 18 – Conjunto de valores para cada parâmetro do RAG.	42
Tabela 19 – Comparação dos resultados do terceiro experimento	43
Tabela 20 – Melhor resultado RAG (chunk-size: 256, overlap: 16, top-k: 3)	43
Tabela 21 – Comparação dos resultados do quarto experimento	45
Tabela 22 – Resultado gemma3:27b	46
Tabela 23 – Resultado deepseek-r1:70b	46
Tabela 24 – Resultado qwen2.5:72b	47

LISTA DE ABREVIATURAS E SIGLAS

BERT	Bidirectional Encoder Representations from Transformers
CIGA	Consórcio de Inovação na Gestão Pública
DOM/SC	Diário Oficial dos Municípios de Santa Catarina
GPT	Generative Pre-trained Transformer
JSON	JavaScript Object Notation
LLM	Large Language Model
ML	Machine Learning
NER	Named Entity Recognition
NLP	Natural Language Processing
PLN	Processamento de Linguagem Natural
RAG	Retrieval-Augmented Generation
REGEX	Regular Expression

SUMÁRIO

1	INTRODUÇÃO	13
1.1	OBJETIVOS	15
1.1.1	Objetivo Geral	15
1.1.2	Objetivos Específicos	15
1.2	ORGANIZAÇÃO DO TRABALHO	15
2	FUNDAMENTAÇÃO TEÓRICA	16
2.1	LICITAÇÕES PÚBLICAS	16
2.1.1	Licitações em Santa Catarina	17
2.2	NAMED ENTITY RECOGNITION	18
2.3	LARGE LANGUAGE MODELS	18
2.3.1	Introdução dos Modelos Selecionados	19
2.3.1.1	LLaMA 3.3	19
2.3.1.2	Qwen 2.5	19
2.3.1.3	DeepSeek R1	20
2.3.1.4	Gemma 3	20
2.4	SAÍDAS ESTRUTURADAS	20
2.5	TÉCNICAS DE PROMPT	21
2.5.1	Introdução às Técnicas Selecionadas	21
2.6	RETRIEVAL-AUGMENTED GENERATION	22
3	TRABALHOS RELACIONADOS	23
3.1	ABORDAGENS BASEADAS EM REGRAS E PRIMEIROS TRANSFOR- MERS	23
3.2	LARGE LANGUAGE MODELS	24
3.3	LARGE LANGUAGE MODELS COM RETRIEVAL-AUGMENTED GE- NERATION	24
3.4	DIFERENCIAL DESTE TRABALHO	24
4	FLUXO PARA EXTRAÇÃO DE DADOS	26
4.1	EXPLORAÇÃO DOS DADOS	27
4.1.1	Origem	27
4.1.2	Análise dos Documentos	28
4.1.3	Características do Conjunto de Dados	30
4.2	EXTRAÇÃO DOS DADOS	32
4.2.1	Configurações	34
4.3	MÉTRICAS DE AVALIAÇÃO	35
5	RESULTADOS	37
5.1	EXPERIMENTO COM ENGENHARIA DE PROMPT	37
5.2	EXPERIMENTO COM EXEMPLOS	40

5.3	EXPERIMENTO COM RAG	41
5.4	EXPERIMENTO COM DIFERENTES LLMS	45
5.5	DISCUSSÃO DOS RESULTADOS	46
5.5.1	Análise dos Erros	47
6	CONCLUSÃO	49
	REFERÊNCIAS	50
	APÊNDICE A – MELHOR PROMPT	55
	APÊNDICE B – ARTIGO DO TCC	59
	APÊNDICE C – CÓDIGO FONTE	70

1 INTRODUÇÃO

A transparência nas contratações públicas é um direito constitucional que só se concretiza quando cidadãos, órgãos de controle e a sociedade conseguem fiscalizar cada etapa do processo licitatório (BRASIL, 1988). Entretanto, o volume crescente de editais, avisos, homologações e demais atos de licitação publicados diariamente inviabiliza qualquer verificação exclusivamente manual. Em Santa Catarina, por exemplo, o Diário Oficial dos Municípios de Santa Catarina (DOM/SC)¹ já registra milhares de publicações em um único dia, sobrecarregando auditores e comprometendo a detecção de irregularidades (CIGA, 2023). Esse cenário impõe um desafio técnico relevante: grande parte das informações está contida em campos textuais extensos e pouco estruturados, o que dificulta sua transformação em dados utilizáveis por sistemas automatizados.

Para enfrentar tal desafio, a área de Extração de Informações vem evoluindo rapidamente. Soluções iniciais basearam-se em Named Entity Recognition (NER) com base em padrões e expressões regulares, exigindo manutenção contínua e específica para cada base documental (KOWSRIHAWAT; VATEEKUL, 2015). Em seguida, modelos estatísticos capturaram relações estruturais entre palavras, aumentando a robustez da extração (RUSINOL; BENKHELFALLAH; POULAIN DANDECY, 2013). O advento de arquiteturas baseadas em aprendizado de máquina, como o BERT, reduziu ainda mais a dependência de regras rígidas e melhorou a generalização (HAMDI *et al.*, 2021). Esses avanços beneficiaram textos jurídicos e administrativos, nos quais a consistência terminológica é limitada e as exceções são frequentes (MOENS; UYTENDAELE; DUMORTIER, 1999).

Nos últimos anos, Large Language Models (LLMs) surgiram como alternativa promissora para o Processamento de Linguagem Natural (PLN) jurídico, graças à flexibilidade de adaptação a novos domínios (KATZ *et al.*, 2023). Ainda assim, esses modelos podem perder desempenho em contextos extensos (LIU, N. F. *et al.*, 2023), apresentar dificuldades em domínios altamente especializados (KANDPAL *et al.*, 2023) e gerar alucinações quando o treinamento ou o contexto fornecido são insuficientes (HUANG *et al.*, 2025). Abordagens de RAG atenuam parte desses problemas ao adicionar trechos relevantes ao prompt (GAO *et al.*, 2024), enquanto a exigência de saídas estruturadas, por exemplo, validadas por *JSON Schema*, reduz violações de formato e classes inexistentes (LIU, Y. *et al.*, 2024; LIU, M. X. *et al.*, 2024).

Quanto às publicações nos Diários Oficiais, elas constituem o meio oficial de divulgar atos administrativos, em especial os relativos às licitações: a Lei 8.666/1993 (e, hoje, a Lei 14.133/2021) tornou obrigatório que cada ente federativo publique, em seu diário, um extrato do edital e demais fases do processo, garantindo transparência e amplo acesso público (BRASIL, 2021). Esses avisos resumidos antecipam informações essenciais como objeto, modalidade, prazos, sem reproduzir todo o documento. Em Santa Catarina, tais publicações vêm do DOM/SC, mantido pelo Consórcio de Inovação na Gestão Pública (CIGA) desde 2007, que disponibiliza, além dos diários em formato PDF, as publicações em JSON no seu portal CIGA Dados².

¹ <https://www.diariomunicipal.sc.gov.br/>

² <https://dados.ciga.sc.gov.br/>

Figura 1 – Exemplo do JSON de uma publicação do DOM/SC

```
{
  "codigo": "6320072",
  "titulo": "PROCESSO ADMINISTRATIVO 248/2024",
  "data": "2024-08-15 16:10:00",
  "cod_registro_info_sfinge": "A5BBAA77FB6B4921B634BD6033504887BC00DEFB",
  "municipio": "Arvoredo",
  "entidade": "Prefeitura Municipal de Arvoredo",
  "categoria": "Licitações",
  "link": "https://www.diariomunicipal.sc.gov.br/?q=id:6320072",
  "texto": "<html xmlns=\"http://www.w3.org/1999/xhtml\">\n\n  <head>\n\n    <meta name=\"pdf:PDFVersion\" conter",
  "url": "https://www.diariomunicipal.sc.gov.br/arquivosbd/atos/2024/08/1723748994_homologao_extrato.pdf"
},
```

Fonte: Próprio autor (2025)

Cada objeto JSON corresponde a uma publicação, vide Figura 1, que reúne os seguintes campos: "codigo" representa o identificador único da publicação no portal; "titulo" é o nome do documento (neste caso, um processo administrativo); "data" indica a data de publicação no sistema; "municipio" e "entidade" identificam o ente público responsável; "categoria" classifica o tipo do ato; "link" e "url" direcionam o usuário à publicação original e ao PDF correspondente. O campo "texto" armazena o conteúdo integral em HTML e servirá de entrada para as LLMs extrair automaticamente as variáveis definidas.

Este trabalho, portanto, investiga se LLMs abertos podem automatizar, com confiabilidade, a extração de dados essenciais de publicações feitas no DOM/SC, fortalecendo a transparência pública ao oferecer dados prontos para consulta e auditoria. Destaca-se que as publicações são extratos resumidos, e não documentos licitatórios completos, o que impõe desafios adicionais: ausência frequente de campos relevantes, informações distribuídas de forma inconsistente e publicações de etapas distintas. Para contornar esses obstáculos e garantir aplicabilidade prática, foram selecionados documentos com até 32.000 caracteres, um limite que abrange cerca de 90% de todo o acervo. Neste cenário de documentos relativamente curtos, testaram-se sistematicamente diversos modelos para extrair 11 variáveis definidas, 6 obrigatórias (tipo do documento, número do processo licitatório, município, modalidade, número da modalidade e objeto) e 5 opcionais (formato da modalidade, data de abertura, site do edital, signatário e cargo do signatário), de um conjunto de 100 publicações anotadas manualmente.

O fluxo experimental inclui pré-processamento, elaboração de prompts zero-shot e few-shot, validação automática das saídas e a comparação entre cenários com e sem RAG. Os resultados indicam que um prompt bem estruturado, aliado a saídas estruturadas, recurso dos provedores de LLM que força a resposta em JSON válido segundo um esquema definido, que são validadas, elimina erros de formatação e atinge desempenho competitivo sem necessidade de RAG, sugerindo que, para documentos curtos como os do DOM/SC, o uso de recuperação de contexto pode introduzir ruído e até degradar a extração. Esses achados confirmam o potencial dos LLMs de código aberto como ferramenta prática para fiscalização e fornecem um protocolo reprodutível para estudos futuros.

1.1 OBJETIVOS

1.1.1 Objetivo Geral

Aplicar LLMs para extração de dados de documentos de extratos publicados no DOM/SC e disponibilizados pelo CIGA.

1.1.2 Objetivos Específicos

- Realizar o levantamento e a análise do formato e da estrutura dos documentos de licitação disponíveis no DOM/SC por meio do CIGA;
- Definir as variáveis relevantes a serem extraídas a partir das publicações;
- Avaliar e selecionar modelos de LLMs adequados à tarefa de extração;
- Validar a acurácia da extração por meio de métricas apropriadas, utilizando um conjunto de dados anotado;
- Comparar o desempenho de diferentes abordagens baseadas em LLMs, incluindo variações de engenharia de prompt e uso opcional de RAG.

1.2 ORGANIZAÇÃO DO TRABALHO

Capítulo 2, Fundamentação Teórica, aborda os conceitos legais de licitação, NER, LLM, engenharia de prompts, RAG e geração de saídas estruturadas; Capítulo 3, Trabalhos Relacionados, apresenta um panorama das pesquisas em PLN aplicadas a documentos legais, destacando a contribuição deste estudo; Capítulo 4, Fluxo para Extração de Dados, descreve os documentos analisados, os desafios de padronização e o método experimental adotado; Capítulo 5, Resultados, traz a análise quantitativa e qualitativa dos experimentos, discutindo limitações e erros observados; por fim, Capítulo 6, Conclusão, sintetiza as principais contribuições, limitações e direções para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Esse capítulo traz os principais conceitos e tecnologias que fundamentam a proposta de extração automática de entidades em documentos de licitações públicas do estado de Santa Catarina. Inicialmente, apresenta-se uma visão geral sobre o processo licitatório e sua importância para a administração pública, com ênfase na realidade catarinense. Em seguida, são abordados os fundamentos de NER. O capítulo também explora os avanços recentes em LLMs, com destaque para os modelos utilizados nos experimentos: LLaMA 3.3, Qwen 2.5, DeepSeek R1 e Gemma 3. Além disso, são discutidos os princípios de geração estruturada de saídas, técnicas de prompts, como *One-shot*, *Few-shot* e *Chain-of-thought*, e a abordagem de RAG, que requer embeddings.

2.1 LICITAÇÕES PÚBLICAS

A Constituição Federal de 1988 reconheceu o acesso à informação como um direito fundamental, reafirmando a publicidade como princípio essencial da Administração Pública (BRASIL, 1988). A partir desse marco, iniciativas foram adotadas para ampliar a transparência e disponibilizar dados sobre a gestão e o uso de recursos públicos como a Lei de Licitações e a Lei de Acesso à Informação.

Em conformidade com o artigo 37, inciso XXI, da Constituição Federal, foi instituída a Lei nº 8.666, de 21 de junho de 1993, conhecida como Lei de Licitações (BRASIL, 1993). A licitação é o procedimento administrativo em que se escolhe a proposta mais vantajosa para a contratação de obras, serviços, compras ou alienações, visando assegurar a legalidade e interesse público nas contratações governamentais (BRASIL, 2025). Ainda na legislação, em seu artigo 21, determinava a obrigatoriedade de publicação dos avisos de licitação no Diário Oficial da União, do Estado, do Distrito Federal ou do Município, conforme o ente federativo responsável pelo certame (BRASIL, 1993).

Apesar dos avanços em transparência ativa, ainda faltava uma norma que garantisse o direito de qualquer cidadão solicitar informações específicas de seu interesse (BRASIL, 2013). Essa lacuna foi preenchida com a promulgação da Lei nº 12.527/2011, conhecida como Lei de Acesso à Informação, que consolidou a política de transparência governamental e determinou que órgãos e entidades públicas disponibilizem proativamente informações de interesse público, preferencialmente por meio eletrônico e via internet, sem necessidade de solicitação expressa do cidadão (BRASIL, 2011).

A Lei nº 8.666/1993 esteve em vigor até dezembro de 2023 (BRASIL, 2025) e foi substituída pela Lei nº 14.133, de 1º de abril de 2021, que se manteve a determinação de publicação dos extratos dos documentos licitatórios nos Diários Oficiais correspondentes, substituiu as modalidades, conforme Tabela 1, e oficializou as fases do processo licitatório, conforme o artigo 17:

Art. 17. O processo de licitação observará as seguintes fases, em sequência:

I - preparatória;

- II - de divulgação do edital de licitação;
- III - de apresentação de propostas e lances, quando for o caso;
- IV - de julgamento;
- V - de habilitação;
- VI - recursal;
- VII - de homologação (BRASIL, 2021).

Tabela 1 – Modalidades de Licitação: comparação entre a Lei nº 8.666/1993 e a Lei nº 14.133/2021

Lei nº 8.666/1993	Lei nº 14.133/2021
Concorrência	Concorrência
Tomada de preços	
Convite	
Concurso	Concurso
Leilão	Leilão
Pregão (Lei nº 10.520/2002)	Pregão
RDC (Lei nº 12.462/2011)	
	Diálogo competitivo

Fonte: (OLIVEIRA, 2021)

2.1.1 Licitações em Santa Catarina

Em Santa Catarina, o CIGA, criado em 2007, tem como propósito apoiar a modernização da administração pública municipal por meio do uso de soluções tecnológicas que promovem cidades mais inteligentes e sustentáveis (CIGA, 2025b). Entre suas principais iniciativas, está a responsabilidade pela manutenção e operação do Diário Oficial Eletrônico dos Municípios de Santa Catarina, atualmente utilizado por mais de mil entidades nos mais de 290 municípios do estado (CIGA, 2025b).

Uma das funcionalidades disponibilizadas é a autopublicação, que permite que as próprias entidades públicas realizem diretamente a publicação de seus atos oficiais, destacando-se os benefícios como a flexibilidade de horários, permitindo publicações entre 7h e 21h (CIGA, 2023). Um reflexo do sucesso dessa iniciativa foi o recorde alcançado em maio de 2023, quando, no dia 26, foram registradas mais de 13 mil autopublicações em um único dia (CIGA, 2023).

Os exemplares do Diário Oficial dos Municípios de Santa Catarina estão disponíveis para download gratuito por meio do portal CIGA Dados. O portal reúne diversos conjuntos de dados abertos relacionados às atividades do consórcio, suas entidades parceiras e Diários Oficiais, permitindo que cidadãos, pesquisadores e gestores públicos possam consultar, reutilizar e analisar informações de interesse público (CIGA, 2025a). Esse repositório de informações oficiais representa uma fonte valiosa para tarefas automatizadas de extração de dados.

Figura 2 – Exemplo de NER

No dia **28 de abril de 2024**, a empresa **Nexora Dynamics** revelou a construção de seu novo laboratório de inovação em **Altalegre do Sul**, uma pequena cidade no interior de **Minas Gerais**. A CEO da companhia, **Júlia Noronha**, afirmou que o investimento inicial será de **R\$ 23,5 milhões**, com foco em tecnologias sustentáveis.

Data	Empresa	Localização	Pessoa	Valor
------	---------	-------------	--------	-------

Fonte: Próprio autor (2025)

2.2 NAMED ENTITY RECOGNITION

O NER é uma subárea do campo de Extração de Informações envolvendo PLN(NADEAU; SEKINE, 2007). Seu objetivo principal é identificar e categorizar automaticamente menções a entidades específicas em textos, como nomes de pessoas, organizações, localizações, datas, valores monetários e outras expressões nomeadas, conforme Figura 2 (NADEAU; SEKINE, 2007). O termo *named entity* foi consolidado durante a conferência MUC-6, onde se percebeu que reconhecer nomes próprios e expressões numéricas era crucial para extrair informações estruturadas de textos não estruturados, como notícias ou documentos administrativos (NADEAU; SEKINE, 2007).

Inicialmente, sistemas de NER baseavam-se em regras construídas manualmente, utilizando de *Regular Expression (REGEX)*, mas com o tempo abordagens de *Machine Learning (ML)* tornaram-se predominantes, oferecendo maior flexibilidade e adaptabilidade a diferentes domínios e idiomas (NADEAU; SEKINE, 2007). Apesar do avanço dos modelos de aprendizado de máquina genéricos para NER, que conseguem identificar entidades como datas e organizações de maneira razoável em diversos contextos, esses sistemas enfrentam sérias limitações quando aplicados a domínios especializados (KESHAVARZ *et al.*, 2022). Nessas situações, torna-se necessário adaptar os modelos ou treinar novos a partir de dados anotados manualmente, um processo custoso e que exige conhecimento técnico especializado (KESHAVARZ *et al.*, 2022).

Para contornar essas limitações, pesquisas recentes têm explorado o uso de LLMs na tarefa de NER (WANG *et al.*, 2023). Apesar das diferenças entre as tarefas, modelos como o GPT-NER reformulam a extração de entidades como uma tarefa de geração de texto, o que permite aproveitar o potencial desses modelos (WANG *et al.*, 2023). Experimentos indicam que essa abordagem tem alcançado resultados promissores, sobretudo em cenários com poucos dados anotados (WANG *et al.*, 2023).

