

UNIVERSIDADE DE SANTA CRUZ DO SUL
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS E PROCESSOS
INDUSTRIAIS
ÁREA DE CONCENTRAÇÃO EM CONTROLE E OTIMIZAÇÃO DE PROCESSOS
INDUSTRIAIS
LINHA DE PESQUISA: INOVAÇÃO E TECNOLOGIA EM SISTEMAS E PROCESSOS
INDUSTRIAIS

Patrick Luiz Martini

SISTEMA ANONIMIZADO DE DADOS A PARTIR DE PRONTUÁRIOS DE SAÚDE

Santa Cruz do Sul

2026

Patrick Luiz Martini

SISTEMA DE DADOS ANONIMIZADO A PARTIR DE PRONTUÁRIOS DE SAÚDE

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Sistemas e Processos Industriais – Mestrado e Doutorado, Universidade de Santa Cruz do Sul – UNISC, como requisito parcial para obtenção do título de Mestre em Sistemas e Processos Industriais.

Orientadora: Profa. Dra. Rejane Frozza

Coorientador: Prof. Dr. Leonel Pablo Carvalho Tedesco

Colaboradora: Profa. Dra. Janine Koepp

Colaborador: Prof. Dr. Marcelo Carneiro

Santa Cruz do Sul

2026

SUMÁRIO

1.	INTRODUÇÃO	14
1.1.	Tema e Problemática.....	14
1.2.	Objetivos.....	15
1.3.	Justificativa	16
1.4.	Organização da Dissertação	13
2.	REFERENCIAL TEÓRICO	16
2.1.	Large Language Models (LLMs)	16
2.1.1.	Arquitetura Transformer	17
2.1.2.	Mecanismo Self-Attention	18
2.1.3.	O papel do decoder.....	18
2.1.4.	LLMs e a Proteção de Dados.....	19
2.2.	Anonimização de dados.....	20
2.3.	Legislação sobre dados pessoais.....	21
2.4.	Prontuários eletrônicos de pacientes	22
2.5.	Considerações do Capítulo	23
3.	METODOLOGIA	25
3.1.	Bibliometria Quantitativa.....	25
3.2.	CrITÉrios de busca e seleção	27
3.3.	Seleção de artigos.....	28
3.4.	Bibliometria Qualitativa	29
3.4.1.	Large language models to identify social determinants of health in electronic health records	30
3.4.2.	A large language model for electronic health records	31
3.4.3.	Natural Language Processing in Electronic Health Records in relation to healthcare decision-making: A systematic review.....	31
3.4.4.	A comparison of large language model versus manual chart review for extraction of data elements from the electronic health record.....	32
3.4.5.	Using Natural Language Processing on Electronic Health Records to Enhance Detection and Prediction of Psychosis Risk	33
3.4.6.	Comparativo dos Trabalhos Relacionados	33
3.5.	Caracterização da Pesquisa.....	37
4.	DESENVOLVIMENTO DO SISTEMA	38
4.1.	Seleção do modelo de linguagem (LLM)	38

4.2.	Arquitetura e lógica de processamento	40
4.3.	Arquitetura de dados	41
4.3.1.	Descrição das Tabelas de Origem	45
4.3.2.	Descrição das Tabelas Anonimizadas.....	47
4.4.	Metodologia de Anonimização e Funcionamento.....	50
4.5.	Engenharia de Prompts e Estratégia de Extração	50
4.5.1.	Prompt de Identificação de Termos (NER)	50
4.5.2.	Prompt de Auditoria e Refino (Self-Correction)	52
4.6.	Arquitetura da Solução e Fluxo de Dados.....	53
4.6.1.	Criptografia Determinística e Proteção de Identificadores.....	56
4.6.2.	Pipeline de Transformação e Anonimização.....	57
4.6.3.	Consolidação de Histórico Unificado.....	57
4.7.	Sistema ANONIM - Sistema de Anonimização de Dados.....	57
4.8.	Avaliação do Sistema ANONIM	60
6.	CONCLUSÃO E TRABALHOS FUTUROS.....	62
	REFERÊNCIAS.....	18
	APÊNDICE A - Artigo 1	69
	APÊNDICE B - Artigo 2	80
	APÊNDICE C - Avaliação do Sistema.....	90

