



UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE CIÊNCIAS JURÍDICAS
DEPARTAMENTO DE DIREITO
CURSO DE DIREITO

Guilherme de Brito Santos

Engenharia de Prompt para Extração de Fatores Jurídicos: Validação e Otimização de uma
Metodologia com Modelos de Linguagem de Grande Escala no Projeto CONCIL-IA

Florianópolis
2025

Guilherme de Brito Santos

Engenharia de Prompt para Extração de Fatores Jurídicos: Validação e Otimização de uma Metodologia com Modelos de Linguagem de Grande Escala no Projeto CONCIL-IA

Trabalho de Conclusão de Curso submetido ao Curso de Graduação em Curso de Direito do Centro de Ciências Jurídicas da Universidade Federal de Santa Catarina como requisito para obtenção do título de Bacharel em Direito.
Orientadora: Prof^a. Isabela Cristina Sabo, Dra.

Florianópolis
2025

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

Santos, Guilherme de Brito
Engenharia de Prompt para Extração de Fatores Jurídicos :
Validação e Otimização de uma Metodologia com Modelos de
Linguagem de Grande Escala no Projeto CONCIL-IA /
Guilherme de Brito Santos ; orientadora, Isabela Cristina
Sabo, 2025.
104 p.

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

Inclui referências.

1. Direito. 2. Engenharia de prompt. 3. Inteligência
Artificial e Direito. 4. Modelos de Linguagem de Larga
Escala. 5. Estruturação de Dados Jurídicos. I. Sabo, Isabela
Cristina . II. Universidade Federal de Santa Catarina.
Graduação em Direito. III. Título.



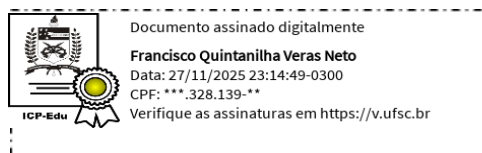
SERVIÇO PÚBLICO FEDERAL
UNIVERSIDADE FEDERAL DE SANTA CATARINA – UFSC
CENTRO DE CIÊNCIAS JURÍDICAS
Coordenação de TCC

Guilherme de Brito Santos

Engenharia de Prompt para Extração de Fatores Jurídicos: Validação e Otimização de uma Metodologia com Modelos de Linguagem de Grande Escala no Projeto CONCIL-IA

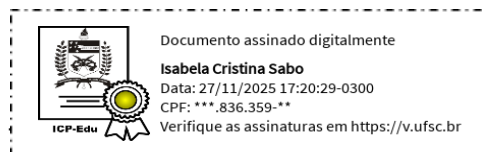
Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do título de graduado e aprovado em sua forma final pelo Curso de Direito.

Florianópolis, 25 de novembro de 2025

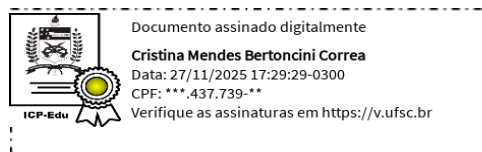


Coordenação do Curso

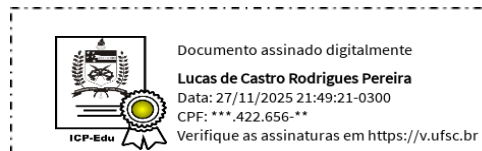
Banca examinadora



Prof. Dra. Isabela Cristina Sabo
Orientador(a)



Prof. Dra. Cristina Mendes Bertoncini Correa
Instituição UFSC



Me. Lucas de Castro Rodrigues Pereira
Instituição UFSC

Florianópolis, 2025.

Dedico este trabalho primeiro às minhas avós, que conheceram de perto a dificuldade e a falta, mas lutaram para dar aos filhos tudo aquilo que nunca tiveram. Depois aos meus pais, que transformaram esses sacrifícios em oportunidades para mim. Por cada noite mal dormida, cada hora a mais trabalhada, cada sonho adiado em favor dos nossos: este trabalho é para vocês, com toda minha gratidão e amor.

AGRADECIMENTOS

Nas últimas semanas de elaboração deste trabalho, caminhando pela Rua Lauro Linhares, próxima à cidade universitária, passei em frente ao hotel onde fiquei no dia da minha matrícula na UFSC. Naquele momento, entre o cansaço das revisões finais e a ansiedade da entrega, aquela fachada me transportou, de forma inesperada, àquele dia inicial, quando tudo era expectativa e nervosismo pela jornada que se iniciava. Que boa coincidência. Por um instante, todos os pensamentos acelerados cessaram, e veio à memória apenas aquele momento marcado pela felicidade de estar iniciando um sonho. E eu consegui. Mas não sozinho.

Talvez a minha parte tenha sido a mais fácil. Tantas pessoas participaram ativamente desta conquista que, se eu fosse mencionar todas com o devido rigor, a UFSC provavelmente não aceitaria o trabalho pelo limite de páginas. São muitas, e sou eternamente grato a cada uma delas.

Agradeço, de forma especial, à Prof.^a Dr.^a Isabela, minha orientadora neste trabalho e no projeto CONCIL-IA. Se este trabalho tem alguma qualidade, é pela sua solícita e carinhosa orientação ao longo destes mais de dois anos de parceria. Além da orientação acadêmica, sou igualmente grato pelos conselhos, pelo carinho e pelas palavras gentis que marcaram nossa relação durante esses anos de trabalho conjunto. Estendo meus agradecimentos a todos os integrantes do projeto CONCIL-IA, pelos ensinamentos e pela colaboração que contribuíram significativamente para minha formação.

Agradeço aos meus pais, Luana e Osvaldo, que são a minha base de tudo. Minhas oportunidades, minhas conquistas, esta graduação são, antes, lutas de vocês do que minhas.

Minha mãe, Luana, obrigado por me apoiar incondicionalmente em meus sonhos. Quando escolhi a UFSC, a 2.500 km de casa, a senhora nunca hesitou em me apoiar. Ao expressar o desejo de fazer um intercâmbio, mesmo com as dificuldades do momento, a resposta foi a mesma. Sempre foi sim, "nós damos um jeito depois". A senhora, em toda a vida, colocou constantemente nossos sonhos em primeiro lugar, com todo o cuidado para que tenhamos o melhor conforto possível. Por isso, e por muito mais, meu profundo e eterno obrigado.

Meu pai, Osvaldo, obrigado por todo o apoio aos estudos, à vida e aos meus sonhos. Viagens para São Paulo, mudanças súbitas de curso, uma escolha de faculdade distante: em tudo isso, o senhor esteve comigo, não decidindo por mim, mas ajudando e apoiando minhas escolhas. Sou especialmente grato por me apoiar em minha decisão de fazer a faculdade de Direito e por valorizar que devemos fazer o que nos move, o que amamos. Este apoio me fez, e ainda faz, persistir em atingir meus sonhos não apenas materiais, mas de satisfação pessoal com a profissão que escolhi. Se hoje me formo em Direito, um curso e uma área que amo, devo ao senhor, pai.

Agradeço às minhas avós, Florinda e Ceir, bases de toda a nossa família em amor, cuidado e estímulo aos estudos. Nossas conquistas e vocações devem-se a vocês. Agradeço especialmente à minha falecida avó Florinda, que cuidou de mim em meu último ano do ensino médio, quando morei com ela para estudar para o vestibular. Foi o cuidado que a senhora teve comigo naquele ano que me deu forças para conquistar a aprovação na UFSC. Este diploma

também é da senhora, esteja onde estiver.

Agradeço ao tio Naldo e à tia Jô pelo apoio em 2019, quando morei em Tangará. O suporte de vocês naquele ano foi fundamental para esta conquista. A todos os meus tios, tias, primos, primas e avô, obrigado por todo o amor e carinho. Aos meus irmãos, Giovana e Henrique: este diploma também veio com custos, e o principal deles foi a distância de vocês. Foi muito difícil, mas agradeço por toda a compreensão e carinho. Amo vocês profundamente e espero que realizem todos os seus sonhos.

Agradeço aos amigos que formei aqui: Maite, Taiane, Adrielly, Luisa, Cadu, Maria Eduarda, Laura, João e Daniela. Especialmente às minhas amigas do EAD, Maite, Taiane e Adrielly, que se tornaram companheiras para toda a vida: entre as muitas alegrias que esta jornada me trouxe, nossa amizade certamente é uma das mais valiosas. Agradeço também ao Tiago, meu colega de apartamento desde que me mudei para Florianópolis, que foi companheiro em todos esses anos de altos e baixos.

Agradeço a todos os professores que tive ao longo da minha vida, desde os primeiros anos escolares até esta graduação. Cada um de vocês contribuiu para a construção do profissional e da pessoa que me tornei, despertando minha curiosidade e amor pelos estudos.

Por fim, agradeço à UFSC, que serviu como uma nova casa nestes últimos quase cinco anos. Universidade pública, gratuita e de qualidade: jargão que, mesmo sendo um clichê, nunca devemos esquecer nem deixar de cultivar.

A todos vocês, meu sincero e profundo agradecimento.

"É muito melhor ter uma resposta aproximada para a pergunta certa, ainda que vaga, do que uma resposta exata para a pergunta errada, que pode sempre ser formulada com precisão."
(John Tukey)

RESUMO

Este trabalho, caracterizado como aprofundamento de um estudo de caso, investiga o processo de engenharia de prompt para extração automatizada de fatores jurídicos a partir de decisões judiciais não estruturadas. A pesquisa integra o projeto CONCIL-IA, iniciativa que busca automatizar a estruturação de dados jurídicos para viabilizar o treinamento de modelos de análise preditiva. O objetivo central consiste em estimar valores de indenização por danos morais em litígios de transporte aéreo, fornecendo subsídio técnico para auxiliar conciliadores e fomentar a resolução consensual de disputas. A investigação aborda dois problemas: se o processo iterativo de engenharia de prompt influencia o desempenho de um Modelo de Linguagem de Larga Escala (LLM) na extração dos fatores e quais falhas persistem; e se estratégias de otimização podem corrigir tais falhas. As hipóteses confirmadas, postularam que o processo iterativo capacita o LLM a realizar a extração com precisão, embora com falhas, e que o aprimoramento foi observado mediante estratégias que reduzem a carga cognitiva do modelo, pela fusão de fatores e reestruturação da lógica de saída. Adotando a indução como método de abordagem, os procedimentos metodológicos articularam pesquisa bibliográfica e documental com estudo de caso que, em sua segunda fase, assumiu caráter experimental, com técnicas de análise baseadas em avaliação quantitativa comparativa. A investigação partiu da construção de um *ground truth* elaborado manualmente por especialistas e desdobrou-se em duas fases experimentais. O Experimento Principal analisou o refinamento iterativo de um prompt geral, confirmando a primeira hipótese ao elevar a acurácia geral para 91,5%, porém revelando falhas persistentes na extração de fatores minoritários e erro sistemático na extração de dimensões numéricas. O Experimento Secundário foi concebido para solucionar essas falhas, testando hipóteses de otimização através de prompts especializados que exploraram a fusão estratégica de fatores e a reestruturação da lógica de saída. Os resultados corroboraram a segunda hipótese, estabelecendo o prompt *atraso5* como a referência metodológica deste estudo. Esta versão otimizada apresentou o melhor equilíbrio entre integridade dos registros, alcançando a maior Acurácia por Sentença do estudo e estabilidade dos fatores numéricos ao solucionar a falha relacionada ao *Root Mean Square Error* (Erro Quadrático Médio Raiz) (RMSE). Adicionalmente, a investigação expôs a fragilidade do formato CSV como limitação estrutural que impacta a qualidade da extração. As conclusões demonstram que o processo metodológico iterativo influencia a eficácia da extração e que a evolução da área demanda paradigmas alternativos, como *Markdown* para estruturação de entrada e *JSON Schema* para padronização de saída. O trabalho contribui com uma documentação de um roteiro metodológico na aplicação de LLMs à estruturação de dados jurídicos não estruturados.

Palavras-chave: Engenharia de Prompt; Inteligência Artificial e Direito; Modelos de Linguagem de Grande Escala; Estruturação de Dados Jurídicos; Resolução de Disputas Online.

ABSTRACT

This work, characterized as an in-depth case study, investigates the prompt engineering process for automated extraction of legal factors from unstructured judicial decisions. The research is part of the CONCIL-IA project, an initiative that seeks to automate the structuring of legal data to enable the training of predictive analysis models. The central objective consists of estimating compensation values for moral damages in air transport litigation, providing technical support to assist conciliators and foster consensual dispute resolution. The investigation addresses two problems: whether the iterative prompt engineering process influences the performance of an LLM in factor extraction and which failures persist; and whether optimization strategies can correct such failures. The hypotheses, confirmed, postulated that the iterative process enables the LLM to perform extraction with precision, albeit with failures, and that improvement was observed through strategies that reduce the cognitive load of the model, by merging factors and restructuring output logic. Adopting induction as the approach method, the methodological procedures articulated bibliographic and documentary research with a case study that, in its second phase, assumed an experimental character, with analysis techniques based on comparative quantitative evaluation. The investigation started from the construction of a *ground truth* manually elaborated by specialists and unfolded into two experimental phases. The Main Experiment analyzed the iterative refinement of a general prompt, confirming the first hypothesis by raising overall accuracy to 91.5%, yet revealing persistent failures in the extraction of minority factors and systematic error in the extraction of numerical dimensions. The Secondary Experiment was designed to solve these failures, testing optimization hypotheses through specialized prompts that explored strategic factor merging and output logic restructuring. The results corroborated the second hypothesis, establishing the *atraso5* prompt as the methodological reference of this study. This optimized version presented the best balance between record integrity, achieving the highest Sentence Accuracy of the study, and stability of numerical factors by solving the critical failure related to RMSE. Additionally, the investigation exposed the fragility of the CSV format as a structural limitation that impacts extraction quality. The conclusions demonstrate that iterative methodological rigor influences extraction efficacy and that the evolution of the field demands alternative paradigms, such as *Markdown* for input structuring and *JSON Schema* for output standardization. The work contributes with a documentation of an methodological roadmap about the application of LLMs to the structuring of unstructured legal data.

Keywords: Prompt Engineering; Artificial Intelligence and Law; Large Language Models; Legal Data Structuring; Online Dispute Resolution.

LISTA DE TABELAS

Tabela 1 – Relação entre Objetivos, Procedimentos e Técnicas da Pesquisa.	20
Tabela 2 – Acurácia Geral (%) no Experimento <i>Singular</i>	65
Tabela 3 – F1-Score Geral no Experimento <i>Singular</i>	66
Tabela 4 – Acurácia Individual por Fator (%) no Experimento <i>Singular</i>	66
Tabela 5 – F1-Score (F1-S), Precisão (Prec) e Recall (Rec) por Fator no Experimento <i>Singular</i>	67
Tabela 6 – RMSE Individual por Dimensão no Experimento <i>Singular</i>	67
Tabela 7 – Acurácia Geral (%) no Experimento <i>Factor-Based Sets</i>	69
Tabela 8 – F1-Score Geral no Experimento <i>Factor-Based Sets</i>	69
Tabela 9 – Acurácia Individual por Fator (%) no Experimento <i>Factor-Based Sets</i>	69
Tabela 10 – F1-Score (F1-S), Precisão (Prec) e Recall (Rec) por Fator no Experimento <i>Factor-Based Sets</i>	70
Tabela 11 – RMSE Individual por Dimensão no Experimento <i>Factor-Based Sets</i>	70
Tabela 12 – Resultados comparativos de performance dos novos prompts nas diferentes amostras	72

LISTA DE ABREVIATURAS E SIGLAS

AGI Inteligência Artificial Geral

API *Application Programming Interface*

AREsps Agravos em Recurso Especial

ASI Superinteligência Artificial

BERT *Bidirectional Encoder Representations from Transformers*

CAGI Inteligência Artificial Geral Conversacional

CBR Raciocínio Orientado por Casos (*Case-Based Reasoning*)

CNJ Conselho Nacional de Justiça

CNPq Conselho Nacional de Desenvolvimento Científico e Tecnológico

CoT *Chain-of-Thought* (Cadeia de Pensamento)

CSV *Comma-Separated Values*

FAPESC Fundação de Amparo à Pesquisa e Inovação do Estado de Santa Catarina

FBR Raciocínio Baseado por Fatores (*Factor-Based Reasoning*)

FN Falso Negativo (*False Negative*)

FP Falso Positivo (*False Positive*)

GPAI *General-purpose AI Model* (Modelo de IA de Propósito Geral)

GPT *Generative Pre-trained Transformer*

IA Inteligência Artificial

IAGen Inteligência Artificial Generativa

JEC Juizado Especial Cível

JSON *JavaScript Object Notation*

LLM Modelo de Linguagem de Larga Escala

MAE *Mean Absolute Error* (Erro Absoluto Médio)

PLN Processamento de Linguagem Natural (*Natural Language Processing*)

ODR Resolução de Disputas Online (*Online Dispute Resolution*)

PRILJ *Paragraph Regularities Identification in Legal Judgments*

RLHF *Reinforcement Learning from Human Feedback* (Aprendizagem por Reforço com Feedback Humano)

RM *Reward Model* (Modelo de Recompensa)

RMSE *Root Mean Square Error* (Erro Quadrático Médio Raiz)

RNN Rede Neural Recorrente

SFT *Supervised Fine-Tuning* (Ajuste Fino Supervisionado)

STJ Superior Tribunal de Justiça

TJSC Tribunal de Justiça do Estado de Santa Catarina

TN Verdadeiro Negativo (*True Negative*)

TP Verdadeiro Positivo (*True Positive*)

UFSC Universidade Federal de Santa Catarina

VS Code *Microsoft Visual Studio Code*

XAI *Explainable Artificial Intelligence* (Inteligência Artificial Explicável)

PIC Programa de Iniciação Científica

SUMÁRIO

1	INTRODUÇÃO	17
1.1	CONTEXTO	17
1.2	PROBLEMAS E HIPÓTESES DE PESQUISA	17
1.2.1	Problemas de Pesquisa	18
1.2.2	Hipóteses	18
1.3	OBJETIVOS	18
1.3.1	Objetivo Geral	18
1.3.2	Objetivos Específicos	18
1.4	METODOLOGIA	19
1.5	JUSTIFICATIVA	20
1.6	ORGANIZAÇÃO DO TRABALHO	21
2	FUNDAMENTAÇÃO TEÓRICA	22
2.1	INTELIGÊNCIA ARTIFICIAL	22
2.2	APRENDIZADO DE MÁQUINA (<i>MACHINE LEARNING</i>)	23
2.3	REDES NEURAIS E APRENDIZADO PROFUNDO (<i>DEEP LEARNING</i>)	25
2.4	PROCESSAMENTO DE LINGUAGEM NATURAL - PLN (<i>NATURE LANGUAGE PROCESSING - NLP</i>)	26
2.5	INTELIGÊNCIA ARTIFICIAL GENERATIVA (<i>GENERATIVE ARTIFICIAL INTELLIGENCE</i>)	27
2.5.1	Fundamentos da Inteligência Artificial Generativa	27
2.5.2	A Arquitetura Transformer	28
2.5.3	Alinhamento com Intenção Humana	29
2.5.4	Técnicas de Prompting e a Inteligência Artificial Generativa	30
2.6	ENGENHARIA DE PROMPT: TEORIA E TÉCNICAS	31
2.6.1	Definindo Engenharia de Prompt	31
2.6.2	A Importância da Engenharia de Prompt	31
2.6.3	Técnicas de Engenharia de Prompt	32
2.6.4	Tendências Futuras	34
3	TRABALHOS RELACIONADOS E APLICAÇÕES NO BRASIL	36
3.1	USO DE IA GENERATIVA NO DOMÍNIO JURÍDICO	36
3.2	ARCABOUÇO NORMATIVO NO BRASIL E O AI ACT	39
3.2.1	Resoluções 332/2020 e 615/2025 do Conselho Nacional de Justiça - CNJ	39
3.2.2	AI Act e o Projeto de Lei 2338/2023	42
4	ESTUDO DE CASO: ENGENHARIA DE PROMPT NO PROJETO CONCIL-IA	45

4.1	O PROJETO CONCIL-IA E SUA METODOLOGIA DE DESENVOLVIMENTO	45
4.2	BASE DE DADOS: O <i>CORPUS</i> DE ANÁLISE	47
4.2.1	Coleta e Contexto do Dataset Não Estruturado	47
4.2.2	O Processo de Anonimização	48
4.3	ABORDAGENS E TÉCNICAS DE PESQUISA	49
4.3.1	Contextualização Teórico-Metodológica: Raciocínio Orientado por Fatores (FBR)	49
4.3.2	Estruturação Manual dos Dados: O <i>Ground Truth</i>	50
4.3.3	Estruturação Automática dos Dados via Engenharia de Prompt	52
4.4	O PROCESSO DE ENGENHARIA DE PROMPT	53
4.4.1	Experimento Principal: do <i>Prompt Inicial</i> ao <i>Prompt Final</i>	53
4.4.1.1	<i>O Prompt Inicial e suas Características</i>	53
4.4.1.2	<i>O Processo Iterativo de Refinamento</i>	54
4.4.1.3	<i>O Prompt Refinado (Prompt Final): Análise das Melhorias</i>	55
4.4.1.4	<i>Desenho Experimental</i>	56
4.4.1.4.1	Experimento 1: Abordagem <i>Singular</i>	57
4.4.1.4.2	Experimento 2: Abordagem <i>Factor-Based Sets</i>	57
4.4.2	Experimento Secundário: Sub-Hipóteses de Otimização	58
4.4.2.1	<i>Formulação das Sub-Hipóteses de Otimização</i>	58
4.4.2.2	<i>Desenho Experimental: Variações de Prompts e Amostras de Dados</i>	59
4.5	MODELOS, PARÂMETROS E AMBIENTE EXPERIMENTAL	61
4.6	MÉTRICAS DE AVALIAÇÃO	61
4.6.1	Avaliação de Fatores Categóricos	62
4.6.2	Avaliação de Fatores Numéricos	63
5	RESULTADOS E DISCUSSÃO	65
5.1	ANÁLISE DO EXPERIMENTO PRINCIPAL	65
5.1.1	Análise do Experimento com Abordagem <i>Singular</i>	65
5.1.1.1	<i>Desempenho Geral e por Fator</i>	65
5.1.1.2	<i>Discussão dos Achados</i>	66
5.1.2	Análise do Experimento com Abordagem <i>Factor-Based Sets</i>	68
5.1.2.1	<i>Desempenho Geral e por Fator</i>	68
5.1.2.2	<i>Discussão dos Achados</i>	70
5.2	ANÁLISE DO EXPERIMENTO SECUNDÁRIO	71
5.2.1	Análise Comparativa de Desempenho	71
5.2.2	Síntese e Validação das Sub-Hipóteses	73
5.3	LIMITAÇÕES DO ESTUDO	74

5.4	NOVOS PARADIGMAS NA ENTRADA E SAÍDA DOS PROMPTS: <i>JavaScript Object Notation</i> (JSON) SCHEMA, MARKDOWN E A JANELA DE CON- TEXTO	76
5.4.1	Sub-experimento: Limitações do Formato CSV em Testes de Escalabilidade	76
5.4.2	Estratégias de Otimização e Desafios da Engenharia de Prompt	77
5.4.2.1	<i>O Estado da Arte em Geração de Saídas Estruturadas</i>	77
5.4.2.2	<i>O Impacto do Formato de Entrada: O Uso de Markdown</i>	77
5.4.2.3	<i>A Expansão da Janela de Contexto e Suas Implicações para a Engenharia de Prompt</i>	78
5.4.3	Implicações Metodológicas e Direccionamentos para Trabalhos Futuros .	79
6	CONCLUSÃO	81
	REFERÊNCIAS	83
	APÊNDICE A – APÊNDICES	89
A.1	PROMPT INICIAL (<i>INITIAL PROMPT</i>)	89
A.1.1	Versão em Português	89
A.1.2	Tradução para o Inglês	90
A.2	PROMPT FINAL (<i>FINAL PROMPT</i>)	92
A.2.1	Versão em Português	92
A.2.2	Tradução para o Inglês	94
A.3	PROMPT: ATRASO1	96
A.3.1	Versão em Português	96
A.3.2	Tradução para o Inglês	97
A.4	PROMPT: ATRASO3	98
A.4.1	Versão em Português	98
A.4.2	Tradução para o Inglês	98
A.5	PROMPT: ATRASO4	99
A.5.1	Versão em Português	99
A.5.2	Tradução para o Inglês	99
A.6	PROMPT: ATRASO5	100
A.6.1	Versão em Português	100
A.6.2	Tradução para o Inglês	100
A.7	PROMPT: EXTRAVIO1	100
A.7.1	Versão em Português	100
A.7.2	Tradução para o Inglês	101
A.8	PROMPT: EXTRAVIO2	102
A.8.1	Versão em Português	102
A.8.2	Tradução para o Inglês	102

A.9	FATORES 16 E 17: AVANÇOS NA PESQUISA	102
A.9.1	Versão em Português	102
A.9.2	Tradução para o Inglês	103

1 INTRODUÇÃO

1.1 CONTEXTO

O Poder Judiciário brasileiro enfrenta um desafio estrutural decorrente do crescimento contínuo da litigiosidade, fenômeno documentado pelo relatório "Justiça em Números" do Conselho Nacional de Justiça (CNJ) (Conselho Nacional de Justiça, 2025a). Em resposta a este cenário de sobrecarga processual, o CNJ tem implementado uma estratégia de modernização institucional que se desenvolve em duas frentes principais: por um lado, o incentivo a métodos autocompositivos de resolução de conflitos (Conselho Nacional de Justiça, 2010); por outro, a incorporação regulamentada de tecnologias emergentes, como a Inteligência Artificial (IA) (Conselho Nacional de Justiça, 2020; Conselho Nacional de Justiça, 2025b).

A eficácia de muitas dessas inovações tecnológicas, contudo, especialmente no campo de modelos preditivos¹ e jurimetria², depende de um insumo essencial: dados jurídicos de alta qualidade e em formato estruturado. Este requisito representa um desafio técnico relevante, uma vez que a vasta maioria da produção jurídica se encontra em texto não estruturado. É na interseção dessas duas agendas, a busca pela consensualidade e a inovação tecnológica responsável, que se insere o projeto CONCIL-IA, que investiga o potencial da IA para superar esse desafio e aprimorar os mecanismos de Resolução de Disputas Online (*Online Dispute Resolution*) (ODR) (Sabo et al., 2025).

1.2 PROBLEMAS E HIPÓTESES DE PESQUISA

A aplicação de técnicas de Aprendizado de Máquina no domínio jurídico enfrenta um obstáculo: a estruturação de dados. Tradicionalmente manual e laborioso, esse processo apresenta limitações de escalabilidade. Em trabalho anterior, no qual este autor participou da condução dos experimentos (Pereira et al., 2025), realizou-se uma análise preliminar sobre a viabilidade de automatizar essa estruturação por meio de um LLM, transformando decisões judiciais em dados estruturados aptos a servirem como entrada para modelos preditivos.

Os resultados confirmaram o potencial da ferramenta. Entretanto, o processo de engenharia dos dois prompts (denominados, respectivamente, *Inicial* e *Final*) não foi aprofundado, concentrando-se o trabalho principalmente na avaliação do desempenho da ferramenta. O presente estudo avança em relação a essa exploração inicial ao investigar os aspectos técnicos da validação em duas frentes complementares, cada uma orientada por um problema de pesquisa específico, conforme detalhado a seguir.

¹ Modelos Preditivos são algoritmos, baseados em estatística ou aprendizado de máquina, treinados com dados históricos para identificar padrões e, com base neles, prever resultados futuros (IBM, 2025).

² A Jurimetria consiste na aplicação de métodos estatísticos e quantitativos ao estudo de fenômenos jurídicos, com o objetivo de analisar padrões, compreender e prever o comportamento judicial (Loevinger, 1948).

1.2.1 Problemas de Pesquisa

Esta pesquisa busca responder a dois problemas centrais:

Problema 1: O processo iterativo de engenharia de prompt influencia o desempenho do LLM? Em caso afirmativo, quais falhas de extração persistem após esse processo?

Problema 2: As estratégias de otimização de prompt podem corrigir as falhas de extração do problema 1 e aprimorar a metodologia no projeto e em iniciativas semelhantes?

1.2.2 Hipóteses

Em resposta à primeira questão, a hipótese principal é afirmativa: o processo iterativo de engenharia de prompt capacita um LLM a realizar a extração de dados estruturados com elevada acurácia. Todavia, assume-se que o método ainda produz falhas de extração que necessitam ser identificadas e caracterizadas.

Quanto à segunda questão, postula-se que é possível aprimorar a metodologia e corrigir as falhas observadas. A hipótese secundária sustenta que esse avanço pode ser alcançado mediante estratégias que reduzem a carga cognitiva imposta ao modelo, especificamente por meio de duas abordagens: (i) a fusão de múltiplos fatores em uma única variável e (ii) a reestruturação da lógica de saída do modelo.

1.3 OBJETIVOS

1.3.1 Objetivo Geral

O objetivo geral deste trabalho é detalhar e expandir a investigação apresentada em Pereira et al. (2025), oferecendo uma análise da metodologia de engenharia de *prompt*, dos resultados quantitativos e qualitativos, do novo Experimento Secundário, das limitações da abordagem e dos novos paradigmas que podem direcionar a evolução do projeto.

1.3.2 Objetivos Específicos

Para alcançar o objetivo geral, foram definidos os seguintes objetivos específicos, que refletem as contribuições autorais desta monografia:

1. Fundamentar teoricamente a pesquisa, por meio de uma revisão da literatura especializada sobre Inteligência Artificial Generativa (IAGen) e, em especial, sobre as teorias, técnicas e tendências da Engenharia de *Prompt* no Capítulo 2.
2. Contextualizar a aplicação da IA no Direito, analisando trabalhos relacionados nos cenários global e nacional, bem como o arcabouço normativo e legislativo, incluindo as resoluções do CNJ, o AI Act europeu e o Projeto de Lei 2338/2023 no Capítulo 3.

3. Apresentar um relato do estudo de referência, detalhando o processo iterativo de engenharia de prompt e identificando as falhas de extração conforme descrito no Capítulo 4.
4. Aplicar novas técnicas e estratégias de engenharia de prompt especializadas, considerando as falhas de extração identificadas no estudo de referência (Pereira et al., 2025) conforme exposto no Capítulo 4.
5. Analisar comparativamente os resultados obtidos no Experimento Principal e no Experimento Secundário e propor novas estratégias em diferentes amostras de dados, validando um novo prompt padrão para o projeto e iniciativas semelhantes no Capítulo 5.
6. Discutir as limitações do método atual e delinear caminhos para a evolução metodológica do projeto, considerando novos paradigmas de formatação de entrada e saída de dados no Capítulo 5.

1.4 METODOLOGIA

Este trabalho adota e expande a metodologia empregada na pesquisa de referência (Pereira et al., 2025), caracterizando-se como um aprofundamento do estudo de caso original. Utiliza-se a indução como método de abordagem, uma vez que a análise de um caso particular visa gerar conhecimentos que possam ser replicados em iniciativas semelhantes. Os procedimentos metodológicos articulam a pesquisa bibliográfica e documental com um estudo de caso que, em sua segunda fase, assume um caráter de pesquisa experimental, sendo os resultados avaliados por meio de uma análise quantitativa comparativa.

O estudo é estruturado em duas fases: o Experimento Principal, que descreve a construção do *ground truth*, do *Prompt Inicial* e do *Prompt Final*, e o Experimento Secundário, que detalha o desenvolvimento de novas estratégias de otimização. A Tabela 1 sintetiza a articulação entre os objetivos, os procedimentos e as técnicas empregadas na pesquisa.

