UNIDADE DE ENSINO SUPERIOR DOM BOSCO
CURSO DE ENGENHARIA DE SOFTWARE

Emanuel Luis Carvalho de Sousa Filho

Sistema Inteligente para Análise de Similaridade em Laudos de Necrópsia Utilizando Modelos de Linguagem e Embeddings Semânticos

São Luís
2025

Emanuel Luis Carvalho de Sousa Filho

Sistema Inteligente para Análise de Similaridade em Laudos de Necrópsia Utilizando Modelos de Linguagem e Embeddings Semânticos

Monografia apresentada ao Curso de Graduação de Engenharia de Software da Unidade de Ensino Superior Dom Bosco como requisito parcial para a obtenção do grau de Bacharel em Engenharia de Software.

Orientador: Prof. Dr. Giovanni Lucca França da Silva

São Luís

2025

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AIH	Autorização de Internação Hospitalar
ANN	*Approximate Nearest Neighbor*
BERT	*Bidirectional Encoder Representations from Transformers*
BLEU	*Bilingual Evaluation Understudy*
BoW	*Bag of Words*
CBOW	*Continuous Bag of Words*
CID-10	Classificação Internacional de Doenças 10ª revisão
CNN	Convolutional Neural Network
CoT	*Chain-of-Thought*
DL	*Deep Learning*
ELMo	*Embeddings from Language Model*
ETL	Extract Transform Load
FAISS	*Facebook AI Similarity Search*
GloVe	*Global Vectors for Word Representation*
GPT	*Generative Pre-trained Transformer*
GPU	Unidade de Processamento Gráfico
GRU	*Gradient Rectified Unlearning*
IA	Inteligência Artificial
LDA	*Latent Dirichlet Allocation*
LGPD	Lei Geral de Proteção de Dado
LLaMA	*Large Language Model Meta AI*
LLM	*Large Language Model*
LoRA	Low-Rank Adaptation
LSTM	*Long Short-Term Memory Networks*

ML	*Machine Learning*
NMF	*Non-Negative Matrix Factorization*
NPMI	*Normalized Pointwise Mutual Information*
OOV	*Out of Vocabulary*
PEFT	*Parameter-Éfficient Fine-Tuning*
PLN	Processamento de Linguagem Natural
QA	*Question & Answer*
RAG	Recuperação Aumentada por Geração
RBC	Raciocínio Baseado em Casos
RLHF	*Reinforcement Learning from Human Feedback*
RNN	Rede Neural Recorrente
ROUGE	*Recall-Oriented Understudy for Gisting Evaluation*
SBERT	*Sentence-BERT*
SG	*Skip-gram*
SNOMED-CT	Systematized Nomenclature of Medicine - Clinical Terms
T5	Text-to-Text Transfer Transformer
TF-IDF	*Term Frequency-Inverse Document Frequency*

SUMÁRIO

1 INTRODUÇÃO

Os laudos de necrópsia desempenham um papel central tanto no sistema de justiça quanto na saúde pública, atuando como instrumentos técnicos capazes de identificar o corpo, esclarecer a causa da morte e caracterizar lesões. Para juízes, promotores, advogados e investigadores, esses documentos fornecem a base para decisões judiciais e garantem a preservação da cadeia de custódia (França, 2017; Brasil, 2024b). Paralelamente, no âmbito da saúde pública, especialmente por meio dos Serviços de Verificação de Óbito (SVOs), os laudos oferecem dados confiáveis para sistemas de informação sobre mortalidade, apoiando a vigilância epidemiológica e a formulação de políticas de saúde (Brasil, 2006; Brasil, 2025). Dessa forma, eles constituem um elo entre o direito e a saúde coletiva, sendo fundamentais para decisões precisas e para a proteção social.

Apesar de sua relevância, a análise desses documentos enfrenta desafios significativos. A qualidade e padronização dos laudos podem ser comprometidas por vícios, imprecisões e subjetividade, enquanto a interpretação da linguagem técnica e a dispersão de informações dificultam a adoção de tecnologias, como a Inteligência Artificial (IA) (Ferreira, Pinheiro e Fernandes, 2025; Pereira, 2013; Rego, 2025). Tais obstáculos impactam não apenas a eficácia das decisões judiciais e das políticas de saúde, mas também levantam questões éticas e sociais, reforçando desigualdades e comprometendo a confiança institucional (Amado, 2024; França, 2017; Silva e Pazin-Filho, 2025).

Na prática pericial, essas dificuldades se traduzem em problemas concretos. A falta de padronização e clareza nos relatórios, combinada com déficits de profissionais e a insuficiência de formação contínua, compromete a qualidade técnica das conclusões (França, 2017; Pereira, 2013). Além disso, os institutos médico-legais operam sobrecarregados, com recursos limitados e metodologias muitas vezes ultrapassadas, fragilizando a confiabilidade dos resultados. A integração de novas tecnologias também enfrenta desafios éticos e a necessidade de validação científica (Ferreira, Pinheiro e Fernandes, 2025; Silva e Pazin-Filho, 2025).

As consequências para a sociedade são diretas e significativas. Morosidade e falhas na análise de laudos podem atrasar investigações, gerar erros judiciais e abalar a confiança pública na justiça (França, 2017). No âmbito da saúde pública, a ausência de dados conclusivos inviabiliza políticas eficazes e pode retardar diagnósticos e tratamentos (Pereira, 2013). Além disso, vieses em sistemas automatizados e a incerteza quanto à responsabilidade legal reforçam dilemas éticos e desigualdades sociais (Barros et al., 2025; Nascimento et al., 2024). Assim, a eficiência pericial não é apenas uma questão técnica, mas um elemento crucial para a justiça, a saúde coletiva e a confiança institucional.

Diante desse cenário, a IA surge como uma ferramenta promissora para apoiar a análise de documentos médico-legais. Por meio do PLN, algoritmos são capazes de compreender,

interpretar e manipular textos complexos, permitindo extrair informações relevantes de laudos de necrópsia e codificar automaticamente causas de morte segundo padrões como o CID-10 (Piraianu et al., 2023; Ferreira et al., 2023; Rai e Rai, 2022). Essa abordagem é especialmente relevante, considerando que os laudos combinam terminologia formal e narrativa, exigindo métodos capazes de estruturar dados e subsidiar decisões judiciais e médicas (Farias e Pinho, 2016; Manaka, Zyl e Kar, 2022).

Modelos pré-treinados, como BERT e ELMo, aliados a técnicas de reconhecimento de entidades nomeadas (NER), ampliam a capacidade de extração e classificação automatizada das informações contidas nos laudos. Complementarmente, LLMs (ou Modelos de Linguagem de Grande Porte), baseados em arquiteturas *transformer*, oferecem compreensão semântica profunda, raciocínio complexo e processamento de longos contextos, tornando-os adequados para sumarização, classificação e análise de textos técnico-científicos (Nascimento et al., 2024; Wang et al., 2023; Yin et al., 2025). Contudo, seu uso isolado apresenta limitações, incluindo imprecisões factuais e a geração de informações incorretas (*hallucinations*) (Jurafsky e Martin, 2025; Xiong et al., 2024).

Para mitigar essas limitações, a arquitetura de RAG combina LLMs com mecanismos de recuperação de informações externas, fornecendo contexto adicional e documentos confiáveis antes da geração de respostas (Amugongo et al., 2025; Rego, 2025; Zhang e Zhang, 2025). Essa integração aumenta a precisão, garante rastreabilidade e transparência, oferecendo uma solução promissora para a análise automatizada de laudos de necrópsia, um campo ainda pouco explorado, mas de elevada relevância jurídico-médica.

A problemática central deste estudo reside na dificuldade da Medicina Legal em comparar e identificar laudos de necrópsia devido à variabilidade terminológica, subjetividade dos peritos e elevado volume de documentos. A ausência de padronização compromete a precisão, confiabilidade e objetividade das análises, impactando diretamente a investigação das causas de morte, a gestão de informações em serviços periciais e a eficácia dos processos judiciais (Pereira, 2013; França, 2017; Brasil, 2024b).

1.1 Justificativa

A justificativa da pesquisa sobre o desenvolvimento de um sistema inteligente para a análise de similaridade em laudos de necrópsia encontra sustentação em múltiplas dimensões, abrangendo relevância social, científica e prática. No contexto da medicina legal, o crescente volume de laudos, aliado à diversidade terminológica e à complexidade da redação técnico-científica, torna a identificação de casos semelhantes um processo moroso, sujeito a falhas humanas e dependente da experiência individual dos peritos (Farias e Pinho, 2016; França, 2017; Pereira, 2013). Nesse cenário, a aplicação de técnicas de inteligência artificial, como LLMs e *embeddings* semânticos, oferece uma solução capaz de transformar textos complexos em representações computacionais precisas e comparáveis, criando um modelo escalável e

reprodutível para análise de similaridade (Ferreira, Pinheiro e Fernandes, 2025; Lorenzi, 1998). Socialmente, essa abordagem contribui para a melhoria da eficiência pericial, reduzindo a sobrecarga de trabalho manual, minimizando erros e padronizando análises, de modo que as informações críticas sobre causas de morte possam ser registradas com maior confiabilidade e rapidamente incorporadas a investigações e políticas de saúde pública (Barros et al., 2025; Pereira, 2013; Rajasekar e Vezhaventhan, 2024).

Do ponto de vista científico, o estudo oferece avanços metodológicos significativos. A combinação de *embeddings* semânticos com LLMs permite não apenas a detecção de similaridades entre documentos complexos, mas também a geração de interpretações em linguagem natural, fornecendo justificativas compreensíveis para decisões periciais (Barros et al., 2025; Ferreira, Pinheiro e Fernandes, 2025; Nascimento et al., 2024). Isso cria oportunidades para o desenvolvimento de novos métodos de recuperação de informações e comparação semântica em contextos clínicos e legais, servindo como referência para futuras aplicações de inteligência artificial na saúde e perícia forense (Lorenzi, 1998; Silva e Pazin-Filho, 2025). Além disso, a automatização do processo promove transparência e confiabilidade nas análises, possibilitando a identificação rápida de padrões recorrentes em causas de morte, contribuindo para políticas preventivas, gestão eficiente de recursos e maior segurança na tomada de decisão médico-legal (Amado, 2024; Pereira, 2013).

Dessa forma, a pesquisa justifica-se pela convergência entre inovação tecnológica, impacto social e avanço do conhecimento científico, demonstrando como ferramentas de inteligência artificial podem integrar eficiência, precisão e explicabilidade à prática pericial, com respaldo em evidências da literatura técnico-científica e estudos prévios em inteligência artificial aplicada à medicina legal.

1.2 Objetivos

Desenvolver uma metodologia para a análise e classificação de laudos de necrópsia, utilizando *embeddings* semânticos e LLMs integrados à arquitetura RAG. Para isso, propõe-se a criação de um sistema capaz de identificar similaridades entre lesões, recuperar documentos relevantes e gerar interpretações em linguagem natural, permitindo comparar diferentes modelos de LLMs e validar os resultados com base na análise de peritos humanos.

Destacam-se como objetivos específicos deste trabalho:

- Implementar um pipeline de pré-processamento e vetorização de laudos de necrópsia, gerando embeddings semânticos para representar as lesões descritas.
- Desenvolver um módulo de recuperação de documentos similares utilizando um banco de dados vetorial e métricas de similaridade.

- Integrar o módulo de recuperação a um LLM através da arquitetura RAG, permitindo a geração de justificativas em linguagem natural para os casos similares recuperados..
- Avaliar comparativamente o desempenho de diferentes LLMs (ex: GPT-4, LLaMA 3, Claude) na tarefa de explicitação das similaridades, validando a qualidade das explicações com um perito médico-legista.

1.3 Hipóteses

- Hipótese Nula (H0): Um sistema baseado em *embeddings* semânticos e LLMs integrados à arquitetura RAG não é capaz de identificar de forma precisa e consistente laudos de necrópsia semelhantes, apresentando resultados que não se aproximam da análise realizada por peritos humanos.
- Hipótese Principal (H1): A utilização combinada de *embeddings* semânticos e LLMs com uso de RAG permite identificar e classificar laudos de necrópsia semelhantes com precisão e consistência, aproximando os resultados da análise realizada por peritos humanos.
- Hipótese Alternativa (H2): A aplicação do LLM para análise interpretativa dos casos recuperados melhora a compreensão das similaridades, fornecendo explicações em linguagem natural que apoiam a tomada de decisão pericial.

1.4 Organização do Trabalho

Este trabalho está estruturado em seis capítulos, organizados para apresentar o conteúdo de forma clara e sequencial.

O Capítulo 2, Trabalhos Relacionados revisa metodologias existentes na aplicação de Inteligência Artificial para análise de documentos médico-legais, com foco em técnicas de Processamento de Linguagem Natural (PLN), Modelos de Linguagem de Grande Porte (LLMs), *embeddings* semânticos e Recuperação de Casos (RAG). São discutidos estudos sobre extração automática de informações, análise de similaridade de laudos e recuperação de casos em contextos forenses e clínicos, destacando as contribuições, limitações e resultados obtidos por abordagens similares. Esse capítulo fornece o panorama necessário para compreender os desafios da identificação automática de laudos de necrópsia semelhantes e embasa a relevância do desenvolvimento de sistemas inteligentes para apoio pericial.

O Capítulo 3, Fundamentação Teórica apresenta os conceitos essenciais para o desenvolvimento do sistema proposto, incluindo PLN, vetorização de textos, *embeddings* semânticos, representação computacional de informações complexas e o papel de LLMs na interpretação contextual de laudos médico-legais. Além disso, discute-se a aplicação de RAG e técnicas de análise de similaridade em domínios de alta criticidade, evidenciando as vantagens de combinar representações vetoriais com geração interpretativa em linguagem natural. O capítulo fornece

o embasamento necessário para compreender como essas tecnologias podem ser aplicadas para identificar e classificar automaticamente laudos semelhantes, alinhando-se aos objetivos de desenvolver um sistema preciso, escalável e interpretável.

No Capítulo 4, Metodologia Proposta, são detalhadas as etapas de desenvolvimento do sistema inteligente, desde a coleta, pré-processamento e normalização dos laudos de necrópsia, passando pela geração de *embeddings* semânticos e armazenamento em banco de dados vetorial, até a aplicação do LLM para análise interpretativa dos casos recuperados. São descritos os procedimentos para classificar automaticamente laudos semelhantes, gerar relatórios em linguagem natural e validar a precisão e confiabilidade do sistema, comparando os resultados com análises humanas. Essa metodologia está alinhada aos objetivos gerais e específicos, abordando a criação de um sistema capaz de apoiar peritos e operadores do direito na identificação de padrões médico-legais.

O Capítulo 5, Resultados e Discussão apresenta a avaliação do sistema, analisando métricas de desempenho como precisão, revocação e F1-score, e discute a capacidade do modelo em identificar laudos semelhantes de maneira confiável. Também é abordada a interpretação fornecida pelo LLM sobre os casos recuperados, demonstrando como o sistema oferece justificativas em linguagem natural para apoiar decisões periciais. O capítulo destaca o potencial da ferramenta, suas limitações e as implicações práticas na automação da análise de laudos, na eficiência do trabalho pericial e na padronização das análises médico-legais.

Por fim, o Capítulo 6, Conclusão reúne as considerações finais sobre o trabalho, evidenciando suas contribuições acadêmicas, sociais e científicas. Ressalta-se a relevância da pesquisa para a melhoria da eficiência pericial, a redução de erros humanos, a análise de grandes volumes de dados e a identificação de padrões críticos em investigações médico-legais. O capítulo apresenta propostas de trabalhos futuros e destaca o impacto do sistema na integração de tecnologia avançada à prática médico-legal, fortalecendo a confiabilidade, a interpretabilidade e a escalabilidade das análises de laudos de necrópsia.

2 TRABALHOS RELACIONADOS

Neste capítulo são apresentados resumos dos trabalhos relacionados à análise de laudos de necrópsia utilizando técnicas de processamento de linguagem natural. Nos últimos anos, diversas pesquisas têm sido desenvolvidas com o objetivo de melhorar a precisão e a eficiência na identificação de similaridades entre laudos médicos. Historicamente, a análise dependia de métodos manuais ou regras rígidas de indexação, mas atualmente soluções baseadas em *embeddings* semânticos, ontologias e modelos de linguagem de grande porte (LLMs) têm sido introduzidas. No contexto dos laudos de necrópsia, tais técnicas são particularmente relevantes, uma vez que lidam com documentos técnicos e sensíveis, fundamentais para o trabalho pericial. Assim, os trabalhos a seguir são organizados de forma a destacar as contribuições que ajudam a compreender limites, possibilidades e aplicações práticas, servindo como base para a metodologia proposta neste estudo.

O primeiro estudo aqui considerado projetou e implementou um protótipo baseado em Raciocínio Baseado em Casos (RBC) para reaproveitar conhecimento histórico de internações e apoiar decisões operacionais sobre o bloqueio de AIH's a proposta seguiu o ciclo clássico do RBC (representação, indexação, recuperação, reutilização, revisão e retenção), criou uma meta-estrutura (tabela "Problemas") e contou com indexação manual orientada por especialistas, o que viabilizou a automação parcial, mas também deixou explícito que a eficácia da recuperação depende criticamente da qualidade da indexação — apontada pelo autor como o principal gargalo do sistema (Lorenzi, 1998).

Em sequência, uma dissertação que avaliou um protótipo RBC aplicado à recuperação de jurisprudência reforçou esse diagnóstico ao incorporar um cálculo de similaridade (baseado em contagem de palavras) e indexação especializada sobre 250 casos: houve ganho de precisão frente ao modelo clássico, porém a avaliação destacou limitações de escalabilidade e generalização devido à indexação manual e à amostra restrita, recomendando a ampliação do universo de casos e a automação da indexação (Oliveira, 2011).

Uma contribuição metodológica relevante veio da tese que propôs algoritmos de Propagação em Grafos Bipartidos (PBG, TPBG, oPBG): ao representar documento–termo em grafos e propagar rótulos, os autores demonstraram competitividade frente a LDA/NMF e ganhos em tarefas semissupervisionadas, utilizando pré-processamento clássico e avaliando tópicos por NPMI; a principal limitação identificada foi a complexidade matemática do arcabouço e a necessidade de heurísticas para detectar deriva conceitual em fluxos textuais (Faleiros, 2016).

No mesmo ano, um estudo de modelagem de conhecimento elaborou a "Ontomédico legal" usando *Methontology* e *Protégé* a partir de 300 laudos médico-legais; a ontologia formalizou classes, propriedades e relações, melhorou a organização e agilidade na recuperação

da informação, mas também evidenciou a dificuldade de conciliar a terminologia formal com a miscelânea de linguagem natural presente nos laudos, apontando para a necessidade de validação contínua por peritos (Farias e Pinho, 2016).

Três anos depois, um estudo sobre resumo automático de textos jurídicos que combina grafos de sentenças, *embeddings* especializados (*Lex2Vec*) e seleção (*k-means*) demonstrou que a união entre representações vetoriais e estruturas gráficas favorece a extração de resumos relevantes, embora a ambiguidade semântica exija vocabulário controlado e curadoria — uma estratégia que pode inspirar um módulo de sumarização e explicação para apoiar a triagem pericial em laudos de necrópsia (Sousa e Prata, 2019).

Avançando para aplicações em texto clínico, um estudo que treinou redes neurais convolucionais com cerca de 30.000 prontuários em português aplicou técnicas de PLN no pré-processamento e alcançou resultados práticos (F-score $\approx$ 63,9%; precisão $\approx$ 72,7%), com desempenho particularmente alto em classes bem representadas, mas mostrou fragilidade perante variação textual e classes raras — lições que indicam que a simplicidade de modelos extractivos pode ser insuficiente para a heterogeneidade dos laudos de necrópsia (Rocha et al., 2022).

Em contexto de interoperabilidade, o projeto de mapeamento entre CID-10 e SNOMED-CT organizou processos ETL (*White Rabbit*, *OpenRefine*, *Pentaho*) e validação por especialistas, mapeando inicialmente 212 termos e revelando problemas recorrentes como termos sem correspondência exata e aproximações semânticas que demandam curadoria humana; esse trabalho sublinha a importância de incorporar camadas de normalização terminológica quando se pretende alinhar texto livre a códigos clínicos (Gualdani et al., 2024).

Uma investigação aplicada à área jurídica usou Similaridade Textual Semântica com *Word2Vec* (*skip-gram*, incluindo modelos pré-treinados *LegalNLP*) e *clustering* (*k-means*) sobre aproximadamente 3.160 ementas, produzindo agrupamentos úteis para identificar divergências jurisprudenciais; os autores salientam, entretanto, que ementas curtas são mais facilmente representadas por *embeddings*, enquanto laudos longos e heterogêneos exigem estratégias como *chunking*, agregação de *embeddings* e sumarização para garantir representações comparáveis (Castro e Neves, 2024).

No plano da privacidade e uso ético de textos clínicos, um estudo observacional aplicou um modelo NER generalista (GLiNER) sem *fine-tuning* a 27.540 resumos de alta (2017–2023), avaliando efeitos da anonimização com ROUGE, BLEU, BERTScore e revisão humana; os resultados mostraram baixa taxa de falha (0,5%) e preservação aceitável da utilidade textual (BERTScore mediano $\approx$ 0,76), ao passo que destacaram a necessidade de definir pontos de corte e limites de comprimento que equilibrem privacidade e utilidade (Silva e Pazin-Filho, 2025).