RESUMO

A crescente digitalização dos serviços de saúde e a adoção de prontuários eletrônicos por instituições de saúde têm gerado um volume expressivo de dados clínicos, representando um recurso valioso para pesquisas acadêmicas e o avanço da medicina baseada em evidências. A promulgação da Lei Geral de Proteção de Dados (LGPD) consolidou um marco regulatório fundamental para a proteção da privacidade dos cidadãos, estabelecendo diretrizes para o tratamento de dados pessoais sensíveis, incluindo aqueles referentes à saúde. Nesse cenário, torna-se essencial o desenvolvimento de soluções tecnológicas que permitam o uso secundário de dados clínicos para fins de pesquisa, assegurando plena conformidade com os princípios de proteção de dados estabelecidos pela legislação. Nesse contexto, este trabalho tem como objetivo desenvolver um sistema de anonimização de prontuários de saúde, viabilizando seu uso seguro e regulamentado em pesquisas acadêmicas, em conformidade com a LGPD. A metodologia adotada combinou uma revisão sistemática da literatura, seguindo o método PRISMA, com a implementação de um sistema baseado em Modelos de Linguagem de Grande Escala (LLMs) e técnicas de Processamento de Linguagem Natural (PLN). A seleção do modelo de linguagem considerou requisitos de suporte ao idioma português, privacidade de dados e capacidade de processamento local, resultando na adoção do modelo *Gemma 3 4B*. O sistema implementa um *pipeline* de anonimização em dois estágios, utilizando técnicas de Reconhecimento de Entidades Nomeadas (NER) para identificar e suprimir cinco categorias de identificadores pessoais presentes em prontuários: nomes de pessoas, idades, datas de atendimento, localizações geográficas e organizações. Para dados estruturados, aplica-se criptografia determinística (BLAKE2b) para pseudonimização de identificadores e transformação de datas, preservando as relações temporais e a consistência dos dados. Complementarmente, um *prompt* de auditoria baseado na LGPD atua como camada de verificação para mitigar falsos negativos. O resultado é uma base de dados anonimizada que protege a privacidade dos pacientes, enquanto preserva a utilidade dos dados para análises acadêmicas, contribuindo para o avanço da pesquisa na área da saúde ao facilitar o compartilhamento seguro de informações entre instituições e pesquisadores.

Palavras-chave: Processamento de Linguagem Natural, Anonimização de Dados, Prontuários de Saúde, Lei Geral de Proteção de Dados; Modelos de Linguagem de Grande Escala.

ABSTRACT

The increasing digitalization of healthcare services and the adoption of electronic health records by healthcare institutions have generated a significant volume of clinical data, representing a valuable resource for academic research and the advancement of evidence-based medicine. The enactment of the General Data Protection Law (GDPL) established a fundamental regulatory framework for protecting citizens' privacy, setting guidelines for the processing of sensitive personal data, including health-related information. In this scenario, it is essential to develop technological solutions that allow for the secondary use of clinical data for research purposes while ensuring full compliance with the data protection principles established by legislation. In this context, this study aims to develop a system for the anonymization of health records, enabling their safe and regulated use in academic research in compliance with the LGPD. The adopted methodology combined a systematic literature review, following the PRISMA method, with the implementation of a system based on Large Language Models (LLMs) and Natural Language Processing (NLP) techniques. The selection of the language model considered requirements for Portuguese language support, data privacy, and local processing capacity, resulting in the adoption of the Gemma 3 4B model. The system implements a two-stage anonymization pipeline, using Named Entity Recognition (NER) techniques to identify and suppress five categories of personal identifiers present in medical records: personal names, ages, service dates, geographic locations, and organizations. For structured data, deterministic encryption (BLAKE2b) is applied for the pseudonymization of identifiers and date transformation, preserving temporal relationships and data consistency. Additionally, an LGPD-based audit prompt acts as a verification layer to mitigate false negatives. The result is an anonymized database that protects patient privacy while preserving data utility for academic analysis, contributing to the advancement of healthcare research by facilitating the secure sharing of information between institutions and researchers.

