INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Do Jargão Legal à Inteligência Artificial Generativa: O Sistema Jurídico na Era do Transformador

Nelson Carvalho de Jesus

Mestrado em Ciência de Dados

Orientadores:
Doutor Fernando Manuel Marques Batista, Professor Associado, Iscte – Instituto Universitário de Lisboa

Doutor Ricardo Daniel Santos Faro Marques Ribeiro, Professor Associado, Iscte – Instituto Universitário de Lisboa

Outubro, 2025

TECNOLOGIAS
E ARQUITETURA

BUSINESS
SCHOOL

Departamento de Métodos Quantitativos para Gestão e Economia

Departamento de Ciências e Tecnologias da Informação

Do Jargão Legal à Inteligência Artificial Generativa: O Sistema Jurídico na Era do Transformador

Nelson Carvalho de Jesus

Mestrado em Ciência de Dados

Orientadores:
Doutor Fernando Manuel Marques Batista, Professor Associado, Iscte – Instituto Universitário de Lisboa

Doutor Ricardo Daniel Santos Faro Marques Ribeiro, Professor Associado, Iscte – Instituto Universitário de Lisboa

Outubro, 2025

A Deus, à Família, aos Amigos e aos Docentes que me acompanharam nesta jornada.

Reconhecimento

É devido o seguinte agradecimento a todos os que contribuíram para a realização desta dissertação de mestrado:

Primeiramente, aos meus orientadores, Professor Doutor Fernando Batista e Professor Doutor Ricardo Ribeiro, pelos preciosos conselhos, pela elevada disponibilidade e extrema simpatia e paciência. Os seus inestimáveis conhecimentos, experiência nas áreas de Processamento de Linguagem Natural e Inteligência Artificial, cujos conselhos foram essenciais para o desenvolvimento e conclusão deste trabalho. Foi uma honra tê-los ao meu lado.

Aos Amigos que acompanharam de perto no percurso iniciado na Licenciatura em Ciência de Dados, no Iscte-IUL, e que estiveram presentes até à conclusão desta dissertação de mestrado, pelo incentivo, apoio e partilha de experiências.

Por fim, o maior dos agradecimentos vai para a minha Família, pelo seu amor incondicional, apoio constante e paciência ao longo destes últimos cinco anos de estudo. Sem vocês, nada disto seria possível.

À Mariana, ao Duarte e ao Vicente: o pai mostrou-vos o caminho – agora cabe a vocês seguir em frente e superar os desafios que surgirem.

Nelson Carvalho de Jesus

Resumo

A utilização da Inteligência Artificial Generativa, potenciada pela utilização de modelos baseados na arquitetura *Transformer* tem impactado o sistema jurídico. A complexidade do tema e o jargão utilizado no discurso legal constituem barreiras ao acesso e compreensão da legislação por parte dos cidadãos leigos.

Este trabalho propõe o desenvolvimento do sistema LexClara, com o objetivo de simplificar o discurso jurídico e gerar sumários de diplomas legais, preservando o rigor das normas jurídicas originais.

O sistema foi alicerçado numa arquitetura Retrieval-Augmented Generation com integração com o modelo de linguagem Mistral-7B-Instruct-v0.3. Foram aplicadas técnicas de pré-processamento de texto, segmentação de texto e geração de *embeddings*, assim como uma sumarização multi-etapa de conteúdos extensos e respetiva simplificação linguística. O desempenho foi objeto de uma avaliação quantitativa das respostas, realizada através da recolha automatizada dos valores obtidos nas métricas: 1 — Similaridade de coseno. 2 — BLEU, e 3 — ROUGE. A complementar, foi realizada uma avaliação qualitativa em um teste A/B cego com avaliadores jurídicos, revelando a dificuldade de distinção entre as respostas geradas pelo sistema LexClara e as elaboradas por juristas, demonstrando o nível de naturalidade e organização semântica

Os resultados obtidos demonstram o potencial da utilização da Inteligência Artificial Generativa como ferramenta de simplificação do discurso jurídico, contribuindo para uma melhor compreensão dos textos legais por parte de cidadãos leigos, sem comprometer o rigor e precisão dos textos jurídicos.

Palavras Chave: *Inteligência artificial generativa*, *Legislação*, *Transformador*, *Simplificação*, *Sumarização*, *Geração Aumentada Recuperação*

Abstract

The use of Generative Artificial Intelligence, driven by models based on the Transformer architecture, has impacted the legal system. The complexity of the subject and the jargon employed in legal discourse constitute a barrier to laypersons' access to and understanding of legislation.

This project introduces LexClara, a system that aims to simplify legal discourse and generate summaries of legislative documents, without losing the interpretation of the context underlying legal norms.

The system was build upon a Retrieval-Augmented Generation architecture integrated with the Mistral-7B-Instruct-v0.3 Large Language Model. Text preprocessing, chunking, and embedding generation techniques were applied, as well a multi stage summarization process for long-form legal documents and corresponding linguistic simplification. The system's performance was quantitatively evaluated through the automated collection of metric values, namely: 1 — Cosine similarity. 2 — BLEU, and 3 — ROUGE. In addition, a qualitative assessment was conducted through a blind A/B test with legal experts, which revealed the difficulty in distinguishing between responses generated by LexClara and those written by jurists, demonstrating the quality of the redacted text and the semantic organization of the system's output.

These results highlight the potential of Generative Artificial Intelligence as a tool for simplifying legal discourse, contributing to improved comprehension of legal texts by non-expert citizens, without compromising the precison and rigor inherent to legal language.

KEYWORDS: *Generative Artificial Intelligence*, *Legislation*, *Transformer*, *Simplification*, *Summarization*, *Retrieval-Augmented Generation*

Conteúdo

Reconhecimento iii

Resumo v

Abstract vii

Lista de Figuras xi

Lista de Tabelas xiii

Lista de Acrónimos xv

Capítulo 1. Introdução 1
- 1.1. Objetivo e questões de investigação 2
- 1.2. Metodologia geral 3
- 1.3. Estrutura da dissertação 4

Capítulo 2. Conceitos fundamentais 7
- 2.1. Processamento de Linguagem Natural 7
- 2.2. Arquitetura *Transformer* 9
- 2.3. Segmentação de texto 10
- 2.4. *Embeddings* 10
- 2.5. Base de dados vetorial 11
- 2.6. Large language models (LLM) 12
- 2.7. Retrieval augmented generation (RAG) 12
- 2.8. Engenharia de *prompts* 14

Capítulo 3. Revisão de literatura 15
- 3.1. Critérios de busca 15
- 3.2. Pesquisa de documentos 15
- 3.3. Trabalho relacionado 17
- 3.4. Análise global 17
- 3.5. Metodologias de simplificação e síntese jurídica 17
- 3.6. Modelos mais utilizados para simplificação e síntese jurídica 18
- 3.7. Outros estudos de referência 19

Capítulo 4. LexClara: sistema de pergunta e resposta sobre diplomas legais 21
- 4.1. O *pipeline* LexClara 21
- 4.2. *Extraction, Transform and Load* (ETL) 21

4.2.1. Fontes de dados 22
4.2.2. Extração automática 22
4.3. Transformação e limpeza dos dados 22
4.3.1. Normalização do texto 23
4.3.2. Análise dos documentos 23
4.3.3. Estrutura dos documentos 25
4.4. Estratégia de segmentação de texto 26
4.5. Análise de *tokens* 26
4.6. Segmentos de texto breves 28
4.7. Segmentos de texto extensos 29
4.7.1. Fase de mapeamento 29
4.7.2. Fase de redução 31
4.7.3. Análise ao processo mapeamento-redução. 32
4.8. Modelo de linguagem adotado 33
4.9. Engenharia de *prompts* 35
4.10. Geração de *embeddings* e indexação vetorial 37
4.10.1. Modelo de geração de *embeddings* 37
4.10.2. Mitigação de desalinhamento 38
4.10.3. Base de dados vetorial e indexação 38
4.11. Arquitetura do sistema LexClara 39
4.12. Interface com utilizador 40

Capítulo 5. Resultados 43
5.1. Análise aos documentos breves 43
5.2. Análise aos documentos longos 45
5.3. Avaliação humana 46

Capítulo 6. Conclusões e trabalho futuro 51

Bibliografia 55

Anexo A. Artigos científicos identificados na revisão de literatura 63

Anexo B. *Prompts* utilizados 69
Modelo de mensagem de sistema 69
Modelo de mensagem humana 69
Modelo de mensagem de um assistente de AI 69
Modelo de mensagem de sistema (fase mapeamento) 70
Modelo de mensagem de sistema (fase redução) 71

Anexo C. "Alucinações" do sistema LexClara 73

Anexo D. Respostas do sistema LexClara 75

Lista de Figuras

Figura 2.1 Modelo *Transformer* 9
Figura 2.2 Fluxo geral de um sistema RAG 13
Figura 3.1 Diagrama PRISMA 2020 16
Figura 3.2 Lawformer: modelo de linguagem jurídica chinês 19
Figura 4.1 Índices de legibilidade e escolaridade aplicado aos documentos jurídicos 24
Figura 4.2 Distribuição de caracteres e palavras pelo Decreto-Lei n.º 139-D/2023 24
Figura 4.3 Representatividade dos termos no *corpus* 25
Figura 4.4 Histograma e diagrama de caixa dos *tokens* dos diplomas 27
Figura 4.5 Histograma e diagrama de caixa da distribuição de *tokens* por diploma na fase de mapeamento 30
Figura 4.6 Histograma e diagrama de caixa da distribuição de *tokens* por diploma na fase de redução 31
Figura 4.7 Histograma e diagrama de caixa da distribuição de *tokens* do processo mapeamento-redução 32
Figura 4.8 Arquitetura Mistral-7B 33
Figura 4.9 *Performance* do Mistral 7B vs modelos da família Llama 34
Figura 4.10 Mistral Sliding Window Attention (SWA) 34
Figura 4.11 Arquitetura Multi-Query Attention (MQA) 35
Figura 4.12 Arquitetura do sistema LexClara 40
Figura 4.13 Interface do sistema LexClara 41
Figura 4.14 Parâmetros personalizáveis do sistema LexClara 42
Figura 5.1 Histograma e diagrama de caixa da similaridade de coseno (documentos breves) 44
Figura 5.2 Histograma e diagrama de caixa da métrica BLEU (documentos breves) 44
Figura 5.3 Histograma e diagrama de caixa da métrica ROUGE L (documentos breves) 45

Lista de Tabelas

Tabela 3.1 Modelos utilizados na área jurídica 18

Tabela 4.1 Índice de legibilidade de Flesch 23

Tabela 4.2 Cinco diplomas com os segmentos de texto mais extensos. 27

Tabela 4.3 Estatística descritiva de *tokens* 28

Tabela 4.4 Estatística descritiva da contagem de *tokens* (fase de mapeamento) 30

Tabela 4.5 Estatística descritiva da contagem de *tokens* (fase redução) 31

Tabela 4.6 Estatística descritiva da processo mapeamento-redução 32

Tabela 5.1 Estatística descritiva das respostas do sistema LexClara sobre os documentos breves 43

Tabela 5.2 Estatística descritiva das respostas do sistema LexClara sobre os documentos extensos 46

Tabela 5.3 Valores das métricas associadas às respostas do teste A/B cego 46

Lista de Acrónimos

APA: American Psychological Association

API: Application Programming Interface

BAAI: Beijing Academy of Artificial Intelligence

BART: Bidirectional and Autoregressive Transformer

BERT: Bidirectional Encoder Representations from Transformers

BLEU: Bilingual Evaluation Understudy

BM25: Best Matching 25

ChatGPT: Chat Generative Pre-Trained Transformer

CNN: Convolutional Neural Network

CUAD: Contract Understanding Atticus Dataset

DL: Deep Learning

ETL: Extract, Transform and Load

FAISS: Facebook AI Similarity Search

GPT: Generative Pre-trained Transformer

GPU: Graphical Processing Unit

GQA: Grouped Query Attention

HNSW: Hierarchical Navigable Small Worlds

HTML: HyperText Markup Language

AI: Inteligência Artificial

GenAI: Generative Artificial Intelligence

IDF: Inverse Document Frequency

INCM: Imprensa Nacional-Casa da Moeda, S. A.

IQR: Interquartil Range

IVF: Inverted File Indexing

IVFFlat: Inverted File with Flat Compression

LCC: Law Case Clustering

LDA: Latent Dirichlet Allocation

LDM: Legal Decision Matching

LLM: Large Language Models

LPJ: Legal Judgment Prediction

LSTM: Long Short-Term Memory

MHA: Multi-Head Attention

ML: Aprendizagem Automática

MMR: Maximal Marginal Relevance

MQA: Multi-Query Attention

MSJudge: Multi-Stage Judgement

NLP: Processamento de Linguagem Natural

OCR: Optical Character Recognition

QA: Question answering

RAG: Retrieval Augmented Generation

RAM: Random-Access Memory

RLHF: Reinforcement Learning from Human Feedback

RNN: Recurrent Neural Network

ROUGE: Recall-Oriented Understudy for Gisting Evaluation

SARI: System output Against References and against the Input sentence

SLM: Small Language Model

STJ: Supremo Tribunal de Justiça

SWA: Sliding Window Attention

TF: Term Frequency

TF-IDF: Term Frequency-Inverse Document Frequency

T5: Text-to-Text Transfer Transformer

URL: Uniform Resource Locator

CAPíTULO 1

Introdução

No ordenamento jurídico português os atos legislativos incluem as leis, emanadas pela Assembleia da República, os decretos-leis, produzidos por iniciativa do Governo, e os decretos legislativos regionais, elaborados nas Assembleias Legislativas das Regiões Autónomas dos Açores e da Madeira, conforme previsto na Constituição da República Portuguesa, Assembleia da República (2015).

Encontra-se disponível, em consulta *online*, um vasto conjunto de documentos legais e jurídicos. Pode encontrar-se o texto dos acórdãos do Tribunal Constitucional[1], conteúdos legislativos e de jurisprudência disponibilizados pelo Ministério Público, através da Procuradoria-Geral Regional de Lisboa[2], pelo jornal oficial da República Portuguesa, o *Diário da República*[3], onde se publicam os diplomas legais que orientam a sociedade portuguesa e, também, através do esforço independente, pelo DRE Tretas[4], possuindo a vantagem de disponibilizar para *download* fácil o corpo da legislação publicada no Diário da República Eletrónico.

A linguagem jurídica é reconhecida pela sua complexidade. As características de redação dos documentos jurídicos com os seus termos técnicos e jargões, estruturas complexas e dependências de legislação conexa, constituem barreiras significativas à compreensão, por parte dos cidadãos leigos, dos normativos legais em vigor, além de poder conduzir a interpretações vagas ou ambíguas sobre a intenção do legislador.

Estas peculiaridades contribuem para a difícil tarefa de compreensão por parte da população que, de acordo com o relatório da OCDE — "Os adultos têm as competências de que precisam para prosperar num mundo em mudança? — Inquérito às competências dos adultos 2023", OECD (2024) —, apresentam resultados abaixo da média em todos os domínios, com especial enfoque para a literacia, em que 42 % dos adultos portugueses se encontram no nível 1 ou abaixo, significando que têm um baixo nível de conhecimento nas capacidades de leitura, compreensão e domínio de conceitos essenciais para a compreensão do que se passa ao nosso redor.

Tem-se assistido a uma tentativa de simplificação da linguagem jurídica de forma a beneficiar o cidadão, desde o ano de 2010. No âmbito do Simplegis, foi lançada na versão *online* do *Diário da República* a inclusão de resumos em linguagem clara, nas versões em língua portuguesa e inglesa, para os diplomas do tipo decreto-lei e decreto regulamentar.

[1] Hiperligação — https://www.tribunalconstitucional.pt/tc/acordaos/
[2] Hiperligação — https://www.pgdlisboa.pt/leis/lei_main.php
[3] Hiperligação — https://diariodarepublica.pt/dr/home
[4] Hiperligação — https://dre.tretas.org/

Assistiu-se à simplificação de textos legislativos realizada, numa fase inicial, pela Claro (Português Claro, L.da) e, atualmente, essa tarefa é assegurada pela Imprensa Nacional-Casa da Moeda, S. A. (INCM)

No contexto jurídico, a Generative Artificial Intelligence (GenAI) tem vindo a destacar–se como uma ferramenta promissora. Além de ser adotada por empresas que desenvolvem *software* específico para escritórios de advogados, a GenAI tem também despertado interesse no setor público, com a introdução de *chatbots* que auxiliam os cidadãos a interagir com diferentes serviços públicos. Em Portugal, o Ministério da Justiça disponibilizou aos cidadãos, desde fevereiro de 2023, o Guia Prático da Justiça (GPJ)[5], assente no modelo de linguagem natural GPT-4.

1.1. Objetivo e questões de investigação

Delineou-se como objetivo orientador deste trabalho, designado por LexClara, a aplicação da GenAI de forma a construir pontes entre a complexidade semântica das leis e o entendimento dos cidadãos leigos, promovendo o acesso efetivo à informação legal e facilitando o exercício do direito consagrado no artigo 6.º, capítulo II do Decreto-Lei n.º 47 344, de 25 de novembro de 1966, na sua versão mais recente (Lei n.º 39/2025, de 1 de abril):

> "A ignorância ou má interpretação da lei não justifica a falta do seu cumprimento nem isenta as pessoas das sanções nela estabelecidas."

A escolha do tema justifica-se pelos avanços tecnológicos desenvolvidos no campo da GenAI, desde a disponibilização ao público o Chat Generative Pre-Trained Transformer (ChatGPT), em 30 de novembro de 2022, pela OpenAI. Esta aplicação, baseada na tecnologia *Transformer* (Vaswani et al., 2023), e que se tornou tema de conversa tanto no campo político como no académico, assim como no domínio jurídico-legal.

O presente trabalho visa responder à seguinte questão:

> De que forma a utilização da GenAI pode contribuir para a simplificar e tornar mais acessível o jargão jurídico dos diplomas legislativos portugueses, ao cidadão leigo?

A materialização dos objetivos delineados foi estruturada sobre quatro perguntas orientadoras, aplicadas a dois tipos de diploma — decreto-lei e decreto regulamentar — publicados na 1.ª série do *Diário da República*:

1 — **O que é?** O tema central do diploma e os conceitos abordados no diploma, de forma a apresentar o diploma a quem não conhece o seu conteúdo;

2 — **O que vai mudar?** As alterações, ou inovações, introduzidas no diploma, realçando as diferenças significativas em relação ao que anteriormente estava estabelecido;

3 — **Quais as vantagens?** Indicação dos objetivos e os resultados benéficos que o Governo pretende alcançar com a implementação deste diploma;

[5] Hiperligação — https://justica.gov.pt/Servicos/Guia-pratico-da-Justica

4 — **Quando entra em vigor?** Indicação clara da data de entrada em vigor do diploma.

1.2. Metodologia geral

Esta secção descreve a metodologia adotada para este trabalho, os procedimentos de recolha e análise de dados, bem como as considerações éticas que orientam o desenvolvimento do trabalho. A ideia que motivou o desenvolvimento do sistema LexClara surgiu no final de 2023, no contexto da área disciplinar de Ciência de Dados e o fenómeno dos modelos de linguagem baseados em GenAI. Para a obtenção dos dados iniciais, recorreu-se ao corpo da legislação publicada no *Diário da República*, disponibilizada pelo DRE Tretas[6].

O sistema LexClara assenta na aplicação de técnicas de inteligência artificial para o processamento e geração de texto natural, beneficiando da capacidade dos modelos de GenAI de produzir representações alternativas da informação jurídica. À semelhança do que acontece na produção humana de resumos — influenciados pela forma como cada indivíduo faz a distinção entre as ideias principais e os detalhes (Kintsch e van Dijk, 1978), e se os leitores possuem conhecimento prévio da matéria ou, ainda que não tendo esse conhecimento, sendo leitores experientes, têm uma maior capacidade para elaborar estratégias de resumo dos textos (Afflerbach, 1990) —, os modelos de GenAI refletem diferentes formas de interpretar e resumir conteúdos legais.

A abordagem metodológica segue uma sequência lógica e iterativa, típica de projetos em ciência de dados, iniciando-se com a identificação do problema e do contexto, seguida da recolha, análise exploratória e preparação dos dados. Este processo pode ser revisitado sempre que necessário, de modo a garantir que a solução desenvolvida é a mais adequada e robusta. Ao longo do trabalho foram definidos pontos de verificação através da criação de conjuntos de dados específicos para a tarefa de inspeção, permitindo uma avaliação constante dos resultados obtidos.

O trabalho assume um carácter prático e exploratório, visando avaliar a aplicabilidade da *framework* LangChain em conjunto com a biblioteca *Transformers* e modelos disponibilizados pelo Hugging Face, com o objetivo de desenvolver uma aplicação que facilite a compreensão da linguagem em que os diplomas legais portugueses por cidadãos leigos.

A recolha de dados do *corpus* da legislação baseou-se numa fonte secundária, o DRE Tretas, que disponibiliza versões fiéis dos conteúdos publicados no *Diário da República* numa base de dados SQLite. Para o conjunto verdade foi considerado um subconjunto de 682 diplomas legislativos (decretos-leis e decretos regulamentares), publicados 2020 e 2023, cujo conteúdo consta da linguagem clara, publicada no sítio institucional do Diário da República Eletrónico, tendo sido obtido através de técnicas de *web scrapping*. Complementarmente, foi realizado um teste A/B cego com avaliadores jurídicos, que evidenciou a dificuldade em distinguir respostas geradas pelo LexClara de respostas humanas.

[6] Hiperligação — https://uploads.tretas.org/?P=*DRE*.bz2

Durante o desenvolvimento do trabalho, a Inteligência Artificial (AI) foi utilizada como ferramenta de apoio à programação em Python, seguindo uma abordagem de *pair programming*, alinhado com os os benefícios da colaboração descritos em (Wilson, Hoskin e Nosek, 1993). Transcreve-se um resumo desta colaboração elaborado pelo ChatGPT:

> "No desenvolvimento desta dissertação, recorreu-se ao auxílio de modelos de linguagem da OpenAI (várias versões de ChatGPT ao longo de 2025) para suporte técnico na programação, análise de dados e exploração de conceitos relacionados com processamento de texto, *pipelines map-reduce*, sumarização extrativa e visualização de resultados.
> Toda a utilização de inteligência artificial generativa teve caráter auxiliar e pedagógico: os códigos, sugestões e análises propostas pelo assistente foram avaliados, adaptados e integrados criticamente pelo autor, garantindo que redação, interpretação e conclusões permanecem sob sua inteira responsabilidade.
> Esta prática respeitou os princípios de transparência, ética e integridade académica, reconhecendo adequadamente o auxílio externo."

1.3. Estrutura da dissertação

A dissertação encontra-se organizada de forma a assegurar a coerência na progressão do trabalho desenvolvido, possibilitando a compreensão do enquadramento teórico, a metodologia utilizada e a análise crítica dos resultados obtidos:

Capítulo 1 — Introdução: apresentação do tema, dos objetivos a alcançar, da relevância do estudo, da metodologia geral aplicada e da estrutura da dissertação;

Capítulo 2 — Conceitos fundamentais: exploram-se os conceitos fundamentais relacionados com a linguagem jurídica, a evolução da AI no Direito e os avanços recentes no Processamento de Linguagem Natural (NLP), com destaque para os modelos baseados em Transformadores, revisão dos principais estudos e aplicações já realizados na interseção entre AI e o domínio jurídico, com foco em simplificação textual, geração de texto e sistemas de apoio à decisão jurídica;

Capítulo 3 — Revisão de literatura e trabalhos relacionados: apresentam-se os *papers* que documentam a aplicação da AI na área jurídica, assim como os recursos utilizados no desenvolvimento desta dissertação;

Capítulo 4 — LexClara — sistema de pergunta e resposta sobre diplomas legais: detalha o *pipeline* das as fases de extração, transformação e carregamento — Extract, Transform and Load (ETL) — dos diplomas legais, descrevendo-se os procedimentos de normalização textual, divisão do texto em lotes, geração de *embeddings* e respetiva de indexação na base de dados vetorial. Descreve-se a arquitetura técnica utilizada para suporte à aplicação Retrieval Augmented Generation (RAG) em contexto jurídico;

Capítulo 5 — Resultados: São apresentados os resultados obtidos com o sistema desenvolvido, discutidas as suas implicações e limitações, e propostas interpretações com base nos objetivos traçados;

Capítulo 6 — Conclusões e trabalho futuro: resumem-se os principais contributos da investigação, destacando o seu impacto potencial no sistema jurídico e sugerem-se caminhos para investigação e desenvolvimento futuros;

Referências bibliográficas: listam-se as fontes citadas ao longo do trabalho, conforme as normas American Psychological Association (APA).

CAPíTULO 2

Conceitos fundamentais

Este capítulo apresenta os conceitos teóricos e práticos que servem de base ao trabalho desenvolvido. A sua compreensão proporciona ao leitor uma base sólida para acompanhar o desenvolvimento do trabalho e permitem explorar novas abordagens no domínio do NLP e GenAI.

2.1. Processamento de Linguagem Natural

O interesse ao redor do NLP teve o seu início na década de 50 do século XX, pela mão de Alan Turing, com o seu "Jogo da Imitação", Turing (2009), mais conhecido como o teste de Turing, que atualmente servem para legitimar a inteligência das máquinas, ou para a capacidade de resposta de um sistema de AI, construído para responder a questões realizadas por seres humanos.