Por fim, complementando a linha tecnológica, uma investigação sobre LLMs e arquite-

turas RAG combinou extração semântica, indexação vetorial (FAISS), *embeddings* (all-MiniLM-L6-v2) e LLMs (GPT-4/Gemini) para análise de bulas, mostrando que RAG efetivamente recupera contexto documental e permite a geração fundamentada de respostas — ainda que os autores ressaltem desafios de validação regulatória em cenários sensíveis — o que aponta para RAG como a arquitetura adequada para recuperar e contextualizar trechos de laudos com rastreabilidade das fontes (Rego, 2025).

Este trabalho amplia os estudos já existentes ao direcionar o foco especificamente para a análise de similaridade de lesões descritas em laudos de necrópsia, um campo pouco explorado na literatura. A proposta integra técnicas que antes foram tratadas de maneira isolada — como *embeddings* semânticos, indexação vetorial, ontologias e mapeamentos terminológicos — em um sistema voltado para capturar as nuances da linguagem médico-legal e garantir a consistência conceitual. Ao incorporar arquiteturas modernas como RAG e LLMs, o modelo não apenas recupera informações relevantes, mas também gera explicações contextualizadas que fortalecem o processo pericial. Somam-se a isso mecanismos de anonimização e validação por especialistas, que asseguram a confiabilidade e a adequação ética no uso dos dados. Nesse sentido, a pesquisa se destaca por oferecer uma solução inovadora e aplicada, capaz de preencher uma lacuna metodológica e apoiar, de forma efetiva, a análise e comparação de lesões em laudos de necrópsia.

3 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta a fundamentação teórica empregada no desenvolvimento deste trabalho, sendo essencial para a compreensão das técnicas utilizadas na metodologia proposta para a análise e classificação de laudos de necrópsia.

3.1 Medicina Legal

A Medicina Legal constitui uma disciplina que aplica conhecimentos médicos e científicos às demandas do Direito e da Justiça. Ela não se limita a uma especialidade médica, mas se caracteriza como ciência, técnica e arte voltadas à elucidação de fatos jurídicos (França, 2017). Historicamente, autores clássicos enfatizam seu papel judicial: Paré descreveu-a como "a arte de fazer relatórios em juízo", enquanto Foderé e Orfila a definiram como a aplicação de conhecimentos médicos à interpretação e elaboração das leis (França, 2017; Coêlho, 2011). Peixoto e Fávero reforçam seu caráter técnico-científico voltado para o Direito, indicando sua relevância social e legal (França, 2017).

3.1.1 Perícia Médico-Legal

A perícia médico-legal é o ato técnico-científico mediante o qual se aplicam conhecimentos médicos para esclarecer fatos de interesse jurídico (Nai e Maia, 2015; Brasil, 2025; França, 2017). Ela se distingue do ato terapêutico, sendo avaliativa e elucidativa, e visa subsidiar decisões judiciais, administrativas ou policiais. A perícia oficial é realizada por órgãos estatais, como os Institutos Médico-Legais (IMLs), ou por peritos designados, com foco em esclarecer a ocorrência de fatos que envolvam vida, saúde ou integridade de pessoas (Brasil, 2025).

O médico-legista atua como analista técnico, sem vínculo com interesses particulares das partes envolvidas, e não como médico assistencial, garantindo a imparcialidade da avaliação (Pereira, 2013; França, 2017).

3.1.2 Laudo de Necrópsia

O laudo médico-legal é um documento técnico que relata detalhadamente a perícia, servindo como base para o processo judicial (Farias e Pinho, 2016; Brasil, 2025; Pereira, 2013).

Sua estrutura padrão inclui:

1. **Preâmbulo:** Identificação do perito, data, hora, local da perícia e qualificação do examinado.
2. **Quesitos:** Perguntas formuladas pela autoridade solicitante.

3. **Histórico:** Contextualização do caso e eventos prévios.
4. **Descrição:** Registro minucioso dos achados, utilizando terminologia anatômica, imagens, fotografias e gráficos (França, 2017; Brasil, 2024b).
5. **Conclusões:** Síntese fundamentada dos achados.
6. **Respostas aos quesitos:** Justificativa técnica das respostas, possibilitando avaliação judicial (Pereira, 2013).

Laudos podem abranger diversos tipos de perícias, incluindo cadavérica, sexológica, odontolegal e traumatológica (Farias e Pinho, 2016).

A clareza e completude do laudo são essenciais, pois a necropsia é irreversível (França, 2017). A descrição técnica deve ser objetiva, sistemática e detalhada, permitindo que o juiz ou outro destinatário compreenda a natureza das lesões sem margem para ambiguidades (França, 2017; Brasil, 2024b). Laudos incompletos ou confusos podem gerar nulidades, atrasos processuais e questionamentos sobre a integridade da prova (Miziara, Miziara e Munoz, 2012).

Um laudo organizado garante confiabilidade e completude, permitindo conclusões sólidas, evitando perícias repetidas e erros na investigação (França, 2017), além de subsidiar as decisões judiciais ao constituir a base do julgamento, possibilitando ao juiz avaliar adequadamente os fatos (Pereira, 2013; França, 2017). Além disso, apresenta relevância social e administrativa, contribuindo para políticas públicas, para a correta identificação de vítimas e para a padronização de informações, evitando a subutilização de dados em sistemas como o Pro-Aim (Pereira, 2013; Brasil, 2024b). Em casos sensíveis, como os de feminicídio, esses laudos devem receber atenção especial, garantindo prioridade no processamento (Brasil, 2024a).

3.1.3 Importância da Análise Automatizada

A complexidade dos laudos periciais, aliada à necessidade de padronização, torna a análise desses documentos um campo propício para o suporte de Inteligência Artificial (IA). Sistemas automatizados podem auxiliar na identificação de padrões, detecção de inconsistências, padronização terminológica e extração de informações críticas, aumentando a eficiência e a confiabilidade dos processos periciais.

A documentação pericial, muitas vezes extensa e detalhada, dificulta a gestão, recuperação e interpretação dos laudos. Nesse contexto, a aplicação de IA e PLN contribui para a organização e análise eficiente de grandes volumes de dados, reduzindo o tempo de investigação e aumentando a produtividade do Judiciário (Farias e Pinho, 2016; Ferreira, Pinheiro e Fernandes, 2025; Sousa e Prata, 2019).

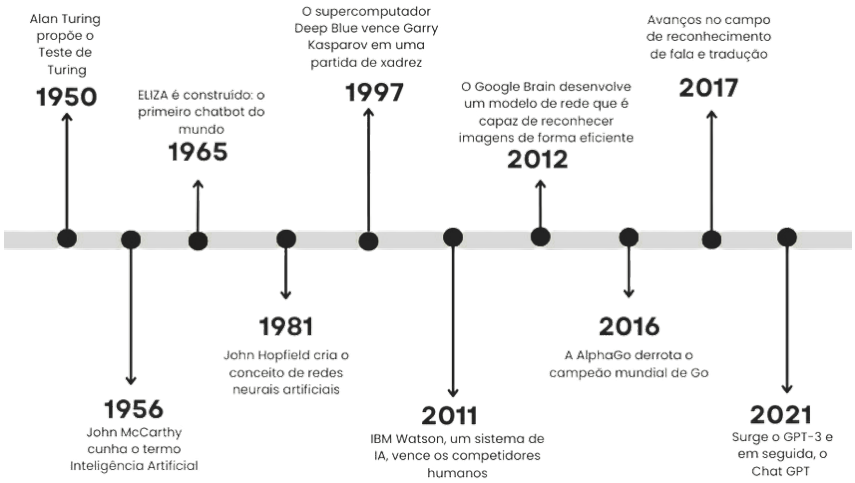
Além disso, essas tecnologias promovem uniformidade e redução da subjetividade, padronizando termos e conceitos por meio de sistemas como SNOMED CT, CID-10 e ontologias

médicas, tornando os laudos consistentes e legíveis por máquinas (Maciel, Ferreira e Marin, 2018; Farias e Pinho, 2016).

Técnicas como o Raciocínio Baseado em Casos (RBC) permitem a comparação e reutilização de conhecimento, medindo similaridade entre laudos, identificando padrões e reaproveitando experiências anteriores (Lorenzi, 1998; Oliveira, 2011; Castro e Neves, 2024). O uso dessas ferramentas aumenta a precisão diagnóstica, acelera o processamento de informações e possibilita a descoberta de evidências que poderiam passar despercebidas, fortalecendo a confiabilidade da Medicina Legal (Ferreira, Pinheiro e Fernandes, 2025).

3.2 Inteligência Artificial

Figura 1 – Evolução da Inteligência artificial ao longo dos anos



Fonte – Adaptada de Mídia Market (2024)

A Inteligência Artificial (IA) é um campo da ciência da computação voltado ao desenvolvimento de sistemas capazes de executar tarefas que requerem inteligência humana, como raciocínio, aprendizado, percepção e compreensão de linguagem (Hulsen, 2023; Rego, 2025). Essencialmente, a IA imita funções cognitivas humanas, interpretando dados, aprendendo com eles e adaptando-se para atingir objetivos específicos (Piraianu et al., 2023; Harsh et al., 2025). Seus sistemas processam grandes volumes de informações, reconhecendo padrões complexos em conjuntos multidimensionais e multimodais, utilizando técnicas como aprendizado de máquina, redes neurais e aprendizado profundo (Bajwa et al., 2021).

O conceito de IA surgiu na década de 1950, quando Alan Turing propôs o teste que leva seu nome e John McCarthy cunhou o termo em 1956, definindo-a como a ciência e engenharia de construir máquinas inteligentes (Volonnino, 2024; Qi, Zhou e Yu, 2024). Desde então, sua evolução passou por fases distintas: a consolidação inicial (1950-1980) com sistemas baseados em regras e raciocínio simbólico; a adoção prática em saúde e engenharia (1990),

com robôs cirúrgicos e sistemas de apoio clínico (Nascimento et al., 2024); o surgimento de aprendizado profundo e redes neurais avançadas (2010); e, mais recentemente, os Grandes Modelos de Linguagem (LLMs), como GPT-3 e GPT-4, capazes de interpretar contexto, gerar textos coerentes e aprender a partir de exemplos limitados (Ghanta et al., 2024; Rego, 2025).

A IA apresenta aplicações amplas em diversos setores, incluindo econometria, indústria automotiva, sistemas de recomendação, reconhecimento facial e veículos autônomos (Hulsen, 2023; Sousa e Prata, 2019). Em alguns domínios, seu desempenho já supera o humano, refletindo-se na crescente presença de assistentes virtuais em atividades cotidianas (Volonnino, 2024).

3.2.1 Aprendizado de Máquina

O Aprendizado de Máquina (ML) é um subcampo da Inteligência Artificial (IA) que se dedica à criação de algoritmos capazes de aprender a partir de dados e realizar previsões ou decisões com base nesse aprendizado (Malik et al., 2024; Nascimento et al., 2024; Lee, Britto e Diwan, 2024; Bajwa et al., 2021; Piraianu et al., 2023). Diferentemente da IA clássica, representada pelos sistemas especialistas, que dependia de regras explícitas fornecidas por especialistas para cada situação possível (Lefèvre e Tournois, 2023), o ML identifica padrões e correlações nos dados para gerar predições ou classificações (Malik et al., 2024).

Abordagens como o Raciocínio Baseado em Casos (RBC) também se diferenciam da IA tradicional, pois centram-se no conhecimento específico de exemplos concretos, em vez de regras genéricas (Oliveira, 2011).

O ML envolve técnicas estatísticas e computacionais que permitem aos algoritmos aprender com experiências anteriores, ajustando seus modelos conforme novos dados se tornam disponíveis (Bajwa et al., 2021; Piraianu et al., 2023, 2023). Entre as principais modalidades de aprendizado destacam-se o supervisionado, o não supervisionado e o por reforço, com variantes como semi-supervisionado, auto-supervisionado e multi-instância.

Entre as principais modalidades de aprendizado de máquina, destaca-se o supervisionado, no qual o modelo recebe um conjunto de dados previamente rotulado e ajusta suas previsões com base em sinais de correção fornecidos (Qi, Zhou e Yu, 2024; Jurafsky e Martin, 2025). Essa abordagem é amplamente aplicada na medicina, incluindo a predição de desfechos clínicos, a medicina de precisão e a classificação diagnóstica, em que a variável de interesse é conhecida antecipadamente (Piraianu et al., 2023).

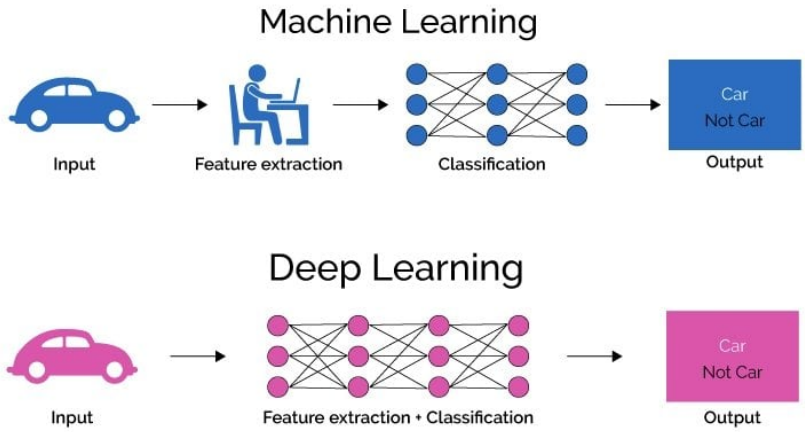
Já o aprendizado não supervisionado permite que o modelo identifique padrões em dados não rotulados, sem qualquer orientação explícita (Qi, Zhou e Yu, 2024; Lee, Britto e Diwan, 2024). Essa modalidade é especialmente útil em tarefas como o agrupamento de documentos médicos e a análise de tópicos latentes em coleções de textos clínicos (Faleiros, 2016).

O aprendizado por reforço baseia-se na interação do modelo com o ambiente, aprendendo a partir de recompensas ou penalidades (Lee, Britto e Diwan, 2024; Barros et al., 2025). Em contextos médicos, essa abordagem ainda emergente tem aplicações, por exemplo, na mitigação de vieses durante o treinamento de modelos (Barros et al., 2025).

Por fim, o aprendizado semi-supervisionado combina dados rotulados e não rotulados para aumentar a eficácia do modelo (Faleiros, 2016). Essa técnica tem sido aplicada, por exemplo, na classificação transdutiva de textos médicos, permitindo aproveitar melhor grandes volumes de dados parcialmente rotulados.

3.2.2 Aprendizado Profundo

Figura 2 – Machine Learning vs. Deep Learning



Fonte – Boesch (2023)

O *Deep Learning* é uma subárea do Aprendizado de Máquina (ML), caracterizada pela profundidade de suas redes neurais e pela capacidade de aprender representações hierárquicas e complexas dos dados (Piraianu et al., 2023). Como ilustrado na Figura 2, enquanto o ML clássico depende de atributos pré-definidos e engenharia manual de características, o DL é capaz de extrair automaticamente representações de alto nível diretamente de entradas brutas, reduzindo a necessidade de intervenção humana (Jurafsky e Martin, 2025). Essas redes são compostas por múltiplas camadas sucessivas, cada uma apta a gerar representações intermediárias que contribuem para o reconhecimento progressivo de padrões complexos (Piraianu et al., 2023; Nascimento et al., 2024).

No contexto do PLN, as arquiteturas de redes neurais evoluíram significativamente, desde Redes Neurais Recorrentes (RNNs) até Modelos de Linguagem de Grande Escala (LLMs) baseados em *transformers*. As RNNs foram desenvolvidas para lidar com dados sequenciais, preservando a memória de entradas anteriores, sendo aprimoradas pelas variantes LSTM e GRU,

que aumentam a eficiência em tarefas de reconhecimento de entidades e outras aplicações de PLN (Ghanta et al., 2024; Zhou et al., 2024; Lee et al., 2019).

Redes Neurais Convolucionais (CNNs), originalmente aplicadas a imagens, também mostraram eficácia em PLN quando combinadas com *embeddings* de palavras, identificando padrões locais e mantendo eficiência computacional. Essa abordagem possibilita aprender relações textuais complexas de forma precisa (Zhou et al., 2024; Raza e Schwartz, 2023).

A arquitetura *Transformer*, introduzida em 2017, tornou-se a base dos LLMs modernos. Por meio de mecanismos de autoatenção, processa sequências em paralelo, capturando relações contextuais entre palavras e aumentando a eficiência no treinamento e análise de textos longos (Ghanta et al., 2024; Zhou et al., 2024; Sivarajkumar et al., 2024; Garcia-Carmona et al., 2025; Rego, 2025).

O uso de técnicas de *Deep Learning* é pertinente para a análise automática de laudos de necrópsia, que consistem em textos extensos, complexos e não estruturados. Ele permite extrair informações críticas e automatizar tarefas que exigiriam grande esforço manual. Com o PLN, dados clínicos podem ser organizados, e informações essenciais identificadas, apoiando decisões médicas e periciais (Piraianu et al., 2023; Elhaddad e Hamam, 2024).

Além disso, modelos de DL possibilitam automatizar a codificação de diagnósticos segundo a CID-10 a partir de descrições livres, aumentando a padronização e a confiabilidade estatística (Duarte et al., 2018; Coutinho e Martins, 2022). Também é possível prever eventos fatais, como overdoses, a partir das narrativas de autópsia (Tang et al., 2023).

LLMs capturam padrões linguísticos e relações semânticas complexas, permitindo descrições detalhadas de lesões e avaliações periciais (Garcia-Carmona et al., 2025; Brasil, 2025; Farias e Pinho, 2016). Além disso, o DL oferece escalabilidade e eficiência no processamento de grandes volumes de dados textuais, sendo aplicável à radiologia forense e à análise de laudos médicos (Bajwa et al., 2021; Ferreira, Pinheiro e Fernandes, 2025).

Dessa forma, o *Deep Learning* constitui a base técnica essencial para sistemas inteligentes de análise automatizada de laudos periciais, integrando interpretação, classificação e extração de informações de forma precisa, escalável e confiável.

3.3 Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN) é um campo da inteligência artificial que visa permitir que computadores compreendam, interpretem, gerem e manipulem a linguagem humana de maneira automatizada. Segundo Zhou et al. (2024), trata-se de um subcampo da ciência da computação voltado para a análise e compreensão de textos em linguagem natural, enquanto Almuhana e Abbas (2022) enfatiza que a área possibilita aos computadores acessar o significado contido em entradas linguísticas humanas. De acordo com Busch et al. (2024), o PLN integra linguística e lógica computacional, transformando dados textuais em representações

compreensíveis para máquinas. Na prática, algoritmos e modelos desenvolvidos para o PLN permitem a interpretação, criação e manipulação de textos, oferecendo funcionalidade e naturalidade na interação entre humanos e sistemas computacionais (Saini et al., 2024; Rego, 2025). O campo, também denominado *Linguística Computacional*, tem promovido avanços significativos em tecnologias contemporâneas (Sousa e Prata, 2019).

Nesse contexto, o principal objetivo do PLN é converter textos não estruturados em representações estruturadas que possam ser processadas por algoritmos. Para isso, aplicam-se técnicas de pré-processamento textual, como tokenização, lematização, remoção de *stopwords*, normalização de caracteres e anonimização de dados sensíveis (Santos et al., 2024; Silva e Pazin-Filho, 2025; Zhou et al., 2024). A tokenização, por exemplo, divide o texto em unidades significativas chamadas tokens, permitindo que cada sentença seja representada como uma sequência de termos (Sousa e Prata, 2019). Técnicas adicionais, como *stemming* e conversão para minúsculas, reduzem a variabilidade morfológica, enquanto a remoção de caracteres especiais, números e pontuações simplifica a análise computacional. Posteriormente, os tokens são transformados em vetores, criando matrizes que permitem capturar relações linguísticas e contextuais por meio de algoritmos de aprendizado (Sousa e Prata, 2019).

Historicamente, os fundamentos do PLN remontam à Antiguidade, com o trabalho de Pāṇini sobre a gramática do Sânscrito, que antecipou princípios de teorias formais de linguagens (Jurafsky e Martin, 2025). No século XX, o desenvolvimento da linguística computacional foi marcado pelo debate entre abordagens estatísticas, baseadas em modelos de probabilidade, como os n-gramas propostos por Shannon (1948), e críticas como as de Chomsky, que questionava a capacidade desses métodos de explicar a competência linguística humana (Jurafsky e Martin, 2025). Nas décadas seguintes, consolidaram-se modelos probabilísticos mais sofisticados, como o *Latent Dirichlet Allocation*, aplicado à mineração de textos (Faleiros, 2016). Mais recentemente, a ascensão do *deep learning* e das arquiteturas baseadas em *transformers* inaugurou uma nova fase, com modelos de larga escala, como GPT-4 e Gemini, capazes de interpretar contextos complexos com elevada precisão (Liu et al., 2025).