2.3 LARGE LANGUAGE MODELS

Em 2017, Vaswani *et al.* (2017) publicaram o artigo *Attention Is All You Need*, no qual introduziram a arquitetura *Transformer*, um modelo de redes neurais baseado exclusivamente em mecanismos de atenção, eliminando o uso de redes recorrentes ou convolucionais. Essa

arquitetura deu origem as LLMs, que são redes neurais com bilhões de parâmetros treinadas sobre grandes volumes de texto, com o objetivo de adquirir capacidades avançadas de compreensão e geração de linguagem natural (MINAEE *et al.*, 2024). Dentre os modelos construídos com base nessa arquitetura, destacam-se o Bidirectional Encoder Representations from Transformers (BERT) e Generative Pre-trained Transformer (GPT), que representam, respectivamente, os *encoders* e *decoders* da arquitetura Transformer (GHOJOGH; GHODSI, 2020).

2.3.1 Introdução dos Modelos Selecionados

Neste trabalho, foram selecionados modelos GPT de linguagem de código aberto que representam o estado da arte em tarefas de compreensão e geração de linguagem natural. Os modelos foram escolhidos por sua disponibilidade pública, tamanho de janela de contexto, e desempenho competitivo em *benchmarks* e capacidade de adaptação a tarefas específicas, como NER. Todos os modelos analisados estão disponíveis na plataforma Ollama¹, que é um projeto aberto que permite baixar, gerenciar e executar LLM inteiramente local, sem depender da nuvem (OLLAMA, 2025). o que facilita sua experimentação e integração em ambientes locais de execução.

Tabela 2 – Modelos de linguagem selecionados: comparação de parâmetros e capacidade de contexto

Modelo	Parâmetros	Janela de contexto (Tokens)
LLaMA 3.3	70B	128K
Qwen 2.5	72B	128K
DeepSeek R1	70B	128K
Gemma 3	27B	128K

Fonte: Próprio autor (2025)

2.3.1.1 LLaMA 3.3

O Llama 3 foi desenvolvida pela equipe de inteligência artificial da Meta como uma evolução direta dos modelos Llama anteriores. Essa família de modelos é composta por modelos com 8, 70 e 405 bilhões de parâmetros, sendo o maior deles com janela de contexto de até 128 mil *tokens* (GRATTAFIORI *et al.*, 2024). Os modelos foram treinados com grandes volumes de dados multilíngue, mais de 15 trilhões de *tokens*, e apresentam capacidades nativas em múltiplos idiomas, programação, raciocínio e uso de ferramentas (GRATTAFIORI *et al.*, 2024). O modelo Llama 3.3 70B, ajustado para seguir instruções, se aproxima do desempenho do Llama 3.1 405B, mesmo sendo significativamente menor em tamanho (META, 2025).

2.3.1.2 Qwen 2.5

O Qwen 2.5 foi desenvolvido pela equipe da Alibaba Cloud e apresenta melhorias significativas em relação às anteriores, com destaque para o aumento do volume de dados de

¹ <https://ollama.com/>

pré-treinamento, de 7 para 18 trilhões de tokens, em um conjunto de dados multilíngue, o que amplia sua capacidade de compreensão e geração em diversos idiomas. Além disso, esses modelos suportam janelas de contexto de até 128 mil tokens. Os modelos da família Qwen 2.5 estão disponíveis em versões base e ajustadas por instruções, com tamanhos que variam de 0.5B a 72B de parâmetros (QWEN *et al.*, 2025).

2.3.1.3 DeepSeek R1

O DeepSeek R1 é uma família de modelos voltada para tarefas de raciocínio. No seu artigo, a equipe do DeepSeek demonstrou que é possível transferir as capacidades de raciocínio de modelos maiores para modelos menores por meio de uma técnica eficaz de destilação. Utilizando dados gerados pelo DeepSeek-R1, aproximadamente 800 mil amostras, foram ajustados diretamente modelos de código aberto amplamente utilizados, como o Qwen2.5 e o LLaMA 3, sem a necessidade de etapas adicionais. O modelo DeepSeek-R1-Distill-Llama-70B é o resultado disso e conta com o contexto de 128K e 70B de parâmetros (DEEPSEEK-AI *et al.*, 2025).

2.3.1.4 Gemma 3

O Gemma 3 é uma família de modelos leves e de código aberto desenvolvida pela Google DeepMind, com versões que variam de 1 a 27 bilhões de parâmetros. Esta versão incorpora capacidades multimodais, como compreensão visual, além de suporte ampliado a múltiplos idiomas e janelas de contexto estendidas de até 128 mil tokens. Os modelos Gemma 3 foram treinados com técnicas de destilação e apresentam desempenho superior em relação à versão anterior, Gemma 2, tanto nas versões pré-treinadas quanto ajustadas por instruções. Destaque para as melhorias em tarefas de matemática, diálogo, seguimento de instruções e uso multilíngue, com o modelo Gemma3-27B-IT atingindo desempenho comparável ao Gemini-1.5-Pro (TEAM *et al.*, 2025).

2.4 SAÍDAS ESTRUTURADAS

As saídas estruturadas são um recurso de provedores de LLMs que assegura que as respostas dos modelos sigam rigorosamente um formato definido em JSON, por meio de um JSON Schema fornecido nas chamadas. Em vez de gerar saídas livres, o modelo é orientado a produzir respostas estruturadas e válidas, facilitando a integração com sistemas e eliminando a necessidade de pós-processamento (POKRASS *et al.*, 2024). O JSON Schema, por sua vez, é uma linguagem declarativa para descrever a estrutura, os tipos e as restrições de dados JSON, conforme Figura 3. Com validadores apropriados, é possível garantir a conformidade dos dados com o esquema (JSON-SCHEMA-ORG, 2025). Para alcançar esse objetivo, o llama.cpp, que é o *backend* usado pelo Ollama, utiliza o GBNF (GGML Backus-Naur Form), uma gramática formal que restringe a geração do modelo apenas a sequências válidas. Durante a inferência,

Figura 3 – Exemplo de JSON schema

```
1 {
2   "$id": "https://example.com/person.schema.json",
3   "$schema": "https://json-schema.org/draft/2020-12/schema",
4   "title": "Pessoa",
5   "type": "object",
6   "properties": {
7     "nome": {
8       "type": "string",
9       "description": "O primeiro nome da pessoa."
10    },
11    "sobrenome": {
12      "type": "string",
13      "description": "O sobrenome da pessoa."
14    },
15    "idade": {
16      "description": "Idade em anos, que deve ser igual ou maior que
17        zero.",
18      "type": "integer",
19      "minimum": 0
20    }
21  }
```

Fonte: Próprio autor (2025)

essas gramáticas são convertidas em autômatos que controlam os tokens permitidos, garantindo saídas corretas mesmo em estruturas aninhadas ou complexas (GGML-ORG, 2025).

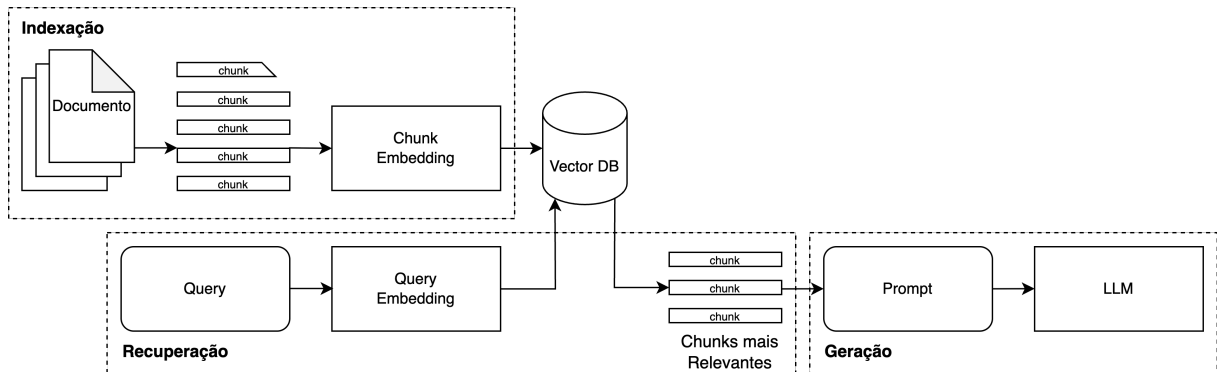
2.5 TÉCNICAS DE PROMPT

Os modelos LLM funcionam a partir de comandos chamados *prompts*, que podem ser textos, imagens, áudios ou vídeos. Compreender como estruturar, avaliar e aplicar corretamente os prompts é essencial para utilizar os modelos de linguagem de forma eficiente, pois, na prática, prompts bem formulados tendem a gerar resultados superiores, o que tem impulsionado o surgimento de diversas técnicas e um crescente volume de pesquisas voltadas ao aprimoramento desse processo (SCHULHOFF *et al.*, 2025). A seguir, são apresentadas breves explicações sobre algumas técnicas de prompt usadas no trabalho.

2.5.1 Introdução às Técnicas Seleccionadas

Role Prompting consiste em atribuir um papel específico ao modelo de linguagem dentro do prompt, como, por exemplo, pedir que ele atue como “Madonna” ou um “jornalista de viagens” (SCHULHOFF *et al.*, 2025).

Figura 4 – Exemplo simplificado de RAG



Fonte: Próprio autor (2025)

Zero/One/Few-Shot Prompting são abordagens em que o modelo é orientado a realizar uma tarefa com nenhum, um, ou 2 ou mais exemplos no próprio prompt respectivamente (SCHULHOFF *et al.*, 2025).

Zero Shot Chain-of-Thought A versão mais direta da técnica Chain-of-Thought consiste em adicionar ao prompt frases que induzem o modelo a seguir um raciocínio estruturado, como: Vamos pensar no passo a passo (SCHULHOFF *et al.*, 2025).

2.6 RETRIEVAL-AUGMENTED GENERATION

Embora os LLMs tenham avançado de forma expressiva, eles ainda falham em tarefas que exigem conhecimento especializado ou dados atualizados, produzindo as alucinações. A técnica de RAG ameniza esse problema ao buscar *chunks* semanticamente relevantes em um banco de dados vetorial externo e incorporá-los ao prompt, elevando a precisão das respostas (GAO *et al.*, 2024).

Um processamento com RAG de maneira simplificada envolve três etapas principais: indexação, recuperação e geração, conforme na Figura 4. A indexação inicia com a limpeza e extração de dados brutos em formatos variados os quais são convertidos em texto simples e segmentados em *chunks* menores para respeitar os limites de contexto dos modelos de linguagem. Esses *chunks* são transformados em vetores por meio de um modelo de *embedding* e armazenados em uma base vetorial, possibilitando buscas eficientes por similaridade (WU *et al.*, 2025). A recuperação é quando uma consulta é feita, ela é convertida em vetor e comparada com os vetores indexados para recuperar os *chunks* mais relevantes. Por fim, na geração os *chunks* são incorporados ao *prompt* enviado ao modelo de linguagem. Esse prompt pode incluir tanto os documentos recuperados quanto o histórico da conversa, permitindo ao modelo formular respostas mais precisas (GAO *et al.*, 2024).

3 TRABALHOS RELACIONADOS

A extração de informações varia de abordagens puramente baseadas em regras até soluções que combinam LLM com recuperação (RAG). Este capítulo revisita seis estudos que ilustram esse espectro, analisando metodologia, conjunto de dados, métricas-chave e limitações de cada um. O panorama resultante evidencia lacunas que esta pesquisa busca suprir no contexto das licitações públicas de Santa Catarina.

A Tabela 3 organiza os seis estudos revisados, comparando metodologia, conjunto de dados, resultados quantitativos e limitações principais. Optou-se por concentrar todos os indicadores de desempenho em uma única coluna intitulada Métrica porque cada pesquisa empregou medidas diferentes, por exemplo, alguns autores reportaram Precisão e Revocação, outros apenas Acurácia, e três estudos usaram F1-score, de acordo com a natureza do experimento e dos dados.

3.1 ABORDAGENS BASEADAS EM REGRAS E PRIMEIROS TRANSFORMERS

O estudo de (SLEIMI *et al.*, 2018) propõe, primeiro, um modelo de metadados para análise de requisitos legais e, segundo, regras automatizadas de extração desses metadados baseadas em Natural Language Processing (NLP), avaliadas em um estudo de caso que mostrou alta acurácia das anotações. Na prática, os autores utilizam Tregex em Java para definir regras de extração apoiadas por um pipeline de NLP ajustado para o francês. A avaliação, feita sobre um conjunto-verdade de 350 sentenças das leis de trânsito de Luxemburgo, 200 usadas para elaborar as regras e 150 para teste, registrou precisão de até 97,2% e revocação de até 94,9%. Apesar dos bons resultados, a abordagem depende da elaboração manual das regras, atividade demorada que não se escala facilmente para novos domínios ou idiomas.

Já o DODFMiner, apresentado por (GUIMARÃES *et al.*, 2024), automatiza a segmentação e a extração de informações dos Diários Oficiais do Distrito Federal em três etapas: saneamento do texto, classificação dos atos normativos via expressões regulares e reconhecimento de entidades nomeadas com diferentes arquiteturas de aprendizado de máquina. A base experimental reúne 13010 atos distribuídos em 18 categorias altamente desbalanceadas, a maior contém 2313 instâncias e a menor apenas 20, o que afeta a robustez dos modelos. Ainda assim, a ferramenta alcançou até 85% de F1-score na etapa de NER. A solução também fica limitada ao escopo desenvolvido e as expressões regulares criadas.

Embora as regras manuais e os primeiros Transformers alcancem boa precisão, a forte dependência de especialistas e a baixa portabilidade entre domínios indicam a necessidade de métodos mais flexíveis.

3.2 LARGE LANGUAGE MODELS

Faria, Xie e Steffek (2024) adotam GPT-4 com técnicas de engenharia de prompt para extrair 8 variáveis de 260 decisões do tribunal de trabalho do Reino Unido. Cada arquivo extraído traz a transcrição da sentença e metadados como datas, local da audiência e partes envolvidas. A solução atingiu acurácia acima de 97% para todos os campos. As limitações observadas incluem a omissão de detalhes processuais nos fatos, causando descrições incompletas e alucinação.

Já Hussain e Thomas (2024) avaliam LLaMA 3, Gemma, Mistral e Phi-3, com exemplos few-shot, para rotular 14 entidades jurídicas no conjunto de 16.570 documentos judiciais indianos com instruções para o retorno ser em JSON. O Mistral foi o mais preciso F1-score de 64%, mas a performance ainda é aquém do desejável e o retorno JSON não ser garantido. Os autores apontam que serão necessárias arquiteturas mais robustas para elevar a qualidade do reconhecimento de entidades jurídicas.

3.3 LARGE LANGUAGE MODELS COM RETRIEVAL-AUGMENTED GENERATION

Pereira *et al.* (2024) apresentam o INACIA, um sistema que incorpora LLMs ao fluxo de instrução processual do TCU para automatizar desde a extração de dados básicos até a emissão de recomendações. A solução combina RAG com o GPT-4 e foi validada em 122 casos já julgados, cada um com 33 variáveis anotadas. Nessa avaliação, o sistema alcançou 59 % de precisão, 40 % de revocação e F1-score de 43 %. Entre as limitações destacam-se os módulos da solução não atingiram o estado da arte devido à complexidade da solução de ajustar os parâmetros RAG e a falta de conhecimento jurídico entre a equipe de pesquisa.

Em linha semelhante, o estudo de (AQUINO *et al.*, 2024) propõe um fluxo de trabalho baseado em RAG para extrair dados estruturados de 40 documentos de licitação do Ministério Público de Santa Catarina (MPSC). A pesquisa explora várias combinações de parâmetros de RAG e modelos de linguagem, buscando a configuração ótima para cada campo. Focada em duas variáveis, a melhor combinação atingiu 90 % de acurácia, superando com folga a abordagem por expressões regulares 58,75 %. Contudo, o escopo restrito, pequena amostra, número limitado de variáveis e concentração em um único órgão, reduz a generalização dos resultados e sinaliza a necessidade de testes mais amplos e diversificados.

A incorporação de RAG enriquece as respostas com contexto mais preciso e curto, mas também impõe uma camada adicional de infraestrutura, que necessita de ajuste de parâmetro e depende da qualidade dos retornos e contexto construído.

3.4 DIFERENCIAL DESTES TRABALHOS

Este estudo, dedicado às licitações públicas de Santa Catarina, mostra que uma solução baseada em LLM, testando exclusivamente modelos abertos, atingiu o melhor desempenho sem recorrer a RAG. Como os documentos analisados são relativamente curtos, a etapa de recuperação

não agregou valor e chegou a degradar os resultados. O ganho nas métricas decorreu, em vez disso, da combinação entre uma engenharia de prompt bem estruturada e um JSON Schema que valida a resposta durante a própria geração, garantindo que todos os campos retornem apenas as classes previamente definidas. Essa validação estruturada elimina alucinações de rótulos inexistentes, simplifica a ingestão dos resultados por sistemas subsequentes e reforça a reprodutibilidade do processo.

Tabela 3 – Comparação dos trabalhos relacionados

Trabalho	Metodologia	Conjunto de dados	Métrica	Limitação
(SLEIMI <i>et al.</i> , 2018)	Regras+NLP	350 documentos de trânsito de Luxemburgo	Precisão 97,2% Revocação 94,9%	Falta de flexibilidade com regras
(GUIMARÃES <i>et al.</i> , 2024)	Expressões Regulares+NLP	13010 Atos do Diário Oficial do Distrito Federal	F1-score 85,1%	Regras fixas e dados desbalanceados
(FARIA; XIE; STEFFEK, 2024)	GPT-4	260 decisões do Tribunal do Trabalho do Reino Unido	Acurácia 97%	Omissão de detalhes processuais e alucinações e modelo fechado
(HUSSAIN; THOMAS, 2024)	LLMs	16.570 documentos judiciais indianos	F1-score 64%	JSON nem sempre correto
(PEREIRA <i>et al.</i> , 2024)	LLMs+RAG	122 casos julgados do Tribunal de Contas da União	F1-score 42,9%	Integração complexa e complexidade da configuração RAG
(AQUINO <i>et al.</i> , 2024)	LLMs+RAG	40 documentos jurídicos do Ministério Público de Santa Catarina	Acurácia 90%	Dataset e hardware limitados

Fonte: Próprio autor (2025)

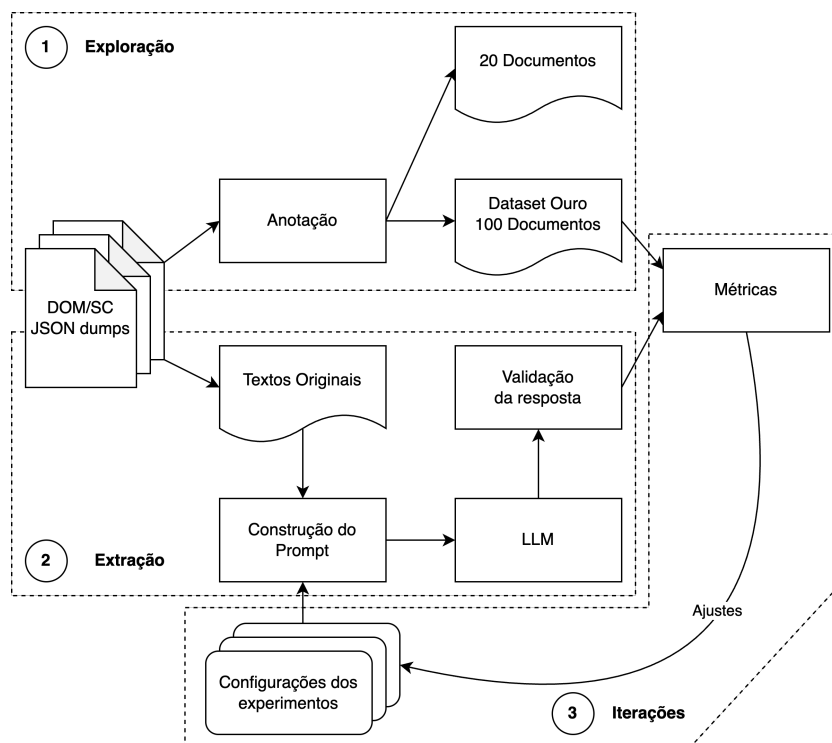
4 FLUXO PARA EXTRAÇÃO DE DADOS

O desenvolvimento do trabalho foi dividido em três etapas principais: a primeira consistiu na exploração dos dados; a segunda, na construção de um fluxo de experimentos; e a terceira, em iterações baseadas na avaliação de métricas e no refinamento das configurações dos experimentos. O objetivo central foi encontrar a solução mais eficaz, utilizando LLM, para a extração de campos previamente definidos durante a etapa de exploração.