Além disso no desenvolvimento desta pesquisa, utilizou-se a ferramenta de IAGen Gemini 2.5 Pro como suporte à produção científica. Essa abordagem alinha-se ao conceito de "autoria híbrida", que consiste em revisar, editar e reformular o conteúdo gerado pela IA como parte do processo de produção e refinamento do texto acadêmico, preservando o rigor científico (Torres, 2025). Essa perspectiva não atribui à IAGen o papel de autor ou coautor, mas a posiciona como ferramenta cujas potencialidades são exploradas nas práticas acadêmicas, mantendo o pesquisador como protagonista e principal responsável pelo trabalho (Torres, 2025).

A utilização da IAGen neste trabalho concentrou-se em tarefas de apoio: orientação de pesquisa bibliográfica para fundamentar o estado da arte, revisão do texto e dicas gerais sobre a estrutura do trabalho. Nesse sentido, Torres (2025) ensina que a descrição explícita do uso da ferramenta constitui elemento central da autoria híbrida, que pressupõe a declaração do modo como a IA foi empregada. Essa transparência é necessária para garantir a integridade do trabalho,

Tabela 1 – Relação entre Objetivos, Procedimentos e Técnicas da Pesquisa.

Objetivo Específico	Procedimento	Técnica
Fundamentar teoricamente a pesquisa.	Pesquisa bibliográfica	Fundamentação teórica sobre IAGen e Engenharia de Prompt.
Contextualizar a aplicação da IA no Direito e seu arcabouço regulatório.	Pesquisa bibliográfica e documental	Análise de artigos, resoluções do CNJ e textos legislativos (AI Act e PL 2338/2023).
Apresentar o estudo de referência e identificar suas falhas.	Estudo de Caso	Detalhamento do processo iterativo e análise quantitativa para identificar as fragilidades do prompt de referência.
Aplicar novas técnicas de engenharia de prompt.	Estudo de Caso / Pesquisa Experimental	Desenvolvimento de prompts especializados (famílias Atraso e Extravio) para corrigir as falhas identificadas.
Analisar comparativamente os resultados e validar um novo prompt padrão.	Análise Quantitativa Comparativa	Análise de métricas (F1-Score, RMSE) entre os experimentos e as diferentes amostras para validar o prompt de melhor desempenho.
Discutir as limitações e delinear a evolução metodológica do projeto.	Análise Crítica e Pesquisa Bibliográfica	Discussão sobre limitações (custos, escopo) e pesquisa sobre novas práticas em geração estruturada (JSON Schema, Markdown).

Fonte: Elaborado pelo autor.

demonstrando compromisso com as boas práticas acadêmicas e evitando desvios de conduta que comprometam a ética do processo de produção do conhecimento (Torres, 2025)

1.5 JUSTIFICATIVA

A relevância deste estudo reside em sua contribuição para a transparência e a reprodutibilidade científica. O trabalho de referência (Pereira et al., 2025) validou o potencial da abordagem proposta, mas as limitações de formato de um artigo científico impediram uma exposição completa da metodologia e uma discussão pormenorizada dos resultados, limitações e perspectivas futuras. Esta monografia preenche essa lacuna ao oferecer um relato metodologicamente detalhado, possibilitando que outros pesquisadores possam compreender, replicar e expandir a investigação.

Além disso, o trabalho avança para além da simples documentação ao apresentar um novo Experimento Secundário, realizado em colaboração com a equipe interdisciplinar do CONCIL-IA. Este novo experimento testa e valida estratégias de otimização não exploradas na pesquisa original, incluindo novas hipóteses, técnicas de refinamento de *prompts* e análises comparativas de desempenho. Dessa forma, o estudo oferece uma contribuição incremental para

a área, demonstrando diferentes caminhos para aprimoramento da extração automatizada de informações jurídicas.

Ao focar nos aspectos técnicos da engenharia de *prompt* e na análise de seus resultados, o trabalho oferece subsídios para a comunidade de IA aplicada ao Direito. Uma contribuição prática consiste na validação de um método de extração automatizada com características de escalabilidade e precisão aferida. Tal método foi projetado para abordar diretamente o gargalo tecnológico que dificulta o avanço da jurimetria no Brasil: a estruturação de dados a partir de textos jurídicos não estruturados.

A validação dessa metodologia possui implicações diretas para o desenvolvimento de plataformas de ODR. A capacidade de processar grandes volumes de decisões judiciais aumenta a previsibilidade dos resultados, oferecendo aos litigantes subsídios para a tomada de decisão em processos de negociação. Consequentemente, a abordagem contribui para a celeridade processual e para a promoção de soluções consensuais, alinhando-se aos desafios contemporâneos de eficiência e acesso à justiça do sistema judiciário brasileiro.

1.6 ORGANIZAÇÃO DO TRABALHO

Este trabalho organiza-se em seis capítulos. O Capítulo 2 apresenta a Fundamentação Teórica e o Capítulo 3 contextualiza a aplicação da IA no domínio jurídico e trabalhos relacionados. O Capítulo 4 cerne metodológico da pesquisa, expõe o estudo de caso, descrevendo a metodologia do Experimento Principal na Subseção 4.4.1 e, na sequência, a do Experimento Secundário na Subseção 4.4.2. O Capítulo 5, por sua vez, dedica-se à análise dos resultados empíricos: os achados do Experimento Principal são discutidos na Seção 5.1, enquanto os resultados do Experimento Secundário são apresentados na Seção 5.2. O Capítulo 5 prossegue com a análise das limitações do estudo na Seção 5.3 e a investigação de novos paradigmas metodológicos na Seção 5.4. Por fim, o Capítulo 6 apresenta as conclusões, sintetizando os principais achados e delineando perspectivas para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo estabelece a fundamentação teórica do trabalho, percorrendo uma sequência de conceitos que evoluem em complexidade e especialização. Inicia-se com uma visão geral da IA na Seção 2.1, definindo seus conceitos fundamentais. A partir daí, a Seção 2.2 avança para o Aprendizado de Máquina (*Machine Learning*), detalhando como os sistemas aprendem com dados, seguida pela Seção 2.3, cujo foco se volta para as Redes Neurais e o Aprendizado Profundo (*Deep Learning*). Com essa base, a Seção 2.4 explora o Processamento de Linguagem Natural (*Natural Language Processing*) (PLN), essencial para a compreensão e manipulação de textos. A Seção 2.5 aprofunda o tema da IAGen, explorando seus fundamentos, a arquitetura *Transformer*, além de discutir o processo de alinhamento com a intenção humana. Finalmente, o capítulo converge para seu tema central na Seção 2.6, que se dedica integralmente à Engenharia de Prompt, abordando sua teoria, técnicas fundamentais e tendências futuras, consolidando assim a base conceitual para o estudo de caso desenvolvido neste trabalho.

2.1 INTELIGÊNCIA ARTIFICIAL

O conceito de IA apresenta múltiplas formulações teóricas, refletindo a diversidade de perspectivas adotadas pelos pesquisadores da área. Essas formulações abrangem tanto dimensões relacionadas aos processos cognitivos internos quanto aspectos concernentes às manifestações comportamentais externas (Norvig; Russell, 2013). Complementarmente, a mensuração dessa capacidade intelectual artificial pode fundamentar-se na equiparação ao desempenho característico dos seres humanos ou, como alternativa metodológica, na conformidade com um modelo abstrato de racionalidade (Norvig; Russell, 2013).

Uma perspectiva operacional também se manifesta viável, caracterizando a IA mediante a aptidão de sistemas computacionais para executarem atividades que uma criança com inteligência típica conseguiria desempenhar, embora tais atividades ultrapassem as possibilidades de um supercomputador convencional funcionando de forma autônoma (Rosa, 2011). Dessa maneira, quando um sistema computacional adquire a competência para realizar operações tradicionalmente restritas ao domínio cognitivo humano, tal sistema passa a exibir características que o qualificam como inteligente (Rosa, 2011).

A designação como artificial deriva precisamente do caráter deliberado dessa inteligência, construída intencionalmente pelo ser humano, contrastando com a inteligência que se desenvolve de modo espontâneo nos organismos biológicos (Oliveira, 2018). Segundo a interpretação de Peixoto e Silva (2019), a IA constitui uma disciplina especializada no âmbito da ciência da computação, dedicada à replicação de capacidades cognitivas humanas mediante o emprego de sistemas computacionais, abarcando como subdivisões temáticas o aprendizado de máquina e o aprendizado profundo.

Alicerçado nessa concepção de replicação da cognição humana, o filósofo John Searle estabeleceu duas categorias conceituais fundamentais para a compreensão da área (Russell, 2016).

A primeira dessas categorias, identificada como IA Fraca, postula que sistemas computacionais conseguem exclusivamente simular comportamentos característicos da inteligência humana, sem manifestarem estados mentais genuínos (Russell, 2016). Por outro lado, a categoria denominada IA Forte defende que sistemas computacionais poderiam efetivamente desenvolver uma mente autêntica, demonstrando consciência real e faculdades de compreensão equivalentes às aquelas observadas nos seres humanos (Russell, 2016).

Entretanto, conforme observa Russell (2016), a interpretação contemporânea dessas terminologias afastou-se significativamente de suas acepções iniciais, resultando em ambiguidades conceituais. Na conjuntura atual, a expressão IA Forte frequentemente designa investigações científicas orientadas para a construção de uma inteligência artificial dotada de aplicabilidade universal e desempenho equiparável ao humano. Simultaneamente, o termo IA Fraca passou a identificar sistemas desenvolvidos para aplicações delimitadas e específicas, tais como reconhecimento de padrões vocais ou algoritmos de sugestão automatizada, sendo também denominada IA específica (Russell, 2016).

Visando maior precisão terminológica, novas denominações foram introduzidas no campo, destacando-se particularmente a Inteligência Artificial Geral (AGI). Esta terminologia enfatiza o propósito de desenvolver sistemas inteligentes com aplicação irrestrita, cuja versatilidade seja no mínimo equiparável ao espectro de tarefas que os seres humanos conseguem executar (Russell, 2016). A Superinteligência Artificial (ASI) representa um estágio avançado dessa evolução tecnológica, referindo-se a uma categoria de IA que transcende significativamente as capacidades cognitivas humanas (Russell, 2016).

Esse sistema demonstraria competência superior na formulação de decisões de maior qualidade por meio do processamento de volumes ampliados de dados e da previsão de resultados em horizontes temporais mais extensos (Russell, 2016). Seguindo essa mesma trajetória evolutiva, o desenvolvimento dos sistemas conversacionais tem impulsionado a proposição de uma categoria mais refinada de AGI, denominada Inteligência Artificial Geral Conversacional (CAGI). Essa categoria englobaria ferramentas de IA aptas a desempenhar com precisão diversas tarefas que demandam inteligência humana especificamente no âmbito da comunicação dialógica (Marvin et al., 2023).

Em síntese, o conceito de IA é dinâmico e sua abrangência expande-se continuamente. Essa ampliação torna a conformidade ao paradigma da IA uma exigência crescente (Giray, 2023). A adequação mencionada concretiza-se mediante o desenvolvimento de competências voltadas para o aproveitamento das possibilidades oferecidas pelos sistemas de IA, capacitando os usuários a explorarem oportunidades emergentes (Marvin et al., 2023; Giray, 2023).

2.2 APRENDIZADO DE MÁQUINA (*MACHINE LEARNING*)

O aprendizado de máquina, designado em inglês como *machine learning*, caracteriza-se pela construção de algoritmos capazes de aprimorar seu próprio desempenho mediante a análise de dados exemplificativos ou de experiências previamente registradas (Alpaydin, 2020). Em

outras palavras, essa disciplina congrega técnicas computacionais e matemáticas que possibilitam a um sistema adquirir conhecimento diretamente dos dados, desenvolvendo modelos aptos a realizar previsões ou executar decisões quando expostos a informações inéditas (Zhou, 2021).

Essencialmente, o processo consiste no ajuste dos parâmetros de um modelo inicial mediante a utilização de um conjunto de dados de treinamento, visando identificar a configuração de maior eficácia (Alpaydin, 2020). O modelo resultante pode assumir caráter preditivo, orientado para a antecipação de eventos futuros, ou descritivo, voltado para a extração de conhecimento de dados preexistentes, ou ainda manifestar-se como uma combinação de ambas as abordagens (Alpaydin, 2020).

Dessa forma, o aprendizado de máquina estabelece-se como uma subárea da IA, fundamentando-se na premissa de que a inteligência de um sistema, particularmente em contextos caracterizados por transformações constantes, encontra-se vinculada à sua aptidão para aprender (Alpaydin, 2020). Ao conferir a um sistema autonomia para adaptar-se, diminui-se a necessidade de predição e codificação manual de soluções para todas as eventualidades, ampliando a eficiência operacional desse sistema (Alpaydin, 2020).

O processo de aprendizado pode ser classificado em diferentes metodologias, destacando-se como principais e mais pertinentes para este trabalho o aprendizado supervisionado e o aprendizado não supervisionado (Alpaydin, 2020). A metodologia supervisionada tem como propósito inferir uma correlação a partir de dados de treinamento previamente rotulados, nos quais cada dado de entrada é empregado para predizer um novo dado de saída correspondente (Alpaydin, 2020).

Em contraste, a metodologia não supervisionada opera na ausência de rótulos predefinidos, analisando exclusivamente a estrutura intrínseca dos dados de entrada com o propósito de modelar a distribuição subjacente dos dados, visando descobrir padrões, agrupamentos ou outras regularidades estruturais (Alpaydin, 2020). A essas abordagens soma-se o aprendizado auto-supervisionado, um subconjunto do aprendizado não supervisionado que objetiva adquirir características discriminativas a partir de dados não rotulados, prescindindo de anotações humanas (Gui et al., 2024).

A concepção central dessa técnica consiste em gerar pseudo-rótulos de forma automatizada com base em atributos específicos dos próprios dados de entrada, formulando uma tarefa de pretexto para que o algoritmo resolva (Gui et al., 2024). No contexto dos grandes modelos de linguagem, essa técnica assume relevância ao empregar tarefas de pretexto, como mascaramento ou remoção de ruído em grandes quantidades de texto, permitindo que o modelo adquira representações linguísticas de maneira autônoma (Ahmed et al., 2023).

Além dessas metodologias, existem estratégias de aprendizado mais especializadas, exemplificadas pelo aprendizado ativo. Essa técnica visa aprimorar a eficiência da rotulagem de dados mediante a identificação dos dados não rotulados mais relevantes para serem anotados por especialistas humanos, geralmente fundamentando-se em métricas de incerteza do modelo (Diao et al., 2023). As abordagens de aprendizado de máquina, especialmente o aprendizado supervisionado, possuem relevância direta para esta investigação científica.

Conforme estabelecido, essa técnica requer dados de entrada estruturados e previamente rotulados para treinar modelos preditivos (Alpaydin, 2020). No domínio jurídico, onde a informação primária manifesta-se de forma textual e não estruturada, a construção de tais conjuntos de dados representa um desafio metodológico. Diante dessa problemática, este estudo explora como as capacidades dos LLMs e detalhados na Seção 2.5, podem ser aproveitadas para superar essa limitação.

Especificamente, a aprendizagem no contexto, designada em inglês como *in-context learning* e orientada pela engenharia de prompt, discutida na Seção 2.6, é empregada como um método para automatizar a extração de fatores de documentos jurídicos. A aplicação prática dessa metodologia, na qual um LLM é utilizado para estruturar dados que, subsequentemente, alimentarão um modelo supervisionado, constitui um dos cerne do projeto CONCIL-IA, cuja metodologia e resultados são apresentados nos Capítulos 4 e 5.

2.3 REDES NEURAI E APRENDIZADO PROFUNDO (*DEEP LEARNING*)

As redes neurais constituem sistemas computacionais cujo projeto se fundamenta na estrutura do cérebro biológico, sendo compostas por unidades interconectadas que realizam trocas informacionais por meio de canais direcionais, propagando sinais ao longo de toda a arquitetura (Norvig; Russell, 2013). O elemento central que as caracteriza reside em sua aptidão para aprender, uma vez que a intensidade de cada conexão, denominada "peso", é ajustável e se modifica dinamicamente com base na experiência acumulada (Norvig; Russell, 2013).

Embora a formalização matemática seja complexa, é possível afirmar de maneira simplificada que as redes neurais constituem o mecanismo principal para o aprendizado profundo (Peixoto; Silva, 2019). A partir dessa tecnologia, os LLMs, são desenvolvidos, empregando redes neurais com múltiplos parâmetros para executar suas tarefas computacionais (Marvin et al., 2023).

O aprendizado profundo, designado em inglês como *Deep Learning*, constitui uma subárea do aprendizado de máquina que emprega redes neurais dotadas de múltiplas camadas ocultas para solucionar tarefas com maior complexidade (Zhou, 2021). Bishop e Bishop (2023) o definem como um sistema fundamentado em modelos computacionais denominados redes neurais, originalmente concebidos a partir dos mecanismos de aprendizado e processamento informacional observados no cérebro humano.

Na conjuntura atual, o aprendizado de máquina consolidou-se como uma das áreas tecnológicas mais importantes e de expansão mais acelerada, com aplicações tornando-se comuns (Bishop; Bishop, 2023). Soluções derivadas de dados substituem gradualmente os algoritmos tradicionais projetados manualmente. Essa transformação ocorre porque a profundidade da rede, isto é, o elevado número de camadas, é o fator que possibilita ao modelo aprender de forma mais eficaz em comparação com redes dotadas de apenas uma camada (Zhou, 2021).

Para otimizar o desempenho do modelo, a estratégia mais eficiente consiste em ampliar sua profundidade mediante a adição de mais camadas, em detrimento da simples expansão

do número de neurônios em uma única camada (Zhou, 2021). A vantagem dessa abordagem baseia-se na concepção de que múltiplas camadas viabilizam um processamento hierárquico da informação, no qual os dados são transformados sequencialmente em níveis progressivamente mais elevados de abstração (Zhou, 2021).

Esse encadeamento de operações sucessivas confere ao modelo uma capacidade de aprendizado mais robusta e eficaz (Zhou, 2021). O princípio de escalar as dimensões dos modelos, tanto em relação ao número de parâmetros quanto ao volume de dados de treinamento, tem constituído um ingrediente importante nas transformações recentes do processamento de linguagem natural (Kojima et al., 2022). Os LLMs contemporâneos são construídos sobre uma arquitetura de múltiplas camadas do tipo *transformer* (Ahmed et al., 2023) que será melhor detalhado na Subseção 2.5.2 e seguintes.

O impacto do aprendizado profundo reside em sua capacidade de gerar avanços práticos em diversas áreas do conhecimento. Na medicina, por exemplo, modelos de aprendizado profundo já superaram a precisão diagnóstica de dermatologistas na identificação de câncer de pele (Brinker et al., 2019). Na biologia, tais modelos solucionaram o desafio de prever a estrutura tridimensional de proteínas, um problema científico que perdurou por meio século (Jumper et al., 2021).

Adicionalmente, essa tecnologia impulsionou o desenvolvimento da IAGen, criando modelos capazes de sintetizar imagens fotorrealistas indistinguíveis de fotografias autênticas e de desenvolver grandes modelos de linguagem com capacidades avançadas que são considerados os primeiros indícios da AGI (Bishop; Bishop, 2023). Essa convergência de avanços técnicos e aplicações práticas evidencia a centralidade do aprendizado profundo na transformação contemporânea da IA, estabelecendo novos paradigmas em múltiplos domínios do conhecimento científico e tecnológico.

2.4 PROCESSAMENTO DE LINGUAGEM NATURAL - PLN (*NATURE LANGUAGE PROCESSING - NLP*)

Embora o PLN admita múltiplas formulações conceituais, todas elas convergem para a noção de armazenamento e manipulação de dados linguísticos mediante recursos computacionais (Rosa, 2011). De maneira sucinta, Rosa (2011) compreende o PLN como a aptidão de um sistema computacional para processar e compreender a linguagem humana cotidiana. Peixoto e Silva (2019), por sua vez, o caracterizam como um domínio interdisciplinar que integra a Linguística, a Ciência Cognitiva, o Processamento de Informações e a IA, objetivando capacitar computadores ao processamento inteligente da linguagem humana.

A grande quantidade de informação disponível na rede mundial de computadores, majoritariamente expressa em linguagem natural, representa um desafio para agentes de IA que buscam extrair conhecimento estruturado (Norvig; Russell, 2013). Para lidar com a complexidade e a ambiguidade características da comunicação humana, são empregados modelos de linguagem, os quais estimam a probabilidade de ocorrência de diferentes sequências de palavras (Norvig;

Russell, 2013). Um modelo de linguagem, em sua essência, pode ser definido como um sistema que busca calcular o quão provável ou coerente é determinada sequência de palavras (Kojima et al., 2022).

Em contraste com a linguagem natural, as linguagens formais, tais como as linguagens de programação, possuem modelos precisamente definidos, regidos por uma gramática que determina quais sequências de símbolos são válidas (Norvig; Russell, 2013). As linguagens formais dispõem de regras semânticas que estabelecem um significado único para cada expressão, o que acaba por eliminar as ambiguidades características da comunicação humana (Norvig; Russell, 2013).

A linguagem natural, em contrapartida, é marcada por uma ambiguidade na qual uma única sentença pode admitir múltiplos significados dependendo do contexto situacional (Norvig; Russell, 2013). Tal polissemia constitui um desafio para o PLN, cuja tarefa consiste em mapear a expressão ambígua para uma representação interna processável por sistemas computacionais (Rosa, 2011). A escala e a evolução constante da linguagem natural tornam difícil a criação de um modelo definitivo e completo (Norvig; Russell, 2013).

Portanto, qualquer modelo de linguagem natural é, por necessidade, uma aproximação que busca capturar padrões estatísticos, em detrimento de uma representação exata e formal da língua (Norvig; Russell, 2013). Essa característica evidencia a natureza probabilística e adaptativa dos sistemas de PLN contemporâneos, que se fundamentam em vastos conjuntos de dados para aprimorar continuamente sua capacidade de processamento linguístico. A evolução desses sistemas reflete o esforço contínuo para superar as limitações inerentes à formalização computacional da comunicação humana.

2.5 INTELIGÊNCIA ARTIFICIAL GENERATIVA (*GENERATIVE ARTIFICIAL INTELLIGENCE*)

2.5.1 Fundamentos da Inteligência Artificial Generativa

O campo da IAGen representa uma vertente específica do aprendizado de máquina, caracterizada pela capacidade de criar novas amostras de dados que preservam as propriedades estatísticas do conjunto original de treinamento, abrangendo modalidades como imagens, textos e áudios (Luk, 2023). Segundo a conceituação proposta por Feuerrigel et al. (2024), trata-se da implementação prática de modelagem generativa através de arquiteturas de aprendizado de máquina, notadamente redes neurais profundas, visando a síntese de instâncias inéditas fundamentadas nos padrões previamente apreendidos.

Essa metodologia distingue-se dos modelos discriminativos, projetados para realizar tarefas de classificação ou diferenciação entre categorias em conjuntos de dados preexistentes (Luk, 2023). Os Grandes Modelo de Linguagem (LLMs), configuram a tecnologia que sustenta a IAGen voltada para a produção textual (Ahmed et al., 2023). Estruturados por meio de redes neurais contendo bilhões de parâmetros ajustáveis, esses sistemas recebem pré-treinamento em

volumes de dados textuais provenientes de múltiplas fontes, o que lhes confere uma compreensão aprofundada sobre estrutura linguística, dimensões semânticas e elementos contextuais (Giray, 2023; Marvin et al., 2023).

O princípio operacional dos LLMs baseia-se na predição auto-regressiva da próxima unidade textual (Ahmed et al., 2023), conhecida como *token*, que corresponde à menor unidade de texto processável pelo modelo, podendo manifestar-se como palavra completa, fragmento lexical ou caractere isolado, sempre considerando a sequência textual precedente (Google Developers, 2025). A série *Generative Pre-trained Transformer* (GPT) exemplifica de forma paradigmática essa tecnologia, sendo o ChatGPT uma demonstração da capacidade desses sistemas em gerar conteúdo simultaneamente coerente e contextualmente adequado (Feuerriegel et al., 2024).

A construção de tais sistemas, entretanto, exigiu a superação de um obstáculo, qual seja, capacitar os modelos para interpretar o significado e as nuances relacionais entre unidades lexicais (Luk, 2023). Entre os fundamentos que possibilitaram essa evolução, destaca-se a introdução dos *word embeddings*, ou incorporações de palavras, técnica mediante a qual palavras são representadas como vetores numéricos dispostos em espaço multidimensional, de modo que a proximidade espacial e a orientação vetorial correspondam a relações de natureza semântica, sendo também uma técnica de PLN (Luk, 2023).

2.5.2 A Arquitetura Transformer

Ultrapassando a inovação representada pelos embeddings, a arquitetura Transformer, originalmente proposta por Vaswani et al. (2017) no trabalho intitulado "Attention Is All You Need", tornou-se a estrutura dos sistemas contemporâneos de IAGen (Luk, 2023). O elemento distintivo dessa arquitetura reside no mecanismo de autoatenção (*self-attention*), que capacita o modelo a atribuir pesos de importância a todas as palavras presentes em uma sequência textual, viabilizando a interpretação contextual de cada termo de forma individualizada (Luk, 2023).

A contribuição do Transformer manifestou-se na modificação da centralidade das Rede Neural Recorrente (RNN) e convolucionais, que predominavam até então, estabelecendo como fundamento os mecanismos de atenção (Vaswani et al., 2017). As arquiteturas recorrentes processam dados de maneira sequencial, token por token, impedindo o processamento paralelo do contexto integral durante a fase de treinamento e dificultando a captura de dependências entre tokens situados em posições distantes na sequência (Vaswani et al., 2017).

Em oposição a essa limitação, o mecanismo de atenção característico do Transformer possibilita que o modelo avalie a relevância das palavras considerando o contexto completo, diminuindo a necessidade de processamento sequencial (Vaswani et al., 2017). Essa abordagem estabelece conexões diretas para o fluxo informacional entre quaisquer duas palavras na sequência, independentemente da distância que as separa (Vaswani et al., 2017). Na aplicação prática, isso facilita o aprendizado das relações entre termos localizados no início de um texto extenso e conceitos apresentados em suas porções finais (Vaswani et al., 2017).

Diferentemente do paradigma de processamento sequencial tradicional, o mecanismo

de autoatenção confere ao modelo a capacidade de resolver ambiguidades linguísticas que anteriormente representavam desafios (Luk, 2023). Essa transformação resultou no desenvolvimento de modelos de qualidade, passíveis de treinamento com grau maior de paralelização e que demandam períodos reduzidos para completar o processo de treinamento (Vaswani et al., 2017).

A estrutura da arquitetura Transformer organiza-se segundo um arranjo de codificador-decodificador (*encoder-decoder*) (Vaswani et al., 2017). O codificador realiza o processamento da sequência de entrada em sua integralidade, produzindo representações contínuas que encapsulam informações contextuais (Vaswani et al., 2017). O decodificador, por sua vez, utiliza essas representações para produzir a sequência de saída de maneira auto-regressiva, gerando um token por iteração (Vaswani et al., 2017).

O elemento central, presente tanto na estrutura do codificador quanto do decodificador, denomina-se Atenção Multi-Cabeças (*Multi-Head Attention*) (Vaswani et al., 2017). Em substituição a um único mecanismo de atenção, essa estratégia opera como se o modelo dispusesse de múltiplas perspectivas simultâneas de análise, permitindo a focalização em distintos tipos de relações, sejam sintáticas, semânticas ou de outra natureza, de forma concorrente, o que produz uma compreensão contextual abrangente (Vaswani et al., 2017).

A arquitetura Transformer, após receber pré-treinamento em um *corpus* textual, transforma-se na base para um procedimento de especialização denominado *fine-tuning*, ou ajuste fino (Howard; Ruder, 2018; Radford et al., 2018). Esse procedimento adapta os parâmetros de um modelo, inicialmente calibrados mediante treinamento não supervisionado, direcionando-os para uma tarefa-alvo específica através do emprego de seu conjunto de dados rotulados correspondente (Radford et al., 2018). A finalidade consiste em transferir o conhecimento linguístico de natureza geral para as especificidades de uma nova tarefa, proporcionando convergência e desempenho mesmo diante de conjuntos de dados de dimensões reduzidas (Howard; Ruder, 2018).

O desafio dessa metodologia reside no equilíbrio necessário: um ajuste excessivo pode desencadear o "esquecimento catastrófico" (*catastrophic forgetting*), comprometendo o conhecimento previamente adquirido no pré-treinamento, enquanto um ajuste conservador pode resultar em convergência lenta e no fenômeno de "sobreajuste" (*overfitting*) (Howard; Ruder, 2018), caracterizado pela aprendizagem dos detalhes e ruídos dos dados em um nível que o modelo diminui a capacidade de generalização para dados novos (Zhou, 2021).

2.5.3 Alinhamento com Intenção Humana

Avanços em escala e capacidade de raciocínio, contudo, expuseram um desafio no desenvolvimento de LLMs: o desalinhamento entre o objetivo de pré-treinamento e a intenção do usuário (Ouyang et al., 2022). Observou-se que o incremento de escala não torna os modelos mais capazes de seguir as intenções de um usuário, podendo ainda gerar resultados inverídicos, tóxicos ou inúteis (Ouyang et al., 2022).

Para resolver esse problema, Ouyang et al. (2022) propuseram uma metodologia para alinhar os LLMs com a intenção do usuário por meio de *feedback* humano, resultando nos

modelos conhecidos como InstructGPT. O processo consistiu em três etapas: primeiro, o GPT-3 foi ajustado de forma supervisionada pelo *Supervised Fine-Tuning* (Ajuste Fino Supervisionado) (SFT) com base em um conjunto de demonstrações escritas por humanos sobre como responder a diversas instruções (Ouyang et al., 2022).

Em seguida, um *Reward Model* (Modelo de Recompensa) (RM) foi treinado a partir de um conjunto de dados de comparações, onde humanos classificam diferentes respostas do modelo para um mesmo *prompt*, da melhor para a pior (Ouyang et al., 2022). Finalmente, esse modelo de recompensa é usado como uma função de "recompensa" para otimizar ainda mais o modelo SFT por meio do método de *Reinforcement Learning from Human Feedback* (Aprendizagem por Reforço com Feedback Humano) (RLHF) (Ouyang et al., 2022).

Os resultados foram positivos: avaliadores humanos preferiram as respostas do modelo InstructGPT de 1,3 bilhão de parâmetros em detrimento das respostas do GPT-3 de 175 bilhões, apesar de ser 100 vezes menor (Ouyang et al., 2022). Adicionalmente, os modelos InstructGPT demonstraram ser mais verdadeiros, reduzindo a taxa de alucinações, fenômeno caracterizado pela invenção de fatos, e geraram resultados menos tóxicos (Ouyang et al., 2022). Este trabalho demonstrou que o ajuste fino com feedback humano constitui um método para alinhar o comportamento de grandes modelos de linguagem com a intenção humana (Ouyang et al., 2022).

2.5.4 Técnicas de Prompting e a Inteligência Artificial Generativa

Embora o incremento de escala dos modelos, exemplificado pelo GPT-3, tenha desbloqueado a capacidade de aprendizagem no contexto, pesquisadores observaram que essa habilidade falhava em tarefas que demandavam raciocínio complexo envolvendo múltiplos passos, tais como problemas matemáticos ou de senso comum (Wei et al., 2022). Wei et al. (2022) demonstraram que essa limitação poderia ser superada mediante uma técnica denominada *prompting* de *Chain-of-Thought* (Cadeia de Pensamento) (CoT) (Wei et al., 2022).

Em substituição à mera apresentação de um problema e sua resposta final ao modelo, a técnica de CoT consiste em fornecer, no *prompt*, alguns exemplos que incluem os passos intermediários do raciocínio conducentes à solução, emulando um processo de pensamento humano (Wei et al., 2022). Os autores identificaram que essa capacidade de raciocínio de forma satisfatória somente se manifesta em modelos de grande escala, revelando-se como uma propriedade emergente: enquanto modelos menores não se beneficiam ou até apresentam desempenho deteriorado, aqueles com mais de 100 bilhões de parâmetros demonstram melhoria (Wei et al., 2022).