Apesar dos avanços, o PLN enfrenta desafios significativos relacionados à complexidade da linguagem humana. A ambiguidade sintática e semântica, apontada por Jurafsky e Martin (2025), é um obstáculo importante, já que uma mesma sentença pode ter múltiplas interpretações. A polissemia, a dependência do contexto e a diversidade linguística tornam a compreensão automatizada difícil, tornando o progresso mais lento em relação a outras técnicas computacionais (Zhou et al., 2024). Em domínios especializados, como o jurídico ou médico-legal, a terminologia específica e a variação interpretativa intensificam essas dificuldades (Sousa e Prata, 2019). Além disso, consultas ambíguas, abreviações, gírias e linguagem informal, comuns em redes sociais, aumentam ainda mais a complexidade (Santos et al., 2024; Zhang e Zhang, 2025).

Outro aspecto relevante envolve questões éticas e de viés algorítmico. Estudos indicam

que vieses raciais e de gênero podem impactar o desempenho de modelos, tornando necessária a garantia de representatividade adequada nos dados de treinamento (Barros et al., 2025; Ferreira, Pinheiro e Fernandes, 2025). No contexto médico, a utilização de algoritmos como suporte a decisões clínicas levanta dúvidas sobre a responsabilidade entre profissionais de saúde e desenvolvedores de modelos de aprendizado de máquina (Nascimento et al., 2024).

Em síntese, o Processamento de Linguagem Natural fornece ferramentas capazes de permitir que sistemas computacionais compreendam e manipulem a linguagem humana em diferentes níveis — morfológico, sintático e semântico. Ao mesmo tempo, a área enfrenta desafios técnicos, contextuais e éticos decorrentes da complexidade da linguagem e da diversidade dos dados, como ambiguidades, polissemia, variações terminológicas e potenciais vieses nos modelos. A integração de técnicas de pré-processamento, representações clássicas de texto, como BoW e TF-IDF, e *embeddings* contextuais possibilita a realização de tarefas avançadas de análise, classificação e interpretação de textos complexos, incluindo documentos médico-legais e outros textos técnico-científicos.

3.3.1 Principais Técnicas

O PLN envolve diversas técnicas que permitem transformar textos em informações computacionais, possibilitando sua análise e interpretação por sistemas inteligentes. Entre essas técnicas, o pré-processamento textual representa a etapa inicial, sendo essencial para lidar com a natureza não estruturada dos dados (Santos et al., 2024). Além de uniformizar os textos, essas técnicas reduzem ambiguidades e preparam os dados para representações numéricas e modelagem semântica.

3.3.1.1 Pré-Processamento

O pré-processamento inclui procedimentos como tokenização, lematização e remoção de *stopwords*. A tokenização consiste na divisão do texto em unidades menores, denominadas *tokens*, que podem ser palavras, frases ou caracteres. Esse processo é fundamental para que cada termo seja tratado individualmente e possa ser analisado e representado computacionalmente (Saini et al., 2024).

A lematização e o *stemming* reduzem palavras a suas formas básicas ou radicais, eliminando variações morfológicas e afixos. Essa simplificação garante maior uniformidade, permitindo que termos com significados semelhantes, como "correr" e "correndo", sejam interpretados como equivalentes (Sousa e Prata, 2019). Paralelamente, a remoção de *stopwords* retira palavras comuns que ocorrem com frequência, mas pouco contribuem para o significado do texto. Essa etapa diminui o esforço computacional e melhora a extração de informações relevantes (Sousa e Prata, 2019).

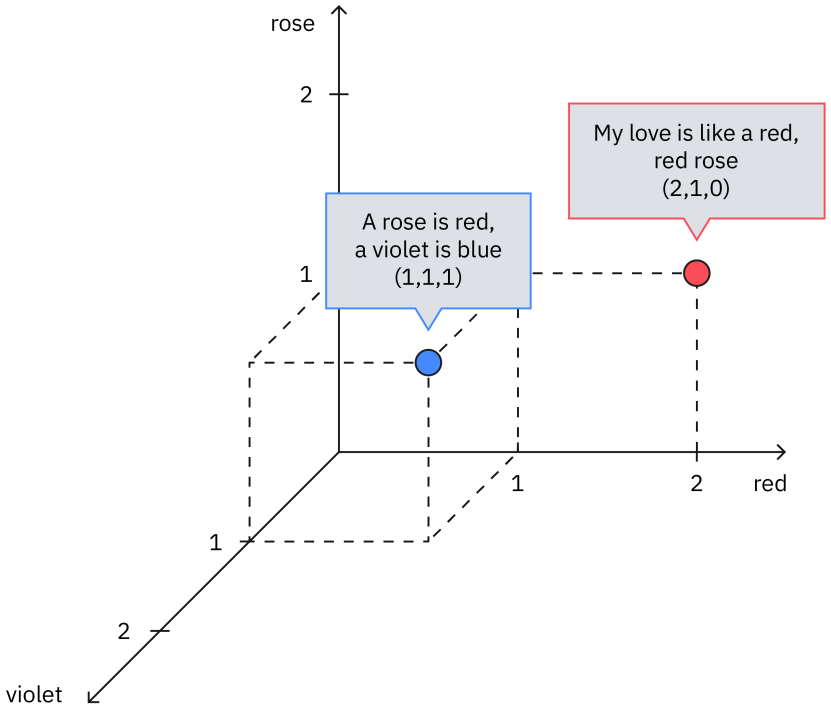
Além dessas etapas, o pré-processamento pode incluir a normalização de letras para minúsculas, a eliminação de pontuações, números e caracteres especiais, ajustando o texto

para uma forma mais consistente e fácil de ser manipulada pelos algoritmos (Saini et al., 2024; Sousa e Prata, 2019).

3.3.1.2 Representações Clássicas de Texto

Após o pré-processamento, os textos podem ser transformados em representações numéricas que permitem o processamento computacional. Entre os métodos tradicionais, destacam-se o *Bag of Words* (BoW) e o TF-IDF.

Figura 3 – Representação vetorial do *Bag of Words*



Fonte – Murel e Kavlakoglu (2024)

O *Bag of Words* representa o documento como um conjunto de palavras, preservando a frequência de ocorrência, mas sem considerar a ordem ou o contexto gramatical (Mujtaba et al., 2018a). Essa abordagem permite a construção de vetores que descrevem a presença de palavras no texto (conforme a Figura 3), sendo simples e eficiente para tarefas de classificação e recuperação de informações (Reys et al., 2020; Permata, Rendika e Julianty, 2025).

O TF-IDF aprimora o BoW ao ponderar os termos de acordo com sua frequência no documento e sua raridade no corpus, valorizando palavras discriminativas e reduzindo a importância de termos comuns (Guti'errez et al., 2022; Mujtaba et al., 2018b; Sivarajkumar et al., 2024). Dessa forma, palavras que aparecem muitas vezes em um documento, mas são raras no corpus, recebem maior peso, contribuindo para uma análise mais precisa do conteúdo textual (Reys et al., 2020).

3.4 Large Language Models

A evolução dos modelos de linguagem reflete um progresso contínuo na capacidade de capturar e gerar linguagem natural, partindo de abordagens estatísticas simples até arquiteturas neurais complexas capazes de compreender contextos amplos e sutis nuances semânticas. Os *Large Language Models* (LLMs) constituem a expressão mais avançada dessa evolução, representando modelos de Inteligência Artificial, geralmente baseados em transformadores, treinados com bilhões de parâmetros e grandes volumes de dados textuais. Esses modelos têm a capacidade de aprender relações linguísticas complexas e desempenhar diversas tarefas de processamento de linguagem natural, incluindo geração de texto, tradução, *question answering* e interação conversacional (Busch et al., 2024; Liu et al., 2025; Yin et al., 2025). A arquitetura *Transformer*, proposta por Vaswani em 2017, é a base predominante dos LLMs modernos, permitindo que o modelo processe sequências de texto em paralelo e capture relações contextuais profundas por meio de mecanismos de *self-attention* (Ghanta et al., 2024; Yang et al., 2025).

Historicamente, os modelos de linguagem iniciaram com abordagens estatísticas como os *n-grams*, que estimavam a probabilidade de uma palavra com base nas anteriores. Embora amplamente utilizados por décadas em tarefas de reconhecimento de fala e tradução automática, esses modelos apresentavam limitações significativas, como o crescimento exponencial de parâmetros e a incapacidade de generalizar além de sequências de treinamento idênticas (Busch et al., 2024; Jurafsky e Martin, 2025). A insuficiência de dados e a esparsidade nas sequências textuais também restringiam a precisão dessas abordagens, especialmente em domínios com vocabulário especializado.

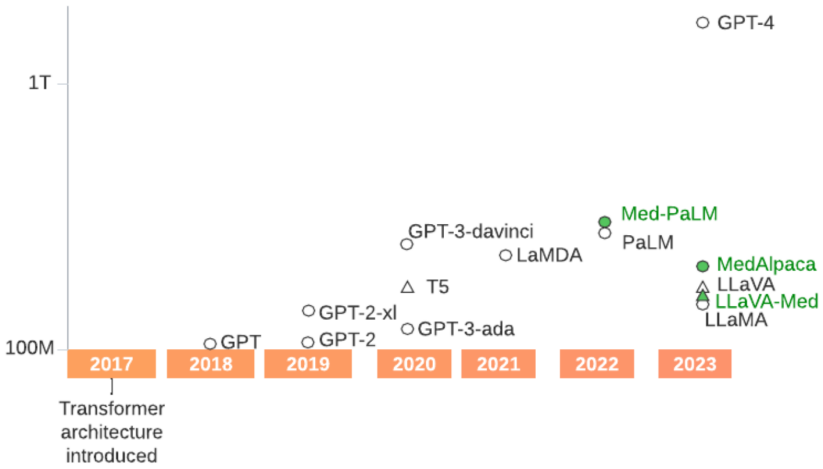
A introdução das RNNs e, posteriormente, das LSTMs trouxe avanços importantes no início dos anos 2000, ao permitir que modelos aprendessem dependências de longo alcance em textos e realizassem predições mais robustas em tarefas diversas de PLN (Busch et al., 2024; Jurafsky e Martin, 2025). Ainda assim, o processamento sequencial das RNNs e LSTMs limitava a velocidade e a escalabilidade, prejudicando o desempenho em textos longos ou conjuntos de dados extensos (Ghanta et al., 2024).

3.4.1 Arquitetura Transformer

A chegada da arquitetura *Transformer* representou uma ruptura nessa sequência evolutiva. Ao adotar mecanismos de *self-attention*, é possível identificar partes relevantes do texto simultaneamente, permitindo processamentos paralelos e capturando relações contextuais complexas em janelas de contexto amplas (Busch et al., 2024; Jurafsky e Martin, 2025). Essa inovação resolveu limitações críticas dos modelos anteriores, incluindo a incapacidade de lidar com grandes volumes de dados de forma eficiente e a dificuldade em generalizar para contextos mais longos. Como resultado, os LLMs baseados em *Transformers*, como GPT-3, LLaMA e variantes voltadas para domínios específicos como BioBERT e ClinicalBERT, atingem níveis de compreensão semântica e geração textual anteriormente inacessíveis, projetando palavras em

espaços vetoriais contínuos onde contextos semelhantes produzem representações próximas (Sarker et al., 2024; Jurafsky e Martin, 2025; Sivarajkumar et al., 2024). É possível observar como a complexidade dos modelos de LLMs aumentou ao longo dos anos conforme ilustrado na Figura 4, que demonstra o crescimento exponencial no número de parâmetros e o surgimento de modelos especializados em domínios biomédicos.

Figura 4 – Evolução dos LLMs ao longo do tempo. Eixo vertical em escala logarítmica indica o tamanho dos parâmetros. Modelos verdes foram pré-treinados em dados biomédicos. Triângulos indicam modelos generativos com arquitetura encoder-decoder; os demais são apenas decoder.



Fonte – Sarker et al. (2024)

Portanto, a trajetória dos modelos de linguagem evidencia uma progressão do simples cálculo de frequências com *n-grams*, passando por redes neurais sequenciais como RNNs e LSTMs, até a sofisticada arquitetura Transformer, que habilita os LLMs modernos a superar restrições de generalização, contexto e paralelização, consolidando-se como a base das aplicações mais avançadas em processamento de linguagem natural (Busch et al., 2024).

3.4.1.1 Funcionamento

A arquitetura *Transformer* representa uma evolução fundamental no processamento de linguagem natural, oferecendo soluções às limitações de modelos sequenciais tradicionais, como RNNs e CNNs. Seu principal avanço reside no mecanismo de atenção (attention), que permite ao modelo atribuir diferentes níveis de importância a elementos de uma sequência, construindo representações contextuais ricas e capturando relações de longo alcance ao longo de grandes trechos de texto (Jurafsky e Martin, 2025; Pinto-Coelho, 2023).

O mecanismo de atenção (à esquerda na Figura 5) possibilita que o modelo avalie a relevância de cada *token* em relação aos demais, permitindo que cada elemento da sequência seja contextualizado de acordo com o conjunto completo de dados de entrada. No caso do

self-attention, cada *token* atua simultaneamente como *query* (consulta), *key* (chave) e *value* (valor), criando vetores que representam tanto o elemento em si quanto suas relações com os demais *tokens* da sequência. Os scores de atenção são calculados pelo produto escalar entre Q e K, normalizados pelo *softmax* (conforme fórmula 3.1), e utilizados para ponderar os valores V, resultando em uma representação final que sintetiza a informação mais relevante (Ghanta et al., 2024; Pinto-Coelho, 2023).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.1}$$

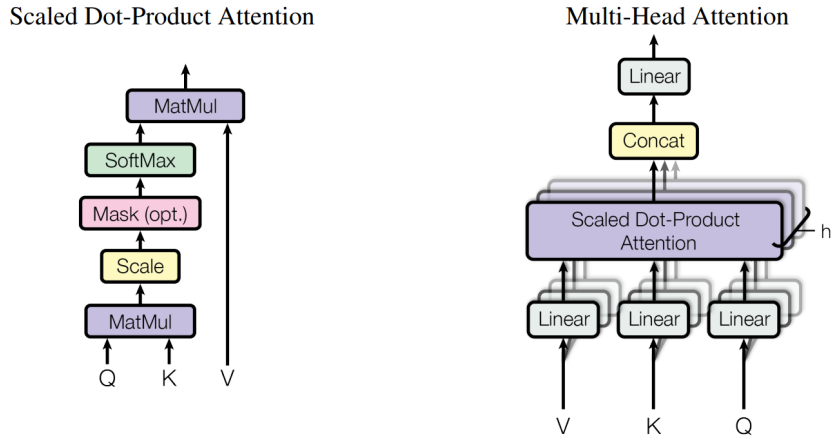


Figura 5 – (à esquerda) Atenção por Produto Escalar Escalado. (à direita) Atenção Multi-Cabeças, composta por várias camadas de atenção executadas em paralelo.

Fonte – Vaswani et al. (2017)

Esse mecanismo permite que o modelo capture dependências de longo alcance, essencial para compreender textos extensos, e supera a limitação de processamento sequencial das redes anteriores. Além disso, o *multi-head attention* (à direita na Figura 5) distribui a análise do contexto em diferentes subespaços, permitindo que cada cabeça foque em padrões específicos ou relações semânticas distintas, enriquecendo a representação final do token (Jurafsky e Martin, 2025). Essa arquitetura foi proposta por Vaswani et al. (2017) conforme a Figura 6.

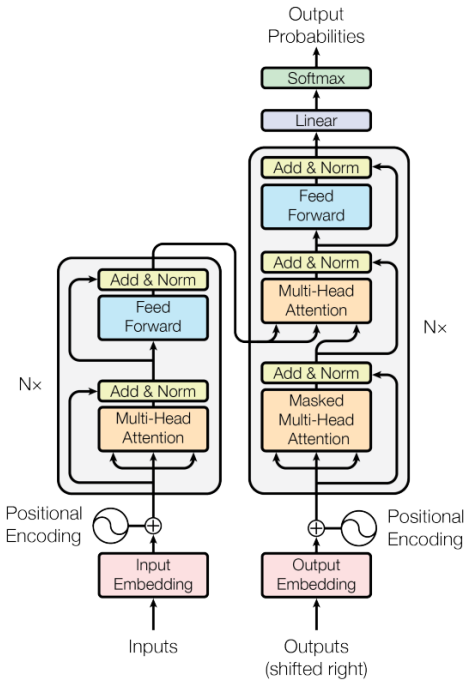


Figura 6 – Arquitetura base de um modelo Transformer

Fonte – Vaswani et al. (2017)

Na Figura 6 o módulo à esquerda representa o *encoding*, enquanto o módulo à direita trata-se do *decoder*. Essa diagramação serve como fundamento para outras configurações de *Transformers*. Sendo as três principais as seguintes: *encoder-decoder*, *decoder-only* e *encoder-only*. Cada uma dessas variantes possui características próprias, voltadas para diferentes tipos de tarefas em NLP.

3.4.1.2 Encoder-Decoder

Os modelos *encoder-decoder* são projetados para mapear sequências de entrada em sequências de saída, que podem variar em comprimento ou tipo de *token*, sendo ideais para tarefas de tradução automática ou reconhecimento de fala. O codificador (*encoder*) transforma a entrada em uma representação contextualizada (H_{enc}), enquanto o decodificador (*decoder*) gera a saída palavra por palavra, condicionando-se tanto à representação do codificador quanto aos *tokens* já produzidos. O *decoder* inclui uma camada de atenção cruzada (*cross-attention*), que permite integrar todas as informações processadas pelo codificador em cada passo de geração (Jurafsky e Martin, 2025), conforme representado na Figura 6.

Essa configuração é útil quando a saída não possui uma correspondência direta com a entrada, como ocorre em tradução de idiomas, onde a ordem e o número de palavras diferem, ou em reconhecimento automático de fala, em que os *tokens* de áudio precisam ser mapeados para palavras escritas (Jurafsky e Martin, 2025).

3.4.1.3 Decoder-Only

Os modelos *decoder-only* seguem um fluxo causal, ou seja, cada *token* gerado depende apenas dos *tokens* anteriores, funcionando de maneira autorregressiva. Essa abordagem é a base dos grandes modelos de linguagem, como GPT, Claude e LLaMA, e é especialmente eficaz em tarefas de geração de texto, sumarização e diálogos interativos. A previsibilidade unidirecional permite que cada *token* seja gerado de forma coerente e sequencial, construindo saídas contínuas e coesas (Mukund e Easwarakumar, 2025; Jurafsky e Martin, 2025).

Esses modelos não necessitam de codificação da entrada separada, como no *encoder-decoder*, tornando-os mais simples para tarefas puramente generativas, mas limitados quando a tarefa exige mapeamento complexo entre diferentes tipos de *tokens*.

3.4.1.4 Encoder-Only

Diferente das duas arquiteturas anteriores, os modelos *encoder-only* não são geradores de texto, mas sim analisadores. Eles produzem representações contextualizadas de cada *token* da entrada, levando em conta o contexto completo, tanto à esquerda quanto à direita do *token*. Essa bidirecionalidade permite que modelos como BERT realizem tarefas de interpretação, classificação, análise de sentimentos e respostas a perguntas, fornecendo *embeddings* ricos e informativos que representam cada palavra em função do contexto total do documento (Coutinho e Martins, 2022; Jurafsky e Martin, 2025).

3.4.2 Vantagens

Os *Transformers* superam os modelos sequenciais em vários aspectos. Primeiramente, o *self-attention* permite capturar dependências de longo alcance, relacionando *tokens* distantes diretamente, o que é essencial para textos longos e complexos (Pinto-Coelho, 2023; Yuan, 2025). Em segundo lugar, a criação de *embeddings* contextualizados possibilita que cada palavra seja representada de forma distinta conforme o contexto, aumentando a precisão semântica (Coutinho e Martins, 2022; Jurafsky e Martin, 2025).

Além disso, a arquitetura permite a paralelização do processamento, possibilitando treinamentos em larga escala com grande eficiência, superando a limitação das RNNs, que processavam sequências de forma linear (Busch et al., 2024; Ghanta et al., 2024). A ampla janela de contexto, que pode abranger centenas de milhares de *tokens*, garante que informações distantes sejam consideradas na predição de palavras futuras (Jurafsky e Martin, 2025).

Embora o *Transformer* clássico apresente complexidade quadrática em relação ao comprimento da entrada, tornando difícil o processamento de textos muito longos, diversas estratégias foram desenvolvidas. Modelos como Longformer e Big Bird utilizam atenção esparsa e arquiteturas hierárquicas para reduzir a complexidade e permitir o uso eficiente de longas sequências, mantendo precisão e fluência (Mukund e Easwarakumar, 2025; Coutinho e Martins,

2022). Técnicas como atenção mista (*mixed attention*), aplicadas em modelos como FLASH, segmentam a sequência em blocos, combinando atenção local e global, de modo a reduzir a perplexidade sem comprometer a captura de dependências de longo alcance (Niu et al., 2023).

3.4.3 Principais Famílias de LLMs

Os LLMs representam um marco no Processamento de Linguagem Natural, pois se baseiam em arquiteturas *transformer* com bilhões de parâmetros e capacidade de aprender relações linguísticas complexas. Essa característica possibilita compreender e gerar textos de forma altamente contextualizada, o que levou os LLMs a se tornarem centrais em pesquisas e aplicações de inteligência artificial (Busch et al., 2024; Yin et al., 2025).