A etapa de exploração trata da origem dos dados, de suas características, da definição dos campos a serem extraídos e do processo de anotação. Essa etapa foi dividida em dois momentos: primeiro, analisaram-se 20 documentos, nos quais foram observadas as características dos dados e do formato, e anotados alguns campos iniciais. Essa amostra inicial foi importante para definir, de forma simplificada, o processo de anotação dos documentos que compõem o conjunto de dados ouro. Em seguida, foram analisados 100 documentos aleatórios, anotando-se os principais campos previamente definidos, ajustando outros para refletirem melhor o conteúdo dos documentos e refinando a definição das classes de campos disponíveis.

A etapa de extração define os formatos de entrada e saída dos experimentos, assim como seus mecanismos de validação. Também é nela que se realiza a criação dos *prompts*, que podem ser configurados para diferentes tipos de experimentos. Isso inclui variações no texto do *prompt*, no *JSON Schema* da saída, na forma de inserção do contexto e na utilização de técnicas como RAG ou abordagens *one-shot* e *few-shot*.

Figura 5 – Visão geral do desenvolvimento



Fonte: Próprio autor (2025)

Por fim, na etapa de iterações realiza-se a geração das métricas. Os campos extraídos em formato JavaScript Object Notation (JSON) são comparados, campo a campo, com o conjunto de dados ouro, gerando as métricas necessárias para avaliar o desempenho de cada *prompt*. A partir desses resultados, novos experimentos foram elaborados com o objetivo de aprimorar os resultados e identificar as melhores configurações para os *prompts*.

4.1 EXPLORAÇÃO DOS DADOS

4.1.1 Origem

Os documentos analisados correspondem às autopublicações de licitações do DOM/SC¹, administradas pelo CIGA. Essas autopublicações estão disponíveis na página de dados abertos do consórcio². A etapa de coleta foi dispensada, pois os arquivos já haviam sido obtidos pelo projeto CEOS. Os documentos originais eram arquivos com listas de publicações, totalizando 122.532 documentos de licitação. O arquivo original é um JSON que reúne uma lista de autopublicações do Diário Oficial, cada item dessa lista corresponde a um JSON com os campos descritos na Tabela 4 e exemplificado na Figura 6. O campo *categoria*, por exemplo, permite filtrar especificamente as publicações de licitações.

Os documentos foram processados e armazenados em um banco de dados SQLite; selecionaram-se apenas publicações da categoria Licitação, a fim de facilitar a realização de análises subsequentes. Foi realizada uma análise exploratória do tamanho dos textos, com o objetivo de priorizar documentos mais curtos. Para isso, foram filtradas as publicações com até 32 mil caracteres. Essa faixa de tamanho correspondeu a aproximadamente 90% do total de documentos processados, resultando em um conjunto de 110.278 documentos selecionados para as etapas seguintes do trabalho.

Tabela 4 – Campos do arquivo JSON de publicações do DOM

Campo	Breve descrição
codigo	Identificador exclusivo no sistema CIGA.
titulo	Nome do arquivo.
data	Data e hora da publicação.
cod_registro_info_sfinge	Código de protocolo no e-Sfinge.
municipio	Ente federativo responsável.
entidade	Órgão ou unidade emissora.
categoria	Classificação do ato.
link	URL para visualização no portal DOM.
texto	HTML do conteúdo do PDF.
url	Link direto para o PDF oficial.

Fonte: Próprio autor (2025)

¹ <https://www.diariomunicipal.sc.gov.br/>

² <https://dados.ciga.sc.gov.br/>

Figura 6 – Exemplo completo de objeto JSON do DOM

```

1  "codigo": "6611928",
2  "titulo": "ERRATA Nº 001 - PE Nº 155/2024/SMS",
3  "data": "2024-11-12 17:43:19",
4  "cod_registro_info_sfinge": "01A844BB11CE682547D615CA1F6A386E150DE5EF",
5  "municipio": "Palhoça",
6  "entidade": "Prefeitura Municipal de Palhoça",
7  "categoria": "Licitações",
8  "link": "https://www.diariomunicipal.sc.gov.br/?q=id:6611928",
9  "texto": "<html xmlns=\"http://www.w3.org/1999/xhtml\">\n\n  <head><
    meta name=\"pdf:docinfo:producer\" content=\"Microsoft: Print To
    PDF\"/>\n\n    <meta name=\"pdf:docinfo:created\" content
    =\"2024-11-12T18:52:48Z\"/>\n\n    <title>Microsoft Word -
    AVISO DE ERRATA N&ordm; 001 - LAUDA.doc</title>\n\n  </head>\n\n
    <body>\n      <div class=\"page\">\n        <p>\n          ESTADO DE SANTA CATARINA \n\n          PREFEITURA MUNICIPAL DE
          PALHO&cedil;a \n\n          AVISO DE ERRATA N&ordm; 001 \n\n          Modalidade: Preg&
          atilde;o Eletr&ocirc;nico n&ordm; 155/2024/SMS. \n\n          Objeto: aquisi&
          ccedil;&atilde;o de 01 (UM) ve&iacute;culo tipo pickup com tra&
          ccedil;&atilde;o \n\n          4X4, para uso no SAMU, vinculado a Secretaria
          Municipal de \n\n          Sa&uacute;de do munic&iacute;pio de Palho&ccedil;a.
          \n\n          Justificativa: Altera&ccedil;&atilde;o no Termo de Refer&ecirc;
          ncia (Anexo II). \n\n          Nova data de abertura: 28/11/2024 &agrave;s 09
          h00min. \n\n          Maiores informa&ccedil;&otilde;es: Setor de Licita&ccedil;
          &otilde;es da Prefeitura, \n\n          http://palhoca.atende.net;https://www.
          portaldecompraspublicas.\n\n          com.br/, cplpalhoca@gmail.com, (48)
          3220-0300 ramais 1733 \n\n          nou 1804. \n\n          Palho&ccedil;a, 12 de novembro
          de 2024. \n\n          EDUARDO FRECCIA \n\n          Prefeito Municipal \n\n        </p>\n\n      </div>\n\n    </body>\n</html>\n",
10 "url": "https://www.diariomunicipal.sc.gov.br/arquivosbd/atos
    /2024/11/1731444194_aviso_de_errata_n_001__lauda_extrato.pdf"

```

Fonte: Próprio autor (2025)

4.1.2 Análise dos Documentos

A partir desse conjunto, foram selecionados 20 documentos para análise de conteúdo. Eles contêm campos de interesse para a extração de dados. A Figura 8 exibe um Aviso de Licitação com alguns campos definidos, o objetivo foi extrair tudo que possível no texto de uma publicação do DOM, além daqueles já disponíveis no JSON inicial; tais campos possibilitam identificar o aviso, esclarecer sua modalidade e respectivos prazos, descrever o objeto, atribuir responsabilidades aos signatários e indicar o endereço eletrônico do edital completo para conferência. Os campos definidos estão listados e descritos abaixo.

- **Tipo do documento:** permite identificar o tipo de publicação; não é padronizado por lei e varia conforme o órgão responsável;

Figura 7 – Exemplo do conjunto de dados ouro anotado

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q			
1	Código	Título	Tipo do Documento	Nº do Processo	Data de Registro	DOM	cof. registro	info. sfings	Município	Entidade	Categoria	DOM	Modalidade	Formato	Nº do Processo	Objeto	Data Abertura	Nº de Informações	Signatário	Cargo do Signatário
2	6320072	PROCESSO ADM Homologação	248/2024	2024-08-15 16:01	A5B8A477F6B4921B63	Arvoredo				Prefeitura Munic Licitações			Dispensa de Licit Presencial		203/2024	CABO FLAT E CIL NULL	2024-08-15 16:01	NULL	NEURI MENEZES	Prefeito
3	4269056	AVISO DE LICITA AVISO DE LICITAÇÃO	145/2022	2022-10-24 09:01	884004E9048E5E3E098	Princesa				Prefeitura munic Licitações			Tomada de Preço NULL		19/2022	CONTRATAÇÃO (2022-11-16T08:10:00Z	https://princesa	MICHELE CRISTU	Presidente	
4	6603591	PUBLICAÇÃO DA HOM	Homologação	23/2024	2024-11-08 11:14	NULL				Canoinhas			Dispensa de Licit Eletrônico		16/2024	aquisição de fari NULL	NULL	JULIANE MUCHE	Secretária de Adminis	
5	5096837	TERMO DE HOM	Homologação	19/2023	2023-08-28 07:11	341F8CD8467871A58A1C	Mondai			Prefeitura munic Licitações			Pregão Eletrônico		4/2023	Registro de preço NULL	NULL	NULL	Assinatura do Respons	
6	4948849	PREGÃO ELETRONICO	Ata de Registro	08/2023	2023-07-11 09:01	80C61CF890943576AA09	Lauro Muller			Prefeitura munic Licitações			Pregão Eletrônico		068/2023	O objeto desta li 2023-07-24T09:10:00Z	www.lauronull	Diezom Silva da	Pregoeiro	
7	4863701	RESULTADO HOM	Homologação	172/2023	2023-06-05 07:54	64431841725557F8C1	Palhoça			Prefeitura Munic Licitações			Pregão Eletrônico		172/2023	Aquisição e inste NULL	www.palhoca.sc	ANDRÉ JOSÉ SIL	Secretária de Adminis	
8	4458462	PROD LIC PREF	de Aviso de Licitação	001/2023	2023-01-06 16:20	875ECD890CA768689D	Iguaçu			Prefeitura munic Licitações			Pregão Presencial		001/2023	Registro de preço 2023-01-18T08:10:00Z	NULL	CLORI PEROZA	Prefeita Municipal	
9	6589293	HOMOLOGAÇÃO	Homologação	249/2024	2024-11-05 14:00	NULL				Vieira			Pregão Eletrônico		113/2024	REGISTRO DE PR NULL	NULL	DORIVAL CARLO	PREFEITO MUNICIPAL	
10	4219193	AVISO DE LICITA AVISO DE LICITAÇÃO	169/2022	2022-09-28 16:50	7F2DC70D2E6165D5544C	São Lourenço de				Prefeitura munic Licitações			Pregão Presencial		093/2022	CONTRATAÇÃO (2022-10-13 08:00:00Z	www.saolouren	RAFAEL CALEFFI	Prefeito Municipal	
11	6425283	EXTRATO DE INE	Aviso de Licitação	64/2024	2024-09-13 17:40	A4FD8AC901180ED81C1J	Passo de Torres			Prefeitura munic Licitações			Inexigibilidade NULL		64/2024	contratação de i NULL	NULL	Altamir Catel Ca	Prefeito em Exercício	
12	6597942	LICITAÇÃO Nº 54	Adjucação	54/2024	2024-11-07 17:20	NULL				São Joaquin			Pregão NULL		10/2024	Contratação de i NULL	NULL	Giovani Nunes	Autoridade Compete	
13	6056089	TERMO DE HOM	Homologação	39/2024	2024-06-06 12:20	A4CD02A594CB90862F0	Balneário Arroio			Prefeitura munic Licitações			Pregão Eletrônico		9/2024	TABELA DE BANC NULL	NULL	NULL	NULL	
14	4094200	EXTRATO DO PRI	Aditivo	250/2021	2022-08-08 16:50	NULL				Jaraguá do Sul			Pregão Presencial		250/2021	Constitui objeto NULL	NULL	OTONIEL DA SILVA	Secretário Municipal c	
15	6179767	EXTRATO DO CO	Contrato	64/2024	2024-07-10 10:20	FCA11E4E7CC18348A70E	Jaguariuna			Prefeitura munic Licitações			Dispensa de Licit Presencial		36/2024	CONTRATAÇÃO (NULL	NULL	LAERTE SILVA DC	PREFEITO MUNICIPAL	
16	5660607	LICITAÇÃO: CON	Aviso de Licitação	01/2024	2024-02-23 09:50	62E4A9DFE51DCC34541F	Praia Grande			Prefeitura munic Licitações			Concorrência Eletrônico		01/2024	EXECUÇÃO DE O 2024-03-15T08:10:00Z	WWW.BLL.COM	FLADIMIR SCHAI	AGENTE DE CONTRAT	
17	3659370	EXTRATO DE COF	Contrato	020/2022	2022-03-04 11:10	EECAD8513E5B5AAE85AC	Porto União			Prefeitura munic Licitações			Pregão Eletrônico		009/2022	aquisição de tint NULL	NULL	Eliseu Mibach	Prefeito Municipal	
18	3642062	PROCESSO HOM	Homologação	157/2021	2022-02-24 13:10	NULL				Descanso			Prefeitura Munic Licitações		50/2021	SERVÇOS DE RE NULL	NULL	SADI INACIO BOI	Prefeito Municipal	
19	4137608	EXTRATO DE HOM	Homologação	136/2022	2022-08-26 13:20	D29C8363D5D2D11690E24	Indaial			Prefeitura Munic Licitações			Pregão Eletrônico		058/2022	AQUISIÇÃO DE C NULL	NULL	MARCIO MOISES	Secretário de Educaçã	

Fonte: Próprio autor (2025)

- **Município:** identifica um dos municípios de Santa Catarina;
- **Data de abertura:** refere-se ao início da sessão ou, em documentos antigos, à abertura dos envelopes; nem sempre está presente e aparece mais em avisos e editais;
- **Número do processo licitatório:** número interno do processo, comum a todos os tipos de licitação; às vezes igual ao número da modalidade;
- **Modalidade:** conforme categorias definidas por lei, como pregão, concorrência, tomada de preços, entre outras;
- **Formato da modalidade:** distingue se a modalidade é presencial ou eletrônica, é tratada separadamente porque nem sempre está explícita;
- **Número da modalidade:** número sequencial da modalidade específica, como o número do pregão ou da concorrência;
- **Objeto:** descrição do item a ser contratado ou adquirido;
- **Site do edital:** endereço eletrônico onde o edital completo pode ser consultado; presente apenas em alguns documentos;
- **Signatário:** pessoa que assina o documento, podendo ser o prefeito, secretário, pregoeiro, ou outro representante oficial;
- **Cargo do signatário:** função ocupada pelo signatário.

Após a análise exploratória, selecionou-se uma amostra aleatória de 100 extratos licitatórios, formando o conjunto de dados ouro mostrado na Figura 7. A tabela contém os campos originais e recebeu, por anotação manual, os demais atributos necessários à avaliação. Essa rotulagem possibilitou reconhecer padrões recorrentes, e estabelecem a base de referência para medir acurácia, precisão, revocação e F1-score.

Figura 8 – Exemplo de Documento de Licitação

ESTADO DE SANTA CATARINA
MUNICÍPIO DE SÃO LOURENÇO DO OESTE – SC
AVISO DE LICITAÇÃO

O MUNICÍPIO DE **SÃO LOURENÇO DO OESTE-SC**, ATRAVÉS DO SR. PREFEITO, torna público para conhecimento dos interessados, que realizará **Sessão de Licitação no dia 13/10/2022, às 08h00min**, no Centro Administrativo no endereço, Rua Duque de Caxias, 789, conforme especificado no **PROCESSO LICITATÓRIO Nº 169/2022. SISTEMA REGISTRO DE PREÇOS. MODALIDADE PREGÃO PRESENCIAL Nº 093/2022.** **OBJETO: CONTRATAÇÃO DE EMPRESA ESPECIALIZADA PARA LOCAÇÃO DE IMPRESSORAS MULTIFUNCIONAIS MONOCROMÁTICAS, INCLUINDO O FORNECIMENTO DE SUPRIMENTOS (EXCETO PAPEL) E OS SERVIÇOS DE MANUTENÇÃO E ASSISTÊNCIA TÉCNICA, A SEREM INSTALADAS NAS SECRETARIAS MUNICIPAIS DE SÃO LOURENÇO DO OESTE-SC E FUNDO MUNICIPAL DE SAÚDE.** Informações e/ou cópia da íntegra deste Edital: fone (49) 3344-8500 – ramal 8502. Site www.saolourenco.sc.gov.br.

São Lourenço do Oeste – SC, 28 de setembro de 2022.

RAFAEL CALEFFI
Prefeito Municipal

Fonte: DOM/SC (2025)

4.1.3 Características do Conjunto de Dados

Esta subseção descreve o perfil de 100 publicações de licitações extraídas de diários oficiais catarinenses. O conjunto é composto por 11 atributos, sendo que 6 deles estão sempre presentes, como o número do processo licitatório, número da modalidade, modalidade, tipo de documento, município e objeto, garantindo identificação clara e contexto mínimo de cada licitação. Os demais atributos surgem de forma intermitente: formato da modalidade, data de abertura, site do edital, signatário e cargo do signatário, que enriquecem a transparência quando disponíveis.

Conforme Tabela 5, entre os 100 documentos, predominam Homologações (35%) e Avisos de Licitação (30%), seguidos por Editais (6%) e uma variedade de outros tipos que ocorrem de forma esporádica (1 a 4%). Esse perfil indica que a amostra foca principalmente nas fases iniciais e finais do processo licitatório, com poucos registros de atos intermediários como Julgamento ou Resultado.

Os 100 documentos estão distribuídos por 61 municípios, de acordo com Tabela 6, o que demonstra ampla cobertura geográfica. Palhoça (6), Araranguá (4) e Canoinhas (4) concentram o maior número de publicações, ao passo que a maioria dos demais municípios aparece uma ou duas vezes.

No que se refere à modalidade, destacado pela Tabela 7, predomina o Pregão (48%), seguido pelas contratações diretas Dispensa de Licitação (23%) e Inexigibilidade (19%). Modalidades de maior complexidade processual, como Tomada de Preços (6%) e Concorrência (4%), apresentam participação menor.

Quanto ao formato da sessão (Tabela 8), 45% dos registros não informam esse campo.