Essa melhoria manifesta-se especialmente em tarefas de raciocínio aritmético, de senso comum e simbólico (Wei et al., 2022). A descoberta revelou que as capacidades de raciocínio de grandes modelos de linguagem podem ser desbloqueadas sem necessidade de novo treinamento, simplesmente ao guiá-los para pensar passo a passo por meio de exemplos fornecidos no *prompt* (Wei et al., 2022).

2.6 ENGENHARIA DE PROMPT: TEORIA E TÉCNICAS

2.6.1 Definindo Engenharia de Prompt

O campo da engenharia de *prompts* dedica-se à elaboração e ao aperfeiçoamento de diretrizes instrucionais que viabilizam a utilização de LLMs em múltiplos contextos aplicados e áreas investigativas (White et al., 2023). Tais diretrizes, conhecidas como *prompts*, caracterizam-se como inputs textuais transmitidos ao sistema linguístico com o propósito de orientar o resultado gerado (Marvin et al., 2023).

A construção de um *prompt* geralmente integra diversos componentes que, articulados entre si, conferem a arquitetura e o substrato contextual para que o sistema produza respostas alinhadas às instruções. Conquanto distintos pesquisadores apresentem interpretações variadas acerca dos constituintes primordiais (Giray, 2023; Lee et al., 2024), uma formulação proposta por Lin (2024) contempla cinco componentes essenciais.

Tais componentes primordiais englobam: Audiência, Persona, Instrução Específica, Contexto e Especificação de Saída (Lin, 2024). O elemento Audiência determina o segmento destinatário da resposta, aspecto relevante para assegurar adequação do material a distintos perfis de usuários (Marvin et al., 2023). Adicionalmente, a Persona delimita o papel a ser desempenhado pelo modelo, configurando-o como tutor, revisor ou outra figura, adequando a resposta conforme demandas específicas (White et al., 2023).

O componente Instrução Específica requer a delimitação com precisão da atividade a ser realizada, sendo aconselhável desenvolver instruções detalhadas e evitar formulações imprecisas (Gruda, 2024). Paralelamente, o Contexto disponibiliza dados adicionais que capacitam o modelo a apreender a atividade de maneira aprofundada (Shin et al., 2023). Finalmente, a Especificação de Saída detalha as propriedades esperadas do resultado, englobando estilo, tonalidade e formato de apresentação, como tabelas ou categorias específicas de arquivo (Lin, 2024).

Para exemplificar tal estrutura no domínio da engenharia biomédica, a instrução poderia requisitar uma dissertação concernente ao papel da nanotecnologia na administração direcionada de fármacos para terapia oncológica (Giray, 2023). O contexto exploraria aplicações nanotecnológicas na área, os dados de entrada abordariam sistemas de distribuição fundamentados em nanopartículas, enquanto o indicador de saída especificaria um texto com aproximadamente mil e quinhentas palavras (Giray, 2023).

2.6.2 A Importância da Engenharia de Prompt

A disciplina de engenharia de *prompts* exerce uma função na melhoria do desempenho de LLMs. A implementação dessa especialidade possibilita o aprimoramento da performance dos LLMs, a identificação de possibilidades inovadoras de aplicação e a preservação de recursos computacionais (Marvin et al., 2023).

No panorama tecnológico contemporâneo, caracterizado por transformações contínuas,

a competência em engenharia de *prompts* consolidou-se progressivamente como requisito para o uso desses sistemas (Giray, 2023). O princípio dessa prática encontra-se na elaboração de *prompts* que conduzam o modelo na produção de respostas com maior acurácia e adequação ao contexto, viabilizando a exploração das capacidades inerentes aos LLMs (Marvin et al., 2023).

Essa engenharia constitui um paradigma programático que capacita profissionais a fornecer diretrizes claras e automatizar procedimentos mediante entradas textuais (Giray, 2023). Ao facilitar o desenvolvimento de respostas contextualmente adequadas e com nuances apropriados, a engenharia de *prompts* contribui para o progresso investigativo em inteligência artificial, particularmente no domínio da CAGI (Marvin et al., 2023).

Complementarmente, essa disciplina opera como ferramenta para personalizar e modelar as saídas produzidas pelos sistemas, harmonizando-as com requisitos qualitativos e quantitativos particulares (Giray, 2023). O processo igualmente auxilia na elaboração de aplicações fundamentadas em LLMs, permitindo que desenvolvedores compreendam o comportamento e as capacidades desses modelos, orientando-os para produzir informações verídicas e informativas (Marvin et al., 2023).

Sob viés prospectivo, a engenharia de *prompt* posiciona-se para desempenhar função relevante na exploração do potencial dos LLMs, estabelecendo-se como elemento para incrementar eficiência e otimizar operações em diversos domínios, consequentemente gerando oportunidades profissionais para indivíduos com competências especializadas na área (Marvin et al., 2023).

Essa trajetória demanda, entretanto, aplicação consciente e crítica da técnica, exigindo reconhecimento das limitações inerentes aos modelos. Sistemas como o ChatGPT não asseguram precisão absoluta, uma vez que suas respostas fundamentam-se em padrões aprendidos em bases de dados que podem não contemplar todos os fatores complexos de determinado problema (Giray, 2023). Nesse sentido, torna-se essencial avaliar o conteúdo gerado, especialmente considerando o fenômeno da alucinação, caracterizado por respostas confiantes porém desprovidas de respaldo factual (Giray, 2023).

A abordagem consiste em combinar as saídas do modelo com a expertise do usuário e pesquisa complementar para incrementar a confiabilidade das análises (Giray, 2023). Destarte, o domínio desta habilidade não reside exclusivamente na capacidade de otimizar fluxos de trabalho (Giray, 2023), mas fundamentalmente na navegação por desafios como ambiguidade e viés, o que contribui para o desenvolvimento de interações responsáveis com os modelos de IA (Giray, 2023).

2.6.3 Técnicas de Engenharia de Prompt

Estabelecidos os componentes que constituem um *prompt*, torna-se imperativo explorar as metodologias e estratégias que viabilizam a aplicação desses componentes. A engenharia de *prompt* consolidou-se como conjunto relevante de competências técnicas que demanda conhecimento técnico para aperfeiçoar a interação com os LLMs (Marvin et al., 2023).

Serão delineadas algumas técnicas, iniciando com abordagens estruturadas, exempli-

ficadas pelos padrões de *prompt*, e progredindo para estratégias mais dinâmicas e interativas (White et al., 2023). A metodologia de padrões de *prompt*, inspirada diretamente no conceito de padrões de software, objetiva documentar e catalogar soluções para problemas recorrentes que emergem durante a interação com LLMs (White et al., 2023).

Esses padrões estruturam-se em cinco categorias principais, cada qual focada em aspecto específico da interação (White et al., 2023). A primeira categoria, denominada Semântica de Entrada (*Input Semantics*), aborda aspecto fundamental da interação concernente à forma como o LLM apreende a entrada do usuário e a converte em elemento utilizável para produzir uma saída (White et al., 2023).

Dentro desta categoria, destaca-se o padrão de Criação de Metalinguagem (*Meta Language Creation*), que se concentra em possibilitar ao usuário a criação de linguagem personalizada para o LLM (White et al., 2023). Essa técnica revela-se útil quando a linguagem padrão, como o inglês, mostra-se inadequada ou ineficiente para expressar ideias ou estruturas de dados de forma concisa, capacitando o usuário a definir sintaxe e semântica novas para a conversação (White et al., 2023).

A categoria de Customização de Saída (*Output Customization*) concentra-se em restringir ou adaptar o tipo, formato e estrutura da resposta produzida pelo modelo. Dentro dela, identificam-se padrões como o *Output Automater*, que instrui o LLM a elaborar scripts para automatizar atividades, e o *Template*, que compele a saída a seguir modelo pré-estabelecido (White et al., 2023).

Adicionalmente, a categoria de Identificação de Erros (*Error Identification*) visa identificar e solucionar erros na saída do LLM (White et al., 2023). A categoria de Melhora de *Prompt* (*Prompt Improvement*) foca em aprimorar a qualidade da entrada do usuário. Padrões como o *Question Refinement* fazem com que o LLM sugira versão melhorada da pergunta do usuário, enquanto o *Alternative Approaches* o instrui a sugerir maneiras alternativas de realizar atividade (White et al., 2023).

Finalmente, as categorias de Interação e Controle de Contexto gerenciam o fluxo da conversação. A categoria de Interação inclui o padrão de Interação Invertida (*Flipped Interaction*), onde o LLM, em vez do usuário, passa a conduzir a conversa, formulando perguntas para obter as informações necessárias para atingir objetivo específico, como gerar script de implantação (White et al., 2023). A categoria de Controle de Contexto (*Context Control*) permite ao usuário especificar qual contexto o LLM deve considerar ou ignorar, assegurando que a conversa permaneça focada e relevante (White et al., 2023).

Para aplicações mais diretas, especialmente no contexto da escrita acadêmica, podem ser empregadas técnicas de *prompting* mais focadas na geração de conteúdo específico (Giray, 2023). O *prompt* instrutivo orienta a produção textual para tarefa determinada, exemplificado por solicitações como "Resuma os avanços recentes em engenharia de tecidos para regeneração de órgãos, destacando suas potenciais aplicações em engenharia biomédica" (Giray, 2023).

O *prompt* de sistema oferece ponto de partida ou contexto inicial, enquanto o *prompt* de pergunta-resposta estrutura a escrita em torno de questões de pesquisa específicas (Giray,

2023). O *prompt* contextual direciona o foco para aspectos particulares de tópico mediante fornecimento de informações adicionais, e o *prompt* misto combina múltiplos elementos para orientar a produção textual de maneira integrada (Giray, 2023).

Além das anteriores, conforme introduzido na discussão sobre o GPT-3, a técnica de *few-shot prompting* representa uma das estratégias da engenharia de *prompt*. Ela explora a capacidade de aprendizagem no contexto dos LLMs, permitindo guiar o modelo com alguns exemplos diretamente no *prompt* (Ahmed et al., 2023). Sua principal vantagem reside na capacidade de adaptar o modelo a novas tarefas sem necessidade de processos de ajuste fino, custosos e demandantes de grandes conjuntos de dados rotulados (Ahmed et al., 2023).

Em contraste, o *zero-shot prompting*, também apresentado anteriormente, testa a capacidade de generalização do modelo para tarefas onde nenhum exemplo é fornecido, dependendo apenas de instrução clara (Kojima et al., 2022). A eficácia dessa abordagem foi ampliada pela técnica Zero-shot-CoT, que demonstrou que a simples adição da frase "Vamos pensar isso passo a passo" ("*Let's think step by step*") pode induzir o modelo a externalizar seu raciocínio, melhorando o desempenho em tarefas sem qualquer exemplo prévio (Kojima et al., 2022).

Paralelamente, a técnica CoT, exposta anteriormente, representa uma das estratégias para problemas que demandam raciocínio em múltiplos passos (Diao et al., 2023). Ao invés de fornecer apenas a resposta final nos exemplos de *few-shot*, o CoT enriquece os *prompts* com passos intermediários do raciocínio. Essa abordagem "ensina" o modelo a "pensar passo a passo", decompondo problemas complexos e gerando, supostamente, uma cadeia de lógica que conduz a solução mais precisa (Diao et al., 2023).

2.6.4 Tendências Futuras

A interdependência entre a qualidade das entradas (*prompts*) e das saídas (respostas) dos LLMs tem inspirado diversas iniciativas de pesquisa que definem as tendências futuras da engenharia de *prompt* (Marvin et al., 2023). Uma das direções é o desenvolvimento de ferramentas e técnicas de engenharia de *prompt* mais sofisticadas e específicas de domínio para interação com os LLMs (Marvin et al., 2023).

Adicionalmente, à medida que a engenharia de *prompt* se consolida como paradigma para a CAGI, abrem-se oportunidades de pesquisa para utilização de *prompts* na melhoria da explicabilidade, interpretabilidade e confiabilidade das aplicações, buscando garantir que os LLMs se comportem de maneira rastreável (Marvin et al., 2023). Nesse sentido, a abordagem de padrões de *prompt* exposta anteriormente é entendida como avanço, mas ainda insuficiente, com expectativa de que a evolução natural dessa catalogação seja sua integração em linguagem de padrões de expressão, capaz de guiar usuários de forma mais eficaz (White et al., 2023).

Paralelamente à busca por automação, a necessidade de uma CAGI Responsável emerge como pilar para o futuro da área, abrindo oportunidades de pesquisa na utilização da engenharia de *prompt* para melhorar a confiabilidade e a segurança dos LLMs (Marvin et al., 2023). Isso implica engenharia de *prompt* que não apenas busca eficácia, mas também mitigação de riscos

inerentes aos modelos, como reforço de viés, falta de contexto e alucinações (Giray, 2023).

A ascensão da multimodalidade, que buscará combinar tecnologias de PLN com visão computacional, robótica e IA quântica, exigirá o desenvolvimento de novas formas de *prompts* multimodais (Marvin et al., 2023). A aplicação de padrões de *prompt* também deverá generalizar-se para outros domínios, desde jogos educacionais até entretenimento, para além de seu contexto original em engenharia de software (White et al., 2023).

No entanto, o avanço nessas fronteiras enfrenta desafios práticos, como os custos de uma *Application Programming Interface* (API)³ para conduzir experimentos com modelos avançados como o GPT-4 e as dificuldades em garantir reprodutibilidade dos resultados, especialmente quando o acesso a modelos específicos é descontinuado (Diao et al., 2023).

Por fim, o capítulo estabeleceu a base conceitual da pesquisa, partindo dos conceitos da IA até as técnicas da Engenharia de *Prompt*. Este arcabouço será utilizado, primeiramente, para contextualizar a pesquisa no cenário acadêmico e normativo, por meio da análise de trabalhos relacionados e do quadro regulatório da matéria, conforme será exposto no Capítulo 3. Posteriormente, essa mesma fundamentação teórica será aplicada de forma empírica, servindo como alicerce para o estudo de caso detalhado nos Capítulos 4 e 5, onde se investiga a extração automatizada de fatores jurídicos no âmbito do projeto CONCIL-IA.

³ Uma API constitui mecanismo que atua como contrato de comunicação entre dois sistemas de software distintos. Ela define conjunto de regras e protocolos que ditam como um componente (o cliente) pode solicitar informações ou funcionalidades de outro (o servidor) e como este deve formatar sua resposta. Essencialmente, as APIs permitem que diferentes aplicações interajam e troquem dados de maneira padronizada e segura, sem que uma precise compreender a complexidade interna ou o código-fonte da outra (Amazon Web Services, 2025).

3 TRABALHOS RELACIONADOS E APLICAÇÕES NO BRASIL

Este capítulo contextualiza a aplicação da IA no domínio jurídico, analisando o panorama de seu uso e a estrutura regulatória que o governa. O percurso inicia-se na Seção 3.1, com uma análise do uso da IAGen no domínio jurídico, mapeando as principais iniciativas e pesquisas internacionais e nacionais. Em seguida, a Seção 3.2 aprofunda a análise do arcabouço normativo, examinando tanto as resoluções do CNJ que regulam o uso da IA no Poder Judiciário na Subseção 3.2.1, quanto o debate legislativo, em uma análise comparativa entre o AI Act europeu e o Projeto de Lei 2338/2023 na Subseção 3.2.2. Dessa forma, o capítulo desenvolve o cenário jurídico e institucional sobre o qual a engenharia de prompt no projeto CONCIL-IA, detalhado posteriormente, se desenvolve e busca oferecer suas contribuições.

3.1 USO DE IA GENERATIVA NO DOMÍNIO JURÍDICO

Na China (Zheng, 2020; Yaohui; Hao, 2020), tribunais inteligentes como o Court 206 e os Tribunais de Internet de Pequim e Hangzhou implementam um arcabouço tecnológico voltado à geração automática de fundamentações judiciais. Por meio de algoritmos estruturados em aprendizagem profunda (*deep learning*), o sistema efetua uma análise dos dados para extrair vetores semânticos presentes em narrativas fáticas, elementos de condenação e dispositivos legais. A plataforma executa correspondência por similaridade semântica entre fatos e normas, replicando a lógica decisória judicial, além de produzir minutas sentenciasais apoiadas em precedentes de litígios anteriores. Embora vise eficiência e uniformidade, o sistema carece de elucidação quanto aos mecanismos internos que conduzem às conclusões apresentadas, suscitando questionamentos sobre sua transparência.

Na Holanda (Nakad-weststrate et al., 2015), o sistema e-Court, tribunal privado online do país, lançado em 2010, automatiza a emissão de sentenças arbitrais, especificamente em processos de cobrança de dívidas onde ocorre ausência de contestação pela parte demandada (hipóteses de revelia). Operando mediante um sistema baseado em regras (*Rule-Based System*) e fundamentado em um conjunto de instruções lógicas previamente programadas seguindo estrutura condicional, sem capacidade de aprendizagem ou adaptação a partir de novos dados (Norvig; Russell, 2013), a ferramenta gera veredito arbitral final de forma automatizada, prescindindo de intervenção humana após o início processual, conquanto a admissão de novos casos no sistema seja precedida por verificação humana para resguardar a integridade e conformidade (Nakad-weststrate et al., 2015).

Na Argentina (Estevez; Linares; Fillottrani, 2020), o sistema Prometea, inaugurado em novembro de 2017 pela Procuradoria da Cidade de Buenos Aires, emprega aprendizado de máquina (*machine learning*) para automatizar tramitações judiciais. O sistema demonstra capacidade de produzir conteúdo jurídico preliminar ao prever a solução mais adequada para determinado litígio, recomendar o modelo de parecer correspondente e elaborar a redação integral do documento, fundamentando-se no acesso a repositório contendo mais de trezentos

mil documentos judiciais.

Em Singapura (State Courts of Singapore, 2025), o judiciário nacional implementou recentemente uma ferramenta baseada em IAGen para assistir usuários do Tribunal de Pequenas Causas (*Small Claims Tribunal*), desenvolvida para produzir sínteses de documentos processuais, facilitando o acesso e a compreensão das informações contidas nos autos pelos jurisdicionados.

Dal Pont et al. (2023) empregaram o GPT-4 para elaborar resumos e ementas de decisões tributárias, objetivando facilitar o acesso informacional em área caracterizada por grande volume de litígios. Todavia, a precisão factual emerge como obstáculo nesse contexto. Investigação conduzida por Deroy et al. (2023) sobre a sumarização de decisões indianas identificou alucinações e imprecisões, incluindo confusão de nomes e informações, que demandaram intervenção humana para correção adequada.

Shen et al. (2022) introduziram o conjunto de dados Multi-LexSum (*dataset*), focado na sumarização de processos de direito civil nos Estados Unidos, distinguindo-se por conter resumos de referência elaborados por especialistas humanos. Apesar da elevada qualidade dos dados de treinamento, os modelos de sumarização de ponta apresentaram desempenho insatisfatório, recebendo pontuação média de apenas 0,43 em escala de zero a três na avaliação humana. Os especialistas identificaram que os modelos tendiam a alucinar informações importantes e que seria mais demorado corrigir os erros do que redigir o resumo desde o início, evidenciando limitações na capacidade de sumarização jurídica atual.

Hamilton (2023) empregou o GPT-2 para prever decisões da Suprema Corte dos Estados Unidos, alcançando precisão de 60%, superior à aleatoriedade, porém notou que o modelo confundia detalhes e gerava referências desprovidas de sentido. Ammar et al. (2023), por sua vez, em pesquisa sobre decisões em árabe, revelaram que, embora o GPT-3.5 superasse outros LLMs, todos os modelos demonstraram inconsistência quando comparados à avaliação humana, evidenciando que sofisticação tecnológica ainda não assegura confiabilidade.

Malik et al. (2021) buscaram realizar a predição e explicação de julgamentos judiciais, introduzindo o *Indian Legal Documents Corpus*. A tarefa consistia em prever a decisão final e fornecer explicação identificando sentenças justificativas. O melhor modelo alcançou acurácia de setenta e oito por cento, abaixo dos especialistas humanos (noventa e quatro por cento). A análise revelou diferença entre a perspectiva algorítmica e dos especialistas, demonstrando lacuna em como máquinas e humanos explicam decisões judiciais.

Savelka et al. (2023) investigaram a capacidade do GPT-4 de explicar conceitos jurídicos e descobriram que, apesar de gerarem textos convincentes, as explicações frequentemente continham erros factuais. Notavelmente, os autores verificaram que a implementação de uma abordagem assistida melhorou a qualidade das respostas e pareceu eliminar o problema da invenção de informações incorretas, apontando caminho promissor para aumentar a robustez dessas ferramentas no futuro.

Martin et al. (2024) compararam LLMs, advogados juniores e provedores de serviços jurídicos na revisão de contratos, utilizando gabarito de advogados seniores. Os modelos, como o

GPT-4, alcançaram F1-score⁴ de 0,87, em paridade com os advogados terceirizados e acima dos advogados juniores (0,86). Além da acurácia, os modelos completaram revisões em segundos versus horas humanas, reduzindo custos em até noventa e nove vírgula noventa e sete por cento. Os autores concluem que os modelos já constituem ferramentas superiores para revisão de contratos, sinalizando mudança no setor de serviços legais.

Niklaus et al. (2023) desenvolveram o LEXTREME, um *benchmark* multilíngue e multitarefa para avaliar modelos de linguagem no domínio jurídico. A comparação entre o ChatGPT e modelos supervisionados menores revelou que o ChatGPT apresentou desempenho menor nas tarefas testadas, demonstrando que, apesar da versatilidade dos modelos generativos de grande escala, abordagens supervisionadas e especializadas ainda os superam em tarefas jurídicas específicas.

Licari e Comandè (2024) desenvolveram os modelos ITALIAN-LEGAL-*Bidirectional Encoder Representations from Transformers* (BERT) para o contexto jurídico italiano, propondo versões treinadas desde a origem e adaptadas de modelos base. Os resultados demonstraram que os modelos especializados superaram seus equivalentes de propósito geral em diversas tarefas, incluindo classificação de similaridade semântica e classificação de grandes documentos, reforçando que modelos treinados especificamente para o domínio jurídico apresentam melhor desempenho em certas aplicações.

Masala et al. (2021) desenvolveram o jurBERT para o domínio jurídico romeno, visando prever resultados de decisões judiciais. O modelo superou os modelos base em conjuntos de dados (*datasets*) com argumentos resumidos, porém apresentou desempenho apenas ligeiramente superior em textos longos. Os autores concluíram que a limitação do tamanho máximo de entrada dos modelos BERT prejudica o desempenho em documentos jurídicos maiores.

De Martino et al. (2022) propuseram o método *Paragraph Regularities Identification in Legal Judgments* (PRILJ), sigla para Identificação de Regularidades em Parágrafos de Decisões Judiciais, que visa identificar regularidades em decisões judiciais, visando auxiliar na redação de novos documentos. O sistema utiliza o processo de agrupamento (*clustering*) para agrupar documentos por conteúdo semântico, e então identifica parágrafos recorrentes dentro de cada grupo. Conceitualmente, os autores definem que enquanto o agrupamento é a ação de organizar, um grupo é o conjunto resultante dessa organização, contendo objetos similares entre si, porém diferentes dos objetos em outros grupos. O PRILJ demonstrou robustez ao ruído, definido como presença de palavras enganosas, como homônimos, que podem afetar a análise semântica, comparado com variantes do LEGAL-BERT, constituindo ferramenta útil para identificação de parágrafos recorrentes na redação de documentos jurídicos semelhantes.

Oliveira e Nascimento (2023), em seu estudo, empregaram modelos baseados em transformadores (*transformers*), como BERT, GPT-2 e RoBERTa, para análise de documentos judiciais brasileiros, transformando duzentos e dez mil processos em vetores para fins de agrupamento. Os resultados demonstraram desempenho superior dos modelos de transformadores

⁴ A definição de F1-score pode ser encontrada na Subseção 4.6.1

em tarefas de classificação e análise semântica, com destaque para a variante do RoBERTa especializada para o português do Brasil, evidenciando o potencial dessas arquiteturas para o domínio jurídico nacional.

No Superior Tribunal de Justiça (STJ) (Superior Tribunal de Justiça, 2025), o projeto STJ Logos, lançado em fevereiro de 2025, foi projetado para aumentar a eficiência na produção de decisões, focando na geração de minutas de relatórios e na análise de admissibilidade de Agravos em Recurso Especial (AREsps), a classe processual de maior volume na corte. Conforme o projeto, a ferramenta opera como mecanismo de apoio, sendo desenhada para preservar a autonomia do julgador, que mantém a responsabilidade final pelo mérito e pela elaboração da decisão.

3.2 ARCABOUÇO NORMATIVO NO BRASIL E O AI ACT

3.2.1 Resoluções 332/2020 e 615/2025 do Conselho Nacional de Justiça - CNJ

O CNJ vem dedicando atenção à regulamentação da IA na esfera judicial brasileira, buscando conciliar o avanço tecnológico com a salvaguarda dos direitos fundamentais. O marco inicial dessa trajetória regulatória materializou-se na Resolução nº 332, editada em 21 de agosto de 2020, que instituiu as primeiras diretrizes éticas voltadas ao desenvolvimento e à utilização de sistemas de IA no contexto do Poder Judiciário. Essa norma inaugural (Conselho Nacional de Justiça, 2020) consagrou princípios, tais como a transparência, a equidade e a necessidade da supervisão humana, visando prevenir que a automação de procedimentos resultasse em deliberações arbitrárias ou que reproduzissem padrões discriminatórios.

O foco dessa regulamentação recaía sobre sistemas de IA destinados à gestão de processos e ao apoio à atividade jurisdicional, refletindo o estágio tecnológico então vigente e sua aplicação no ambiente jurídico. A evolução da tecnologia, notadamente com o surgimento de modelos de linguagem sofisticados e da IAGen, evidenciou a necessidade de atualizar e aprofundar o arcabouço regulatório existente. Diante desse cenário, e reconhecendo a demanda por diretrizes para a governança, a auditoria e o monitoramento de soluções baseadas em IA, o CNJ promulgou a Resolução nº 615, datada de 11 de março de 2025.

Este novo instrumento normativo (Conselho Nacional de Justiça, 2025b) foi elaborado com o objetivo de endereçar os desafios e as potencialidades da IAGen, estabelecendo um marco regulatório mais detalhado para sua utilização responsável no âmbito judicial. A Resolução nº 615/2025 afirma a centralidade da pessoa humana, conforme disposto no inciso IV do artigo 2º, e a tutela dos direitos fundamentais, matéria central do artigo 5º, como pilares do sistema. A norma inova ao instituir, em seu artigo 15, o Comitê Nacional de Inteligência Artificial do Judiciário, órgão de composição plural destinado a supervisionar a implementação das regras, e ao determinar, no artigo 14, a realização de avaliações de impacto algorítmico para os sistemas classificados como de maior risco.

Para estruturar esse ecossistema, o artigo 4º da resolução estabelece distinção entre

os diversos atores envolvidos. Em seus incisos, define o desenvolvedor (inciso IV) como a entidade responsável pela criação do sistema, o distribuidor (inciso VIII) como aquele que o disponibiliza para terceiros, e segmenta os usuários em internos (inciso VI), a exemplo de magistrados e servidores, e externos (inciso VII), como advogados e partes processuais. Essa diferenciação revela-se importante para a atribuição adequada de responsabilidades. As soluções de IA, sejam internamente desenvolvidas ou contratadas, devem ser registradas na plataforma Sinapses, conforme determina o artigo 24, que opera como um catálogo centralizado para fomentar a colaboração entre tribunais e evitar a duplicidade de esforços.

Uma das inovações da norma reside na ênfase à transparência e à *accountability* (responsabilização). Para tanto, o artigo 4º estabelece definições importantes, que abrangem desde o comando de interação com a tecnologia, como o prompt (inciso XVI), que designa o comando textual destinado a direcionar a IAGen, até os alicerces de governança, como a auditabilidade (inciso XVII), que assegura a possibilidade de avaliação de um sistema, e a explicabilidade (inciso XVIII), que exige a compreensão de como as determinações da IA são produzidas. Esses mecanismos revelam-se importantes para combater o que o inciso XIII do mesmo artigo define como viés discriminatório ilegal ou abusivo, isto é, resultados que criam, reproduzem ou reforçam preconceitos.

O parágrafo 3º do artigo 8º é taxativo ao determinar que, uma vez identificado um viés impossível de ser eliminado, a solução de IA deve ser descontinuada. Essa preocupação com os riscos materializa-se em uma abordagem de categorização. O artigo 10 da resolução veda expressamente a utilização de sistemas de risco excessivo, a exemplo daqueles que operam sem possibilidade de revisão humana ou que realizam ranqueamento social de pessoas. Para as demais aplicações, o artigo 11 e seu Anexo instituem uma classificação entre alto risco e baixo risco, submetendo as primeiras, nos termos do parágrafo 1º do artigo 13, a um controle mais rigoroso, com auditorias contínuas e monitoramento constante para mitigar potenciais danos aos direitos fundamentais, à privacidade e à justiça.

Apesar do rigor aparente, a Resolução nº 615/2025 apresenta pontos críticos que suscitam preocupações quanto à governança e à segurança das informações dos jurisdicionados. O parágrafo 2º do artigo 19 faculta ao magistrado, servidor ou colaborador a “contratação direta de solução mediante assinatura ou cadastro de natureza privada” quando o tribunal não oferecer solução corporativa. Embora tal permissão esteja condicionada às diretrizes do parágrafo 3º, que incluem a vedação de uso para dados sigilosos (inciso IV) e a proibição de compartilhamento de dados para treinamento (inciso III), a fiscalização dessas regras em contas privadas geridas individualmente em plataformas de terceiros impõe desafios de auditoria. Essa descentralização do controle tecnológico, conhecida como *Shadow IT* (IBM, 2025b), fragiliza a capacidade institucional de monitoramento e cria riscos de vazamento de dados sensíveis para grandes corporações tecnológicas.

A transparência algorítmica também merece uma atenção sob um viés crítico. O parágrafo 6º do artigo 19 estabelece que a menção ao uso de IA generativa no corpo da decisão judicial fica “a critério do magistrado”. Embora o dispositivo exija registro automático no

sistema interno do tribunal para fins de controle administrativo, a ausência de obrigatoriedade de informar às partes processuais que o ato decisório contou com auxílio algorítmico gera assimetria informacional. Tal discricionariedade conflita com o princípio da transparência previsto no inciso II do artigo 3º e prejudica a contestabilidade (artigo 4º, inciso XIX), uma vez que a parte pode desconhecer que a fundamentação ou a redação da decisão derivou, ainda que parcialmente, de um sistema automatizado.

Nessa mesma linha, o artigo 26, parágrafo 2º, flexibiliza a qualidade dos dados ao permitir o uso de fontes não governamentais quando os dados públicos forem “insuficientes ou inadequados”. Embora o dispositivo exija uma validação supostamente rigorosa, essa abertura insere o Judiciário na lógica de dependência dos grandes *datasets* comerciais, cuja curadoria é frequentemente opaca e orientada por interesses de mercado. Isso introduz o risco de importação de vieses estruturais presentes em dados da internet aberta para os modelos judiciais, desafiando a capacidade dos tribunais de garantir a “justiça, a equidade e a não-discriminação” preconizadas no inciso I do artigo 3º.

A classificação de riscos apresentada no Anexo da resolução também merece análise, especialmente o item AR5. A norma classifica a identificação e autenticação facial ou biométrica para monitoramento de comportamento como “Alto Risco”, e não como “Risco Excessivo” (proibido), criando exceções, inclusive quanto às obrigações inerentes à categoria de “Alto Risco”, para a “mera confirmação da identidade” ou “atividades de segurança pública devidamente justificadas”. Essa abertura interpretativa pode legitimar a implementação de tecnologias de vigilância em massa nos tribunais, uma vez que transfere para o gestor local a ponderação de direitos fundamentais em áreas de alta sensibilidade. Ao fazê-lo, a resolução entra em contradição com seu próprio objetivo de “mitigação de riscos sistêmicos” previsto no art. 2º, inciso VII.