3.4.3.1 Generative Pre-trained Transformer

A família GPT, desenvolvida pela OpenAI, é caracterizada por uma arquitetura *decoder-only* e funciona de maneira autorregressiva, prevendo a próxima palavra com base no contexto anterior. Essa estrutura a torna altamente eficiente em tarefas de geração de texto natural (Jurafsky e Martin, 2025; Liu et al., 2025).

O GPT-4, lançado em 2023, ultrapassou a marca de um trilhão de parâmetros e introduziu avanços em precisão, redução de vieses e capacidade multimodal, permitindo trabalhar simultaneamente com texto e imagem. Em 2024, o GPT-4o trouxe maior eficiência energética, custos de execução reduzidos e ampliação da janela de contexto para até 128 mil *tokens*, consolidando sua posição como modelo de uso generalista e de alta performance (Sarker et al., 2024; Garcia-Carmona et al., 2025; Ghanta et al., 2024).

No campo da saúde, a família GPT tem sido aplicada em sumarização de prontuários médicos, geração de relatórios clínicos, extração de informações em grandes bases de dados e simulação de diálogos médico-paciente, o que auxilia tanto em atividades administrativas quanto em apoio a diagnósticos. Essa adaptabilidade demonstra seu potencial em contextos biomédicos diversos (Sivarajkumar et al., 2024; Tortora, 2024).

3.4.3.2 Large Language Model Meta AI

O LLaMA, desenvolvido pela Meta, consolidou-se como um modelo de código aberto, o que possibilitou ampla adoção acadêmica e comercial. Sua principal característica é o equilíbrio entre desempenho e acessibilidade, permitindo que instituições adaptem o modelo a domínios específicos por meio de técnicas de ajuste fino (Sarker et al., 2024; Garcia-Carmona et al., 2025).

O LLaMA 2, lançado em 2023, foi disponibilizado em versões de até 70 bilhões de parâmetros. Em LLaMA, o LLaMA 3 expandiu ainda mais, alcançando configurações de até

405 bilhões, tornando-se competitivo em relação aos modelos proprietários de larga escala (Garcia-Carmona et al., 2025; Liu et al., 2025).

Na saúde, a família LLaMA vem sendo usada como base para modelos especializados em biomedicina, análise de literatura científica, interpretação de exames textuais e apoio à pesquisa clínica. Sua natureza aberta facilita a criação de soluções customizadas para hospitais e centros de pesquisa, especialmente em contextos onde é necessário controlar os dados de forma local (Shen et al., 2025; Sivarajkumar et al., 2024).

3.4.3.3 Bidirectional Encoder Representations from Transformers

O BERT, criado pela Google em 2018, introduziu representações bidirecionais de linguagem a partir de uma arquitetura *encoder-only*. Diferentemente dos modelos autorregressivos, o BERT foca na compreensão contextual do texto e é amplamente usado em classificação, análise de sentimentos e sistemas de perguntas e respostas (Adhikary et al., 2024; Busch et al., 2024).

As versões originais incluem o BERT-Base, com 110 milhões de parâmetros, e o BERT-Large, com 340 milhões, marcando um avanço expressivo na época. Embora menores em escala que os LLMs mais recentes, o impacto conceitual do BERT foi profundo, influenciando o desenvolvimento de novos modelos e variantes (Jurafsky e Martin, 2025; Khattak et al., 2019).

Na saúde, variantes especializadas como o BioBERT e o ClinicalBERT foram adaptadas a textos biomédicos e clínicos, sendo utilizadas em extração de entidades médicas, análise de prontuários, mineração de literatura científica e apoio à tomada de decisões médicas. Essas aplicações têm fortalecido a pesquisa em biomedicina e a gestão hospitalar (Amugongo et al., 2025; Sarker et al., 2024; Sivarajkumar et al., 2024).

3.4.3.4 Text-to-Text Transfer Transformer

O T5, lançado pela Google em 2019, propôs um paradigma inovador ao transformar todas as tarefas de PLN em problemas de entrada e saída textual. Baseado em uma arquitetura *encoder-decoder*, alcançou até 11 bilhões de parâmetros e mostrou grande versatilidade em diferentes contextos (Sarker et al., 2024; Jurafsky e Martin, 2025).

Entre suas evoluções destaca-se o Flan-T5, que aprimorou a capacidade de generalização para múltiplas tarefas, tornando o modelo eficiente em tradução, sumarização, reconhecimento de entidades e resposta a perguntas. Essa abordagem unificada favorece a adaptação a cenários diversos, sem a necessidade de arquiteturas altamente específicas (Bora e Cuayahuitl, 2024; Lopez et al., 2025).

Na saúde, o T5 tem sido explorado para a identificação automática de entidades médicas, classificação de prontuários clínicos e sumarização de grandes volumes de dados. Sua capacidade de unificar diferentes tarefas de PLN torna-o particularmente útil em sistemas hospitalares que

exigem análise rápida e organizada de múltiplos tipos de informação (Sivarajkumar et al., 2024; Yang et al., 2022).

3.4.3.5 Comparativo

Tabela 1 – Comparativo entre as famílias de LLMs

Família	**Arquitetura**	**Nº Params.**	**Características**	**Aplicações na área da saúde**	**Ref.**
GPT (OpenAI)	Decoder-only (autorregressivo)	GPT-4: >1 trilhão; GPT-4o: até 128k tokens de contexto	Geração de texto, modelos multimodais (texto+imagem), ampla janela de contexto	Sumarização de prontuários, relatórios clínicos, apoio em diagnósticos, simulação de diálogos médico-paciente	(Sarker et al., 2024; Garcia-Carmona et al., 2025; Tortora, 2024)
LLaMA (Meta)	Decoder-only, código aberto	LLaMA 2: até 70B; LLaMA 3: até 405B	Código aberto, ajustável a domínios específicos, alta escalabilidade	Modelos biomédicos customizados, análise de literatura médica, interpretação de exames, pesquisa clínica	(Sarker et al., 2024; Garcia-Carmona et al., 2025; Shen et al., 2025)
BERT (Google)	Encoder-only (bidirecional)	BERT-Base: 110M; BERT-Large: 340M	Representações contextuais, *embeddings* bidirecionais, base para variantes especializadas	BioBERT e ClinicalBERT para análise de prontuários, extração de entidades médicas, mineração de literatura biomédica	(Amugongo et al., 2025; Busch et al., 2024; Sivarajkumar et al., 2024)
T5 (Google)	Encoder-decoder (seq2seq)	T5-11B: até 11 bilhões	Paradigma unificado "texto para texto", generalização multitarefa (Flan-T5)	Identificação de entidades médicas, classificação de textos clínicos, sumarização de dados hospitalares	(Sarker et al., 2024; Lopez et al., 2025; Yang et al., 2022)

A progressão dos modelos de linguagem revela um movimento de crescente sofisticação, evoluindo de arquiteturas bidirecionais como o BERT até modelos multimodais como o GPT-4o. Enquanto o BERT e suas variantes se destacam pela análise textual profunda e mineração de informações médicas, os GPT e T5 oferecem maior flexibilidade em geração de relatórios e interação com profissionais. Já o LLaMA, devido à sua abertura, representa uma opção estratégica para customizações específicas em ambientes de saúde.

A Tabela 1 sintetiza as principais famílias de LLMs, destacando suas arquiteturas, número de parâmetros, características específicas, aplicações na área da saúde e referências associadas. Nela, é possível observar que modelos como o GPT se sobressaem na geração de texto e em contextos multimodais, enquanto o BERT oferece representações contextuais profundas para análise de prontuários. O LLaMA, por sua vez, possibilita ajustes em domínios específicos, e o T5 se apresenta como uma solução versátil para tarefas de texto a texto, incluindo sumarização e classificação de dados clínicos.

Dessa forma, os LLMs se consolidam como ferramentas para apoiar tanto a pesquisa biomédica quanto a prática clínica, oferecendo soluções que vão desde a automatização de

relatórios até a interpretação contextual de grandes volumes de dados médicos (Shen et al., 2025; Sivarajkumar et al., 2024; Yang et al., 2022).

3.4.4 Limitações e Desafios dos LLMs em Sistemas de Apoio à Decisão

As limitações e desafios associados ao uso de LLMs têm impacto direto sobre a confiabilidade, a segurança e a aplicabilidade prática de Sistemas de Apoio à Decisão (SADs), especialmente em contextos críticos como a saúde e a justiça. Entre os problemas mais relevantes estão as alucinações, os vieses e os desafios relacionados à escalabilidade e ao custo computacional. Cada um desses aspectos interfere na qualidade das respostas geradas, na equidade das recomendações e na aceitação legal e ética dos sistemas.

Um dos riscos mais conhecidos é o fenômeno das alucinações, no qual os LLMs produzem informações linguística e gramaticalmente coerentes, mas que são factualmente incorretas ou fabricadas (Rego, 2025; Yin et al., 2025). Essa característica decorre da natureza probabilística desses modelos (Liu et al., 2025) e pode ter consequências significativas. Na área da saúde, conselhos incorretos gerados por LLMs podem representar um perigo direto ao paciente, comprometendo a segurança clínica e a confiabilidade do sistema (Sarker et al., 2024; Jurafsky e Martin, 2025). Em investigações forenses, informações alucinadas podem levar à criação de pistas falsas ou à introdução de evidências juridicamente inadmissíveis, prejudicando a integridade do processo (Yin et al., 2025). Embora técnicas de mitigação, como a Geração Aumentada por Recuperação (RAG), possam reduzir parcialmente esse risco, a eficácia dessas abordagens depende da qualidade e atualidade das fontes utilizadas (Zhang e Zhang, 2025).

Outro desafio crítico refere-se aos vieses. LLMs aprendem a partir de grandes conjuntos de dados históricos, o que faz com que possam refletir e até amplificar preconceitos presentes nesses dados (Barros et al., 2025; Su et al., 2025). Na prática clínica, isso pode resultar em diagnósticos enviesados e recomendações desiguais, afetando desproporcionalmente grupos minoritários e ampliando disparidades existentes (Su et al., 2025). Em contextos forenses, os modelos podem priorizar certos tipos de evidência ou reforçar estereótipos, comprometendo a imparcialidade do processo (Jurafsky e Martin, 2025; Yin et al., 2025). Um aspecto adicional é o chamado viés de automação, no qual profissionais humanos podem passar a confiar excessivamente nas recomendações do modelo, reduzindo a verificação crítica e colocando em risco a segurança dos pacientes (Nascimento et al., 2024).

Além de alucinações e vieses, os LLMs apresentam desafios relacionados à escalabilidade e ao custo computacional. Modelos de grande porte, como o MedPaLM e o LLaMA-v2 70B, exigem recursos de hardware substanciais, incluindo múltiplas GPUs, para treinamento e inferência eficientes (Sarker et al., 2024). A arquitetura transformer, base da maioria dos LLMs, possui complexidade quadrática no mecanismo de autoatenção (self-attention), o que torna dispendioso processar longas sequências de texto, como históricos clínicos extensos (Jurafsky e Martin, 2025). A limitação da janela de contexto também contribui para perda de informações

importantes, comprometendo o desempenho em tarefas de raciocínio mais complexas (Lopez et al., 2025; Su et al., 2025). Essas restrições dificultam o uso dos modelos em tempo real ou em ambientes com recursos limitados.

O impacto dessas limitações se estende à confiabilidade dos sistemas de apoio à decisão. Alucinações reduzem a segurança e a precisão das recomendações, enquanto vieses comprometem a imparcialidade e a equidade das decisões. A opacidade dos LLMs, característica conhecida como black-box, torna os processos de tomada de decisão menos transparentes, o que prejudica a confiança de profissionais e usuários, além de limitar a aceitabilidade legal das evidências produzidas (Ferreira, Pinheiro e Fernandes, 2025; Malik et al., 2024; Sivarajkumar et al., 2024). Limitações de contexto, degradação do desempenho e ausência de reprodutibilidade acrescentam riscos adicionais, especialmente em aplicações médicas e forenses (Lopez et al., 2025; Vaid et al., 2024; Yin et al., 2025).

Em resposta a esses desafios, diversas estratégias têm sido propostas. Modelos menores e mais eficientes, técnicas de quantização, LoRA e otimizações de treinamento reduzem a pegada computacional e permitem um uso mais viável dos sistemas (Sarker et al., 2024; Faleiros, 2016). No entanto, mesmo com tais abordagens, permanecem limitações estruturais que precisam ser consideradas no planejamento e na implementação de SADs.

Dessa forma, as limitações intrínsecas dos LLMs exigem atenção constante. Para que sistemas de apoio à decisão mantenham confiabilidade, segurança e imparcialidade, é fundamental combinar mitigação de riscos, avaliação contínua dos modelos e entendimento claro de suas restrições técnicas e éticas. Dessa forma é possível utilizar LLMs de maneira segura e responsável em contextos críticos.

3.4.5 Técnicas de Otimização e Especialização de LLMs

Conforme discutido na subseção anterior, os LLMs apresentam limitações significativas que impactam sua confiabilidade, segurança e aplicabilidade em Sistemas de Apoio à Decisão, sobretudo em contextos médico-legais e jurídicos. Fenômenos como alucinações, vieses e alto custo computacional reduzem a robustez das soluções propostas e limitam a aceitação prática desses modelos (Jurafsky e Martin, 2025; Yin et al., 2025; Su et al., 2025). Para mitigar tais problemas, diferentes técnicas de especialização e otimização têm sido desenvolvidas, visando tanto aprimorar a qualidade das respostas quanto adequar os modelos a cenários específicos, reduzindo riscos e ampliando a eficiência de uso.

Entre as estratégias mais relevantes estão a *engenharia de prompt* e o *fine-tuning*, que atuam em níveis distintos de intervenção. A primeira, por meio da formulação cuidadosa de instruções textuais, explora a capacidade do modelo de ajustar seu comportamento sem necessidade de retrainamento ou grande custo computacional (Zhang e Zhang, 2025; Liu et al., 2025). Já o *fine-tuning*, ao adaptar parâmetros internos do modelo com base em dados adicionais, permite incorporar terminologias especializadas e adequar-se a domínios restritos,

como os setores médico e jurídico (Garcia-Carmona et al., 2025; Mukund e Easwarakumar, 2025). Ambas as técnicas, portanto, emergem como respostas práticas aos desafios previamente identificados, ainda que apresentem limitações próprias que exigem análise crítica.

Nas subseções seguintes, serão detalhados os fundamentos, aplicações e limitações da *engenharia de prompt* e do *fine-tuning*, destacando seu papel na mitigação das restrições estruturais dos LLMs e seu potencial de integração em sistemas de apoio à decisão críticos.

3.4.5.1 Engenharia de Prompt

A engenharia de prompt, também chamada de *prompt engineering*, constitui-se no processo de concepção e formulação de instruções textuais destinadas a orientar Modelos de LLMs para a realização de tarefas específicas. Segundo Jurafsky e Martin (2025), trata-se de um processo de elaboração de *prompts* eficazes, em que a simples definição do enunciado influencia diretamente a qualidade da resposta gerada. Essa prática, como reforça Zhang e Zhang (2025), possibilita que o modelo produza resultados mais adequados ao usuário a partir do cuidado no desenho e na construção da entrada textual.

O *prompt* em si é uma sequência de texto que, ao ser submetida ao modelo, desencadeia a geração de tokens condicionados à instrução inicial, permitindo ao sistema avançar progressivamente na produção da resposta (Jurafsky e Martin, 2025). Nesse sentido, a engenharia de prompt configura-se como um recurso fundamental para potencializar aplicações atuais dos LLMs, por incluir no enunciado elementos como contexto, instruções específicas e exemplos, aprimorando a adequação das respostas (Liu et al., 2025).

Entre suas vantagens, destacam-se a facilidade de adaptação a novas tarefas sem necessidade de retrainamento, os baixos requisitos computacionais e a possibilidade de melhorar a consistência das saídas, sobretudo quando o enunciado é claro e explícito (Zhang e Zhang, 2025). Estratégias diversas podem ser exploradas para esse fim. Prompts bem estruturados reduzem ambiguidades linguísticas e aumentam a precisão, ao passo que o *few-shot prompting* se mostra eficaz em cenários com escassez de dados anotados, proporcionando ao modelo exemplos mínimos de como resolver determinada tarefa (Jurafsky e Martin, 2025; Zhang e Zhang, 2025).

Outra técnica relevante é o *Chain-of-Thought* (CoT), que induz o modelo a seguir uma linha de raciocínio passo a passo, o que melhora sua capacidade em resolver problemas complexos e de múltiplas etapas (Zhang et al., 2024; Zhou et al., 2025). Já os *role-playing prompts* permitem atribuir ao modelo papéis específicos, como o de um especialista clínico, adaptando o estilo de resposta às necessidades de determinado contexto (Ke et al., 2025). Além disso, arquiteturas avançadas como o RAG possibilitam ao LLM consultar bases de conhecimento externas, aumentando a confiabilidade, a rastreabilidade das respostas e reduzindo o risco de alucinações (Rego, 2025).

Apesar dos avanços, a aplicação dessa técnica em contextos especializados, como a análise médico-legal, apresenta limitações significativas. Um dos desafios mais evidentes é a ocorrência de *hallucination bias*, em que o modelo gera informações não fundamentadas em fatos (Ghanta et al., 2024). Em ambientes jurídicos ou forenses, tal inconsistência compromete a confiabilidade, pois respostas imprecisas podem levar a interpretações equivocadas (Beauchemin, Khoury e Gagnon, 2024). Soma-se a isso o problema da ausência de explicabilidade, já que os LLMs operam em grande parte como caixas-pretas, dificultando a rastreabilidade e a aceitação de suas inferências em tribunais ou perícias (Ferreira, Pinheiro e Fernandes, 2025; Yin et al., 2025).

Outra limitação importante refere-se à carência de conhecimento específico do domínio, já que os LLMs são treinados em dados amplos e heterogêneos, mas nem sempre especializados, o que os torna menos sensíveis às nuances técnicas de documentos legais e médicos (Yin et al., 2025). Além disso, há o risco de vieses algorítmicos, pois os modelos podem reproduzir desigualdades ou estereótipos presentes nos dados de treinamento, comprometendo a imparcialidade exigida em análises forenses (Ferreira, Pinheiro e Fernandes, 2025).

Os desafios éticos e regulatórios também não podem ser negligenciados. Questões relativas à privacidade, à proteção de dados sensíveis e ao uso legítimo das informações médicas surgem de forma recorrente, sobretudo quando tais tecnologias são aplicadas em investigações legais (Ferreira, Pinheiro e Fernandes, 2025). Há ainda a dificuldade prática de criar e adaptar *prompts* eficazes, visto que enunciados eficientes em um cenário podem ser ineficazes em outro, exigindo conhecimento especializado e constante revisão (Busch et al., 2024).

Por fim, mesmo quando se recorre a técnicas como RAG para enriquecer as respostas, a qualidade da recuperação do contexto pode comprometer o resultado: se os fragmentos relevantes não forem integralmente recuperados, a resposta tende a ser incompleta ou incoerente (Amugongo et al., 2025). Assim, a supervisão humana permanece necessária, seja para validar a adequação dos resultados, seja para garantir que os erros do modelo não comprometam análises em cenários de alta responsabilidade (Yin et al., 2025).

Em síntese, a engenharia de *prompt* apresenta-se como uma técnica versátil, eficiente e de fácil implementação para melhorar a interação com LLMs, mas sua utilização em domínios sensíveis requer cautela redobrada, integração de abordagens híbridas e constante validação por especialistas.

3.4.5.2 Fine-Tuning

O *fine-tuning* (ajuste fino) em modelos de linguagem consiste em adaptar um modelo previamente treinado para lidar com tarefas ou domínios específicos, ajustando seus parâmetros com base em novos dados rotulados. Essa técnica, conforme observado em Jurafsky e Martin (2025), refere-se à continuação do treinamento de um modelo pré-existente, explorando dados adicionais para adequar sua capacidade de representação a contextos particulares. Tal

paradigma mostrou-se consolidado tanto no processamento de linguagem natural quanto em áreas correlatas, como a visão computacional (Rasmy et al., 2020).

Tradicionalmente, o *full fine-tuning* envolve a atualização de todos os pesos do modelo, processo robusto, mas altamente custoso em termos computacionais. Métodos recentes, como o LoRA, introduzem matrizes de baixa dimensionalidade em camadas congeladas, reduzindo drasticamente a quantidade de parâmetros a serem ajustados sem perda expressiva de desempenho (Mukund e Easwarakumar, 2025). De forma semelhante, técnicas de **adapter tuning** possibilitam acoplar módulos específicos ao modelo, permitindo especializações sem necessidade de retreinar toda a arquitetura. Essas estratégias, classificadas como PEFT, têm sido particularmente relevantes para modelos de grande escala.