Tabela 5 – Tipo de Documento e frequência

Tipo de Documento	N
Homologação	35
Aviso de Licitação	30
Ata de Registro de Preços	3
Adjudicação	2
Aditivo	3
Contrato	4
Julgamento	1
Errata	3
Ata de Recebimento e Abertura	3
Apostilamento	2
Edital	6
Aviso de Suspensão	1
Aviso de Cancelamento	1
Resultado	2
Ratificação	3
Anulação	1

Fonte: Próprio autor (2025)

Tabela 6 – Município e frequência

Município	N	Município	N	Município	N
Arvoredo	1	Princesa	1	Canoinhas	4
Mondaí	2	Lauro Müller	1	Palhoça	6
Ipuacú	1	Videira	2	São Lourenço do Oeste	1
Passo de Torres	1	São Joaquim	3	Balneário Arroio do Silva	1
Jaraguá do Sul	1	Jaguaruna	3	Praia Grande	1
Porto União	2	Descanso	1	Indaial	1
Alto Bela Vista	1	Capivari de Baixo	1	Modelo	1
Chapadão do Lageado	1	Nova Itaberaba	2	Herval d'Oeste	2
Tubarão	1	Concórdia	2	Santiago do Sul	1
Santo Amaro da Imperatriz	1	Nova Erechim	1	Romelândia	1
Ituporanga	2	Fraiburgo	3	Rio dos Cedros	1
São Bento do Sul	1	Irati	1	Matos Costa	1
Rio do Sul	2	Xaxim	1	Orleans	1
Garopaba	1	Abdon Batista	1	Rio do Oeste	2
Pouso Redondo	1	Braço do Norte	2	Rancho Queimado	1
São Martinho	2	Vitor Meireles	1	Chapecó	1
Armazém	1	Belmonte	1	São João do Sul	1
Águas de Chapecó	1	Araranguá	4	Palmeira	1
Jaborá	1	Agronômica	1	Tunápolis	1
Treze de Maio	1	Petrolândia	1	Caibi	2
Lindóia do Sul	1	Otacílio Costa	1	Ponte Serrada	3
Campo Alegre	1	São João do Oeste	1	Coronel Freitas	1
Timbó Grande	1	Guaraciaba	1	Salto Veloso	1

Fonte: Próprio autor (2025)

Tabela 7 – Modalidade e frequência

Modalidade	N
Dispensa de Licitação	23
Tomada de Preços	6
Pregão	48
Inexigibilidade	19
Concorrência	4

Fonte: Próprio autor (2025)

Tabela 8 – Formato da Modalidade e frequência

Formato da Modalidade	N
Presencial	29
Eletrônico	26
Ausente (NULL)	45

Fonte: Próprio autor (2025)

Tabela 9 – Completude dos outros campos

Campo	Preenchido	Ausente (NULL)
Número do Processo Licitatório	100	0
Número da Modalidade	100	0
Objeto	100	0
Data de Abertura	23	77
Site do Edital	31	69
Signatário	86	14
Cargo do Signatário	86	14

Fonte: Próprio autor (2025)

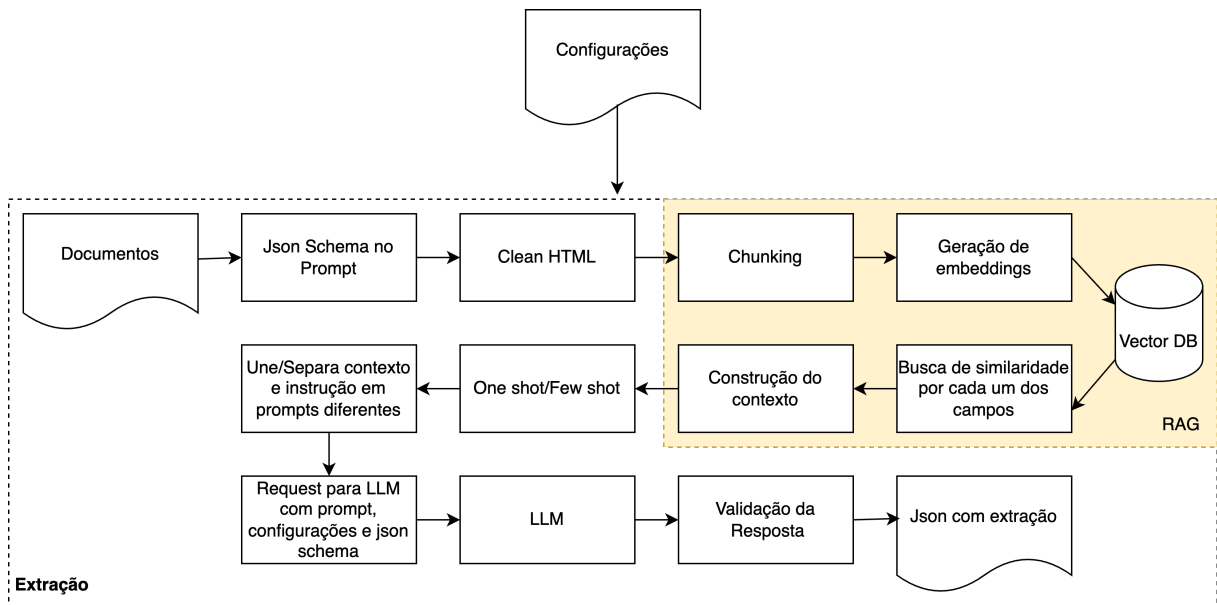
Entre os casos explicitados, observa-se leve predominância de sessões presenciais (29%) sobre eletrônicas (26%).

Por fim, a Tabela 9 sintetiza a completude de campos essenciais. Todos os documentos contêm número do processo, modalidade e descrição do objeto, garantindo identificação única e contexto mínimo. Em contrapartida, Data de Abertura (23%) e Site do Edital (31%) apresentam baixa taxa de preenchimento. Signatário e respectivo cargo figuram em 86% dos casos, proporcionando razoável rastreabilidade dos responsáveis.

4.2 EXTRAÇÃO DOS DADOS

Com os documentos selecionados e os campos de interesse definidos, implementou-se um fluxo de extração conforme a Figura 9. O servidor Ollama utilizado conta com os modelos LLM. O texto bruto de cada arquivo JSON do DOM é inicialmente pré-processado segundo as configurações de prompt, nas quais se decide o uso de RAG em que é usado o LangChain, a remoção de tags HTML e a adoção de estratégias one-shot ou few-shot. Em seguida, o conteúdo é

Figura 9 – Pipeline de extração com a definição das etapas por configuração



Fonte: Próprio autor (2025)

enviado através do Client Ollama a uma LLM acompanhado do JSON Schema, gerado pela classe Pydantic, que especifica a estrutura esperada da resposta. Por fim, a saída gerada é validada contra a classe Pydantic correspondente, garantindo conformidade e integridade dos dados extraídos.

O processo inicia-se com a ingestão dos textos brutos do JSON. A partir desse ponto, uma definição centralizada, carregada a partir de um arquivo de configurações, determina se cada módulo será executado, bem como os valores de todos os parâmetros. Essa abordagem garante rastreabilidade total dos experimentos e permite a replicação exata de cada execução.

Na fase de preparação do prompt, o esquema JSON que descreve a estrutura de saída é incorporado ao texto de sistema, assegurando que o modelo tenha, de antemão, o formato preciso a ser gerado.

Quando indicado na configuração, o corpo do documento é limpo: todas as marcações HTML e caracteres especiais são removidos, e o espaçamento é normalizado, preservando apenas o conteúdo relevante.

Caso o RAG esteja ativado, o texto limpo é dividido em blocos de tamanho e sobreposição controlados. Cada bloco é transformado em vetor semântico e indexado em um banco vetorial temporário. Em seguida, para cada campo alvo, realizam-se buscas de similaridade que retornam apenas os trechos mais pertinentes; esses trechos formam um contexto direcionado que acompanha as instruções no prompt final.

O sistema decide, conforme parametrização, se incluirá exemplos zero-shot, one-shot ou few-shot, bem como se todo o conteúdo será enviado em uma única mensagem ou dividido em papéis distintos como System e User.

A consulta ao modelo de linguagem é efetuada em *stream*, recebendo os trechos conforme são gerados, permitindo que a resposta seja monitorada em tempo real. O serviço recebe, além

do prompt, um esquema de saída, JSON Schema, de modo que o retorno já venha estruturado. Ao final, a resposta é submetida a uma validação da classe pydantic equivalente ao schema: como a geração do modelo sempre respeita o schema a validação é apenas para manusear o resultado com a classe equivalente ao schema.

Por fim, todas as execuções são temporizadas e registradas, viabilizando comparações objetivas entre diferentes combinações de parâmetros. Dessa forma, com exceção dos documentos de entrada e do JSON resultante, produto final do fluxo, cada decisão tomada pelo pipeline é completamente governada pela configuração global, o que confere ao sistema alto grau de reprodutibilidade, controle experimental e confiabilidade dos dados extraídos.

4.2.1 Configurações

Tabela 10 – Campos de configuração de cada experimento

Categoria	Campo	Descrição
Metadados	id	Identificador inteiro que distingue unicamente o experimento.
	description	Breve resumo do propósito ou da principal característica do ensaio.
Estratégia de execução	single_prompt	Define se a interação com o LLM ocorre em uma única chamada ou em duas chamadas uma de sistema para regras e outra de usuário.
	schema_version	Versão do esquema Pydantic usada no JSON Schema para garantir a geração do modelo.
	clean_context	Indica se o texto deve ter remoção de HTML e outros caracteres.
	shot	Abordagem de exemplo utilizada (e.g. <i>zero-shot</i> , <i>few-shot</i>).
	prompt	Instruções enviadas ao LLM, contendo os campos a extrair.
	model	Nome ou identificador do modelo selecionado (p. ex. <code>llama3.3:70b</code>).
RAG	chunk_size	Tamanho dos fragmentos de texto utilizados na recuperação.
	overlap	Número de tokens sobrepostos entre fragmentos consecutivos.
	top_k	Quantidade de passagens mais relevantes recuperadas para compor o <i>prompt</i> final.

Fonte: Próprio autor (2025)

Os parâmetros de cada experimento, resumidos na Tabela 10, são declarados em arquivos YAML, conforme exemplificado na Figura 10. Manter todas as definições de modelo, técnica de disparo (*shot*), limpeza de contexto e ajustes de RAG em um único artefato faz com que

Figura 10 – Exemplo de arquivo YAML de configuração de experimento

```
1 - id: 0
2   schema_version: 1
3   description: Extração básica de licitação
4   single_prompt: true
5   clean_context: true
6   shot: zero_shot
7   prompt: |-
8     Extraia as seguintes informações do documento de licitação...
9     Retorne em Json
10  model: llama3.3:70b
11  rag_configuration:
12    chunk-size: 512
13    overlap: 64
14    top-k: 3
```

Fonte: Próprio autor (2025)

o desenho experimental possa ser modificado sem alterar uma linha de código, preservando rastreabilidade e favorecendo reprodutibilidade.

Para assegurar a integridade dessas configurações, cada arquivo YAML é carregado por classes tipadas em Pydantic. Nesse instante o sistema verifica a presença de todos os campos obrigatórios, a coerência de tipos e requisitos semânticos — como descrições únicas e sequência crescente de identificadores. Qualquer violação dispara uma exceção e interrompe a execução, impedindo que combinações inválidas produzam resultados inconsistentes.

O YAML foi escolhido em vez de JSON porque oferece legibilidade superior em cenários que exigem edição frequente: aceita comentários nativos, suporta blocos multilinhas com a sintaxe (-) e dispensa a necessidade de escapar caracteres especiais. Essa combinação de clareza, anotações embutidas e suporte natural a textos extensos torna-o a opção mais prática para iterar rapidamente em configurações de experimentos com LLMs.

4.3 MÉTRICAS DE AVALIAÇÃO

A avaliação da extração foi dividida em dois grupos de campos enums (multiclasse) e não-enums, para que cada categoria receba métricas condizentes com o seu perfil:

- **Campos não-enum** (data de abertura, número do processo licitatório, número da modalidade, objeto, site do edital, signatário e cargo do signatário): aplicam-se as métricas padrão de acurácia, precisão, revocação e F1-score.
- **Campos enum** (tipo do documento, município, modalidade e formato da modalidade): a acurácia foi calculada de modo convencional, ao passo que precisão, revocação e F1-score

foram obtidos por meio da média macro, isto é, a média aritmética das métricas individuais de cada classe.

Em ambos os grupos, quando o conjunto ouro continha o rótulo `null`, as extrações correspondentes foram incluídas no cálculo da acurácia, mas excluídas dos cálculos de precisão, revocação e F1-score, a fim de evitar distorções nessas métricas. Para fins de clareza, adotam-se as seguintes definições:

- True Positive (TP) – a predição coincide com o valor do conjunto-ouro e esse valor não é `null`.
- False Positive (FP) – a predição fornece um valor diferente do conjunto-ouro ou prevê algo quando o rótulo é `null`.
- False Negative (FN) – o sistema não prevê valor (`null`) enquanto o conjunto-ouro contém um valor válido.
- True Negative (TN) – tanto o sistema quanto o conjunto-ouro indicam `null`.

As datas foram padronizadas no formato ISO YYYY-MM-DDThh:mm. Paralelamente, todo o conteúdo textual foi convertido para ASCII e submetido à remoção de pontuação e de espaços em excesso, mitigando penalidades decorrentes de variações meramente superficiais; além disso, para o campo objeto utilizou-se a distância de Levenshtein com limiar de 0,8, enquanto os demais campos exigiram correspondência exata.

5 RESULTADOS

Os resultados foram organizados em quatro experimentos sucessivos, cada um com o objetivo de refinar progressivamente o desempenho do sistema de extração de dados. O Llama-3.3-70B foi adotado como modelo de referência nos experimentos por três razões práticas: (i) apresenta desempenho de ponta em benchmarks públicos; (ii) está amplamente disponível no ecossistema de código aberto; e (iii) reduz o tempo de iteração durante o ajuste de prompts e a validação automática.

No primeiro experimento com engenharia de prompt, partiu-se de um prompt elementar que apenas solicitava a identificação dos campos delineados em um JSON Schema mínimo. A cada iteração, note-se que um único experimento é composto por várias iterações consecutivas, esse prompt foi enriquecido por meio da atribuição de um papel ao modelo, da inclusão de instruções específicas para cada entidade, de expressões regulares destinadas à validação e da definição explícita do formato de saída. Essa estratégia incremental possibilitou observar, de forma controlada, o impacto de cada ajuste sobre as métricas de acurácia, precisão, revocação e F1-score.

O segundo experimento com exemplos concentrou-se na inserção de exemplos. Comparou-se o desempenho do primeiro experimento e testaram-se abordagens *one-shot*, *few-shot* e casos pontuais cuidadosamente selecionados, a fim de verificar em que medida a demonstração explícita de padrões de resposta poderia orientar o modelo a produzir saídas mais consistentes.

Os resultados desse experimento serviram de base para o terceiro com RAG, dedicado à integração de *retrieval-augmented generation* (RAG). Nesse experimento, foram empregadas representações vetoriais obtidas com BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020), variando-se parâmetros de *chunk size*, *overlap* e *top-k*. O objetivo foi mensurar o ganho de contexto proporcionado pela recuperação de passagens relevantes e seu reflexo nas mesmas métricas avaliadas anteriormente.

No quarto experimento com LLMs, o melhor resultado do segundo experimento, obtido com exemplos e sem RAG, foi avaliado em diferentes modelos, incluindo Llama (linha de base), Gemma, Qwen e Deepseek. Essa avaliação entre modelos permitiu identificar diferenças de desempenho entre as LLMs, com melhores métricas em alguns campos e piores em outros.

5.1 EXPERIMENTO COM ENGENHARIA DE PROMPT

O primeiro experimento realizado consistiu na aplicação de técnicas básicas de engenharia de prompt. Inicialmente, foi utilizado um prompt simples, com a instrução direta, encontrada na Tabela 11. Nesse estágio inicial, não foi realizada nenhuma limpeza prévia do texto de entrada, e as instruções foram separadas do contexto em prompts distintos.

Apesar da simplicidade, este primeiro prompt já conseguiu realizar extrações básicas, o que foi possível graças ao uso do JSON Schema. Contudo, nesse cenário inicial, o JSON Schema atuava apenas como uma gramática definindo a geração da resposta, resultando em extrações

incompletas e pouco precisas devido à falta de especificidade nas instruções fornecidas à LLM.

A partir disso, foram realizadas diversas iterações para melhorar o desempenho da extração. Cada nova iteração envolveu a inclusão progressiva de elementos mais detalhados no prompt, tais como definição explícita do papel desempenhado pelo modelo, especificação clara dos campos a serem extraídos, definição rigorosa do formato de saída (JSON), tipos esperados para cada campo, aplicação de expressões regulares (regex) e regras adicionais. Essas regras adicionais tinham como objetivo evitar inferências indesejadas por parte do modelo, garantir que o texto original não fosse alterado e assegurar clareza na distinção entre campos obrigatórios e opcionais.

Em paralelo, também foram feitas melhorias no JSON Schema, incorporando, principalmente, expressões regulares (regex). A inclusão do regex no JSON Schema foi particularmente benéfica para campos com formatos pré-definidos, enquanto para campos multiclasse limitados a um conjunto específico de valores foi empregada a restrição do tipo "enum".

Tabela 11 – Configuração 0 do primeiro experimento

Campo	Descrição
ID	0
Versão do Schema	0
Descrição	Prompt contendo apenas uma instrução direta
Prompt único?	Não
Contexto limpo?	Não
Modelo	llama3.3:70b
Prompt utilizado	<i>Extraia as informações do documento de licitação.</i>

Fonte: Próprio autor (2025)

Tabela 12 – Configuração 9 do primeiro experimento (melhor resultado)

Campo	Descrição
ID	9
Versão do Schema	1
Descrição	Todas as características
Prompt único?	Sim
Contexto limpo?	Sim
Modelo	llama3.3:70b
Prompt utilizado	Apêndice A, mas sem os exemplos

Fonte: Próprio autor (2025)

A configuração que apresentou o melhor desempenho é a descrita na Tabela 12 que consolidou instruções detalhadas em um único prompt, aliadas à limpeza prévia do contexto textual. Os resultados alcançados por esse prompt são demonstrados nas tabelas e métricas correspondentes. Ao todo, foram realizados dez experimentos diferentes, sendo reportadas neste trabalho apenas as métricas do primeiro e do último experimento, visando destacar a evolução significativa obtida com o processo iterativo descrito.

Tabela 13 – Comparação dos resultados do primeiro experimento

Métrica	Configuração 0	Configuração 9	Variação absoluta
Acurácia média	57,09%	89,45%	+32,36 p.p.
F1-score médio	60,95%	87,29%	+26,34 p.p.