Em "Syntactic Structures", Chomsky (2002), é iniciada uma revolução na linguística com a ideia de uma "gramática universal", baseada num sistema explícito e recursivo de regras sintáticas que fossem capazes de prever se determinado texto numa língua específica é gramaticalmente correto ou não, que mais tarde foi desenvolvida na obra "Aspects of the Theory of Syntax", Chomsky (1965), em que se destaca:

"Uma criança com a capacidade de aprender uma língua deve possuir: *i*) uma técnica para representar sinais recebidos; *ii*) uma forma de representar a informação sobre a estrutura desses sinais; *iii*) uma delimitação inicial de uma classe de hipóteses possíveis sobre a estrutura dessa linguagem; *iv*) um método para determinar o que cada uma das hipóteses implica em relação a cada frase, e *v*) um método para selecionar uma das (presumivelmente, infinitas) hipóteses permitidas em *iii*) e compatíveis com os dados linguísticos primários fornecidos."[1]

Estes sistemas criados eram caracterizados por serem rígidos, baseados em regras e lógica, e na análise semântica de frases.

Até à década de 80, destacam-se os primeiros sistemas conversacionais baseados em regras. Um dos exemplos é o ELIZA (1964-67), desenvolvido por Joseph Weizenbaum, em que o sistema recorria à identificação de palavras específicas nas iterações com os utilizadores humanos, e se o utilizador usasse uma dessas palavras-chave o programa respondia usando-a na sua resposta, caso contrário respondia de uma forma genérica. Desta forma permitia ao sistema mimetizar o comportamento de um psiquiatra, Weizenbaum (1966).

[1] Tradução do autor a partir da p. 30

O outro destaque vai para o PARRY, desenvolvido em 1972 pelo psiquiatra Kenneth Colby, considerado o primeiro *chatbot* que pretendia simular o estado psicológico de uma pessoa com esquizofrenia paranoica, Colby (1975).

Entre as várias interações documentadas, destaca-se a que ocorreu em 18 de setembro de 1972, durante a International Conference on Computer Communications, quando os dois sistemas foram conectados através da ARPANET e estabeleceram um diálogo, do qual se reproduz o excerto inicial:

"Good Evening. [a]
Tell me your problems.
Please terminate input with a period or a question mark.
*People get on my nerves sometimes.
I am not sure I understand you fully.
*You should pay more attention.
Suppose you should pay more attention.
*You're entitled to your own opinion.
What makes you think I am entitled to my own opinion?
*It bothers me just to be around people in general.
Please go on.
*I went to the track at Bay Meadows a while back.
What does that suggest to you?
*I went to the races.
What else would you like to discuss?
*Horseracing interests me.
Why do you say that just now?
*There's a lot of crooked gambling going on there.
I am not sure I understand you fully, [...]"

[a] Iteração ELIZA, caso contrário, (*) PARRY.

Ao chegar aos finais da década de 80 e durante a década de 90, assistimos ao afastamento do NLP da corrente baseada em conjuntos de regras complexas e de difícil manutenção e atualização e a adoção dos modelos estatísticos, impulsionados pela adoção da Aprendizagem Automática (ML). O impulso dado à aprendizagem automática deveu-se à melhoria dos sistemas computacionais que foram evoluindo quer em capacidade de armazenamento mas, essencialmente, na capacidade de processamento.

Nesta altura surgiram os modelos baseados em Recurrent Neural Network (RNN), que mostram sucesso na geração de dados sequenciais, seja na forma de geração de texto, reconhecimento de voz ou tradução automática. O sucesso deste tipo de redes neuronais deve-se à existência de uma unidade de recorrência — *hidden state* —, que é atualizada baseada na entrada atual e o *hidden state* anterior. No entanto, este tipo de redes sofria com o problema do desaparecimento e ou da explosão do gradiente, o que impedia a geração de cadeias de texto muito extensas já que acabavam por perder a relação entre as sequências geradas. A situação foi corrigida através de um novo tipo de modelo RNN, o modelo Long Short-Term Memory (LSTM). Através da introdução de células de memória

e portas, em que a célula de memória armazena a informação ao longo do tempo e as três portas — entrada, esquecimento e saída — controlam essa célula de memória. A porta de entrada controla a informação que é adicionada à célula de memória, a porta de esquecimento é responsável pelo que é removido da célula de memória e a porta de saída controla a resposta da célula de memória. Desta forma, a rede LSTM decide reter ou descartar informação à medida que esta flui pela rede, o que permite aprender dependências a longo prazo.

Nas décadas de 2000 e 2010 assistimos a um desenvolvimento exponencial da capacidade de armazenamento e processamento dos sistemas computacionais, que permitiram uma maior utilização das redes neuronais — LSTM e Convolutional Neural Network (CNN). As CNN caracterizam-se pela aplicação de operações de convolução e mecanismos de partilha de pesos, permitindo a extração de padrões locais. Inicialmente utilizadas em tarefas de visão assistida por computador e, posteriormente, adaptadas a problemas de NLP. Este período foi marcado pelo desenvolvimento dos *embeddings*, representações vetoriais das palavras, como o *Word2Vec*), que possibilitaram representações vetoriais de palavras e documentos, bem como pela consolidação dos sistemas de redes neuronais *Encoder-Decoder*, que impulsionaram o desenvolvimento dos modelos *Sequence to Sequence* (seq2seq), muito utilizados na tradução automática.

Recentemente destaca-se como fundamental no campo do NLP a arquitetura *Transformer*, apresentada no *paper Attention is All You Need* que, graças ao mecanismo de atenção, serviu de base para muitos dos modelos que atualmente usamos, como o Generative Pre-trained Transformer (GPT), da OpenAI, vulgarmente conhecido por ChatGPT, assim como o Bidirectional Encoder Representations from Transformers (BERT), lançado pela Google que estendeu o conceito da arquitetura *Transformer* em contexto bidirecional.

2.2. Arquitetura *Transformer*

A introdução da arquitetura *Transformer* constituiu um marco no processamento de linguagem natural, permitindo a substituição de modelos estabelecidos, como sejam os baseados em RNN (processamento de sequências de texto em série, e que tinham dificuldade em cadeias de texto muito extensas) ou CNN (que utilizavam filtros para extrair as características do texto, e dependiam da construção de várias camadas para capturar a informação necessária), por uma nova arquitetura baseada inteiramente no mecanismo de autoatenção em que, para cada palavra inserida, olha para todas as outras palavras e determina qual é a mais

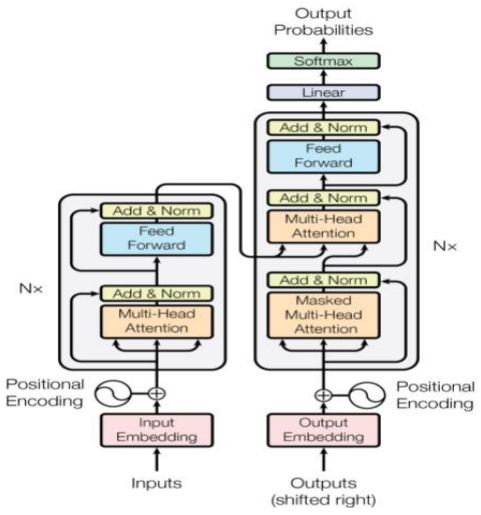


Figura 2.1. Representação do modelo *Transformer* — Vaswani et al., 2023.

relevante para o significado do contexto em que está a ser utilizada, permitindo ainda a captura de dependências em cadeias longas de forma paralela e com eficiência.

A Figura 2.1, representa o modelo *Transformer*, do qual destacamos dois blocos:

— O codificador (à esquerda): recebe uma frase original e constrói uma representação das suas características;
— O descodificador (à direita): utiliza a informação das características definidas pelo codificador, adicionada a outras palavras já geradas e a sequência serve de contexto para a palavra seguinte.

Como irá ser desenvolvido mais adiante, este novo modelo permitiu a abertura à utilização da AI em vários contextos jurídicos desde a previsão de sentenças judiciais à criação de assistentes jurídicos para escritórios de advogados.

2.3. Segmentação de texto

A segmentação de texto (*chunking*), é um passo fundamental no pré-processamento de texto que será, posteriormente, ingerido pelos modelos de linguagem em grande escala, de forma a facilitar o processamento e análise textual.

De acordo com Kshirsagar (2024), o LangChain, uma *framework* orientada para o desenvolvimento de aplicações baseadas em Large Language Models (LLM), que disponibiliza diversos métodos para a segmentação de texto, dos quais se destacam: 1 — Divisão de texto por tamanho fixo (com e sem sobreposição de caracteres); 2 — Segmentação de texto de acordo com o conteúdo, e 3 — Divisão semântica do texto. Considerando que a seleção de um modelo de divisão de texto é crucial para a otimização de um modelo RAG, o autor conclui que a forma de aplicação de cada uma das técnicas depende do sistema que se está a desenvolver, assim como o tipo de documentos a processar.

A segmentação de texto tem um papel relevante nas arquiteturas RAG, uma vez que existem documentos legislativos extensos que excedem a janela de contexto dos modelos de *embeddings*. As ferramentas de segmentação de texto do LangChain permitem a divisão de documentos em segmentos de texto mais facilmente processados pelos modelos de *embeddings*, sendo possível configurar a sobreposição entre segmentos de texto, assegurando que o contexto não se perde entre cada divisão.

2.4. *Embeddings*

Os *embeddings* correspondem a representações vetoriais de diferentes tipos de dados, incluindo texto, imagem e áudio, nas quais a informação é convertida em vetores numéricos num espaço multidimensional, permitindo capturar padrões e relações entre os dados representados.

No caso do texto, esta representação vetorial traduz palavras, frases ou documentos, em valores numéricos num espaço de alta dimensão. Como resultado, palavras que surgem em contextos semelhantes têm tendência a apresentar vetores próximos entre si, como acontece com termos semanticamente relacionados, p. ex. cão e gato ou maçã e laranja.

Esta codificação de palavras em *arrays* numéricos permite que os modelos de linguagem se possam aperceber das relações entre as palavras, como sinónimos ou antónimos, e até contexto.

Os *embeddings*, de uma forma geral, são criados da seguinte forma:

1 — Escolha de um modelo de *embeddings* que seja mais adequado aos dados ou tarefa que pretende realizar;
2 — Deve-se preparar os dados de forma compatível como o modelo de *embeddings* selecionado;
3 — Se usarmos um modelo pré-treinado devemos carregar os pesos e respetiva arquitetura. Se estiver a treinar um novo modelo, devemos preparar a entrada de dados para o algoritmo;
4 — Se estiver a utilizar um modelo de geração de *embeddings*, introduza uma palavra de forma a obter a sua representação vetorial;
5 — Os *embeddings* agora gerados constituem as características que serão utilizadas nos modelos de linguagem natural, para pesquisa de similaridade de texto ou até sistemas de recomendação.

2.5. Base de dados vetorial

As bases de dados vetoriais são os repositórios fundamentais para os vetores de *embeddings*, ou os *arrays* numéricos representativos da informação relevante a ser utilizada pelos LLMs. Esta estrutura permite uma busca não só pelo conceito de similaridade, através da similaridade de coseno, mas também pela reordenação de documentos e seleção de passagens para sumarização da informação selecionada, mantendo a redução da redundância e relevância da consulta original, caso seja utilizado o Maximal Marginal Relevance (MMR), Carbonell e Goldstein (1998).

Encontram-se disponíveis várias bases de dados vetoriais, apresentado diferentes inovações tecnológicas. Para soluções em código aberto, robustas e suportadas por uma base de dados relacional, existe a extensão pgvector, que permite adicionar a busca vetorial ao PostgreSQL.

O pgvector disponibiliza suporte a diferentes métodos de indexação, nomeadamente Hierarchical Navigable Small Worlds (HNSW), Malkov e Yashunin (2020), baseado numa estrutura de grafos, em que a distância ótima de similaridade entre vetores corresponde à aresta mais curta, ao mesmo tempo que assegura que cada ponto pode ser alcançado por um número mínimo de saltos, e Inverted File with Flat Compression (IVFFlat), Emanuilov e Dimov (2024), que assenta no agrupamento de vetores semelhantes.Neste último método, é utilizado um algoritmo como o *k-means* para identificar os centróides, posteriormente associa cada vetor ao seu centróide mais próximo. Após esta fase é realizado um índice invertido em que o centróide é a chave e o valor é uma lista de vetores associado a esse determinado centróide.

Temos ainda o Facebook AI Similarity Search (FAISS) em que, utilizando um Inverted File Indexing (IVF) reduz a dimensionalidade dos dados, e com a otimização para Graphical

Processing Unit (GPU) melhora em grande parte o desempenho da pesquisa. O Milvus que, utilizando o algoritmo Best Matching 25 (BM25) permite classificar os documentos que mais se adequam a determinada pesquisa, considerando um conjunto de fatores como Term Frequency (TF), Inverse Document Frequency (IDF) e normalização do tamanho do documento e, no final, baseando-se nas contribuições é atribuído um valor a um documento em relação a uma determinada pesquisa. Outras soluções disponíveis são Chroma, MongoDB Atlas ou Pinecone.

2.6. Large language models (LLM)

Os modelos de linguagem de grande escala definem-se por um conjunto de funções que recebem do utilizador um determinado texto (existem modelos mais recentes que são multimodais que recebem como entrada não só texto, mas também áudio e imagem) ou instruções (*prompt*) e, por sua vez, o modelo produz uma saída adequada a essas instruções fornecidas (resposta).

Por detrás desta definição algo simplista esconde-se um conjunto de tarefas não só complexas, mas também computacionalmente exigentes, como sejam a tokenização, ou a segmentação de texto em intervalos menores, que podem ser parte de uma palavra ou até pontuação, e a geração de *embeddings*, representações vetoriais das palavras, que permitem capturar o contexto, a relação entre os *tokens* (unidades de texto) e até a posição dos mesmos em determinada frase. Seguidamente entra uma das tecnologias mais inovadoras, a arquitetura *Transformer*. O seu mecanismo de autoatenção permite que o modelo olhe para outras representações onde a palavra se encontra representada e escolha, com base numa distribuição de probabilidade, a palavra seguinte que mais se adequa ao contexto, a previsão do próximo *token* que, após um ciclo de geração de previsões, constrói a resposta.

A informação utilizada para processamento e treino do modelo é obtida pela digitalização de livros, *webscrapping* de páginas da Internet, artigos, etc., conteúdos em diversos formatos, de modo a capturar o máximo de informação semântica, a gramática ou até estilos de escrita que, dependendo da originalidade do utilizador, podem ser utilizados para a redação de contratos de aquisição de terras ao estilo do escritor Edgal Allan Poe, Albany Law School (2023).

Em alguns modelos, como seja o Llama2, Touvron et al. (2023) ou o GPT-4, OpenAI et al. (2024), após o treino é utilizado o Reinforcement Learning from Human Feedback (RLHF) de forma a aprimorar as respostas mais alinhadas com as expectativas humanas.

2.7. Retrieval augmented generation (RAG)

Os LLMs têm se destacado na tarefa de geração de texto. Exemplos são os modelos Mistral ou Llama, que são treinados sobre conjuntos de texto sem qualquer marcação, permitindo a criação de textos literários e até código de programação. No entanto, têm se verificado comportamentos indesejáveis, os quais são identificados como alucinações, a geração de factos que não encontram paralelo em dados reais e que pode levar à disseminação de falsas respostas jurídicas, Dahl et al. (2024) a utilizadores das diversas aplicações suportadas

por sistemas de AI. Outra das limitações é a falha na atualização da informação aos dados utilizados para o treino do modelo, o que leva a que as respostas sejam corretas, mas desatualizadas em relação ao momento atual: um LLM treinado com todos os dados disponíveis na Internet relativos ao ano de 2024, não é capaz de responder quem foi o vencedor do campeonato de futebol de 2024/25 (Sporting Clube de Portugal).

A Figura 2.2 ilustra os seguintes componentes de um sistema RAG:

1 — Recolha de informação estruturada ou não estruturada;
2 — Segmentação de texto;
3 — Geração de *embeddings*;
4 — Armazenamento dos *embeddings* numa base de dados vetorial;
5 — Iteração do utilizador com o sistema;
6 — O *retriever* que pesquisa a entrada de texto;
7 — O LLM que recebe a informação recolhida pelo *retriever*;
8 — A resposta elaborada pelo LLM à entrada de texto colocada pelo utilizador.

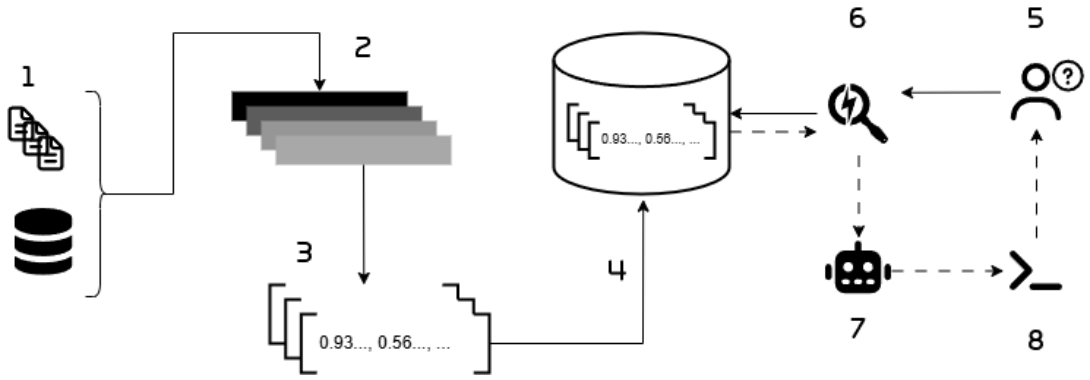


Figura 2.2. Fluxo geral de um sistema RAG (imagem do autor).

De forma a endereçar as duas problemáticas anteriormente indicadas, a arquitetura RAG, Lewis et al. (2020), propõe a utilização de um LLM em combinação com um mecanismo externo dinâmico de recuperação da informação, como p. ex. bases de dados vetoriais, motores de busca ou repositórios institucionais, permitindo o acesso a conhecimento atualizado e ou de um domínio específico. Os resultados obtidos demonstram a eficácia na geração de texto mais factual e específico, sem a necessidade da execução de um novo processo de treino do modelo e, superando as limitações do conhecimento pré-existente e da falsa factualidade apresentada e, representando um avanço significativo na AI e NLP.

Existem ainda desafios na implementação deste tipo de sistema plasmados em Gu (2025), nomeadamente escalabilidade diretamente relacionada com o aumento do volume da informação disponível, a formulação variedade semântica das perguntas realizadas ao LLM, em que uma pergunta pode que pode influenciar a qualidade da resposta, e a latência em que para o objetivo de uma maior qualidade na resposta, tendo como compromisso assumido um maior tempo de geração da mesma.

2.8. Engenharia de *prompts*

A engenharia de *prompts* é um processo que tem como foco a elaboração e otimização de instruções eficazes para LLMs. Tem como objetivos orientar o modelo na obtenção de uma resposta precisa, minimizando as respostas erróneas ou imprecisas (as alucinações), e potenciando a relevância, a exatidão e o formato da resposta apresentada ao utilizador. Uma engenharia de *prompts* efetiva requer uma compreensão da componente técnica dos LLMs, assim como as especificidades da intenção do utilizador.

Em Brown et al. (2020), onde foi publicitado o modelo GPT-3, são apresentados os conceitos de *zero-shot* e *few-shot prompting*, existindo outras técnicas como *role-based prompting*, Kong et al. (2024) ou *chain-of-thought prompting*, Wei et al. (2023).

CAPíTULO 3

Revisão de literatura

A metodologia utilizada para a realização da revisão da literatura assenta na "Declaração PRISMA 2020: uma Diretriz Atualizada para Publicação de Revisões Sistemáticas", Page et al. (2024), que permite a transparência, a qualidade e a reprodutibilidade dos passos realizados, permitindo aos leitores a análise crítica e a inspeção dos passos realizados.

A presente revisão de literatura assentou na exploração do uso de NLP e demais avanços da GenAI na área jurídica, que se encontram disponíveis na base de dados bibliográfica Scopus[1].

3.1. Critérios de busca

Como referência tecnológica tomou-se a arquitetura *Transformer*, que provocou uma revolução no campo do NLP ao introduzir o mecanismo de atenção, tornando-se a base de vários modelos de linguagem como o BERT, Devlin et al. (2019), ou o GPT, Radford et al. (2018). Este mecanismo de atenção é fulcral para a eficiência da arquitetura *Transformer*, já que permite que o modelo avalie a relevância de cada palavra numa sequência em relação às demais, independentemente da posição em que se encontra.

Com o lançamento do serviço de inteligência artificial generativa, o ChatGPT, pela OpenAI, em 2022, e a sua ampla adoção — quer pelo público em atividades criativas ou quotidianas, quer pelas empresas na automação de serviços ou no atendimento ao cliente —, este evento tornou-se um marco temporal na evolução dos LLMs, fomentando a produção científica e o desenvolvimento de aplicações baseadas em GenAI.

Considerando o curto espaço temporal definido, 2022 a 2024, observamos que os documentos selecionados possuem um reduzido número de citações, que são reflexo da sua novidade. No entanto, destacamos que esse facto não compromete de forma alguma a sua relevância, uma vez que o trabalho plasmado nesses documentos representa os mais recentes avanços científicos, que ainda não estão consolidados, na respetiva área de estudo. A sua inclusão nesta revisão permite uma visão atualizada com as tendências da área, permitindo que se analise a sua evolução e impacto futuro.

3.2. Pesquisa de documentos

A pesquisa realizada na base de dados Scopus possui a seguinte composição:

> (TITLE-ABS-KEY ("legislation" OR "jurisprudence" OR "legal documents") AND TITLE-ABS-KEY ("text simplification" OR "question answering" OR "information extraction" OR

[1] http://www.scopus.com.

"chatbot" OR "dialogue systems" OR "summarisation") AND TITLE-ABS-KEY ("transformer" OR "GPT" OR "generative artificial intelligence")) AND PUBYEAR > 2022 AND PUBYEAR < 2026

Consistindo no agrupamento dos seguintes itens:

1 — O domínio temático dentro do contexto jurídico, "TITLE-ABS-KEY ("legislation" OR "jurisprudence" OR "legal documents")".

2 — As áreas de aplicação da AI e do NLP, "TITLE-ABS-KEY ("text simplification" OR "question answering" OR "information extraction" OR "chatbot" OR "dialogue systems" OR "summarisation")".

3 — As tecnologias na área da AI, focadas no processamento de linguagem natural, "TITLE-ABS-KEY ("transformer" OR "GPT" OR "generative artificial intelligence")".

4 — O período temporal definido, "PUBYEAR > 2022 AND PUBYEAR < 2026".

Para a organização das diferentes fases do processo de recolha, análise e triagem de documentos para este trabalho foi utilizado o *software* de recolha e gestão de documentos científicos Zotero.

A Figura 3.1 mostra as diferentes fases do processo de recolha e documentação do processo de revisão de literatura.

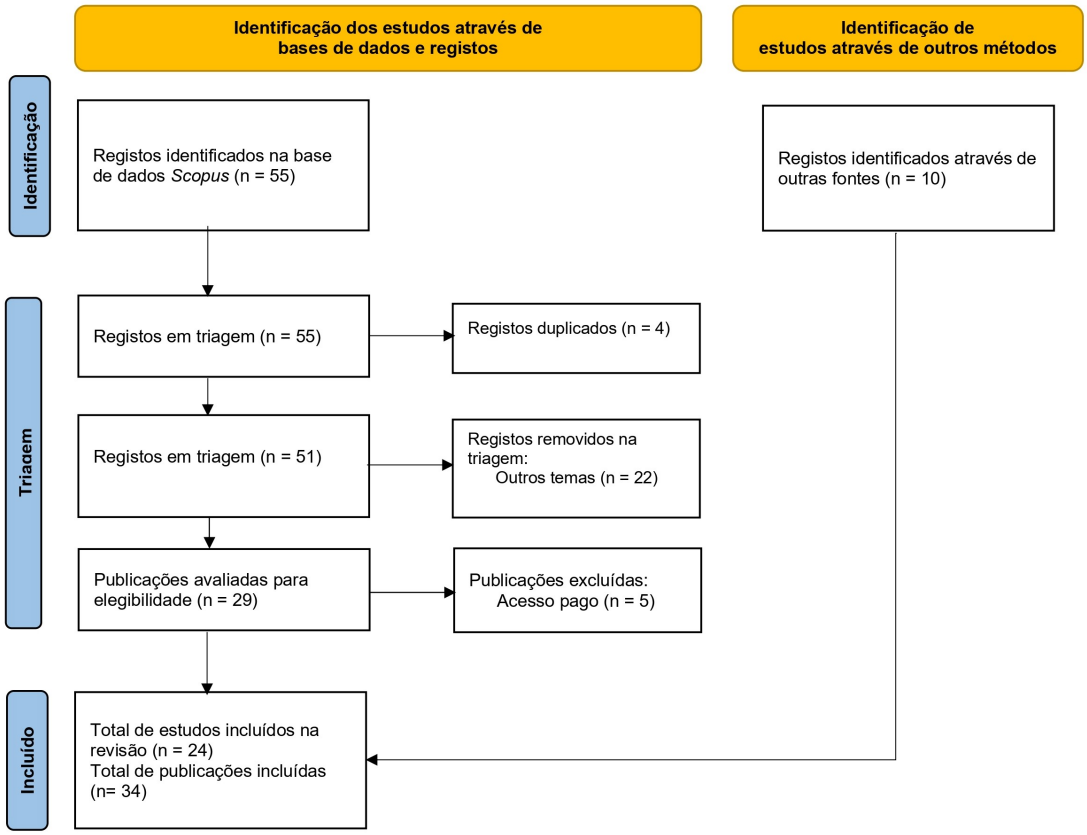


Figura 3.1. Diagrama PRISMA 2020

Adaptado pelo autor a partir da Declaração PRISMA 2020 (ESS Jean Piaget - Vila Nova de Gaia - Portugal).