Keywords: Natural Language Processing; Data Anonymization; Health Records; General Data Protection Law (LGPD); Large Language Models.

1. INTRODUÇÃO

A crescente digitalização dos serviços de saúde e a adoção de Prontuários Eletrônicos de Saúde (PES) por hospitais e clínicas têm gerado um volume massivo de dados clínicos. Essas informações tornaram-se um recurso inestimável para a pesquisa acadêmica, impulsionando, por exemplo, estudos epidemiológicos, o treinamento de modelos preditivos e o avanço da medicina baseada em evidências (Kim *et al.*, 2019).

No entanto, a promulgação da Lei Geral de Proteção de Dados Pessoais (LGPD) (Brasil, 2018) introduziu um novo paradigma regulatório. Embora fundamental para resguardar os direitos de privacidade dos cidadãos, agora reconhecidos como um direito fundamental (Brasil, 2022), a legislação impõe desafios operacionais complexos sobre como as instituições realizam o tratamento e o compartilhamento de dados sensíveis. Na prática, o receio de sanções e a falta de mecanismos técnicos adequados frequentemente restringem ou inviabilizam o acesso de pesquisadores a esses repositórios, criando um gargalo significativo para a condução de estudos na área da saúde (Carvalho; Freitag; Santos, 2022).

Diante desse cenário, a problemática desta proposta consiste em desenvolver um sistema que viabilize a disponibilização de dados de prontuários eletrônicos para pesquisas na área da saúde, em conformidade com a LGPD.

1.1. Tema e Problemática

A LGPD impõe desafios em relação a pensar como podem ser realizados os acessos a dados de pacientes, a fim de realizar pesquisas científicas que contribuem para o avanço na área da saúde. Avanços, por exemplo, na compreensão de perfis de pacientes relacionados a certos diagnósticos de doenças; análises preditivas de diagnósticos; identificação de padrões em conjuntos de dados; entre outros.

Os registros eletrônicos em saúde (*Electronic Health Records* - EHR) representam uma fonte valiosa para a pesquisa clínica. Contudo, a natureza sensível dos dados de saúde apresenta desafios significativos de segurança e confidencialidade para sua manipulação, exigindo uso de métodos que possibilitem a desidentificação das informações pessoais. A desidentificação visa garantir que, durante a troca de informações, tanto interna quanto externamente, os dados sensíveis dos pacientes permaneçam protegidos, atendendo às regulamentações governamentais, como a LGPD (Brasil, 2018; Azzouzi *et al.*, 2024).

Atualmente, muitas instituições de saúde têm revisado suas políticas de compartilhamento de dados não estruturados com pesquisadores externos, devido às

dificuldades relacionadas à segurança de informações sensíveis (Abdalla *et al.*, 2020). A LGPD define claramente o conjunto de informações que exigem proteção, reforçando a necessidade de garantir a confidencialidade dos dados. Quando aplicadas adequadamente, técnicas de desidentificação podem representar uma solução eficiente para possibilitar a realização de pesquisas com dados anonimizados, sem violar a privacidade dos indivíduos (Raso *et al.*, 2024).

Diante deste cenário, identificam-se três principais causas da problemática deste tema de pesquisa:

- Uso de registros eletrônicos e sistemas de gerenciamento para a saúde.
- A promulgação da LGPD e suas restrições ao uso e compartilhamento de dados pessoais.
- A necessidade de viabilizar pesquisas acadêmicas que dependem do uso de dados de saúde.

Nesse contexto, esta proposta busca investigar e desenvolver um sistema que envolva técnicas de análise de dados para identificar e anonimizar informações sensíveis presentes em prontuários eletrônicos. Busca-se responder a seguinte pergunta de pesquisa: “Como desenvolver um sistema que viabilize a utilização de dados anonimizados em pesquisas na área da saúde, respeitando as exigências da LGPD?”.