Fonte: Próprio autor (2025)

Tabela 14 – Resultados da configuração 0 do primeiro experimento

Campo	Acurácia	Precisão	Revocação	F1-score
Tipo do Documento	65.00%	63.62%	67.57%	65.54%
Número do Processo Licitatório	39.00%	39.00%	39.00%	39.00%
Município	91.00%	92.75%	90.82%	91.78%
Modalidade	97.00%	99.17%	96.84%	97.99%
Formato da Modalidade	56.00%	66.72%	98.08%	79.41%
Número da Modalidade	46.00%	46.00%	46.00%	46.00%
Objeto	65.00%	65.00%	65.00%	65.00%
Data de Abertura	1.00%	0.00%	0.00%	0.00%
Site do Edital	25.00%	21.05%	64.52%	31.75%
Signatário	68.00%	68.00%	79.07%	73.12%
Cargo do Signatário	75.00%	76.29%	86.05%	80.87%

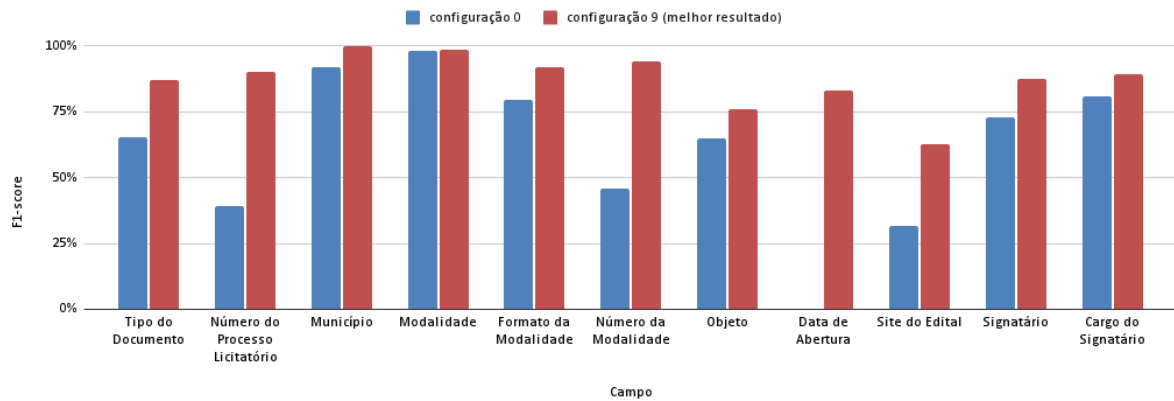
Fonte: Próprio autor (2025)

Tabela 15 – Resultado da configuração 9 do primeiro experimento

Campo	Acurácia	Precisão	Revocação	F1-score
Tipo do Documento	85.00%	83.14%	91.03%	86.91%
Número do Processo Licitatório	90.00%	90.00%	90.00%	90.00%
Município	100.00%	100.00%	100.00%	100.00%
Modalidade	98.00%	98.40%	98.53%	98.47%
Formato da Modalidade	92.00%	92.72%	91.18%	91.94%
Número da Modalidade	94.00%	94.00%	94.00%	94.00%
Objeto	76.00%	76.00%	76.00%	76.00%
Data de Abertura	93.00%	80.00%	86.96%	83.33%
Site do Edital	82.00%	58.33%	67.74%	62.69%
Signatário	85.00%	84.78%	90.70%	87.64%
Cargo do Signatário	89.00%	91.46%	87.21%	89.29%

Fonte: Próprio autor (2025)

Figura 11 – F1-score dos experimentos 0 e 9



Fonte: Próprio autor (2025)

Comparando os resultados apresentados na Tabela 14 e na Tabela 15 observa-se um avanço expressivo em todas as métricas principais. A acurácia média dos onze campos passou de aproximadamente 57,09% para 89,45%, enquanto o F1-score médio evoluiu de 60,95% para 87,29%, conforme pode ser visto na Tabela 13. Analisando o gráfico Figura 11 Destacam-se ganhos em campos originalmente críticos: Data de Abertura saltou de 1% para 93% de acurácia e Número do Processo Licitatório de 39% para 90%. Campos já fortes, como Município e Modalidade, atingiram ou se mantiveram acima de 98% de F1-score, com Município chegando a desempenho perfeito em todas as métricas. Esses resultados corroboram que os refinamentos introduzidos tornaram o sistema substancialmente mais robusto e confiável.

5.2 EXPERIMENTO COM EXEMPLOS

Este experimento da pesquisa foi concebido para investigar a influência de exemplos na tarefa de extração. Para isso, delinearam-se doze configurações de prompt que progridem em complexidade e variam tanto na quantidade quanto no tipo de exemplos fornecidos ao modelo de linguagem:

A configuração de linha de base corresponde ao melhor resultado do primeiro experimento. A configuração 1 (one-shot) acrescenta um exemplo completo para orientar formato e conteúdo, enquanto a configuração 2 exibe somente o JSON de saída desse exemplo, evitando o viés do texto de entrada. As configurações 3 e 4 (few-shot) utilizam múltiplos exemplos, primeiro com texto e JSON completos, depois somente com os JSONs. Já as configurações 5 e 6 (dynamic one-shot) selecionam em tempo de execução o exemplo mais semelhante ao documento analisado. As configurações 7 a 10 empregam exemplos altamente específicos, focados em estruturas ou campos particulares, a fim de verificar se a proximidade semântica melhora o desempenho. Por fim, a configuração 11 (chain of thought) faz o modelo "pensar no passo a passo".

A configuração 4, *few-shot* de saída, com os exemplos no prompt do Apêndice A, mostrou-se a alternativa mais eficaz para a extração de entidades em documentos de licitação. Ao suprimir

o texto de entrada dos exemplos e manter somente o JSON de referência, diminuiu-se o consumo da janela de contexto, permitindo alocar mais tokens ao documento analisado e evitando cortes de informações relevantes, sobretudo em documentos volumosos.

Os resultados confirmam essa vantagem, conforme Tabela 16. O F1-score médio passou de 87,29% no cenário de linha de base para 89,17,% na configuração 4 *few-shot* de saída, indicando um avanço na média mesmo com pequenas oscilações por campo. Conforme Tabela 17 e Figura 12 foram observadas melhorias expressivas em “Tipo do Documento” (+4,3 p.p.), “Objeto” (+5,0 p.p.), “Site do Edital” (+4,0 p.p.), “Signatário” (+5,5 p.p.) e “Cargo do Signatário” (+3,6 p.p.). A estabilidade manteve-se em “Número da Modalidade” e “Município”, que conservaram respectivamente 94 % e 100 % de F1-score. As ligeiras reduções em “Modalidade” (-0,5 p.p.), “Formato da Modalidade” (-0,3 p.p.) e “Data de Abertura” (-1,9 p.p.).

Em síntese, a elevação média do F1-score e o ganho substancial em campos como “Objeto” sustentam a recomendação do *few-shot* de saída apenas como configuração preferencial para tarefas de extração estruturada.

Tabela 16 – Comparação dos resultados do segundo experimento

Métrica	Linha de base	Configuração 4	Variação absoluta
Acurácia média	89,45%	91,00%	+1,55 p.p.
F1-score médio	87,29%	89,17%	+1,88 p.p.

Fonte: Próprio autor (2025)

Tabela 17 – Resultados do Few-shot (saída apenas)

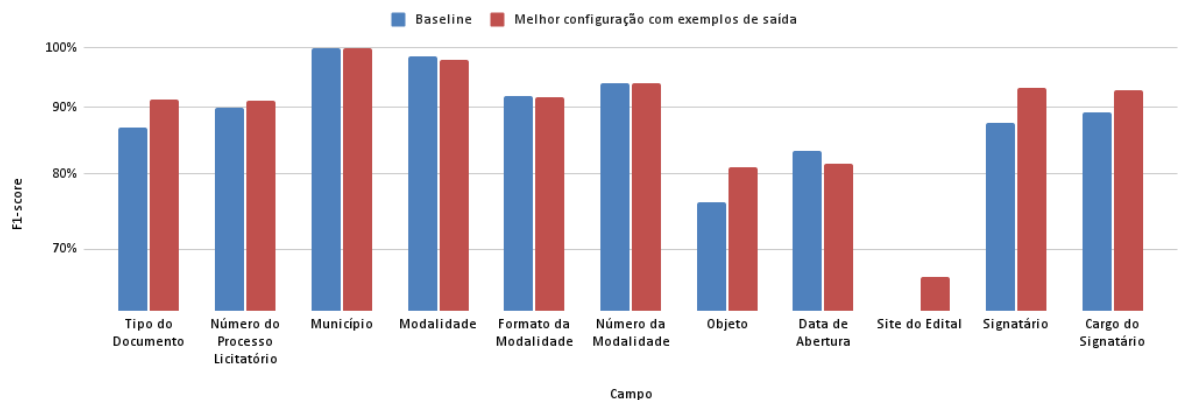
Campo	Acurácia	Precisão	Revocação	F1-score
Tipo do Documento	89.00%	88.89%	93.71%	91.23%
Número do Processo Licitatório	91.00%	91.00%	91.00%	91.00%
Município	100.00%	100.00%	100.00%	100.00%
Modalidade	97.00%	98.40%	97.48%	97.94%
Formato da Modalidade	92.00%	92.55%	90.78%	91.66%
Número da Modalidade	94.00%	94.00%	94.00%	94.00%
Objeto	81.00%	81.00%	81.00%	81.00%
Data de Abertura	90.00%	70.97%	95.65%	81.48%
Site do Edital	83.00%	58.54%	77.42%	66.67%
Signatário	92.00%	92.05%	94.19%	93.10%
Cargo do Signatário	92.00%	95.12%	90.70%	92.86%

Fonte: Próprio autor (2025)

5.3 EXPERIMENTO COM RAG

Neste experimento aplicou-se a estratégia de Retrieval-Augmented Generation (RAG) sobre o embedding *bert-large-portuguese-cased* (SOUZA; NOGUEIRA; LOTUFO, 2020), utilizando o *RecursiveCharacterTextSplitter* do *LangChain* e o *ChromaDB*

Figura 12 – Linha de base e melhor resultado com exemplos de saída



Fonte: Próprio autor (2025)

para busca por similaridade de cosseno. Os parâmetros avaliados estão descritos na Tabela 18. Totalizando 36 Configurações distintas.

Figura 13 – Contexto bruto abreviado

```
1 # Contexto
2 ## Nome do Documento
3 PROCESSO ADMINISTRATIVO 248/2024
4
5 ## Texto do Documento
6 MUNICIPIO DE ARVOREDO
7 Compras e Contratos
8
9 Termo Homologação - Termo de Homologação
10 Entidade - Processo Administrativo - Minuta - Licitação: 2185 Ano -
   Minuta - Licitação: 2024 Número -
11
12 Minuta - Licitação: 259 codigoCliente: 2185 anoMinuta: 2024
   cotaCredenciamento: Não
13 ...
```

Fonte: Próprio autor (2025)

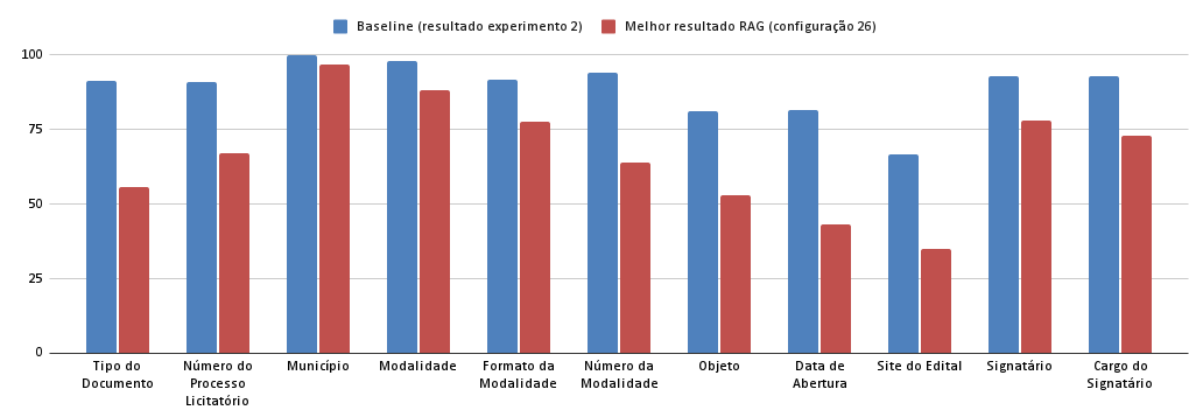
Tabela 18 – Conjunto de valores para cada parâmetro do RAG.

Chunk size	Chunk overlap	Top K
64, 128, 256	16, 32, 64	1, 3, 5, 10

Fonte: Próprio autor (2025)

As tabelas 13 e 15 mostram, respectivamente, o contexto original do documento e o contexto fragmentado pela estratégia RAG. Observa-se que, à medida que o *chunk size* e o *top-K*

Figura 14 – Comparação entre linha de base e melhor resultado RAG



Fonte: Próprio autor (2025)

Tabela 19 – Comparação dos resultados do terceiro experimento

Métrica	Linha de base	Melhor configuração (26) RAG	Variação absoluta
Acurácia média	91,00 %	63,55 %	-27,45 p.p.
F1-score médio	89,17 %	66,54 %	-22,63 p.p.

Fonte: Próprio autor (2025)

Tabela 20 – Melhor resultado RAG (chunk-size: 256, overlap: 16, top-k: 3)

Campo	Acurácia	Precisão	Revocação	F1-score
Tipo do Documento	38.00%	55.49%	55.73%	55.61%
Número do Processo Licitatório	67.00%	67.00%	67.00%	67.00%
Município	97.00%	97.10%	96.62%	96.86%
Modalidade	89.00%	92.09%	84.77%	88.28%
Formato da Modalidade	60.00%	68.06%	90.58%	77.72%
Número da Modalidade	64.00%	64.00%	64.00%	64.00%
Objeto	53.00%	53.00%	53.00%	53.00%
Data de Abertura	43.00%	27.85%	95.65%	43.14%
Site do Edital	42.00%	24.68%	61.29%	35.19%
Signatário	76.00%	75.27%	81.40%	78.21%
Cargo do Signatário	70.00%	69.47%	76.74%	72.93%

Fonte: Próprio autor (2025)

Figura 15 – Contexto abreviado construído com RAG do documento

```
1 # Contexto
2 ## Nome do Documento
3 PROCESSO ADMINISTRATIVO 248/2024
4
5 ## Texto do Documento
6 ### Tipo do documento
7 MUNICIPIO DE ARVOREDO
8 Compras e Contratos
9
10 Termo Homologação - Termo de Homologação
11 Entidade - Processo Administrativo - Minuta - Licitação: 2185
12 Ano - Minuta - Licitação: 2024
13 Número -
14
15 ### Número do processo administrativo
16 Minuta - Licitação: 259 codigoCliente: 2185 anoMinuta: 2024
    cotaCredenciamento: Não
17
18 Pág 1 / 1
19
20 IPM Sistemas Ltda
21 Atende.Net - WCO v:2015.04
22
23 ### Município
24 Total do Fornecedor: R$460,00
25
26 Arvoredo, 15 de agosto de 2024.
27 ...
```

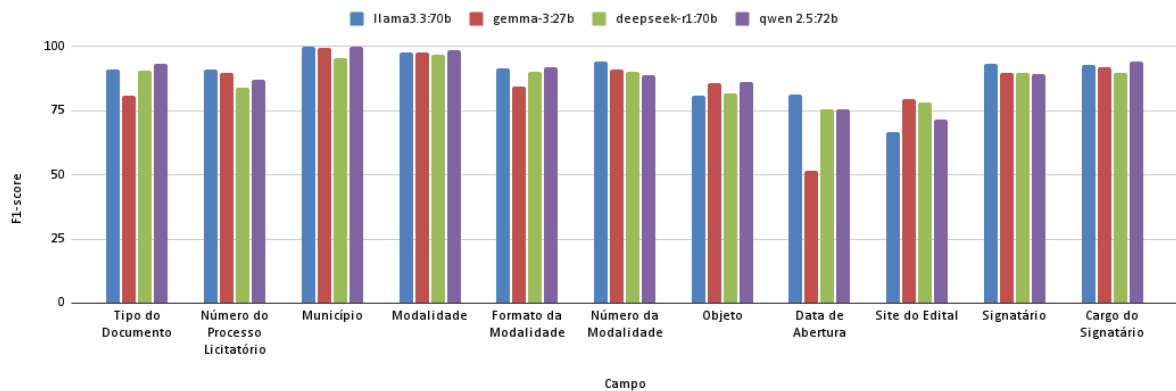
Fonte: Próprio autor (2025)

aumentam, o trecho recuperado fica mais extenso e complexo, enquanto valores menores podem omitir informações cruciais.

Essa variação no contexto impacta diretamente a qualidade da extração estruturada, resultando em degradação de performance. Conforme evidenciado na Tabela 19 e Figura 14, mesmo a melhor configuração de RAG, que foi a 26, (chunk-size: 256, overlap: 16, top-K: 3) fica significativamente abaixo da linha de base, com queda de 27,45 pontos percentuais na acurácia média e de 22,63 pontos percentuais no F1-score médio. Detalhes da melhor configuração estão na Tabela 20, onde se observa que campos com forte consistência no cenário sem RAG (por exemplo, “Município”) mantêm desempenho estável, ao passo que os outros campos sofrem maior degradação.

Em resumo, para documentos curtos como os deste estudo, a etapa adicional de recuperação não traz benefício e, ao contrário, pode prejudicar o desempenho, pois o texto original já contém o contexto essencial para a extração.

Figura 16 – Resultado do quarto experimento com diferentes LLMs



Fonte: Próprio autor (2025)

5.4 EXPERIMENTO COM DIFERENTES LLMS

Foram comparados os modelos gemma-3:27b, deepseek-r1:70b e qwen 2.5:72b com a linha de base llama3.3:70b da Tabela 17. Os resultados completos constam das Tabelas 17 a 24. A média do F1-score por modelo apresenta o llama3.3 com desempenho médio de 89,2%, seguido por qwen 2.5 com 88,8%, deepseek-r1 com 87,5% e gemma-3 com 85,6%. A proximidade entre a linha de base e o qwen 2.5 indica que um prompt cuidadosamente calibrado pode oferecer resultados comparáveis.

A avaliação campo a campo evidencia variações relevantes, conforme pode ser visto na Figura 16. O atributo Município obtém 100% de F1 tanto no llama3.3 quanto no qwen 2.5, demonstrando que todos os modelos capturam bem essa entidade. Em Modalidade, o qwen 2.5 atinge 98,85% de F1, superando o deepseek-r1 (96,73%) e o llama3.3 (97,94%). Para Formato da Modalidade, observa-se incremento gradual: gemma-3 com 84,23%, deepseek-r1 com 90,22%, llama3.3 com 91,66% e qwen 2.5 com 92,07%. O campo Data de Abertura mantém-se o mais suscetível a ruído; llama3.3 registra 81,48% de F1, gemma-3 cai para 51,61% e deepseek-r1 e qwen 2.5 alcançam 75,47%. Em Site do Edital, gemma-3 alcançou 79,37% de F1; llama3.3 obtém 66,67%, situando os demais modelos entre esses valores. Nos atributos relacionados ao signatário, o qwen 2.5 destaca-se com 94,05% em Cargo do Signatário e 89,14% em Signatário.

Tabela 21 – Comparação dos resultados do quarto experimento

Métrica	llama3.3:70b	gemma-3:27b	deepseek-r1:70b	qwen 2.5:72b
Acurácia média	91,00%	88,36%	88,90%	90,81%
F1-score médio	89,17%	85,59%	87,50%	88,76%

Fonte: Próprio autor (2025)

Dessas observações e da Tabela 21 emergem duas conclusões principais. Primeiro, o llama3.3 demonstra robustez e desempenho consistente em diversos campos, reforçando a eficácia de um prompt cuidadosamente estruturado. Segundo, cada modelo apresenta vantagens

específicas conforme o campo avaliado, de modo que nenhum domina todas as métricas.