3.3. Trabalho relacionado

As descobertas alcançadas no decorrer da revisão da literatura foram organizadas em três categorias: 1 — A secção 3.4 apresenta uma análise abrangente dos resultados da pesquisa; 2 — A secção 3.5 diz respeito às metodologias aplicadas à simplificação e síntese de textos jurídicos, e 3 — A secção 3.6 aborda um comparativo de modelos e algoritmos empregues no domínio jurídico.

Ao resumir as evidências recolhidas e realizar uma análise crítica das metodologias utilizadas, esta secção tem como objetivo apresentar uma visão geral do estado da arte na simplificação de textos jurídicos.

3.4. Análise global

Dos 24 artigos científicos resultantes da pesquisa na base de dados Scopus, conforme descrito na secção 3.2, verificou-se uma predominância de de artigos de conferência, 21, e somente 3 artigos publicados em revistas científicas (*journals*). Os artigos de conferências refletem os trabalhos inovadores e em desenvolvimento na área jurídica, podendo apresentar novas descobertas, melhores práticas ou novas metodologias na utilização da AI no contexto jurídico.

A lista completa dos artigos científicos identificados, todos em acesso aberto mediante a utilização da Virtual Private Network (VPN) do Iscte-IUL, encontra-se disponível no Anexo A.

3.5. Metodologias de simplificação e síntese jurídica

Na análise realizada por Deepika et al. (2024) são mencionadas várias técnicas utilizadas para a extração e sumarização de conteúdo, como sejam o modelo Pegasus, baseado na tecnologia *Transformer*, em que o pré-treino realizado permite uma melhoria significativa dos resultados. Sobre a mesma tecnologia, são testados os modelos BERT e Bidirectional and Autoregressive Transformer (BART), Lewis et al. (2019) e o modelo baseado em grafos, TextRank, utilizado para processamento de texto que acabou por obter bons resultados face aos modelos mais evoluídos para a tarefa específica de extração de conhecimento.

No Supremo Tribunal de Justiça (STJ) português, a sumarização de documentação legal é realizada manualmente pelos juízes, o que torna a tarefa morosa. Para a agilização desta tarefa, Dias, Ribeiro e Pinto (2024), aplicaram diferentes técnicas de sumarização como o LexRank, que se mostrou útil na classificação de frases, o SeburnSum para definir o melhor resumo e o mBART para gerar um sumário mais coerente e fluente já que no seu desenvolvimento, pela Meta AI, foi adaptada a arquitetura BART para cenário multilingue.

Ao abordar os desafios colocados aos idiomas com poucos recursos no cenário da GenAI, a aplicação da aprendizagem por transferência na simplificação de textos jurídicos russos, Athugodage, Mitrofanove e Gudkov (2024), surge como um avanço no sentido de expandir o alcance e a acessibilidade de tecnologias avançadas de NLP, indo além de idiomas amplamente estudados, como o inglês. Nesta sua investigação, devido aos modelos Text-to-Text Transfer Transformer (T5) e GPT terem um desempenho insatisfatório na

tarefa de sumarização de textos em língua russa, os mesmos foram objeto de um *fine-tuning*, originando os modelos T5-RLS2000, GPT-simplifier-large-text, e GPT-simplifier25, baseados num *corpus* de documentos legais russos. A investigação apresentada abre caminho para que o método apresentado possa ser aplicado em outras línguas, disponibilizando a leitura de documentos legais a uma maior audiência.

O desenvolvimento de novas metodologias aplicadas à geração de conjuntos de dados elevada qualidade tem sido objeto de investigação recente. Zmiycharov, Tsonkov e Koychev (2024), elaboraram um conjunto de dados de legislação europeia em 24 línguas, recorrendo a uma abordagem em dois passos recorrendo ao TextRank e ao GPT, a qual demonstrou potencial para se tornar uma técnica eficaz de sumarização automática.

As metodologias e os avanços descritos ao longo desta subsecção, revelam os esforços da comunidade científica em superar os obstáculos com que se deparam no processo de simplificação e sumarização de textos jurídicos. Constatamos que a adaptação de modelos existentes, como o T5-RLS2000 ou GPT-simplifier-large-text, ou pela criação de novos conjuntos de dados, como o EurLexSummarization, demonstram o interesse de proporcionar uma abertura ao conhecimento dos documentos legais. Estas iniciativas permitem redefinir os limites da utilização da GenAI no contexto de simplificação e síntese jurídica, orientando a utilização futura do NLP para uma maior eficácia e eficiência de processos e uma democratização no acesso ao conhecimento jurídico.

3.6. Modelos mais utilizados para simplificação e síntese jurídica

Com base nos documentos da revisão de literatura, elaborou-se um *top* cinco dos modelos mais utilizados para as tarefas de simplificação e síntese jurídica — Tabela 3.1.

Modelo	Número de referências	Principais aplicações
BERT	6	Classificação de textos, sumarização jurídica
BART	4	Sumarização legislativa, geração de texto simplificado
Pegasus	3	Sumarização abstrativa de documentos jurídicos
GPT	3	Sumarização
Legal Pegasus	3	Tradução e simplificação de textos jurídicos

Tabela 3.1. Modelos utilizados na área jurídica

Como critério de ordenação contabilizou-se o número de vezes que cada modelo foi referenciado como base tecnológica no desenvolvimento da investigação, destacando a sua utilização em aplicações como simplificação, sumarização ou geração de texto, conforme indicado nos estudos de referência.

O modelo BART foi utilizado na tarefa de sumarização extrativa de texto em "Abstractive Text Summarization: A Transformer Based Approach", uma abordagem semelhante à sumarização humana, que permite reter o contexto, ou a ideia, de um documento fonte

e gerar um resultado que, potencialmente, contém novas frases que poderão não estar presentes na fonte.

3.7. Outros estudos de referência

A aplicação da tecnologia *Transformer* permitiu a inclusão da AI na área jurídica, com progressos nas seguintes áreas: previsão de sentenças — Legal Judgment Prediction (LPJ), sumarização e classificação de documentos jurídicos, perguntas e respostas sobre o contexto jurídico e RAG.

Destaca-se a aplicação na *framework* Multi-Stage Judgement (MSJudge), Ma et al. (2021), na área de previsão de sentenças judiciais, baseado em factos criminais. No desenvolvimento desta *framework* foi necessária a criação de um novo conjunto de dados representativo do "ciclo de vida" de um caso, utilizados para pré-treino de documentos extensos, ao qual se aplicou um modelo de aprendizagem multitarefa muito comum nas tarefas de processamento de linguagem natural, mencionando como exemplos de sucesso os trabalhos realizados por Guo, Pasunuru e Bansal (2018), e Nishida et al. (2019). Menciona-se o LegalDuet, Xu et al. (2025), um modelo de pré-treino contínuo elaborado para potenciar os resultados de modelos pré-treinados de linguagem natural como sejam o BERT-xs, Zhong et al. (2019) e o BERT-Chinese, Devlin et al. (2019), na deteção de factos criminais. Esta melhoria é conseguida pela adoção de um mecanismo de aprendizagem dual baseado na divisão entre Law Case Clustering (LCC) que, ao agrupar casos similares, ajuda o modelo de linguagem natural na identificação de subtis diferenças entre os mesmos, e o Legal Decision Matching (LDM) que permite o alinhamento dos factos criminais com as respetivas decisões.

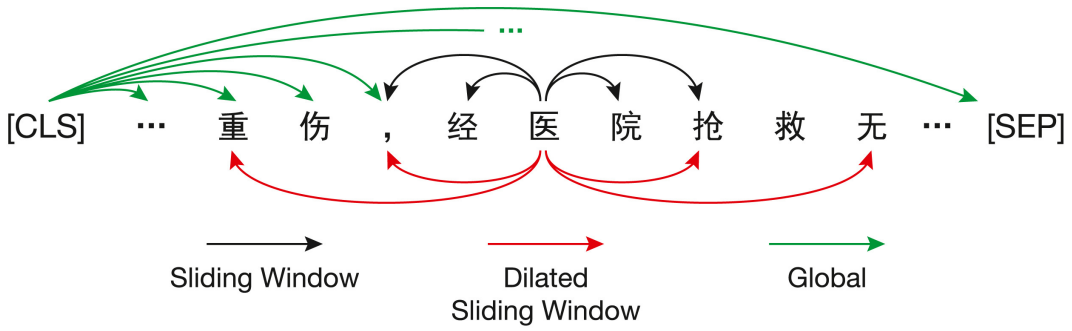


Figura 3.2. Lawformer: combinação dos três tipos de mecanismo de atenção — Xiao et al., 2021.

Realça-se o trabalho realizado com o Lawformer, Xiao et al. (2021) — Figura 3.2 —, na criação de um modelo para a compreensão de documentos legais extensos, utilizando como base o codificador Longformer — Beltagy, Peters e Cohan (2020) —, introduzindo três tipos de mecanismos de atenção para processamento de sequências de texto longas: 1 — *sliding window attention*: em que somente é realizado um cálculo das pontuações de atenção entre os *tokens* mais próximos; 2 — *dilated sliding window attention*: a janela de

contexto pode ser dilatada de modo a abarcar um texto mais longo, e 3 — *global attention*: utilizada nos específicos em que se pretende obter a totalidade da sequência de forma a maximizar a informação recolhida.

Considerando que grande parte dos documentos jurídicos é de grande complexidade — com termos específicos, documentos extensos e com inúmeras citações de legislação e de decisões —, exigindo dos modelos uma garantia de fidelidade e coesão ao conteúdo transmitido. Nesse sentido, a compreensão do texto legal de acordo com o enquadramento legislativo é fulcral, tendo Long et al. (2018) apresentado um novo modelo, o AutoJudge, que consegue captar as complexas relações entre argumentos, factos e leis.

Os documentos legais carregam consigo um grau de complexidade elevado, muito devido à linguagem formal utilizada, à sua extensão, à utilização de remissões encadeadas, constituindo um obstáculo à necessária clareza e sumarização das ideias principais da letra da lei perante o cidadão. De modo a inverter esta situação, destacam-se os trabalhos realizados na sumarização de documentos legislativos europeus por Zmiycharov, Tsonkov e Koychev (2024), que recorrendo à criação de um conjunto de dados recolhido dos sumários da legislação europeia — responsabilidade do Serviço de Publicações da União Europeia —, devidamente ajustado para o requisito funcional da sumarização de textos, e a um modelo T5 devidamente afinado para a tarefa de sumarização conseguiu melhores resultados que um modelo mais específico, o LEGAL-BERT, Chalkidis et al. (2020). Os avanços na área jurídica têm-se inspirado na área da investigação operacional, através de abordagens não supervisionadas de sumarização extrativa de casos jurídicos. Exemplo disso é a criação do algoritmo DELSumm, Bhattacharya et al. (2021), desenvolvido com base um conjunto de dados de casos indianos — abrangendo questões civis como criminais, incluindo ordens judiciais e sentenças. O modelo foi aprimorado com contributos de especialistas da área jurídica, que ajudaram a definir os parâmetros da função objetivo, permitindo obter melhores resultados que alguns métodos supervisionados treinados em grande volumes de documentos, como sejam o SummaRuNNer, Nallapati, Zhai e Zhou (2016), o BERTSUM, Liu (2019) ou Gist, Liu e Chen (2019) que, sob determinadas condições, os modelos extrativos ainda se revelam competitivos.

CAPíTULO 4

LexClara: sistema de pergunta e resposta sobre diplomas legais

O LexClara foi desenvolvido para lidar com um desafio: simplificar a terminologia técnica e estrutural da legislação, que dificulta a compreensão por parte do cidadão leigo.

A metodologia apresentada nesta dissertação encontra-se alinhada com esse objetivo, propondo a simplificação da linguagem de diplomas legislativos do ordenamento jurídico português através da resposta a quatro perguntas, garantido que em cada fase do processo a integridade semântica e o contexto são preservados.

Neste trabalho, a *framework* LangChain é utilizado na orquestração do sistema LexClara, permitindo a integração dos diversos componentes necessários ao funcionamento de um sistema RAG. Através desta *framework* em código aberto é possível articular o processamento de documentos legislativos, a recuperação de informação e a geração de respostas pelo LLM.

A *framework* LangChain suporta as seguintes funcionalidades:

- Segmentadores de texto;
- Processadores de documentos;
- *Prompts*;
- Modelos de linguagem;
- Bases de dados vetoriais;
- Cadeias, e
- Agentes.

Na implementação do LexClara, complementando o LangChain, são utilizados modelos disponibilizados na plataforma Hugging Face. Recorreu-se a um LLM pré-treinado para a tarefa de pergunta-resposta, a um modelo de geração de *embeddings* e também à biblioteca *Transformers*, suportando diferentes fases de um sistema RAG.

4.1. O *pipeline* LexClara

O *pipeline* implementado para o LexClara segue as etapas normais de um processo ETL, seguido de uma segmentação de texto (*chunking*), geração das representações vetoriais do texto (*embeddings*), indexação a uma base de dados vetorial e, por fim, a integração com um sistema de RAG. Este alinhamento garante que os dados se encontram normalizados, organizados e devidamente otimizados para a recuperação contextual e geração de texto mais acessível pelo LLM.

4.2. *Extraction, Transform and Load* (ETL)

Esta secção descreve o processo de extração dos diplomas legais que compõem o *corpus* deste trabalho, constituindo a base de conhecimento sobre a qual assenta o sistema de

processamento e recuperação de informação RAG, LexClara. São identificadas as fontes de dados utilizadas, assim como o detalhe da extração automática. É, também, analisada a estrutura de documentos e os principais tipos de diploma — decreto-lei e decreto regulamentar —, de forma a assegurar uma segmentação textual coerente, fundamental para as etapas posteriores de transformação e indexação vetorial.

4.2.1. Fontes de dados

A validação da origem dos dados é uma tarefa crucial para um projeto de ciência de dados. Nesse sentido, face à entidade editora do *Diário da República*, a INCM, não disponibilizar um histórico dos diplomas publicados, recorreu-se a uma fonte de informação secundária, o DRE Tretas para a obtenção dos documentos que serviram para a constituição do *corpus* utilizado neste trabalho.

Os dados da legislação portuguesa publicados desde 1910, recolhidos com periodicidade semanal, encontram-se disponíveis em diversos formatos[1]: SQLite (utilizado neste trabalho), JSON e um *dump* PostgresSQL.

Para este trabalho, recorreu-se à técnica de *webscrapping* para extrair do sítio oficial do *Diário da República*, o conteúdo da linguagem natural dos diplomas legais. Este conteúdo serve como conjunto verdade para a avaliação das respostas geradas pelo LexClara, permitindo comparar o desempenho do modelo com a informação elaborada por juristas. A GenAI é utilizada para simplificar a linguagem destes documentos, tornando-os acessíveis ao cidadão leigo.

4.2.2. Extração automática

O processo de extração de dados processa-se de forma automática, desde a ligação ao sítio DRE Tretas, a partir do qual é extraída a versão mais recente das publicações do *Diário da República*, até ao processo de descompactação dos dados.

Para este trabalho, a extração dos dados foi definida com base na data de publicação dos diplomas entre 2020 e 2023. Esta delimitação temporal deveu-se a uma questão de uniformização dos Uniform Resource Locator (URL) dos documentos publicados em *Diário da República*, tendo sido desenvolvidos os seguintes conjuntos de dados:

1 — Extração dos URL dos diários e suplementos, tendo como objetivo a construção do conjunto de dados dos conteúdos da linguagem clara, publicados no sítio do *Diário da República*;

2 — Criação do *corpus* da legislação, correspondente ao conteúdo dos diplomas tipificados como decreto-lei ou decreto regulamentar;

3 — Conteúdo da linguagem clara, que irá servir de conjunto verdade para avaliar a resposta do LLM.

4.3. Transformação e limpeza dos dados

A seguinte secção descreve o processo de transformação e limpeza do conteúdo textual, etapa essencial para garantir a consistência e legibilidade do *corpus* legislativo. São abordadas

[1] Hiperligação — https://uploads.tretas.org/?P=*DRE*.bz2

as técnicas de normalização do conteúdo textual, a análise à estrutura dos documentos e a estratégia se segmentação de textos implementada para manter a coerência entre os textos jurídicos. Aborda-se com detalhe a análise de *tokens* e da distinção entre segmentos de texto breves e extensos, assim como a descrição do processo de mapeamento-redução, que otimiza a reagregação dos conteúdos dos segmentos extensos.

4.3.1. Normalização do texto

O conteúdo do conjunto de dados, originalmente em formato HyperText Markup Language (HTML), foi extraído e convertido para texto através da utilização da biblioteca Selenium, permitindo a análise estatística — como a contagem de palavras e caracteres — e semântica, através de medidas como o Term Frequency-Inverse Document Frequency (TF-IDF) e o Índice de Legibilidade de Flesch.

Considerando que os LLM são treinados sobre grandes conjuntos de dados estruturados e não estruturados e, por esse motivo, possuem uma diversidade sintática e uma riqueza de vocabulário, a normalização do texto deve preservar essas características, garantindo que informações semânticas relevantes se mantenham e que o sistema RAG continue a gerar respostas precisas.

4.3.2. Análise dos documentos

Foram analisados 682 diplomas segundo duas abordagens. Uma qualitativa, que foca na complexidade dos documentos, utilizando-se uma adaptação do índice de legibilidade de Flesch (Flesch, 1948), para a língua portuguesa do Brasil baseada na fórmula apresentada em Moreno et al., 2022[2]:

$$226 - 1.04 \times \left(\frac{\text{n.º palavras}}{\text{Qnt. de frases}} \right) - 72 \times \left(\frac{\text{Qnt. de sílabas}}{\text{Qnt. de palavras}} \right)$$

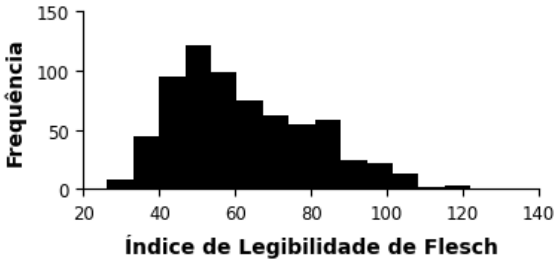
O índice de legibilidade de Flesch calcula a legibilidade média de cada parágrafo do conjunto de dados, permitindo aferir a complexidade dos documentos jurídicos do ponto de vista do cidadão leigo.

Na Tabela 4.1, apresenta-se a escala do índice de legibilidade de Flesch.

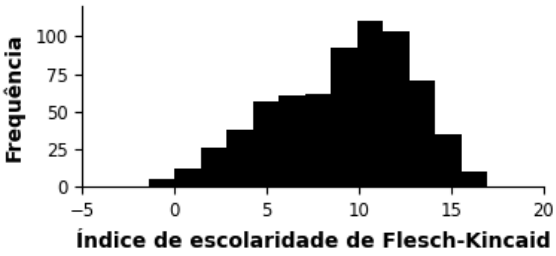
Valores	Descrição
100 - 90	Muito fácil de ler. Fácil de entender para um aluno médio de 11 anos.
90 – 80	Fácil de ler. Português coloquial.
80 – 70	Bastante fácil de ler.
70 – 60	Português claro. Fácil de entender para alunos de 13 a 15 anos.
60 – 50	Bastante difícil de ler.
50 – 30	Difícil de ler.
30 – 10	Muito difícil de ler. Melhor compreendido por universitários.
10 – 0	Extremamente difícil de ler. Melhor compreendido por universitários.

Tabela 4.1. Índice de legibilidade de Flesch

[2] Hiperligação — https://legibilidade.com/sobre



(a) Índice legibilidade de Flesch.



(b) Índice escolaridade Flesch-Kincaid.

Figura 4.1. Índices de legibilidade e escolaridade aplicado aos documentos jurídicos.

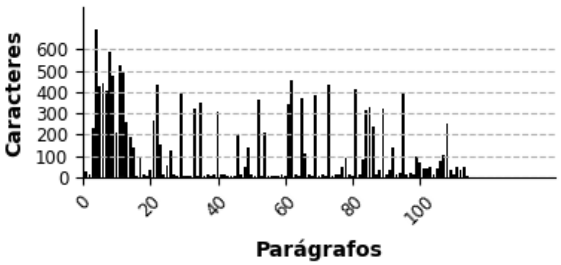
Verifica-se na Figura 4.1a, que os valores de médios de legibilidade dos parágrafos encontram-se no intervalo de 50 a 60, indicando que a legibilidade é complexa, sendo melhor compreendidos por cidadãos com maiores níveis de escolaridade. Esta observação é confirmada pela distribuição apresentada na Figura 4.1b, correspondente ao Índice de escolaridade de Flesch-Kincaid, o qual estima o nível de escolaridade necessário para a compreensão do texto, combinando o comprimento médio dos parágrafos com a complexidade das palavras. Verifica-se que o pico da distribuição se situa no nível 11, estendendo-se até ao nível 15, significando que os textos analisados exigem, na sua maioria, um nível de escolaridade equivalente ao secundário e, em alguns casos, ao ensino superior, o que implica um esforço acrescido de interpretação por parte do leitor.

É esta complexidade semântica que o sistema LexClara procura mitigar, recorrendo à GenAI de modo a simplificar o discurso jurídico, sem comprometer a exatidão e rigor legislativo.

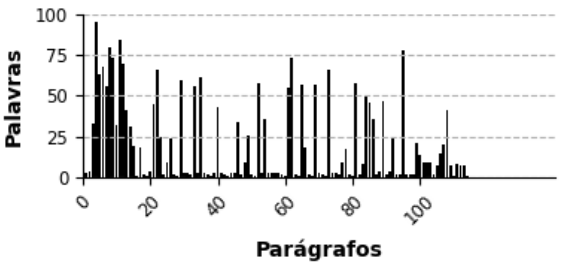
No cálculo do número de classes do histograma foi utilizada a regra de Doane. A escolha deveu-se a dois fatores: 1 — A amostra tem uma assimetria > 0 (0,59) (cauda à direita), e 2 — Possui um melhor comportamento em relação à regra de Sturges (da qual é uma evolução), ao introduzir um fator de correção que tem em conta a assimetria, podendo apresentar um maior número de classes.

A abordagem quantitativa baseou-se em técnicas estatísticas como o TF-IDF e o recurso à contagem do número de palavras em cada parágrafo de um determinado diploma.

A contagem de palavras por parágrafo foi realizada de forma a poder-se ter uma visão da distribuição do número de palavras ao longo de um diploma:



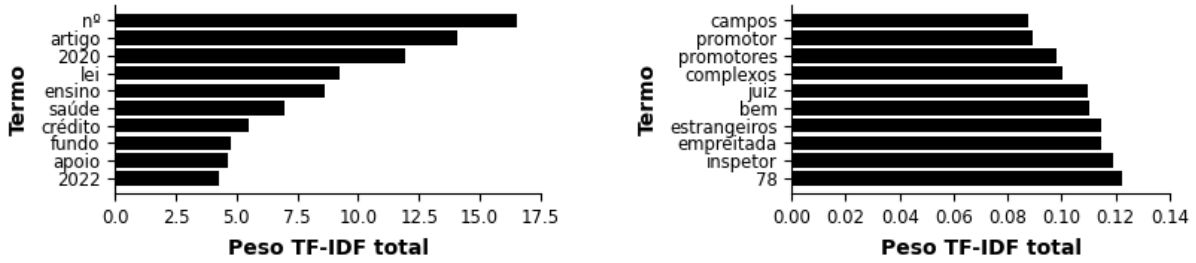
(a) Distribuição de caracteres.



(b) Distribuição de palavras.

Figura 4.2. Distribuição de caracteres e palavras pelo Decreto-Lei n.º 139-D/2023.

O objetivo desta análise passa por identificar e quantificar a existência de padrões de densidade da informação ao longo de um documento, servindo de base para a possível aplicação de um outro método de segmentação de texto.



(a) Termos mais representativos no *corpus*.

(b) Termos menos representativos no *corpus*.

Figura 4.3. Representatividade dos termos no *corpus*.

O TF-IDF surgiu no contexto de recuperação da informação, evoluindo através dos trabalhos realizados por Jones, 2004, em que introduziu o conceito de "especificidade do termo", em que o termo mais relevante deveria ser o que é referenciado o menor número de vezes em toda a coleção, Salton e Buckley, 1988, em que colocaram à prova diversas técnicas de ponderação de termos, inclusive o IDF, e demonstraram a sua eficácia.

O cálculo da soma global dos pesos TF-IDF de todos os documentos tem como objetivo identificar os termos mais relevantes no *corpus* em estudo. Os resultados evidenciam a importância que certos termos na temática nos documentos analisados.

4.3.3. Estrutura dos documentos

A análise à estrutura dos documentos foi realizada com base na Lei n.º 74/98, de 11 de novembro[3], no que diz respeito à tipologia, e às divisões constantes nos diplomas: 1 — Capítulos. 2 — Artigos. 3 — Alíneas, e 4 — Subalíneas.