1.2. Objetivos

O objetivo geral é projetar e desenvolver um sistema capaz de gerar uma base de dados anonimizada a partir de prontuários de saúde, possibilitando a realização de pesquisas acadêmicas na área, em conformidade com a Lei Geral de Proteção de Dados.

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

- Estudar diferentes processos de anonimização e limpeza de dados, para identificar a melhor abordagem e técnica a serem utilizadas.
- Modelar o sistema de análise de prontuários eletrônicos para a geração de conhecimento útil, abrangendo tanto informações estruturadas, como informações do paciente, quanto não estruturadas, como registros de evolução do paciente.
- Desenvolver a solução proposta e avaliar, de forma empírica, o modelo de anonimização, analisando tanto a conformidade com a LGPD quanto a manutenção da utilidade dos dados para fins de pesquisa acadêmica.

1.3. Justificativa

O uso crescente de prontuários eletrônicos de saúde por instituições hospitalares e clínicas tem gerado um grande volume de dados valiosos para pesquisas científicas e clínicas. No entanto, a implementação da LGPD, em 2018, impôs novos desafios quanto ao uso e compartilhamento de informações sensíveis dos pacientes, restringindo o acesso aos dados necessários para o desenvolvimento de pesquisas na área da saúde.

No contexto de prontuários eletrônicos, informações sensíveis compreendem todo dado pessoal que, conforme definido pela LGPD (Art. 5º, inciso II), diz respeito, entre outros, a aspectos referentes à saúde de uma pessoa natural. Na prática clínica, essas informações incluem, por exemplo: dados de identificação direta do paciente (nomes, sobrenomes e iniciais), dados demográficos (idade, data de nascimento), dados de localização (endereço, cidade, bairro), vínculos organizacionais (hospitais, clínicas, operadoras de saúde), além do próprio conteúdo clínico dos registros (diagnósticos, prescrições, resultados de exames e evoluções médicas). A combinação desses elementos, mesmo quando isoladamente pareçam inofensivos, pode permitir a re-identificação do indivíduo, o que evidencia a necessidade de técnicas abrangentes de anonimização que atuem sobre múltiplas categorias de dados simultaneamente.

Essa problemática ganha ainda mais relevância quando alinhada aos Objetivos de Desenvolvimento Sustentável (ODS) da Agenda 2030, propostos pela Organização das Nações Unidas (Guia Agenda Brasil 2030, 2020). Especificamente, este estudo insere-se no escopo do ODS 3 – Saúde e Bem-Estar, cujo propósito é “assegurar uma vida saudável e promover o bem-estar para todos, em todas as idades”. O ODS 3 possui 13 metas específicas que abrangem diversas áreas da saúde pública, desde a redução da mortalidade materna até o aumento do financiamento e a capacitação de profissionais da saúde. O atingimento de tais metas, no entanto, depende estritamente do avanço da pesquisa científica e da medicina baseada em evidências. Neste cenário, o presente trabalho atua como um viabilizador tecnológico: ao propor uma ferramenta que permite o uso secundário seguro de dados clínicos massivos presentes em prontuários eletrônicos (Kim *et al.*, 2019), o sistema instrumentaliza pesquisadores da saúde com as informações necessárias para conduzir estudos epidemiológicos e clínicos essenciais para a concretização das metas estabelecidas pela agenda global.

Estudos apontam que, mesmo quando há mecanismos de proteção como a anonimização de dados, o público possui lacunas no entendimento das garantias de privacidade, o que pode afetar o consentimento informado para o uso secundário de informações em pesquisas (Corman *et al.*, 2022). De maneira similar, percepções negativas sobre segurança e privacidade afetam a adoção de prontuários eletrônicos compartilhados, como demonstrado por um estudo realizado

em Hong Kong, que identificou a necessidade de medidas eficazes de proteção e controle para fortalecer a confiança dos pacientes (ONG, 2023).