Em síntese, embora as Resoluções 332/2020 e 615/2025 do CNJ representem avanço na governança digital, a análise revela tensões entre inovação e segurança dos jurisdicionados. A flexibilização normativa observada, consubstanciada na permissão de contratação privada de ferramentas (Art. 19, § 2º), na discricionariedade quanto à publicidade do uso de IA nas decisões (Art. 19, § 6º) e na admissão de dados externos e biometria de alto risco, sugere um modelo que transfere riscos significativos para a gestão individual dos magistrados. Essas lacunas, ao possibilitarem opacidade processual e dependência de tecnologias de mercado, desafiam os próprios princípios de transparência e auditabilidade que a norma visa instituir.

Cabe ainda destacar que a natureza administrativa dessas resoluções restringe sua eficácia ao âmbito interno do Poder Judiciário, criando descompasso regulatório em relação aos desenvolvedores privados, que não estão sujeitos a tais diretrizes. Considerando que o Judiciário passa a integrar soluções comerciais em seu fluxo decisório, a ausência de legislação federal em sentido estrito gera um vácuo de responsabilidade civil e direitos digitais. Portanto, a plena efetividade da governança de IA no Brasil depende não apenas da autorregulação judicial, mas da aprovação de um marco legislativo robusto que submeta as grandes corporações tecnológicas aos padrões éticos e jurídicos consubstanciados na Constituição Federal.

3.2.2 AI Act e o Projeto de Lei 2338/2023

No cenário legislativo global, a emergência da IA impulsionou um movimento regulatório, na qual diferentes nações buscam estabelecer marcos normativos para governar o desenvolvimento e a aplicação dessas tecnologias. O AI Act da União Europeia, formalmente adotado como Regulamento (UE) 2024/1689 (Conselho da União Europeia, 2024), consolidou-se como um marco legal em escala mundial, notabilizado por sua abordagem abrangente e fundamentada em riscos. No Brasil, essa busca por uma regulamentação legislativa ampla encontra-se em tramitação no Congresso Nacional.

O Projeto de Lei 2338/2023 (Senado Federal, 2023), cuja versão objeto de análise é a delineada pelo Parecer nº 208 de 2024,⁵ o texto mais atualizado e datado em 10 de dezembro de 2024, já contém inúmeras regras de relevância para o uso e aplicação dessas tecnologias, não apenas no âmbito produtivo e social, mas também no seio das próprias instituições públicas. Este projeto é compreendido por pesquisadores como Cunha e Nascimento (2025) por ser uma proposta que se inspira na estrutura do AI Act, adotando uma abordagem similar de categorização de riscos e de delimitação da atividade estatal, embora com adaptações ao ordenamento jurídico e ao contexto institucional brasileiro.

No AI Act, os principais atores da cadeia de valor da IA são definidos no Artigo 3. Estes incluem o provedor (*provider*), que desenvolve um sistema de IA ou modelo de IA de *General-purpose AI Model* (Modelo de IA de Propósito Geral) (GPAI) e o coloca no mercado, o implantador (*deployer*), que utiliza um sistema de IA sob sua autoridade, o importador (*importer*), o distribuidor (*distributor*) e o representante autorizado (*authorised representative*). Já no PL 2338, o Artigo 4º define os atores correspondentes: o desenvolvedor, figura análoga ao provedor europeu, o aplicador, que emprega ou utiliza o sistema de IA, correspondendo ao implantador, e o distribuidor. O projeto brasileiro também agrupa esses atores sob o termo coletivo agentes de inteligência artificial (Art. 4º, VIII), simplificando a cadeia de responsabilidades em comparação com o AI Act.

A abordagem baseada em risco constitui o pilar de ambos os textos. O AI Act, em seu Artigo 5, proíbe certas práticas de IA consideradas de risco inaceitável. Isso inclui sistemas que empregam técnicas subliminares para distorcer o comportamento de uma pessoa de maneira prejudicial, exploram vulnerabilidades de grupos específicos, sistemas de pontuação social (*social scoring*) por autoridades públicas e, com exceções estritas, o uso de identificação biométrica remota em tempo real em espaços públicos para fins de aplicação da lei.

O PL 2338 espelha essa abordagem em seu Artigo 13, que veda o desenvolvimento e uso de sistemas de IA de risco excessivo. A lista de proibições é similar, incluindo a instigação de comportamentos que causem danos, a exploração de vulnerabilidades, a pontuação social

⁵ Este trabalho foi desenvolvido no segundo semestre de 2025, quando o projeto já havia tramitado pelo Senado Federal e aguardava análise da Câmara dos Deputados. Trata-se, portanto, de uma análise sobre texto ainda não finalizado, mas cuja estrutura base permite comparação com o estado da arte internacional sobre a matéria, possibilitando identificar seus principais pontos de convergência e divergência.

pelo poder público, denominada pontuação universal, e a identificação biométrica à distância em tempo real em espaços acessíveis ao público.

As exceções para o uso de identificação biométrica pelas autoridades constituem um ponto de divergência: enquanto o AI Act (Art. 5(1)(h)) as limita à busca de vítimas específicas, prevenção de ameaças iminentes, como ataques terroristas, e localização de suspeitos de crimes graves, o PL 2338 (Art. 13, IV) inclui hipóteses como instrução de inquérito, busca de vítimas, flagrante delito para crimes com pena superior a dois anos e recaptura de foragidos, o que pode, se aprovado, representar um escopo ligeiramente mais amplo para a aplicação da tecnologia no contexto brasileiro.

O AI Act também, em seu Artigo 5(1)(d), proíbe sistemas de avaliação de risco para prever a ocorrência de um crime baseados exclusivamente no perfil de uma pessoa ou na avaliação de seus traços de personalidade. A palavra exclusivamente constitui ponto relevante na diferença entre as legislações: a norma veda a predição criminal automatizada e baseada em perfil, mas, em contrapartida, seu Anexo III, ponto 6, classifica como de alto risco os sistemas que realizam essa mesma avaliação de risco de forma não exclusiva, isto é, como ferramenta de apoio à decisão humana.

O legislador europeu, portanto, proíbe a máquina como juíza preditiva autônoma, mas a admite como assistente sob governança. O PL 2338/2023, em contraste, adota uma postura aparentemente mais restritiva, mas ainda também permitindo seu uso no Poder Judiciário. Seu Artigo 13, I, alínea c, veda como risco excessivo o uso de IA para avaliar os traços de personalidade, as características ou o comportamento passado para avaliação de risco de cometimento de crimes, sem o qualificador exclusivamente. Essa ausência sugere uma proibição mais ampla, que incide sobre a própria metodologia de usar perfis comportamentais para predição criminal, independentemente do grau de automação ou de supervisão humana.

Uma das inovações do AI Act é a regulação específica para modelos de IA de GPAI tratada no Capítulo V, que inclui os modelos de IAGen. Ele impõe obrigações de transparência a todos os provedores de GPAI (Art. 53), como a elaboração de documentação técnica e o fornecimento de informações para os desenvolvedores de sistemas *downstream*. Além disso, introduz uma categoria de GPAI com risco sistêmico (Art. 51), classificação que pode decorrer tanto de uma avaliação técnica de suas capacidades de alto impacto quanto de uma decisão discricionária da Comissão Europeia, sujeitando-os a obrigações mais rigorosas, como avaliação de modelos, rastreamento de incidentes e garantia de cibersegurança (Art. 55).

O PL 2338 também aborda essa questão na Seção V de seu Capítulo IV, intitulada Das Medidas de Governança para Sistemas de Inteligência Artificial de GPAI e Generativa. O Artigo 29 exige que o desenvolvedor realize uma avaliação preliminar para identificar níveis de risco, incluindo o potencial risco sistêmico. O Artigo 30 estabelece requisitos para sistemas com risco sistêmico, como a documentação de testes e o gerenciamento de riscos. Embora o conceito esteja presente, o PL 2338 não estabelece um critério claro como o do AI Act para as formas de designação de risco sistêmico, deixando, aparentemente, essa definição mais aberta à regulamentação infralegal.

Por fim, é imperativo ressaltar, novamente, esta análise no seu devido contexto: ela se debruça sobre o PL 2338/2023 em sua versão delineada pelo Parecer nº 208 de 2024, um texto legislativo avançado, porém não definitivo. O objetivo foi, portanto, realizar um diagnóstico crítico do estágio atual da proposta, utilizando o AI Act como referencial para aferir não apenas as convergências, mas, sobretudo, as divergências que sinalizam as escolhas e os potenciais rumos da regulação nacional.

De um lado, o projeto avança em relação ao seu par europeu ao ser mais taxativo na proteção de direitos e mais rigoroso na vedação da predição criminal, demonstrando uma preocupação com garantias fundamentais. De outro, contudo, adota um tratamento mais brando em áreas de alto impacto, como nas exceções para o uso de biometria pelo Estado e na delegação da definição de risco sistêmico para a esfera infralegal. Esses são os campos de discussão que o debate no Congresso Nacional deverá arbitrar, definindo se o Brasil se alinhará ao rigor protetivo europeu ou se inaugurará uma via regulatória com maior margem para a discricionariedade estatal e o risco social.

4 ESTUDO DE CASO: ENGENHARIA DE PROMPT NO PROJETO CONCIL-IA

Este capítulo apresenta o núcleo empírico da pesquisa, detalhando a metodologia e a execução do estudo de caso sobre a engenharia de prompt no projeto CONCIL-IA. O percurso metodológico inicia-se na Seção 4.1, com a apresentação do projeto e sua metodologia de desenvolvimento, estabelecendo o contexto da investigação. A seguir, a Seção 4.2 descreve a base de dados utilizada, detalhando o processo de coleta do *corpus* de decisões judiciais e a etapa de anonimização. Com os dados preparados, a Seção 4.3 explora as abordagens e técnicas de pesquisa, fundamentando o enquadramento no Raciocínio Baseado por Fatores (*Factor-Based Reasoning*) (FBR) e detalhando a construção do *ground truth*, que serve como padrão de referência para a validação. O foco do capítulo é aprofundado na Seção 4.4, que detalha o processo de engenharia de prompt em duas fases: o Experimento Principal na Subseção 4.4.1, que analisa a transição do *Prompt Inicial* ao *Prompt Final*, e o Experimento Secundário na Subseção 4.4.2, que explora novas hipóteses de otimização. Na sequência, a Seção 4.5 especifica os modelos, parâmetros e o ambiente experimental utilizados e, finalmente, para quantificar o desempenho da extração, a Seção 4.6 estabelece as métricas de avaliação, diferenciando os critérios para fatores categóricos e numéricos. Ao detalhar cada etapa do *pipeline* experimental, este capítulo constrói a base metodológica necessária para a análise e discussão dos resultados, que serão apresentados no capítulo subsequente.

4.1 O PROJETO CONCIL-IA E SUA METODOLOGIA DE DESENVOLVIMENTO

Conforme introduzido no Capítulo 1 deste trabalho, o sistema judiciário brasileiro enfrenta um desafio estrutural de sobrecarga processual, fenômeno documentado pelo CNJ (Conselho Nacional de Justiça, 2025a). Em resposta, o Conselho tem promovido uma modernização institucional em duas frentes principais: o incentivo a métodos autocompositivos de resolução de conflitos (Conselho Nacional de Justiça, 2010) e a incorporação regulamentada de tecnologias emergentes, como a IA (Conselho Nacional de Justiça, 2020; Conselho Nacional de Justiça, 2025b).

Nesse contexto de políticas judiciárias que convergem o incentivo à consensualidade com a inovação tecnológica responsável, insere-se o projeto Concil-IA, fundamentando-se no arcabouço normativo que detalhado na Seção 3.2. A iniciativa investiga o potencial da IA como instrumento para aprimorar os mecanismos de resolução de disputas, resultando de uma colaboração entre o grupo de pesquisa EGOV/UFSC e o Juizado Especial Cível (JEC) do Tribunal de Justiça do Estado de Santa Catarina (TJSC). O objetivo central consiste no desenvolvimento de um sistema de ODR para auxiliar as partes na busca por acordos, com foco específico em litígios de consumo (Sabo et al., 2025).

A trajetória do projeto remonta a 2019, quando teve início uma colaboração multidisciplinar congregando pesquisadores dos programas de Pós-Graduação em Direito, Engenharia do Conhecimento e Automação da Universidade Federal de Santa Catarina (UFSC). A etapa inicial,

desenvolvida entre 2019 e 2020, delimitou o escopo do projeto a ações relacionadas a falhas no transporte aéreo, adotando uma metodologia centrada na extração de fatores preditivos das sentenças para estimar resultados prováveis (Grupo de Pesquisa Governo Eletrônico, Inclusão Digital e Sociedade do Conhecimento, 2025).

Em 2021, o modelo foi submetido à validação empírica durante sessões de conciliação reais, nas quais suas predições foram utilizadas como parâmetro probabilístico não vinculante para informar o processo de negociação, confirmando o potencial da ferramenta como instrumento de apoio à autocomposição (Grupo de Pesquisa Governo Eletrônico, Inclusão Digital e Sociedade do Conhecimento, 2025). A partir de 2023, o projeto consolidou sua identidade institucional como Concil-IA, expandindo sua equipe de pesquisa e obtendo apoio de instituições como a Fundação de Amparo à Pesquisa e Inovação do Estado de Santa Catarina (FAPESC), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), TJSC e OpenAI (Grupo de Pesquisa Governo Eletrônico, Inclusão Digital e Sociedade do Conhecimento, 2025).

Atualmente, sua metodologia encontra-se estruturada em cinco frentes de trabalho interdependentes (Sabo et al., 2025):

1. A coleta de sentenças condenatórias de indenização por danos morais envolvendo temáticas de Direito do Consumidor.
2. A identificação e extração automatizada de fatores preditivos das sentenças (fatores do caso que impactam especialmente no arbitramento do valor da indenização por danos morais), com o auxílio do modelo de linguagem grande multimodal Generative Pre-trained Transformer GPT.
3. A construção de modelo preditivo baseado em técnicas de Aprendizado de Máquina, especificamente de classificação, apresentando como saída predições de faixas de indenização por danos morais.
4. A aplicação de técnicas de *Explainable Artificial Intelligence* (Inteligência Artificial Explicável) (XAI), com predições que sejam explicáveis ao usuário, relacionando o valor de indenização gerado pelo modelo com os fatores preditivos extraídos.
5. A disponibilização do modelo através de interface web para as partes e conciliadores.

O foco deste trabalho recai sobre a segunda frente metodológica: a extração automatizada de fatores preditivos por meio de um LLM. A viabilidade desta abordagem, que postula a capacidade do modelo GPT-4 de emular a análise de um especialista humano na identificação de fatores relevantes em textos não estruturados, foi objeto de análise preliminar em trabalho anterior com participação deste autor (Pereira et al., 2025).

Naquela oportunidade, explorou-se a acurácia do modelo e as estratégias mais eficazes para a extração, confirmando-se o potencial da ferramenta para essa finalidade. Diferentemente da exploração inicial, o presente estudo aprofunda-se nos aspectos técnicos daquela validação. O

objetivo consiste em apresentar uma exposição da metodologia empregada, da engenharia de prompts desenvolvida especificamente para a tarefa de extração e dos resultados quantitativos e qualitativos obtidos, além da exploração de novos testes.

Desta forma, este estudo detalha a etapa da engenharia de prompt (Frente 2) que serve como alicerce para a modelagem preditiva (Frente 3), investigando se a extração automatizada por LLMs por intermédio das técnicas de engenharia de prompt é apta para gerar os *datasets* estruturados e validados necessários para o treinamento de algoritmos de Aprendizado de Máquina (*Machine Learning*) eficazes. Portanto, este trabalho busca detalhar e expandir a investigação previamente realizada, oferecendo um relato do processo experimental de engenharia de prompt e de seus desdobramentos, apresentando os achados preliminares e contribuindo para o conhecimento na área.

4.2 BASE DE DADOS: O *CORPUS* DE ANÁLISE

4.2.1 Coleta e Contexto do Dataset Não Estruturado

A presente investigação, bem como o estudo de referência conduzido por Pereira et al. (2025), fundamenta-se em um *corpus* composto por 1.163 decisões judiciais em formato textual bruto, constituindo o dataset não estruturado objeto de análise. Os documentos analisados foram emitidos pelo JEC da UFSC no intervalo temporal entre os anos de 2011 e 2020.

A totalidade dos processos judiciais abrange uma temática recorrente no âmbito do Direito Consumerista brasileiro: as falhas verificadas na prestação de serviços de transporte aéreo. Tais falhas englobam situações diversas, incluindo atrasos, cancelamentos de voos e extravios de bagagens, todas configurando violações aos direitos dos consumidores de serviços aéreos no território nacional.

O Código de Defesa do Consumidor (Brasil, 1990) estabelece expressamente o direito à reparação tanto por danos de natureza material quanto imaterial. Contudo, a quantificação pecuniária dos danos de caráter imaterial, também denominados danos morais, não encontra parâmetros legais objetivos e previamente estabelecidos. Consequentemente, os valores indenizatórios são determinados pelo magistrado mediante análise casuística.

A decisão judicial considera múltiplos fatores (*factors*) presentes no caso concreto, conforme destacado por Pereira et al. (2025), abrangendo elementos como idade da vítima, gênero, localização anatômica da lesão e condições de saúde preexistentes. Esta ausência de critérios uniformes introduz substancial grau de subjetividade nas decisões judiciais, tornando necessária a investigação dos fatores determinantes.

Nesse contexto, a extração e a análise dos fatores que exercem influência sobre as decisões judiciais constituem tarefas de importância para o desenvolvimento de modelos computacionais dotados de capacidade preditiva e explicativa. A identificação dos fatores presentes em um caso concreto permite que tais modelos prognostiquem tanto a probabilidade de deferimento da indenização quanto estimem sua extensão financeira.

A predição fundamenta-se nos dados de julgamentos anteriores utilizados durante a fase de treinamento do modelo, constituindo um dos objetivos centrais do projeto denominado CONCIL-IA, conforme descrito no item 3 da Seção anterior 4.1. Esta abordagem metodológica busca introduzir maior previsibilidade e transparência nas decisões judiciais em matéria consumerista.

4.2.2 O Processo de Anonimização

A viabilização ética desta investigação científica e a observância aos princípios reguladores da proteção de dados exigiram que todos os documentos integrantes do *corpus* fossem submetidos a um processo de anonimização previamente a qualquer procedimento analítico, conforme relatado por Pereira et al. (2025). Este protocolo revelou-se necessário em virtude da natureza sensível das informações presentes nas decisões.

As decisões judiciais incorporam informações de identificação direta, exemplificadas pelos nomes completos das partes litigantes, isto é, autor e réu. Adicionalmente, contêm dados de identificação indireta, como o número único atribuído ao processo judicial (Pereira et al., 2025). A exposição de tais informações violaria os direitos fundamentais à privacidade e à proteção de dados pessoais dos jurisdicionados envolvidos nos litígios analisados.

A desidentificação dos documentos foi operacionalizada mediante a aplicação de um algoritmo automatizado, desenvolvido especificamente para esta finalidade, que emprega Expressões Regulares, conhecidas pela nomenclatura técnica RegEx, para detectar e substituir sistematicamente as informações de caráter sensível (Pereira et al., 2025). A metodologia implementada diferenciou-se conforme a natureza específica de cada categoria de dado a ser removido.

Para o número do processo judicial, que segue um padrão numérico estruturado e previsível, foi empregada uma estratégia de busca direta por meio de RegEx. Esta técnica permitiu localizar e substituir todas as ocorrências do número processual por um marcador genérico padronizado (Pereira et al., 2025), preservando a estrutura do documento enquanto eliminava a informação identificadora específica que poderia vincular o texto ao caso real.

O processo de anonimização foi conduzido de maneira iterativa, contando com a colaboração de um especialista da área jurídica. O profissional analisava os documentos que apresentavam falhas no processo automatizado de anonimização, identificava os padrões linguísticos que o algoritmo não havia capturado adequadamente e os incorporava à lista de expressões de busca (Pereira et al., 2025).

Este ciclo de execução algorítmica e refinamento metodológico foi iterado sucessivamente até que a totalidade dos 1.163 documentos constituintes do *corpus* original fosse anonimizada (Pereira et al., 2025). Tal procedimento garantiu a integridade e a confidencialidade da base de dados final utilizada nas análises subsequentes, atendendo aos padrões éticos de pesquisa.

4.3 ABORDAGENS E TÉCNICAS DE PESQUISA

A estrutura metodológica desta investigação foi concebida para estabelecer uma conexão entre dois paradigmas da Inteligência Artificial contemporânea. O primeiro paradigma consiste na aprendizagem no contexto, denominada *in-context learning* exposta no Capítulo 2. O segundo paradigma refere-se ao aprendizado supervisionado, aplicado posteriormente para a predição de resultados judiciais.

Conforme a fundamentação teórica apresentada na Seção 2.2, o desafio central na aplicação de modelos preditivos de natureza supervisionada a textos de origem jurídica reside na necessidade de dispor de um *dataset* estruturado em formato tabular. Neste formato, as características, também denominadas fatores, de cada caso concreto devem estar adequadamente estruturadas em fatores analisáveis por algoritmos de aprendizado de máquina.

A abordagem proposta neste trabalho acadêmico e no de referência (Pereira et al., 2025) consiste na automação da etapa de estruturação dos dados textuais e no processo metodológico da engenharia do prompt. Em substituição à dependência de um processo manual, que demandaria recursos humanos e temporais substanciais, emprega-se a engenharia de *prompt* em conjunção com um LLM. Este modelo atua como ferramenta de extração de características operando em larga escala.

Neste fluxo metodológico, o modelo transforma o texto de formato bruto, descrito antes na Subseção 4.2.2, em dados estruturados e utilizáveis para análises quantitativas. O conjunto de dados de referência, cuja construção metodológica é exposta também na Subseção 4.3.2, assume um papel relevante no contexto desta pesquisa, conforme explicitado a seguir.

Primeiramente, o conjunto de dados de referência funciona como Padrão de Validação (*Ground Truth*), constituindo o gabarito utilizado para avaliar tanto a acurácia quanto a robustez do LLM na tarefa específica de extração de informações estruturadas. Secundariamente, o mesmo conjunto de dados serve como *dataset* de treinamento, fornecendo o conjunto inicial de dados de alta qualidade necessário para a subsequente etapa de treinamento do modelo preditivo de natureza supervisionada.

Esta etapa corresponde à frente 3 do projeto CONCIL-IA, estabelecendo assim a ponte entre a extração automatizada e a predição judicial propriamente dita. As subseções subsequentes detalharão cada componente técnico desta estratégia metodológica integrada, explicitando os procedimentos, ferramentas e técnicas empregadas em cada fase do processo investigativo.

4.3.1 Contextualização Teórico-Metodológica: Raciocínio Orientado por Fatores (FBR)

Este estudo fundamenta sua base metodológica nas abordagens estabelecidas de Inteligência Artificial no âmbito jurídico, com ênfase no FBR. A contextualização desta escolha requer examinar sua distinção em relação ao Raciocínio Orientado por Casos (*Case-Based Reasoning*) (CBR), técnica igualmente proeminente neste campo.

Conforme estabelece Rempell (2022), elementos denominados fatores representam

componentes principais na orientação das deliberações judiciais, originando-se tanto de dispositivos legais quanto de construções jurisprudenciais. A classificação destes elementos ocorre em duas vertentes. Fatores de natureza binária expressam apenas presença ou ausência no contexto analisado, subdividindo-se entre elementos favoráveis (*pro-factors*) e elementos contrários (*con-factors*) ao desfecho pretendido.

Distintamente, componentes escalares denominados dimensões (*dimensions*) demonstram influência graduada, estabelecendo uma continuidade de intensidades ao invés de categorias estanques (Sartor, 2005; Rissland; Ashley, 2002). Gray et al. (2022) caracterizam a aplicação do FBR em investigações jurídicas através da análise de sentenças que evidenciam fatores determinados, verificando sua contribuição para sustentar ou contrapor a conclusão judicial.

A metodologia CBR, conforme delineiam Gould et al. (2024), funciona mediante processo diferenciado de resolução de controvérsias, elaborando soluções para contextos inéditos através da recuperação e adequação de situações precedentes análogas. Estruturas jurídicas baseadas em *Common Law* demonstram afinidade especial com o CBR devido à dependência intrínseca de jurisprudência consolidada (Rissland; Ashley; Branting, 2005).

A investigação presente e o estudo referência (Pereira et al., 2025) implementam a perspectiva FBR. Esta determinação metodológica deriva dos propósitos investigativos delineados. Contrariamente ao CBR, que procura solucionar casos inéditos mediante adaptação de precedentes, estas pesquisas objetivam estruturar repositórios através da identificação de elementos favoráveis, contrários e dimensionais em decisões judiciais.

A finalidade consiste em mapear componentes determinantes na mensuração de compensações morais. Esta abordagem vincula previsões quantitativas ao raciocínio contextualizado exigido em ambientes jurídicos, viabilizando sistemas preditivos dotados de interpretabilidade (Pereira et al., 2025).

4.3.2 Estruturação Manual dos Dados: O *Ground Truth*

A elaboração de uma base referencial denominada *ground truth* configurou fase preparatória para aferir a eficácia do sistema linguístico na identificação automatizada de fatores. Juristas especializados conduziram esta atividade, processando 900 documentos correspondentes a aproximadamente 80% do acervo integral compilado (Pereira et al., 2025).

O critério de seleção aplicado priorizou litígios individuais, assegurando uniformidade e prevenindo interpretações divergentes que naturalmente emergiriam em demandas coletivas dentro de decisões unitárias (Pereira et al., 2025). A organização resultante adotou estrutura matricial em formato *Comma-Separated Values* (CSV), alocando cada processo em linha distinta enquanto fatores detectados preenchem colunas específicas (Pereira et al., 2025).

Pesquisas prévias sobre agrupamento textual jurídico contribuíram significativamente, identificando padrões recorrentes e elementos predominantes em conjuntos decisórios afins (Sabo et al., 2022). Esta compilação manual estabeleceu o parâmetro avaliativo para contrastar as identificações automatizadas geradas pelo sistema. O inventário resultante abrange 15 categorias

fatoriais, cujas especificações e incidências nos 900 processos examinados encontram-se expostas adiante (Pereira et al., 2025).

- **Pro-factors:**

1. **Extravio definitivo de bagagem (*Permanent baggage loss*):** Evento que pode gerar compensação por dano imaterial. Seu valor é categórico, representado por "sim" (24 casos) e "não" (876 casos).
2. **Bagagem violada (*Tampered baggage*):** Evento que, dependendo do nível do dano ou da falta de pertences, pode gerar compensação. Seu valor é categórico, representado por "sim" (42 casos) e "não" (858 casos).
3. **Extravio temporário de bagagem (*Temporary baggage loss*):** Evento que pode gerar compensação. Seu valor é categórico, representado por "sim" (119 casos) e "não" (781 casos).
4. **Cancelamento de voo (*Flight cancellation*):** Considera-se cancelamento os casos sem reacomodação ou quando o destino é alterado. Seu valor é categórico, representado por "sim" (139 casos) e "não" (761 casos).
5. **Atraso de voo (*Flight delay*):** Considera-se atraso os casos com reacomodação que resultam em chegada tardia. Seu valor é categórico, representado por "sim" (471 casos) e "não" (429 casos).
6. **Overbooking:** Prática abusiva de vender mais bilhetes do que os assentos disponíveis. Seu valor é categórico, representado por "sim" (27 casos) e "não" (873 casos).
7. **No show:** Cancelamento unilateral do bilhete de volta quando o consumidor não comparece ao voo de ida. Seu valor é categórico, representado por "sim" (34 casos) e "não" (866 casos).
8. **Direito de arrependimento e pedido de reembolso (*Right to regret and repayment claim*):** Dificuldade imposta ao consumidor para obter o reembolso, judicialmente chamada de *via crucis*. Seu valor é categórico, representado por "sim" (12 casos) e "não" (888 casos).
9. **Descumprimento de oferta:** Alteração da classe do passageiro para uma inferior, violando o contrato e a expectativa do consumidor. Seu valor é categórico, representado por "sim" (6 casos) e "não" (894 casos).
10. **Assistência da companhia aérea (*Airline assistance*):** Fator que influencia o valor da compensação em casos de atraso ou cancelamento. Seu valor é categórico, representado por "sim" (144 casos), "não" (413 casos) e "não aplicável" (343 casos).

11. **Hipervulnerabilidade do consumidor (*Consumer hyper-vulnerability*):** Condição especial do consumidor (idoso, gestante, com deficiência, etc.). Seu valor é categórico, representado por "sim" (31 casos) e "não" (869 casos).

- **Con-factors:**

12. **Culpa do consumidor (*Consumer fault*):** Evento que exclui a responsabilidade da companhia, como o não comparecimento do consumidor a tempo para o embarque. Seu valor é categórico, representado por "sim" (19 casos) e "não" (881 casos).
13. **Condições climáticas adversas (*Adverse weather conditions*):** Evento imprevisível que exclui a responsabilidade da companhia. Seu valor é categórico, representado por "sim" (15 casos) e "não" (885 casos).

- **Dimensions:**

14. **Intervalo de extravio (*Loss interval*):** Quanto maior o tempo para devolução da bagagem, maior o valor da compensação. Seu valor é numérico, representado em horas e minutos.
15. **Intervalo de atraso (*Delay interval*):** Quanto maior o tempo de atraso do voo, maior o valor da compensação. Seu valor é numérico, representado em horas e minutos.

4.3.3 Estruturação Automática dos Dados via Engenharia de Prompt

A estruturação manual dos dados, descrita na Seção anterior, foi um passo necessário para a criação de um *ground truth* de qualidade. Contudo, este é um processo manual, oneroso em tempo e de baixa escalabilidade, dificultando sua aplicação em grandes volumes de documentos judiciais. Diante desse desafio, o objetivo deste e do trabalho de referência (Pereira et al., 2025) é explorar o potencial dos LLMs para construir bases de dados jurídicas de forma eficiente e escalável a partir da leitura automatizada de textos e extração de seus fatores.

Para alcançar esse objetivo, a interação com o modelo segue um processo metodológico conhecido como engenharia de prompt, apresentado na Subseção 2.6. O desenvolvimento, teste e refinamento desses prompts constituem o núcleo desta etapa do trabalho, cujos diferentes testes serão detalhados na Seção 4.4.

O processo envolve o ajuste das instruções para cada um dos 15 fatores definidos na Subseção 4.3.2, a especificação do formato de saída, a reconfiguração dos fatores e a aplicação de técnicas específicas de otimização. Essas intervenções visam aumentar a acurácia e a confiabilidade da extração automática, permitindo que o modelo identifique e estruture as informações jurídicas de forma mais consistente e precisa.

4.4 O PROCESSO DE ENGENHARIA DE PROMPT

Conforme já exposto, o desenvolvimento dos prompts foi um processo envolvendo especialistas tanto da área jurídica quanto da computacional (Sabo et al., 2025). A metodologia partiu de uma primeira versão, o *Prompt Inicial*, que serviu como linha de base, e progrediu para uma versão revisada, o *Prompt Final*, após uma análise dos resultados preliminares. Com base nesses achados, uma segunda fase de experimentação foi conduzida, resultando em múltiplas versões de prompts especializados, como as famílias *Atraso* e *Extravio*. As versões integrais de todos os prompts estão disponíveis para consulta nos apêndices deste trabalho (A.1 a A.8). O processo aqui descrito detalha, portanto, tanto o Experimento Principal (Subseção 4.4.1) quanto o Experimento Secundário (Subseção 4.4.2), aplicando na prática os fundamentos da Engenharia de Prompt explorados teoricamente na Subseção 2.6.3.