Decreto-lei

No artigo 12.º podemos encontrar a estrutura que orienta a publicação dos diplomas, do tipo decreto-lei, no *Diário da República*:

> "Artigo 12.º
> **Diplomas legislativos do Governo**
> 1 — Os decretos-leis obedecem ao formulário seguinte:
> *a*) Decretos-leis previstos na alínea *a*) do n.º 1 do artigo 198.º da Constituição:
> «Nos termos da alínea *a*) do n.º 1 do artigo 198.º da Constituição, o Governo decreta o seguinte: (Segue-se o texto.)»
> *b*) Decretos-leis previstos na alínea *b*) do n.º 1 do artigo 198.º da Constituição:
> «No uso da autorização legislativa concedida pelo artigo ... da Lei n.º .../..., de ... de ..., e nos termos da alínea *b*) do n.º 1 do artigo 198.º da Constituição, o Governo decreta o seguinte: (Segue-se o texto.)» *d*) Decretos-leis previstos no n.º 2 do artigo 198.º da Constituição:
> «Nos termos do disposto no n.º 2 do artigo 198.º da Constituição, o Governo decreta o seguinte: (Segue-se o texto.)»

[3] Hiperligação — https://diariodarepublica.pt/dr/legislacao-consolidada/lei/1998-34448175

2 —- Após o texto seguem-se, sucessivamente, a data de aprovação em Conselho de Ministros, a assinatura do Primeiro-Ministro e dos ministros competentes, a data da promulgação, a ordem de publicação e a assinatura do Presidente da República, a data da referenda e a assinatura do Primeiro-Ministro. [. . .]"

Decreto regulamentar

No mesmo diploma podemos encontrar, no seu artigo 14.º a seguinte construção relativa ao diploma decreto regulamentar:

"Artigo 14.º
Outros diplomas do Governo
1 — Os outros diplomas do Governo obedecem ao formulário seguinte:
a) Decretos regulamentares:
«Nos termos da alínea *c*) do artigo 199.º da Constituição e . . .
(segue-se a identificação do ato legislativo a regulamentar), o Governo decreta o seguinte:
(Segue-se o texto.)»
«Nos termos da alínea *g*) do artigo 199.º da Constituição, o Governo decreta o seguinte:
(Segue-se o texto.)» [. . .]"

4.4. Estratégia de segmentação de texto

Após análise aos documentos presentes no conjunto de dados, decidiu-se pela segmentação de textos entre "preambulo" e "capitulo".

Ao preâmbulo corresponde o texto do diploma que antecede o capítulo, em documentos que possuam esta divisão ou, todos os documentos que não possuem capítulos. Para a divisão por capítulos, considerou-se que um conceito, ou ideia jurídica, estaria encapsulado em cada capítulo o que capacita a busca realizada à base de dados de acordo com a similaridade do conceito passado na pergunta.

No âmbito do sistema LexClara, esta divisão de texto pode dar origem a segmentos demasiado longos para o modelo de *embeddings*, cuja janela de contexto define a extensão máxima a ser processada. Estes modelos convertem a linguagem natural em representações vetoriais passíveis de compreensão pela máquina, permitindo que o sistema retenha a informação semântica essencial durante a normalização e análise de documentos jurídicos.

4.5. Análise de *tokens*

Previamente à analise da contabilização de *tokens* deve-se ter conhecimento das definições de "dimensão" e "sequência", duas das características mais importantes de um modelo de geração de *embeddings* que, em conjunto, definem a forma como o modelo processa e representa o texto.

A dimensão, p. ex., 1024, refere-se ao tamanho do vetor de saída. É a forma final da representação do texto num formato vetorial. Quando um modelo de *embeddings* processa uma palavra como "decreto-lei", não a devolve como texto, mas convertido numa representação numérica de alta dimensão, cuja dimensionalidade é definida pelo próprio modelo.

A sequência, por exemplo 8192, representa o limite máximo de *tokens* que o modelo consegue processar numa única vez, possuindo outras designações como "contexto" ou *max_seq_length*. Os modelos de *embeddings* possuem um limite máximo de *tokens* (unidades de palavra) que conseguem processar de uma única vez. Se se tentar processar um texto com 10 000 *tokens*, o modelo irá ignorar todo o conteúdo após os 8192 *tokens*. Este valor é importante, pois determina quão longos podem ser os documentos que o modelo consegue processar sem perder informação.

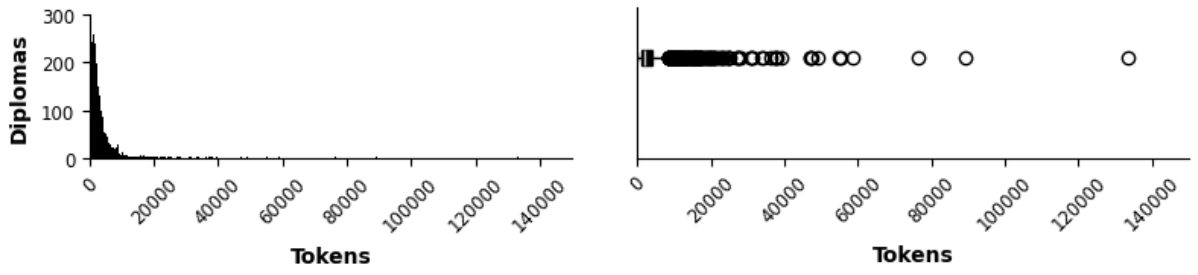
Verifica-se que 202 segmentos de texto gerados pela divisão em "preambulo" e "capitulo", 9,6 % do total de 2112 segmentos de texto, ultrapassavam os 8000 *tokens* da janela de contexto do gerador de *embeddings*.

Apresenta-se na Tabela 4.2 os cinco diplomas com maior número de *tokens* que, em conjunto, representam cerca de 92 segmentos de texto.

Diploma	**Número de *tokens***
Decreto-Lei n.º 9/2021, de 29 de janeiro	133 373
Decreto-Lei n.º 93/2020, de 3 de novembro	89 349
Decreto-Lei n.º 23/2021, de 23 de março	76 611
Decreto-Lei n.º 102-B/2020, de 9 de dezembro	58 522
Decreto-Lei n.º 15/2022, de 14 de janeiro	55 261

Tabela 4.2. Cinco diplomas com os segmentos de texto mais extensos.

A análise de *tokens* apresenta um histograma, Figura 4.4a, com a distribuição numérica de *tokens* por diploma, ao qual foi aplicado a regra de Freedman–Diaconis (Freedman e Diaconis, 1981) para o cálculo da largura dos intervalos de classes (*bins*). Esta regra é menos sensível aos valores extremos (*outliers*), já que é baseada na Interquartil Range (IQR), que é a diferença entre os terceiro quartil (Q_3) e o primeiro quartil (Q_1).



(a) Histograma. **(b)** Diagrama de caixa.

Figura 4.4. Histograma e diagrama de caixa dos *tokens* dos diplomas.

No histograma, Figura 4.4a, constata-se um enviesamento à direita, indicador de uma maior frequência de valores até aos 5000 *tokens*, permitindo o processamento direto de *embeddings* dos segmentos de texto resultantes da aplicação da estratégia de divisão inicialmente definida [por preâmbulo e capítulo(s)].

Na análise ao diagrama de caixa, Figura 4.4b, verifica-se que existiam no conjunto de dados segmentos de texto com até um máximo de 133 373 *tokens* e um elevado número de

valores extremos. Surgiu, então, a necessidade de aprofundar a análise estatística realizada aos *tokens* de forma a elaborar estratégia para ultrapassar o impacto na utilização do modelo gerador de *embeddings*, considerando a diferença do número de *tokens* entre a capacidade de sequenciação do modelo e o número de *tokens* gerados em cada segmento de texto.

Complementa-se esta análise com um resumo dos cinco números, Tabela 4.3, aos quais foram adicionados os valores da contagem, da média e do desvio padrão:

Estatística	**Valor**
Mínimo	27,00
Primeiro quartil (Q_1: 25%)	1065,25
Mediana (Q_2: 50%)	2088,50
Terceiro quartil (Q_3: 75%)	4012,50
Máximo	133 373,00
Contagem	2114,00
Média (μ)	3691,44
Desvio padrão (σ)	6320,43

Tabela 4.3. Estatística descritiva de *tokens*. *Nota.* — Valores apresentados com duas casas decimais significativas.

Como resultado desta análise aos documentos foi elaborada uma estratégia que permitiu tratar de diplomas, cujos os seus segmentos de texto não ultrapassavam os 8000 *tokens*, e os diplomas em que, pelo menos um dos seus segmentos de texto ultrapassavam o valor limite definido anteriormente.

A estratégia desenvolvida assentou na utilização de um processo de mapeamento e redução, concebido para o tratamento dos segmentos de texto extensos. A aplicação deste novo processo originou três conjuntos de documentos, cada um destinado a finalidades específicas: 1 — diplomas de menor extensão ($\leq$ 8000 *tokens*), utilizados diretamente para a geração de *embeddings*; 2 — diplomas que apresentam segmentos de texto extensos, selecionados para tratamento na fase de mapeamento – neste âmbito foram incluídos os cinco diplomas mencionados na Tabela 4.2; e 3 — um conjunto de dados adicional, constituído pelos diplomas que combinam segmentos extensos e breves, sendo estes últimos tratados de forma semelhante aos diplomas compostos apenas por segmentos breves.

4.6. Segmentos de texto breves

A definição de segmentos de texto breves é aplicada aos diplomas cujos textos não ultrapassam os 8000 *tokens*, correspondendo à janela de contexto do processamento do gerador de *embeddings* definido no âmbito deste trabalho.

Para estes segmentos o processo de gravação na base de dados vetorial é linear, já que a geração de *embeddings* e gravação ocorre num passo, uma vez que já se encontram previamente registados num conjunto de dados.

O conjunto de dados dos documentos breves é constituído por 557 diplomas, perfazendo 1250 segmentos de texto breves, ou seja, 81,67 % dos diplomas selecionados para este trabalho e 59,19 % dos segmentos de texto gerados.

4.7. Segmentos de texto extensos

O conceito de segmentos de texto extensos continua a ser baseado no número de *tokens* mas, desta vez, aplicado a diplomas que possuam, pelo menos, uma segmentação de texto em número superior a 8000 *tokens*.

Os segmentos de texto extensos necessitam de um procedimento adicional de mapeamento e redução da extensão dos mesmos, de forma a serem conformes ao limite de *tokens* definido pelo modelo gerador de *embeddings*.

O conjunto de dados dos documentos extensos é constituído por 125 diplomas e 862 segmentos de texto, 18,33 % dos diplomas do trabalho e 40,81 % dos segmentos de texto gerados.

A sumarização é uma técnica de NLP, que permite reduzir documentos extensos, em blocos de texto de menor dimensão, ainda assim mantendo o contexto e veracidade do conteúdo.

Existem duas grandes estratégias para a sumarização de documentos: a sumarização extrativa e a sumarização abstrativa.

A sumarização extrativa consiste na recolha de termos ou frases do documento, e na geração de um sumário. Por norma não gera novo texto, mas sim reagrupa os termos ou frases mais representativos do texto num sumário compacto. Por outro lado, a sumarização abstrativa, cria novas frases que não se encontram no texto original. Para esse efeito usam modelos de linguagem natural avançados (como as redes neuronais ou modelos de aprendizagem profunda), que possuem uma capacidade "compreender o significado do texto" e o apresentar ao utilizador numa forma mais concisa, clara e eficaz.

Utilizou-se, para a sumarização dos segmentos mais longos, uma abordagem abstrativa, recorrendo ao mapeamento-redução, para um processamento de grande volume mais eficaz, em combinação com um LLM, para a geração de um novo texto coeso e bem articulado.

Procedeu-se da seguinte forma: o corpo do texto segmentado foi dividido em pedaços de menor dimensão, os quais foram sumarizados independentemente (fase de mapeamento) e, posteriormente, foram combinados num único sumário (fase de redução). Em cada uma das fases foi utilizado o LLM Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) de forma a reter as ideias principais, apresentando-as de forma mais sucinta que no texto original.

4.7.1. Fase de mapeamento

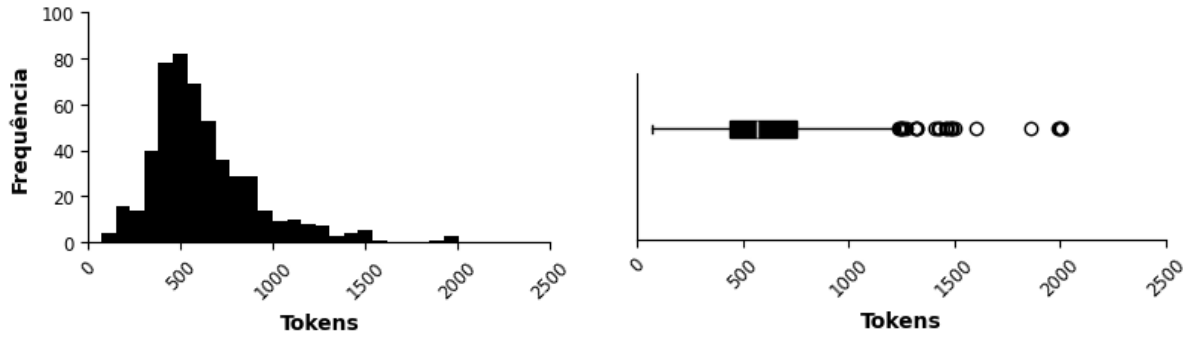
Para a fase de mapeamento foi utilizado um conjunto de dados composto pelos cinco diplomas com segmentos de texto mais extensos.

Foi definido um *prompt* específico para a fase de mapeamento[4], com o objetivo de orientar o LLM na tarefa de sumarização dos segmentos de texto. Este *prompt* detalha

[4] Ver formulação no Anexo B.

um conjunto de condições a cumprir, garantindo que o conteúdo seja condensado sem comprometer a exatidão ou coerência do texto.

Os diplomas foram previamente segmentados em blocos de 5000 caracteres, com uma sobreposição de 200 caracteres, de forma preservar o contexto entre os segmentos consecutivos. A partir dos cinco diplomas mais extensos, Tabela 4.2, obteve-se um conjunto de 92 segmentos distintos, aos quais foi aplicada a função de mapeamento, originando 515 novos segmentos de menor dimensão.



(a) Histograma. **(b)** Diagrama de caixa.

Figura 4.5. Histograma e diagrama de caixa da contagem de *tokens* (mapeamento).

Na análise ao histograma da distribuição de *tokens*, Figura 4.5a, destaca-se a assimetria à direita e um pico entre os 500-600 *tokens*. O diagrama de caixa, Figura 4.5b, complementa esta análise ao evidenciar uma distribuição quase simétrica entre os quartis Q_1 e Q_3 e a respetiva mediana (Q_2), assim como os limites superiores e inferiores (os segmentos de reta verticais, *whisker*), muito semelhantes.

Apesar da divisão dos segmentos de texto por um número uniforme de caracteres, subsistem alguns valores extremos, entre os 1250 e 2000 *tokens*, que não serão alvo de tratamento adicional.

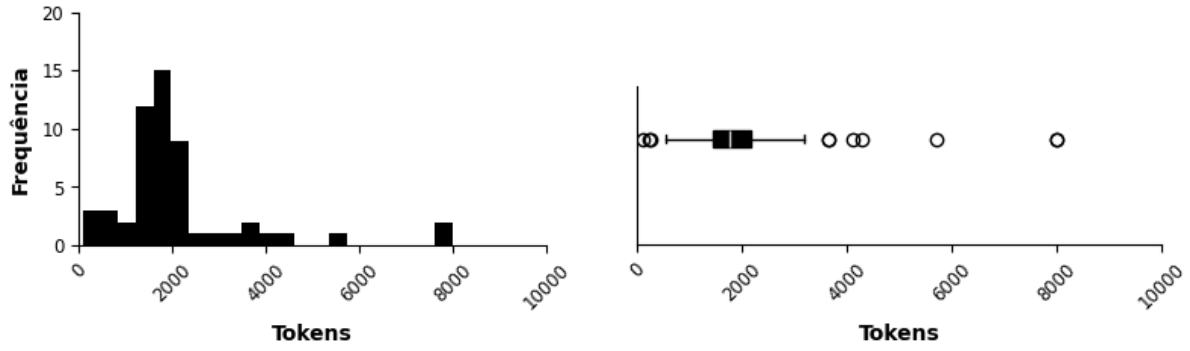
A Tabela 4.4 apresenta as estatísticas resultantes da fase de mapeamento:

Estatística	**Valor**
Mínimo	70,00
Primeiro quartil (Q_1: 25%)	437,00
Mediana (Q_2: 50%)	565,00
Terceiro quartil (Q_3: 75%)	756,50
Máximo	2001,00
Contagem	515,00
Média (μ)	624,22
Desvio padrão (σ)	292,05

Tabela 4.4. Estatística descritiva da contagem de *tokens* (mapeamento). *Nota.* — Valores apresentados com duas casas decimais significativas.

4.7.2. Fase de redução

A fase de redução consiste na agregação de vários segmentos de texto obtidos na fase de mapeamento e reduzi-los ainda a um menor número, também com a intervenção do LLM e sem perder contexto. Para esta fase o critério de geração de geração máximo de *tokens* foi limitado a 8000. O *prompt* utilizado na fase de redução[5] foi construído para ser um pouco mais permissivo em relação ao número total de frases geradas, resultando em 54 resumos. A aplicação desta estratégia permitiu uma redução de 89,51% no número de segmentos de texto gerados, mantendo o limite de *tokens* para a geração de *embeddings*.



(a) Histograma. **(b)** Diagrama de caixa.

Figura 4.6. Histograma e diagrama de caixa da contagem de *tokens* (redução).

A análise ao histograma, Figura 4.6a, constata-se a presença de uma distribuição unimodal, com um pico concentrado perto dos 2000 *tokens*. O diagrama de caixa, Figura 4.6b, complementa esta análise ao evidenciar uma concentração nos valores do Q_3, tendo o limite superior uma maior amplitude de distribuição de valores. Verificam-se valores extremos, mas dentro do limite definido de 8000 *tokens*.

A Tabela 4.5 apresenta o valor das estatísticas resultantes da fase de redução:

Estatística	**Valor**
Mínimo	101,00
Primeiro quartil (Q_1: 25%)	1439,25
Mediana (Q_2: 50%)	1775,00
Terceiro quartil (Q_3: 75%)	2157,25
Máximo	7990,00
Contagem	54,00
Média (μ)	2107,20
Desvio padrão (σ)	1535,70

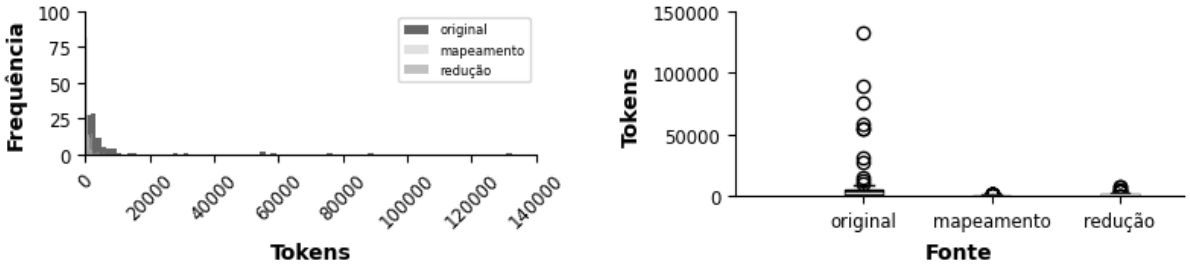
Tabela 4.5. Estatística descritiva da contagem de *tokens* (redução). *Nota.* — Valores apresentados com duas casas decimais significativas.

[5] Ver formulação no Anexo B.

4.7.3. Análise ao processo mapeamento-redução.

O processo de mapeamento-redução teve como objetivo a transformação do *corpus* inicial extenso em segmentos de texto mais reduzidos, através da decomposição e posterior agregação dos segmentos de texto, permitindo a contagem e agregação dos *tokens* de forma eficiente.

Esta fase analítica do processo é fundamental para compreender a composição dos textos e os padrões de distribuição do conjunto de dados. De seguida apresenta-se a distribuição obtida após a aplicação do processo:



(a) Histograma. **(b)** Diagrama de caixa.

Figura 4.7. Histograma e diagrama de caixa da distribuição de *tokens* no processo de mapeamento-redução.

No histograma, Figura 4.7a, na fonte texto original a distribuição da frequência é baixa, enquanto o número de *tokens* é elevado (acima dos 10 000). No processo de mapeamento assiste-se a uma drástica redução do número de *tokens* e ao aumento da frequência dos mesmos, assim como a uma distribuição bimodal. Na fase de redução, por sua vez, e devido à restrição do número de *tokens* se situar de até 8000, a distribuição possui valores de frequências mais elevados, para um menor número de texto.

O diagrama de caixa, Figura 4.7b, compara a distribuição dos valores do número de *tokens* com origem nas três fontes analisadas: 1 — Original; 2 — Mapeamento, e 3 — Redução.

Estatística	**Original**	**Mapeamento**	**Redução**
Mínimo	83,00	70,00	101,00
Primeiro quartil (Q_1: 25%)	1359,75	437,00	1439,25
Mediana (Q_2: 50%)	2780,50	565,00	1775,00
Terceiro quartil (Q_3: 75%)	4988,00	756,50	2157,25
Máximo	132 483,00	2001,00	7999,00
Contagem	92,00	515,00	54,00
Média (μ)	8680,13	624,22	2107,20
Desvio padrão (σ)	20 152,51	292,05	1535,70

Tabela 4.6. Estatística descritiva da processo mapeamento-redução. *Nota.* — Valores apresentados com duas casas decimais significativas.

Constata-se que na fonte original a dispersão de dados é maior, refletida na amplitude interquartil mais elevada e pela presença de valores extremos, Tabela 4.6. É um comportamento consistente as características iniciais do *corpus*, antes da aplicação de qualquer processo de transformação.

Na fase de mapeamento, verifica-se uma reduzida variabilidade, indicando que o processo de reestruturação dos segmentos textuais contribuiu para uma distribuição mais concentrada entre os valores da mediana.

Na etapa de redução nota-se uma maior compressão dos valores, indicativa que a operação de consolidação dos segmentos de texto eliminaram grande parte dos valores extremos que ainda subsistiam das fases anteriores. A mediana manteve-se praticamente inalterada, o que reforça a ideia da consolidação homogénea dos segmentos de texto.

Em geral, este comportamento observado confirma a eficácia progressiva das etapas de mapeamento e redução na homogeneização da distribuição dos dados, ao mesmo tempo que mantém as características estatísticas da amostra original.

4.8. Modelo de linguagem adotado

A escolha do modelo de linguagem para este trabalho baseou-se nos seguintes critérios: 1 — Possibilidade de execução em ambiente local. 2 — Disponibilização em código aberto. 3 — Suporte para a língua portuguesa; e 4 — ser enquadrado na categoria de Small Language Model (SLM).

Com base nos pressupostos definidos, foi selecionado o modelo Mistral-7B-instruct-v0.3. Trata-se de uma versão afinada com base no Mistral-7B-v0.3, treinada para seguir instruções, o que a torna adequada para a arquitetura RAG adotada neste trabalho. Esta terceira versão menor do modelo conta com um vocabulário com com 32 768 *tokens* e suporte à versão do tokenizador utilizado internamente pela Mistral.

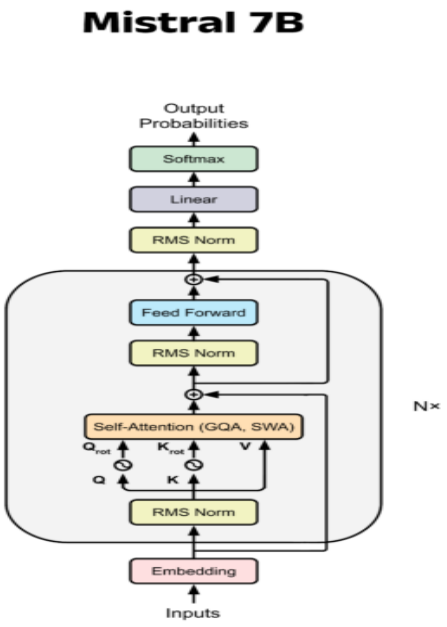


Figura 4.8. Arquitetura Mistral-7B, Smith (2024).

O modelo possui o suporte para a chamada de funções ou Application Programming Interface (API) diretamente a partir dos *prompts*, flexibilizando a iteração com o utilizador na execução de ações externas (p. ex., chamadas a API de meteorologia), ou na geração de respostas estruturadas (p. ex., elaboração de listas ou tabelas). Este treino adicional, baseado num *corpus* constituído por pares de pergunta-resposta, permite melhorar o desempenho do modelo em específicas de compreensão e geração de texto.

O modelo base Mistral-7B, possui ≈ 7000 milhões de parâmetros e assenta numa arquitetura *Transformer* do tipo *decoder only*, Figura 4.8. Em avaliações comparativas, o modelo demonstrou um desempenho superior a modelos da família Llama (1-34B, 2-7B e 2-13B) em vários domínios, com destaque para o raciocínio, matemática e geração de código, Figura 4.9.

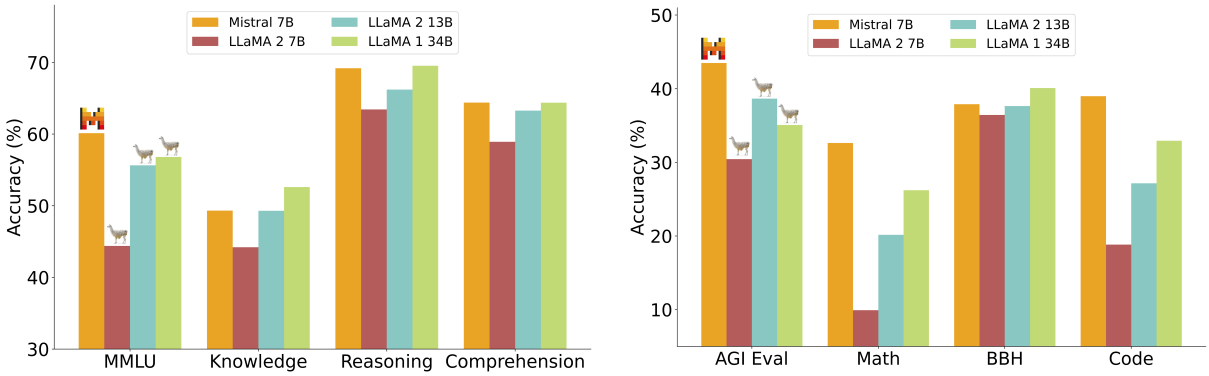


Figura 4.9. *Performance* do Mistral 7B vs modelos da família Llama em vários testes (fonte: Mistral 7B).