Neste contexto, torna-se essencial o desenvolvimento de soluções tecnológicas que garantam a proteção e a privacidade dos dados pessoais, conforme exigido pela LGPD, ao mesmo tempo em que possibilitem sua utilização para fins de pesquisa científica. O desenvolvimento de um sistema que realize a anonimização de dados e informações contidos nos prontuários eletrônicos de saúde surge como uma abordagem viável para equilibrar a proteção de dados com a necessidade de avanço científico. Tal solução possibilita que instituições de saúde compartilhem informações de forma segura e regulamentada, gerando conhecimento essencial para:

- Aprimorar a qualidade dos serviços de saúde por meio de análises seguras e confiáveis.
- Promover avanços científicos na área da saúde, baseados em evidências.

Segundo Azzouzi *et al.* (2024), a extração e desidentificação manual de prontuários apresentam um alto custo e lentidão, enquanto métodos baseados em regras falham em capturar as nuances da linguagem clínica. Contudo, soluções baseadas em *Large Language Models* (LLMs) demonstram ser eficazes na extração de informações não estruturadas de textos médicos, garantindo a proteção da privacidade por meio de abordagens de anonimização avançadas (Wiest *et al.*, 2024).

Desta forma, a solução proposta se justifica ao empregar Modelos de Linguagem de Grande Escala (LLMs) para realizar uma desidentificação automatizada e sensível ao contexto clínico. Ao contrário das técnicas tradicionais de mascaramento e supressão de dado, que apresentam limitações na era de *big data* por degradarem o valor da informação (Gadotti *et al.*, 2024), a aplicação destas arquiteturas avançadas possibilita a anonimização com preservação de utilidade (*utility-preserving anonymization*) (Lee *et al.*, 2017). Consequentemente, o projeto alinha-se ao rigor técnico-jurídico exigido pela LGPD e destrava o acesso seguro aos prontuários eletrônicos, viabilizando a extração de *insights* clínicos que são indispensáveis para apoiar a tomada de decisão médica e o aprimoramento dos serviços de saúde (Hossain *et al.*, 2023).

REFERÊNCIAS

- ABDALLA, Mohamed; ABDALLA, Moustafa; RUDZICZ, Frank; HIRST, Graeme. Using word embeddings to improve the privacy of clinical notes. *Journal of the American Medical Informatics Association*, v. 27, n. 6, p. 901–907, 2020. DOI: <https://doi.org/10.1093/jamia/ocaa038>.
- ATIWA, Sumaia et al. Accelerated hardware implementation of BLAKE2 cryptographic hash for blockchain. In: *IEEE CANADIAN CONFERENCE ON ELECTRICAL AND COMPUTER ENGINEERING (CCECE)*, 2020, London. Proceedings [...]. [S. l.]: IEEE, 2020. p. 1–6. DOI: <https://doi.org/10.1109/CCECE47787.2020.9255752>.
- AZZOUZI, Mohamed El; *et al.* Automatic de-identification of French electronic health records: a cost-effective approach exploiting distant supervision and deep learning models. *BMC Medical Informatics and Decision Making*, v. 24, n. 1, 1 dez. 2024. <https://doi.org/10.1186/s12911-024-02422-5>.
- BRASIL. Emenda Constitucional n. 115, de 10 de fevereiro de 2022. Altera a Constituição Federal para incluir a proteção de dados pessoais entre os direitos e garantias fundamentais e para fixar a competência privativa da União para legislar sobre proteção e tratamento de dados pessoais. 2022. Disponível em: https://www.planalto.gov.br/ccivil_03/constituicao/Emendas/Emc/emc115.htm. Acessado em: 25 nov. 2024.
- BRASIL. Lei n. 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/113709.htm. Acessado em: 24 nov. 2024.
- BRASIL. Instituto Brasileiro de Geografia e Estatística. Objetivos de Desenvolvimento Sustentável. 2024. Disponível em: <https://ods.ibge.gov.br/>
- Brown, Tom; MANN, Benjamin; RYDER, Nick; et al. Language models are few-shot learners. *Advances in neural information processing systems*, v. 33, p. 1877-1901, 2020.
- CHUNG, H. W. et al. Scaling Instruction-Finetuned Language Models. [S. l.]: arXiv, 2022. DOI: 10.48550/ARXIV.2210.11416. Disponível em: <https://arxiv.org/abs/2210.11416>. Acesso em: 3 set. 2025.
- CONSELHO FEDERAL DE MEDICINA. Resolução CFM n. 1.638/2002. Define prontuário médico e torna obrigatória a criação da Comissão de Revisão de Prontuários nas instituições de saúde. Brasília, DF: CFM, 2002. Disponível em: https://sistemas.cfm.org.br/normas/arquivos/resolucoes/BR/2002/1638_2002.pdf. Acesso em: 25 nov. 2024.
- CORMAN, Amy *et al.* Public comprehension of privacy protections applied to health data shared for research: An Australian cross-sectional study. *International Journal of Medical Informatics*, v. 167, p. 104859, 2022. DOI: <https://doi.org/10.1016/j.ijmedinf.2022.104859>
- DE CARVALHO, H.E.R.H.; FREITAG, A. E. B.; DOS SANTOS, D. R.. Impacts of the implementation of the General Law for the protection of personal data in brazil: a bibliometric