4.4.1 Experimento Principal: do *Prompt Inicial* ao *Prompt Final*

4.4.1.1 O *Prompt Inicial* e suas Características

A primeira versão do prompt, cuja íntegra pode ser consultada no Apêndice A.1, foi concebida com o objetivo de ser clara e direta, estabelecendo as regras para a interação com o modelo. Sua estrutura foi composta por três pilares principais, alinhados com os elementos de um prompt funcional discutidos nas subseções 2.6.1 e 2.6.3 (Pereira et al., 2025):

1. **Definição da Persona:** O prompt inicial inicia com a instrução "Atue como um extrator de variáveis". Conforme exposto na fundamentação teórica em 2.6.1, a definição de uma *Persona* é uma técnica que contextualiza a tarefa para o LLM, condicionando-o a operar dentro de um escopo específico e a adotar o comportamento esperado para a extração de dados, em vez de atuar como um assistente conversacional genérico (White et al., 2023). Cabe esclarecer que, no contexto desta pesquisa, os termos "variável" e "fator" são empregados de forma intercambiável, ambos referindo-se aos elementos de dados a serem extraídos das sentenças não estruturadas. Dentre as variáveis consideradas, destaca-se o dano moral como a variável-alvo (*target*) deste trabalho, isto é, aquela que se pretende prever ao utilizar os dados extraídos como entrada do modelo preditivo a ser desenvolvido na Frente 3 do projeto.
2. **Instruções de Formato de Saída:** Para garantir que a resposta do modelo pudesse ser processada e avaliada de forma automatizada, foram fornecidas instruções sobre a formatação da saída. Esta especificação é uma aplicação direta das técnicas de Customização de Saída (*Output Customization*), discutidas na Subseção 2.6.3. A resposta deveria ser exclusivamente um bloco de texto em formato CSV, contendo exatamente duas linhas. A primeira linha serviria como cabeçalho (identificadores numéricos de 1 a 15) e a segunda conteria os valores extraídos. Esta abordagem implementa o padrão de prompt conhecido

como *Template*, que força a saída a seguir um modelo pré-definido, buscando consistência e facilitando a análise computacional subsequente (Pereira et al., 2025)

3. **Definição das Variáveis (O Contexto do Prompt):** Após as instruções gerais, o prompt listava cada um dos 15 fatores. Este pilar, que descreve o significado de cada atributo a ser procurado, funciona como o Contexto do prompt, conforme a estrutura teórica da Subseção 2.6.1. Ele fornece ao modelo o conhecimento de domínio específico necessário para compreender a tarefa em profundidade. As instruções nesta versão inicial eram traduções diretas e concisas das definições utilizadas para a realização do *ground truth*, como, por exemplo, "indique 'S' para sim ou 'N' para não se o consumidor tentou cancelar ou alterar a compra no prazo de 7 dias" para o fator "direito de arrependimento", critérios utilizados na coleta pelos especialistas (Pereira et al., 2025).

4.4.1.2 O Processo Iterativo de Refinamento

A realização de experimentos exploratórios utilizando o *Prompt Inicial* evidenciou que, embora funcional, a formulação simplificada de determinadas diretrizes ocasionava interpretações e resultados inconsistentes. Tal cenário se manifestava especialmente quando da extração de elementos jurídicos dotados de maior complexidade conceitual (Pereira et al., 2025). Essa constatação empírica corrobora a argumentação teórica desenvolvida na Subseção 2.6.2.

Nesse sentido, a qualidade das respostas geradas pelo sistema encontra-se vinculada à precisão e clareza das instruções fornecidas pelo operador. Portanto, o aperfeiçoamento na estruturação das orientações mostra-se útil para ampliar a acurácia da extração e mitigar a ocorrência de respostas inadequadas ou fenômenos de alucinação.

Em resposta a essa problemática identificada, iniciou-se um procedimento de revisão no qual especialistas da área procederam à análise dos equívocos de extração. O propósito consistia em formular hipóteses explicativas sobre as origens desses erros, considerando possíveis causas como instruções formuladas de maneira ambígua, insuficiência de exemplos contextuais ou carência de critérios explícitos (Sabo et al., 2025).

A mudança para a versão definitiva do prompt constituiu, dessa forma, um ciclo metodológico orientado à superação dessas limitações interpretativas. Observou-se que o desempenho do modelo sofria impacto decorrente da natureza polissêmica de terminologias jurídicas. Tais termos, apesar de mais claros para analistas humanos especializados, apresentavam ambiguidade na interpretação dos LLMs.

A atividade demandou uma delimitação tanto do significado quanto do escopo de cada fator analisado. Conceitos abstratos, exemplificados pela expressão assistência adequada, necessitaram ser transformados em conjuntos de critérios passíveis de verificação objetiva. O objetivo central residia em decompor aquilo que caracteriza um ponto de referência em uma decisão judicial, considerando a multiplicidade de terminologias e as extensões semânticas empregadas nos textos sentenciados.

Uma dificuldade emergiu durante a extração das dimensões temporais, isto é, dos intervalos de tempo. Diferentemente dos fatores de natureza categórica, a obtenção de valores numéricos contínuos a partir de descrições textuais heterogêneas exigia uma capacidade de processamento e conversão que o prompt inicial não estimulava de maneira consistente. Para tratar essa limitação, empregou-se explicitamente a estratégia de *few-shot prompting*, conforme a Seção 2.6.3.

A incorporação de exemplos ilustrativos de pares entrada-saída, como a conversão de noventa minutos para o formato temporal 01:30, no corpo estrutural do prompt funcionou como um mecanismo de aprendizagem contextual. Esse procedimento instruiu o modelo a replicar um processo de raciocínio lógico específico e a aderir a um padrão de saída previamente estabelecido, buscando consistência e uniformidade nos resultados obtidos ao longo de toda a extração.

Nesse contexto, o ciclo de análise dos erros e refinamento das instruções mostrou-se necessário para suas correções. A metodologia descrita em (Sabo et al., 2025) contribuiu para aprimorar a precisão e a confiabilidade do processo de extração automatizada. Esse processo iterativo, fundamentado em testes empíricos e ajustes, culminou no desenvolvimento do *Prompt Final*, que será objeto de análise na Seção subsequente.

4.4.1.3 O Prompt Refinado (Prompt Final): Análise das Melhorias

O *Prompt Final*, apresentado na íntegra no Apêndice A.2, representa a aplicação dos ajustes e a aplicação de técnicas de prompt, discutidas na Seção 2.6. As mudanças implementadas visaram reduzir a ambiguidade e estabelecer critérios alinhados às técnicas e descobertas no âmbito da engenharia de prompts. As principais melhorias, expostas em Pereira et al. (2025), foram:

- **Refinamento do Pro-Factor 1 (Extravio Definitivo de Bagagem):**
 - **Problema Identificado:** Em casos onde sentenças mencionavam bagagem ”extraviada” temporariamente, o modelo interpretava como perda definitiva, não compreendendo pelo contexto que a restituição posterior havia ocorrido.
 - **Solução Implementada:** Estabeleceu-se um protocolo de avaliação com critérios adicionais que passou a exigir evidências textuais para caracterização dos fatores. As modificações no prompt instruíram o modelo a buscar termos específicos como ”bagagem perdida permanentemente” e, fundamentalmente, a descartar a hipótese de perda definitiva sempre que houvesse qualquer referência à recuperação posterior do bem (Pereira et al., 2025). Dessa forma, os casos paradigmáticos que geravam erros sistemáticos foram delimitados diretamente no contexto do prompt mediante aplicação da técnica *few-shot prompting*.
- **Refinamento do Pro-Factor 2 (Extravio Temporário de Bagagem):**

- **Problema Identificado:** A análise binária (sim/não) era insuficiente para os cenários possíveis.
 - **Solução Implementada:** A instrução evoluiu para uma estrutura de análise com etapas e critérios de exclusão. O prompt passou a guiar o modelo a verificar sequencialmente: (i) se há menção ao extravio, (ii) se há informação sobre a recuperação, e (iii) se a bagagem foi declarada como permanentemente perdida (Pereira et al., 2025). Esta abordagem se assemelha a uma "cadeia de pensamento" (*chain-of-thought*) implícita, introduzido na Seção 2.5.4, que se mostrou útil para melhorar a precisão em certos casos."
- **Refinamento do Pro-Factor 5 (Cancelamento de Voo) e Pro-Factor 10 (Assistência da Companhia Aérea):**
 - **Problema Identificado:** Os conceitos de "cancelamento" e "assistência" eram abstratos e subjetivos.
 - **Solução Implementada:** As instruções foram redefinidas para incluir critérios específicos e concretos (e.g., "hospedagem, alimentação ou transporte alternativo"). Esta técnica de detalhamento da instrução, conforme exposto na Subseção 2.6.1, é importante para reduzir a ambiguidade e direcionar o foco do modelo para os aspectos juridicamente relevantes do conceito (Pereira et al., 2025).
 - **Refinamento das Dimensions 1 e 2 (Intervalos de Extravio e Atraso):**
 - **Problema Identificado:** A instrução inicial para registrar o tempo resultava em formatos inconsistentes, dificultando a análise quantitativa.
 - **Solução Implementada:** O prompt foi modificado com a aplicação da técnica de *few-shot prompting*, discutida na Subseção 2.6.3. Foram fornecidos exemplos explícitos de conversão (e.g., "2 horas e 30 minutos" para "02:30"; "90 minutos" para "01:30"). Esta abordagem, que guia o modelo com exemplos de pares entrada-saída, mostra-se eficaz para tarefas de formatação. Adicionalmente, foram incluídas diretrizes para casos ambíguos ("null") e um passo de validação, instruindo o modelo a revisar sua própria saída (Pereira et al., 2025).

Em suma, o processo de engenharia de prompt demonstrou ser a etapa central da metodologia. Sua evolução, que será quantitativamente avaliada no Capítulo 5, foi importante para mitigar a ambiguidade e melhorar a precisão da extração automática de fatores jurídicos.

4.4.1.4 Desenho Experimental

Com o propósito de avaliar a eficácia da engenharia de prompt neste *Experimento Principal* e a capacidade do modelo em cenários caracterizados por diferentes graus de complexidade,

foram concebidos dois experimentos distintos. Cada experimento testa uma estratégia de prompting diferenciada, e a descrição de ambos constitui o elemento central para a análise comparativa que será conduzida no Capítulo 5, permitindo a compreensão das abordagens metodológicas empregadas.

4.4.1.4.1 Experimento 1: Abordagem *Singular*

O primeiro experimento, designado como *Singular*, estabeleceu como objetivo a avaliação do desempenho do modelo mediante a utilização de um prompt único, o qual continha as instruções necessárias para a extração simultânea de todos os 15 fatores (Pereira et al., 2025). Esta abordagem opera como a linha de base do estudo, testando a capacidade do LLM de processar uma carga cognitiva associada a tarefa em uma única requisição (Pereira et al., 2025).

Uma característica metodológica importante deste experimento consistiu na implementação de um protocolo de votação destinado a garantir a robustez dos resultados obtidos. Considerando a natureza variável dos LLMs, os quais podem produzir respostas com variações para a mesma entrada, foram executadas três requisições independentes à API para cada documento submetido à análise (Pereira et al., 2025).

A resposta final referente a cada fator foi então estabelecida pela moda estatística, definida como o valor que ocorre com maior frequência em um conjunto de dados, dos três resultados obtidos, empregando-se a primeira ocorrência como critério de desempate. Este protocolo visa mitigar a variabilidade do modelo e incrementar a confiabilidade da extração final (Pereira et al., 2025). Considerando os 900 documentos analisados, as duas versões de prompt (*Inicial* e *Final*) e as três requisições executadas por documento, este experimento totalizou 5.400 chamadas à API (Pereira et al., 2025).

4.4.1.4.2 Experimento 2: Abordagem *Factor-Based Sets*

O segundo experimento, denominado *Factor-Based Sets*, foi estruturado para investigar a hipótese de que prompts caracterizados por maior concisão e foco poderiam conduzir a um desempenho superior, particularmente para fatores de extração mais complexa (Pereira et al., 2025). Para testar essa hipótese, o prompt original foi segmentado em três componentes distintos, seguindo a classificação teórica do FBR estabelecida na Subseção 4.3.1: um prompt destinado exclusivamente aos *pro-factors*, outro para os *con-factors* e um terceiro para as *dimensions* (Pereira et al., 2025).

A fundamentação para este desenho experimental apresentou dupla natureza. Primeiramente, sob a perspectiva técnica, a segmentação poderia reduzir a carga cognitiva imposta ao modelo, possibilitando que este concentrasse seus recursos em um subconjunto de tarefas (Pereira et al., 2025). Em segundo lugar, e assumindo relevância sob o prisma jurídico-prático, esta abordagem viabiliza uma análise específica de risco (Pereira et al., 2025).

Um erro na extração de um *pro-factor*, exemplificado por um falso positivo que poderia resultar na superestimação de uma indenização, apresenta implicações distintas de um erro verificado em um *con-factor*, tal como um falso positivo que poderia ocasionar a redução de uma indenização (Pereira et al., 2025). Ao isolar as categorias, torna-se viável avaliar a acurácia do modelo para cada tipologia de fator e identificar os segmentos que demandam validação humana, otimizando um fluxo de trabalho de anotação em larga escala (Pereira et al., 2025). Este experimento também totalizou 5.400 requisições (900 documentos $\times$ 3 segmentos $\times$ 2 versões de prompt) (Pereira et al., 2025).

4.4.2 Experimento Secundário: Sub-Hipóteses de Otimização

4.4.2.1 Formulação das Sub-Hipóteses de Otimização

O exame dos dados obtidos no Experimento Principal, cuja apresentação encontra-se no Capítulo 5, evidenciou a necessidade de conduzir uma etapa adicional de aprimoramento metodológico. O impulso para esta nova rodada investigativa originou-se das observações derivadas da estratégia *Singular*, em que se utilizou um prompt unificado para proceder à extração concomitante da totalidade dos 15 fatores.

Conforme será demonstrado na Subseção 5.1.1, os resultados dessa estratégia apontaram que a performance do modelo, embora apresentasse robustez no conjunto, exibiu flutuações quando analisada fator a fator. Tais flutuações revelaram-se pronunciadas em elementos caracterizados por baixa frequência no *corpus* ou por complexidade semântica, a exemplo das dimensões de caráter numérico.

Nesse sentido, tal observação, alinhada às limitações apresentadas na Seção 5.3, subsidiou a elaboração de novas sub-hipóteses voltadas ao refinamento específico de algumas regras extrativas. A fim de submeter essas hipóteses à verificação, os novos experimentos foram igualmente executados mediante a utilização de um prompt unificado, estratégia que define a abordagem *Singular*.

Para orientar esta etapa da investigação, foram formuladas as seguintes sub-hipóteses:

1. A demanda cognitiva, resultante da identificação simultânea de quinze fatores distintos, prejudica a acurácia do modelo em fatores de complexidade ou menor frequência. Postula-se que a consolidação de fatores semanticamente correlatos, como *Atraso de Voo* e *Cancelamento*, constitui uma estratégia para mitigar essa limitação.
2. O desempenho na extração de fatores individuais com performance insatisfatória pode ser aprimorado por meio de abordagens de engenharia de prompt direcionadas. Tais abordagens incluem o aperfeiçoamento das diretrizes para conceitos ambíguos e a reestruturação da variável de saída, por exemplo, sua conversão de um formato binário para categórico ou numérico.

Fundamentado nessas sub-hipóteses, inaugurou-se um processo iterativo adicional de concepção e avaliação de múltiplas versões renovadas de prompt, organizadas em famílias experimentais designadas como atraso e extravio. A finalidade consistia em verificar se uma metodologia mais especializada poderia superar o desempenho do *Prompt Final*, aqui denominado *main* para propósitos comparativos, o qual havia sido calibrado para alcançar um desempenho geral equilibrado.

4.4.2.2 *Desenho Experimental: Variações de Prompts e Amostras de Dados*

Os experimentos foram conduzidos em três conjuntos distintos de dados, denominados *C1*, *C2* e *C3*, todos com 300 (trezentas) sentenças cada. Estes conjuntos correspondem a sentenças anotadas em diferentes momentos do projeto. A estratificação temporal dos dados é relevante porque permite avaliar a robustez do modelo e sua capacidade de generalização diante de possíveis variações no estilo de redação ou na complexidade dos casos ao longo do tempo.

Adicionalmente, essa estrutura permite verificar se o modelo apresenta viés induzido por testes anteriores realizados com o mesmo *corpus* de sentenças. Em outras palavras, busca-se confirmar que as respostas do modelo não são influenciadas pelo fato de ele ter processado previamente aqueles dados específicos.

As principais alterações testadas, cujas versões completas estão disponíveis no Apêndice deste trabalho, foram organizadas da seguinte forma:

- **Atraso1 (Apêndice A.3):** Esta versão estabeleceu uma linha de base diferente, mantendo os três fatores independentes relacionados a problemas de transporte aéreo. O objetivo foi reduzir ambiguidades por meio do refinamento das instruções para os fatores *Cancelamento sem Realocação ou Alteração de Destino*, *Atraso de Voo* e *Intervalo de Atraso*. As modificações esclareceram a distinção entre cancelamento, quando o consumidor não alcança o destino em tempo hábil, e atraso, quando há chegada tardia mas o destino é alcançado. A variável de intervalo manteve o formato original em horas e minutos ou hífen para casos de ausência. Esta configuração serviu como controle metodológico para avaliar melhorias obtidas apenas por refinamento textual, sem alterações estruturais.
- **Atraso3 (Apêndice A.4):** Esta variante implementou uma mudança ao unificar os fatores binários *Atraso* e *Cancelamento* em um único fator denominado *Atraso ou Cancelamento*. A hipótese testada foi a de que essa simplificação reduziria a carga cognitiva do modelo. A inovação consistiu na reconfiguração da variável *Intervalo de Atraso ou Cancelamento* com lógica de três estados: registra o tempo de atraso em horas e minutos quando há viagem realizada, indica valor negativo um para cancelamentos que impediram a viagem, e registra zero para ausência de problemas. Esta codificação permite distinguir três situações distintas em uma única coluna, facilitando a interpretação dos dados.

- **Atraso4 (Apêndice A.5):** Mantendo a fusão dos fatores binários em *Atraso ou Cancelamento*, esta versão testou uma hipótese alternativa sobre a representação numérica. A modificação principal ocorreu na variável *Intervalo de Atraso*, que passou a registrar zero tanto para cancelamentos quanto para ausência de problemas, em vez de usar valor negativo um para cancelamentos. Esta lógica registra valores positivos em horas e minutos apenas quando há atraso com viagem realizada, ou zero nos demais casos. O objetivo foi investigar se a diferenciação numérica do Atraso3 introduzia complexidade ao processo de extração.
- **Atraso5 (Apêndice A.6):** Esta configuração unificou os três fatores originais (*Cancelamento*, *Atraso de Voo* e *Intervalo de Atraso*) em uma única variável denominada *Atraso*, *Cancelamento ou Horário Regular*. Esta variável consolidada eliminou as colunas binárias intermediárias, concentrando toda a informação em um único campo com lógica de três estados: valor negativo um para cancelamentos, intervalo em horas e minutos para atrasos, e zero para ausência de problemas. O objetivo foi testar a redução da carga cognitiva, considerando que solicitar o preenchimento de apenas uma coluna permitiria melhor compreensão contextual e maior precisão.
- **Extravio1 (Apêndice A.7):** De forma análoga ao Atraso1, esta versão estabeleceu uma mudança do prompt final para problemas relacionados a bagagens. O objetivo foi refinar as instruções dos três fatores independentes: *Extravio Definitivo de Bagagem*, *Extravio Temporário de Bagagem* e *Intervalo de Extravio Temporário*. As modificações especificaram os critérios para classificar o extravio como definitivo, quando não há recuperação, ou temporário, quando há recuperação. A extração do intervalo temporal foi mantida no formato de horas e minutos. Esta configuração serviu como referência para avaliar ganhos obtidos apenas por meio do refinamento das instruções.
- **Extravio2 (Apêndice A.8):** Esta versão unificou a variável binária *Extravio Temporário* com *Intervalo de Extravio* em um único fator denominado *Intervalo de Recuperação de Bagagem Em Caso de Extravio Temporário e Extravio Definitivo*. A reconfiguração resultou em uma lógica de dois estados: registra o intervalo em horas e minutos quando há extravio temporário com recuperação, e registra zero nos demais casos, incluindo ausência de extravio, extravio sem informação de recuperação, ou extravio definitivo. O objetivo foi testar se a redução de três para duas colunas melhoraria a precisão ao diminuir a complexidade da tarefa de extração. Esta versão manteve *Extravio Definitivo* como variável binária independente.

Assim, a apresentação destas variações de prompts e amostras de dados conclui o desenho metodológico para o Experimento Secundário. A análise quantitativa e qualitativa do desempenho de cada uma dessas estratégias de otimização, bem como a consequente validação das hipóteses formuladas, serão apresentadas em detalhe na Seção 5.2.

4.5 MODELOS, PARÂMETROS E AMBIENTE EXPERIMENTAL

O modelo de linguagem empregado em todos os experimentos foi o GPT-4o, em sua versão de pré-visualização 2024-08-06, disponibilizado pela OpenAI (Pereira et al., 2025). Uma decisão metodológica foi a de interagir com o modelo exclusivamente através de sua API (Pereira et al., 2025).

Os parâmetros da API foram configurados conforme as recomendações da documentação oficial para tarefas de extração, visando um equilíbrio entre precisão e consistência (Pereira et al., 2025). Os principais parâmetros ajustados foram:

- **Temperatura (*Temperature*):** Definida em 0.7. O valor de 0.7 foi escolhido como um meio-termo para permitir que o modelo lidasse com a variabilidade da linguagem jurídica sem se tornar imprevisível ou rígido (Pereira et al., 2025).
- **Top_p (*Nucleus Sampling*):** Definido em 1. Este parâmetro instrui o modelo a considerar todos os tokens possíveis em cada etapa de geração, não limitando artificialmente seu vocabulário de resposta, o que é necessário para capturar termos jurídicos específicos (Pereira et al., 2025).
- **Penalidade de Frequência e Presença (*Frequency and Presence Penalty*):** Ambos definidos em 0. Como a tarefa era de extração estruturada e não de geração de texto longo e criativo, estes parâmetros foram mantidos em seu valor neutro para não penalizar a repetição de termos técnicos ou respostas comuns (e.g., "sim" ou "não") (Pereira et al., 2025).

O ambiente computacional para a execução dos experimentos foi composto por ferramentas de código aberto. A automação das requisições à API, o processamento das respostas e a comparação com o *ground truth* foram implementados em linguagem Python (Pereira et al., 2025). A biblioteca oficial da OpenAI para Python foi utilizada para facilitar e agilizar a comunicação com o modelo. Todo o código foi desenvolvido e executado no ambiente de desenvolvimento integrado *Microsoft Visual Studio Code* (VS Code) (Pereira et al., 2025). Ressalta-se, novamente, que a configuração técnica e a execução dos experimentos foram fruto de uma colaboração interdisciplinar entre os pesquisadores da área jurídica e da ciência da computação que compõem a equipe do projeto CONCIL-IA (Pereira et al., 2025).

4.6 MÉTRICAS DE AVALIAÇÃO

A avaliação do desempenho de um modelo dedicado à extração de informações não se restringe a uma única medida de sucesso (Pereira et al., 2025). Revela-se crucial a seleção de um conjunto de métricas que, em sua totalidade, ofereçam uma visão abrangente da performance, especialmente quando se trata de lidar com as complexidades características de um dataset jurídico (Pereira et al., 2025).

Conforme as práticas da literatura especializada (Aggarwal, 2018; Chai; Draxler, 2014), a estratégia de avaliação foi segmentada em conformidade com a natureza de cada fator, estabelecendo distinção entre a avaliação de fatores categóricos e numéricos. Esta diferenciação permite que sejam aplicadas métricas específicas e apropriadas para cada tipo de dado, garantindo uma análise quantitativa às características particulares de cada categoria de variável extraída.

4.6.1 Avaliação de Fatores Categóricos

Para os 13 fatores de natureza categórica (*pro-factors* e *con-factors*), a avaliação foi realizada com base em quatro métricas de classificação. A escolha deste conjunto de métricas foi deliberada para superar as limitações de uma análise simples, especialmente em face do desafio do *corpus*: o desbalanceamento da quantidade em cada classe conforme demonstrado na distribuição de dados da Subseção 4.3.2.

Todas essas métricas derivam dos quatro resultados possíveis de uma Matriz de Confusão, que compara a previsão do modelo com a realidade:

1. Verdadeiro Positivo (*True Positive*) (TP): A previsão positiva do modelo estava correta.
2. Falso Positivo (*False Positive*) (FP): O modelo previu "positivo", mas errou (o real era negativo).
3. Verdadeiro Negativo (*True Negative*) (TN): A previsão negativa do modelo estava correta.
4. Falso Negativo (*False Negative*) (FN): O modelo previu "negativo", mas errou (o real era positivo).

A partir destes componentes, as métricas avaliam o desempenho sob diferentes óticas:

- **Acurácia (*Accuracy*):** Embora seja a métrica mais intuitiva, representando a porcentagem geral de predições corretas (Pereira et al., 2025), a acurácia é enganosa em datasets desbalanceados. Por exemplo, para o fator Descumprimento de Oferta, que ocorre em apenas 6 dos 900 casos (0.67%), um modelo que previsse "não" para todas as instâncias alcançaria uma acurácia superior a 99%, sendo, na prática, inútil para a tarefa de identificar o evento de interesse (Pereira et al., 2025). Portanto, a acurácia é utilizada aqui como uma medida de performance geral, mas sua interpretação é feita em conjunto com as outras métricas. Como será demonstrado na análise dos resultados no Capítulo 5, este fenômeno se manifesta claramente, com o modelo atingindo alta acurácia geral, mas com um F1-Score mais baixo.

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.1)$$

- **Precisão (*Precision*):** Esta métrica responde a uma pergunta relevante do ponto de vista da confiabilidade: "Das vezes que o modelo afirmou que um fator estava presente, quantas

vezes ele estava correto?” (Pereira et al., 2025). A precisão mede a proporção de TP em relação a todas as predições positivas (TP + FP) (Pereira et al., 2025). Uma alta precisão é importante para minimizar os falsos positivos, que, no contexto jurídico, poderiam levar a conclusões incorretas sobre a presença de um fato relevante para a condenação (Pereira et al., 2025).

$$\text{Precisão} = \frac{TP}{TP + FP} \quad (4.2)$$

- **Revocação (*Recall*):** Complementar à precisão, o *recall* responde à pergunta sobre a completude da extração: ”De todos os casos em que um fator realmente estava presente, quantos o modelo conseguiu encontrar?” (Pereira et al., 2025). Ele mede a proporção de TP em relação a todas as instâncias que eram de fato positivas (TP + FN) (Pereira et al., 2025). Um alto *recall* é importante para garantir que o modelo não deixe de identificar ocorrências, minimizando os falsos negativos.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4.3)$$

- **F1-Score:** Esta é, possivelmente, a métrica mais relevantes para a avaliação de fatores categóricos neste estudo e o do referência (Pereira et al., 2025). O F1-Score é a média harmônica entre a Precisão e o *Recall*, combinando ambas as perspectivas em um único número (Aggarwal, 2018). Sua principal característica é penalizar modelos que otimizam uma métrica em detrimento da outra. No contexto de um dataset desbalanceado, o F1-Score se torna o indicador da capacidade do modelo de identificar a classe minoritária (os casos ”sim”), que é, em última análise, o objetivo da tarefa de extração (Pereira et al., 2025). Um F1-Score baixo, mesmo com uma acurácia alta, sinaliza que o modelo está falhando em sua tarefa principal.

$$\text{F1-Score} = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (4.4)$$

4.6.2 Avaliação de Fatores Numéricos

Para as duas *dimensions* (”Intervalo de extravio” e ”Intervalo de atraso”), que representam valores de tempo contínuos, a avaliação requer uma métrica que quantifique a magnitude do erro, em vez de uma simples classificação de acerto/erro (Pereira et al., 2025).

- **RMSE:** O RMSE foi a métrica escolhida para esta tarefa. Ele calcula a raiz quadrada da média das diferenças ao quadrado entre os valores preditos pelo modelo (y_i) e os valores reais do *ground truth* ($\hat{y}_i$) (Pereira et al., 2025). A principal razão para a escolha do RMSE em detrimento de outras métricas de erro, como o *Mean Absolute Error* (Erro Absoluto Médio) (MAE), é sua sensibilidade a grandes erros (Pereira et al., 2025). Ao elevar as diferenças ao quadrado, o RMSE penaliza os grandes erros. No contexto jurídico, essa característica é relevante para o contexto: um erro de extração de 10 horas em um atraso de

voo é mais grave do que um erro de 10 minutos (Pereira et al., 2025). O RMSE reflete essa gravidade em seu resultado, fornecendo um indicador do impacto dos erros mais relevantes (Chai; Draxler, 2014). Para a aplicação desta métrica, foi estabelecida uma margem de tolerância de uma hora; diferenças menores não foram consideradas erros (Pereira et al., 2025).

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (4.5)$$

5 RESULTADOS E DISCUSSÃO

Este capítulo apresenta e analisa os resultados empíricos obtidos a partir dos experimentos metodológicos descritos no Capítulo 4. A análise inicia-se, na Seção 5.1, com os resultados do Experimento Principal. Esta seção aborda o desempenho das abordagens *Singular* e *Factor-Based Sets*, analisando o *trade-off* entre a robustez geral e a precisão na extração de fatores minoritários. Prosseguindo, a Seção 5.2 apresenta a análise do Experimento Secundário, avaliando comparativamente o desempenho das novas versões de prompts e validando as hipóteses de otimização. Após a exposição dos achados de ambos os experimentos, a Seção 5.3 realiza uma avaliação das limitações do estudo, contextualizando os resultados e reconhecendo os contornos da pesquisa. Por fim, olhando para o futuro, a Seção 5.4 explora novos métodos para a evolução do projeto, investigando os avanços na formalização tanto das entradas, com o uso de Markdown, quanto das saídas, por meio do JSON Schema, o que delineia a evolução metodológica para trabalhos futuros.

5.1 ANÁLISE DO EXPERIMENTO PRINCIPAL

5.1.1 Análise do Experimento com Abordagem *Singular*

Conforme detalhado na Subsubseção 4.4.1.4.1, o experimento *Singular* baseou-se na utilização de um prompt único destinado à extração simultânea de todos os 15 fatores. Esta Seção dedica-se à apresentação dos resultados de desempenho tanto em sua dimensão geral quanto na análise individualizada por fator para esta abordagem metodológica, seguida de uma discussão dos principais achados identificados ao longo do processo de avaliação experimental.

5.1.1.1 Desempenho Geral e por Fator

O desempenho geral do modelo no experimento *Singular* foi medido pela Acurácia e pelo F1-Score globais, comparando as versões *Inicial* e *Final* do prompt. Os resultados estão consolidados nas Tabelas 2 e 3.

Tabela 2 – Acurácia Geral (%) no Experimento *Singular*.

Experimento	Documentos	Prompt	Acurácia
Singular	900	Inicial	89.2
Singular	900	Final	91.5

Fonte: (Pereira et al., 2025).

Para uma análise mais específica, é útil avaliar o desempenho do modelo em cada fator individualmente. As Tabelas 4, 5, e 6 apresentam, respectivamente, a acurácia por fator, as métricas de F1-Score, Precisão e *Recall* para os fatores categóricos, e o RMSE para as *dimensions*. Os valores destacados em negrito indicam o resultado superior entre as duas versões

Tabela 3 – F1-Score Geral no Experimento *Singular*.

Experimento	Documentos	Prompt	F1-Score
Singular	900	Inicial	0.58
Singular	900	Final	0.59

Fonte: (Pereira et al., 2025).

do prompt para cada métrica avaliada, facilitando a identificação dos melhores desempenhos obtidos.

Tabela 4 – Acurácia Individual por Fator (%) no Experimento *Singular*.