Tabela 22 – Resultado gemma3:27b

Campo	Acurácia	Precisão	Revocação	F1-score
Tipo do Documento	86.00%	78.85%	82.83%	80.79%
Número do Processo Licitatório	89.00%	90.82%	89.00%	89.90%
Município	98.00%	100.00%	98.79%	99.39%
Modalidade	97.00%	97.14%	98.30%	97.72%
Formato da Modalidade	86.00%	89.38%	79.64%	84.23%
Número da Modalidade	90.00%	91.84%	90.00%	90.91%
Objeto	85.00%	86.73%	85.00%	85.86%
Data de Abertura	76.00%	41.03%	69.57%	51.61%
Site do Edital	89.00%	78.12%	80.65%	79.37%
Signatário	86.00%	88.64%	90.70%	89.66%
Cargo do Signatário	90.00%	96.20%	88.37%	92.12%

Fonte: Próprio autor (2025)

Tabela 23 – Resultado deepseek-r1:70b

Campo	Acurácia	Precisão	Revocação	F1-score
Tipo do Documento	89.00%	93.12%	88.05%	90.51%
Número do Processo Licitatório	84.00%	84.00%	84.00%	84.00%
Município	97.00%	95.36%	95.65%	95.51%
Modalidade	95.00%	97.93%	95.56%	96.73%
Formato da Modalidade	90.00%	87.86%	92.71%	90.22%
Número da Modalidade	90.00%	90.00%	90.00%	90.00%
Objeto	82.00%	82.00%	82.00%	82.00%
Data de Abertura	88.00%	66.67%	86.96%	75.47%
Site do Edital	86.00%	67.44%	93.55%	78.38%
Signatário	88.00%	87.78%	91.86%	89.77%
Cargo do Signatário	89.00%	91.57%	88.37%	89.94%

Fonte: Próprio autor (2025)

5.5 DISCUSSÃO DOS RESULTADOS

Os experimentos confirmam que a qualidade do prompt exerce impacto decisivo sobre a extração estruturada. A evolução do zero-shot para o few-shot de saída apenas elevou o F1-score médio de 87,29% para 89,17%, com ganhos expressivos em campos como Objeto (+5,0 p.p.) e Tipo do Documento (+4,3 p.p.). Esse avanço decorre da apresentação explícita do formato de resposta. Adicionalmente, a omissão do texto de entrada nos exemplos evitou ruído e manteve o foco na estrutura.

Em contraste, a introdução do RAG não trouxe benefícios. Mesmo o melhor arranjo de chunk size, overlap e top-k ficou abaixo da configuração few-shot pura, indicando que

Tabela 24 – Resultado qwen2.5:72b

Campo	Acurácia	Precisão	Revocação	F1-score
Tipo do Documento	92.00%	94.31%	92.19%	93.24%
Número do Processo Licitatório	87.00%	87.00%	87.00%	87.00%
Município	100.00%	100.00%	100.00%	100.00%
Modalidade	98.00%	99.17%	98.53%	98.85%
Formato da Modalidade	93.00%	89.66%	94.63%	92.07%
Número da Modalidade	89.00%	89.00%	89.00%	89.00%
Objeto	86.00%	86.00%	86.00%	86.00%
Data de Abertura	89.00%	66.67%	86.96%	75.47%
Site do Edital	86.00%	66.67%	77.42%	71.64%
Signatário	87.00%	87.64%	90.70%	89.14%
Cargo do Signatário	92.00%	96.34%	91.86%	94.05%

Fonte: Próprio autor (2025)

a recuperação de passagens similares não acrescentou contexto relevante para documentos relativamente curtos. Na prática, o mecanismo de RAG aumentou a complexidade do contexto, sem proporcionar sinal adicional suficiente para melhorar as previsões; nos casos em que chunks ou top-k foram reduzidos, observou-se perda de partes essenciais do texto.

Por fim, a comparação entre as diferentes LLMs llama3.3 (linha de base), gemma-3, deepseek-r1 e qwen 2.5 revela desempenho amplamente semelhante. A variação do F1-score médio permaneceu dentro de 3,6 p.p., com o llama3.3 registrando 89,17% e o gemma-3, 85,59 %. Embora haja flutuações pontuais por campo (por exemplo, o qwen 2.5 atinge 98,85% em Modalidade, enquanto o gemma-3 lidera em Site do Edital com 79,37%), nenhum modelo supera os demais de forma consistente. Assim, a escolha da arquitetura pode priorizar fatores como tempo de inferência ou custo computacional, pois a simples adoção de outro modelo não garante saltos significativos de qualidade.

Em síntese, o estudo mostra que a engenharia de prompt e a escolha criteriosa de exemplos oferecem ganhos mais substanciais do que o uso de RAG ou a simples troca de LLMs, sobretudo quando os modelos já apresentam desempenhos similares. A seguir, a próxima subseção aprofunda a análise dos principais erros identificados.

5.5.1 Análise dos Erros

A análise qualitativa dos resultados revela padrões de falha nos campos avaliados. No Objeto, o erro mais frequente consiste no truncamento da descrição, que leva o modelo a devolver apenas a porção inicial do texto, reduzindo o F1-score mesmo quando a primeira metade está correta. Os campos Data de Abertura e Site do Edital apresentam forte instabilidade: a amostra de exemplos válidos é pequena e qualquer cadeia numérica é, por vezes, interpretada como data real, enquanto endereços parciais ou abreviados confundem a detecção de URLs, provocando oscilações superiores a 10 p.p. entre execuções. Já os identificadores numéricos, Número do

Processo Licitatório e Número da Modalidade, são propensos a trocas mútuas, pois compartilham o padrão “123/2025”.

Quanto aos campos de múltiplas classes, o Tipo do Documento tem troca entre rótulos semelhantes, como “Homologação” e “Adjudicação”. A Modalidade preserva alta acurácia, mas ainda sofre substituição esporádica entre termos próximos, como “Pregão” e “Tomada de Preços”. No campo Formato da Modalidade, os mesmos tipos de erros acontecem quando se troca “Presencial” por “Eletrônico”. Por fim, nos atributos de assinatura, ocorrem três tipos de deslize: o nome do Signatário pode vir duplicado ou com sobrenomes truncados; o Cargo do Signatário às vezes incorpora o nome da pessoa ou, inversamente, carece de parte do título; e ambos podem ser trocados quando o documento lista mais de um responsável.

Em resumo, tratam-se de falhas pontuais que podem ser resolvidas com mais exemplos anotados, validações, e a própria evolução das LLM; trabalhos futuros que combinem essas técnicas devem reduzir ou eliminar esses erros residuais.

6 CONCLUSÃO

Este trabalho demonstrou que LLM, quando combinados a engenharia de prompts, saídas estruturadas, e exemplos de saída, são capazes de extrair informações com elevado grau de precisão a partir de documentos de licitações municipais catarinenses. Os resultados atingiram F1-score médio de quase 90%, atingindo resultados alinhados ao que se observa na literatura da área, superando desafios recorrentes do processamento de linguagem natural em documentos administrativos e mostrando que soluções baseadas em LLM podem automatizar tarefas que tradicionalmente exigiam revisão humana especializada.

Entre as principais descobertas, destaca-se que a adição de exemplos no prompt (few-shot) elevou consistentemente os índices de acurácia e F1-score, especialmente para campos como Objeto e Tipo de Documento. O uso de diferentes modelos de LLM open-source não alterou substancialmente o desempenho global, sugerindo que a engenharia de prompt exerce impacto mais decisivo que a troca de modelo. Por outro lado, a adoção de RAG não apenas deixou de agregar valor em documentos curtos, como em muitos casos reduziu a performance, indicando que a recuperação de contexto deve ser usada com cautela e apenas em cenários em que o tamanho e a dispersão do conteúdo justifiquem a complexidade adicional.

Apesar dos avanços, o trabalho apresenta limitações relevantes. O conjunto de dados experimental limitou-se a 100 documentos, o que restringe a generalização estatística dos achados. Certos campos avaliados, como Data de Abertura e Site do Edital, estiveram sub-representados, impactando a robustez das métricas nesses casos. Persistiram erros típicos de truncamento de informações, confusão entre campos numéricos, dificuldades com casos de múltiplos signatários e ainda não se tratou de documentos que excedem o limite de 32 mil caracteres. A extração de valores compostos e documentos acima de 32 mil caracteres permanecem desafios abertos.

Para avançar no estado da arte, recomenda-se que estudos futuros ampliem o corpus de avaliação para abranger maior diversidade documental e aumentem a representatividade de todos os campos. A utilização adaptativa de RAG, aplicada apenas a documentos realmente extensos ou fragmentados, também merece investigação.

Conclui-se que, dentro do escopo proposto, a automação baseada em LLMs para extração estruturada em licitações públicas é não apenas viável, mas representa um avanço significativo em direção a processos mais transparentes, eficientes e auditáveis no setor público brasileiro. Com o amadurecimento das técnicas e expansão do corpus, espera-se que soluções desse tipo se consolidem como padrão em iniciativas de governo digital, fiscalização e controle social.

REFERÊNCIAS

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.

BRASIL. CONSTITUIÇÃO DA REPÚBLICA FEDERATIVA DO BRASIL DE 1988.

Brasília, DF: Presidência da República, 5 out. 1988. Disponível em:

https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm.

Acesso em: 18 mai. 2025.

_____. **LEI Nº 12.527, DE 18 DE NOVEMBRO DE 2011.** Brasília, DF: Presidência da República, 18 nov. 2011. Regula o acesso a informações previsto no inciso XXXIII do art. 5º, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal; altera a Lei nº 8.112, de 11 de dezembro de 1990; revoga a Lei nº 11.111, de 5 de maio de 2005, e dispositivos da Lei nº 8.159, de 8 de janeiro de 1991; e dá outras providências. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm.

Acesso em: 18 mai. 2025.

_____. **LEI Nº 12.527: LEI DE ACESSO À INFORMAÇÃO Poder Executivo Federal 2011-2012.** Brasília, DF: Controladoria-Geral da União, mai. 2013. Disponível em:

<https://www.gov.br/acessoainformacao/pt-br/central-de-conteudo/publicacoes/arquivos/balanco1ano.pdf>. Acesso em: 22 mai. 2025.

_____. **LEI Nº 14.133, DE 1º DE ABRIL DE 2021.** Brasília, DF: Presidência da República, 1 abr. 2021. Lei de Licitações e Contratos Administrativos. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/l14133.htm.

Acesso em: 18 mai. 2025.

_____. **LEI Nº 8.666, DE 21 DE JUNHO DE 1993.** Brasília, DF: Presidência da República, 21 jun. 1993. Regulamenta o art. 37, inciso XXI, da Constituição Federal, institui normas para licitações e contratos da Administração Pública e dá outras providências.

Disponível em: https://www.planalto.gov.br/ccivil_03/leis/l8666cons.htm.

Acesso em: 18 mai. 2025.

_____. **Licitações e contratações.** Brasília, DF: Controladoria-Geral da União.

Disponível em: <https://portal.datransparencia.gov.br/entenda-a-gestao-publica/licitacoes-e-contratacoes>. Acesso em: 22 mai. 2025.

CIGA. Dados. Florianópolis, SC: Consórcio de Inovação na Gestão Pública. Disponível em: <https://dados.ciga.sc.gov.br/about>. Acesso em: 18 mai. 2025.

_____. **Quem Somos.** Florianópolis, SC: Consórcio de Inovação na Gestão Pública.

Disponível em: <https://consorciociga.gov.br/quem-somos/>. Acesso em: 18 mai. 2025.

CIGA. **Recorde**. Florianópolis, SC: Consórcio de Inovação na Gestão Pública, 7 jun. 2023. Disponível em: <https://consorciociga.gov.br/ciga-celebra-recorde-de-autopublicacoes-no-diario-oficial-dos-municipios-de-santa-catarina/>. Acesso em: 18 mai. 2025.

DEEPSEEK-AI *et al.* **DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning**. [S.l.: s.n.], 2025. arXiv: 2501.12948 [cs.CL]. Disponível em: <https://arxiv.org/abs/2501.12948>.

FARIA, Joana Ribeiro de; XIE, Huiyuan; STEFFEK, Felix. **Automatic Information Extraction From Employment Tribunal Judgements Using Large Language Models**. [S.l.: s.n.], 2024. arXiv: 2403.12936 [cs.CL]. Disponível em: <https://arxiv.org/abs/2403.12936>.

GAO, Yunfan *et al.* **Retrieval-Augmented Generation for Large Language Models: A Survey**. [S.l.: s.n.], 2024. arXiv: 2312.10997 [cs.CL]. Disponível em: <https://arxiv.org/abs/2312.10997>.

GGML-ORG. **llama.cpp - GBNF Grammars Guide**. [S.l.: s.n.], 2025. <https://github.com/ggml-org/llama.cpp/blob/master/grammars/README.md>. Acesso em: 18 mai. 2025.

GHOJOGH, Benyamin; GHODSI, Ali. Attention Mechanism, Transformers, BERT, and GPT: Tutorial and Survey. working paper or preprint. [S.l.], dez. 2020. Disponível em: <https://hal.science/hal-04637647>.

GRATTAFIORI, Aaron *et al.* **The Llama 3 Herd of Models**. [S.l.: s.n.], 2024. arXiv: 2407.21783 [cs.AI]. Disponível em: <https://arxiv.org/abs/2407.21783>.

GUIMARÃES, Gabriel M.C. *et al.* DODFMiner: An automated tool for Named Entity Recognition from Official Gazettes. **Neurocomputing**, v. 568, p. 127064, 2024. ISSN 0925-2312. DOI: <https://doi.org/10.1016/j.neucom.2023.127064>. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0925231223011876>.

HAMDI, Ahmed *et al.* Information extraction from invoices. In: SPRINGER. INTERNATIONAL Conference on Document Analysis and Recognition. [S.l.: s.n.], 2021. P. 699–714.

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 (ACM), v. 43, n. 2, p. 1–55, jan. 2025. ISSN 1558-2868. DOI: 10.1145/3703155. Disponível em: <http://dx.doi.org/10.1145/3703155>.

HUSSAIN, Atin Sakkeer; THOMAS, Anu. **Large Language Models for Judicial Entity Extraction: A Comparative Study**. [S.l.: s.n.], 2024. arXiv: 2407.05786 [cs.CL]. Disponível em: <https://arxiv.org/abs/2407.05786>.

JSON-SCHEMA-ORG. **What is JSON Schema?** [S.l.: s.n.], 2025.

<https://json-schema.org/overview/what-is-jsonschema>. Acesso em: 18 mai. 2025.

KANDPAL, Nikhil *et al.* **Large Language Models Struggle to Learn Long-Tail Knowledge.**

[S.l.: s.n.], 2023. arXiv: 2211.08411 [cs.CL]. Disponível em:

<https://arxiv.org/abs/2211.08411>.

KATZ, Daniel Martin *et al.* **Natural Language Processing in the Legal Domain.** [S.l.: s.n.],

2023. arXiv: 2302.12039 [cs.CL]. Disponível em:

<https://arxiv.org/abs/2302.12039>.

KESHAVARZ, Hossein *et al.* Named Entity Recognition in Long Documents: An End-to-end Case Study in the Legal Domain. *In: 2022 IEEE International Conference on Big Data (Big Data)*. Osaka, Japan: IEEE, 2022. DOI: 10.1109/BigData55660.2022.10020873.

Disponível em: <https://doi.org/10.1109/BigData55660.2022.10020873>.

KOWSRIHAWAT, Kankawin; VATEEKUL, Peerapon. An information extraction framework for legal documents: A case study of thai supreme court verdicts. *In: IEEE. 2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE)*. [S.l.: s.n.], 2015. P. 275–280.

LIU, Michael Xieyang *et al.* “We Need Structured Output”: Towards User-centered Constraints on Large Language Model Output. *In: EXTENDED Abstracts of the CHI Conference on Human Factors in Computing Systems*. [S.l.]: ACM, mai. 2024. (CHI ’24), p. 1–9. DOI: 10.1145/3613905.3650756. Disponível em:

<http://dx.doi.org/10.1145/3613905.3650756>.

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>.

LIU, Yu *et al.* Are LLMs good at structured outputs? A benchmark for evaluating structured output capabilities in LLMs. **Information Processing Management**, v. 61, n. 5, p. 103809, 2024. ISSN 0306-4573. DOI: <https://doi.org/10.1016/j.ipm.2024.103809>.

Disponível em:

<https://www.sciencedirect.com/science/article/pii/S0306457324001687>.

META. **Llama 3.3 Model Card.** [S.l.: s.n.], 2025. Disponível em:

https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/.

Acesso em: 18 mai. 2025.

MINAEE, Shervin *et al.* **Large Language Models: A Survey.** arXiv:2402.06196 [cs.CL]. 2024. arXiv: 2402.06196 [cs.CL]. Disponível em:

<https://doi.org/10.48550/arXiv.2402.06196>.

MOENS, MARIE-FRANCINE; UYTENDAELE, CAROLINE; DUMORTIER, JOS.

Information extraction from legal texts: the potential of discourse analysis. **International**

Journal of Human-Computer Studies, v. 51, n. 6, p. 1155–1171, 1999. ISSN 1071-5819. DOI: <https://doi.org/10.1006/ijhc.1999.0296>. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1071581999902963>.

NADEAU, David; SEKINE, Satoshi. A survey of named entity recognition and classification. **Lingvisticae Investigationes**, v. 30, n. 1, p. 3–26, 2007. ISSN 0378-4169. DOI: 10.1075/li.30.1.03nad. Disponível em: <https://doi.org/10.1075/li.30.1.03nad>.

OLIVEIRA, Murillo Preve Cardoso de. **Quais são as modalidades de licitação?** São Paulo, SP: Schiefler Advocacia, 2021. Disponível em: <https://schiefler.adv.br/modalidades-de-licitacao/>. Acesso em: 22 mai. 2025.

OLLAMA. **Ollama**. Runtime open-source para executar modelos de linguagem localmente. 2025. Disponível em: <https://ollama.com/>. Acesso em: 7 jul. 2025.

PEREIRA, Jayr *et al.* **INACIA: Integrating Large Language Models in Brazilian Audit Courts: Opportunities and Challenges**. [S.l.: s.n.], 2024. arXiv: 2401.05273 [cs.CL]. Disponível em: <https://arxiv.org/abs/2401.05273>.

POKRASS, Michelle *et al.* **Introducing Structured Outputs in the API**. [S.l.: s.n.], ago. 2024. Accessed: 2025-05-25. Disponível em: <https://openai.com/index/introducing-structured-outputs-in-the-api/>.

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>.

RUSINOL, Marçal; BENKHELFALLAH, Tayeb; POULAIN DANDECY, Vincent. Field extraction from administrative documents by incremental structural templates. *In: IEEE. 2013 12th International Conference on Document Analysis and Recognition*. [S.l.: s.n.], 2013. P. 1100–1104.

SCHULHOFF, Sander *et al.* **The Prompt Report: A Systematic Survey of Prompt Engineering Techniques**. [S.l.: s.n.], 2025. arXiv: 2406.06608 [cs.CL]. Disponível em: <https://arxiv.org/abs/2406.06608>.

SLEIMI, Amin *et al.* Automated Extraction of Semantic Legal Metadata using Natural Language Processing. *In: 2018 IEEE 26th International Requirements Engineering Conference (RE)*. [S.l.: s.n.], 2018. P. 124–135. DOI: 10.1109/RE.2018.00022.

SOUZA, Fábio; NOGUEIRA, Rodrigo; LOTUFO, Roberto. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. *In: INTELLIGENT Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I*. Rio Grande, Brazil: Springer-Verlag, 2020. P. 403–417. DOI: 10.1007/978-3-030-61377-8_28. Disponível em: https://doi.org/10.1007/978-3-030-61377-8_28.

TEAM, Gemma *et al.* **Gemma 3 Technical Report**. [S.l.: s.n.], 2025. arXiv: 2503.19786 [cs.CL]. Disponível em: <https://arxiv.org/abs/2503.19786>.

VASWANI, Ashish *et al.* **Attention Is All You Need**. arXiv:1706.03762 [cs.CL]. 2017. arXiv: 1706.03762 [cs.CL]. Disponível em: <https://doi.org/10.48550/arXiv.1706.03762>.