analysis. *Revista de Gestão e Secretariado - GESEC*, v. 13, n. 3, p. 1398–1411, 2022. DOI: <https://doi.org/10.7769/gesec.v13i3.1412>.

DEVLIN, Jacob et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES (NAACL-HLT), 2019, Minneapolis. Proceedings [...]. Minneapolis: ACL, 2019. p. 4171–4186. DOI: <https://doi.org/10.18653/v1/N19-1423>.

GADOTTI, Andrea *et al.* Anonymization: The imperfect science of using data while preserving privacy. *Science Advances*, v. 10, n. 29, p. eadn7053, 2024.

GE, Jin et al. A comparison of large language model versus manual chart review for extraction of data elements from the electronic health record. *medRxiv*, 2023. DOI: <https://doi.org/10.1101/2023.08.31.23294924>.

GOOGLE. Flan-T5-XL. In: HUGGING FACE. [S. l.], 28 nov. 2023. Disponível em: <https://huggingface.co/google/flan-t5-xl>. Acesso em: 3 set. 2025.

GOOGLE. Gemma 3 Technical Report. In: HUGGING FACE. [S. l.], 25 mar. 2025. Disponível em: <https://huggingface.co/papers/2503.19786>. Acesso em: 2 nov. 2025.

GUEVARA, Marco et al. Large language models to identify social determinants of health in electronic health records. *npj Digital Medicine*, v. 7, n. 6, 2024. Disponível em: <https://www.nature.com/articles/s41746-023-00970-0>.

GUIA AGENDA BRASIL 2030. Integrando ODS, educação e sociedade. São Paulo: Raquel Cabral; Thiago Gehre, 2020. Disponível em: <https://odsbrasil.gov.br/>. Acesso em: 10 nov. 2025.

HEVNER, Alan R.; MARCH, Salvatore T.; PARK, Jinsoo; RAM, Sudha. Design Science in Information Systems Research. *MIS Quarterly*, v. 28, n. 1, p. 75–105, 2004. DOI: <https://doi.org/10.2307/25148625>.

HOSSAIN, Elias et al. Natural Language Processing in Electronic Health Records in relation to healthcare decision-making: A systematic review. *Computers in Biology and Medicine*, v. 155, n. 106649, 2023. DOI: <https://doi.org/10.1016/j.combiomed.2023.106649>.

IRVING, Jessica et al. Using Natural Language Processing on Electronic Health Records to Enhance Detection and Prediction of Psychosis Risk. *Schizophrenia Bulletin*, v. 47, n. 2, p. 405–414, 2021. DOI: <https://doi.org/10.1093/schbul/sbaa126>.

KAMPELOPOULOS, Dimitrios et al. A review of LLMs and their applications in the architecture, engineering and construction industry. *Artificial Intelligence Review*, v. 58, n. 250, p. 1–46, 2025. DOI: <https://doi.org/10.1007/s10462-025-11241-7>.