Nos domínios médicos e forenses, a adaptação via *fine-tuning* mostra-se especialmente útil. Em aplicações clínicas, o processo possibilita capturar terminologias e contextos próprios da área da saúde, como ressaltado em Garcia-Carmona et al. (2025). Modelos como BioBERT e ClinicalBERT evidenciam a eficácia desse ajuste ao lidarem com literatura biomédica e melhorarem a compreensão semântica de diagnósticos e relatórios (Amugongo et al., 2025). Exemplos mais recentes incluem o Med-PaLM 2, construído sobre a base do PaLM por meio de *fine-tuning* específico (Su et al., 2025). No campo jurídico, adaptações semelhantes resultaram em modelos como Legal-BERT e Legal-PEGASUS, capazes de lidar com a complexidade textual de documentos legais (Mukund e Easwarakumar, 2025). Já em aplicações forenses, destaca-se o Forensic-LLM, ajustado com LoRA para operar em investigações e relatórios especializados (Shen et al., 2025).

Além do *fine-tuning* clássico, práticas como *instruction tuning* e RLHF ampliam a capacidade dos modelos de seguir instruções complexas e alinhar-se às preferências humanas. O primeiro utiliza pares de instrução e resposta para refinar a interpretação do modelo, enquanto o segundo busca aproximar suas saídas a padrões éticos e de segurança em domínios sensíveis (Zhou et al., 2025).

Apesar de seu potencial, o *fine-tuning* em contextos médico-legais traz riscos. O uso de dados sensíveis pode expor informações pessoais de pacientes ou indivíduos investigados, exigindo conformidade com legislações como a LGPD (Rajasekar e Vezhaventhan, 2024). Há ainda a possibilidade de introdução ou amplificação de vieses preexistentes nos dados, o que comprometeria a equidade em diagnósticos clínicos ou julgamentos jurídicos (Ferreira, Pinheiro e Fernandes, 2025; Yin et al., 2025). Outro desafio é o risco de *overfitting*, sobretudo quando o conjunto de dados é restrito ou não representativo, limitando a generalização para novos casos (Garcia-Carmona et al., 2025). Ademais, o custo de anotação de dados especializados e a exigência de infraestrutura computacional podem restringir o acesso a tais tecnologias em regiões com menor desenvolvimento tecnológico (Rajasekar e Vezhaventhan, 2024).

Assim, o *fine-tuning* emerge como estratégia eficaz para aproximar LLMs das necessidades de setores altamente especializados, como a medicina e a ciência forense. Contudo, sua

adoção deve ser acompanhada de mecanismos de mitigação de viés, garantias de privacidade e consideração de custos, de modo a equilibrar os ganhos de desempenho com a responsabilidade ética e social.

3.5 Embeddings Semânticos

Os *embeddings* semânticos constituem representações numéricas de palavras, frases ou documentos que capturam relações semânticas e contextuais de forma distribuída [i, j]. Conforme destacado por Khattak et al. (2019), um *embedding* é um vetor real que representa uma palavra ou expressão com base no contexto em que aparece, permitindo que algoritmos de aprendizado de máquina processem o texto de forma eficiente.

Seu impacto é notório em diversas aplicações de PLN, sobretudo em cenários clínicos e médico-legais, como laudos de necrópsia. Entre suas vantagens, destacam-se a redução da dimensionalidade em relação a métodos clássicos (TF-IDF, BoW), a mitigação de ambiguidades lexicais, a preservação de relações semânticas complexas e a capacidade de representar informações contextuais dinâmicas (Jurafsky e Martin, 2025; Khodadad et al., 2025; Permata, Rendika e Julianty, 2025).

Os *embeddings* podem ser classificados como estáticos, que atribuem vetores fixos a cada palavra, ou contextuais, que ajustam as representações conforme o contexto da sentença ou documento, e também como contextuais ou de sentença, que geram vetores para palavras ou para sentenças/parágrafos inteiros, capturando nuances semânticas mais amplas (Khattak et al., 2019; Khodadad et al., 2025; Reys et al., 2020; Zhu et al., 2023).

3.5.1 Embeddings Estáticos

Os *embeddings* estáticos são representações vetoriais de palavras em que cada palavra do vocabulário recebe um vetor fixo, independentemente do contexto em que aparece no texto. Isso significa que a mesma palavra terá sempre a mesma representação, seja em "banco de praça"ou em "banco financeiro"(Khattak et al., 2019; Reys et al., 2020). Esses vetores são geralmente aprendidos a partir de grandes corpora utilizando técnicas de previsão de palavras ou de coocorrência, como é o caso do Word2Vec, GloVe e FastText, permitindo que palavras semanticamente próximas fiquem próximas no espaço vetorial. Essa característica facilita a captura de relações semânticas globais, possibilitando, por exemplo, operações de analogia como “rei” - “homem” + “mulher” $\approx$ “rainha”.

Entre as principais vantagens dos *embeddings* estáticos, destacam-se a simplicidade, o baixo custo computacional e a facilidade de integração em modelos tradicionais de aprendizado de máquina. Eles são particularmente eficientes em tarefas em que o sentido geral das palavras é suficiente, como classificação de texto ou análise de sentimentos (Reys et al., 2020). No entanto, apresentam limitações significativas, principalmente a incapacidade de diferenciar

palavras polissemânticas, pois não consideram o contexto específico da ocorrência da palavra, o que reduz seu desempenho em tarefas que exigem compreensão mais profunda da linguagem natural (Khattak et al., 2019).

Além disso, variações como o FastText tentam mitigar algumas limitações representando palavras como somas de vetores de subpalavras, melhorando o tratamento de palavras raras ou desconhecidas. Apesar dessas extensões, *embeddings* estáticos permanecem mais limitados em comparação com *embeddings* contextuais, que ajustam a representação de cada palavra de acordo com o contexto da sentença ou documento (Khodadad et al., 2025; Zhu et al., 2023).

3.5.1.1 Word2Vec

O Word2Vec é uma abordagem amplamente utilizada para a geração de *embeddings* estáticos, consistindo em um modelo de aprendizado que transforma cada palavra em um vetor fixo em um espaço contínuo de alta dimensão, de modo que termos semanticamente próximos fiquem representados por vetores próximos (Khattak et al., 2019; Reys et al., 2020). O modelo oferece duas arquiteturas principais: *Continuous Bag of Words* (CBOW) e *Skip-gram* (SG).

Reys et al. (2020) e Khattak et al. (2019) descrevem o funcionamento das duas arquiteturas da seguinte forma:

- CBOW: Prediz uma palavra central com base nas palavras que a antecedem e sucedem, ou seja, a partir do contexto.
- SG: Prediz palavras do contexto a partir de uma palavra foco.

Essa estrutura permite capturar relações semânticas e sintáticas entre palavras, possibilitando, por exemplo, operações de analogia, como "rei- "homem"+ "mulher" "rainha". Por se tratar de *embeddings* estáticos, cada palavra recebe sempre a mesma representação, independentemente do contexto em que aparece, o que favorece aplicações em tarefas de processamento de linguagem natural, como classificação de texto, análise de sentimentos e recuperação de informação, mas limita o tratamento de palavras polissemânticas, que podem ter significados diferentes em contextos distintos (Khattak et al., 2019; Reys et al., 2020).

3.5.1.2 GloVe

O *Global Vectors for Word Representation* (GloVe) é outro modelo amplamente utilizado para a geração de *embeddings* estáticos, desenvolvido com o objetivo de combinar as vantagens das abordagens baseadas em matrizes de coocorrência global de palavras e das redes neurais preditivas (Khattak et al., 2019; Reys et al., 2020). O modelo cria vetores de palavras a partir de estatísticas de coocorrência em grandes corpora, representando cada termo em um espaço contínuo de alta dimensão de modo que relações semânticas e sintáticas entre palavras sejam refletidas nas distâncias e direções dos vetores.

Entre suas características técnicas, Reys et al. (2020) e Khattak et al. (2019) destacam:

- Construção de uma matriz de coocorrência global: registra quantas vezes cada palavra aparece no contexto de outra palavra no corpus.
- Fatoração da matriz de coocorrência: o GloVe aplica técnicas de decomposição para gerar vetores densos que preservam informações semânticas e sintáticas importantes.

Como *embeddings* estáticos, os vetores gerados pelo GloVe mantêm uma representação fixa para cada palavra, independentemente do contexto em que ela aparece. Isso facilita seu uso em tarefas de PLN, como análise de sentimentos, classificação de textos e recuperação de informação, mas apresenta limitações no tratamento de palavras polissemânticas, que podem ter múltiplos significados dependendo do contexto (Khattak et al., 2019; Reys et al., 2020). Além disso, GloVe permite capturar operações de analogia semântica, similarmente ao Word2Vec.

3.5.2 Embeddings Contextuais e de Sentença

Os *embeddings* contextuais e de sentença representam uma evolução significativa em relação aos *embeddings* estáticos, ao capturar o significado das palavras com base no contexto em que aparecem. Diferentemente dos *embeddings* estáticos, onde cada palavra tem uma representação fixa, os *embeddings* contextuais ajustam os vetores conforme a sentença ou documento, permitindo que a mesma palavra tenha diferentes representações dependendo do seu uso (Khattak et al., 2019; Khodadad et al., 2025). Isso é particularmente útil para lidar com palavras polissemânticas e expressões idiomáticas, que podem variar seu significado conforme o contexto. Modelos baseados em arquiteturas de transformadores, como BERT, GPT e seus derivados, são amplamente utilizados para gerar esses *embeddings* contextuais. Eles utilizam mecanismos de atenção para considerar todas as palavras na sentença ao gerar a representação de cada palavra, capturando relações semânticas complexas e nuances contextuais (Khattak et al., 2019; Khodadad et al., 2025; Zhu et al., 2023).

3.5.2.1 Sentence-BERT

O Sentence-BERT (SBERT) é uma adaptação do modelo BERT para a geração de *embeddings* de sentenças, projetado para capturar similaridades semânticas entre textos mais longos, como frases ou parágrafos (Khodadad et al., 2025; Zhu et al., 2023). Diferentemente do BERT original, que gera *embeddings* para palavras individuais, o SBERT utiliza uma arquitetura *Siamese* ou *Triplet*, onde duas ou mais sentenças são processadas simultaneamente para aprender representações vetoriais que refletem suas relações semânticas. Através de técnicas de pooling, o SBERT produz um vetor fixo para cada sentença, permitindo medir a similaridade entre elas usando métricas como a similaridade do cosseno.

A vantagem do SBERT reside na sua capacidade de capturar o contexto dinâmico e as relações semânticas completas entre sentenças, superando as limitações dos *embeddings* estáticos como o Word2Vec, especialmente em tarefas que envolvem compreensão de textos mais longos e complexos (Khodadad et al., 2025; Zhu et al., 2023). SBERT adapta modelos BERT para gerar vetores de sentenças semanticamente coesos, permitindo medir similaridade entre textos mais longos:

3.5.2.2 OpenAI Embeddings

Os *OpenAI Embeddings* são representações vetoriais geradas por modelos de linguagem da OpenAI, projetadas para capturar a semântica de palavras, sentenças ou documentos inteiros em um espaço contínuo de alta dimensão (Khodadad et al., 2025; Zhu et al., 2023). Diferentemente de *embeddings* estáticos tradicionais, como Word2Vec e GloVe, os OpenAI *Embeddings* podem refletir o contexto em que uma palavra ou expressão aparece, embora também possam ser utilizados como representações fixas para frases ou documentos. Cada entrada textual é convertida em um vetor numérico que captura nuances semânticas e relações de similaridade com outros vetores, permitindo que esses *embeddings* sejam aplicados em tarefas de busca semântica, agrupamento de documentos, detecção de similaridade, recomendação e classificação de texto (Khodadad et al., 2025). A capacidade de adaptação ao contexto permite superar limitações de polissemia comuns em *embeddings* estáticos, de modo que palavras com significados diferentes em contextos distintos podem ser representadas de forma distinta. Essa flexibilidade amplia o desempenho em tarefas de Processamento de Linguagem Natural, especialmente em sistemas de recuperação de informação e em arquiteturas de RAG, nas quais a similaridade semântica entre consultas e documentos é um fator crítico para a qualidade dos resultados (Khodadad et al., 2025; Zhu et al., 2023).

3.5.2.3 Modelos Especializados: BioBERT e ClinicalBERT

O avanço do PLN no campo da saúde e das ciências biomédicas motivou o surgimento de modelos especializados, entre os quais se destacam o BioBERT e o ClinicalBERT. Ambos são adaptações do BERT original, mas treinados em corpora específicos, de modo a captar nuances da linguagem médica e clínica que os modelos de uso geral não conseguem representar adequadamente. Enquanto o BERT genérico foi pré-treinado apenas em textos de domínio aberto, como Wikipedia e BooksCorpus (Lee et al., 2019), as versões especializadas utilizaram grandes bases biomédicas e clínicas, o que lhes conferiu melhor desempenho em tarefas de extração de informação no setor da saúde (Amugongo et al., 2025).

O BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) manteve a estrutura arquitetural do BERT, mas, após uma inicialização com seus pesos, foi submetido a um pré-treinamento adicional em extensos corpora biomédicos, como abstracts do PubMed e artigos completos do PubMed Central (Lee et al., 2019). Essa

adaptação permitiu que o modelo entendesse termos técnicos, abreviações e entidades próprias do vocabulário científico, alcançando desempenhos superiores em tarefas como reconhecimento de entidades biomédicas, extração de relações e resposta a perguntas no domínio médico (Lee et al., 2019). Com versões que chegam a mais de 340 milhões de parâmetros (Khattak et al., 2019), o BioBERT tornou-se referência em mineração de textos biomédicos, consolidando-se como um recurso de alto impacto para aplicações em biotecnologia e saúde (Kishore e Bodapati, 2025).

O ClinicalBERT, por sua vez, foi desenvolvido com foco em textos clínicos, treinado sobre milhões de notas médicas do banco MIMIC-III, um dos maiores repositórios públicos de registros clínicos (Khattak et al., 2019). Seguindo as mesmas tarefas de pré-treinamento do BERT original, como a predição de tokens mascarados e a predição da próxima sentença, o modelo foi adaptado para compreender anotações de prontuários, evoluções clínicas e descrições de procedimentos (Kishore e Bodapati, 2025). Sua arquitetura base possui cerca de 110 milhões de parâmetros e foi concebida para oferecer representações profundas de linguagem clínica, permitindo não apenas a extração de entidades, mas também a identificação de correlações entre diagnósticos e tratamentos, além da possibilidade de apoiar previsões sobre desfechos clínicos (Khattak et al., 2019; Kishore e Bodapati, 2025).

A principal diferença entre esses modelos e os LLMs de uso geral está, portanto, na especialização de domínio. Como argumentam os estudos, modelos genéricos apresentam desempenho limitado em mineração de textos biomédicos devido à alta densidade de termos técnicos e expressões específicas, ao passo que versões ajustadas em corpora especializados alcançam melhorias expressivas em precisão e recall (Khodadad et al., 2025; Lee et al., 2019). Isso os torna ferramentas valiosas para contextos em que a linguagem é técnica e pouco acessível, como nos laudos clínicos e médico-legais.

No âmbito da análise de laudos de necrópsia, a utilização do BioBERT e do ClinicalBERT mostra-se particularmente promissora. A capacidade de reconhecimento de entidades médicas pode apoiar a extração de termos anatômicos, diagnósticos e circunstâncias descritas nos documentos, enquanto a codificação automática pode auxiliar na atribuição de classificações padronizadas, como os códigos da CID-10, a partir das descrições textuais de causa mortis e achados necroscópicos (Coutinho e Martins, 2022). Além disso, a organização e recuperação de informações a partir de relatórios extensos pode ser facilitada pela integração de ontologias, acelerando fluxos periciais e otimizando a interpretação dos dados (Farias e Pinho, 2016; Rego, 2025).

Entretanto, a aplicação desses modelos ao contexto forense em português enfrenta limitações relevantes. Embora já existam experiências com ajustes de modelos ao domínio clínico em língua portuguesa, ainda são escassos os trabalhos voltados ao processamento de textos médico-legais nacionais (Coutinho e Martins, 2022). Além disso, desafios relacionados à interpretabilidade, ao risco de vieses e à falta de dados anotados em português restringem a

utilização direta desses modelos, exigindo adaptações metodológicas e éticas específicas para o campo jurídico-forense (Cho et al., 2024; Yin et al., 2025).

Assim, BioBERT e ClinicalBERT representam avanços relevantes no processamento de linguagem especializada, oferecendo ganhos expressivos em precisão e compreensão contextual em domínios médicos e biomédicos. No entanto, a sua aplicação a laudos de necrópsia e documentos forenses demanda esforços adicionais de adaptação linguística, além de cuidados éticos e técnicos para garantir confiabilidade e validade nos cenários práticos.

3.5.3 Captura de Relações Semânticas e Contextuais

Os *embeddings* capturam semântica e contexto transformando unidades textuais em vetores densos, onde a proximidade vetorial indica similaridade (Jurafsky e Martin, 2025; Khattak et al., 2019). Modelos estáticos preservam relações locais ou globais entre palavras, enquanto modelos contextuais adaptam os vetores a cada ocorrência da palavra no texto.

A similaridade entre textos pode ser quantificada usando métricas como o cosseno do ângulo entre vetores (*cosine similarity*), o que fundamenta aplicações em clustering, recuperação de informação e raciocínio baseado em casos clínicos (Faleiros, 2016; Garcia-Carmona et al., 2025).

3.5.4 Vantagens, Aplicações e Desafios

Os *embeddings* oferecem diversas vantagens em relação às representações clássicas de texto, como TF-IDF ou Bag-of-Words. Entre os principais benefícios destacam-se a captura de contexto e significado, incorporando a ordem das palavras e relações semânticas (Noll et al., 2025; Reys et al., 2020), a utilização de vetores densos de baixa dimensão, que favorece a eficiência computacional e a generalização (Guti'errez et al., 2022; Jurafsky e Martin, 2025), e a redução do esforço em *feature engineering*, já que os *embeddings* aprendem diretamente a partir do corpus sem necessidade de extração manual de características (Khattak et al., 2019).

Na aplicação a laudos de necrópsia, os *embeddings* possibilitam a codificação automatizada de textos livres, como a atribuição de códigos ICD-10, utilizando *embeddings* contextuais para capturar nuances médicas (Coutinho e Martins, 2022; Khodadad et al., 2025). Eles também permitem a captura de contextos forenses complexos, com modelos como Doc2Vec e ClinicalBERT representando relações entre termos técnicos e descrições de lesões (Mujtaba et al. (2018a); 281), e a integração de dados textuais com conceitos clínicos e códigos UMLS ou ICD-10, enriquecendo a análise semântica (36; 37). Além disso, oferecem suporte a sistemas de IA para classificação, recuperação e raciocínio em radiologia forense e análise de laudos médico-legais (6, 9; 16; 38).

No entanto, seu uso apresenta desafios. A qualidade do texto, frequentemente marcada por erros, abreviações e frases curtas, pode prejudicar o treinamento dos *embeddings* (Ferreira

et al., 2023). Termos fora do vocabulário (OOV) e variantes linguísticas exigem mecanismos específicos para tratamento (Duarte et al., 2018). A interpretação e validação dos resultados devem atender critérios de explicabilidade e conformidade legal (Barros et al., 2025; Ferreira, Pinheiro e Fernandes, 2025), enquanto polissemia e ambiguidade técnica podem limitar o desempenho em textos extensos ou com linguagem especializada dispersa (Khattak et al., 2019; Pereira, 2013; Rego, 2025).

3.5.5 Comparação de Embeddings Semânticos

Para consolidar as informações apresentadas sobre diferentes técnicas de *embeddings* semânticos, a Tabela 2 resume os principais tipos, princípios, vantagens, limitações e aplicações de cada abordagem.

Tabela 2 – Comparativo entre técnicas de embeddings semânticos

Tipo	**Técnica**	**Princípio**	**Vantagens**	**Limitações**	**Aplicações**
Estático	Word2Vec	CBOW / Skip-gram; predição local	Vetor fixo por palavra; captura relações sintáticas e semânticas locais	Não captura contexto dinâmico; limitações em frases longas	Similaridade lexical, recuperação simples de termos, clustering de palavras
Estático	GloVe	Co-ocorrência global; matriz TCM	Captura relações globais entre palavras; vetores densos	Não contextual; depende da estatística global do corpus	Análise semântica de corpus grande, representação de vocabulário amplo
Contextual	Sentence-BERT	Redes Siamese/Triplet sobre BERT; pooling de sentenças	Captura contexto dinâmico; vetor semântico por sentença	Requer maior processamento; depende de pré-treinamento	Similaridade de sentenças, clustering, RAG, avaliação de relevância
Contextual/Denso	OpenAI Embeddings	Transformers proprietários; aprendizado contrastivo	Otimizado para recuperação; suporte a alta dimensionalidade e textos longos	Proprietário; uso dependente de API; complexidade computacional	Sistemas RAG, classificação semântica, busca semântica em grandes volumes de texto

3.5.6 Métricas de Similaridade e Avaliação

A avaliação da similaridade entre documentos e a eficácia de sistemas de recuperação de informação requer métricas quantitativas capazes de mensurar tanto a proximidade semântica entre representações vetoriais quanto o desempenho do sistema na identificação de documentos relevantes (Jurafsky e Martin, 2025; Silva et al., 2020). Neste contexto, métricas de similaridade como a Similaridade de Cosseno fornecem uma medida de quão próximos estão os vetores de características extraídos de textos, sendo amplamente utilizadas em tarefas de Processamento de Linguagem Natural e análise de *embeddings* (Permata, Rendika e Julianty, 2025; Oyelade e Ezugwu, 2020).