WANG, Shuhe *et al.* GPT-NER: Named Entity Recognition via Large Language Models. **arXiv preprint arXiv:2304.10428**, v. v4, 2023. Last revised 7 Oct 2023. arXiv: 2304.10428 [cs.CL]. Disponível em: <https://doi.org/10.48550/arXiv.2304.10428>.

WU, Shangyu *et al.* **Retrieval-Augmented Generation for Natural Language Processing: A Survey**. [S.l.: s.n.], 2025. arXiv: 2407.13193 [cs.CL]. Disponível em: <https://arxiv.org/abs/2407.13193>.

APÊNDICE A – MELHOR PROMPT

Você é um especialista em licitação brasileira.

Extraia as seguintes entidades do documento de licitação no # Contexto

- Tipo do documento [enum] ^(Apostilamento|Anulação|Ata de Registro de Preços|Ata de Recebimento e Abertura|Adjudicação|Aviso de Licitação|Aviso de Suspensão|Aviso de Cancelamento|Contrato|Edital|Errata|Homologação|Aditivo|Julgamento|Resultado|Ratificação)\$
- Número do processo administrativo [string] ^\d+/\d+\$
- Município [enum] ^(Abdon Batista|Abelardo Luz|Agrolândia|Agronômica|Água Doce|Águas de Chapecó|Águas Frias|Águas Mornas|Alfredo Wagner|Alto Bela Vista|Anchieta|Angelina|Anita Garibaldi|Anitápolis|Antônio Carlos|Apiúna|Araranguá|Araquari|Araranguá|Armazém|Arroio Trinta|Arvoredo|Ascurra|Atalanta|Aurora|Balneário Arroio do Silva|Balneário Barra do Sul|Balneário Camboriú|Balneário Gaivota|Balneário Piçarras|Balneário Rincão|Bandeirante|Barra Bonita|Barra Velha|Bela Vista do Toldo|Belmonte|Benedito Novo|Biguaçu|Blumenau|Bocaina do Sul|Bom Jardim da Serra|Bom Jesus|Bom Jesus do Oeste|Bom Retiro|Bombinhas|Botuverá|Braço do Norte|Braço do Trombudo|Brunópolis|Brusque|Caçador|Caibi|Calmon|Camboriú|Campo Alegre|Campo Belo do Sul|Campo Erê|Campos Novos|Canelinha|Canoinhas|Capão Alto|Capinzal|Capivari de Baixo|Catanduvas|Caxambu do Sul|Celso Ramos|Cerro Negro|Chapadão do Lageado|Chapecó|Cocal do Sul|Concórdia|Cordilheira Alta|Coronel Freitas|Coronel Martins|Correia Pinto|Corupá|Criciúma|Cunha Porã|Cunhataí|Curitibanos|Descanso|Dionísio Cerqueira|Dona Emma|Doutor Pedrinho|Entre Rios|Ermo|Erval Velho|Faxinal dos Guedes|Flor do Sertão|Florianópolis|Formosa do Sul|Forquilha|Fraiburgo|Frei Rogério|Galvão|Garopaba|Garuva|Gaspar|Governador Celso Ramos|Grão-Pará|Gravatal|Guabiruba|Guaraciaba|Guaramirim|Guarujá do Sul|Guatambu|Herval do Oeste|Ibiam|Ibicaré|Ibirama|Içara|Ilhota|Imaruí|Imbituba|Imbuia|Indaial|Iomerê|Ipira|Iporã do Oeste|Ipuçu|Ipumirim|Iraceminha|Irani|Irati|Irineópolis|Itá|Itaiópolis|Itajaí|Itapema|Itapiranga|Itapoá|Ituporanga|Jaborá|Jacinto Machado|Jaguaruna|Jaraguá do Sul|Jardinópolis|Joaçaba|Joinville|José Boiteux|Jupiá|Lacerdópolis|Lages|Laguna|Lajeado Grande|Laurentino|Lauro Müller|Lebon Régis|Leoberto Leal|Lindóia do Sul|Lontras|Luiz Alves|Luzerna|Macieira|Mafra|Major Gercino|Major Vieira|Maracajá|Maravilha|Marema|Massaranduba|Matos Costa|Meleiro|Mirim Doce|Modelo|Mondaí|Monte Carlo|Monte Castelo|Morro da Fumaça|Morro Grande|Navegantes|Nova Erechim|Nova Itaberaba|Nova Trento|Nova Veneza|Novo Horizonte|Orleans|Otacílio Costa|Ouro|Ouro Verde|Paial|Painel|Palhoça|Palma Sola|Palmeira|Palmitos|Papanduva|Paraíso|Passo de Torres|Passos Maia|Paulo Lopes|Pedras Grandes|Penha|Peritiba|Pescaria Brava|Petrolândia|Pinhalzinho|Pinheiro Preto|

Piratuba|Planalto Alegre|Pomerode|Ponte Alta|Ponte Alta do Norte|Ponte Serrada|Porto Belo|Porto União|Pouso Redondo|Praia Grande|Presidente Castello Branco|Presidente Getúlio|Presidente Nereu|Princesa|Quilombo|Rancho Queimado|Rio das Antas|Rio do Campo|Rio do Oeste|Rio do Sul|Rio dos Cedros|Rio Fortuna|Rio Negrinho|Rio Rufino|Riqueza|Rodeio|Romelândia|Salete|Saltinho|Salto Veloso|Sangão|Santa Cecília|Santa Helena|Santa Rosa de Lima|Santa Rosa do Sul|Santa Terezinha|Santa Terezinha do Progresso|Santiago do Sul|Santo Amaro da Imperatriz|São Bento do Sul|São Bernardino|São Bonifácio|São Carlos|São Cristóvão do Sul|São Domingos|São Francisco do Sul|São João Batista|São João do Itaperiú|São João do Oeste|São João do Sul|São Joaquim|São José|São José do Cedro|São José do Cerrito|São Lourenço do Oeste|São Ludgero|São Martinho|São Miguel da Boa Vista|São Miguel do Oeste|São Pedro de Alcântara|Saudades|Schroeder|Seara|Serra Alta|Siderópolis|Sombrio|Sul Brasil|Taió|Tangará|Tigrinhos|Tijucas|Timbé do Sul|Timbó|Timbó Grande|Três Barras|Treviso|Treze de Maio|Treze Tílias|Trombudo Central|Tubarão|Tunápolis|Turvo|União do Oeste|Urubici|Urupema|Urussanga|Vargeão|Vargem|Vargem Bonita|Vidal Ramos|Videira|Vitor Meireles|Witmarsum|Xanxerê|Xavantina|Xaxim|Zortéa)\$

- Modalidade [enum] ^(Concorrência|Concurso|Convite|Credenciamento|Diálogo Competitivo|Dispensa de Licitação|Inexigibilidade|Leilão|Regime Diferenciado de Contratações|Pregão|Tomada de Preços)\$
- Formato da modalidade [enum|null] ^(Presencial|Eletrônico)\$
- Número da modalidade [string] ^\d+/\d+\$
- Objeto [string] ^.+ \$
- Data de abertura [string|null] \d{4}-\d{2}-\d{2}T\d{2}:\d{2}:\d{2}\$
- Site do edital [string|null] ^(https?:\\\/|www\.)[^s]+\$
- Nome do Signatário [string|null] ^.+ \$
- Cargo do signatário [string|null] ^.+ \$

Regras

- Sem inferência: extraia apenas informações explícitas. Não deduza, interprete ou complete.
- Extração exata: mantenha o texto como está, sem alterar ortografia, pontuação ou capitalização.
- Campos opcionais: se a informação não estiver clara ou presente, retorne null.
- Formatos obrigatórios: respeite os formatos definidos.

Retorne no formato JSON.

Exemplos

Saída

```
{
  "tipo_do_documento": "Homologação",
  "numero_do_processo_licitatório": "3706/2023",
  "município": "Canoinhas",
  "modalidade": "Pregão",
  "formato_da_modalidade": "Eletrônico",
  "numero_da_modalidade": "89/2023",
  "objeto": "MURO PRÉ- FABRICADO M PROPRIA",
  "data_de_abertura": null,
  "site_do_edital": null,
  "signatário": "RAFAEL ROTTILI ROEDER",
  "cargo_do_signatário": "Secretário de Planejamento"
}
```

Saída

```
{
  "tipo_do_documento": "Aviso de Licitação",
  "numero_do_processo_licitatório": "6/2023",
  "município": "Navegantes",
  "modalidade": "Pregão",
  "formato_da_modalidade": "Eletrônico",
  "numero_da_modalidade": "6/2023",
  "objeto": "Pregão Eletrônico para Registro de Preço visando a contratação de empresa especializada em prestação de serviço na manutenção corretiva, preventiva e estética na área mecânica, funilaria, pintura(c/ adesivação), elétrica(eletrônica), tapeçaria e borracharia de veículos automotores com fornecimento e substituição de peças, materiais e acessórios em estado novo, originais dos fabricantes das marcas dos veículos e com desconto de 5% (cinco por cento) sobre os preços das tabelas e catálogos dos fabricantes ou revendas autorizadas, para atender a manutenção da frota das Secretarias, Fundos, Fundações, Polícia Militar e Corpo de Bombeiros Militar de Navegantes",
  "data_de_abertura": "2023-05-31T14:00",
  "site_do_edital": "www.navegantes.sc.gov.br",
  "signatário": "Libardoni Fronza",
  "cargo_do_signatário": "Prefeito"
}
```

Saída

```
{
  "nome_do_documento": "Dispensa de Licitação",
  "numero_do_processo_licitatório": "55/2024",
```

```
"município": "Ipumirim",  
"modalidade": "Dispensa de Licitação",  
"formato_da_modalidade": null,  
"número_da_modalidade": "24/2024",  
"objeto": "Aquisição de sabonete líquido neutro para atender as  
necessidades da Secretaria de Educação, Cultura e Esportes."  
"data_de_abertura": null,  
"site_do_edital": null,  
"signatário": null,  
"cargo_do_signatário": null  
}
```

Contexto

{CONTEXT0}

APÊNDICE B – ARTIGO DO TCC

Extração de Dados a Partir de Publicações do Diário Oficial dos Municípios de Santa Catarina usando Large Language Models

Felipe Valentin Nascimento¹, Carina Friedrich Dorneles²

¹ Departamento de Informática e Estatística (INE)
Universidade Federal de Santa Catarina (UFSC)
Florianópolis – SC – Brasil

Resumo. A extração automática de dados estruturados é essencial para tornar documentos úteis em análises computacionais e para apoiar políticas públicas baseadas em evidências. No Diário Oficial dos Municípios de Santa Catarina (DOM/SC), embora parte das licitações já seja disponibilizada em JSON, o campo de texto livre concentra as informações mais relevantes e permanece pouco explorado. Este estudo investiga o emprego de Large Language Models (LLMs) para extrair esses dados. Após inspecionar mais de 110 000 publicações, construímos um conjunto ouro de 100 documentos com 11 variáveis anotadas (6 obrigatórias e 5 opcionais). O fluxo experimental engloba pré-processamento, engenharia de prompts, uso opcional de Retrieval-Augmented Generation (RAG) e validação automática por JSON Schema. O prompt final foi testado em quatro modelos abertos de última geração. A combinação de few-shot com saída JSON válido zerou erros de formatação e atingiu F1 médio de 89,2 %, com melhor desempenho do Llama-3.3-70B. Concluímos que, para textos curtos, LLMs com prompts bem elaborados dispensam RAG e podem fortalecer a transparência e a fiscalização das contratações públicas.

Abstract. Extracting structured data is crucial for computational analysis and evidence-based policymaking. In the Diário Oficial dos Municípios de Santa Catarina (DOM/SC), most key information still resides in an unstructured free-text field. This work explores the use of Large Language Models (LLMs) to convert that text into structured data. After scanning over 110 000 procurement notices, we built a 100-document gold corpus annotated with 11 variables (6 mandatory, 5 optional). The experimental pipeline comprises preprocessing, prompt engineering, optional Retrieval-Augmented Generation (RAG), and automatic validation through a JSON Schema. The final prompt was evaluated on four state-of-the-art open models. Few-shot prompting with enforced JSON output eliminated formatting errors and yielded an average F1-score of 89.2 %, led by Llama-3.3-70B. Our results show that well-designed prompts allow LLMs to reliably extract entities from short procurement documents without RAG, thereby enhancing transparency and oversight in public contracting.

1. Introdução

A transparência nas contratações públicas é um pilar para a fiscalização do processo licitatório por cidadãos e órgãos de controle [Brasil 1988]. Contudo, o volume crescente

de publicações diárias, como as do Diário Oficial dos Municípios de Santa Catarina (DOM/SC), torna a verificação manual impraticável [CIGA 2025]. Um desafio técnico reside no fato de que informações cruciais estão em campos textuais pouco estruturados, dificultando a automação.

A área de Extração de Informações tem evoluído de soluções baseadas em regras e expressões regulares para modelos estatísticos e, mais recentemente, para arquiteturas de aprendizado de máquina como o BERT [Hamdi et al. 2021]. Os Large Language Models (LLMs) emergiram como uma alternativa flexível, embora apresentem desafios como a geração de alucinações e a dificuldade com domínios especializados [Katz et al. 2023]. Técnicas como Retrieval-Augmented Generation (RAG) [Gao et al. 2024] e o uso de saídas estruturadas com JSON Schema visam mitigar esses problemas.

Em Santa Catarina, as publicações do DOM/SC, mantido pelo Consórcio de Inovação na Gestão Pública (CIGA), são disponibilizadas em formato JSON. No entanto, o campo "texto" armazena o conteúdo integral em HTML, de onde as informações relevantes precisam ser extraídas (ver figura 1). Este trabalho investiga o uso de LLMs de código aberto para extrair, de forma confiável, dados essenciais de extratos de licitações do DOM/SC. Testamos sistematicamente quatro modelos para recuperar 11 variáveis em um conjunto de 100 publicações anotadas manualmente.

```
{
  "codigo": "6320072",
  "titulo": "PROCESSO ADMINISTRATIVO 248/2024",
  "data": "2024-08-15 16:10:00",
  "cod_registro_info_sfinge": "A5BBAA77FB6B4921B634BD6033504887BC00DEFB",
  "municipio": "Arvoredo",
  "entidade": "Prefeitura Municipal de Arvoredo",
  "categoria": "Licitações",
  "link": "https://www.diariomunicipal.sc.gov.br/?q=id:6320072",
  "texto": "<html xmlns='http://www.w3.org/1999/xhtml'>\n\n  <head>\n\n    <meta name='pdf:PDFVersion' conter",
  "url": "https://www.diariomunicipal.sc.gov.br/arquivosbd/atos/2024/08/1723748994_homologao_extrato.pdf"
},
```

Figura 1. Estrutura JSON de uma publicação do DOM/SC, evidenciando o campo "texto" com o HTML completo da licitação

O objetivo geral é aplicar LLMs para a extração de dados de extratos publicados no DOM/SC. Os objetivos específicos incluem: analisar a estrutura dos documentos, definir as variáveis a serem extraídas, selecionar e avaliar modelos de LLMs, validar a acurácia da extração e comparar diferentes abordagens, incluindo engenharia de prompt e o uso de RAG.

2. Fundamentação Teórica

Esta seção resume os conceitos que embasam a extração automática de entidades em licitações catarinenses, mantendo apenas o essencial para o limite de páginas do SBC.

2.1. Transparência e Licitações Públicas

A Constituição de 1988 consagrou a publicidade como princípio da Administração, reforçado pela Lei nº 8.666/1993 (revogada pela Lei nº 14.133/2021) e pela Lei de Acesso à Informação [Brasil 1988, Brasil 2011, Brasil 2021]. Em Santa Catarina, o

CIGA mantém o *Diário Oficial dos Municípios* (DOM/SC), que registrou mais de 13 000 publicações em um único dia [CIGA 2025]. O volume inviabiliza auditoria manual e motiva soluções de PLN.

2.2. Named Entity Recognition (NER)

NER identifica e classifica menções (datas, locais, valores etc.) em textos [Nadeau and Sekine 2007]. Abordagens baseadas em *regex* evoluíram para modelos de aprendizado de máquina; porém, domínios especializados exigem dados anotados, elevando custos [Keshavarz et al. 2022]. Experimentos recentes com LLM reformulam NER como geração condicional, alcançando bons resultados mesmo com poucos exemplos [Wang et al. 2023].

2.3. Large Language Models

Desde o *Transformer* [Vaswani et al. 2017], LLM escalam para bilhões de parâmetros, exibindo raciocínio e geração de texto de alta qualidade [Minaee et al. 2024]. Este trabalho avalia quatro modelos de código aberto disponíveis no Ollama, listados na Tabela 1. Todos suportam janelas de 128 k tokens, permitindo processar trechos integrais de extratos licitatórios.

Modelo	Parâmetros	Contexto
LLaMA 3.3	70B	128 k
Qwen 2.5	72B	128 k
DeepSeek R1	70B	128 k
Gemma 3	27B	128 k

Tabela 1. Modelos avaliados

2.4. Saídas Estruturadas e Engenharia de Prompt

Forçar a saída em JSON, via JSON Schema e gramáticas GBNF [ggml org 2025] garante validade e elimina pós-processamento. Técnicas como *Role*, *Zero/One/Few-Shot* e *Chain-of-Thought* permitem adaptar o comportamento do modelo [Schulhoff et al. 2025]. Na prática, *few-shot* com exemplos de saída foi a estratégia mais eficaz neste estudo.

2.5. Retrieval-Augmented Generation

O RAG combina busca vetorial e geração condicionada para mitigar alucinações [Gao et al. 2024].

3. Trabalhos Relacionados

A extração de informações em documentos legais evoluiu significativamente. Abordagens iniciais utilizavam regras manuais e NLP, como o trabalho de [Sleimi et al. 2018], que alcançou alta precisão (97,2%) na análise de leis de trânsito, mas com a limitação da escalabilidade. O DODFMiner [Guimarães et al. 2024] automatizou a extração em Diários Oficiais do Distrito Federal com expressões regulares e aprendizado de máquina, atingindo 85% de F1-score, porém limitado ao escopo desenvolvido.

Com o advento dos LLMs, surgiram novas abordagens. [de Faria et al. 2024] usaram GPT-4 para extrair dados de decisões judiciais com acurácia superior a 97%, embora

tenham observado omissões e alucinações. [Hussain and Thomas 2024] avaliaram diversos LLMs na rotulagem de entidades jurídicas, com o modelo Mistral atingindo 64% de F1-score, indicando a necessidade de arquiteturas mais robustas.

A técnica de RAG foi explorada para enriquecer o contexto. O sistema INACIA [Pereira et al. 2024] combinou RAG e GPT-4 para processos do TCU, alcançando um F1-score de 43%, com a complexidade da configuração do RAG sendo uma limitação. [Aquino et al. 2024] também usaram RAG em documentos de licitação do MPSC, atingindo 90% de acurácia na extração de duas variáveis, mas com um escopo limitado de dados.

Este trabalho se diferencia por focar em modelos de LLM abertos para extrair dados de documentos de licitação relativamente curtos do DOM/SC. Demonstramos que uma engenharia de prompt estruturada, combinada com a validação de saída via JSON Schema, alcança alto desempenho sem a necessidade de RAG, que em nossos testes chegou a degradar os resultados. Esta abordagem elimina alucinações de rótulos e simplifica a integração dos dados.

4. Fluxo para Extração de Dados

O desenvolvimento do trabalho foi estruturado em três etapas: exploração dos dados, construção de um fluxo de experimentos e iterações para refinar os resultados, conforme ilustrado na Figura 2.