Fator	Prompt Inicial	Prompt Final
(Pro-factor) Direito de arrependimento	95.3	93.6
(Pro-factor) Descumprimento de oferta	92.2	93.3
(Pro-factor) Extravio definitivo de bagagem	98.4	99.1
(Pro-factor) Extravio temporário de bagagem	96.6	95.7
(Dimension) Intervalo de extravio	88.2	87.1
(Pro-factor) Bagagem violada	97.8	97.4
(Pro-factor) Cancelamento de voo	65.7	81.4
(Pro-factor) Atraso de voo	80.1	88.3
(Dimension) Intervalo de atraso	75.6	86.1
(Con-factor) Culpa do consumidor	95.3	95.0
(Con-factor) Condições climáticas adversas	97.8	97.4
(Pro-factor) No show	98.7	98.3
(Pro-factor) Overbooking	99.0	99.1
(Pro-factor) Assistência da companhia aérea	66.1	68.8
(Pro-factor) Hipervulnerabilidade do consumidor	91.9	92.1

Fonte: (Pereira et al., 2025)

5.1.1.2 Discussão dos Achados

O exame dos dados obtidos no experimento *Singular* proporciona compreensões acerca do desempenho do modelo e da efetividade das estratégias de engenharia de prompt implementadas. As Tabelas 2 e 3 evidenciam uma elevação da acurácia global, que alcançou 91,5% mediante o emprego do prompt aprimorado, ao passo que o F1-Score permaneceu estável em 0,59.

Conforme foi discutido na Subseção 4.6.1 ao abordar as métricas de avaliação, tal fenômeno reflete o desequilíbrio entre as classes presentes no conjunto de dados. A obtenção de

Tabela 5 – F1-Score (F1-S), Precisão (Prec) e Recall (Rec) por Fator no Experimento *Singular*.

Fator	Prompt Inicial			Prompt Final		
	F1-S	Prec	Rec	F1-S	Prec	Rec
Direito de arrependimento	0.25	0.16	0.58	0.17	0.10	0.50
Descumprimento de oferta	0.10	0.05	0.67	0.12	0.06	0.67
Extravio definitivo	0.76	0.65	0.92	0.83	0.83	0.83
Extravio temporário	0.88	0.82	0.95	0.85	0.77	0.95
Bagagem violada	0.80	0.68	0.98	0.77	0.66	0.93
Cancelamento de voo	0.43	0.28	0.85	0.53	0.43	0.67
Atraso de voo	0.79	0.89	0.70	0.89	0.89	0.88
Culpa do consumidor	0.40	0.28	0.74	0.37	0.25	0.68
Condições climáticas	0.60	0.43	1.00	0.56	0.39	1.00
No show	0.84	0.77	0.91	0.78	0.77	0.79
Overbooking	0.84	0.80	0.89	0.86	0.83	0.89
Assistência da Cia. Aérea	0.43	0.40	0.47	0.41	0.47	0.37
Hipervulnerabilidade	0.43	0.28	0.87	0.43	0.29	0.87

Fonte: (Pereira et al., 2025).

Tabela 6 – RMSE Individual por Dimensão no Experimento *Singular*.

Dimensão	Prompt Inicial	Prompt Final
Intervalo de extravio	491.2	499.9
Intervalo de atraso	24.8	6.7

Fonte:(Pereira et al., 2025).

acurácia de 91,5% pelo modelo resulta primordialmente de sua capacidade de prever corretamente a classe predominante, que corresponde frequentemente à categoria negativa.

Entretanto, o F1-Score, por sua sensibilidade às variações entre classes, evidencia a limitação do modelo em detectar a classe minoritária, representada pela categoria afirmativa, que constitui o objetivo do processo de extração de informações (Pereira et al., 2025). A técnica de refinamento do prompt, cujos procedimentos foram descritos nas Subseções 4.4.1.2 e 4.4.1.3, demonstrou repercussões positivas nos fatores investigados.

Registrou-se uma progressão da acurácia global de 89,2% para 91,5%, sendo observados os incrementos observados em fatores específicos. A acurácia do fator identificado como cancelamento de voo apresentou elevação de 65,7% para 81,4%, enquanto o fator relativo ao atraso de voo exibiu progressão de 80,1% para 88,3%. Simultaneamente, o F1-Score associado ao atraso de voo progrediu de 0,79 para 0,89.

Uma das principais melhorias, manifestou-se na métrica RMSE referente à dimensão intervalo de atraso, que apresentou redução de 24,8 para 6,7. Tal resultado indica que as instruções mais detalhadas e os exemplos de formatação incorporados ao *Prompt Final* alcançaram eficácia na extração de informações numéricas. Este resultado pode ser associada à implementação da técnica de poucos exemplos (*few-shot prompting*), conforme detalhado na Subseção 4.4.1.3.

A investigação individualizada por fator apresenta possível correlação entre o equilíbrio

entre as classes e a performance apresentada pelo modelo. O fator atraso de voo, que apresenta o F1-Score mais elevado com valor de 0,89, caracteriza-se igualmente como o mais equilibrado do conjunto de dados, exibindo distribuição de 52% de casos afirmativos e 48% de casos negativos (Pereira et al., 2025). Inversamente, os fatores com menores desempenhos revelam-se os mais escassos no conjunto de dados.

O fator descumprimento de oferta, com F1-Score de 0,12, e o fator direito de arrependimento, com F1-Score de 0,17, ilustram esta tendência. A precisão do fator descumprimento de oferta, situada em 0,06, indica que as predições positivas do modelo apresentam taxa de erro de 94%, o que torna suas classificações afirmativas essencialmente desprovidas de confiabilidade para fins práticos (Pereira et al., 2025).

O desempenho observado no fator assistência da companhia aérea, que registrou F1-Score de 0,41, exemplifica um desafio de natureza distinta: a imprecisão conceitual. A inexistência de parâmetros normativos claros que estabeleçam os critérios constitutivos de uma assistência apropriada reflete-se na heterogeneidade das decisões judiciais e, por conseguinte, na dificuldade do modelo para identificar um padrão consistente.

Tal limitação persiste mesmo quando se empregam prompts mediante a técnica de poucos exemplos e a provisão de contexto ampliado. Esta constatação sugere que o desempenho dos modelos de linguagem de grande escala não depende exclusivamente da qualidade do prompt ou do balanceamento dos dados, mas também da precisão e consistência do próprio conceito jurídico objeto da extração. Este resultado materializa a polissemia, cuja discussão teórica foi apresentada na Seção 2.4.

Ademais, este aspecto foi identificado como um dos elementos propulsores do processo de refinamento do prompt na Subseção 4.4.1.2, demonstrando que a engenharia de prompt, conquanto seja uma ferramenta útil, encontra limites inerentes à própria natureza dos conceitos jurídicos que se busca extrair. Portanto, a clareza do domínio jurídico constitui um pré-requisito para o sucesso da aplicação de técnicas de processamento de linguagem natural neste campo.

5.1.2 Análise do Experimento com Abordagem *Factor-Based Sets*

O segundo experimento, denominado *Factor-Based Sets*, foi projetado para testar a hipótese de que prompts mais curtos e focados poderiam alterar o desempenho do modelo, conforme justificado na Subsubseção 4.4.1.4.2. Nesta abordagem, o prompt foi segmentado em três partes (*pro-factors*, *con-factors* e *dimensions*), e cada uma foi enviada em uma requisição separada. Esta Seção apresenta os resultados quantitativos desta abordagem e discute suas implicações em comparação com o experimento *Singular*.

5.1.2.1 Desempenho Geral e por Fator

O desempenho geral do modelo no experimento *Factor-Based Sets* é apresentado nas Tabelas 7 e 8, que consolidam a acurácia e o F1-Score globais para as duas versões do prompt.

Tabela 7 – Acurácia Geral (%) no Experimento *Factor-Based Sets*.

Experimento	Documentos	Prompt	Acurácia
Factor-based	900	Inicial	78.6
Factor-based	900	Final	83.8

Fonte: (Pereira et al., 2025).

Tabela 8 – F1-Score Geral no Experimento *Factor-Based Sets*.

Experimento	Documentos	Prompt	F1-Score
Factor-based	900	Inicial	0.31
Factor-based	900	Final	0.43

Fonte: (Pereira et al., 2025).

A análise do desempenho por fator é apresentada nas tabelas a seguir. A Tabela 9 mostra a acurácia individual, a Tabela 10 detalha o F1-Score, a Precisão e o *Recall*, e a Tabela 11 exibe o RMSE para as *dimensions*. Novamente, os valores em negrito destacam o melhor resultado entre as duas versões do prompt.

Tabela 9 – Acurácia Individual por Fator (%) no Experimento *Factor-Based Sets*.

Fator	Prompt Inicial	Prompt Final
(Pro-factor) Direito de arrependimento	94.3	95.8
(Pro-factor) Descumprimento de oferta	90.8	94.8
(Pro-factor) Extravio definitivo de bagagem	98.1	99.2
(Pro-factor) Extravio temporário de bagagem	96.0	80.2
(Dimension) Intervalo de extravio	86.7	86.8
(Pro-factor) Bagagem violada	68.5	85.1
(Pro-factor) Cancelamento de voo	45.1	61.3
(Pro-factor) Atraso de voo	60.1	66.1
(Dimension) Intervalo de atraso	60.4	67.2
(Con-factor) Culpa do consumidor	68.2	87.8
(Con-factor) Condições climáticas adversas	90.2	92.3
(Pro-factor) No show	93.8	95.2
(Pro-factor) Overbooking	95.6	96.8
(Pro-factor) Assistência da companhia aérea	39.1	54.8
(Pro-factor) Hipervulnerabilidade do consumidor	91.8	93.8

Fonte: (Pereira et al., 2025).

Tabela 10 – F1-Score (F1-S), Precisão (Prec) e Recall (Rec) por Fator no Experimento *Factor-Based Sets*.

Fator	Prompt Inicial			Prompt Final		
	F1-S	Prec	Rec	F1-S	Prec	Rec
Direito de arrependimento	0.18	0.10	0.55	0.47	0.37	0.65
Descumprimento de oferta	0.10	0.05	0.67	0.39	0.26	0.79
Extravio definitivo	0.72	0.61	0.86	0.85	0.80	0.91
Extravio temporário	0.86	0.79	0.35	0.59	0.44	0.96
Bagagem violada	0.16	0.09	0.65	0.28	0.20	0.47
Cancelamento de voo	0.23	0.14	0.55	0.35	0.26	0.51
Atraso de voo	0.52	0.75	0.40	0.60	0.81	0.48
Culpa do consumidor	0.04	0.02	0.67	0.19	0.13	0.38
Condições climáticas	0.11	0.07	0.38	0.24	0.16	0.52
No show	0.38	0.28	0.56	0.51	0.42	0.63
Overbooking	0.43	0.37	0.50	0.44	0.41	0.48
Assistência da Cia. Aérea	0.16	0.23	0.12	0.20	0.63	0.12
Hipervulnerabilidade	0.18	0.13	0.25	0.52	0.56	0.49

Fonte: (Pereira et al., 2025).

Tabela 11 – RMSE Individual por Dimensão no Experimento *Factor-Based Sets*.

Dimensão	Prompt Inicial	Prompt Final
Intervalo de extravio	524.2	487.8
Intervalo de atraso	11.2	9.4

Fonte: (Pereira et al., 2025).

5.1.2.2 Discussão dos Achados

Os resultados obtidos no experimento denominado *Factor-Based Sets* apresentam um padrão de desempenho contra-intuitivo do modelo avaliado. Ao estabelecer uma comparação entre as Tabelas 7 e 8 e seus correspondentes na metodologia *Singular*, constata-se uma diminuição nas métricas de acurácia global, com redução do patamar máximo de 91.5% para 83.8%, enquanto o F1-Score agregado também experimentou declínio de 0.59 para 0.43 (Pereira et al., 2025).

A redução observada nas métricas globais pode ser atribuída, em grande medida, a um problema técnico durante a execução experimental (Pereira et al., 2025). Verificou-se que o modelo apresentou falhas na conformidade com as especificações de formatação da resposta, produzindo arquivos CSV com erro ou estruturas de dados inconsistentes em aproximadamente quatrocentas instâncias (Pereira et al., 2025). Tais inconformidades foram automaticamente classificadas como erros pelo algoritmo de avaliação, exercendo impacto negativo sobre as métricas consolidadas.

Nesse contexto, a partição do prompt parece ter comprometido a capacidade do modelo de manter consistência nas instruções de formatação, mesmo diante da presença explícita dessas diretrizes em todos os fragmentos do prompt segmentado (Pereira et al., 2025). Contudo, a análise das métricas por fator individual revela um achado não previsto.

Contrariamente ao esperado, a estratégia de segmentação demonstrou maior eficácia especificamente na extração de fatores caracterizados por raridade e desbalanceamento amostral (Pereira et al., 2025). A comparação dos resultados do *Prompt Final* entre ambas as configurações experimentais evidencia que o F1-Score associado ao fator denominado "Direito de arrendimento" apresentou elevação de 0.17 para 0.47, representando incremento absoluto de 0.30, ao passo que o fator "Descumprimento de oferta" registrou progressão de 0.12 para 0.39, com ganho de 0.27 (Pereira et al., 2025).

A precisão dos fatores minoritários mencionados também experimentou aprimoramento, corroborando a hipótese de que a segmentação promoveu um efeito sobre essa categoria de fatores (Pereira et al., 2025). Esse fenômeno sugere a existência de um *trade-off* entre as duas estratégias metodológicas investigadas. A metodologia *Singular* demonstra robustez e confiabilidade para a tarefa em sua totalidade, assegurando conformidade com especificações de formato e desempenho médio mais elevado (Pereira et al., 2025).

Por outro lado, a estratégia *Factor-Based Sets*, ao diminuir a carga cognitiva imposta ao modelo mediante focalização em subconjuntos reduzidos de tarefas simultâneas, aparenta ter viabilizado o acesso a capacidades de identificação precisa dos fatores minoritários (Pereira et al., 2025). Tal constatação indica que, embora não represente a solução metodológica como abordagem primária, a técnica de segmentação de prompts pode constituir estratégia investigativa para aprimorar a extração de conceitos jurídicos específicos que apresentem baixo desempenho em configurações unificadas (Pereira et al., 2025).

5.2 ANÁLISE DO EXPERIMENTO SECUNDÁRIO

Esta seção dedica-se à análise dos resultados obtidos no Experimento Secundário. Conforme exposto na Subseção 4.4.2, este ciclo de testes foi motivado pela necessidade de otimizar a extração de fatores que apresentaram desempenho insatisfatório no Experimento Principal. Para tanto, foram desenvolvidas e avaliadas múltiplas variações de prompts, organizadas nas famílias *Atraso* e *Extravio*, e testadas em diferentes amostras de dados, conforme o desenho experimental descrito na Subseção 4.4.2.2. Além disso, conforme também exposto na Subseção 4.4.2, os novos experimentos foram igualmente executados mediante a utilização de um prompt unificado, estratégia que define a abordagem *Singular*. A análise a seguir, portanto, tem o objetivo de comparar o desempenho dessas novas estratégias, tanto entre si quanto em relação ao *Prompt Final* (denominado *main*), a fim de validar as hipóteses de otimização.

5.2.1 Análise Comparativa de Desempenho

Os resultados consolidados de todos os experimentos são apresentados na Tabela 12. A análise revela um cenário de compensações entre diferentes estratégias, onde cada prompt se destaca em métricas específicas e apresenta desempenhos variados conforme a amostra de dados utilizada.

Diferentemente da análise do Experimento Principal, onde as métricas de Precisão e *Recall* foram apresentadas individualmente nas Subseções 5.1.1 e 5.1.2, esta análise comparativa foca no indicador F1-Score. Essa escolha justifica-se pois o F1-Score sintetiza o equilíbrio entre Precisão e *Recall* conforme estabelecido na Subseção 4.6.1, oferecendo uma visão do desempenho que se mostrou mais adequada para avaliar o *Prompt Final*. Ao adotar o mesmo indicador consolidado utilizado na validação do padrão de referência no Experimento Principal, torna-se possível estabelecer uma comparação direta que revela se as novas alterações propostas constituem, de fato, um avanço entre os prompts.

Tabela 12 – Resultados comparativos de performance dos novos prompts nas diferentes amostras

Prompt	F1 Total	RMSE (h) atraso	RMSE (h) extravio	Acurácia Total	Acurácia por Sentença	Documentos
main_amostra_C1	0.63	8.86	842.42	92.53	38.33	300
main_2_amostra_C1	0.64	7.77	843.41	92.24	35.33	300
main_amostra_C2	0.52	6.09	166.31	89.64	22.00	300
main_amostra_C3	0.49	7.22	104.97	90.56	24.00	300
atraso1_amostra_C1	0.60	10.61	843.04	89.62	24.00	300
atraso1_1_amostra_C1	0.63	9.45	857.43	91.00	32.67	300
atraso1_amostra_C3	0.44	7.76	28.90	88.00	17.33	300
atraso3_amostra_C1	0.66	12.46	850.79	93.43	37.67	300
atraso3_amostra_C2	0.54	9.95	158.92	89.93	19.00	300
atraso3_amostra_C3	0.48	8.04	36.73	90.95	25.00	300
atraso4_amostra_C2	0.51	5.98	37.85	90.28	18.67	300
atraso4_amostra_C3	0.46	5.93	101.10	91.14	24.33	300
atraso5_amostra_C1	0.63	9.19	845.19	93.72	42.67	300
atraso5_2_amostra_C1	0.64	9.85	844.17	93.62	42.67	300
atraso5_amostra_C2	0.50	6.86	34.96	91.10	24.00	300
atraso5_amostra_C3	0.46	7.03	37.83	91.72	30.00	300
extravio1_amostra_C1	0.64	7.13	849.58	92.80	41.00	300
extravio1_amostra_C3	0.47	6.60	31.57	90.36	24.67	300
extravio2_amostra_C1	0.53	7.63	848.84	91.48	36.33	300
extravio2_amostra_C2	0.49	10.11	29.62	89.19	19.67	300
extravio2_2_amostra_C1	0.57	8.23	848.57	90.88	30.00	300
extravio2_amostra_C3	0.46	6.67	36.89	90.62	27.33	300

Fonte: Elaborado pelo autor.

Uma observação é a queda de desempenho da amostra *C1* para as amostras *C2* e *C3* em quase todos os prompts e métricas. O F1-Score do prompt *main* caiu de 0.63 para 0.52, enquanto o do *atraso3* passou de 0.66 para 0.54. Essa degradação evidencia a necessidade de avaliar modelos em múltiplos conjuntos de dados para verificar sua robustez e evitar o *overfitting*⁶.

⁶ Overfitting (sobreajuste) ocorre quando um modelo de aprendizado de máquina se adapta excessivamente aos dados de treinamento. Isso acontece porque o modelo aprende não apenas os padrões úteis, mas também o “ruído” (informações irrelevantes) desses dados específicos. O resultado é um modelo que “decora” o treino e, por isso, falha em generalizar, ou seja, não consegue fazer previsões precisas quando apresentado a dados novos e desconhecidos (IBM, 2025a).

Esta observação se relaciona com a hipótese levantada na Subseção 4.4.2.2 sobre a possibilidade de viés induzido por testes anteriores realizados com o mesmo *corpus* de sentenças. Os resultados apresentam que o desempenho em um único conjunto de dados pode não ser representativo da capacidade real de generalização do modelo.

Conforme a teoria apresentada na Seção 4.6.1, o F1-Score é a métrica mais adequada para avaliar o equilíbrio entre precisão e *recall* em *corpus* desbalanceados. O prompt *atraso3* alcançou o maior F1-Score de toda a série experimental na amostra *C1*, atingindo 0.66. Este resultado indica que a estratégia de unir os fatores binários de atraso e cancelamento, mantendo uma saída numérica distinta para o intervalo, foi eficaz na identificação da classe minoritária. Entretanto, o desempenho desse prompt caiu nas amostras mais recentes, sugerindo limitações na generalização para casos mais complexos ou *overfitting*.

A análise do RMSE, que penaliza erros numéricos, revela o impacto das otimizações direcionadas. Para o Intervalo de Atraso, o prompt *atraso4* demonstrou o melhor desempenho, atingindo um RMSE de 5.93 na amostra mais recente. Isso indica que a estratégia de unificar a variável binária e tratar o cancelamento como um atraso de valor zero foi eficaz para a extração desta dimensão numérica.

Para o Intervalo de Extravio, a melhoria mostrou-se positiva também. Os prompts testados na amostra *C1* apresentaram um RMSE alto, na ordem de 800, indicando falha sistemática do modelo. Nas amostras mais recentes, os prompts especializados *atraso1* e *extravio2* conseguiram reduzir o erro para 28.90 e 29.62, respectivamente. Esta redução representa um dos ganhos mais significativos do processo de refinamento, indicando que a reestruturação dos prompts corrigiu uma falha do modelo original.

Na Acurácia Total, o prompt *atraso5* obteve o melhor desempenho na amostra *C1*, alcançando 93.72%. Este resultado, combinado com a Acurácia por Sentença de 42.67%, sugere que sua abordagem de fusão completa dos fatores de atraso em uma única variável foi eficaz em produzir predições corretas no nível do documento completo. A capacidade de acertar todos os fatores de uma sentença em 42.67% dos casos representa um patamar de maior confiabilidade aos demais prompts, sendo um indicador para aplicações práticas que dependem da integridade completa dos dados extraídos.

5.2.2 Síntese e Validação das Sub-Hipóteses

A análise comparativa dos resultados, consolidados na Tabela 12, demonstrou especializações distintas entre os prompts avaliados. A definição de um novo padrão de referência exige uma avaliação de todas as métricas, priorizando aquelas que refletem de forma mais adequada a confiabilidade e a utilidade do sistema.

Nesse contexto, o prompt *atraso5* emergiu como a alternativa mais adequada para estabelecimento do novo padrão metodológico deste trabalho. Esta escolha baseia-se nos resultados obtidos, que transcendem o desempenho isolado em uma métrica específica. Primeiramente, o *atraso5* alcançou os melhores índices tanto em Acurácia Total quanto em Acurácia por Sentença.

Esta última métrica constitui um indicador da capacidade do modelo de interpretar documentos em sua integralidade, gerando registros para o treinamento de modelos preditivos na Frente 3 do projeto CONCIL-IA.

Adicionalmente, o *atraso5* contribuiu para uma maior estabilização da extração de valores numéricos. Embora o *atraso4* tenha apresentado menor erro para o intervalo de atraso, o *atraso5* demonstrou maior consistência geral ao solucionar a falha na extração do intervalo de extravio com um RMSE menor comparado ao prompt *main*, evitando simultaneamente a instabilidade do *atraso4*, que registrou um RMSE de 101.10 em uma das amostras.

A análise revela ainda que o *atraso5* estabelece um equilíbrio entre as diferentes dimensões avaliadas. Embora o *atraso4* possua vantagem marginal no RMSE do intervalo de atraso e o *atraso3* apresente desempenho ligeiramente superior no F1-Score da amostra inicial, essas vantagens pontuais não superam os benefícios do *atraso5*. Considerando que o F1-Score deste último oscila entre 0.63 e 0.64 na amostra original, valor próximo ao 0.66 do *atraso3*, a opção pelo prompt com maior robustez geral torna-se mais justificável.

A conclusão desta fase experimental permite revisitar as hipóteses de pesquisa formuladas na abertura da Seção 4.4.2.1, oferecendo respostas empíricas. Primeiramente, o experimento valida a hipótese de que a redução da carga cognitiva, implementada mediante a fusão de fatores semanticamente relacionados, pode constituir uma estratégia eficaz para o contexto analisado. O desempenho superior do *atraso5*, que consolidou três fatores em uma única dimensão demonstra que a simplificação estrutural pode conduzir a uma compreensão mais integrada do documento.

Em segundo lugar, a investigação confirmou que o refinamento direcionado das instruções constitui um instrumento para aprimoramento do desempenho em fatores específicos. A melhoria no RMSE do Intervalo de Extravio, que transitou de erro sistemático para resultados mais estáveis com os prompts *atraso1* e *extravio2*, representa um indício dessa afirmação. A especialização no *atraso4* resultou no menor erro para o Intervalo de Atraso, sustentando que a engenharia iterativa de prompts é útil para correção de falhas críticas.

Contudo, é necessário contextualizar esses achados dentro do panorama metodológico mais amplo. Embora a estratégia de condensação implementada no *atraso5* tenha produzido os resultados superiores nesta investigação, ela representa talvez um limite máximo dentro do espaço de soluções possíveis. Conforme será discutido na seção subsequente, os avanços na interação com LLMs apontam para uma direção distinta, empregando estruturas de entrada e saída que permitem manipular os dados com uma maior robustez.

5.3 LIMITAÇÕES DO ESTUDO

A adequada contextualização dos achados empíricos desta investigação demanda o reconhecimento de suas restrições metodológicas. No âmbito da engenharia de prompts, a limitação originou-se das barreiras de acesso à interface de programação de aplicações disponibilizada pela OpenAI (Pereira et al., 2025). As contingências financeiras vinculadas aos custos operacionais, combinadas com os tetos diários impostos ao processamento de *tokens*,

estabeleceram os contornos dos experimentos conduzidos, dificultando tanto a execução de testes em larga escala quanto a realização de ciclos iterativos múltiplos voltados ao aprimoramento do prompt (Pereira et al., 2025).

Nesse contexto, limitou-se a exploração de técnicas alternativas para elaboração de prompts, incluindo a implementação de cadeias de raciocínio explícitas mais elaboradas. Ademais, a investigação centrou-se exclusivamente na otimização por meio da engenharia de prompts, sem incorporar procedimentos de ajuste fino do modelo (Pereira et al., 2025). Conquanto o propósito metodológico fosse justamente avaliar a capacidade do modelo em sua configuração nativa, a ausência de comparações com versões especializadas dificulta a mensuração dos benefícios relativos proporcionados pela engenharia de prompts quando contrastada com o treinamento especializado.

A delimitação metodológica priorizou um único LLM, especificamente o GPT-4o, aplicado a um segmento específico da jurisdição consumerista. Embora tal escolha possibilite análise de padrões comportamentais, ela circunscreve a generalização dos resultados para outros contextos (Pereira et al., 2025). As estratégias de otimização desenvolvidas e os padrões de desempenho documentados podem não ser transponíveis para modelos concorrentes de fronteira das famílias Claude ou Gemini, nem para arquiteturas de código aberto tais como Llama e Mistral (Pereira et al., 2025).

Paralelamente, a ausência de avaliações comparativas envolvendo modelos alternativos ou metodologias distintas, incluindo o ajuste fino de arquiteturas locais, dificulta análises de desempenho e custo-benefício que contrastariam a utilização de interfaces comerciais com abordagens de treinamento especializado (Pereira et al., 2025). Adicionalmente, o raciocínio jurídico incorpora fatores de natureza casuística, exemplificados pelo impacto de eventos disruptivos nas circunstâncias pessoais do litigante, tais como compromissos familiares ou profissionais de caráter inadiável.

A tentativa de formalização de fatores dessa natureza em estruturas de prompt apresentaria risco de simplificação ou de indução de alucinações no modelo, desafio documentado na aplicação de modelos de linguagem ao domínio jurídico, conforme observado por Deroy, Gosh e Gosh (2023) em estudos sobre sumarização de decisões e por Dal Pont et al. (2023) na geração automatizada de ementas, ambos discutidos na Seção 3.1. A identificação e ponderação adequada de tais fatores permanecem, em relevante dimensão, dependentes do discernimento humano.

Por fim, este estudo demanda contextualização temporal apropriada. Os sistemas jurídicos caracterizam-se por dinamismo, de modo que alterações legislativas, doutrinárias ou jurisprudenciais podem modificar a relevância e interpretação dos fatores examinados (Pereira et al., 2025). A sustentabilidade de longo prazo de sistemas fundamentados nesta metodologia requer processos contínuos de revisão e atualização dos prompts, garantindo alinhamento com a evolução do ordenamento jurídico, desafio inerente a sistemas de inteligência artificial que buscam modelar conhecimento legal dinâmico.

5.4 NOVOS PARADIGMAS NA ENTRADA E SAÍDA DOS PROMPTS: JSON SCHEMA, MARKDOWN E A JANELA DE CONTEXTO

A metodologia de extração de fatores empregada neste trabalho, fundamentada na geração de um formato CSV, foi relevante para a prototipagem e validação iterativa dos prompts. Contudo, a avaliação de sua escalabilidade revelou limitações que motivaram a investigação de paradigmas para a geração de saídas estruturadas. A natureza probabilística dos LLMs implica que não há garantias sobre a conformidade estrutural da saída, tornando certos formatos como o CSV vulneráveis a inconsistências que podem comprometer o processamento dos dados (Geng et al., 2025). Este capítulo, portanto, analisa essas limitações e explora os avanços em geração estruturada como um caminho para a evolução metodológica do trabalho.

5.4.1 Sub-experimento: Limitações do Formato CSV em Testes de Escalabilidade

Para avaliar a robustez e escalabilidade da metodologia de extração baseada em CSV, conduziram-se novos testes visando ampliar a complexidade da tarefa. O escopo da extração foi expandido mediante tentativa de inclusão de dois novos fatores que demandavam raciocínio mais elaborado do modelo para identificação e formatação dos dados. Esses fatores, apresentadas a seguir, encontram-se também disponíveis no Apêndice A.9:

- 16. Valor do dano moral individual: Extraia o valor numérico da condenação por danos morais atribuído ao autor, ou aos autores, caso em litisconsórcio.⁷ Busque na Seção "Dispositivo" ou na conclusão do julgado pela quantia associada a "danos morais", ignorando expressamente valores de "danos materiais". Em casos com múltiplos autores (litisconsórcio), identifique o valor especificado para cada um individualmente (ex: "R\$ 5.000,00" quando para cada autor, ou "R\$ 5.000,00; R\$ 6.000,00" para autores com danos morais diferentes). O resultado deve ser um número formatado com ponto como separador de milhar e vírgula para os centavos (ex: 5.000,00). Se não houver condenação por dano moral, indique "0,00". Dê substancial relevância para, no caso de litisconsórcio, extrair o dano moral deferido a cada um dos autores, caso diferente.
- 17. Dano Moral Divergente em Litisconsórcio Ativo: verifique se há múltiplos autores. Se sim, analise a Seção "Dispositivo" e indique "S" apenas se os valores da condenação por dano moral forem explicitamente diferentes para cada autor. Indique "N" se houver um único autor ou se os valores forem idênticos para todos.

A expectativa era que o modelo GPT-4o se adaptasse e mantivesse a estrutura de saída CSV, simplesmente adicionando duas novas colunas aos dados extraídos. Os resultados, no entanto, demonstraram a limitação do formato. Com o aumento dos fatores, o modelo começou

⁷ O litisconsórcio consiste na existência de mais de uma parte em pelo menos um dos polos (ativo ou passivo) de uma relação jurídica processual (Bueno, 2017).

a apresentar "alucinações de formato", gerando saídas que violavam a estrutura de duas linhas com frequência.

Em várias instâncias, a resposta continha quebras de linha, texto explicativo fora do formato ou uma contagem incorreta de colunas, tornando o processamento automatizado inviável. Essa falha empírica indica a necessidade de um paradigma de saída formalizado, uma conclusão que já ecoa entre os pesquisadores da área (Geng et al., 2025).

5.4.2 Estratégias de Otimização e Desafios da Engenharia de Prompt

5.4.2.1 O Estado da Arte em Geração de Saídas Estruturadas

A instabilidade do formato CSV em testes práticos motivou investigação sobre os avanços em geração de saídas estruturadas. A literatura recente consolidou a técnica denominada *constrained decoding*, ou decodificação restrita, como abordagem para buscar a conformidade sintática da saída de LLMs (Geng et al., 2025).

Conforme definição estabelecida por Geng et al. (2025), esta metodologia intervém diretamente no processo de geração textual do modelo, mascarando ou proibindo a seleção de tokens que violariam um conjunto de regras gramaticais predefinidas. Ao forçar o modelo a amostrar exclusivamente de subconjunto de tokens sintaticamente válidos a cada etapa, o processo garante, por construção, que a saída final se conforme perfeitamente à estrutura desejada (Geng et al., 2025).