Destacam-se vários avanços tecnológicos integrados no modelo, como a Sliding Window Attention (SWA), Child et al. (2019) — Figura 4.10 —, conceitos introduzidos no Longformer, e a Grouped Query Attention (GQA), Ainslie et al. (2023) — Figura 4.11 —, que contribuem para uma maior eficiência e rapidez na inferência.

A SWA consiste numa otimização do mecanismo de atenção que, ao invés de calcular a atenção entre todos os *tokens* da sequência, o modelo limita a atenção a um número limitado de *tokens* anteriores, dentro de uma janela deslizante de ≈ 4096 *tokens*[6] que, teoricamente, tem um limite de atenção de 131 000 *tokens*. Desta forma, cada *token* retém o seu contexto local, mas a janela desliza à medida que a sequência avança, permitindo a incorporação de um contexto cada vez maior.

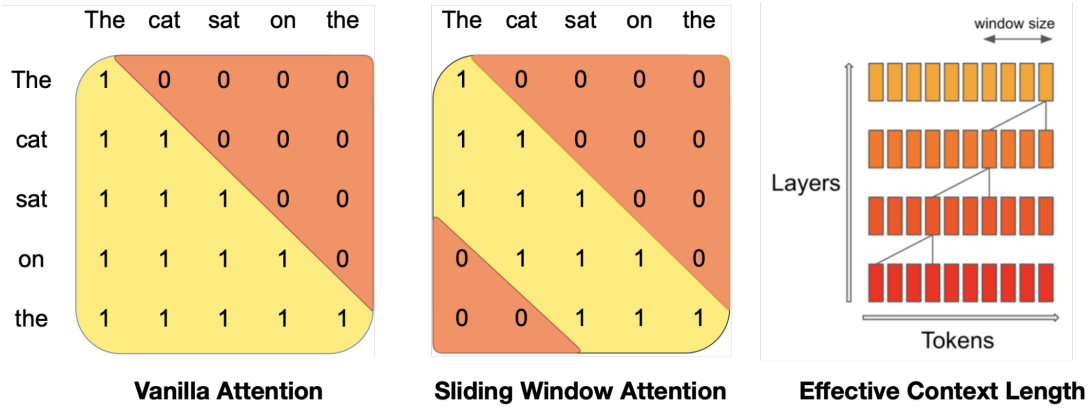


Figura 4.10. Mistral SWA, Jiang et al. (2023)).

[6] A versão utilizada neste trabalho tem uma capacidade de processamento de uma só sequência de 32 768 *tokens*.

A GQA, por sua vez, pretende lidar com a degradação e instabilidade do processo de treino associadas ao uso de Multi-Query Attention (MQA). Para tal, agrupa várias cabeças de *query* de forma a partilharem as mesmas chaves (k) e valores (v), reduzindo o consumo de memória — pela menor necessidade de utilização das matrizes K e V — e aumentando a velocidade de inferência e com uma qualidade comparável à da Multi-Head Attention (MHA).

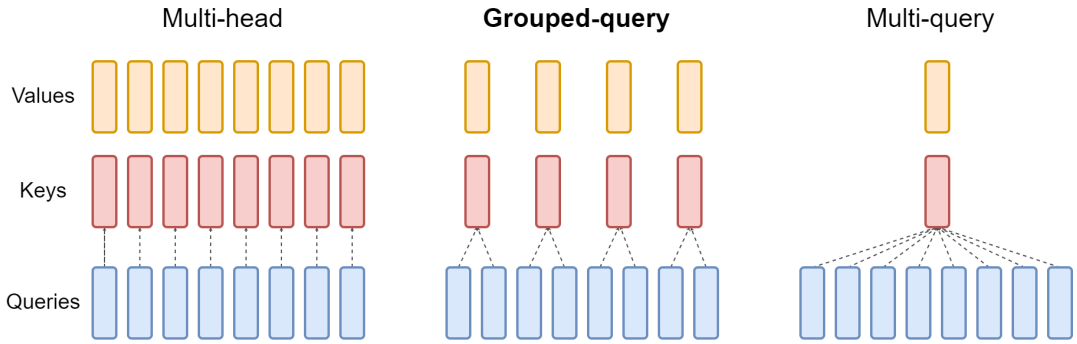


Figura 4.11. Arquitetura MQA, Ainslie et al. (2023)).

4.9. Engenharia de *prompts*

Com a disponibilização ao público do modelo de linguagem natural ChatGPT, assistiu-se a uma nova forma de interação das pessoas e a GenAI, capaz de nos dar respostas criativas, coerentes e cativantes. Contudo, a qualidade das respostas depende em grande medida da forma como as informações são solicitadas. É neste contexto de interação humano-máquina que surge a engenharia de *prompts*, uma abordagem que permite extrair o potencial da GenAI, produzindo resultados precisos, alinhados com as expectativas do utilizador e sem ambiguidades entre questões semelhantes.

Desta forma a engenharia de *prompts* tem um papel relevante na comunicação com LLM, tornando-se uma competência essencial para investigadores, profissionais da área de tecnologia da informação e utilizadores comuns. A aplicação adequada permite otimizar respostas, além de compreender as limitações e possíveis enviesamentos resultantes destas novas tecnologias. A compreensão do funcionamento e a aplicação de boas práticas à construção de *prompts* torna-se, por isso, relevante para o uso correto da GenAI.

Existem seis técnicas de engenharia de *prompt* baseadas em texto, Schulhoff et al. (2025):

- *Zero-Shot prompting*;
- *Few-Shot prompting*;
- *Chain-of-Thought prompting*;
- *Ensembling*;
- *Self-Criticism*;
- *Decomposition*.

O foco é a técnica *Few-Shot prompting*, Brown et al. (2020), utilizada neste trabalho. O *Few-Shot prompting* direciona a resposta do modelo LLM através da exemplificação de

padrões ao invés de instruções explícitas. De uma forma geral os passos a realizar são os seguintes:

Definição da tarefa: consiste na identificação clara do objetivo que o modelo deve cumprir. A redação caracteriza-se por ser assertiva, devendo-se identificar o tipo de informação que o modelo deve extrair ou a tarefa de geração de texto que deve realizar;

Seleção de exemplos: este é o passo central da técnica do *few-shot prompting*. Os exemplos escolhidos, que se recomenda que sejam vários, deverão ser representativos da tarefa que o modelo irá executar;

Consistência de formatação: a formatação dos exemplos deve ser constante. Esta uniformização da estrutura é útil para que o modelo reconheça o padrão existente e o possa seguir na geração de texto;

Estrutura da *prompt*: a *prompt* final deve ser clara. O inicio deverá contar com uma breve instrução da tarefa, seguido de alguns exemplos selecionados e terminando numa nova "entrada", de modo a que o modelo possa prosseguir com a geração da "saída";

Refinamento: é importante avaliar a qualidade da resposta gerada pelo modelo. No caso da resposta não seja satisfatória, poder-se-á proceder ao refinamento do *prompt*, podendo implicar a alteração dos exemplos fornecidos, ou mesmo da sua ordem, ou até proceder à adição de mais exemplos, caso a tarefa seja mais complexa.

O *prompt* indicado ao LLM segue a seguinte estrutura:

Modelo de mensagem de sistema: [7] a biblioteca LangChain possui uma função útil para a definição de modelos de mensagens de sistema — *SystemMessagePrompt-Template*. A mensagem de sistema permite enviar instruções ao modelo, nomeadamente aos baseados em *chat*, de forma a definir o seu comportamento ou o contexto onde vai atuar. É a primeira instrução que o modelo recebe, ainda antes de iniciar a interação com o utilizador e permite modelar a forma de resposta do LLM;

Modelo de mensagem humana: [8] enquanto o modelo de mensagem de sistema, define as regras e o comportamento do LLM, a função *SystemMessagePromptTemplate* é utilizada para definir uma *prompt* característica de uma mensagem do utilizador humano, neste âmbito, a pergunta que o modelo de GenAI deve responder. Para manter um dinamismo na geração das perguntas, e respetivas respostas como iremos ver mais à frente, sempre que é realizada uma chamada ao modelo existe uma função que permite gerar um conjunto de quatro perguntas, baseadas num conjunto verdade, construído a partir do conteúdo da linguagem clara extraído diretamente dos diplomas com as tipologias decreto-lei e decreto regulamentar, publicados no sítio oficial do Diário da República Eletrónico;

[7] Ver formulação no Anexo B
[8] Ver formulação no Anexo B

Modelo de mensagem de um assistente de AI: [9] como o nome indica, esta função — *AIMessagePromptTemplate* — representa a resposta de um assistente de AI numa iteração com o utilizador humano. A utilidade deste modelo de resposta está contida no conceito, ou técnica, de *Few-Shot prompting*. Esta técnica permite passar ao modelo um ou mais exemplos, que em conjunto constituem um padrão de conversa entre um utilizador humano e um assistente digital, de forma a que este último compreenda não só o tipo de resposta que se pretende obter, mas também um determinado estilo de resposta ou formato de saída, p. ex., pode ser desejável apresentar uma resposta do assistente digital no formato de listagem.

4.10. Geração de *embeddings* e indexação vetorial

Esta secção descreve a transformação dos textos jurídicos em representações vetoriais, *embeddings*, que permitem capturar as relações semânticas entre palavras e documentos. Aborda-se o modelo de linguagem adotado, o método de vetorização e a estratégias de mitigação do desalinhamento linguístico. Também é abordada a estrutura de armazenamento vetorial, que assegura uma recuperação eficiente da informação.

4.10.1. Modelo de geração de *embeddings*

O modelo para a geração de *embeddings* escolhido foi o BAAI/bge-m3, ou "M3-Embedding" (Chen et al., 2024), desenvolvido por elementos da Beijing Academy of Artificial Intelligence (BAAI), tendo em mente a unificação de três dimensões principais:

1 — Multilingue, disponibiliza suporte a mais de 100 línguas, entre as quais a portuguesa.

2 — Multifunção de recuperação, o gerador de *embeddings* unifica as funcionalidades mais comuns de recuperação num único modelo: *i*) *dense retrieval*;*ii*) *sparse retrieval*, e *iii*) *multi-vec retrieval*.

3 — Multi-granularidade (contexto longo), permitindo o processamento de texto com diferentes granularidades de entrada: *i*) *sentence-level*; *ii*) *passage level*, e *iii*) *doc-level*, podendo ir até 8192 *tokens*.

Na vertente técnica, destaca-se o vetor de *embedding* de 1024 dimensões, o comprimento máximo de entrada de até 8192 *tokens*, o que permite o processamento de um amplo contexto estimado em cerca de 15 a 25 pp. de texto. Acima deste valor o processamento pode impactar a eficiência do processamento e o consumo de memória. É recomendada a adoção do uso de uma abordagem híbrida de recuperação da informação (combinação *dense* e *sparse retrieval*), em conjunto com um mecanismo de *re-ranking* de forma a melhorar a precisão.

Em relação ao suporte multilingue, o desempenho do modelo depende muito da qualidade e disponibilidade dos dados para cada língua, o que pode resultar em valores de *performance* inferiores.

[9] Ver formulação no Anexo B

Os factores que levaram à escolha deste modelo foram a janela para a geração de *embeddings*, já que permite o processamento em uma só fase dos diplomas tipificados como breves, assim como os diplomas extensos, aos quais foram aplicados os processos de mapeamento e redução para o número de *tokens* de até 8000, e ao suporte para a língua portuguesa, aliado à *performance* e consumos de Random-Access Memory (RAM).

4.10.2. Mitigação de desalinhamento

O possível desalinhamento entre o modelo gerador de *embeddings* e o LLM constitui um obstáculo, uma vez que o modelo de *embeddings* não foi utilizado durante o treino do LLM. Esta situação pode pode resultar em representações vetoriais dos documentos que não refletem o espaço semântico esperado pelo LLM. DEsta forma, os documentos recuperados pelo mecanismo de busca vetorial podem ser semanticamente próximos no espaço dos *embeddings*, mas menos relevantes para o LLM, comprometendo a eficácia da recuperação da informação e a integração do contexto no sistema RAG.

Também exerce influência a utilização de modelo de geração de *embeddings* multilingue, que recorre a um tokenizador igualmente multilingue. No entanto, devido à diversidade de línguas que procura abranger, nem sempre consegue capturar eficazmente certas informações relevantes, como a riqueza do vocabulário e o equilíbrio de representatividade entre as línguas, de modo a evitar que o vocabulário mais extenso de uma delas seja um fator de domínio linguístico.

Um dos caminhos passava pela implementação de um *reranking* dos segmentos de texto. No entanto, de forma a garantir uma maior exatidão optou-se pela filtragem dos segmentos de texto que correspondam a um único diploma legislativo, que o utilizador especifica na questão colocada. Ao selecionar todos os segmentos gerados a um único documento, eliminamos a possibilidade de ruído, permitindo que a resposta final elaborada pelo LLM seja mais direta e precisa.

Caso o utilizador coloque uma questão que requeira a comparação ou a integração de conceitos entre mais que um diploma, esta orientação não tem sucesso. No entanto, futuras investigações poderão explorar estratégias que possam lidar com perguntas que requeiram a consulta de múltiplos documentos.

4.10.3. Base de dados vetorial e indexação

Para o sistema de busca e recuperação de dados, a escolha da base de dados vetorial recaiu sobre a ChromaDB.

A decisão foi baseada em várias características que a tornam uma solução robusta e adaptável, tanto para o ambiente de desenvolvimento local quanto para futuras implementações em larga escala.

A facilidade de uso e a leveza da ChromaDB permitiram uma integração suave e uma prototipagem rápida. Por ser uma base de dados em código aberto, oferece a flexibilidade e a transparência que são fundamentais para o desenvolvimento de trabalhos académicos e de código aberto. Além disso, a capacidade de ser embutida na aplicação a desenvolver

localmente elimina a dependência de um servidor externo durante a fase de desenvolvimento, simplificando bastante o processo de instalação e configuração.

A arquitetura da ChromaDB foi elaborada para maximizar o armazenamento e a pesquisa de *embeddings*, as representações vetoriais dos dados. Para garantir que os vetores mais relevantes sejam recuperados de forma eficiente e precisa em resposta a uma consulta, a base de dados faz uso de algoritmos de pesquisa de vizinhos mais próximos, como o HNSW.

Para além dos fatores técnicos, a ChromaDB revelou-se uma escolha robusta pela sua capacidade de escalabilidade e suporte ao ambiente em nuvem, permitindo que a solução evolua, acompanhando o aumento do volume de dados. Por último, a integração nativa com *frameworks* de desenvolvimento como o LangChain foi um fator determinante, pois a compatibilidade simplificou a criação do *pipeline* de processamento e a interação entre o modelo de linguagem e a base de dados vetorial.

4.11. Arquitetura do sistema LexClara

Esta secção detalha a arquitetura do sistema LexClara, a estrutura de trabalho que suporta a capacidade de recuperação de informação e geração de respostas do trabalho. Para alcançar o objetivo de alcançar o equilíbrio entre a simplificação das respostas e a contextualização aplicada à área jurídica, optou-se por uma arquitetura RAG.

A arquitetura proposta, Figura 4.12, foi concebida de forma modular, permitindo a avaliação de cada um dos componentes de forma independente.

Apresenta-se a estrutura dos principais módulos que compõem o sistema RAG desenvolvido:

- A ingestão de dados realiza-se através da extração de dados do sítio do *Diário da República* (1A), para a recolha dos conteúdos da linguagem clara. Assim como a extração da base de dados do *corpus* da legislação a partir do sítio do DRE Tretas (1B);
- A divisão em segmentos de texto (2A) foi definida em preâmbulo e capítulos, de forma a manter os conceitos legislativos concentrados;
- Se os segmentos de texto são extensos, >8000 *tokens*, é ativado o processo de mapeamento-redução (3) de forma a gerar novos segmentos de texto (2B) dentro do limite da dimensão do modelo de geração de *embeddings*. Este processo tem a intervenção do LLM Mistral-7B-instruct-v0.3 (9A);
- No caso dos segmentos de texto sejam breves, o modelo gerador de *embeddings* (4) converte para o formato vetorial o conteúdo legislativo (5), sendo estes valores gravados para um conjunto de dados para posterior inspeção, assim como são armazenados na base de dados vetorial (6);
- A base de dados vetorial (6) é carregada com o conteúdo do diploma, o(s) respetivo(s) vetor(es) e uma metadata associada, correspondente à identificação do diploma;

- O utilizador (7) ao colocar uma questão esta é enviada para o *retriever* (8), que por sua vez consulta a base de dados vetorial e a resposta que recebe é, por sua vez, passada ao LLM Mistral-7B-instruct-v0.3 (9B) que a simplifica e a envia para o *prompt* (10) do utilizador.

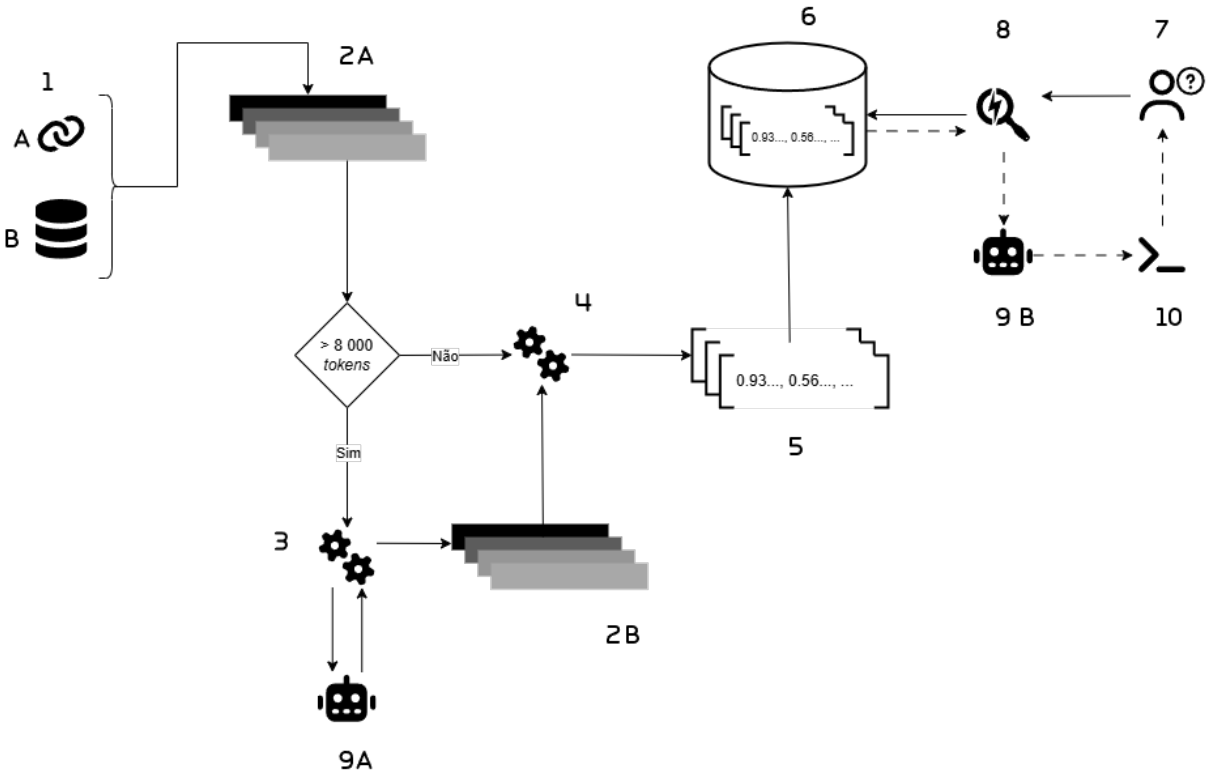


Figura 4.12. Arquitetura do sistema RAG, LexClara (imegem do autor).

4.12. Interface com utilizador

A presente secção descreve os controlos e parâmetros ajustáveis da interface desenvolvida para a iteração do utilizador com o LLM.

O utilizador pode selecionar um diploma aleatório através do botão "Usar Exemplo Aleatório". Desta forma, é disponibilizada uma questão presente no conjunto de dados da linguagem natural no campo "Pergunta", assim como a resposta que foi elaborada pelo diário oficial no campo "Resposta Esperada (para avaliação)". A articulação dos valores destes dois campos permite o cálculo das métricas de avaliação utilizadas para a validação do modelo, as quais podem ser registadas através do botão "Gravar métricas", Figura 4.13.

Importa salientar que, nesta fase de desenvolvimento, o utilizado não pode inserir texto livre no campo "Pergunta", estando limitado à seleção de questões presentes no conjunot de dados. Esta restrição visa garantir a uniformização do processo de avaliação automática. A possibilidade de introdução de questões livres constitui um desenvolvimento futuro do LexClara.

O utilizador tem também a possibilidade de configurar as variáveis de forma manual, nomeadamente as variáveis que controlam o comportamento de geração da resposta do LLM, permitindo, p. ex., afinar o grau de criatividade da resposta gerada.

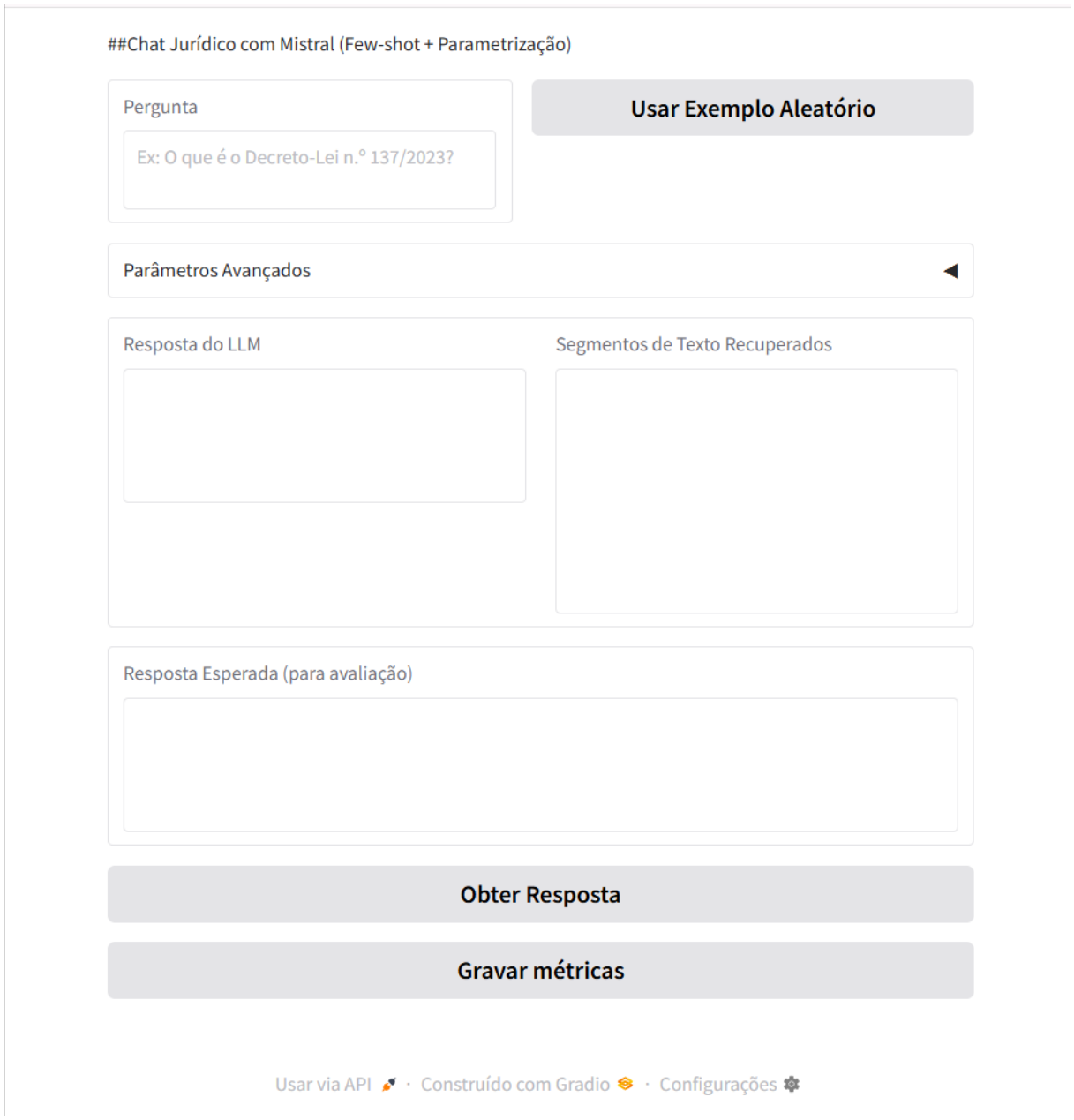


Figura 4.13. Interface do sistema RAG, LexClara (imagem do autor).