Por outro lado, métricas de avaliação clássicas, como Precisão, Revocação (*Recall*) e F1-Score, permitem quantificar a qualidade de sistemas de recuperação de informação e classificação, a partir da relação entre acertos, erros e omissões (Silva et al., 2020; Wang et al., 2024). Essas métricas são fundamentais para balancear a capacidade do sistema de

identificar documentos relevantes sem incluir resultados irrelevantes, sendo aplicadas em diferentes domínios, inclusive no campo médico-legal (Reys et al., 2020; Tang et al., 2023).

Assim, Similaridade de Cosseno e métricas de avaliação formam abordagens complementares: enquanto a primeira captura relações semânticas entre documentos, as demais fornecem indicadores objetivos sobre a eficácia do sistema em selecionar e recuperar informações (Jurafsky e Martin, 2025; Rego, 2025). Essa combinação é especialmente relevante em contextos críticos, como a análise de laudos médico-legais, onde decisões apoiadas por resultados automatizados podem gerar implicações práticas e éticas (Barros et al., 2025; Ferreira, Pinheiro e Fernandes, 2025). Dessa forma, esta subseção apresenta, de forma detalhada, essas métricas, destacando suas definições, aplicações e limitações.

3.5.6.1 Similaridade de Cosseno

A Similaridade de Cosseno é uma métrica amplamente utilizada para mensurar a proximidade entre vetores de características em tarefas de PLN, especialmente quando se trabalha com *embeddings* de palavras ou documentos (Permata, Rendika e Julianty, 2025). Matemáticamente, dado dois vetores ($\vec{A}$) e ($\vec{B}$) no espaço de *embedding*, a Similaridade de Cosseno é definida como o cosseno do ângulo (θ) entre eles, conforme a seguinte expressão:

$$sim_{\cos}(\vec{A}, \vec{B}) = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\|_2 \cdot \|\vec{B}\|_2} = \frac{\sum_{i=1}^{n} A_i \cdot B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{3.2}$$

Onde (n) representa a dimensionalidade dos vetores, ($\vec{A}$) e ($\vec{B}$) são os elementos correspondentes de cada vetor, e (||) denota a norma euclidiana do vetor (Jurafsky e Martin, 2025). Valores de Similaridade de Cosseno próximos de 1 indicam alta proximidade semântica, enquanto valores menores sugerem divergência entre os vetores (Oyelade e Ezugwu, 2020).

Em aplicações práticas, como na análise de laudos médico-legais, a Similaridade de Cosseno permite comparar *embeddings* de documentos para identificar correspondências semânticas, independentemente da magnitude dos vetores. Essa abordagem tem sido utilizada em sistemas como o BERTScore, que emprega a Similaridade de Cosseno para calcular métricas de avaliação de qualidade textual, como *precision*, *recall* e *F1-Score* entre *token embeddings* (Sohn et al., 2024).

A métrica se destaca por ser computacionalmente eficiente e por capturar relações de orientação entre vetores, o que a torna mais robusta que medidas de distância absoluta, como a Distância Euclidiana, especialmente em espaços de alta dimensionalidade (Kishore e Bodapati, 2025). Além disso, possibilita uma avaliação "suave" de similaridade, superando limitações de correspondência literal e permitindo capturar significado semântico de forma mais precisa (Jurafsky e Martin, 2025; Sohn et al., 2024).

Entretanto, a Similaridade de Cosseno apresenta limitações em contextos complexos.

Ela não incorpora nuances contextuais detalhadas, subestimando a similaridade de termos frequentes e sendo inadequada para dados heterogêneos, como séries temporais ou imagens, sem pré-processamento específico (Sivarajkumar et al., 2024; Yan e Cheng, 2024). Em domínios técnico-jurídicos e médicos, adaptações como o *soft cosine* podem ser necessárias para considerar relações entre termos semanticamente correlatos, garantindo medidas mais representativas de similaridade textual (Castro e Neves, 2024).

Portanto, a Similaridade de Cosseno constitui uma métrica fundamental e eficiente para comparação de *embeddings*, mas seu uso em contextos técnicos complexos deve ser complementado por estratégias adicionais, garantindo maior fidelidade semântica e relevância nos resultados (Rego, 2025).

3.5.7 Precisão, Revocação e F1-Score

As métricas de Precisão, Revocação (*Recall*) e F1-Score são amplamente empregadas para avaliar o desempenho de sistemas de recuperação de informação e classificação de documentos, incluindo aplicações em laudos médico-legais. Cada uma delas captura aspectos distintos da qualidade do sistema, sendo derivadas da contagem de acertos (*true positives*, TP), erros (*false positives*, FP) e omissões (*false negatives*, FN) (Jurafsky e Martin, 2025; Silva et al., 2020).

No presente contexto, consideramos acertos (TP) como os documentos relevantes corretamente recuperados, enquanto os erros (FP) correspondem a documentos não relevantes incluídos na recuperação. As omissões (FN) representam documentos relevantes não recuperados pelo sistema.

A Precisão mede a proporção de resultados recuperados que são efetivamente relevantes, representando a confiabilidade do sistema em evitar falsos positivos. Em termos matemáticos, é definida como:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.3}$$

Alta precisão indica que a maior parte dos documentos retornados pelo sistema é relevante, aspecto importante em cenários clínicos ou legais nos quais falsos positivos podem gerar custos ou implicações significativas (Ebietomere e Ekuobase, 2019; Wang et al., 2024; Oyelade e Ezugwu, 2020).

A Revocação (*Recall*) avalia a proporção de documentos relevantes que o sistema conseguiu recuperar em relação ao total de documentos relevantes disponíveis:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.4}$$

Essa métrica reflete a capacidade do sistema de não omitir documentos importantes, sendo fundamental quando a perda de informações relevantes pode comprometer decisões ou análises críticas (Silva et al., 2020; Wang et al., 2024).

O F1-Score combina Precisão e Revocação em uma única métrica, utilizando a média harmônica, oferecendo equilíbrio entre os dois aspectos:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.5}$$

Valores mais altos de F1-Score indicam melhor desempenho global, enquanto valores baixos revelam deficiências tanto na precisão quanto na revocação (Reys et al., 2020; Tang et al., 2023; Khanbhai et al., 2021).

Em sistemas de análise de laudos médico-legais, essas métricas refletem diferentes prioridades operacionais. Priorizar Precisão reduz falsos positivos, garantindo que documentos classificados como relevantes sejam realmente pertinentes, mas pode levar à perda de documentos relevantes (*false negatives*). Em contrapartida, priorizar Revocação assegura que a maior parte dos documentos relevantes seja recuperada, embora aumente a probabilidade de incluir documentos irrelevantes (Silva et al., 2020; Tang et al., 2023). O *trade-off* entre essas métricas pode ser visualizado por curvas ROC, que ilustram a relação entre sensibilidade (*recall*) e a taxa de falsos positivos, auxiliando na escolha do ponto de operação adequado para o sistema (Alkan et al., 2025).

Além de considerações técnicas, a escolha e ajuste dessas métricas têm implicações éticas e de justiça, especialmente em contextos clínicos e médico-legais. Sistemas de aprendizado de máquina podem introduzir vieses que afetam desproporcionalmente grupos minoritários, tornando essencial equilibrar Precisão e Revocação de forma a minimizar impactos adversos (Barros et al., 2025; Ferreira, Pinheiro e Fernandes, 2025).

Em resumo, Precisão, Revocação e F1-Score oferecem uma visão abrangente da eficácia de sistemas de recuperação e classificação de documentos, sendo indispensáveis para medir a confiabilidade, a completude e o equilíbrio do desempenho, sobretudo em domínios onde decisões erradas podem ter consequências críticas.

3.6 Retrieval-Augmented Generation

Modelos de linguagem de grande escala armazenam conhecimento em parâmetros que, embora extensos, são estáticos e sujeitos a imprecisões. O Retrieval-Augmented Generation (RAG) surge como resposta a essa limitação ao permitir que o processo de geração seja condicionado por evidências externas recuperadas de coleções atualizadas. Essa estratégia permite ancorar respostas em textos verificáveis, reduzindo a dependência exclusiva da memória paramétrica do modelo e ampliando sua aplicabilidade em domínios que exigem maior fidelidade

factual. (Jurafsky e Martin, 2025; Yang et al., 2025)

O *pipeline* RAG compreende três etapas:

1. indexação — segmentação dos documentos em trechos (*chunks*) e transformação desses trechos em vetores por meio de um codificador;
2. recuperação — busca por similaridade entre o vetor da consulta e os vetores indexados;
3. geração — síntese linguística condicionada nas passagens selecionadas, inseridas no *prompt* ou no prefixo do gerador.

Duas formulações têm sido enfatizadas: a RAG-Sequence, que assume um documento latente único para gerar toda a sequência alvo, e a RAG-Token, que permite marginalizar documentos distintos para cada *token*, conferindo maior flexibilidade na composição de respostas a partir de múltiplas fontes (Lewis et al., 2020). A escolha entre as variantes implica *trade-offs* computacionais e de decodificação, bem como diferenças na maneira como a probabilidade de saída é aproximada e combinada com evidências externas. (Lewis et al., 2020)

Ao fornecer contextos atualizados e verificáveis, o RAG melhora a factualidade das respostas e oferece rastreabilidade, permitindo que a saída seja acompanhada pelos trechos de origem para verificação humana (Jurafsky e Martin, 2025; Rego, 2025). Estudos comparativos mostram ganhos em tarefas de QA e sumarização quando se combina LLMs com recuperação densa, indicando redução nas ocorrências de informações inventadas em relação a geradores não aumentados (Mukund e Easwarakumar, 2025; Lewis et al., 2020). Entretanto, a mitigação das alucinações não é absoluta; falhas de recuperação ou fontes não confiáveis podem perpetuar erros (Bora e Cuayahuitl, 2024; Zhang e Zhang, 2025).

Em contextos médico-legais, onde a confiabilidade documental e a rastreabilidade são essenciais, o RAG apresenta aplicações relevantes: interpretação automatizada de laudos e textos técnicos, apoio à elaboração de relatórios estruturados, recuperação de casos clínicos para raciocínio baseado em experiência e suporte a análises imagéticas quando integrado a pipelines multimodais (Rego, 2025). Além disso, o uso de RAG permite confrontar conclusões periciais com normativas (Brasil, 2025) e bases de referência, contribuindo para a fundamentação técnica das decisões.

A adoção do RAG em domínios sensíveis esbarra em obstáculos técnicos e éticos. Tecnicamente, a presença de ruído, vieses do *retriever* e limitações na indexação podem comprometer a qualidade do contexto recuperado, levando a respostas inconsistentes ou errôneas (Zhang e Zhang, 2025). Eticamente, persistem preocupações quanto a vieses nos dados, transparência de modelos proprietários e responsabilidade por decisões assistidas por IA — temas especialmente delicados em perícias médicas e forenses. A ausência de padronização e métricas específicas para verificar factualidade e relevância em saúde aumenta a complexidade de avaliação (Ferreira, Pinheiro e Fernandes, 2025; Ke et al., 2025; Nascimento et al., 2024).

Embeddings constituem o núcleo da recuperação densa: consultas e trechos são codificados em vetores que permitem medidas de similaridade semântica superiores às abordagens baseadas exclusivamente em léxico (Noll et al., 2025; Zhang e Zhang, 2025). A combinação de técnicas densas e esparsas em estratégias híbridas tende a otimizar precisão e cobertura, desde que o índice vetorial (por exemplo FAISS) e o modelo de embedding sejam adequadamente calibrados para o domínio (Yang et al., 2025; Zhang et al., 2024).

O RAG representa um avanço metodológico capaz de tornar modelos de linguagem mais ancorados e verificáveis em domínios que exigem precisão documental. Seu sucesso prático depende, contudo, de uma engenharia de recuperação robusta, de curadoria rigorosa das fontes e de um arcabouço ético-regulatório que permita integrar tais sistemas em processos periciais e clínicos com transparência e responsabilidade (Amugongo et al., 2025; Garcia-Carmona et al., 2025; Rego, 2025).

3.6.1 Bancos de Dados Vetoriais

Bancos de dados vetoriais, também chamados de índices vetoriais, são estruturas projetadas para armazenar e recuperar representações numéricas de dados textuais, conhecidas como *embeddings*. Esses vetores traduzem informações linguísticas em formas matemáticas de alta dimensionalidade, permitindo que sistemas computacionais comparem e recuperem conteúdos não por mera coincidência de palavras, mas por proximidade semântica. Em arquiteturas do tipo *Retrieval-Augmented Generation* (RAG), esses bancos atuam como componente central, integrando busca semântica por *embeddings* e geração de linguagem natural contextualizada (Rego, 2025).

O processo de indexação normalmente envolve a segmentação de documentos, a conversão de cada trecho em vetores de *embeddings* por meio de modelos pré-treinados e a posterior inserção desses vetores em um índice vetorial. Essa estrutura permite consultas rápidas e eficientes, nas quais uma *query* também é transformada em vetor e comparada com os demais para identificar conteúdos mais próximos em termos de significado (Rego, 2025).

Entre as principais soluções disponíveis destacam-se o FAISS, biblioteca amplamente utilizada em pesquisas e aplicações práticas, notadamente em tarefas de indexação vetorial de larga escala; o Pinecone, serviço em nuvem que oferece escalabilidade e gerenciamento simplificado de índices vetoriais; o Weaviate, plataforma orientada a grafos e com integração nativa a LLMs; e o Milvus, sistema de código aberto voltado para grandes volumes de dados. No caso do FAISS, há registros explícitos de seu uso em sistemas RAG para viabilizar buscas por similaridade semântica (Rego, 2025), enquanto Pinecone e Weaviate diferenciam-se pela oferta de infraestrutura e integração, ainda que não tenham sido detalhados nas fontes utilizadas.

Um ponto central na eficiência desses bancos é a aplicação de algoritmos de busca por vizinhos aproximados, conhecidos como ANN. Essa técnica reduz a complexidade da comparação em coleções extensas, permitindo que a busca não dependa de cálculos exatos de distância

entre todos os vetores, mas sim de aproximações suficientemente boas para manter a relevância dos resultados. Essa abordagem é fundamental em contextos com grandes volumes de dados, como coleções médico-legais, nas quais a recuperação rápida de documentos semelhantes é um requisito essencial para apoiar processos de análise (Rego, 2025).

Apesar das vantagens, a adoção de bancos vetoriais em sistemas aplicados a laudos médico-legais apresenta desafios. Do ponto de vista técnico, a escalabilidade e a indexação são apontadas como gargalos recorrentes, também observados em sistemas de Raciocínio Baseado em Casos, nos quais a qualidade da indexação define diretamente a efetividade da recuperação (Lorenzi, 1998). Além disso, surgem preocupações relacionadas ao custo de armazenamento e à latência em consultas sobre grandes bases de dados. Em paralelo, aspectos éticos e legais merecem destaque, pois o uso de *embeddings* em documentos sensíveis suscita discussões sobre privacidade, representatividade dos dados e aceitabilidade jurídica de evidências geradas por sistemas computacionais (Barros et al., 2025; Ferreira, Pinheiro e Fernandes, 2025; Rajasekar e Vezhaventhan, 2024; Rocha, Nahim e Lemos, 2024).

Assim, bancos de dados vetoriais configuram-se como ferramentas indispensáveis na busca semântica de grandes volumes textuais e na sustentação de sistemas RAG, mas seu uso em domínios críticos como o médico-legal exige não apenas soluções técnicas robustas para indexação e escalabilidade, como também protocolos éticos e legais que garantam confiabilidade e segurança na aplicação.

Tabela 3 – Comparativo entre soluções de bancos de dados vetoriais

Solução	Tipo/Disponibilidade	Principais Vantagens	Limitações	Cenários Indicados
FAISS	Biblioteca open-source (Meta/Facebook)	▪ Alta performance em busca por similaridade; ▪ Suporte a índices para milhões de vetores; ▪ Amplo uso acadêmico e experimental.	▪ Não é um banco de dados completo, mas uma biblioteca; ▪ Exige gerenciamento manual de escalabilidade e persistência.	▪ Pesquisa acadêmica; ▪ Protótipos e sistemas locais de médio porte.
Pinecone	Plataforma SaaS em nuvem	▪ Totalmente gerenciado (infraestrutura, escalabilidade e segurança); ▪ Baixa latência mesmo em grandes volumes; ▪ Integrações nativas com LLMs.	▪ Dependência de serviço pago em nuvem; ▪ Menor flexibilidade em customização.	▪ Sistemas corporativos em produção; ▪ Casos que exigem disponibilidade global e manutenção mínima.
Weaviate	Banco de dados vetorial open-source com suporte a nuvem	▪ Suporte a busca híbrida (vetorial + palavras-chave); ▪ Integração com modelos de linguagem; ▪ Estrutura orientada a grafos.	▪ Curva de aprendizado mais complexa; ▪ Menor maturidade em ambientes corporativos que o Pinecone.	▪ Aplicações que requerem flexibilidade; ▪ Sistemas que combinam semântica e contexto relacional.
Milvus	Banco de dados vetorial open-source (Zilliz)	▪ Alta escalabilidade e suporte a bilhões de vetores; ▪ Comunidade ativa e suporte comercial via Zilliz Cloud; ▪ Bom desempenho para Big Data.	▪ Configuração e manutenção complexas em ambientes locais; ▪ Requer infraestrutura robusta.	▪ Aplicações industriais; ▪ Processamento massivo de dados multimodais.

O comparativo mostra que o FAISS é mais indicado para pesquisas e protótipos, oferecendo alta performance, mas sem recursos nativos de escalabilidade. O Pinecone, por ser um serviço em nuvem, é adequado a ambientes corporativos que exigem praticidade e baixa latência, embora implique custos e dependência externa. O Weaviate agrega flexibilidade ao combinar busca vetorial e semântica em estrutura de grafos, enquanto o Milvus é voltado a cenários de grande escala, suportando bilhões de vetores, mas com maior complexidade de implantação. Assim, a escolha depende do equilíbrio entre desempenho, custo e nível de

controle da infraestrutura.

3.6.2 Utilização de Bancos de Dados Não Relacionais para Busca e Armazenamento

Os bancos de dados não relacionais (NoSQL) representam uma alternativa aos sistemas tradicionais relacionais (SQL), distinguindo-se por sua flexibilidade de esquemas, escalabilidade horizontal e capacidade de manipular dados semi-estruturados ou não estruturados, como textos, embeddings e metadados (Rodrigues et al., 2024; Faleiros, 2016). Diferentemente dos bancos relacionais, que estruturam informações em tabelas fixas, os bancos NoSQL permitem modelagens mais adaptáveis, essenciais para grandes volumes de dados heterogêneos (Barros et al., 2018; Lorenzi, 1998).

No contexto da análise de laudos médico-legais, essa flexibilidade é particularmente relevante. Laudos são documentos longos, ricos em informações contextuais, cujo conteúdo textual e metadados associados não se encaixam de maneira eficiente em esquemas relacionais rígidos. Além disso, a necessidade de consultas rápidas para busca semântica e integração de embeddings torna os bancos NoSQL adequados para suportar sistemas de Recuperação Aumentada por Geração (RAG) e Processamento de Linguagem Natural (PLN) (Hoppe et al., 2021; Rego, 2025).

3.6.2.1 Tipos de Bancos NoSQL e Aplicações

Os bancos de dados NoSQL englobam diferentes arquiteturas, cada uma projetada para atender necessidades específicas de armazenamento e recuperação de dados. Dependendo do tipo de dados, do volume e do padrão de consultas, certas soluções se mostram mais eficientes do que outras. Entre as principais categorias estão os bancos key-value, orientados a documentos e orientados a colunas, além de motores especializados em indexação vetorial (Barros et al., 2018; Lorenzi, 1998; Rego, 2025). Cada uma dessas abordagens apresenta características próprias de escalabilidade, flexibilidade e desempenho, sendo aplicáveis a diferentes cenários, como o armazenamento de documentos médico-legais e a implementação de sistemas de busca semântica e híbrida (Hoppe et al., 2021; Yang et al., 2025).

3.6.2.1.1 Key-Value Stores

Os bancos do tipo key-value, como o Redis, oferecem armazenamento simples e rápido, em que cada chave única aponta para um valor associado. Essa arquitetura é útil para indexação rápida de metadados ou embeddings em memória, permitindo respostas de baixa latência. Sistemas que combinam Redis com índices vetoriais, como FAISS, conseguem executar buscas semânticas eficientes, garantindo a integração entre desempenho e escalabilidade (Rego, 2025).

3.6.2.1.2 Document Stores

Os bancos orientados a documentos armazenam informações em formatos como JSON ou BSON, permitindo que documentos complexos, como laudos clínicos ou médico-legais, sejam armazenados integralmente com seus metadados. O MongoDB, por exemplo, facilita consultas flexíveis e indexações em grande escala. Já o Elasticsearch e o OpenSearch combinam características de banco de documentos com motores de busca, possibilitando tanto dense retrieval via embeddings quanto buscas por palavras-chave (BM25), oferecendo suporte a sistemas híbridos de RAG (Hoppe et al., 2021; Sivarajkumar et al., 2024).