4.1. Exploração e Preparação dos Dados

Os dados são provenientes das autopublicações de licitações do DOM/SC, disponibilizadas pelo CIGA. De um total de 122.532 documentos, foram selecionados 110.278 da categoria "Licitação" com até 32 mil caracteres, o que representa cerca de 90% do acervo. A partir de uma análise inicial, foram definidas 11 variáveis para extração, sendo 6 obrigatórias (tipo do documento, número do processo licitatório, município, modalidade, número da modalidade e objeto) e 5 opcionais (formato da modalidade, data de abertura, site do edital, signatário e cargo do signatário).

Um conjunto de dados ouro com 100 publicações foi anotado manualmente para servir de base de referência para a avaliação dos modelos. Este conjunto de dados possui uma ampla cobertura geográfica com 61 municípios e diversas modalidades de licitação, com predomínio de Pregão (48%), Dispensa de Licitação (23%) e Inexigibilidade (19%).

4.2. Pipeline Experimental

Foi implementado um fluxo de extração utilizando o servidor Ollama com os modelos de LLM. O processo envolve as seguintes etapas:

1. **Pré-processamento:** O texto bruto de cada documento é pré-processado, com a remoção opcional de tags HTML e a aplicação de estratégias como *one-shot* ou *few-shot*.
2. **Construção do Prompt:** O prompt é construído dinamicamente, incorporando o esquema JSON da saída, o contexto (texto do documento) e, opcionalmente, exemplos.

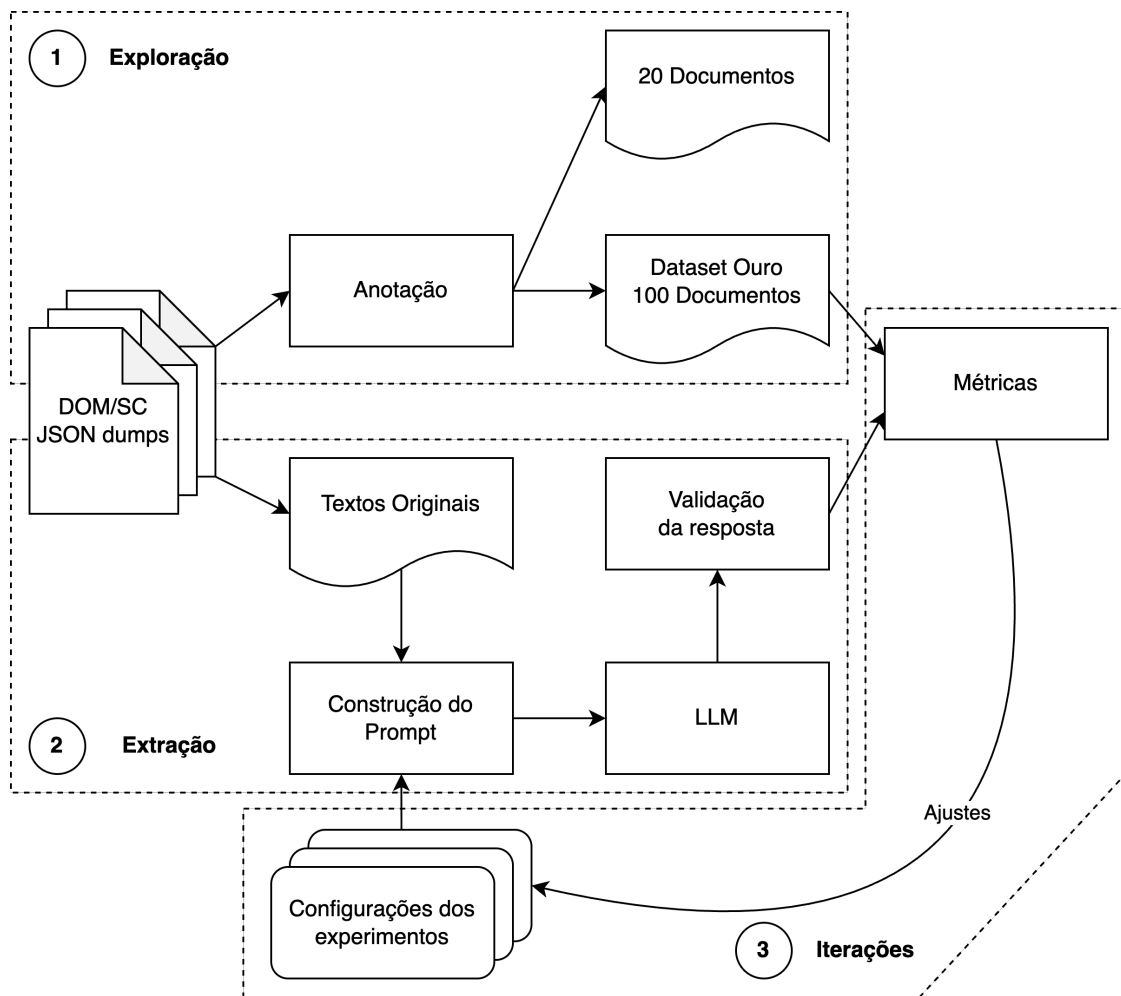


Figura 2. Diagrama do desenvolvimento

3. **Execução com ou sem RAG:** A configuração do experimento determina se a técnica de RAG será utilizada. Se ativada, o texto é dividido em *chunks*, transformado em vetores com um modelo de *embedding* (BERTimbau [Souza et al. 2020]), e os trechos mais relevantes são recuperados para compor o prompt.
4. **Geração e Validação:** A consulta é enviada ao LLM com um JSON Schema para forçar a saída estruturada. A resposta é então validada por uma classe Pydantic correspondente, garantindo a integridade dos dados.

Os experimentos foram gerenciados por arquivos de configuração YAML, garantindo a rastreabilidade e reprodutibilidade.

4.3. Métricas de Avaliação

A avaliação foi realizada comparando as extrações com o conjunto de dados ouro. Foram utilizadas as métricas de acurácia, precisão, revocação e F1-score. Para campos categóricos (*enum*), foi calculada a média macro dessas métricas. As datas foram normalizadas para o formato ISO, e os textos foram normalizados (conversão para ASCII, remoção de pontuação e espaços extras) para mitigar variações superficiais. Para o campo

”objeto”, foi usada a distância de Levenshtein com um limiar de 0,8, enquanto os demais campos exigiram correspondência exata.

5. Resultados e Discussão

Os experimentos foram conduzidos de forma incremental, utilizando o Llama 3.3-70B [Meta 2025] como modelo de referência inicial.

5.1. Engenharia de Prompt

Os primeiros experimentos dedicaram-se a refinar o prompt, que evoluiu de uma instrução sucinta para uma versão completa com definição de papel, regras explícitas e validação por *JSON Schema*. O *prompt* simples obteve F1 médio de 60,95 %. Após nove ciclos de ajustes, o *prompt* enriquecido elevou o F1 médio a 87,29 % (+26,34 p.p.) e a acurácia média a 89,45 % (+32,36 p.p.). Os ganhos foram especialmente expressivos nos campos *Data de Abertura* e *Número do Processo Licitatório*, enquanto *Município* atingiu desempenho perfeito. A Tabela 2 sintetiza essa evolução.

Métrica	Prompt simples	Prompt enriquecido	Δ absoluto
Acurácia média	57,09%	89,45%	+32,36 p.p.
F1-score médio	60,95%	87,29%	+26,34 p.p.

Tabela 2. Evolução do desempenho durante a engenharia de prompt

5.2. Inclusão de Exemplos (Few-shot)

O segundo experimento, ilustrado na Figura 3, avaliou o efeito de inserir exemplos na mensagem enviada ao modelo. A configuração que apresentou o melhor resultado foi o *few-shot* contendo somente a saída-referência (JSON), sem o texto de entrada dos exemplos. Essa estratégia elevou o F1-score médio de 87,29 % para 89,17 % (+1,88 p.p.) e a acurácia média de 89,45 % para 91,00 % (+1,55 p.p.), conforme resumido na Tabela 3. Os ganhos mais expressivos ocorreram nos campos *Objeto* (+5,0 p.p.) e *Signatário* (+5,5 p.p.), detalhados na Tabela 4. Esses resultados confirmam que fornecer apenas a estrutura de saída esperada é suficiente para orientar o modelo, dispensando a inclusão do texto de entrada nos exemplos.

Métrica	Linha de base	Few-shot	Variação
Acurácia média	89,45%	91,00%	+1,55 p.p.
F1-score médio	87,29%	89,17%	+1,88 p.p.

Tabela 3. Comparação de desempenho entre a linha de base e a configuração *few-shot* (saída apenas)

5.3. Experimento com RAG

No terceiro experimento avaliamos o uso de *Retrieval-Augmented Generation* (RAG) em 36 combinações de *chunk size*, *overlap* e *top-k*. Em vez de aprimorar o resultado, todas as variações reduziram o desempenho: a melhor configuração (ID 26) registrou F1 médio de 66,54 %, 22,63 p.p. abaixo da linha de base do segundo experimento. Isso indica que, para os documentos curtos estudados, o texto integral já contém contexto suficiente e que a fragmentação e a recuperação introduzidas pelo RAG acrescentam ruído e complexidade desnecessários. A Tabela 5 resume esse efeito negativo.

Você é um especialista em licitação brasileira.

Extraia as seguintes entidades do documento de licitação no # Contexto

- Tipo do documento [enum] ...
- Número do processo administrativo [string] ^\d+/\d+\$
- Município [enum] ...
- Modalidade [enum] ...
- Formato da modalidade [enum|null] ...
- Número da modalidade [string] ^\d+/\d+\$
- Objeto [string] ^.+
- Data de abertura [string|null] \d{4}-\d{2}-\d{2}T\d{2}:\d{2}:\d{2}\$
- Site do edital [string|null] ^(\https?:\/\/|www\.)(^\s)+\$
- Nome do Signatário [string|null] ^.+
- Cargo do signatário [string|null] ^.+

Regras

...

Retorne no formato JSON.

Exemplos

{EXEMPLOS}

Contexto

{CONTEXT0}

,

Figura 3. Melhor prompt simplificado

Campo	Acurácia	Precisão	Revocação	F1-score
Tipo do Documento	89,00%	88,89%	93,71%	91,23%
Número do Processo Licitatório	91,00%	91,00%	91,00%	91,00%
Município	100,00%	100,00%	100,00%	100,00%
Modalidade	97,00%	98,40%	97,48%	97,94%
Formato da Modalidade	92,00%	92,55%	90,78%	91,66%
Número da Modalidade	94,00%	94,00%	94,00%	94,00%
Objeto	81,00%	81,00%	81,00%	81,00%
Data de Abertura	90,00%	70,97%	95,65%	81,48%
Site do Edital	83,00%	58,54%	77,42%	66,67%
Signatário	92,00%	92,05%	94,19%	93,10%
Cargo do Signatário	92,00%	95,12%	90,70%	92,86%

Tabela 4. Desempenho detalhado por campo, configuração *few-shot* (saída apenas)

Métrica	Linha de base	RAG (conf. 26)	Δ absoluto
Acurácia média	91,00 %	63,55 %	-27,45 p.p.
F1-score médio	89,17 %	66,54 %	-22,63 p.p.

Tabela 5. Impacto do RAG no desempenho de extração

5.4. Experimento com LLMs

O melhor *prompt* (*few-shot* com saída JSON) foi aplicado a quatro modelos abertos de última geração: Llama-3.3-70B, Qwen-2.5-72B , DeepSeek R1-70B e Gemma-3-27B. Todos produziram resultados muito próximos, com F1 médio dentro de um intervalo de 3,6 p.p. (Gemma-3: 85,6 % Llama-3.3: 89,2 %). O Qwen-2.5 liderou em *Modalidade* (98,85 %) e *Cargo do Signatário* (94,05 %), enquanto o Gemma-3 obteve a melhor pontuação para *Site do Edital* (79,37 %). A pequena variação entre modelos sugere que a engenharia de *prompt* teve impacto maior que a escolha do LLM.

Modelo	Acurácia média	F1-score médio
Llama-3.3-70B	91,00 %	89,17 %
Qwen-2.5-72B	90,81 %	88,76 %
DeepSeek R1-70B	88,90 %	87,50 %
Gemma-3-27B	88,36 %	85,59 %

Tabela 6. Desempenho médio dos LLMs com o *prompt* final

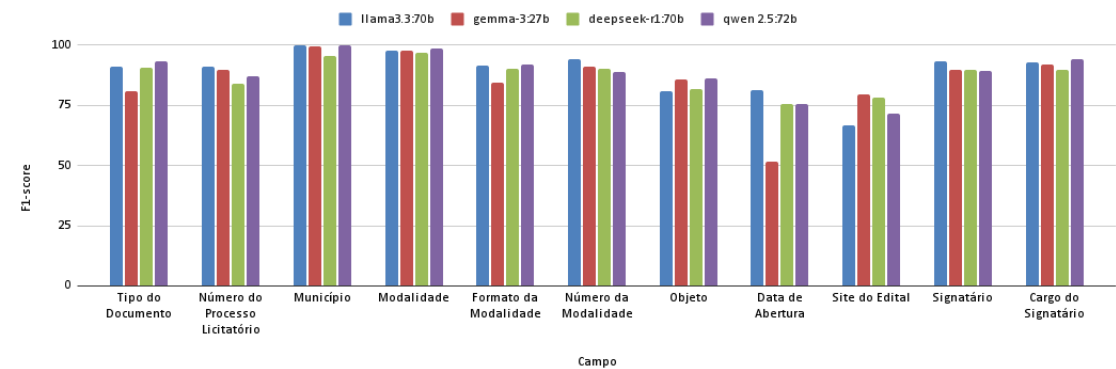


Figura 4. Distribuição dos F1-scores por campo para cada modelo

5.5. Análise de Erros

A análise qualitativa dos erros revelou padrões recorrentes:

- **Objeto:** Truncamento da descrição, resultando em extrações parciais.
- **Data de Abertura e Site do Edital:** Instabilidade devido à baixa representatividade no dataset e à confusão com outros padrões numéricos ou textuais.
- **Campos Numéricos:** Troca entre "Número do Processo Licitatório" e "Número da Modalidade" devido ao padrão similar (ex: "123/2025").
- **Campos Categóricos:** Troca entre rótulos semelhantes, como "Homologação" e "Adjudicação".
- **Signatário e Cargo:** Extrações incompletas, duplicação de nomes ou troca de informações entre os dois campos.

6. Conclusão

Este trabalho demonstrou a viabilidade e a eficácia do uso de Large Language Models de código aberto para a extração estruturada de dados de publicações de licitações do

DOM/SC, alcançando um F1-score médio de quase 90%. A combinação de engenharia de prompt, exemplos *few-shot* (apenas saída) e saídas forçadas em JSON se mostrou uma estratégia robusta, superando tanto a simples troca de modelos quanto o uso de RAG para os documentos curtos analisados. A Figura 3 apresenta o *prompt* que obteve o melhor desempenho, servindo como referência para futuras replicações.

As principais contribuições são: (1) a demonstração de que LLMs abertos podem atingir alta performance em tarefas de extração em documentos administrativos; (2) a constatação de que, em textos curtos, a engenharia de prompt é mais crucial que a arquitetura do modelo ou técnicas de RAG; e (3) a proposição de um fluxo de trabalho reprodutível para a extração de dados que fortalece a transparência e a fiscalização de contratações públicas.

Como limitações, o conjunto de dados de validação é restrito a 100 documentos, e certos campos estavam sub-representados. Erros como truncamento e confusão entre campos persistem.

Para trabalhos futuros, recomenda-se a expansão do corpus de avaliação, a investigação do uso adaptativo de RAG apenas para documentos longos e o desenvolvimento de estratégias para lidar com campos compostos e múltiplos signatários. Conclui-se que a automação baseada em LLMs é um avanço significativo para a eficiência e audibilidade no setor público.

Referências

- Aquino, I., dos Santos, M. M., Dorneles, C., and Carvalho, J. T. (2024). Extracting information from brazilian legal documents with retrieval augmented generation. In *Anais Estendidos do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 280–287, Porto Alegre, RS, Brasil. SBC.
- Brasil (1988). Constituição da república federativa do brasil de 1988.
- Brasil (2011). Lei nº 12.527, de 18 de novembro de 2011. Regula o acesso a informações previsto no inciso XXXIII do art. 5º, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal; altera a Lei nº 8.112, de 11 de dezembro de 1990; revoga a Lei nº 11.111, de 5 de maio de 2005, e dispositivos da Lei nº 8.159, de 8 de janeiro de 1991; e dá outras providências.
- Brasil (2021). Lei nº 14.133, de 1º de abril de 2021. Lei de Licitações e Contratos Administrativos.
- CIGA (2025). Recorde.
- de Faria, J. R., Xie, H., and Steffek, F. (2024). Automatic information extraction from employment tribunal judgements using large language models.
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., and Wang, H. (2024). Retrieval-augmented generation for large language models: A survey.
- ggml org (2025). llama.cpp - gbnf grammars guide. <https://github.com/ggml-org/llama.cpp/blob/master/grammars/README.md>.
- Guimarães, G. M., da Silva, F. X., Queiroz, A. L., Marcacini, R. M., Faleiros, T. P., Borges, V. R., and Garcia, L. P. (2024). Dodfminer: An automated tool for named entity recognition from official gazettes. *Neurocomputing*, 568:127064.

- Hamdi, A., Carel, E., Joseph, A., Coustaty, M., and Doucet, A. (2021). Information extraction from invoices. In *International Conference on Document Analysis and Recognition*, pages 699–714. Springer.
- Hussain, A. S. and Thomas, A. (2024). Large language models for judicial entity extraction: A comparative study.
- Katz, D. M., Hartung, D., Gerlach, L., Jana, A., and II, M. J. B. (2023). Natural language processing in the legal domain.
- Keshavarz, H., Vagena, Z., Kouki, P., Fountalis, I., Mabrouki, M., and Belaweid, A. (2022). Named entity recognition in long documents: An end-to-end case study in the legal domain. In *2022 IEEE International Conference on Big Data (Big Data)*, Osaka, Japan. IEEE.
- Meta (2025). Llama 3.3 model card.
- Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., and Gao, J. (2024). Large language models: A survey. arXiv:2402.06196 [cs.CL].
- Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticæ Investigationes*, 30(1):3–26.
- Pereira, J., Assumpcao, A., Trecenti, J., Airoso, L., Lente, C., Cléto, J., Dobins, G., Nogueira, R., Mitchell, L., and Lotufo, R. (2024). Inacia: Integrating large language models in brazilian audit courts: Opportunities and challenges.
- Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., Costa, H. D., Gupta, S., Rogers, M. L., Goncarencu, I., Sarli, G., Galynker, I., Peskoff, D., Carpuat, M., White, J., Anadkat, S., Hoyle, A., and Resnik, P. (2025). The prompt report: A systematic survey of prompt engineering techniques.
- Sleimi, A., Sannier, N., Sabetzadeh, M., Briand, L., and Dann, J. (2018). Automated extraction of semantic legal metadata using natural language processing. In *2018 IEEE 26th International Requirements Engineering Conference (RE)*, pages 124–135.
- Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I*, page 403–417, Berlin, Heidelberg. Springer-Verlag.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. arXiv:1706.03762 [cs.CL].
- Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., and Wang, G. (2023). Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*, v4. Last revised 7 Oct 2023.

APÊNDICE C – CÓDIGO FONTE

O repositório que reúne os códigos-fonte e os resultados do experimento está disponível em: <https://github.com/felipevalentin/tcc>