Como o modo de amostragem (*sampling*) está ativo nas definições do LLM, a resposta obtida depende da definição destes parâmetros, que controlam, p. ex., a aleatoriedade e ou extensão do texto gerado.

Em seguida apresenta-se uma breve descrição dos hiperparâmetros disponibilizados no interface, Figura 4.14:

- **Temperatura**: controla o nível de aleatoriedade na geração de texto. Valores mais baixos (p. ex.: 0.0 – 0.3) tornam as respostas mais determinísticas e focadas, enquanto valores mais altos (p. ex.: 0.8 – 1.0) promovem maior diversidade e criatividade;

- **Top-p (*nucleus sampling*)**: define o limite cumulativo de probabilidade para a seleção das próximas palavras, p. ex., um valor de $top_p = 0{,}9$ restringe a escolha às palavras cujo somatório de probabilidades atinge 90%;
- **Top-k**: determina o número máximo de *tokens* candidatos considerados em cada etapa de geração, p. ex., $top_k = 50$, apenas as 50 palavras (ou caracteres) mais prováveis são elegíveis para a próxima posição. Um valor menor de k, p. ex., 5 – 20 a saída é mais previsível e factual, ao invés, p. ex., 50 – 100 a resposta é mais imprevisível e com palavras menos prováveis, gerando um texto com maior criatividade;
- **Max tokens**: especifica o número máximo de *tokens* que o modelo pode gerar em cada resposta. Este limite previne saídas demasiado extensas e permite controlar o custo computacional do processo de inferência;
- **Repetition penalty**: Aplica uma penalização a *tokens* que já foram utilizados na sequência gerada, reduzindo a probabilidade de repetições redundantes. Valores ligeiramente superiores a 1, p. ex., 1,1 – 1,3 são eficazes para melhorar legibilidade das respostas, evitando repetições de palavras ou expressões.

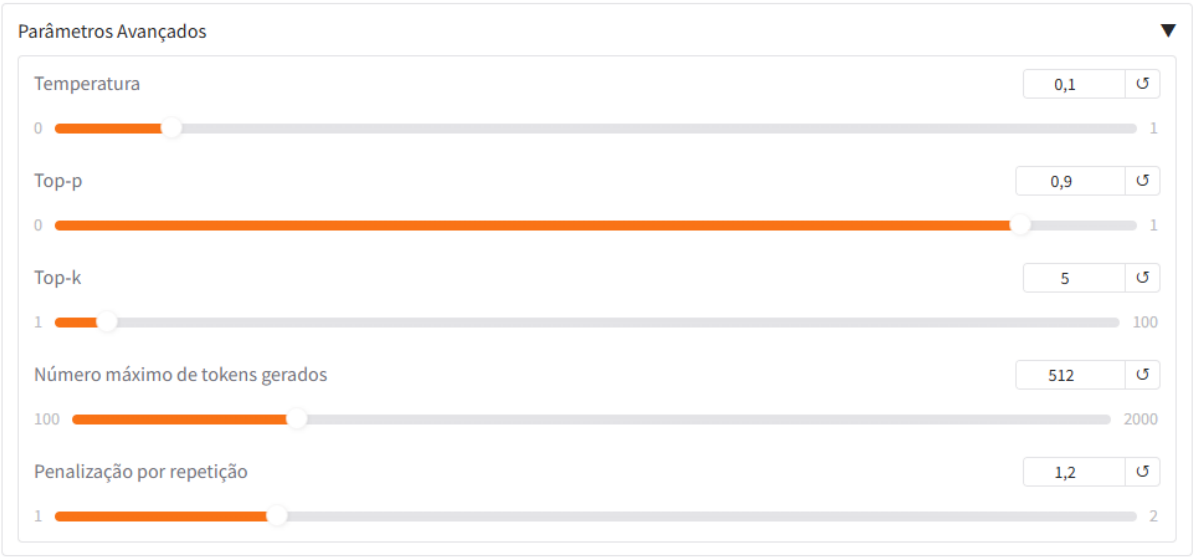


Figura 4.14. Parâmetros personalizáveis do sistema RAG, LexClara (imagem do autor).

Destaca-se a possibilidade de gravar um conjunto de métricas de avaliação da qualidade da resposta gerada, as quais são: 1 — A similaridade de coseno. 2 — A métrica Bilingual Evaluation Understudy (BLEU) , e 3 — A métrica Recall-Oriented Understudy for Gisting Evaluation (ROUGE). Os resultados destas métricas são utilizados exclusivamente para a análise e validação das respostas no âmbito desta dissertação.

CAPíTULO 5

Resultados

O presente capítulo é dedicado à apresentação dos resultados obtidos na utilização da aplicação desenvolvida no decorrer deste trabalho.

A análise de resultados encontra-se dividida em três partes: 1 — Foca no desempenho do sistema com documentos breves, caracterizados por um menor volume de texto. 2 — Detalha os resultados obtidos no processamento dos documentos longos, que resultaram da execução do processo de mapeamento-redução, essencial para gerir o volume de dados, e 3 — Avaliação humana, teste A/B realizado por profissionais da área jurídica.

Para avaliar a qualidade das respostas geradas pelo LLM em relação ao conjunto verdade, selecionaram-se três métricas utilizadas em tarefas de avaliação automática de modelos de linguagem natural, de acordo com a revisão da literatura:

Similaridade de coseno: mede a proximidade semântica entre o texto gerado pelo LLM e o respetivo conjunto verdade, permitindo quantificar a semelhança semântica entre ambos;

BLEU: o foco desta métrica é a precisão, sendo utilizada para comparar uma sequência de palavras em comum, n-gramas, entre o texto gerado e o conjunto-verdade;

ROUGE: aplicado à sumarização de textos, esta métrica avalia se o texto gerado pelo modelo capta as ideias principais do conjunto verdade.

A utilização destas métricas, na análise quantitativa, permite o reconhecimento de padrões e a capacidade do sistema LexClara em gerar conteúdo relevante e coerente.

5.1. Análise aos documentos breves

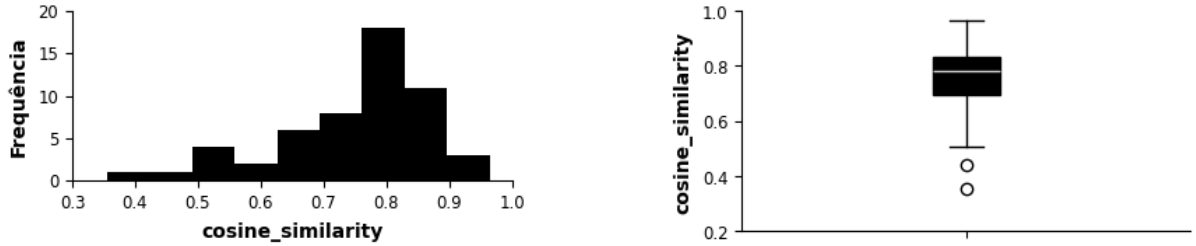
Os resultados apresentados nesta secção resultam da recolha de 54 observações, relativas às respostas realizadas ao sistema LexClara.

Estatística	Similaridade coseno	BLEU	ROUGE L
Média (μ)	0,75	0,18	0,36
Mediana (Q_2: 50%)	0,78	0,13	0,34
Desvio padrão (σ)	0,12	0,16	0,19
Mínimo	0,36	—	0,04
Máximo	0,96	0,57	0,78

Nota. — Valores apresentados com duas casas decimais significativas.

Tabela 5.1. Estatística descritiva das respostas do sistema LexClara.

Verifica-se na Tabela 5.1 que a similaridade de coseno[1] apresenta valores de média (μ) elevados (0,75), com um máximo de 0,96. Por sua vez a métrica BLEU (0,18) indica uma baixa semelhança entre o texto gerado pelo LexClara e o texto do conjunto verdade (o texto correspondente à resposta que se encontra publicado no *Diário da República*). Em relação à métrica ROUGE L, um valor de média (μ) 0,36 é indicativo que, $\approx \frac{1}{3}$ das palavras aparecem em ordem semelhante, indicando que o modelo tem alguma capacidade de capturar a estrutura da frase de referência (não significando que o conteúdo esteja errado), atingindo um valor máximo de 0,78.

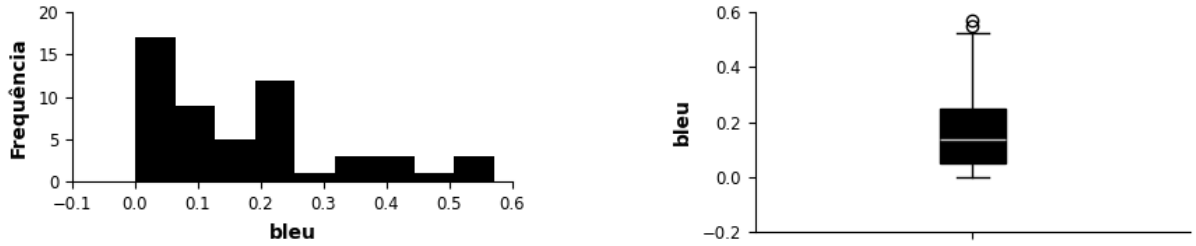


(a) Histograma. **(b)** Diagrama de caixa.

Figura 5.1. Histograma e diagrama de caixa da similaridade de coseno.

A análise ao histograma, Figura 5.1a, revela-nos uma distribuição unimodal, com um pico centrado entre os 0,75-0,85 e assimetria à esquerda, reveladora de uma maior ocorrência de valores mais elevados da similaridade de coseno.

Em relação ao diagrama de caixa, Figura 5.1b, verifica-se que limite mínimo se situa acima dos 0,5, com uma mediana perto dos 0,8 (reforçando a assimetria verificada no histograma), e o primeiro quartil ($Q1$) a mostrar uma maior variabilidade das observações em relação ao quartil terceiro ($Q3$), apresentando alguns valores extremos abaixo do limite inferior.

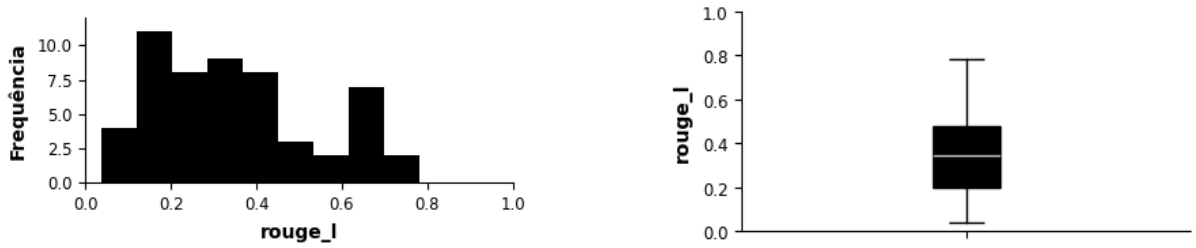


(a) Histograma. **(b)** Diagrama de caixa.

Figura 5.2. Histograma e diagrama de caixa da métrica BLEU.

Em relação ao diagrama de caixa, Figura 5.2b, verifica-se que limite máximo se situa perto dos 0,5, com uma mediana situada nos 0,12, e o terceiro quartil ($Q3$) a mostrar uma maior variabilidade das observações em relação ao quartil primeiro ($Q1$), apresentando alguns valores extremos.

[1] Anexo D.



(a) Histograma. **(b)** Diagrama de caixa.

Figura 5.3. Histograma e diagrama de caixa da métrica ROUGE L.

Na análise à métrica ROUGE L o histograma, Figura 5.3a, apresenta características bimodais, com os picos a se situarem nos intervalos de 0,15-0,2 e 0,6-0,7 e uma ligeira assimetria à direita.

O diagrama de caixa, Figura 5.3b, realça os baixos valores obtidos na métrica ROUGE L, com uma mediana a situar-se abaixo dos 0,4, uma dispersão equilibrada das observações pelos quartis $Q1$ e $Q3$, um limite superior perto dos 0,8 e sem a presença de valores extremos.

A utilização conjunta das métricas similaridade de coseno, BLEU e ROUGE L permitem obter um perfil da qualidade de resposta do sistema LexClara, face ao conjunto de referência constituído por respostas elaboradas por juristas e ou linguistas, em termos de organização semântica, rigor lexical e integridade da estrutura das frases. Tendo como base os valores da média (μ) de cada uma das métricas:

- Similaridade de coseno: 0,75;
- BLEU: 0,18;
- ROUGE L: 0,36.

A similaridade de coseno é consistente com uma resposta gerada coerente e alinhada com o contexto do texto de referência, sendo capaz de capturar o sentido da resposta esperada. Em contraste a baixa pontuação da métrica BLEU, que mede afere a precisão exata de frases através da sobreposição de *n-gramas* contíguos, mostra um conteúdo, ou vocabulário diferente do texto de referência. O valor ROUGE L indica uma *performance* razoável, indicando que a resposta consegue capturar o conteúdo essencial, mas falhando na coesão da frase, uma vez que há informação que não é capturada na ordem correta.

Em conclusão, o modelo compreende o que precisa ”dizer”, mas não usa o texto de referência (que não pertence ao conjunto de dados carregado), cumprindo o objetivo de preservação do significado das frases, em detrimento de uma linguagem, ou terminologia, mais técnica. Dessa forma cumpre o objetivo de simplificação da linguagem para o cidadão leigo na área jurídica.

5.2. Análise aos documentos longos

A análise aos documentos longos teve como base um conjunto de 13 observações, correspondentes aos diplomas constantes na Tabela 4.2.

Estatística	Similaridade de coseno	BLEU	ROUGE L
Média (μ)	0,62	0,03	0,18
Mediana (Q_2: 50%)	0,63	—	0,10
Desvio padrão (σ)	0,18	0,07	0,18
Mínimo	0,29	—	—
Máximo	0,89	0,19	0,52

Nota. — Valores apresentados com duas casas decimais significativas.

Tabela 5.2. Estatística descritiva das respostas do sistema LexClara.

Na Tabela 5.2 constata-se os seguintes resultados para a média: 1 — Similaridade de coseno: 0,62. 2 — BLEU: 0,03, e 3 — ROUGE L: 0,18. Estes valores indicam que: As respostas captam o contexto correto do tema, mas não de uma forma eficaz (1), o modelo responde com diferentes palavras e numa ordem incorreta face ao conjunto de referência (2 e 3).

À pergunta: "Quando entra em vigor o diploma Decreto-Lei n.º 15/2022, de 14 de janeiro?". Obtém-se como resposta do sistema LexClara: "O diploma Decreto-Lei n.º 15/2022, de 14 de janeiro **entrou em vigor no dia 14 de abril de 2022**.", que se encontra incorreta. Sendo a resposta correta: "Este decreto-lei **entra em vigor no dia 15 de janeiro de 2022**."

Como possíveis causas para estes resultados destacam-se: 1 — O reduzido conjunto de documentos que foi carregado, podendo não estar a fornecer contexto suficiente para que o *retriever* devolva segmentos de texto relevantes para o LLM gerar respostas corretas, e 2 — O processo de mapeamento-redução pode estar a descartar informação relevante para as questões colocadas.

5.3. Avaliação humana

As respostas foram selecionadas a partir dos 1.º e 4.º melhores resultados da similaridade de coseno, valores presentes na Tabela 5.3, obtidos do conjunto de avaliação de respostas breves do sistema LexClara. Esta seleção intencional incluiu casos representativos de elevada correspondência semântica, utilizados no teste A/B com avaliadores jurídicos, permitindo analisar o desempenho do modelo. Seguem-se as métricas associadas a cada resposta:

Resposta	Similaridade de coseno	BLEU	ROUGE L
1.ª	0,96	0,57	0,78
4.ª	0,87	0,25	0,45

Nota. — Valores apresentados com duas casas decimais significativas.

Tabela 5.3. Valores das métricas associadas às respostas do teste A/B cego.

A avaliação de qualidade das respostas foi conduzida através de um teste A/B cego, realizado por dois avaliadores, designados por avaliador jurídico 1 (AJ1) e avaliador jurídico 2 (AJ2), ambos possuindo vasta experiência na área jurídica.

Em cada pergunta foram apresentadas duas respostas candidatas — uma gerada pelo sistema LexClara e a outra publicada na linguagem clara no Diário da República Eletrónico. Não foi facilitada aos avaliadores jurídicos a informação da origem de cada resposta. Solicitou-se aos avaliadores jurídicos que realizassem uma análise qualitativa de cada resposta, identificando, qual havia sido gerada pelo sistema LexClara e qual correspondia à redação humana elaborada por juristas.

Teste enviado ao AJ1:

— **Pergunta**: "O que vai mudar com o diploma Decreto-Lei n.º 63/2023, de 31 de julho?"

— **Resposta 1**: 'Passa a ser possível aos titulares de título de condução caducado por via legal proceder à sua revalidação sem submissão a exame especial.
Esta revalidação fica condicionada à apresentação de atestado médico para os condutores com mais de 60 anos.
O regime previsto no presente decreto-lei aplica-se aos títulos de condução emitidos antes de 1 de janeiro de 2008, que habilitem à condução de veículos das categorias AM, A1, A2, A, B1, B, BE e de veículos agrícolas.
O regime excecional previsto no presente decreto-lei aplica-se por um período de um ano a contar do dia da sua entrada em vigor."

— **Resposta 2**: 'Este decreto-lei cria um regime extraordinário e temporário de revalidação de títulos de condução. Os titulares de títulos de condução caducados por via legal poderão proceder à sua revalidação sem submissão a exame especial. Esta revalidação está condicionada à apresentação de atestado médico para os condutores com mais de 60 anos. Este regime aplica-se aos títulos de condução emitidos antes de 1 de janeiro de 2008, cujos prazos de validade constantes dos respetivos documentos físicos não correspondam ao prazo legalmente previsto e em vigor, e que habilitem à condução de veículos das categorias AM, A1, A2, A, B1, B, BE e de veículos agrícolas. Este regime aplica-se por um período de um ano a contar do dia da sua entrada em vigor."

Avaliação do AJ1:

"Li com atenção o Decreto-Lei n.º 63/2023, de 31 de julho, publicado no *Diário da República* (hiperligação aqui) e, a partir da contextualização do diploma, é possível perceber algumas nuances que ajudam a distinguir a mão de quem redigiu cada resposta.

Eu li as perguntas e respostas. A primeira parece mais simples, clara e objetiva. Já a segunda mais complexa, mais "mais completa". A Resposta 1 tem um estilo mais telegráfico e direto, típico de sistemas automatizados ou de uma síntese puramente informativa. Limita-se a reproduzir as linhas centrais do diploma, com construções sintáticas mais secas e sem preocupação com clareza narrativa ou contextualização. Já a Resposta 2 apresenta um tom mais técnico e interpretativo, introduzindo expressões como "cria um regime extraordinário e temporário", "cujos prazos de

validade constantes dos respetivos documentos físicos não correspondam ao prazo legalmente previsto", entre outras. Essa estrutura é característica de quem domina a técnica legislativa e faz leitura jurídica do texto — o que sugere autoria humana, provavelmente de um jurista ou técnico especializado.

Além disso, a Resposta 2 reproduz com mais precisão o conteúdo normativo do diploma, destacando condições e enquadramentos específicos que a primeira simplesmente omite. Em termos de qualidade redacional e densidade jurídica, a segunda é mais completa e fiel ao espírito do decreto-lei.

Em suma:

Resposta 1: estilo sintético e informativo → provavelmente gerada por sistema ou linguagem clara.

Resposta 2: estilo técnico, jurídico e contextualizado → redigida por jurista ou técnico legislativo."

Teste enviado ao AJ2:

— **Pergunta**: "O que vai mudar com o diploma Decreto-Lei n.º 134/2023, de 28 de dezembro?"

— **Resposta 1**: "Com o diploma Decreto-Lei n.º 134/2023, de 28 de dezembro, é aprovado o prémio salarial de valorização das qualificações no mercado de trabalho. Este prémio é destinado a jovens trabalhadores que apresentem declaração do imposto sobre o rendimento das pessoas singulares (IRS) em território nacional detentores do grau académico de licenciado ou de mestre, ou graus académicos estrangeiros reconhecidos com o nível, objetivos e natureza idêntico aos graus portugueses de licenciado ou de mestre. Os licenciados e mestres passam a ter direito a receber anualmente um prémio salarial no valor de 697 euros por cada ano de licenciatura e 1500 euros por cada ano de mestrado. Este incentivo é atribuído pelo número de anos de trabalho equivalente à duração regular do ciclo de estudos concluído. Os licenciados e mestres que tenham obtido o grau académico em data anterior a 2023 têm direito ao prémio salarial desde que o número de anos subsequente à atribuição do grau académico elegível seja inferior ao número de anos do ciclo de estudos. Estes beneficiários têm direito a receber o valor do prémio pelo número de anos remanescente. A atribuição do prémio salarial depende da verificação cumulativa dos seguintes requisitos: a) sejam titulares de grau académico de licenciado ou mestre, ou de grau académico estrangeiro reconhecido, com data de atribuição no ano de 2023 ou seguintes, inclusive, considerando-se como tal o primeiro grau académico obtido pelo beneficiário atribuído por instituições de ensino superior nacionais, públicas ou privadas, ou reconhecido em Portugal; b) tenham auferido rendimentos de categoria a ou b do IRS; c) tenham, no ano de atribuição do prémio salarial, até 35 anos de idade, inclusive; d) sejam residentes em território nacional, nos termos e para os efeitos do disposto no artigo 19.º do decreto-lei n.º 398/98, de 17 de dezembro, na sua redação atual; e) tenham a sua situação tributária e contributiva regularizada

perante a autoridade tributária e a segurança social.

— **Resposta 2**: "Os licenciados e mestres passam a ter direito a receber anualmente um prémio salarial no valor de 697 euros por cada ano de licenciatura e 1500 euros por cada ano de mestrado.

O prémio salarial é atribuído anualmente, pelo número de anos de trabalho equivalente à duração regular do ciclo de estudos concluído.

Para a atribuição do prémio é necessário reunir as seguintes condições:

Ter até 35 anos de idade;

Ser residente em território nacional;

Ser contribuinte com rendimentos nas categorias A e/ou B do IRS; e

Ter obtido, nos anos de 2023 e seguintes, grau de licenciado e/ou mestre, em qualquer área científica (inclui graus académicos estrangeiros equivalentes aos graus portugueses). Os beneficiários que tenham concluído o grau académico antes de 2023 têm direito ao prémio pelo número de anos remanescente, desde que o número de anos subsequente à atribuição daquele grau académico seja inferior ao número de anos equivalentes ao ciclo de estudos respetivo.

As condições são verificadas anualmente.

O prémio salarial é requerido pelo beneficiário em formulário eletrónico e pago pela Autoridade Tributária e Aduaneira, por transferência bancária para o IBAN constante dos respetivos sistemas."

Avaliação do AJ2:

"A meu sentir que ambas respostas estão corretas, sendo a resposta numero um com mais elementos técnicos e robustos, escritos com uma maior formalidade enquanto a resposta n.º 2 (sistema) é escrita de forma mais objetiva, pontual e resumida.

A explicação que o cidadão compreende melhor é a resposta n.º 2. O cidadão não tem conhecimento jurídico técnico, a preocupação e o foco são o resultado prático, para ele pouco importa a citação do número da lei por exemplo. Enquanto que para nós juristas a resposta prática e resumida faltam elementos importantes, o número da lei por exemplo é importante para saber quais pessoas receberiam o benefício em um interstício de tempo, analisar se já está em vigor ou contados da publicação etc. É como um químico dizendo para uma pessoa: «o uso da enzima rara proteolítica, dissolve a proteína e ajudará com a cicatrização da ferida". O doente quer apenas saber que use creme de papaína x dias, x vezes que resolve.»"

Resultados:

Teste enviado ao AJ1: Resposta n.º 1 elaborada por jurista e resposta n.º 2 elaborada pelo sistema LexClara;

Teste enviado ao AJ2: Resposta n.º 1 elaborada pelo sistema LexClara e resposta n.º 2 elaborada por jurista.

Conclui-se que após o término do teste A/B cego ambos avaliadores jurídicos não foram capazes de distinguir, com exatidão, qual das respostas tinha sido gerada pelo sistema LexClara e qual correspondia à redação humana. Este resultado é relevante, pois demonstra que a simplificação automática da linguagem jurídica pode ter uma qualidade comparável à redação elaborada por humanos, demonstrando o potencial do sistema LexClara como ferramenta de simplificação do discurso jurídico.

CAPíTULO 6

Conclusões e trabalho futuro

O principal objetivo deste trabalho consistiu no desenvolvimento de um *chatbot* jurídico, suportado por uma arquitetura RAG, com capacidade de simplificar a linguagem jurídica, direcionado ao cidadão leigo, fornecendo respostas claras e objetivas com base nas quatro perguntas que contextualizam um diploma legislativo.

A execução do trabalho implicou a orquestração de diversas etapas, como a extração, a limpeza e a segmentação dos diplomas legais publicados no Diário da República Eletrónico. Nas fases de processamento da informação foi necessário aplicar a abordagem de processamento por lotes, especialmente na etapa de geração de *embeddings* (modelo BAAI/bge-m3) e respetiva indexação na base de dados vetorial (ChromaDB), assim como na fase de mapeamento e redução, de forma assegurar um ambiente computacional responsivo.

Para a geração de respostas recorreu-se ao modelo de linguagem Mistral-7B-Instruct-v0.3, disponibilizado via Hugging Face e coordenado pela *framework* LangChain, sendo disponibilizado ao utilizador uma interface em Gradio.