Essa abordagem híbrida é essencial em contextos técnico-especializados, nos quais a busca puramente semântica pode não capturar termos médicos ou legais pouco frequentes. A combinação de buscas vetoriais e lexicais maximiza a relevância dos resultados, permitindo que o sistema considere tanto similaridade semântica quanto precisão lexical (Rego, 2025; Yang et al., 2025).

3.6.2.2 Column-Oriented Stores

Os bancos orientados a colunas, como o Cassandra, são indicados para cenários que exigem alta escalabilidade e tolerância a falhas, sendo capazes de processar grandes volumes de dados distribuídos. Sua estrutura é eficiente para consultas analíticas sobre conjuntos extensos de documentos e metadados, suportando operações de leitura e escrita de forma distribuída (Lorenzi, 1998).

3.6.2.3 Motores de Busca e Indexação Vetorial

Soluções como Elasticsearch e FAISS ilustram a integração entre bancos NoSQL e indexação vetorial. O Elasticsearch fornece um sistema distribuído de busca de texto completo, com suporte nativo a métodos de recuperação densos, permitindo consultas rápidas em grandes volumes de documentos (Hoppe et al., 2021; Sivarajkumar et al., 2024; Yang et al., 2025). Por sua vez, o FAISS oferece índices vetoriais de alta performance para embeddings, possibilitando buscas por similaridade semântica precisas (Rego, 2025). Essa combinação resulta em sistemas escaláveis, capazes de integrar armazenamento, metadados e recuperação semântica de maneira eficiente.

3.6.3 Integração e Benefícios dos Bancos NoSQL

O uso de bancos NoSQL em sistemas médico-legais permite:

- Armazenamento flexível: Capacidade de armazenar documentos complexos, embeddings e metadados em formatos adaptáveis, facilitando a modelagem de dados heterogêneos (Rodrigues et al., 2024; Faleiros, 2016).

- Escalabilidade: Suporte a crescimento horizontal, essencial para lidar com grandes volumes de laudos e registros clínicos (Lorenzi, 1998; Sivarajkumar et al., 2024).
- Consultas rápidas: Indexação eficiente e suporte a buscas híbridas (semânticas e lexicais), otimizando a recuperação de informações relevantes (Hoppe et al., 2021; Rego, 2025).
- Integração com motores de busca: Combinação com soluções como Elasticsearch e FAISS para maximizar a eficiência na recuperação de documentos (Hoppe et al., 2021; Sivarajkumar et al., 2024; Yang et al., 2025).

Embora bancos vetoriais dedicados como Pinecone ou Weaviate ofereçam desempenho especializado em busca de similaridade, a integração com bancos NoSQL permite combinar escalabilidade, gerenciamento de metadados e busca híbrida, fornecendo uma base robusta para sistemas RAG aplicados a documentos médico-legais (Yang et al., 2025).

REFERÊNCIAS

ADHIKARY, Subinay et al. A case study for automated attribute extraction from legal documents using large language models. **Artificial Intelligence and Law**, Springer Science and Business Media LLC, nov. 2024. ISSN 1572-8382. Disponível em: <http://dx.doi.org/10.1007/s10506-024-09425-7>. Citado na página 30.

ALKAN, Muhammet et al. **Artificial Intelligence-Driven Clinical Decision Support Systems**. arXiv, 2025. Disponível em: <https://arxiv.org/abs/2501.09628>. Citado na página 46.

ALMUHANA, Hasanen Abdul-Jawad Hussain; ABBAS, Hawraa Hassan. Classification of specialities in textual medical reports based on natural language processing and feature selection. **Indonesian Journal of Electrical Engineering and Computer Science**, Institute of Advanced Engineering and Science, v. 27, n. 1, p. 163, jul. 2022. ISSN 2502-4752. Disponível em: <http://dx.doi.org/10.11591/ijeecs.v27.i1.pp163-170>. Citado na página 20.

AMADO, Thiago Campos. Bioetica e inovacoes tecnologicas na saude: Desafios eticos e legais na era da inteligencia artificial, bioimpressao e telemedicina. **Revista Contemporanea**, Brazilian Journals, v. 4, n. 10, p. e6358, out. 2024. ISSN 2764-7757. Disponível em: <http://dx.doi.org/10.56083/RCV4N10-204>. Citado 2 vezes nas páginas 7 e 9.

AMUGONGO, Lameck Mbangula et al. Retrieval augmented generation for large language models in healthcare: A systematic review. **PLOS Digital Health**, Public Library of Science (PLoS), v. 4, n. 6, p. e0000877, jun. 2025. ISSN 2767-3170. Disponível em: <http://dx.doi.org/10.1371/journal.pdig.0000877>. Citado 7 vezes nas páginas 8, 30, 31, 35, 36, 40 e 48.

BAJWA, Junaid et al. Artificial intelligence in healthcare: transforming the practice of medicine. **Future Healthcare Journal**, Elsevier BV, v. 8, n. 2, p. e188aEUR"e194, jul. 2021. ISSN 2514-6645. Disponível em: <http://dx.doi.org/10.7861/fhj.2021-0095>. Citado 3 vezes nas páginas 17, 18 e 20.

BARROS, Bianca Matos de et al. Justica algoritmica na saude: Uma revisao sobre deteccao e avaliacao de impactos dos vieses em aprendizado de maquina. In: **Anais do XXV Simposio Brasileiro de Computacao Aplicada a Saude (SBCAS 2025)**. Sociedade Brasileira de Computacao - SBC, 2025. (SBCAS 2025), p. 665–676. Disponível em: <http://dx.doi.org/10.5753/sbcas.2025.7711>. Citado 9 vezes nas páginas 7, 9, 19, 22, 32, 43, 44, 46 e 49.

BARROS, Rhuan et al. Case law analysis with machine learning in brazilian court. In: _____. **Recent Trends and Future Technology in Applied Intelligence**. Springer International Publishing, 2018. p. 857–868. ISBN 9783319920580. Disponível em: <http://dx.doi.org/10.1007/978-3-319-92058-0_82>. Citado na página 51.

BEAUCHEMIN, David; KHOURY, Richard; GAGNON, Zachary. Quebec automobile insurance question-answering with retrieval-augmented generation. In: **Proceedings of the Natural Legal Language Processing Workshop 2024**. Association for Computational Linguistics, 2024. p. 48aEUR"60. Disponível em: <http://dx.doi.org/10.18653/v1/2024.nllp-1.5>. Citado na página 35.

BOESCH, Gaudenz. **Deep Learning vs Machine Learning: Key Differences Explained**. 2023. <https://viso.ai/deep-learning/deep-learning-vs-machine-learning/>. Acesso em: 28 set. 2025. Citado na página 19.

BORA, Arunabh; CUAYAHUITL, Heriberto. Systematic analysis of retrieval-augmented generation-based llms for medical chatbot applications. **Machine Learning and Knowledge Extraction**, MDPI AG, v. 6, n. 4, p. 2355–2374, out. 2024. ISSN 2504-4990. Disponível em: <http://dx.doi.org/10.3390/make6040116>. Citado 2 vezes nas páginas 30 e 47.

BRASIL, Conselho Federal de Medicina (CFM). **Resolução CFM n.º 2.430, de 21 de maio de 2025**. 2025. Conselho Federal de Medicina (CFM). Regulamenta o ato médico pericial e a produção da prova técnica médica; revoga as Resoluções CFM n.º 1.497/1998 e n.º 2.325/2022. Disponível em: <https://sistemas.cfm.org.br/normas/visualizar/resolucoes/BR/2025/2430>. Citado 3 vezes nas páginas 15, 20 e 47.

BRASIL, Ministério da Justiça e Segurança Pública. **Protocolo Nacional de Investigação e Perícias nos Crimes de Feminicídio**. 2024. Documento oficial do Ministério da Justiça e Segurança Pública. Regulamentado pela Portaria nº340, de 22 de junho de 2020; desclassificado o sigilo em 2024. Disponível em: <https://dspace.mj.gov.br/bitstream/1/12487/1/Protocolo%20Nacional%20de%20Investiga%C3%A7%C3%A3o%20e%20Per%C3%ADcias%20nos%20Crimes%20de%20Feminic%C3%ADdio.pdf>. Citado na página 16.

BRASIL, Ministério da Justiça e Segurança Pública (MJSP). **POPs – Perícia Criminal: POPs Medicina Legal (Perícia Criminal 2024 – Vol. 7)**. 2024. Documento oficial. Volume 7, Procedimentos Operacionais Padrão. Disponível em: <https://www.gov.br/mj/pt-br/assuntos/sua-seguranca/seguranca-publica/analise-e-pesquisa/pop/pops-pericia-criminal-2024-medicina-legal-vol-7-pdf.pdf>. Citado 3 vezes nas páginas 7, 8 e 16.

BRASIL, Ministério da Saúde. **Portaria n.º 1.405, de 29 de junho de 2006: institui a Rede Nacional de Serviços de Verificação de Óbito e Esclarecimento da Causa Mortis (SVO)**. 2006. Diário Oficial da União; institui a Rede Nacional de SVO. Disponível em: <https://bvsms.saude.gov.br/bvs/saudelegis/gm/2006/prt1405_29_06_2006.html>. Citado na página 7.

BRASIL, Ministério da Saúde. **Portaria GM/MS n.º 7.236, de 16 de junho de 2025**. 2025. Diário Oficial da União; dispõe sobre normas de saúde pública. Disponível em: <https://bvsms.saude.gov.br/bvs/saudelegis/gm/2025/prt7236_17_06_2025.html>. Citado na página 7.

BUSCH, Felix et al. Large language models for structured reporting in radiology: past, present, and future. **European Radiology**, Springer Science and Business Media LLC, v. 35, n. 5, p. 2589aEUR"2602, out. 2024. ISSN 1432-1084. Disponível em: <http://dx.doi.org/10.1007/s00330-024-11107-6>. Citado 8 vezes nas páginas 20, 24, 25, 28, 29, 30, 31 e 35.

CASTRO, Marcella Queiroz de; NEVES, Ana Regia. Pln e seguranca juridica: Identificacao de divergencias jurisprudenciais com processamento de linguagem natural. In: **Anais do XV Simposio Brasileiro de Tecnologia da Informacao e da Linguagem Humana (STIL 2024)**. Sociedade Brasileira de Computacao, 2024. (STIL 2024), p. 451–456. Disponível em: <http://dx.doi.org/10.5753/stil.2024.245333>. Citado 3 vezes nas páginas 13, 17 e 45.

CHO, Ha Na et al. Task-specific transformer-based language models in health care: Scoping review. **JMIR Medical Informatics**, JMIR Publications Inc., v. 12, p. e49724, nov. 2024. ISSN 2291-9694. Disponível em: <http://dx.doi.org/10.2196/49724>. Citado na página 42.

COÊLHO, Bruna Fernandes. Histórico da medicina legal no brasil. **REDU - Revista Direito UNIFACS**, n. 132, 2011. ISSN 1808-4435. Disponível em: <https://revistas.unifacs.br/index.php/redu/article/viewFile/1505/1188>. Citado na página 15.

COUTINHO, Isabel; MARTINS, Bruno. Transformer-based models for icd-10 coding of death certificates with portuguese text. **Journal of Biomedical Informatics**, Elsevier BV, v. 136, p. 104232, dez. 2022. ISSN 1532-0464. Disponível em: <http://dx.doi.org/10.1016/j.jbi.2022.104232>. Citado 5 vezes nas páginas 20, 28, 29, 41 e 42.

DUARTE, Francisco et al. Deep neural models for icd-10 coding of death certificates and autopsy reports in free-text. **Journal of Biomedical Informatics**, Elsevier BV, v. 80, p. 64–77, abr. 2018. ISSN 1532-0464. Disponível em: <http://dx.doi.org/10.1016/j.jbi.2018.02.011>. Citado 2 vezes nas páginas 20 e 43.

EBIETOMERE, Esingbemi Princewill; EKUOBASE, Godspower Osaretin. A semantic retrieval system for case law. **Applied Computer Systems**, Walter de Gruyter GmbH, v. 24, n. 1, p. 38aEUR"48, maio 2019. ISSN 2255-8691. Disponível em: <http://dx.doi.org/10.2478/acss-2019-0006>. Citado na página 45.

ELHADDAD, Malek; HAMAM, Sara. Ai-driven clinical decision support systems: An ongoing pursuit of potential. **Cureus**, Springer Science and Business Media LLC, abr. 2024. ISSN 2168-8184. Disponível em: <http://dx.doi.org/10.7759/cureus.57728>. Citado na página 20.

FALEIROS, Thiago de Paulo. **Propagacao em grafos bipartidos para extracao de topicos em fluxo de documentos textuais**. Tese (Doutorado) — Universidade de Sao Paulo, Agencia USP de Gestao da Informacao Academica (AGUIA), 2016. Disponível em: <http://dx.doi.org/10.11606/T.55.2016.tde-10112016-105854>. Citado 8 vezes nas páginas 12, 18, 19, 21, 33, 42, 51 e 52.

FARIAS, Karla Meneses; PINHO, Fábio Assis. Ontologias como ferramenta de organização e representação do conhecimento: um olhar sobre os laudos médico-legais. **Informação em Pauta**, v. 1, n. 2, p. 41–65, 2016. ISSN 2525-3468. Disponível em: <https://dialnet.unirioja.es/servlet/articulo?codigo=6254116>. Citado 7 vezes nas páginas 8, 13, 15, 16, 17, 20 e 41.

FERREIRA, Fabricio Alves; PINHEIRO, Iege; FERNANDES, Cristiane Braga. O uso da ferramenta de inteligencia artificial na radiologia forense - revisao. **STUDIES IN HEALTH SCIENCES**, Brazilian Journals, v. 6, n. 3, p. e19872, set. 2025. ISSN 2764-0884. Disponível em: <http://dx.doi.org/10.54022/shsv6n3-054>. Citado 14 vezes nas páginas 7, 9, 16, 17, 20, 22, 33, 35, 36, 43, 44, 46, 47 e 49.

FERREIRA, Patricia Pita et al. Real-time classification of causes of death using ai: Sensitivity analysis. **JMIR AI**, JMIR Publications Inc., v. 2, p. e40965, nov. 2023. ISSN 2817-1705. Disponível em: <http://dx.doi.org/10.2196/40965>. Citado 2 vezes nas páginas 8 e 43.

FRANÇA, Genival Veloso de. **Medicina Legal**. 11. ed. Rio de Janeiro: Guanabara Koogan, 2017. 684 p. ISBN 978-85-277-3227-7. Disponível em: <https://saude.ufpr.br/wp-content/uploads/2021/09/Medicina%20Legal%20Genival%20Veloso%202017.pdf>. Citado 4 vezes nas páginas 7, 8, 15 e 16.

GARCIA-CARMONA, Angel Manuel et al. Leveraging large language models for accurate retrieval of patient information from medical reports: Systematic evaluation study. **JMIR AI**, JMIR Publications Inc., v. 4, p. e68776, jul. 2025. ISSN 2817-1705. Disponível em: <http://dx.doi.org/10.2196/68776>. Citado 8 vezes nas páginas 20, 29, 30, 31, 34, 36, 42 e 48.

GHANTA, Sai Nikhila et al. Applications of chatgpt in heart failure prevention, diagnosis, management, and research: A narrative review. **Diagnostics**, MDPI AG, v. 14, n. 21, p. 2393, out. 2024. ISSN 2075-4418. Disponível em: <http://dx.doi.org/10.3390/diagnostics14212393>. Citado 7 vezes nas páginas 18, 20, 24, 26, 28, 29 e 35.

GUALDANI, Fabricio Amadeu et al. Modelo de mapeamento semantico entre as terminologias de saude cid-10 e snomed-ct. **Em Questao**, FapUNIFESP (SciELO), v. 30, 2024. ISSN 1807-8893. Disponível em: <http://dx.doi.org/10.1590/1808-5245.30.134988>. Citado na página 13.

GUTI'ERREZ, Luis Felipe et al. Similarity analysis of federal reserve statements using document embeddings: the great recession vs. covid-19. **SN Business & Economics**, Springer Science and Business Media LLC, v. 2, n. 7, jun. 2022. ISSN 2662-9399. Disponível em: <http://dx.doi.org/10.1007/s43546-022-00248-9>. Citado 2 vezes nas páginas 23 e 42.

HARSH, Dr et al. Revolutionizing justice: How artificial intelligence is transforming autopsies and crime scene investigations. **International Journal of Environmental Sciences**, Academic Science Publications and Distributions, p. 1366aEUR"1369, ago. 2025. ISSN 2229-7359. Disponível em: <http://dx.doi.org/10.64252/2hntc327>. Citado na página 17.

HOPPE, Christoph et al. Towards intelligent legal advisors for document retrieval and question-answering in german legal documents. In: **2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)**. IEEE, 2021. p. 29aEUR"32. Disponível em: <http://dx.doi.org/10.1109/AIKE52691.2021.00011>. Citado 3 vezes nas páginas 51, 52 e 53.

HULSEN, Tim. Explainable artificial intelligence (xai): Concepts and challenges in healthcare. **AI**, MDPI AG, v. 4, n. 3, p. 652aEUR"666, ago. 2023. ISSN 2673-2688. Disponível em: <http://dx.doi.org/10.3390/ai4030034>. Citado 2 vezes nas páginas 17 e 18.

JURAFSKY, Daniel; MARTIN, James H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. 3rd. ed. Pearson, 2025. Online manuscript released August 24, 2025. Disponível em: <https://web.stanford.edu/~jurafsky/slp3/>. Citado 21 vezes nas páginas 8, 18, 19, 21, 24, 25, 26, 27, 28, 29, 30, 32, 33, 34, 35, 37, 42, 43, 44, 45 e 47.

KE, Yu He et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. **npj Digital Medicine**, Springer Science and Business Media LLC, v. 8, n. 1, abr. 2025. ISSN 2398-6352. Disponível em: <http://dx.doi.org/10.1038/s41746-025-01519-z>. Citado 2 vezes nas páginas 34 e 47.

KHANBHAI, Mustafa et al. Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. **BMJ Health & Care Informatics**, BMJ, v. 28, n. 1, p. e100262, mar. 2021. ISSN 2632-1009. Disponível em: <http://dx.doi.org/10.1136/bmjhci-2020-100262>. Citado na página 46.

KHATTAK, Faiza Khan et al. A survey of word embeddings for clinical text. **Journal of Biomedical Informatics**, Elsevier BV, v. 100, p. 100057, 2019. ISSN 1532-0464. Disponível em: <http://dx.doi.org/10.1016/j.yjbinx.2019.100057>. Citado 7 vezes nas páginas 30, 37, 38, 39, 41, 42 e 43.

KHODADAD, Mohammad et al. **Towards Domain Specification of Embedding Models in Medicine**. arXiv, 2025. Disponível em: <https://arxiv.org/abs/2507.19407>. Citado 6 vezes nas páginas 37, 38, 39, 40, 41 e 42.

KISHORE, Majji Venkata; BODAPATI, Prajna. High-performance semantic similarity analysis for medical research documents using transformer models (biobert/clinicalbert) with wmd/wms. **Journal of Theoretical and Applied Information Technology**, Little Lion Scientific, v. 103, n. 7, p. 2842–2853, abr. 2025. ISSN 1992-8645. Disponível em: <https://www.jatit.org/volumes/Vol103No7/18Vol103No7.pdf>. Citado 2 vezes nas páginas 41 e 44.

LEE, Craig; BRITTO, Shawn; DIWAN, Khaled. Evaluating the impact of artificial intelligence (ai) on clinical documentation efficiency and accuracy across clinical settings: A scoping review. **Cureus**, Springer Science and Business Media LLC, nov. 2024. ISSN 2168-8184. Disponível em: <http://dx.doi.org/10.7759/cureus.73994>. Citado 2 vezes nas páginas 18 e 19.

LEE, Jinhyuk et al. Biobert: a pre-trained biomedical language representation model for biomedical text mining. **Bioinformatics**, Oxford University Press (OUP), v. 36, n. 4, p. 1234aEUR"1240, set. 2019. ISSN 1367-4811. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/btz682>. Citado 3 vezes nas páginas 20, 40 e 41.

LEFÈVRE, Thomas; TOURNOIS, Laurent. Artificial intelligence and diagnostics in medicine and forensic science. **Diagnostics**, MDPI AG, v. 13, n. 23, p. 3554, nov. 2023. ISSN 2075-4418. Disponível em: <http://dx.doi.org/10.3390/diagnostics13233554>. Citado na página 18.

LEWIS, Patrick et al. **Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks**. arXiv, 2020. Disponível em: <https://arxiv.org/abs/2005.11401>. Citado na página 47.

LIU, Yanyi et al. Reducing hallucinations of large language models via hierarchical semantic piece. **Complex & Intelligent Systems**, Springer Science and Business Media LLC, v. 11, n. 5, abr. 2025. ISSN 2198-6053. Disponível em: <http://dx.doi.org/10.1007/s40747-025-01833-9>. Citado 7 vezes nas páginas 21, 24, 29, 30, 32, 33 e 34.

LOPEZ, Ivan et al. Clinical entity augmented retrieval for clinical information extraction. **npj Digital Medicine**, Springer Science and Business Media LLC, v. 8, n. 1, jan. 2025. ISSN 2398-6352. Disponível em: <http://dx.doi.org/10.1038/s41746-024-01377-1>. Citado 3 vezes nas páginas 30, 31 e 33.