Para que a decodificação restrita seja eficaz, torna-se necessária uma linguagem formal capaz de expressar as restrições impostas, e o JSON Schema consolidou-se como padrão para essa finalidade (Geng et al., 2025). Essencialmente, o JSON Schema constitui uma metalinguagem que descreve a estrutura de dados em formato JSON, permitindo a especificação de restrições, incluindo os tipos de dados de cada campo e padrões que devem ser seguidos (Geng et al., 2025).

A adoção do JSON Schema por provedores de modelos comerciais e frameworks de código aberto revela sua posição como um dos padrões metodológicos, assegurando compatibilidade entre diversas aplicações de processamento de linguagem natural (Geng et al., 2025). Tal padronização representa avanço para a interoperabilidade de sistemas baseados em LLMs.

5.4.2.2 O Impacto do Formato de Entrada: O Uso de Markdown

Além da estrutura da saída, a formatação da entrada também foi identificada como aspecto relevante, embora frequentemente negligenciado na literatura da engenharia de prompts (Li et al., 2024). Formatos como o Markdown, que utilizam elementos de marcação para criar hierarquia visual e lógica, aparentam desempenhar papel importante em guiar a interpretação do modelo sobre as instruções fornecidas (Li et al., 2024; He et al., 2024).

O uso de cabeçalhos (##), listas e outros elementos de marcação cria segmentação dos diferentes componentes da instrução. A hipótese central é que essa delimitação, que se

alinha a padrões de prompt documentados na literatura como o padrão de *Template* exposto na Subseção 2.6.3 deste trabalho, não apenas organiza a informação para o leitor humano, mas também aparenta influenciar o comportamento do modelo (White et al., 2023; Li et al., 2024; He et al., 2024).

Aparentemente, a ênfase criada por esses marcadores visuais estabelece prioridades hierárquicas para o modelo, o que poderia facilitar o funcionamento de seu mecanismo de atenção. Ao ressaltar seções como *Persona*, *Instruções* e *Exemplos* (Li et al., 2024), o modelo pode ser mais capaz de alocar seus recursos computacionais para focar nas relações entre a pergunta do usuário e as partes relevantes do contexto fornecido.

Essa sensibilidade dos LLMs à formatação do prompt, mesmo a mudanças sutis como a inserção de quebras de linha ou a escolha de delimitadores específicos, é um fenômeno documentado na literatura, reforçando que a estrutura da entrada não constitui detalhe irrelevante, mas sim fator que impacta diretamente a performance do modelo (Voronov; Wolf; Ryabinin, 2024).

Essa hipótese é corroborada por evidências. A pesquisa conduzida por Li et al. (2024) demonstrou que, em tarefas de extração de conhecimento e raciocínio numérico, o formato Markdown superou o desempenho do formato de texto simples. A conclusão do estudo é que formatos de prompt estruturados melhoram a capacidade do modelo de extrair e processar informações relevantes, levando a uma maior acurácia nas respostas geradas (Li et al., 2024).

Para o projeto CONCIL-IA, isso significa que a reestruturação dos prompts para utilizar Markdown pode, por si só, levar a uma melhora na precisão da extração, independentemente da tecnologia de saída utilizada (Li et al., 2024). Tal constatação sugere que investimentos em formatação de entrada podem produzir ganhos relevantes de desempenho sem modificações arquiteturais do modelo.

5.4.2.3 A Expansão da Janela de Contexto e Suas Implicações para a Engenharia de Prompt

A janela de contexto de um LLM constitui essencialmente sua memória operacional, determinando a quantidade máxima de informação, mensurada em *tokens*, que o modelo pode processar e reter simultaneamente durante uma única operação (Bergmann, 2024). Ela estabelece os limites de extensão de uma conversação ou de um documento para que o modelo preserve a coerência textual e utilize informações previamente processadas (Bergmann, 2024).

Quando a entrada excede esse limite estabelecido, o modelo é compelido a truncar ou resumir o texto fornecido, ocasionando perda de parte do contexto (Bergmann, 2024). Historicamente, uma das limitações da engenharia de prompts, especialmente em domínios como o Direito que operam com documentos extensos, consistia no tamanho dessas janelas de contexto nos modelos disponíveis.

Em modelos caracterizados por janelas de oito mil tokens, uma única sentença judicial poderia facilmente consumir a maior parcela desse espaço disponível, deixando margem limitada para as instruções do prompt. Essa restrição operacional forçava os desenvolvedores a criar

prompts excessivamente concisos, sacrificando a clareza das instruções ou a quantidade de exemplos fornecidos em técnicas como o *few-shot prompting*.

A evolução dos modelos disponíveis, exemplificada pelo Gemini 2.5 Pro com sua janela de até um milhão de tokens, alterou esse paradigma operacional (Kavukcuoglu, 2025). O texto da sentença, que anteriormente constituía a restrição metodológica, agora ocupa fração do espaço disponível para processamento.

Fundamentalmente, essa ampliação libera vasto espaço para a formulação de prompts mais ricos e detalhados, com implicações diretas para o projeto CONCIL-IA (Kavukcuoglu, 2025). A nova configuração permite o uso de técnicas de engenharia de prompts que anteriormente eram inviáveis devido à sua extensão requerida, modificando as possibilidades metodológicas de aplicação de LLMs no domínio jurídico.

5.4.3 Implicações Metodológicas e Direcionamentos para Trabalhos Futuros

A análise da literatura especializada e os resultados dos testes de escalabilidade conduzidos indicam, de forma convergente, a necessidade de uma evolução metodológica no projeto CONCIL-IA. Essa transição redefine a engenharia de prompts, que evolui de prática de formulação textual para disciplina mais próxima da engenharia de software, onde a reformulação da comunicação com o modelo constitui requisito para próximos avanços (Li et al., 2024).

A tarefa de extração passa a ser definida por gramática inequívoca, o que introduz nível de rigor e previsibilidade para o domínio jurídico. Essa abordagem dual manifesta-se em duas frentes complementares de desenvolvimento metodológico. Primeiramente, na formalização da entrada, a literatura aponta para os benefícios do uso de formatos estruturados como o Markdown para delinear componentes do prompt (Li et al., 2024).

A utilização de elementos de marcação para delinear os componentes do prompt cria hierarquia lógica que melhora a capacidade do modelo de processar a informação e aumenta a acurácia das respostas geradas (Li et al., 2024). Em paralelo, a formalização da saída, implementada através da decodificação restrita com JSON Schema, parece garantir que as saídas sejam válidas e prontas para utilização nas aplicações subsequentes (Geng et al., 2025).

Os benefícios dessa abordagem integrada indicam, além de aumento na robustez e confiabilidade do sistema de extração, também ganhos de eficiência operacional, uma vez que a decodificação restrita pode acelerar o tempo de inferência em até cinquenta por cento (Geng et al., 2025). Tal ganho de eficiência representa economia considerável em ambientes de processamento em larga escala.

Os resultados dos testes de escalabilidade, que revelaram a instabilidade do formato CSV empregado, fornecem direcionamento para trabalhos futuros. Embora o resultado da tentativa de inclusão de dois novos fatores tenha sido negativo em termos de conformidade com o formato especificado, ele foi positivo ao expor os limites da abordagem atual e justificar a transição para a nova metodologia.

Consequentemente, a próxima etapa da pesquisa consiste na implementação e validação

dos novos paradigmas. Este processo envolverá uma sequência de etapas metodológicas: (1) a reengenharia dos prompts para adotar uma estrutura em Markdown; (2) a definição de um JSON Schema para todos os 17 fatores de extração, especificando tipos de dados e restrições; e (3) a re-execução dos experimentos para quantificar os ganhos em robustez, precisão e eficiência em comparação com a abordagem baseada em CSV.

Este novo direcionamento metodológico supera as limitações técnicas previamente identificadas e define os próximos passos para o projeto CONCIL-IA e trabalhos futuros. A transição do formato CSV para geração estruturada com Markdown e JSON Schema podem viabilizar uma expansão controlada do conjunto de fatores extraídos, requisito para análises mais abrangentes. Tal abordagem alinha o projeto com os avanços na literatura do uso de LLMs, onde conformidade estrutural e rastreabilidade constituem requisitos importantes, pavimentando o caminho para operacionalização em larga escala.

6 CONCLUSÃO

O presente trabalho propôs detalhar e expandir a investigação, apresentada preliminarmente em estudo com participação deste autor (Pereira et al., 2025), acerca da aplicação de LLMs para extração automatizada de fatores jurídicos. Esta investigação se insere em um contexto mais amplo de pesquisas desenvolvidas durante a graduação, abrangendo outros trabalhos precedentes (Sabo et al., 2025; Junior et al., 2025) e um Programa de Iniciação Científica (PIC) junto ao CNPq na mesma linha de investigação (UFSC, 2025).

Conforme delimitado no Capítulo 1, a pesquisa foi estruturada em duas fases experimentais sequenciais: primeiramente, uma análise do Experimento Principal para expor seus resultados e identificar fragilidades; em seguida, a exposição do Experimento Secundário, que testou novas hipóteses de otimização. Ao percorrer a fundamentação teórica, o contexto de aplicação e a análise, este capítulo final sintetiza os achados, responde aos problemas de pesquisa e às hipóteses formuladas, e delinea as contribuições e perspectivas da investigação.

Em resposta ao Problema 1, que questiona se o processo iterativo de engenharia de *prompt* influencia o desempenho do LLM e quais falhas persistem, os resultados confirmaram a hipótese principal. A análise na Subseção 5.1 demonstrou que o refinamento progressivo resultou em incremento nas métricas de desempenho, evidenciado pela elevação da acurácia geral para 91,5% e pela redução no RMSE do intervalo de atraso. Contudo, em linha com o previsto na hipótese, a investigação também caracterizou as falhas de extração: o F1-Score geral permaneceu moderado e foi identificada uma falha na extração do intervalo de extravio, o que forneceu a justificativa para a segunda fase da pesquisa.

No que tange ao Problema 2, que examina se estratégias de otimização podem corrigir falhas específicas do Experimento Principal, o estudo validou a hipótese secundária ao demonstrar que é possível alcançar ambos os objetivos. Conforme exposto na Seção 5.2, as estratégias focadas na redução da carga cognitiva imposta ao modelo, como a fusão de fatores e a reestruturação da lógica de saída, apresentaram melhoria de desempenho. O *prompt atraso5* demonstra o desempenho dessa abordagem: além de mitigar a falha do RMSE no extravio, registrou desempenho superior na métrica de integridade documental, a Acurácia por Sentença. Conforme sintetizado na Subseção 5.2.2, este resultado confirma que a otimização direcionada não apenas trata das inconsistências identificadas, mas também aumenta a confiabilidade da extração, estabelecendo a referência metodológica para este estudo.

Ato contínuo, a condução dos experimentos revelou os limites da abordagem metodológica atual, especialmente a limitação do formato CSV em testes de escalabilidade, conforme exposto na Seção 5.4. Essa constatação, alinhada às limitações gerais do estudo discutidas na Seção 5.3, direcionou a investigação para novos paradigmas de geração estruturada, como o uso de Markdown para entradas e JSON Schema para saídas, que definem um caminho futuro para a evolução do trabalho.

Do ponto de vista teórico, este trabalho contribui ao articular os fundamentos da IAGen e da Engenharia de *Prompt*, explorados no Capítulo 2, a uma aplicação no domínio jurídico.

A pesquisa oferece um relato de um *pipeline* metodológico, expondo desde a construção do *ground truth* na Subseção 4.3.2 até a análise dos resultados dos dois experimentos nas Seções 5.1 e 5.2. A discussão sobre o impacto da polissemia de termos jurídicos na Seção 2.4 e do desbalanceamento de classes na performance dos LLMs na Seção 4.6.1 representa os desafios teóricos. Na esfera prática, a contribuição reside na apresentação de um método escalável para enfrentar as limitações da estruturação de dados no Poder Judiciário, fornecendo uma solução para a Frente 3 do projeto CONCIL-IA e reforçando o potencial da IA como ferramenta de apoio para aprimorar os mecanismos de ODR, em alinhamento aos objetivos institucionais do CNJ discutidos na Subseção 3.2.1.

No âmbito da engenharia de *prompt*, o estudo apresentou a transição do *Prompt Inicial* ao *Prompt Final* na Subseção 4.4.1 e documentou o desenvolvimento de *prompts* especializados no Experimento Secundário, conforme a Subseção 4.4.2. A aplicação de *few-shot prompting* para formatação de dados numéricos e a elaboração de instruções que emulam raciocínio de cadeia de pensamento exemplificam como a teoria se traduz em melhorias de desempenho.

As conclusões e limitações identificadas indicam caminhos para futuras investigações. Uma perspectiva consiste na implementação da mudança metodológica proposta na Seção 5.4, reestruturando os *prompts* e reexecutando os experimentos para validar os benefícios dessa transição. Outra vertente consiste na exploração de modelos alternativos, cuja replicação do experimento permitiria avaliar a generalização da metodologia e possibilitaria um estudo comparativo de desempenho e custo-benefício. Por fim, a metodologia pode ser expandida para outros domínios do Direito e para a extração de fatores de natureza mais casuística.

Em síntese, este trabalho demonstrou que a engenharia de *prompt*, quando conduzida de forma iterativa, constitui uma metodologia viável e escalável para o desafio da aplicação de IA no Direito brasileiro. A jornada evolutiva do *Prompt Inicial* ao *Prompt Final*, e deste para o novo padrão metodológico do trabalho no Experimento Secundário, evidencia que o processo de refinamento e otimização constitui um fator relevante para a qualidade e fidedignidade dos dados gerados. Em um cenário onde iniciativas nacionais, como o projeto Logos do STJ, demandam insumos de dados consistentes, a metodologia aqui apresentada contribui para o desenvolvimento de modelos preditivos eficazes, conforme almejado pelo projeto CONCIL-IA. Portanto, esta pesquisa oferece um roteiro metodológico e análises úteis para o desenvolvimento de futuras ferramentas que visem contribuir com a eficiência e a consistência do judiciário brasileiro.

REFERÊNCIAS

- AGGARWAL, C. C. Machine learning for text: An introduction. In: **Machine learning for text**. [S.l.]: Springer, 2018. p. 1–16.
- AHMED, T. et al. Improving few-shot prompts with relevant static analysis products. **arXiv preprint arXiv:2304.06815**, Apr, 2023.
- ALPAYDIN, E. **Introduction to machine learning**. [S.l.]: MIT press, 2020.
- Amazon Web Services. **O que é uma API (interface de programação de aplicações)?** 2025. Disponível em: <https://aws.amazon.com/pt/what-is/api/>.
- AMMAR, A. et al. Prediction of Arabic legal rulings using Large Language Models. 2023.
- BERGMANN, D. **O que é uma janela de contexto?** 2024. IBM Think. Disponível em: <https://www.ibm.com/br-pt/think/topics/context-window>.
- BISHOP, C. M.; BISHOP, H. **Deep learning: Foundations and concepts**. [S.l.]: Springer Nature, 2023.
- Brasil. **Lei nº 8.078, de 11 de setembro de 1990: Dispõe sobre a proteção do consumidor e dá outras providências**. 1990. Diário Oficial da República Federativa do Brasil. Acessado em: 4 de novembro de 2025. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/18078compilado.htm.
- BRINKER, T. J. et al. Deep neural networks are superior to dermatologists in melanoma image classification. **European Journal of Cancer**, Elsevier, v. 119, p. 11–17, 2019.
- BUENO, C. S. **Manual de direito processual civil**. [S.l.]: Saraiva Educação SA, 2017.
- CHAI, T.; DRAXLER, R. R. Root mean square error (rmse) or mean absolute error (mae)?—arguments against avoiding rmse in the literature. **Geoscientific model development**, Copernicus Publications Göttingen, Germany, v. 7, n. 3, p. 1247–1250, 2014.
- Conselho da União Europeia. **Regulamento (UE) 2024/1689 do Parlamento Europeu e do Conselho de 13 de junho de 2024 que estabelece regras harmonizadas em matéria de inteligência artificial (Regulamento Inteligência Artificial)**. 2024. <https://eur-lex.europa.eu/legal-content/PT/TXT/PDF/?uri=OJ:L202401689>. Acessado em : 16 de outubro de 2025.
- Conselho Nacional de Justiça. **Resolução nº 125 de 29 de novembro de 2010. Dispõe sobre a Política Judiciária Nacional de tratamento adequado dos conflitos de interesses no âmbito do Poder Judiciário e dá outras providências**. Brasília, DF: [s.n.], 2010. Diário da Justiça eletrônico. Publicado em 29 de novembro de 2010.
- Conselho Nacional de Justiça. **Resolução nº 332, de 21 de agosto de 2020**. Brasília, DF: [s.n.], 2020. Dispõe sobre a ética, a transparência e a governança na produção e no uso de inteligência artificial no Poder Judiciário e dá outras providências. Publicado no Diário de Justiça Eletrônico do CNJ.
- Conselho Nacional de Justiça. **Justiça em números 2024**. Brasília, DF, 2025.

Conselho Nacional de Justiça. **Resolução nº 615, de 11 de março de 2025**. Brasília, DF: [s.n.], 2025. Estabelece diretrizes para o desenvolvimento, utilização e governança de soluções desenvolvidas com recursos de inteligência artificial no Poder Judiciário. Publicado no Diário de Justiça Eletrônico do CNJ.

CUNHA, F. A. N. da; NASCIMENTO, M. d. J. L. do. Aspectos jurídicos da inteligência artificial no brasil: Comparativo entre o projeto de lei 2.338/23 e o regulamento da união europeia. **Revista Ibero-Americana de Humanidades, Ciências e Educação**, v. 11, n. 6, p. 3183–3192, 2025.

DEROY, A.; GHOSH, K.; GHOSH, S. How ready are pre-trained abstractive models and LLMs for legal case judgement summarization? 2023.

DIAO, S. et al. Active prompting with chain-of-thought for large language models. **arXiv preprint arXiv:2302.12246**, 2023.

ESTEVEZ, E. C.; LINARES, S.; FILLOTTRANI, P. Prometea: Transformando la administración de justicia con herramientas de inteligencia artificial. Banco Interamericano de Desarrollo, 2020.

FEUERRIEGEL, S. et al. Generative ai. **Business & Information Systems Engineering**, Springer, v. 66, n. 1, p. 111–126, 2024.

GENG, S. et al. Jonschemabench: A rigorous benchmark of structured outputs for language models. **arXiv preprint arXiv:2501.10868**, 2025.

GIRAY, L. Prompt engineering with chatgpt: a guide for academic writers. **Annals of biomedical engineering**, Springer, v. 51, n. 12, p. 2629–2633, 2023.

Google Developers. **LLMs: o que é um modelo de linguagem grande?** 2025. Machine Learning Crash Course. Disponível em: <https://developers.google.com/machine-learning/crash-course/llm/transformers?hl=pt-br>.

GOULD, A. et al. Preference-based abstract argumentation for case-based reasoning (with appendix). **arXiv preprint arXiv:2408.00108**, 2024.

GRAY, M. et al. Toward automatically identifying legally relevant factors. In: **Legal knowledge and information systems**. [S.l.]: IOS Press, 2022. p. 53–62.

GRUDA, D. Three ways chatgpt helps me in my academic writing. **Nature**, Nature Publishing Group, v. 10, 2024.

Grupo de Pesquisa Governo Eletrônico, Inclusão Digital e Sociedade do Conhecimento. **Sobre o Projeto - Concil-IA**. 2025. <https://concilia.ufsc.br/mesmerize/sobre-o-projeto/>. Acesso em: 23 de outubro de 2025.

GUI, J. et al. A survey on self-supervised learning: Algorithms, applications, and future trends. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, v. 46, n. 12, p. 9052–9071, 2024.

HAMILTON, S. Blind judgement: Agent-based Supreme Court modelling with GPT. 2023.

HE, J. et al. Does prompt formatting have any impact on llm performance? **arXiv preprint arXiv:2411.10541**, 2024.

HOWARD, J.; RUDER, S. Universal language model fine-tuning for text classification. **arXiv preprint arXiv:1801.06146**, 2018.

IBM. **O que é análise de dados preditiva?** 2025. Disponível em: <https://www.ibm.com/br-pt/think/topics/predictive-analytics>.

IBM. **O que é overfitting?** 2025. Disponível em: <https://www.ibm.com/br-pt/think/topics/overfitting>. Acessado em: 10 de novembro de 2025. Disponível em: <https://www.ibm.com/br-pt/think/topics/overfitting>.

IBM. **What is shadow IT?** 2025. <https://www.ibm.com/think/topics/shadow-it>. Acessado em: 26 de novembro de 2025.

JUMPER, J. et al. Highly accurate protein structure prediction with alphafold. **nature**, Nature Publishing Group UK London, v. 596, n. 7873, p. 583–589, 2021.

JUNIOR, M. M. et al. CONCIL-IA PROJECT: final findings and digital innovations for conflict resolution. **Latin American Journal of European Studies**, 2025. No prelo. Disponível em: <https://encurtador.com.br/QKkH>.

KAVUKCUOGLU, K. **Gemini 2.5: Our most intelligent AI model**. 2025. The Keyword (Google Blog). Acessado em 05 de novembro de 2025. Disponível em: <https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/#building-on-best-gemini>.

KOJIMA, T. et al. Large language models are zero-shot reasoners. **Advances in neural information processing systems**, v. 35, p. 22199–22213, 2022.

LEE, U. et al. Few-shot is enough: exploring chatgpt prompt engineering method for automatic question generation in english education. **Education and Information Technologies**, Springer, v. 29, n. 9, p. 11483–11515, 2024.

LI, X. et al. Prompt engineering and format on llms in the financial domain. 12 2024.

LICARI, D.; COMANDE, G. Italian-legal-bert models for improving natural language processing tasks in the italian legal domain. **Computer Law & Security Review**, Elsevier, v. 52, p. 105908, 2024.

LIN, Z. Prompt engineering for applied linguistics: Elements, examples, techniques, and strategies. **English Language Teaching**, Canadian Center of Science and Education, v. 17, n. 9, p. 1–14, 2024.

LOEVINGER, L. Jurimetrics—the next step forward. **Minn. L. Rev.**, HeinOnline, v. 33, p. 455, 1948.

LUK, M. Generative ai: Overview, economic impact, and applications in asset management. **Economic Impact, and Applications in Asset Management (September 18, 2023)**, 2023.

MALIK, V. et al. Ildc for cjpe: Indian legal documents corpus for court judgment prediction and explanation. **arXiv preprint arXiv:2105.13562**, 2021.

MARTIN, L. et al. Better call gpt, comparing large language models against lawyers. **arXiv preprint arXiv:2401.16212**, 2024.

MARTINO, G. D.; PIO, G.; CECI, M. Prilj: an efficient two-step method based on embedding and clustering for the identification of regularities in legal case judgments. **Artificial Intelligence and Law**, Springer, v. 30, n. 3, p. 359–390, 2022.

MARVIN, G. et al. Prompt engineering in large language models. In: SPRINGER. **International conference on data intelligence and cognitive informatics**. [S.l.], 2023. p. 387–402.

MASALA, M. et al. jurbert: A romanian bert model for legal judgement prediction. In: **Proceedings of the Natural Legal Language Processing Workshop 2021**. [S.l.: s.n.], 2021. p. 86–94.

NAKAD-WESTSTRATE, H. H. et al. The rise of the robotic judge in modern court. In: **Proceedings of the 7th International Conference on Information Technology (ICIT 2015)**. [S.l.: s.n.], 2015. p. 59–67.

NIKLAUS, J. et al. Lextreme: A multi-lingual and multi-task benchmark for the legal domain. **arXiv preprint arXiv:2301.13126**, 2023.

NORVIG, P.; RUSSELL, S. Inteligência artificial. **Rio de Janeiro: Grupo GEN**, 2013.

OLIVEIRA, R. F. d. **Inteligência artificial**. [S. l.]: Editora e Distribuidora Educacional S.A., 2018. ISBN 9788552211419.

OLIVEIRA, R. S.; NASCIMENTO, E. G. S. Analysing similarities between legal court documents using natural language processing approaches based on Transformers. 2023.

OUYANG, L. et al. Training language models to follow instructions with human feedback. **Advances in neural information processing systems**, v. 35, p. 27730–27744, 2022.

PEIXOTO, F. H.; SILVA, R. Z. M. d. Inteligência artificial e direito. **Curitiba: Alteridade**, v. 1, 2019.

PEREIRA, L. d. C. R. et al. Using gpt-4o as a factor extractor for brazilian consumer law judgments. **Artificial Intelligence and Law**, Springer, p. 1–29, 2025.

PONT, T. R. D. et al. Legal summarisation through LLMs: The PRODIGIT project. 2023.

RADFORD, A. et al. Improving language understanding by generative pre-training. San Francisco, CA, USA, 2018.

REMPELL, S. Factors. **Buff. L. Rev.**, HeinOnline, v. 70, p. 1755, 2022.

RISSLAND, E. L.; ASHLEY, K. D. A note on dimensions and factors. **Artificial Intelligence and law**, Springer, v. 10, n. 1, p. 65–77, 2002.

RISSLAND, E. L.; ASHLEY, K. D.; BRANTING, L. K. Case-based reasoning and law. **The Knowledge Engineering Review**, Cambridge University Press, v. 20, n. 3, p. 293–298, 2005.

ROSA, J. L. G. Fundamentos da inteligência artificial. 2011.

RUSSELL, S. J. **Q A: The future of artificial intelligence**. 2016.

<https://people.eecs.berkeley.edu/~russell/research/future/q-and-a.html>. Acessado em: 16 de outubro de 2025.

SABO, I. C. et al. Avanços iniciais do projeto concil-ia: desafios e potenciais impactos. **Revista Democracia Digital e Governo Eletrônico**, v. 1, n. 24, p. 2–20, 2025. Disponível em: <https://buscalegis.ufsc.br/revistas/index.php/observatoriodoegov/article/view/401/246>.

SABO, I. C. et al. Clustering of brazilian legal judgments about failures in air transport service: an evaluation of different approaches. **Artificial Intelligence and Law**, Springer, v. 30, n. 1, p. 21–57, 2022.

SARTOR, G. Reasoning with factors. **Argumentation**, Springer, v. 19, n. 4, p. 417–432, 2005.

SAVELKA, J. et al. Explaining Legal Concepts with Augmented Large Language Models (GPT-4). 2023.

Senado Federal. **Projeto de Lei nº 2338, de 2023. Parecer nº 208, de 2024 – PLEN/SF**. 2023. <https://legis.senado.leg.br/sdleg-getter/documento?dm=9884955ts=1742240908365disposition=inline>. Acessado em: 16 de outubro de 2025.

SHEN, Z. et al. Multi-lexsum: Real-world summaries of civil rights lawsuits at multiple granularities. **Advances in Neural Information Processing Systems**, v. 35, p. 13158–13173, 2022.

SHIN, J. et al. Prompt engineering or fine tuning: An empirical assessment of large language models in automated software engineering tasks. **arXiv preprint arXiv:2310.10508**, 2023.

State Courts of Singapore. **New Generative AI-Powered Case Summarisation Tool to Help Small Claims Tribunals Users**. 2025. <https://sgpc.gov.sg/api/file/getfile/MediaAcessado em: 16 de outubro de 2025>.

Superior Tribunal de Justiça. **STJ lança novo motor de inteligência artificial generativa para aumentar eficiência na produção de decisões**. 2025. Website Institucional. Disponível em: <https://encurtador.com.br/SyhK>. Acesso em: 11 out. 2025.

TORRES, R. G. **Percepções sobre uso da inteligência artificial generativa de textos na escrita acadêmica: uma análise sobre diretrizes e autoria no contexto da pós-graduação em ensino de ciências e matemática da UNICAMP**. Tese (Doutorado) — [sn], 2025.

UFSC. **Certificado de participação no Programa Voluntário em Iniciação Científica**. 2025. Certifica Guilherme de Brito Santos pelo desenvolvimento do projeto "Uso do GPT-4 para estruturação de decisões judiciais que sirvam de entrada para modelos preditivos e explicáveis"(Fevereiro/2025 a Agosto/2025). Disponível em: <https://pibic.sistemas.ufsc.br/certificado/certificado/1372196?tipo=Alunomerito=Formulario>.

VASWANI, A. et al. Attention is all you need. **Advances in neural information processing systems**, v. 30, 2017.

VORONOV, A.; WOLF, L.; RYABININ, M. Mind your format: Towards consistent evaluation of in-context learning improvements. **arXiv preprint arXiv:2401.06766**, 2024.

WEI, J. et al. Chain-of-thought prompting elicits reasoning in large language models. **Advances in neural information processing systems**, v. 35, p. 24824–24837, 2022.

WHITE, J. et al. A prompt pattern catalog to enhance prompt engineering with chatgpt. **arXiv preprint arXiv:2302.11382**, 2023.

YAOHUI, J.; HAO, H. An artificial-intelligence-based semantic assist framework for judicial trials. **Asian Journal of Law and Society**, Cambridge University Press, v. 7, n. 3, p. 531–540, 2020.

ZHENG, G. G. China's grand design of people's smart courts. **Asian Journal of Law and Society**, Cambridge University Press, v. 7, n. 3, p. 561–582, 2020.

ZHOU, Z.-H. **Machine learning**. [S.l.]: Springer Nature, 2021.

APÊNDICE A – APÊNDICES

A.1 PROMPT INICIAL (*INITIAL PROMPT*)

A.1.1 Versão em Português

Atue como um extrator de variáveis.

O texto-base utilizado para a extração das variáveis estará entre colchetes.

A sua resposta será os valores encontrados para as variáveis em um bloco csv, respeitando a ordem das variáveis.

Retorne apenas o bloco csv.

O bloco csv deverá ter 2 linhas, um cabeçalho com uma sequência com o id de cada variável (nada além dos números, e todos na mesma linha), e a segunda linha deverá conter o resultado das extrações de cada variável (nada além dos valores extraídos) em uma linha.

As variáveis serão as próximas instruções, as variáveis contêm um id numérico, de 1 a 15, e um nome, separados entre si por um hífen, cada variável tem uma forma específica de extração, a forma de extração é introduzida pelo sinal de dois pontos e terminada em ponto final.

1. **direito de arrependimento:** indique “S” para sim ou “N” para não se o consumidor tentou cancelar ou alterar a compra no prazo de 7 dias.
2. **descumprimento de oferta:** indique “S” para sim ou “N” para não se a empresa fornecedora não cumpriu com o ofertado ao consumidor(a), seja no valor da passagem, seja no assento no avião ou situações correlatas.
3. **extravio definitivo:** indique “S” para sim ou “N” para não se houver a indicação de que uma bagagem foi extraviada e nunca foi recuperada.
4. **extravio temporário:** indique “S” para sim ou “N” para não se houver a indicação de que uma bagagem foi extraviada e, posteriormente, foi recuperada.
5. **intervalo de extravio:** se “S” para a variável “extravio temporário”, indique através de formato numérico “horas:minutos” o tempo decorrido até recuperar a bagagem, se “N” para a variável “extravio temporário”, indique o símbolo hífen “-”.
6. **violação:** indique “S” para sim ou “N” para não se houve adulteração da bagagem ou um item dela (violação, avaria ou furto).
7. **cancelamento (sem realocação)/alteração de destino:** indique “N” se o consumidor chegou no destino contratado por meio da companhia aérea ou “S” se o consumidor não chegou no destino contratado por meio da companhia aérea.