A avaliação da qualidade das respostas do sistema LexClara foi realizada com base em métricas amplamente utilizadas em NLP: a similaridade de coseno, para verificar a proximidade semântica entre a resposta gerada pelo sistema LexClara e o conjunto verdade), a BLEU, para medir a precisão de n-gramas, e a BLEU, que avalia a mais longa sequência de palavras, não necessariamente consecutivas, mas ainda em ordem, que é partilhada entre o texto gerado e o conjunto verdade.

Os resultados obtidos demonstram que o sistema LexClara ao gerar respostas enquadradas no tema e numa linguagem clara, simplificando o acesso à informação jurídica, demonstrando o potencial da utilização da GenAI na área jurídica. As métricas de avaliação — similaridade de coseno (0,96), BLEU (0,72) e ROUGE (0,81) — confirmam a qualidade da resposta nos casos com elevada correspondência semântica. Verificam-se algumas respostas incorretas, ou "alucinadas", que poderão ter origem na escolha tanto do modelo de *embeddings* como do LLM que, embora se apresentem com capacidade multilingue, na qual se inclui a língua portuguesa, a fraca representatividade da mesma, quer em número de vocábulos — resultante da pouca diversidade de textos ou até a inexistência de conjuntos de textos jurídicos nas fases de treino —, prejudica a qualidade a resposta mesmo a questões diretas.

Em relação à qualidade geral das respostas do sistema LexClara, constata-se a necessidade de realizar algumas melhorias.

Tendo em vista a melhoria contínua do sistema, deixam-se várias linhas orientadoras de desenvolvimentos futuros:

- Aplicação da tarefa de afinação (*finetuning*) supervisionada do modelo de *embeddings* com base num conjunto de dados de perguntas e resposta em português jurídico, de forma a melhorar a relação semântica de conceitos (p. ex., "Roubo e assalto são a mesma coisa?") ou o vocabulário jurídico em português (p. ex., "revogação tácita" ou "eficácia retroativa", podem não constar do conjunto de treino). Desta forma, não é necessário retreinar o modelo por completo, bastando para isso ajustar o *projection head* responsável pela geração dos *embeddings*.

  São diversos os ganhos potenciais com este processo, desde o reforço da coerência vetorial, em que os termos semelhantes ficam mais próximos no espaço vetorial — "revogação tácita"[1] ou "eficácia retroativa"[2] —, na melhoria da precisão da recuperação da informação, em que o modelo RAG passa a recuperar passagens relevantes em termos jurídicos — "Roubo e assalto são a mesma coisa?" —, ou a adaptação do modelo ao léxico e semântica da língua portuguesa, passando a reconhecer padrões ou formulações frequentes do português jurídico:
  i) Formulações habituais: "Nos termos da alínea *a*) do n.º 1 do artigo 198.º da Constituição, o Governo decreta o seguinte: [...]";
  ii) Frases passivas: "São revogados os seguintes diplomas: [...]";
- A utilização de um modelo especificamente treinado para a língua portuguesa favorece a compatibilidade semântica e estrutural entre o modelo gerador de *embeddings* e o LLM. Este alinhamento linguístico contribui para uma representação mais fidedigna dos vocábulos, das dependências sintáticas e concordâncias de termos, assegurando a coerência entre as fases de indexação e geração de textual;
- A implementação de um *ensemble* de LLM, combinando um modelo jurídico de menor dimensão — especializado na área jurídica — com um modelo de maiores dimensões, capaz de capturar mais riqueza textual e, consequentemente, melhorar as respostas fornecidas ao utilizador;
- A utilização de um vasto *corpus* de textos legislativos melhora o desempenho do LLM, permitindo ao tokenizador e gerador de *embeddings* capturar de forma mais precisa as variações semânticas e contextuais presentes nos textos. Desta forma o incremento da diversidade textual favorece, também, a melhoria da coocorrência, tornando os conceitos semelhantes mais próximos no espaço vetorial, resultando numa melhoria na qualidade da recuperação de informação;
- O trabalho consistiu na implementação de um sistema de *chatbot*, o LexClara. Estes sistemas de iteração máquina-utilizador (sistema conversacional) baseados em AI, beneficiam com a capacidade de reter as conversas anteriores, preservando o

[1] Ver Decreto-Lei n.º 41/2009, de 11 de fevereiro em: https://diariodarepublica.pt/dr/detalhe/decreto-lei/41-2009-601351
[2] Ver Portatia n.º 127/2017, de 3 de abril em: https://diariodarepublica.pt/dr/detalhe/portaria/127-2017-106804122

contexto e a continuidade do diálogo. Ao reter o conteúdo das iterações anteriores, o modelo compreende melhor o tema e irá adaptar as respostas ao utilizador, tornando esta comunicação mais contextual e contínua;

- É necessária uma análise mais aprofundada ao processo de mapeamento-redução, com o objetivo a avaliar o processo de sumarização realizado pelo LLM Mistral--7B-instruct-v0.3 aos documentos extensos.

Em síntese, o trabalho desenvolvido mostra a viabilidade técnica e o potencial da aplicação de modelos de GenAI ao domínio jurídico português, mostrando que a integração do conhecimento através de técnicas de RAG com a elaboração das respostas através de um LLM geram respostas contextualizadas e acessíveis, constituindo uma nova forma de acesso à informação legislativa.

Bibliografia

Abdallah, A., Piryani, B., & Jatowt, A. (2023). Exploring the State of the Art in Legal QA Systems. *Journal of Big Data*, *10*(1), 127. https://doi.org/10.1186/s40537-023-00802-8

Afflerbach, P. (1990). The Influence of Prior Knowledge and Text Genre on Readers' Prediction Strategies. *Journal of Reading Behavior*, *22*. https://doi.org/10.1080/10862969009547700

Ainslie, J., Lee-Thorp, J., Jong, M. de, Zemlyanskiy, Y., Lebrón, F., & Sanghai, S. (2023, dezembro 23). *GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints*. arXiv: 2305.13245 [cs]. https://doi.org/10.48550/arXiv.2305.13245

Albany Law School. (2023, outubro 30). *A Tale of Property and Dread, ChatGPT Pens Story about Land Contracts (in the style of Edgar Allan Poe) | Albany Law School*. Obtido julho 28, 2025, de https://www.albanylaw.edu/news/tale-property-and-dread-chatgpt-pens-story-about-land-contracts-the-style-edgar-allan-poe

Alsekait, D., Almanaseer, W., Gomaa, I., Zeid, M.-E., Baligh, S., Alaa, H., Hegazy, J., Magdy, J., Eldmrat, A., & Abdelminaam, D. (2024). Harnessing Deep Learning for Effective Extractive Text Summarization: A Comparative Study. *Int. Conf. Intell. Methods, Syst. Appl., IMSA*, 581–588. https://doi.org/10.1109/IMSA61967.2024.10652647

Assembleia da República. (2015, novembro). *Constituição da República Portuguesa Sétima Revisão Constitucional – 2005* (Assembleia da República – Divisão de Edições). Obtido outubro 8, 2025, de https://www.parlamento.pt/arquivodocumentacao/documents/crpviirevisao.pdf

Athugodage, M., Mitrofanove, O., & Gudkov, V. (2024, maio). Transfer Learning for Russian Legal Text Simplification. Em R. Wilkens, R. Cardon, A. Todirascu & N. Gala (Eds.), *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI) @ LREC-COLING 2024* (pp. 59–69). ELRA and ICCL. Obtido setembro 28, 2025, de https://aclanthology.org/2024.readi-1.6/

Beltagy, I., Peters, M. E., & Cohan, A. (2020, dezembro 2). *Longformer: The Long-Document Transformer*. arXiv: 2004.05150 [cs]. https://doi.org/10.48550/arXiv.2004.05150

Bhattacharya, P., Poddar, S., Rudra, K., Ghosh, K., & Ghosh, S. (2021, junho 30). *Incorporating Domain Knowledge for Extractive Summarization of Legal Case Documents*. arXiv: 2106.15876 [cs]. https://doi.org/10.48550/arXiv.2106.15876

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020, julho 22). *Language Models Are Few-Shot Learners*. arXiv: 2005.14165 [cs]. https://doi.org/10.48550/arXiv.2005.14165

Carbonell, J., & Goldstein, J. (1998). The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 335–336. https://doi.org/10.1145/290941.291025

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020, outubro 6). *LEGAL-BERT: The Muppets Straight out of Law School*. arXiv: 2010.02559 [cs]. https://doi.org/10.48550/arXiv.2010.02559

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., & Liu, Z. (2024, junho 28). *BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation*. arXiv: 2402.03216 [cs]. https://doi.org/10.48550/arXiv.2402.03216

Child, R., Gray, S., Radford, A., & Sutskever, I. (2019, abril 23). *Generating Long Sequences with Sparse Transformers*. arXiv: 1904.10509 [cs]. https://doi.org/10.48550/arXiv.1904.10509

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press. https://apps.dtic.mil/sti/tr/pdf/AD0616323.pdf

Chomsky, N. (2002). *Syntactic Structures* (2ª ed.). Mouton de Gruyter. https://www.ling.upenn.edu/courses/ling5700/Chomsky1957.pdf

Colby, K. M. (1975). *Artificial Paranoia: A computer simulation of paranoid processes* (1ª ed.). PERGAMON PRESS INC. https://doi.org/10.1016/C2013-0-02631-X

Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis*, *16*(1), 64–93. https://doi.org/10.1093/jla/laae003

Deepika, R., Das, S., Dey, N. S., Panuganti, J., & Hussain, M. R. (2024). Comprehensive Analysis of Text Summarization Techniques for Legal Documents. *Grenze International Journal of Engineering & Technology (GIJET)*, *10*, 817–824. Obtido setembro 28, 2025, de https://research.ebsco.com/linkprocessor/plink?id=a987f9b8-a04b-3143-bfe5-abfb3a6d760e

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, maio 24). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv: 1810.04805 [cs]. https://doi.org/10.48550/arXiv.1810.04805

Dias, M. R., Ribeiro, R., & Pinto, H. S. (2024). Contributions to Legal Document Summarization: Judgments from the Portuguese Supreme Court of Justice. Em M.

Rodrigues, J. P. Leal & F. Portela (Eds.), *13th Symposium on Languages, Applications and Technologies (SLATE 2024)* (2:1–2:14, Vol. 120). Schloss Dagstuhl – Leibniz-Zentrum für Informatik. https://doi.org/10.4230/OASIcs.SLATE.2024.2

Emanuilov, S., & Dimov, A. (2024). Billion-Scale Similarity Search Using a Hybrid Indexing Approach with Advanced Filtering. *Cybernetics and Information Technologies*, *24*(4), 45–58. https://doi.org/10.2478/cait-2024-0035

Flesch, R. (1948). A New Readability Yardstick. *Journal of Applied Psychology*, *32*(3), 221–233. https://doi.org/10.1037/h0057532

Freedman, D., & Diaconis, P. (1981). On the histogram as a density estimator:L2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *57*(4), 453–476. https://doi.org/10.1007/BF01025868

Ghimire, A., Shrestha, R., & Edwards, J. (2023). Too Legal; Didn't Read (TLDR): Summarization of Court Opinions. *Intermt. Eng., Technol. Comput., IETC*, 164–169. https://doi.org/10.1109/IETC57902.2023.10152119

Gu, J. (2025). A Research of Challenges and Solutions in Retrieval Augmented Generation (RAG) Systems. *Highlights in Science, Engineering and Technology*, *124*, 132–138. https://doi.org/10.54097/364hex16

Guo, H., Pasunuru, R., & Bansal, M. (2018, julho). Soft Layer-Specific Multi-Task Summarization with Entailment and Question Generation. Em I. Gurevych & Y. Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 687–697). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1064

Harikrishnan, K., Malathi, M., & Sundharakumar, K. (2024). Topic-Driven Contractual Language Understanding and Summarization: An Integrated Approach for Simplifying Legal Documents. *Int. Conf. Intell. Technol., CONIT*, 1–6. https://doi.org/10.1109/CONIT61985.2024.10627025

Hemanth Kumar, M., Jayanth, P., & Anand Kumar, M. (2024). Large Language Models for Indian Legal Text Summarisation. *Proc. CONECCT - IEEE Int. Conf. Electron., Comput. Commun. Technol.*, 1–5. https://doi.org/10.1109/CONECCT62155.2024.10677065

Hung, C.-C., Rim, W., Frost, L., Bruckner, L., & Lawrence, C. (2023). Walking a Tightrope - Evaluating Large Language Models in High-Risk Domains. Em Hupkes D., Dankers V., Batsuren K., Sinha K., Kazemnejad A., Christodoulopoulos C., Cotterell R. & Bruni E. (Eds.), *GenBench - GenBench: Workshop Gen, (Benchmarking) NLP, Proc.* (Pp. 193–203). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/2023.genbench-1.8

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023,

outubro 10). *Mistral 7B*. arXiv: 2310.06825 [cs]. https://doi.org/10.48550/arXiv.2310.06825

Jones, K. S. (2004). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *60*(5), 493–502. https://www.staff.city.ac.uk/~sbrp622/idfpapers/ksj_orig.pdf

Kale, A. R., & Deshmukh, P. R. (2024). Abstractive Text Summarization: A Transformer Based Approach. *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, 1–4. https://doi.org/10.1109/I2CT61223.2024.10544120

Kalyan, M., Dwivedi, P., & Sharma, S. (2024). Enhancing Legal Document Summarization for Professionals: An Extractive Approach. *Int. Conf. Comput. Commun. Netw. Technol., ICCCNT*. https://doi.org/10.1109/ICCCNT61001.2024.10725546

Karthick, K., Pooja, T., Oviya, V., Damodharan, J., & Senthamizh Selvi, S. (2024). Interactive Legal Assistance System using Large Language Models. *Int. Conf. I-SMAC (IoT Soc., Mob., Anal. Cloud), I-SMAC - Proc.*, 931–937. https://doi.org/10.1109/I-SMAC61858.2024.10714868

Kintsch, W., & van Dijk, T. A. (1978). Toward a Model of Text Comprehension and Production. *Psychological Review*, *85*(5), 363–394. https://doi.org/10.1037/0033-295X.85.5.363

Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., Zhou, X., Wang, E., & Dong, X. (2024, março 14). *Better Zero-Shot Reasoning with Role-Play Prompting*. arXiv: 2308.07702 [cs]. https://doi.org/10.48550/arXiv.2308.07702

Kshirsagar, A. (2024). Enhancing RAG Performance Through Chunking and Text Splitting Techniques. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *10*(5), 151–158. https://doi.org/10.32628/CSEIT2410593

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019, outubro 29). *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. arXiv: 1910.13461 [cs]. https://doi.org/10.48550/arXiv.1910.13461

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 9459–9474.

Liu, C.-L., & Chen, K.-C. (2019). Extracting the Gist of Chinese Judgments of the Supreme Court. *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, 73–82. https://doi.org/10.1145/3322640.3326715

Liu, Y. (2019, setembro 5). *Fine-Tune BERT for Extractive Summarization*. arXiv: 1903.10318 [cs]. https://doi.org/10.48550/arXiv.1903.10318

Liu, Y.-H., & Lin, Y.-F. (2024). LASG: Streamlining Legal Adjudication with AI-Enabled Summary Generation. Em Shahriar H., Ohsaki H., Sharmin M., Towey D., Majumder AKM.J.A., Hori Y., Yang J.-J., Takemoto M., Sakib N., Banno R. & Ahamed S.I. (Eds.), *Proc. - IEEE Annu. Comput., Softw., Appl. Conf., COMPSAC* (pp. 2242–2247). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/COMPSAC61105.2024.00360

Long, S., Tu, C., Liu, Z., & Sun, M. (2018, setembro 18). *Automatic Judgment Prediction via Legal Reading Comprehension*. arXiv: 1809.06537 [cs]. https://doi.org/10.48550/arXiv.1809.06537

Ma, L., Zhang, Y., Wang, T., Liu, X., Ye, W., Sun, C., & Zhang, S. (2021). Legal Judgment Prediction with Multi-Stage CaseRepresentation Learning in the Real Court Setting, 993–1002. https://doi.org/10.1145/3404835.3462945

Malkov, Y. A., & Yashunin, D. A. (2020). Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(4), 824–836. https://doi.org/10.1109/TPAMI.2018.2889473

Marino, G., Licari, D., Bushipaka, P., Comandé, G., & Cucinotta, T. (2023). Automatic Rhetorical Roles Classification for Legal Documents using LEGAL-TransformerOverBERT. Em Lagioia F., Mumford J., Odekerken D. & Westermann H. (Eds.), *CEUR Workshop Proc.* (Pp. 28–36, Vol. 3441). CEUR-WS.

Mihalcea, R., & Tarau, P. (2004, julho). TextRank: Bringing Order into Text. Em D. Lin & D. Wu (Eds.), *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (pp. 404–411). Association for Computational Linguistics. Obtido setembro 28, 2025, de https://aclanthology.org/W04-3252/

Moreno, G. C. d. L., de Souza, M. P. M., Hein, N., & Hein, A. K. (2022). *ALT: Um Software Para Análise de Legibilidade de Textos Em Língua Portuguesa* (3). https://doi.org/10.48550/ARXIV.2203.12135

Nallapati, R., Zhai, F., & Zhou, B. (2016, novembro 14). *SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents*. arXiv: 1611.04230 [cs]. https://doi.org/10.48550/arXiv.1611.04230

Nguyen, H., & Ding, J. (2023). Keyword-based Augmentation Method to Enhance Abstractive Summarization for Legal Documents. *Int. Conf. Artif. Intel. Law, ICAIL - Proc. Conf.*, 437–441. https://doi.org/10.1145/3594536.3595120

Nguyen, H.-T., Phi, M.-K., Ngo, X.-B., Tran, V., Nguyen, L.-M., & Tu, M.-P. (2024). Attentive Deep Neural Networks for Legal Document Retrieval. *Artificial Intelligence and Law*, *32*(1), 57–86. https://doi.org/10.1007/s10506-022-09341-8

Nishida, K., Nishida, K., Nagata, M., Otsuka, A., Saito, I., Asano, H., & Tomita, J. (2019, julho). Answering While Summarizing: Multi-task Learning for Multi-hop QA with Evidence Extraction. Em A. Korhonen, D. Traum & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational*

Linguistics (pp. 2335–2345). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1225

OECD. (2024, dezembro 10). *Do Adults Have the Skills They Need to Thrive in a Changing World?: Survey of Adult Skills 2023*. OECD. Obtido julho 13, 2025, de https://www.oecd.org/en/publications/do-adults-have-the-skills-they-need-to-thrive-in-a-changing-world_b263dc5d-en.html

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., ... Zoph, B. (2024, março 4). *GPT-4 Technical Report*. arXiv: 2303.08774 [cs]. https://doi.org/10.48550/arXiv.2303.08774

Page, M., McKenzie, J., Bossuyt, P., Boutron, I., Hoffman, T., Mulrow, C., Shamseer, L., Tetzlaff, L., Akl, E., Brennan, S., Chou, R., Glanville, J., Grimshaw, J., Hrobjartsson, A., Lalu, M., Li, T., Loder, E., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2024). Declaração PRISMA 2020: Uma Diretriz Atualizada Para Publicação de Revisões Sistemáticas. *Germinare*, *4*, 1–19. https://doi.org/10.5281/zenodo.13271469

Patil, S., Joshi, P., Ingle, A., Jayappa, A., & Ketkar, O. (2023). Text Extraction and Finetuning Transformers for Abstractive Summarisation. *Int. Conf. Comput., Commun., Control Autom., ICCUBEA*, 1–5. https://doi.org/10.1109/ICCUBEA58933.2023.10392203

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*. Improving language understanding with unsupervised learning. https://openai.com/index/language-unsupervised/

Rana, D., & Mehta, R. (2023). Research Challenges for Legal Document Summarization. Em Tomar G.S. & Bansal J. (Eds.), *Proc. - IEEE World Conf. Appl. Intell. Comput., AIC* (pp. 307–312). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/AIC57670.2023.10263906

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Savelka, J. (2023). Unlocking Practical Applications in Legal Domain: Evaluation of GPT for Zero-Shot Semantic Annotation of Legal Texts. *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, 447–451. https://doi.org/10.1145/3594536.3595161

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., ... Resnik, P. (2025, fevereiro 26). *The Prompt Report: A Systematic Survey of Prompt Engineering Techniques*. arXiv: 2406.06608 [cs]. https://doi.org/10.48550/arXiv.2406.06608

Singh, D. (2024). Legal Documents Text Analysis Using Natural Language Processing (NLP). *Int. Conf. Self Sustain. Artif. Intell. Syst., ICSSAS - Proc.*, 1302–1307. https://doi.org/10.1109/ICSSAS64001.2024.10760929

Smith, B. (2024, novembro 26). *Mistral 7B Explained: Towards More Efficient Language Models*. Towards Data Science. Obtido outubro 13, 2025, de https://towardsdatascience.com/mistral-7b-explained-towards-more-efficient-language-models-7f9c6e6b7251/

Takale, S., Payal, S., Jagtap, S., Jagtap, P., & Khan, A. (2023). Legal Data Assistive Tool Using Deep-Learning. *ICSCCC - Int. Conf. Secur. Cyber Comput. Commun.*, 791–796. https://doi.org/10.1109/ICSCCC58608.2023.10176784

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023, julho 19). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. arXiv: 2307.09288 [cs]. https://doi.org/10.48550/arXiv.2307.09288

Turing, A. M. (2009). Computing Machinery and Intelligence. Em R. Epstein, G. Roberts & G. Beber (Eds.), *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer* (pp. 23–65). Springer Netherlands. https://doi.org/10.1007/978-1-4020-6710-5_3

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023, agosto 2). *Attention Is All You Need*. arXiv: 1706.03762 [cs]. https://doi.org/10.48550/arXiv.1706.03762

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023, janeiro 10). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. arXiv: 2201.11903 [cs]. https://doi.org/10.48550/arXiv.2201.11903

Weizenbaum, J. (1966). ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine. *Commun. ACM*, *9*(1), 36–45. https://doi.org/10.1145/365153.365168

Wilson, J. D., Hoskin, N., & Nosek, J. T. (1993). The Benefits of Collaboration for Student Programmers. *Proceedings of the Twenty-Fourth SIGCSE Technical Symposium on Computer Science Education*, 160–164. https://doi.org/10.1145/169070.169383

Xiao, C., Hu, X., Liu, Z., Tu, C., & Sun, M. (2021, maio 9). *Lawformer: A Pre-trained Language Model for Chinese Legal Long Documents*. arXiv: 2105.03887 [cs]. https://doi.org/10.48550/arXiv.2105.03887

Xu, B., Dai, X., Liu, Z., Xie, H., Yi, X., Wang, S., Yan, Y., Yang, L., Gu, Y., & Yu, G. (2025, maio 1). *LegalDuet: Learning Fine-grained Representations for Legal Judgment Prediction via a Dual-View Contrastive Learning*. arXiv: 2401.15371 [cs]. https://doi.org/10.48550/arXiv.2401.15371

Yu, S., Gao, W., Qin, Y., Yang, C., Huang, R., Chen, Y., & Lin, C. (2025). IterSum: Iterative summarization based on document topological structure. *Information Processing and Management*, *62*(1), 103918. https://doi.org/10.1016/j.ipm.2024.103918

Zhong, H., Zhang, Z., Liu, Z., & Sun, M. (2019). *Open Chinese Language Pre-trained Model Zoo* (rel. téc.). https://github.com/thunlp/openclap

Zmiycharov, V., Tsonkov, T., & Koychev, I. (2024). EurLexSummarization – A New Text Summarization Dataset on EU Legislation in 24 Languages with GPT Evaluation. *Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria (CLIB 2024)*, 206–213. Obtido setembro 28, 2025, de https://aclanthology.org/2024.clib-1.22/

ANEXO A

Artigos científicos identificados na revisão de literatura

O presente anexo apresenta a lista completa dos artigos científicos considerados na revisão de literatura para este trabalho. Os artigos foram selecionados com base na sua relevância para o tema e encontram-se disponíveis em acesso aberto, mediante a utilização da Virtual Private Network (VPN) do Iscte-IUL.

(1) **Título:** Abstractive Text Summarization: A Transformer Based Approach
Citação: Kale e Deshmukh (2024)
Resumo: O estudo propõe uma abordagem combinada do uso dos modelos XLNet e BART na sumarização abstrativa de textos jurídicos longos, contribuindo para o avanço da sumarização automática no domínio jurídico.
Local: IEEE 9th International Conference for Convergence in Technology, I2CT — 2024.

(2) **Título:** Attentive Deep Neural Networks for Legal Document Retrieval
Citação: Nguyen et al. (2024)
Resumo: O estudo propõe o uso de redes neuronais profundas com mecanismos de atenção para melhorar a recuperação de textos legais extensos. Apresenta duas arquiteturas — Attentive CNN e Paraformer. Os resultados mostram que os métodos apresentados superam abordagens tradicionais, com o Paraformer a alcançar o melhor desempenho no conjunto de dados COLIEE.
Local: Artificial Intelligence and Law — 2024.

(3) **Título:** Automatic Rhetorical Roles Classification for Legal Documents using LEGAL-TransformerOverBERT
Citação: Marino et al. (2023)
Resumo: O estudo propõe o modelo LEGAL-ToBERT, que combina um codificador *Transformer* com o LEGAL-BERT, para identificar papéis retóricos em decisões judiciais. A abordagem hierárquica melhora significativamente os resultados face aos modelos isolados, captando melhor as relações entre frases longas e complexas em textos jurídicos.
Local: CEUR Workshop Proceedings — 2023.