LORENZI, F. **Uso da metodologia de raciocínio baseado em casos na investigação de irregularidades nas internações hospitalares**. Tese (Dissertação de Mestrado) — Universidade Federal do Rio Grande do Sul, 1998. Disponível em: <http://hdl.handle.net/10183/26552>. Citado 7 vezes nas páginas 9, 12, 17, 49, 51, 52 e 53.

MACIEL, Daiane Aparecida; FERREIRA, Deborah Pimenta; MARIN, Heimar De Fatima. Padroes de terminologias nacionais para procedimentos e intervencoes na saude. **Revista de**

Administracao em Saude, Associacao Brasileira de Medicina Preventia e Administracao em Saude - ABRAMPAS, v. 18, n. 71, jun. 2018. ISSN 2526-3528. Disponível em: <http://dx.doi.org/10.23973/ras.71.111>. Citado na página 17.

MALIK, Dr Pankaj et al. Enhancing forensic analysis of digital evidence using machine learning: Techniques, applications, and challenges. **International Journal of Innovative Research in Engineering & Multidisciplinary Physical Sciences**, International Journal of Innovative Research in Engineering and Multidisciplinary Physical Sciences, v. 12, n. 5, set. 2024. ISSN 2349-7300. Disponível em: <http://dx.doi.org/10.37082/IJIRMPS.v12.i5.230988>. Citado 2 vezes nas páginas 18 e 33.

MANAKA, Thokozile; ZYL, Terence van; KAR, Deepak. **Improving Cause-of-Death Classification from Verbal Autopsy Reports**. arXiv, 2022. Disponível em: <https://arxiv.org/abs/2210.17161>. Citado na página 8.

MIZIARA, Ivan Dieb; MIZIARA, Carmen Silvia MG; MUNOZ, Daniel Romero. A institucionalizacao da medicina legal no brasil. **Saude, Etica & Justica**, Universidade de Sao Paulo, Agencia USP de Gestao da Informacao Academica (AGUIA), v. 17, n. 2, p. 66, dez. 2012. ISSN 1414-218X. Disponível em: <http://dx.doi.org/10.11606/issn.2317-2770.v17i2p66-74>. Citado na página 16.

MUJTABA, Ghulam et al. Classification of forensic autopsy reports through conceptual graph-based document representation model. **Journal of Biomedical Informatics**, Elsevier BV, v. 82, p. 88aEUR"105, jun. 2018. ISSN 1532-0464. Disponível em: <http://dx.doi.org/10.1016/j.jbi.2018.04.013>. Citado 2 vezes nas páginas 23 e 42.

MUJTABA, Ghulam et al. Prediction of cause of death from forensic autopsy reports using text classification techniques: A comparative study. **Journal of Forensic and Legal Medicine**, Elsevier BV, v. 57, p. 41aEUR"50, jul. 2018. ISSN 1752-928X. Disponível em: <http://dx.doi.org/10.1016/j.jflm.2017.07.001>. Citado na página 23.

MUKUND, S Ajay; EASWARAKUMAR, K. S. Optimizing legal text summarization through dynamic retrieval-augmented generation and domain-specific adaptation. **Symmetry**, MDPI AG, v. 17, n. 5, p. 633, abr. 2025. ISSN 2073-8994. Disponível em: <http://dx.doi.org/10.3390/sym17050633>. Citado 5 vezes nas páginas 28, 29, 34, 36 e 47.

MUREL, Jacob; KAVLAKOGLU, Eda. **O que é um saco de palavras?** 2024. <https://www.ibm.com/br-pt/think/topics/bag-of-words>. Acesso em: 28 set. 2025. Citado na página 23.

Mídia Market. **Como funciona a Inteligência Artificial: conheça seu presente, passado e futuro**. 2024. <https://midia.market/conteudos/consumo/como-funciona-a-inteligencia-artificial/>. Acesso em: 28 set. 2025. Citado na página 17.

NAI, Gisele Alborghetti; MAIA, Fernando Cezar Cardoso. A pericia medica na literatura cientifica. **COLLOQUIUM VITAE**, Associacao Prudentina de Educacao e Cultura (APEC), v. 7, n. 2, p. 40–56, ago. 2015. ISSN 1984-6436. Disponível em: <http://dx.doi.org/10.5747/cv.2015.v07.n2.v137>. Citado na página 15.

NASCIMENTO, Sabrina Maciel et al. Inteligencia artificial e suas implicacoes eticas e legais: revisao integrativa. **Revista Bioetica**, FapUNIFESP (SciELO), v. 32, 2024. ISSN 1983-8042. Disponível em: <http://dx.doi.org/10.1590/1983-803420243729PT>. Citado 8 vezes nas páginas 7, 8, 9, 18, 19, 22, 32 e 47.

NIU, Kunying et al. Retrieve and rerank for automated icd coding via contrastive learning. **Journal of Biomedical Informatics**, Elsevier BV, v. 143, p. 104396, jul. 2023. ISSN 1532-0464. Disponível em: <http://dx.doi.org/10.1016/j.jbi.2023.104396>. Citado na página 29.

NOLL, Richard et al. Enhancing diagnostic precision for rare diseases using case-based reasoning. **Journal of the American Medical Informatics Association**, Oxford University Press (OUP), jun. 2025. ISSN 1527-974X. Disponível em: <http://dx.doi.org/10.1093/jamia/ocaf092>. Citado 2 vezes nas páginas 42 e 48.

OLIVEIRA, Symball Rufino de. Recuperação inteligente de jurisprudência: uma avaliação do raciocínio baseado em casos aplicado a recuperação de jurisprudências no tribunal regional eleitoral do distrito federal. **Revista Ibero-Americana de Ciência da Informação**, v. 2, n. 1, abr. 2011. Disponível em: <https://periodicos.unb.br/index.php/RICI/article/view/1408>. Citado 3 vezes nas páginas 12, 17 e 18.

OYELADE, Olaide N.; EZUGWU, Absalom E. A case-based reasoning framework for early detection and diagnosis of novel coronavirus. **Informatics in Medicine Unlocked**, Elsevier BV, v. 20, p. 100395, 2020. ISSN 2352-9148. Disponível em: <http://dx.doi.org/10.1016/j.imu.2020.100395>. Citado 3 vezes nas páginas 43, 44 e 45.

PEREIRA, Daniel de Menezes. **Aspectos historicos e atuais da pericia medico legal e suas possibilidades de evolucao**. Tese (Doutorado) — Universidade de Sao Paulo, Agencia USP de Gestao da Informacao Academica (AGUIA), 2013. Disponível em: <http://dx.doi.org/10.11606/D.2.2013.tde-17122013-081615>. Citado 6 vezes nas páginas 7, 8, 9, 15, 16 e 43.

PERMATA, Regita; RENDIKA; JULIANTY, Luh Candra. Towards an automated essay evaluation system nlp based text embeddings and similarity metrics. **Digital Zone: Jurnal Teknologi Informasi dan Komunikasi**, Universitas Lancang Kuning, v. 16, n. 1, p. 37aEUR"46, jul. 2025. ISSN 2086-4884. Disponível em: <http://dx.doi.org/10.31849/digitalzone.v16i1.26541>. Citado 4 vezes nas páginas 23, 37, 43 e 44.

PINTO-COELHO, Lus. How artificial intelligence is shaping medical imaging technology: A survey of innovations and applications. **Bioengineering**, MDPI AG, v. 10, n. 12, p. 1435, dez. 2023. ISSN 2306-5354. Disponível em: <http://dx.doi.org/10.3390/bioengineering10121435>. Citado 3 vezes nas páginas 25, 26 e 28.

PIRAIANU, Alin-Ionut et al. Enhancing the evidence with algorithms: How artificial intelligence is transforming forensic medicine. **Diagnostics**, MDPI AG, v. 13, n. 18, p. 2992, set. 2023. ISSN 2075-4418. Disponível em: <http://dx.doi.org/10.3390/diagnostics13182992>. Citado 5 vezes nas páginas 8, 17, 18, 19 e 20.

QI, Minni; ZHOU, Dan; YU, Xiaojun. Application and innovation of artificial intelligence in forensic medicine. **Computer and Information Science**, Canadian Center of Science and Education, v. 17, n. 2, p. 52, out. 2024. ISSN 1913-8989. Disponível em: <http://dx.doi.org/10.5539/cis.v17n2p52>. Citado 2 vezes nas páginas 17 e 18.

RAI, Prakhyath; RAI, B Shamantha. Visualized document similarity framework with the aid of knowledge graph. In: **2022 International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)**. IEEE, 2022. p. 36aEUR"39. Disponível em: <http://dx.doi.org/10.1109/DISCOVER55800.2022.9974739>. Citado na página 8.

RAJASEKAR, K. Prabhu; VEZHAVENTHAN, D. Artificial intelligence revolutionizing legal and forensic practices: A comprehensive analysis. In: **2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)**. IEEE, 2024. p. 1aEUR"12. Disponível em: <http://dx.doi.org/10.1109/icccnt61001.2024.10724685>. Citado 3 vezes nas páginas 9, 36 e 49.

RASMY, Laila et al. **Med-BERT: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction**. arXiv, 2020. Disponível em: <https://arxiv.org/abs/2005.12833>. Citado na página 36.

RAZA, Shaina; SCHWARTZ, Brian. Entity and relation extraction from clinical case reports of covid-19: a natural language processing approach. **BMC Medical Informatics and Decision Making**, Springer Science and Business Media LLC, v. 23, n. 1, jan. 2023. ISSN 1472-6947. Disponível em: <http://dx.doi.org/10.1186/s12911-023-02117-3>. Citado na página 20.

REGO, Daniel Meireles do. **Analise Semantica Automatizada com LLM e RAG para Bulas Farmaceuticas**. arXiv, 2025. Disponível em: <https://arxiv.org/abs/2507.21103>. Citado 19 vezes nas páginas 7, 8, 14, 17, 18, 20, 21, 32, 34, 41, 43, 44, 45, 47, 48, 49, 51, 52 e 53.

REYS, Arthur D. et al. Predicting multiple icd-10 codes from brazilian-portuguese clinical notes. arXiv, 2020. Disponível em: <https://arxiv.org/abs/2008.01515>. Citado 7 vezes nas páginas 23, 37, 38, 39, 42, 44 e 46.

ROCHA, Maria Eduarda Albuquerque; NAHIM, Laura Falci; LEMOS, Yara Vieira. Os impactos da inteligencia artificial na medicina sob a otica deontologica e legislativa brasileira. **Brazilian Journal of Health Review**, South Florida Publishing LLC, v. 7, n. 4, p. e72148, ago. 2024. ISSN 2595-6825. Disponível em: <http://dx.doi.org/10.34119/bjhrv7n4ed.espc-012>. Citado na página 49.

ROCHA, Naila da et al. Natural language processing to extract information from portuguese-language medical records. **Data**, MDPI AG, v. 8, n. 1, p. 11, dez. 2022. ISSN 2306-5729. Disponível em: <http://dx.doi.org/10.3390/data8010011>. Citado na página 13.

RODRIGUES, Fillipe Barros et al. Natural language processing applied to forensics information extraction with transformers and graph visualization. **IEEE Transactions on Computational Social Systems**, Institute of Electrical and Electronics Engineers (IEEE), v. 11, n. 4, p. 4727–4743, ago. 2024. ISSN 2373-7476. Disponível em: <http://dx.doi.org/10.1109/TCSS.2022.3159677>. Citado 2 vezes nas páginas 51 e 52.

SAINI, Rajat et al. Clinical analytics of medical data of patients using natural language processing approach. **International Journal of Intelligent Systems and Applications in Engineering**, v. 12, n. 19s, p. 804–813, Mar. 2024. Disponível em: <https://ijisae.org/index.php/IJISAE/article/view/5213>. Citado 3 vezes nas páginas 21, 22 e 23.

SANTOS, Livia A. dos et al. Um pipeline de pre-processamento de dados textuais em portugues para analise de redes sociais. In: **Anais do XV Simposio Brasileiro de Tecnologia da Informacao e da Linguagem Humana (STIL 2024)**. Sociedade Brasileira de Computacao, 2024. (STIL 2024), p. 463–468. Disponível em: <http://dx.doi.org/10.5753/stil.2024.245373>. Citado 2 vezes nas páginas 21 e 22.

SARKER, Abeed et al. Natural language processing for digital health in the era of large language models. **Yearbook of Medical Informatics**, Georg Thieme Verlag KG, v. 33, n. 01, p. 229aEUR"240, ago. 2024. ISSN 2364-0502. Disponível em: <http://dx.doi.org/10.1055/s-0044-1800750>. Citado 6 vezes nas páginas 25, 29, 30, 31, 32 e 33.

SHEN, Chen et al. **FEAT: A Multi-Agent Forensic AI System with Domain-Adapted Large Language Model for Automated Cause-of-Death Analysis**. arXiv, 2025. Disponível em: <https://arxiv.org/abs/2508.07950>. Citado 4 vezes nas páginas 30, 31, 32 e 36.

SILVA, Carolina Giordani da et al. Snomed-ct as a standardized language system model for nursing: an integrative review. **Revista Gaucha de Enfermagem**, FapUNIFESP (SciELO), v. 41, 2020. ISSN 0102-6933. Disponível em: <http://dx.doi.org/10.1590/1983-1447.2020.20190281>. Citado 3 vezes nas páginas 43, 45 e 46.

SILVA, Rildo Pinto da; PAZIN-FILHO, Antonio. Anonimizacao de textos medicos com processamento de linguagem natural. **Journal of Health Informatics**, Sociedade Brasileira de Informatica em Saude, v. 17, p. 1227, maio 2025. ISSN 2175-4411. Disponível em: <http://dx.doi.org/10.59681/2175-4411.v17.2025.1227>. Citado 4 vezes nas páginas 7, 9, 13 e 21.

SIVARAJKUMAR, Sonish et al. Clinical information retrieval: A literature review. **Journal of Healthcare Informatics Research**, Springer Science and Business Media LLC, v. 8, n. 2, p. 313aEUR"352, jan. 2024. ISSN 2509-498X. Disponível em: <http://dx.doi.org/10.1007/s41666-024-00159-4>. Citado 11 vezes nas páginas 20, 23, 25, 29, 30, 31, 32, 33, 45, 52 e 53.

SOHN, Jiwoong et al. **Rationale-Guided Retrieval Augmented Generation for Medical Question Answering**. arXiv, 2024. Disponível em: <https://arxiv.org/abs/2411.00300>. Citado na página 44.

SOUSA, Rogerio Nogueira de; PRATA, David Nadler. Resumo automatico de textos juridicos usando grafos com vocabulario controlado e algoritmo k-means com words embedding. **REVISTA ESMAT**, Even3, v. 11, n. 18, p. 65–80, out. 2019. ISSN 2177-0360. Disponível em: <http://dx.doi.org/10.34060/reesmat.v11i18.304>. Citado 6 vezes nas páginas 13, 16, 18, 21, 22 e 23.

SU, Hankun et al. Large language models in medical diagnostics: Scoping review with bibliometric analysis. **Journal of Medical Internet Research**, JMIR Publications Inc., v. 27, p. e72062, jun. 2025. ISSN 1438-8871. Disponível em: <http://dx.doi.org/10.2196/72062>. Citado 3 vezes nas páginas 32, 33 e 36.

TANG, Leigh Anne et al. Using natural language processing to predict fatal drug overdose from autopsy narrative text: Algorithm development and validation study. **JMIR Public Health and Surveillance**, JMIR Publications Inc., v. 9, p. e45246, maio 2023. ISSN 2369-2960. Disponível em: <http://dx.doi.org/10.2196/45246>. Citado 3 vezes nas páginas 20, 44 e 46.

TORTORA, Leda. Beyond discrimination: Generative ai applications and ethical challenges in forensic psychiatry. **Frontiers in Psychiatry**, Frontiers Media SA, v. 15, mar. 2024. ISSN 1664-0640. Disponível em: <http://dx.doi.org/10.3389/fpsyt.2024.1346059>. Citado 2 vezes nas páginas 29 e 31.

VAID, Akhil et al. Local large language models for privacy-preserving accelerated review of historic echocardiogram reports. **Journal of the American Medical Informatics Association**, Oxford University Press (OUP), v. 31, n. 9, p. 2097aEUR"2102, abr. 2024. ISSN 1527-974X. Disponível em: <http://dx.doi.org/10.1093/jamia/ocae085>. Citado na página 33.

VASWANI, Ashish et al. **Attention Is All You Need**. arXiv, 2017. Disponível em: <https://arxiv.org/abs/1706.03762>. Citado 2 vezes nas páginas 26 e 27.

VOLONNINO, G. Artificial intelligence and future perspectives in forensic medicine: a systematic review. **LA CLINICA TERAPEUTICA**, Societa Editrice Universo, IT, v. 175, n. 3, p. 193–202, maio 2024. ISSN 1972-6007. Disponível em: <https://doi.org/10.7417/CT.2024.5062>. Citado 2 vezes nas páginas 17 e 18.

WANG, Sheng et al. **ChatCAD: Interactive Computer-Aided Diagnosis on Medical Image using Large Language Models**. arXiv, 2023. Disponível em: <https://arxiv.org/abs/2302.07257>. Citado na página 8.

WANG, Song et al. A natural language processing approach to detect inconsistencies in death investigation notes attributing suicide circumstances. **Communications Medicine**, Springer Science and Business Media LLC, v. 4, n. 1, out. 2024. ISSN 2730-664X. Disponível em: <http://dx.doi.org/10.1038/s43856-024-00631-7>. Citado 3 vezes nas páginas 43, 45 e 46.

XIONG, Guangzhi et al. **Benchmarking Retrieval-Augmented Generation for Medicine**. arXiv, 2024. Disponível em: <https://arxiv.org/abs/2402.13178>. Citado na página 8.

YAN, Aijun; CHENG, Zijun. A review of the development and future challenges of case-based reasoning. **Applied Sciences**, MDPI AG, v. 14, n. 16, p. 7130, ago. 2024. ISSN 2076-3417. Disponível em: <http://dx.doi.org/10.3390/app14167130>. Citado na página 45.

YANG, Qimin et al. Dual retrieving and ranking medical large language model with retrieval augmented generation. **Scientific Reports**, Springer Science and Business Media LLC, v. 15, n. 1, maio 2025. ISSN 2045-2322. Disponível em: <http://dx.doi.org/10.1038/s41598-025-00724-w>. Citado 6 vezes nas páginas 24, 47, 48, 51, 52 e 53.

YANG, Xi et al. A large language model for electronic health records. **npj Digital Medicine**, Springer Science and Business Media LLC, v. 5, n. 1, dez. 2022. ISSN 2398-6352. Disponível em: <http://dx.doi.org/10.1038/s41746-022-00742-2>. Citado 2 vezes nas páginas 31 e 32.

YIN, Zhipeng et al. **Digital Forensics in the Age of Large Language Models**. arXiv, 2025. Disponível em: <https://arxiv.org/abs/2504.02963>. Citado 8 vezes nas páginas 8, 24, 29, 32, 33, 35, 36 e 42.

YUAN, Jiayu. Efficient techniques for processing medical texts in legal documents using transformer architecture. MDPI AG, jan. 2025. Disponível em: <http://dx.doi.org/10.20944/preprints202501.0139.v1>. Citado na página 28.

ZHANG, Wan; ZHANG, Jing. Hallucination mitigation for retrieval-augmented large language models: A review. **Mathematics**, MDPI AG, v. 13, n. 5, p. 856, mar. 2025. ISSN 2227-7390. Disponível em: <http://dx.doi.org/10.3390/math13050856>. Citado 7 vezes nas páginas 8, 21, 32, 33, 34, 47 e 48.

ZHANG, Xiaoming et al. **Hierarchical Retrieval-Augmented Generation Model with Rethink for Multi-hop Question Answering**. arXiv, 2024. Disponível em: <https://arxiv.org/abs/2408.11875>. Citado 2 vezes nas páginas 34 e 48.

ZHOU, Binggui et al. Natural language processing for smart healthcare. **IEEE Reviews in Biomedical Engineering**, Institute of Electrical and Electronics Engineers (IEEE), v. 17, p. 4aEUR"18, 2024. ISSN 1941-1189. Disponível em: <http://dx.doi.org/10.1109/RBME.2022.3210270>. Citado 2 vezes nas páginas 20 e 21.

ZHOU, Shuang et al. Large language models for disease diagnosis: a scoping review. **npj Artificial Intelligence**, Springer Science and Business Media LLC, v. 1, n. 1, jun. 2025. ISSN 3005-1460. Disponível em: <http://dx.doi.org/10.1038/s44387-025-00011-z>. Citado 2 vezes nas páginas 34 e 36.

ZHU, Junlin et al. Semantic matching based legal information retrieval system for covid-19 pandemic. **Artificial Intelligence and Law**, Springer Science and Business Media LLC, v. 32, n. 2, p. 397aEUR"426, mar. 2023. ISSN 1572-8382. Disponível em: <http://dx.doi.org/10.1007/s10506-023-09354-x>. Citado 4 vezes nas páginas 37, 38, 39 e 40.