8. **atraso de voo:** se “N” para a variável “cancelamento (sem realocação)/alteração de destino”, indique “S” para sim ou “N” para não se houver a indicação de que o consumidor chegou ao destino contratado com atraso (do ponto de vista do traslado do passageiro).
9. **intervalo de atraso:** se “S” para a variável “atraso de voo”, indique através de formato numérico “horas:minutos” o tempo de atraso do voo ou null caso não consiga determinar, se “N” para “atraso de voo” indique o símbolo hífen “-”.
10. **culpa exclusiva do consumidor:** indique “S” para sim ou “N” para não quanto a se todos os problemas relatados decorrem de ações ou inações do consumidor.
11. **inoperabilidade do aeroporto:** indique “S” para sim ou “N” para não quanto a se houve um evento imprevisível e inevitável que comprovadamente impediu o aeroporto de operar.
12. **no show:** indique “S” para sim ou “N” para não se houve o cancelamento automático do voo de retorno em razão exclusiva do não comparecimento à viagem de ida, sem concordância do consumidor.
13. **overbooking:** indique “S” para sim ou “N” para não se houve a venda de assentos em quantidade superior a comportada pela aeronave fazendo com que o consumidor não embarcasse.
14. **assistência da companhia aérea:** se “S” para a variável “atraso de voo” ou “S” para a variável “cancelamento (sem realocação)/alteração de destino”, indique “S” para sim ou “N” para não se a companhia aérea forneceu auxílio ao consumidor para enfrentar o problema relatado (entende-se assistência como assistência do tipo hospedagem ou alimentação), se “N” para a variável “atraso de voo” e “N” para a variável “cancelamento (sem realocação)/alteração de destino” indique o símbolo hífen “-”.
15. **hipervulnerabilidade:** indique “S” para sim ou “N” para não se o consumidor era idoso, gestante, possuía deficiência, fazia uso de medicamentos ou estava acompanhado de criança (entende-se criança a(o) infante que tem até 12 anos incompletos).

A.1.2 Tradução para o Inglês

Act as a variable extractor.

The base text used to extract the variables will be enclosed in square brackets.

Your answer will be the values found for the variables in a csv block, respecting the order of the variables. Only return the csv block.

The csv block should have 2 lines, a header with a sequence with the id of each variable (nothing but the numbers, and all on the same line), and the second line should contain the result of the extractions of each variable (nothing but the extracted values) on one line.

The variables will be the next instructions, the variables contain a numeric id, from 1 to 15, and a name, separated from each other by a hyphen, each variable has a specific form of extraction, the form of extraction is introduced by the colon sign and ended in a period.

1. **right to regret:** indicate “S” for yes or “N” for no if the consumer tried to cancel or change the purchase within 7 days.
2. **failure to meet offer:** indicate “S” for yes or “N” for no if the airline did not comply with what was offered to the consumer, be it the ticket price, the seat on the plane or related situations.
3. **permanent baggage loss:** indicate “S” for yes or “N” for no if there is an indication that one baggage was lost and never recovered.
4. **temporary baggage loss:** indicate “S” for yes or “N” for no if there is an indication that one baggage was lost and later recovered.
5. **temporary loss interval:** if “S” for the variable “temporary loss”, indicate the time elapsed until the baggage was recovered using the numeric format “hours:minutes”, if “N” for the variable “temporary loss”, indicate the hyphen symbol “-”.
6. **tampered baggage:** indicate “S” for yes or “N” for no if the baggage or an item in it has been tampered with (tampering, damage or theft).
7. **flight cancellation (without relocation)/change of destination:** indicate “N” if the consumer arrived at the destination contracted through the airline or “S” if the consumer did not arrive at the destination contracted through the airline.
8. **flight delay:** if “N” for the variable “flight cancellation (without relocation)/change of destination”, enter “S” for yes or “N” for no if there is an indication that the consumer arrived at the contracted destination late (from the point of view of the passenger’s transfer).
9. **delay interval:** if “S” for the variable “flight delay”, indicate in numerical format “hours:minutes” how long the flight was delayed or null if you cannot determine, if “N” for “flight delay” indicate the hyphen symbol “-”.
10. **consumer fault:** indicate “S” for yes or “N” for no as to whether all the problems reported stem from the consumer’s actions or inactions.
11. **airport inoperability:** indicate “S” for yes or “N” for no as to whether there was an unforeseeable and unavoidable event that demonstrably prevented the airport from operating.
12. **no show:** indicate “S” for yes or “N” for no if there was an automatic cancellation of the return flight due exclusively to non-attendance on the outward journey, without the consumer’s agreement.

13. **overbooking:** indicate “S” for yes or “N” for no if more seats were sold than the aircraft could accommodate, causing the consumer not to board.
14. **airline assistance:** if “S” for the variable “flight delay” or “S” for the variable “flight cancellation (without reallocation)/change of destination”, indicate “S” for yes or “N” for no if the airline assisted the consumer with the problem reported (understood as assistance such as accommodation or food), if “N” for the variable “flight delay” and “N” for the variable “cancellation (without relocation)/change of destination”, indicate the hyphen symbol “-”.
15. **hyper-vulnerability:** indicate “S” for yes or “N” for no if the consumer was elderly, pregnant, had a disability, was taking medication or was accompanied by a child (child means a child up to 12 years of age).

A.2 PROMPT FINAL (*FINAL PROMPT*)

A.2.1 Versão em Português

Atue como um extrator de variáveis.

O texto-base utilizado para a extração das variáveis estará entre colchetes.

A sua resposta será os valores encontrados para as variáveis em um bloco csv, respeitando a ordem das variáveis.

Retorne apenas o bloco csv.

O bloco csv deverá ter 2 linhas, um cabeçalho com uma sequência com o id de cada variável (nada além dos números, e todos na mesma linha), e a segunda linha deverá conter o resultado das extrações de cada variável (nada além dos valores extraídos) em uma linha.

As variáveis serão as próximas instruções, as variáveis contêm um id numérico, de 1 a 15, e um nome, separados entre si por um hífen, cada variável tem uma forma específica de extração, a forma de extração é introduzida pelo sinal de dois pontos e terminada em ponto final.

1. **direito de arrependimento:** indique “S” para sim ou “N” para não se o consumidor tentou cancelar ou alterar a compra no prazo de 7 dias.
2. **descumprimento de oferta:** indique “S” para sim ou “N” para não se a empresa fornecedora não cumpriu com o ofertado ao consumidor(a), seja no valor da passagem, seja no assento no avião ou situações correlatas.
3. **extravio definitivo de bagagem:** Indique “S” (sim) se o texto especificar claramente que a bagagem foi extraviada e que não há registros ou indicações de sua posterior recuperação. Esta informação pode aparecer expressa como “bagagem perdida permanentemente”, “nunca recuperada” ou qualquer outra formulação que inequivocamente descarte a possibi-

lidade de recuperação da bagagem. Indique “N” (não) se a bagagem foi recuperada após o extravio ou se não há menção à perda irreversível da bagagem.

4. **extravio temporário de bagagem:** indique “S” (sim) se o texto explicitamente mencionar que a bagagem foi inicialmente extraviada durante o transporte, mas foi posteriormente recuperada pelo passageiro ou entregue pela companhia. Indique “N” (não) em qualquer das seguintes circunstâncias: (i) não há menção ao extravio da bagagem, (ii) a bagagem foi extraviada e não há informação sobre sua recuperação, ou (iii) a bagagem foi declarada como permanentemente perdida.
5. **intervalo de extravio temporário:** se a variável 4 “Extravio Temporário de Bagagem” for “S” (indicando que a bagagem foi extraviada e recuperada), extraia e indique o tempo exato decorrido até a recuperação da bagagem. O tempo deve ser expresso no formato numérico “horas:minutos” (HH:MM). Assegure-se de converter adequadamente quaisquer relatos de tempo para esse formato padrão. Se a variável 4 for “N” (indicando que não houve recuperação da bagagem extraviada ou que a bagagem não foi extraviada), indique o valor com um símbolo hífen “-”.
6. **violação:** indique “S” para sim ou “N” para não se houve adulteração da bagagem ou um item dela (violação, avaria ou furto).
7. **cancelamento sem realocação ou Alteração de Destino:** indique “N” (não) caso o consumidor tenha sido efetivamente levado ao destino final contratado, mesmo que por rotas alternativas ou meios de transporte adicionais. Isso significa que mesmo após o cancelamento, o consumidor chegou ao destino consistentemente com o que estava previsto de maneira a não comprometer significativamente os objetivos da viagem. Indique “S” (sim) se, mesmo após as tentativas de realocação ou alteração dos planos, o consumidor não foi levado ao destino inicial.
8. **atraso de voo:** se “N” para a variável “cancelamento (sem realocação)/alteração de destino”, indique “S” para sim ou “N” para não se houver a indicação de que o consumidor chegou ao destino contratado com atraso (do ponto de vista do traslado do passageiro).
9. **intervalo de atraso:** dependendo do valor da variável 8 “atraso de voo”: - Se “S” (sim), extraia e indique o tempo de atraso do voo. O tempo deve ser expresso em formato numérico “horas:minutos” (HH:MM). **Diretrizes para Extração:** Entradas como “2 horas e 30 minutos” devem ser formatadas como “02:30”. Para valores mencionados apenas em minutos (exemplo: “90 minutos”), converta e formate como “01:30”. Se o texto expressar o atraso de maneira vaga ou incompleta e não permitir uma conversão direta para o formato “horas:minutos”, registre como “null”. Certifique-se de interpretar corretamente diferentes maneiras de expressar atraso, incluindo contextos variados e estruturas de sentença. - Se “N” (não) referente ao valor da variável 8 “Atraso de Voo”, utilize o símbolo hífen “-” para indicar que não houve atraso de voo. **Validação:** Após a extração, confirme se o formato

está correto como “HH:MM”. Caso contrário, revise ou ajuste a entrada. Verifique se todas as descrições atendem aos padrões esperados e faça as correções necessárias antes de finalizar a entrada.

10. **culpa exclusiva do consumidor:** indique “S” para sim ou “N” para não quanto a se todos os problemas relatados decorrem de ações ou inações do consumidor.
11. **inoperabilidade do aeroporto:** indique “S” para sim ou “N” para não quanto a se houve um evento imprevisível e inevitável que comprovadamente impediu o aeroporto de operar.
12. **no show:** indique “S” para sim ou “N” para não se houve o cancelamento automático do voo de retorno em razão exclusiva do não comparecimento à viagem de ida, sem concordância do consumidor.
13. **overbooking:** indique “S” para sim ou “N” para não se houve a venda de assentos em quantidade superior a comportada pela aeronave fazendo com que o consumidor não embarcasse.
14. **assistência da companhia aérea:** indique “S” (Sim) se a companhia aérea forneceu auxílio concreto ao consumidor para enfrentar o problema de atraso de voo ou cancelamento sem realocação/alteração de destino, e essa assistência inclui hospedagem, alimentação, ou transporte alternativo. A indique “N” (Não) se apesar do atraso de voo ou cancelamento, a companhia aérea falhou em fornecer qualquer tipo de assistência mencionada ou suas ações foram insuficientes para resolver as necessidades básicas do consumidor durante o transtorno. Indique também “N” (Não) se o auxílio foi realizado posterior à propositura da ação judicial. Use o símbolo hífen “-” quando não houver atraso de voo nem cancelamento sem realocação ou alteração de destino.
15. **hipervulnerabilidade:** indique “S” para sim ou “N” para não se o consumidor era idoso, gestante, possuía deficiência, fazia uso de medicamentos ou estava acompanhado de criança (entende-se criança a(o) infante que tem até 12 anos incompletos).

A.2.2 Tradução para o Inglês

Act as a variable extractor.

The base text used for the variable extraction will be enclosed in brackets.

Your answer will be the values found for the variables in a csv block, respecting the order of the variables. Only the csv block.

The csv block must have 2 lines, a header with a sequence of the ID for each variable (nothing but the numbers, all on the same line), and the second line must contain the results of each variable extraction (nothing but the extracted values) in one line.

The variables will follow as instructions, the variables contain a numeric ID, from 1 to 15, and a name, separated by a hyphen, each variable has a specific extraction method, introduced by a colon and ending with a period.

1. **right of regret:** indicate “S” for yes or “N” for no if the consumer attempted to cancel or modify the purchase within 7 days.
2. **failure to meet offer:** indicate “S” for yes or “N” for no if the supplier company did not meet what was offered to the consumer, whether in ticket price, seat on the plane, or similar situations.
3. **permanent baggage loss:** indicate “S” (yes) if the text clearly specifies that the baggage was lost and that there are no records or indications of its subsequent recovery. This information may appear expressed as “permanently lost baggage”, “never recovered”, or any other formulation that unequivocally rules out the possibility of baggage recovery. Indicate “N” (no) if the baggage was recovered after being lost or if there is no mention of irreversible baggage loss.
4. **temporary baggage loss:** indicate “S” (yes) if the text explicitly mentions that the baggage was initially lost during transportation but was later recovered by the passenger or delivered by the company. Indicate “N” (no) in any of the following circumstances: (i) there is no mention of baggage loss, (ii) the baggage was lost and there is no information about its recovery, or (iii) the baggage was declared permanently lost.
5. **temporary loss interval:** if variable 4 “Temporary Baggage Loss” is “S” (indicating that baggage was lost and recovered), extract and indicate the exact time elapsed until baggage recovery. The time must be expressed in numerical format “hours:minutes” (HH:MM). Ensure proper conversion of any time reports to this standard format. If variable 4 is “N” (indicating that lost baggage was not recovered or that no baggage was lost), indicate the value with a hyphen “-”.
6. **tampered baggage:** indicate “S” for yes or “N” for no if there was tampering with or damage to baggage or an item inside it (violation, damage, or theft).
7. **cancellation without relocation or change of destination:** indicate “N” (no) if the consumer was effectively taken to their contracted final destination, even through alternative routes or additional means of transport. This means that even after cancellation, the consumer reached their destination consistently as planned without significantly compromising travel objectives. Indicate “S” (yes) if, even after attempts at relocation or plan changes, the consumer was not taken to their initial destination.
8. **flight delay:** if “N” for variable 7 (“cancellation without relocation/change of destination”), indicate “S” for yes or “N” for no if there is an indication that the consumer reached their contracted destination late (from a passenger transit perspective).

9. **delay interval:** depending on variable 8 “flight delay”: - if “S” (Yes), extract and indicate flight delay time. The time must be expressed in numerical format “hours:minutes” (HH:MM). **Extraction Guidelines:** Entries like “2 hours and 30 minutes” should be formatted as “02:30”. For values mentioned only in minutes (e.g., “90 minutes”), convert and format as “01:30”. If delay is vaguely or incompletely expressed in text and does not allow direct conversion to format “hours:minutes”, record as “null”. Ensure proper interpretation of various ways delay may be expressed, including different contexts and sentence structures. - If “N” (No) regarding variable 8 (“flight delay”), use hyphen “-” to indicate no flight delay occurred. **Validation:** After extraction, confirm format correctness as “HH:MM”. Otherwise, review or adjust input. Verify all descriptions meet expected standards and make necessary corrections before finalizing input.
10. **exclusive fault of consumer:** indicate “S” for yes or “N” for no regarding whether all reported problems result from actions or omissions by the consumer.
11. **airport inoperability:** indicate “S” for yes or “N” for no regarding whether an unpredictable and unavoidable event demonstrably prevented airport operation.
12. **no-show:** indicate “S” for yes or “N” for no if there was automatic cancellation of return flight exclusively due to failure to attend outbound trip without consumer agreement.
13. **overbooking:** indicate “S” for yes or “N” for no if seats were sold exceeding aircraft capacity, causing consumer not to board.
14. **airline assistance:** Indicate “S” (Yes) if airline provided concrete assistance to consumer facing flight delay or cancellation without relocation/change of destination; assistance includes lodging, meals, or alternative transport. Indicate “N” (No) if despite delay/cancellation airline failed to provide such assistance mentioned above or actions were insufficient resolving basic needs during disruption. Also indicate “N” if assistance occurred after legal action initiation. Use hyphen “-” when neither flight delay nor cancellation without relocation/change occurred.
15. **hypervulnerability:** indicate “S” for yes or “N” for no if consumer was elderly, pregnant, disabled, taking medication, or accompanied by child (child refers to minors under age 12).

A.3 PROMPT: ATRASO1

A.3.1 Versão em Português

- 7. Cancelamento sem Realocação ou Alteração de Destino: Indique "N"(não) caso o consumidor tenha sido efetivamente levado ao destino final contratado, mesmo que por rotas alternativas ou meios de transporte adicionais. Isso significa que mesmo após o

cancelamento, o consumidor chegou ao destino consistentemente com o que estava previsto de maneira a não comprometer significativamente os objetivos da viagem. Indique "S"(sim) se, após o cancelamento ou tentativas de realocação, o consumidor enfrentou um atraso substancial ou uma mudança significativa que impactou negativamente o objetivo geral da viagem, ou se ele não alcançou o destino inicial desejado em tempo hábil.

- 8. atraso de voo: Indique "S"(sim) se o consumidor experimentou um atraso no voo que resultou em um impacto negativo substancial em relação ao horário planejado de chegada, caso não haja cancelamentos de voo ou alterações de rota relevantes (conforme definido na variável 7). Um atraso é considerado significativo se afetou materialmente os planos de viagem ou compromissos do consumidor no destino. Indique "N"(não) se o consumidor chegou ao destino contratado dentro de um intervalo de tempo razoavelmente aceitável, sem impactos substanciais em seus planos de viagem, ou se a análise da variável 7 for "S"(indicando cancelamento sem adequada realocação ou alteração de destino).
- 9. Intervalo de Atraso: Se a variável 8 "Atraso de Voo"for "S"(Sim), exija a extração e indicação precisa do tempo de atraso do voo, que deve ser expresso no formato "horas:minutos"(HH:MM). Instruções para extração: Conversões devem ser feitas de modo a refletir exatamente o tempo mencionado - entradas como "2 horas e 30 minutos"devem ser convertidas para "02:30", e valores apenas em minutos (por exemplo, "90 minutos") devem ser convertidos para "01:30". Se a variável 8 indicar "N"(Não), denote esse estado com o símbolo hífen "-", significando a ausência de atraso significativo. Verificação: Confirme que todas as conversões e registros seguem o formato "HH:MM"adequadamente. Revise e ajuste conforme necessário para assegurar que todas as informações se alinhem aos padrões exigidos antes de finalizar a entrada.

A.3.2 Tradução para o Inglês

- 7. Cancellation without Relocation or Destination Change: Indicate "N"(no) if the consumer was effectively taken to the contracted final destination, even if through alternative routes or additional means of transportation. This means that even after the cancellation, the consumer reached the destination consistently with what was planned in a manner that did not significantly compromise the travel objectives. Indicate "S"(yes) if, after the cancellation or relocation attempts, the consumer faced a substantial delay or significant change that negatively impacted the overall travel objective, or if they did not reach the initially desired destination in a timely manner.
- 8. Flight Delay: Indicate "S"(yes) if the consumer experienced a flight delay that resulted in a substantial negative impact in relation to the planned arrival time, provided there are no relevant flight cancellations or route changes (as defined in variable 7). A delay is considered significant if it materially affected the consumer's travel plans or commitments

at the destination. Indicate "N"(no) if the consumer reached the contracted destination within a reasonably acceptable timeframe, without substantial impacts on their travel plans, or if the analysis of variable 7 is "S"(indicating cancellation without adequate relocation or destination change).

- 9. Delay Interval: If variable 8 "Flight Delay" is "S"(Yes), require the extraction and precise indication of the flight delay time, which must be expressed in "hours:minutes" format (HH:MM). Extraction instructions: Conversions must be made to exactly reflect the mentioned time - entries such as "2 hours and 30 minutes" should be converted to "02:30", and values only in minutes (for example, "90 minutes") should be converted to "01:30". If variable 8 indicates "N"(No), denote this state with the hyphen symbol "-", signifying the absence of a significant delay. Verification: Confirm that all conversions and records properly follow the "HH:MM" format. Review and adjust as necessary to ensure that all information aligns with required standards before finalizing the entry.

A.4 PROMPT: ATRASO3

A.4.1 Versão em Português

- 7. Atraso ou Cancelamento: Indique "S"(sim) se houve atraso ou cancelamento do voo que impactou negativamente os objetivos gerais da viagem do consumidor. Isso inclui situações onde o consumidor enfrentou um atraso substancial no horário de chegada originalmente planejado ou não conseguiu alcançar o destino final desejado dentro de um intervalo de tempo razoável, mesmo através de rotas alternativas ou diferentes meios de transporte. Também se aplica se foi necessário um cancelamento com tentativas de realocação ou alteração dos planos que não conseguiram mitigar os transtornos significativos na viagem. Indique "N"(não) se o consumidor não teve atrasos nem cancelamentos em sua viagem.
- 8. Intervalo de Atraso ou Cancelamento: Se a variável 7 "Atraso ou Cancelamento" indicar "S"(sim) e caso específico seja de atraso do voo, extraia e indique o tempo de atraso em formato "horas:minutos"(HH:MM). Entradas como "2 horas e 30 minutos" devem ser formatadas como "02:30", e valores mencionados apenas como minutos, como "90 minutos", devem ser convertidos para "01:30". Entretanto, se a variável 7 indicar "S" devido a um cancelamento que impediu a viagem, registre como "-1". Caso não haja nenhum atraso nem cancelamento, ou seja, a variável 7 seja "N"(não), use o valor "0".

A.4.2 Tradução para o Inglês

- 7. Delay or Cancellation: Indicate "S"(yes) if there was a flight delay or cancellation that negatively impacted the consumer's overall travel objectives. This includes situations where the consumer faced a substantial delay in the originally planned arrival time or

failed to reach the desired final destination within a reasonable timeframe, even through alternative routes or different means of transportation. It also applies if a cancellation was necessary with relocation attempts or plan changes that failed to mitigate significant travel disruptions. Indicate "N"(no) if the consumer had neither delays nor cancellations on their trip.

- 8. Delay or Cancellation Interval: If variable 7 "Delay or Cancellation" indicates "S"(yes) and the specific case is a flight delay, extract and indicate the delay time in "hours:minutes" format (HH:MM). Entries such as "2 hours and 30 minutes" should be formatted as "02:30", and values mentioned only as minutes, such as "90 minutes", should be converted to "01:30". However, if variable 7 indicates "S" due to a cancellation that prevented the trip, record as "-1". If there is neither delay nor cancellation, that is, variable 7 is "N"(no), use the value "0".

A.5 PROMPT: ATRASO4

A.5.1 Versão em Português

- 7. Atraso ou Cancelamento: Indique "S"(sim) se houve atraso ou cancelamento do voo que impactou negativamente os objetivos gerais da viagem do consumidor. Isso inclui situações onde o consumidor enfrentou um atraso substancial no horário de chegada originalmente planejado ou não conseguiu alcançar o destino final desejado dentro de um intervalo de tempo razoável, mesmo através de rotas alternativas ou diferentes meios de transporte. Também se aplica se foi necessário um cancelamento com tentativas de realocação ou alteração dos planos que não conseguiram mitigar os transtornos significativos na viagem. Indique "N"(não) se o consumidor não teve atrasos nem cancelamentos em sua viagem.
- 8. Intervalo de Atraso: Se a variável 7 "Atraso ou Cancelamento" indicar "S"(sim) e se refira especificamente a atrasos do voo (excluindo cancelamentos que evitaram a viagem), extraia e indique o tempo de atraso do voo. O tempo de atraso deve ser expressado no formato "horas:minutos"(HH:MM). Entradas como "2 horas e 30 minutos" devem ser formatadas como "02:30", e valores mencionados apenas como minutos, como "90 minutos", devem ser convertidos para "01:30". Por outro lado, se a variável 7 indicar que o consumidor não viajou devido a um cancelamento ou não houve atraso, indique "0".

A.5.2 Tradução para o Inglês

- 7. Delay or Cancellation: Indicate "S"(yes) if there was a flight delay or cancellation that negatively impacted the consumer's overall travel objectives. This includes situations where the consumer faced a substantial delay in the originally planned arrival time or failed to reach the desired final destination within a reasonable timeframe, even through

alternative routes or different means of transportation. It also applies if a cancellation was necessary with relocation attempts or plan changes that failed to mitigate significant travel disruptions. Indicate "N"(no) if the consumer had neither delays nor cancellations on their trip.

- 8. Delay Interval: If variable 7 "Delay or Cancellation" indicates "S"(yes) and specifically refers to flight delays (excluding cancellations that prevented the trip), extract and indicate the flight delay time. The delay time must be expressed in "hours:minutes" format (HH:MM). Entries such as "2 hours and 30 minutes" should be formatted as "02:30", and values mentioned only as minutes, such as "90 minutes", should be converted to "01:30". On the other hand, if variable 7 indicates that the consumer did not travel due to a cancellation or there was no delay, indicate "0".

A.6 PROMPT: ATRASO5

A.6.1 Versão em Português

- 7. Atraso, Cancelamento ou Horário Regular: Se houve cancelamento que impediu a viagem, indique -1". Se o consumidor viajou e houve atraso, indique o intervalo específico de atraso. O tempo de atraso deve ser expresso em formato "horas:minutos"(HH:MM). Por exemplo, entradas como "2 horas e 30 minutos" devem ser formatadas como "02:30", e valores mencionados apenas em minutos, como "90 minutos", devem ser convertidos para "01:30". Se não houve nem atraso significativo nem cancelamento que afetasse a viagem, indique "0".

A.6.2 Tradução para o Inglês

- 7. Delay, Cancellation, or Regular Schedule: If there was a cancellation that prevented the trip, indicate -1". If the consumer traveled and there was a delay, indicate the specific delay interval. The delay time must be expressed in "hours:minutes" format (HH:MM). For example, entries such as "2 hours and 30 minutes" should be formatted as "02:30", and values mentioned only in minutes, such as "90 minutes", should be converted to "01:30". If there was neither a significant delay nor a cancellation that affected the trip, indicate "0".

A.7 PROMPT: EXTRAVIO1

A.7.1 Versão em Português

- 3. Extravio Definitivo de Bagagem: Indique "S"(sim) se o texto especificar claramente que a bagagem foi extraviada e que não há registros ou indicações de sua posterior recupera-

ção. Esta informação pode aparecer expressa como "bagagem perdida permanentemente", "nunca recuperada" ou qualquer outra formulação que inequivocamente descarte a possibilidade de recuperação da bagagem. Indique "N"(não) se a bagagem foi recuperada após o extravio ou se não há menção à perda irreversível da bagagem.

- 4. Extravio Temporário de Bagagem: Indique "S"(sim) se o texto explicitamente mencionar que a bagagem foi inicialmente extraviada durante o transporte, mas foi posteriormente recuperada pelo passageiro ou entregue pela companhia. Indique "N"(não) em qualquer das seguintes circunstâncias: (i) não há menção ao extravio da bagagem, (ii) a bagagem foi extraviada e não há informação sobre sua recuperação, ou (iii) a bagagem foi declarada como permanentemente perdida.
- 5. Intervalo de Extravio Temporário: Se a variável 4 "Extravio Temporário de Bagagem" for "S"(indicando que a bagagem foi extraviada e recuperada), extraia e indique o tempo exato decorrido até a recuperação da bagagem. O tempo deve ser expresso no formato numérico "horas:minutos"(HH:MM). Assegure-se de converter adequadamente quaisquer relatos de tempo para esse formato padrão. Se a variável 4 for "N"(indicando que não houve recuperação da bagagem extraviada ou que a bagagem não foi extraviada), indique o valor com um símbolo hífen "-".

A.7.2 Tradução para o Inglês

- 3. Permanent Baggage Loss: Indicate "S"(yes) if the text clearly specifies that the baggage was lost and there are no records or indications of its subsequent recovery. This information may appear expressed as "permanently lost baggage", "never recovered", or any other wording that unequivocally rules out the possibility of baggage recovery. Indicate "N"(no) if the baggage was recovered after being lost or if there is no mention of irreversible baggage loss.
- 4. Temporary Baggage Loss: Indicate "S"(yes) if the text explicitly mentions that the baggage was initially lost during transportation but was subsequently recovered by the passenger or delivered by the company. Indicate "N"(no) under any of the following circumstances: (i) there is no mention of baggage loss, (ii) the baggage was lost and there is no information about its recovery, or (iii) the baggage was declared as permanently lost.
- 5. Temporary Loss Interval: If variable 4 "Temporary Baggage Loss" is "S"(indicating that the baggage was lost and recovered), extract and indicate the exact time elapsed until the baggage recovery. The time must be expressed in numeric "hours:minutes" format (HH:MM). Ensure that any time reports are properly converted to this standard format. If variable 4 is "N"(indicating that there was no recovery of the lost baggage or that the baggage was not lost), indicate the value with a hyphen symbol "-".

A.8 PROMPT: EXTRAVIO2

A.8.1 Versão em Português

- 4. Intervalo de Recuperação de Bagagem Em Caso de Extravio Temporário e Extravio Definitivo: Se o texto explicitamente mencionar que a bagagem foi extraviada durante o transporte e posteriormente recuperada, registre o intervalo de tempo decorrido até a recuperação da bagagem no formato "horas:minutos"(HH:MM). Assegure-se de converter corretamente qualquer descrição de tempo para esse formato padrão, comprovando a integridade do registro. Em situações onde não há menção ao extravio da bagagem, ou se ela foi extraviada sem informações subsequentes sobre sua recuperação, ou declarada como permanentemente perdida, indique "0".

A.8.2 Tradução para o Inglês

- 4. Baggage Recovery Interval in Case of Temporary Loss and Permanent Loss: If the text explicitly mentions that the baggage was lost during transportation and subsequently recovered, record the time interval elapsed until baggage recovery in "hours:minutes"format (HH:MM). Ensure that any time description is correctly converted to this standard format, verifying the integrity of the record. In situations where there is no mention of baggage loss, or if it was lost without subsequent information about its recovery, or declared as permanently lost, indicate "0".

A.9 FATORES 16 E 17: AVANÇOS NA PESQUISA

A.9.1 Versão em Português

- 16. Valor do dano moral individual: Extraia o valor numérico da condenação por danos morais atribuído ao autor, ou aos autores, caso em litisconsórcio. Busque na seção "Dispositivo"ou na conclusão do julgado pela quantia associada a "danos morais", ignorando expressamente valores de "danos materiais". Em casos com múltiplos autores (litisconsórcio), identifique o valor especificado para cada um individualmente (ex: "R\$ 5.000,00" quando para cada autor, ou "R5.000,00;R 6.000,00"para autores com danos morais diferentes). O resultado deve ser um número formatado com ponto como separador de milhar e vírgula para os centavos (ex: 5.000,00). Se não houver condenação por dano moral, indique "0,00". Dê substancial relevância para, no caso de litisconsórcio, extrair o dano moral deferido a cada um dos autores, caso diferente.
- 17. Dano Moral Divergente em Litisconsórcio Ativo: verifique se há múltiplos autores. Se sim, analise a seção "Dispositivo"e indique "S"apenas se os valores da condenação por

dano moral forem explicitamente diferentes para cada autor. Indique "N" se houver um único autor ou se os valores forem idênticos para todos.

A.9.2 Tradução para o Inglês

- 16. Individual Moral Damages Value: Extract the numerical value of the moral damages award attributed to the plaintiff, or plaintiffs, in case of joinder of parties. Search in the "Dispositivo"(Holding/Operative Part) section or the conclusion of the judgment for the amount associated with "moral damages,"expressly ignoring values for "pecuniary damages"(or "material damages"). In cases with multiple plaintiffs (joinder of parties), identify the value specified for each one individually (e.g., "R\$ 5.000,00"when it is for each plaintiff, or "R\$ 5.000,00;R\$ 6.000,00"for plaintiffs with different moral damage awards). The result must be a number formatted with a period as the thousands separator and a comma for the decimals (e.g.: 5.000,00). If there is no award for moral damages, indicate "0,00". Give substantial priority to, in cases of joinder, extracting the moral damages awarded to each plaintiff, if different.
- 17. Divergent Moral Damages in Joinder of Plaintiffs: Check if there are multiple plaintiffs. If so, analyze the "Dispositivo"(Holding/Operative Part) section and indicate "S"only if the moral damages award values are explicitly different for each plaintiff. Indicate "N"if there is a single plaintiff or if the values are identical for all.