(4) **Título:** Comprehensive Analysis of Text Summarization Techniques for Legal Documents
Citação: Deepika et al. (2024)
Resumo: O estudo analisa técnicas de sumarização aplicadas a documentos jurídicos indianos, propondo melhorias no desempenho de modelos como BART, Pegasus, BERT e TextRank. A avaliação, baseada em métricas ROUGE, demonstra bons resultados na normalização e sumarização de textos legais.
Local: 15th International Conference on Advances in Computing, Control, and Telecommunication Technologies, ACT — 2024.

(5) **Título:** Contributions to Legal Document Summarization: Judgments from the Portuguese Supreme Court of Justice
Citação: Dias, Ribeiro e Pinto (2024)
Resumo: O estudo investiga técnicas de sumarização automática aplicadas a documentos jurídicos portugueses, com o objetivo de apoiar os juízes do STJ na elaboração de resumos. Propõe-se avaliar diferentes métodos para melhorar a eficiência e a qualidade dos resumos, reduzindo o tempo despendido na análise manual de documentos extensos.
Local: OpenAccess Series in Informatics — 2024.

(6) **Título:** Enhancing Legal Document Summarization for Professionals: An Extractive Approach
Citação: Kalyan, Dwivedi e Sharma (2024)
Resumo: O estudo aborda a sumarização extrativa de documentos jurídicos, visando melhorar a extração de informação relevante. Utiliza um conjunto de dados anotado de decisões judiciais e respetivos resumos, explorando modelos de ML, redes neuronais e *Transformers*. A metodologia inclui a conversão de resumos abstrativos em extrativos e a avaliação com métricas como ROUGE e BERTScore.
Local: 15th International Conference on Computing Communication and Networking Technologies, ICCCNT — 2024.

(7) **Título:** EurLexSummarization - a new text summarization dataset on EU legislation in 24 languages with GPT evaluation
Citação: Zmiycharov, Tsonkov e Koychev (2024)
Resumo: O estudo apresenta um novo conjunto de dados multilingue para sumarização de textos jurídicos. Realizam-se experiências com técnicas extrativas e com modelos GPT. O trabalho contribui para melhorar a recuperação de informação em textos legais extensos.
Local: Proceedings of the International Conference Computational Linguistics in Bulgaria — 2024.

(8) **Título:** Exploring the State of the Art in Legal QA systems
Citação: Abdallah, Piryani e Jatowt (2023)
Resumo: O estudo apresenta uma revisão abrangente sobre sistemas de Question answering (QA) no domínio jurídico, analisando 14 conjuntos de dados de referência e os modelos de Deep Learning (DL) mais recentes. São discutidas arquiteturas, técnicas, desempenho e limitações destes modelos.
Local: Journal of Big Data —- 2023.

(9) **Título:** Harnessing Deep Learning for Effective Extractive Text Summarization: A Comparative Study
Citação: Alsekait et al. (2024)
Resumo: O estudo analisa diferentes métodos de sumarização de textos, incluindo modelos de ML, redes neuronais e modelos *Transformers* como o T5. A investigação envolve aquisição e pré-processamento de conjuntos de dados, treino e teste dos modelos, com avaliação através de métricas como ROUGE, visando comparar o desempenho das abordagens e concluir que a sumarização automática é mais eficaz com NLP, apesar dos desafios impostos pela diversidade de textos e contextos.

Local: 2nd International Conference of Intelligent Methods, Systems and Applications, IMSA — 2024.

(10) **Título:** Interactive Legal Assistance System using Large Language Models
Citação: Karthick et al. (2024)
Resumo: O estudo apresenta um *chatbot* baseado em RAG) para facilitar a compreensão de documentos legais indianos sobre segurança alimentar, direcionado a cidadãos sem formação jurídica. Desenvolvido com modelos de grande escala como GPT e Llama3, o sistema oferece uma interface intuitiva para esclarecer dúvidas ou sumarizar secções específicas. Esta abordagem promove a acessibilidade linguística e aproxima o público das complexidades legais.
Local: 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), I-SMAC 2024 - Proceedings — 2024.

(11) **Título:** IterSum: Iterative summarization based on document topological structure
Citação: Yu et al. (2025)
Resumo: O estudo apresenta o IterSum, uma *framework* de sumarização de textos. O texto é dividido em blocos, gerando resumos iterativos que se combinam progressivamente até formar o resumo final. Validado em nove tipos de conjuntos de dados públicos — incluindo notícias, bases de conhecimento e documentos jurídicos —, o modelo alcançou desempenho de estado da arte medido por *ROUGE*, sendo também confirmado por avaliações humanas.
Local: Information Processing and Management — 2025.

(12) **Título:** Keyword-based Augmentation Method to Enhance Abstractive Summarization for Legal Documents
Citação: Nguyen e Ding (2023)
Resumo: O estudo propõe a inserção de palavras-chave em modelos de ML para melhorar a sumarização de documentos jurídicos longos, guiando o modelo na identificação de informação relevante. Os resultados demonstram que a abordagem baseada em palavras-chave aumenta a qualidade da sumarização.
Local: 19th International Conference on Artificial Intelligence and Law, ICAIL 2023 - Proceedings of the Conference — 2023.

(13) **Título:** Large Language Models for Indian Legal Text Summarisation
Citação: Hemanth Kumar, Jayanth e Anand Kumar (2024)
Resumo: O estudo aborda a sumarização de decisões judiciais na área de NLP jurídico, utilizando modelos supervisionados e não supervisionados, tanto extrativos, como abstrativos, incluindo também Longformer e BERT-Legal, Pegasus. Adicionalmente, foram aplicados LLM com *prompt engineering* para sumarização. O trabalho inclui ainda um chatbot e uma aplicação web que permitem sumarizar documentos legais em múltiplas línguas, tornando o acesso à informação jurídica mais eficiente e acessível.
Local: Proceedings of CONECCT 2024 - 10th IEEE International Conference on Electronics, Computing and Communication Technologies — 2024.

(14) **Título:** LASG: Streamlining Legal Adjudication with AI-Enabled Summary Generation
Citação: Liu e Lin (2024)
Resumo: O estudo apresenta o LASG, uma ferramenta de sumarização automática de

decisões judiciais baseada em IA. Testes demonstram que o método permite identificar rapidamente os pontos-chave das decisões, promovendo maior acessibilidade e eficiência na análise de informação jurídica.

Local: Proceedings - 2024 IEEE 48th Annual Computers, Software, and Applications Conference, COMPSAC — 2024.

(15) **Título:** Legal Data Assistive Tool Using Deep-Learning

Citação: Takale et al. (2023)

Resumo: Ferramenta de sumarização abstrativa de documentos legais usando modelos pré-treinados. Classificação de sentenças com base em palavras-chave, POS tagging, segmentação de texto, posição e centralidade, com foco em entidades críticas (datas, pessoas, atos legais), permitindo a profissionais jurídicos extrair rapidamente os pontos principais dos documentos.

Local: Third International Conference on Secure Cyber Computing and Communication (ICSCCC) — 2023.

(16) **Título:** TextRank: Bringing Order into Text

Citação: Mihalcea e Tarau (2004)

Resumo: TextRank: modelo de *ranking* baseado em grafos para processamento de texto. Aplica extração de palavras-chave e frases de forma não supervisionada, obtendo resultados competitivos em *benchmarks* estabelecidos.

Local: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing — 2004.

(17) **Título:** Legal Documents Text Analysis Using Natural Language Processing (NLP)

Citação: Singh (2024)

Resumo: Este estudo explora o uso de técnicas de NLP para automatizar a análise de documentos legais na Índia, visando reduzir as pendências judiciais. Aplicando recuperação de informação, análise de sentimento, Optical Character Recognition (OCR), modelação de tópicos — Latent Dirichlet Allocation (LDA) —, e sumarização com . O sistema extrai informação-chave e gera resumos concisos, promovendo maior eficiência e equidade na justiça.

Local: 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS) — 2024.

(18) **Título:** Research Challenges for Legal Document Summarization

Citação: Rana e Mehta (2023)

Resumo: Os documentos de decisões judiciais são detalhados e contêm termos e códigos jurídicos, tornando a sua leitura e análise complexa. Este estudo aborda a sumarização automática, comparando métodos tradicionais, abordagens específicas do domínio jurídico e modelos baseados em *Transformers* aplicados a decisões indianas.

Local: IEEE World Conference on Applied Intelligence and Computing (AIC) — 2023.

(19) **Título:** Text Extraction and Finetuning Transformers for Abstractive Summarisation

Citação: Patil et al. (2023)

Resumo: Com o aumento do volume de dados, surgem documentos extensos que consomem tempo e energia para leitura. Este estudo propõe um sumarizador abstrativo de texto baseado em NLP, aplicado a artigos de notícias e documentos legais. O sistema

integra três modelos principais, ajustados com base em conjuntos de dados destes domínios, permitindo obter resumos concisos e relevantes.
Local: 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA) — 2023.

(20) **Título:** Too Legal; Didn't Read (TLDR): Summarization of Court Opinions
Citação: Ghimire, Shrestha e Edwards (2023)
Resumo: O acesso à justiça depende da compreensão de textos legais, que são extensos e complexos. Este estudo propõe uma abordagem de sumarização automática de textos legais baseada em NLP, permitindo que profissionais gerem resumos de forma rápida e eficiente, reduzindo esforço, tempo e custos associados à elaboração manual de resumos.
Local: Intermountain Engineering, Technology and Computing, IETC — 2023.

(21) **Título:** Topic-Driven Contractual Language Understanding and Summarization: An Integrated Approach for Simplifying Legal Documents
Citação: Harikrishnan, Malathi e Sundharakumar (2024)
Resumo: Este estudo apresenta uma ferramenta modular para simplificação e sumarização de contratos legais complexos, utilizando modelos baseados em transformadores. Com base no Contract Understanding Atticus Dataset (CUAD), o sistema segmenta os contratos em sete classes através de uma variante supervisionada do BERTopic, seguido de sumarização com o modelo Legal Pegasus, ajustado ao domínio jurídico. A abordagem integrada permite que utilizadores sem formação jurídica compreendam rapidamente os contratos e tomem decisões informadas perante a crescente complexidade documental.
Local: 4th International Conference on Intelligent Technologies (CONIT) — 2024.

(22) **Título:** Transfer Learning for Russian Legal Text Simplification
Citação: Athugodage, Mitrofanove e Gudkov (2024)
Resumo: Este estudo apresenta resultados inovadores na simplificação de textos jurídicos em russo, introduzindo o primeiro conjunto de dados paralelo baseado em publicações da Rossiyskaya Gazeta Legal Papers. Foram testadas três abordagens envolvendo as arquiteturas T5 e GPT, avaliadas com métricas como ROUGE, System output Against References and against the Input sentence (SARI) e BERTScore, bem como índices de legibilidade Flesch-Kincaid e Gunning Fog. Avaliações humanas, mostraram que o modelo GPT ajustado ao *corpus* específico apresenta melhor desempenho.
Local: Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI) @ LREC-COLING — 2024.

(23) **Título:** Unlocking Practical Applications in Legal Domain: Evaluation of GPT for Zero-Shot Semantic Annotation of Legal Texts
Citação: Savelka (2023)
Resumo: Este estudo avaliou a capacidade de um modelo GPT de última geração para realizar anotação semântica de pequenos trechos de textos jurídicos em configurações *zero-shot*. Os resultados mostram que o modelo apresenta desempenho notável em diferentes tipos de documentos, fornecendo orientações para académicos e profissionais do direito sobre a integração de LLM em fluxos de trabalho envolvendo anotação semântica de textos legais.

Local: ICAIL '23: Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law — 2023.

(24) **Título:** Walking a Tightrope - Evaluating Large Language Models in High-Risk Domain
Citação: Hung et al. (2023)
Resumo: Este estudo analisa o desempenho de LLM ajustados por instrução em domínios de alto risco, nomeadamente jurídico e médico, avaliando precisão factual e segurança. Experiências realizadas em seis conjuntos de dados de NLP, incluindo QA e sumarização, evidenciam limitações dos modelos atuais. Os resultados sublinham a necessidade de aperfeiçoar métricas específicas de domínio, visando melhorar a confiabilidade e segurança dos LLM.
Local: Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP — 2023.

ANEXO B

Prompts utilizados

Em seguida apresentam-se os exemplos de *prompts* utilizados para auxiliar o LLM na resposta ao utilizador:

Modelo de mensagem de sistema

> És um *chatbot* de assistência jurídica que ajuda o utilizador leigo a ter contacto com as leis portuguesas.
> Responde à pergunta com base no contexto legislativo existente na base de dados.
> Escreve frases completas com escrita e pontuação corretas.
> Se o contexto não contiver a resposta, apenas responde que não é possível, de momento, encontrar uma resposta.
> Não inventes respostas que não estejam no contexto.
> Os exemplos de diálogo apresentados servem apenas para demonstrar o formato e o estilo da resposta esperada, não devem ser repetidos.

Modelo de mensagem humana

> **Pergunta 1** — O que é o diploma Decreto-Lei n.º 15/2022?
> **Pergunta 2** — O que vai mudar com o diploma Decreto-Lei n.º 15/2022?
> **Pergunta 3** — Que vantagens traz o diploma Decreto-Lei n.º 15/2022?
> **Pergunta 4** — Quando entra em vigor o diploma Decreto-Lei n.º 15/2022?

Modelo de mensagem de um assistente de AI

As respostas encontram-se alinhadas com o conteúdo presente no resumo em linguagem clara, disponibilizado no Diário da República Eletrónico, e cujas perguntas foram realizadas na descrição anterior:

> **Resposta 1** — "Este decreto-lei estabelece a organização e funcionamento do sistema elétrico nacional (SEN) adequando-se às necessidades e desafios colocados pelos instrumentos estratégicos que norteiam a política energética da União Europeia e de Portugal.";
> **Resposta 2** — "As alterações agora introduzidas por este decreto-lei podem estruturar-se em cinco eixos fundamentais:
>
> ◦ A atividade administrativa de controlo prévio das atividades do SEN:
>
> Concentrar, num só diploma, as matérias relacionadas com a organização e funcionamento do SEN, até agora dispersas por vários diplomas legais;
> Simplificar o funcionamento do SEN, eliminando a distinção entre produção em regime ordinário e de produção em regime especial, com a inevitável eliminação de dois procedimentos distintos de licenciamento da atividade de produção de eletricidade;

∘ O planeamento das redes:

Maximização de todo o potencial de capacidade de receção da rede elétrica de serviço público (RESP), em linha com o interesse público da proteção dos consumidores que suportam os seus custos;
∘ A introdução de mecanismos concorrenciais para o exercício das atividades do SEN:

As atividades de comercializador de último recurso e de agregador de último recurso, bem como as de operador logístico de mudança de comercializador e a de emissão de garantias de origem passam, agora, a ser exercidas mediante licença a atribuir de modo concorrencial e transparente;
∘ A participação ativa dos consumidores, na produção e nos mercados:

Este decreto-lei impõe a instalação de contadores e redes inteligentes e assegura, através da criação da figura do agregador, a eliminação das barreiras à participação nos mercados de eletricidade;
Reforço dos direitos de informação dos consumidores, designadamente, através da concentração da informação essencial, que atualmente se encontra dispersa por várias entidades e diversos locais, no sítio na Internet da Entidade Reguladora dos Serviços Energéticos;
∘ O enquadramento e densificação legislativa de novas realidades como o reequipamento, os híbridos e o armazenamento:

Criação ou densificação do enquadramento jurídico de realidades inovadoras e, bem assim, do estabelecimento de um quadro jurídico adequado aos projetos-piloto de inovação e desenvolvimento através da criação de três Zonas Livres Tecnológicas (ZLT).
Foi desenvolvido o Plano Nacional Energia e Clima 2030 (PNEC 2030), que constitui o principal instrumento de política energética e climática nacional para a próxima década rumo a um futuro neutro em carbono.
O PNEC 2030 estabelece metas, objetivos e respetivas políticas e medidas em matéria de redução de emissões de gases com efeito de estufa.
O Pacto Ecológico Europeu estabeleceu o roteiro para a redução de emissões em, pelo menos, 55% até 2030.

Resposta 3:

"∘ Promove-se ativamente a incorporação de energias de fontes renováveis;
∘ Maior eficiência energética;
∘ Maior segurança energética;
∘ Aumento do mercado interno e investigação, inovação e competitividade, bem como uma abordagem clara para o alcance dos referidos objetivos e metas.";

Resposta 4 – "Este Decreto-Lei entra em vigor no dia 15 de janeiro de 2022."

Modelo de mensagem de sistema (fase mapeamento)

O *prompt* utilizado na fase de mapeamento permite realizar um processo de extração de conteúdo controlado, no qual é indicado ao modelo LLM a identificação de 10 passagens diretamente extraídas do texto segmentado, e que melhor resumem o seu conteúdo. Possui instruções explícitas de modo a que se possa assegurar a fidelidade do conteúdo, assim como a exclusão de parágrafos que contenham a menção "Revogada."

"Extrai até 10 frases do seguinte texto que melhor resumam o seu conteúdo.
REGRAS:
- Só escolhe frases exatas do texto.
- Ignora os parágrafos com: (Revogada.)
- Não inventes ou parafraseies.
Texto:
Resumo extrativo: "

Modelo de mensagem de sistema (fase redução)

Em seguida, foi realizada uma redução global do conjunto de frases selecionadas na fase anterior, com o objetivo de produzir uma sumarização do conteúdo mais objetiva e concisa. Desta forma, apresenta-se o *prompt* utilizado na fase de redução:

"Aqui tens um conjunto de frases selecionadas de várias partes de um documento.
Seleciona até 30 frases que, no conjunto, resumam melhor o documento completo.
Regras:
- Mantém apenas frases já dadas.
Texto:
Resumo global extrativo: "

ANEXO C

"Alucinações" do sistema LexClara

Esta secção tem como objetivo apresentar algumas respostas incorretas, ou que não encontram base no contexto disponibilizado: as "'alucinações".

As imagens seguintes mostram respostas em que o modelo apresentou informação que era inexistente nos documentos disponibilizados ou gerou texto sem qualquer suporte textual.

A análise associada descreve as condições de ocorrência e as possíveis causas e medidas de mitigação que sejam possíveis de implementar.

Caso n.º 1: Decreto-Lei n.º 115/2023, de 15 de dezembro.

Pergunta — Quando entra em vigor o diploma Decreto-Lei n.º 115/2023, de 15 de dezembro?

Resposta do LLM — Este decreto-lei entra em vigor no dia 16 de dezembro de 2023.

Resposta esperada — Este decreto-lei entra em vigor no dia 1 de janeiro de 2024.

Análise:

A resposta do sistema LexClara está errada. A resposta passada ao utilizador tem como base o n.º 2, artigo 11.º:

> "Artigo 11.º
> **Entrada em vigor**
> 1 — [...]
> 2 — O disposto no artigo 3.º produz efeitos no dia seguinte ao da publicação do presente decreto-lei. "

Enquanto que o conteúdo correto encontra-se no n.º 1, artigo 11.º:

> "Artigo 11.º
> **Entrada em vigor e produção de efeitos**
> 1 — O presente decreto-lei entra em vigor no primeiro dia do mês seguinte ao da sua publicação.
> 2 — [...] "

Caso n.º 2: Decreto-Lei n.º 70-B/2021, de 6 de agosto.

Pergunta — Quando entra em vigor o diploma Decreto-Lei n.º 70-B/2021, de 6 de agosto?

Resposta do LLM — O diploma Decreto-Lei n.º 70-B/2012, de 6 de agosto entrou em vigor no dia seguinte à sua publicação, conforme artigo 34º do referido diploma.

Segmentos de Texto Recuperados – Não foram encontrados documentos que correspondem ao critério.

Análise:

Este caso representa uma "alucinação" do sistema LexClara, uma vez que elabora uma resposta verosímil para o utilizador, mas sem qualquer base factual. O campo segmentos de texto recuperados, mostra o(s) segmento(s) de texto que dizem respeito ao diploma mencionado na pergunta (Decreto-Lei n.º 70-B/2021), que se encontram carregados na base de dados vetorial. Constata-se que não existem segmentos que correspondam ao critério diploma. Ainda assim, o modelo oferece-nos uma resposta plausível a essa ausência de contexto.

A análise qualitativa revela que o critério de aceitação das respostas não se deve basear exclusivamente em métricas quantitativas. A validação humana tem um papel relevante na interpretação dos resultados e contribui para o aprimoramento dos modelos de linguagem natural, nomeadamente através da abordagem da aprendizagem por reforço (*reinforcmement learning*).

Em conclusão, as análises anteriores demonstram que as respostas plausíveis podem surgir, mesmo na ausência de base factual, o que reforça a importância de uma validação humana. A harmonização entre as métricas quantitativas e a avaliação qualitativa é essencial para garantir uma coerência e fiabilidade do texto legislativo assistido por GenAI.

ANEXO D

Respostas do sistema LexClara

Segue-se a análise a duas situações de pergunta-resposta obtidas nos testes realizados ao sistema LexClara, recaíndo a escolha da métrica de similaridade de coseno:

Com o valor máximo da similaridade de coseno de 0,96:

Pergunta — O que vai mudar com o diploma Decreto-Lei n.º 63/2023, de 31 de julho?
Resposta gerada — "Este decreto-lei cria um regime extraordinário e temporário de revalidação de títulos de condução. Os titulares de títulos de condução caducados por via legal poderão proceder à sua revalidação sem submissão a exame especial. Esta revalidação está condicionada à apresentação de atestado médico para os condutores com mais de 60 anos. Este regime aplica-se aos títulos de condução emitidos antes de 1 de janeiro de 2008, cujos prazos de validade constantes dos respetivos documentos físicos não correspondam ao prazo legalmente previsto e em vigor, e que habilitem à condução de veículos das categorias AM, A1, A2, A, B1, B, BE e de veículos agrícolas. Este regime aplica-se por um período de um ano a contar do dia da sua entrada em vigor."
Resposta esperada — "Passa a ser possível aos titulares de título de condução caducado por via legal proceder à sua revalidação sem submissão a exame especial.
Esta revalidação fica condicionada à apresentação de atestado médico para os condutores com mais de 60 anos.
O regime previsto no presente decreto-lei aplica-se aos títulos de condução emitidos antes de 1 de janeiro de 2008, que habilitem à condução de veículos das categorias AM, A1, A2, A, B1, B, BE e de veículos agrícolas.
O regime excecional previsto no presente decreto-lei aplica-se por um período de um ano a contar do dia da sua entrada em vigor."

No exemplo apresentado verifica-se uma forte correspondência textual entre a resposta gerada pelo sistema LexClara e o texto de referência. O LLM conseguiu captar com grande precisão o conteúdo essencial do diploma e, cujas diferenças não afetam a exatidão da resposta. Este resultado demonstra a boa capacidade do modelo em reproduzir com exatidão e rigor o conteúdo legislativo.

Com o valor mínimo da similaridade de coseno de 0,36:

Pergunta – O que vai mudar com o diploma Decreto-Lei n.º 27/2021, de 16 de abril?
Resposta gerada – "Não há informação relevante nos diplomas selecionados sobre o Decreto-Lei n.º 27/2021, de 16 de abril."
Resposta esperada – "As instituições de ensino superior têm o direito e o dever de participar em atividades de ligação à sociedade, nomeadamente na divulgação e transferência de

conhecimento, ou de valorização social e económica do conhecimento científico. Esta missão deve ser concretizada em cooperação com as empresas, as autarquias locais, as organizações dos setores social, cultural e de saúde, a Administração Pública, e a rede de centros de gestão direta e participada do Instituto de Emprego e Formação Profissional. São criadas redes colaborativas de instituições de ensino superior em parceria com empregadores, promovendo formação inicial superior, projetos e serviços de inovação e empresariais, entre outros. Cabe à Direção-Geral do Ensino Superior desenvolver anualmente um relatório sobre a oferta e procura de formações curtas de âmbito profissional, designadamente cursos técnicos superiores profissionais e formações pós-graduadas. É atribuído suplemento remuneratório a dirigentes de instituições de investigação e desenvolvimento ou outras unidades orgânicas da instituição de ensino superior, ainda que não autónomas. A elaboração da dissertação ou do trabalho de projeto e a realização do estágio são orientadas por doutores, por especialistas de reconhecida experiência e competência profissional ou por especialistas considerados como tal pelo órgão científico estatutariamente competente do estabelecimento de ensino superior, nacional ou estrangeiro. Pode ser considerado especialista quem seja detentor de um grau académico, e: - Exerça ou tenha exercido profissão na área do ciclo de estudos em causa, possuindo, no mínimo, 10 anos de experiência profissional nessa área; - Apresente um currículo profissional de qualidade e relevância comprovadas, devidamente confirmado e aceite pelo órgão cientificamente competente da instituição de ensino superior; - Não seja titular de contrato por tempo indeterminado com uma instituição de ensino superior. Os membros do júri devem ser especialistas no domínio em que se insere a dissertação, o trabalho de projeto ou o relatório de estágio. O título de especialista é conferido por um certificado onde consta a identificação de todas as instituições de ensino superior que integraram o consórcio."

O valor mínimo revela a divergência entre a resposta gerada e o texto de referência do conjunto verdade, não conseguindo estabelecer uma correspondência com o texto esperado.

Neste exemplo, o diploma não se encontra carregado na base de dados, sendo a resposta do sistema LexClara incorreta. Este enviesamento por omissão pode ser resolvido não só por um carregamento do *corpus* da legislação, mas também por uma afinação adicional ao *prompt*.

Esta comparação dos valores máximo e mínimo da similaridade de coseno revela o comportamento variável das respostas do modelo face à existência, ou não, de informação de suporte, assim como da necessidade de garantir a monitorização e a avaliação da qualidade das respostas, de modo a manter a fiabilidade do sistema em contexto real de consulta por parte do cidadão.