KANWAL, Tehsin; ANJUM, Adeel; KHAN, Abid. Privacy preservation in e-health cloud: taxonomy, privacy requirements, feasibility analysis, and opportunities. *Cluster Computing*, v. 24, n. 1, p. 293–317, 2021. DOI: <https://doi.org/10.1007/s10586-020-03106-1>.

KIM, Ellen *et al.* The Evolving Use of Electronic Health Records (EHR) for Research. *Seminars in Radiation Oncology*, v. 29, n. 4, p. 354–361, 2019. DOI: <https://doi.org/10.1016/j.semradonc.2019.05.010>.

LEE, H; KIM, S; KIM, J W; CHUNG, Y D. Utility-preserving anonymization for health data publishing. *BMC Medical Informatics and Decision Making*, v. 17, 2017. DOI: <https://doi.org/10.1186/s12911-017-0499-0>.

LIBERATI, Alessandro *et al.* The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *PLOS Medicine*, v. 6, n 7, p. e1000100, 21 jul. 2009. DOI: <https://doi.org/10.1371/journal.pmed.1000100>.

MIENYE, Ibomoiye Domor; SWART, Theo G. A comprehensive review of deep learning: architectures, recent advances, and applications. *Information*, v. 15, n. 12, p. 755, nov. 2024. DOI: <https://doi.org/10.3390/info15120755>.

NAÇÕES UNIDAS. Agenda 2030 para o Desenvolvimento Sustentável. [S. l.]: ONU, 15 set. 2015. Disponível em: <https://brasil.un.org/pt-br/91863-agenda-2030-para-o-desenvolvimento-sustentavel>. Acesso em: 10 jun. 2024.

ONG, Rebecca. Factors affecting patient and public perceptions of the adoption of electronic health record sharing: A Hong Kong study. *International Journal of Medical Informatics*, v. 178, p. 105193, 2023. DOI: <https://doi.org/10.1016/j.ijmedinf.2023.105193>

PENG, Cheng *et al.* Model tuning or prompt tuning? A study of large language models for clinical concept and relation extraction. *Journal of Biomedical Informatics*, v. 153, p. 104630, mar. 2024. DOI: <https://doi.org/10.1016/j.jbi.2024.104630>.

RAFFEL, Colin *et al.* Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, v. 21, n. 1, p. 1-67, jun. 2020. Disponível em: <http://jmlr.org/papers/v21/20-074.html>.

RAHMANI, Amir Masoud; HEMMATI, Atefeh; ABBASI, Shirin. The Rise of Large Language Models: Evolution, Applications, and Future Directions. *Engineering Reports*, v. 7, e 70368, p. 1-29, 2025. DOI: <https://doi.org/10.1002/eng2.70368>.

RAIAAN, Mohaimenul Azam Khan *et al.* A review on large language models: architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, v. 12, p. 26839–26874, 2024. DOI: <https://doi.org/10.1109/ACCESS.2024.3365742>.

RASO, Emanuele; LORETI, Pierpaolo; RAVAZIOL, Michele; BRACCIALE, Lorenzo. Anonymization and Pseudonymization of FHIR Resources for Secondary Use of Healthcare Data. *IEEE Access*, v. 12, p. 44929–44939, 2024. DOI: <https://doi.org/10.1109/ACCESS.2024.3381034>.

SHARIF, Arfan. What Is CRUD? Create, Read, Update, and Delete. *CrowdStrike*, 20 dez. 2022. Disponível em: <https://www.crowdstrike.com/en-us/cybersecurity-101/observability/crud/>. Acesso em: 20 abr. 2026.

VASWANI, Ashish et al. Attention is all you need. In: INTERNATIONAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NIPS), 31., 2017, Long Beach. Proceedings [...]. Long Beach: Curran Associates, 2017. p. 5998–6008. DOI: <https://doi.org/10.48550/arxiv.1706.03762>.

VOKINGER, Kerstin N.; STEKHOVEN, Daniel J.; KRAUTHAMMER, Michael. Lost in Anonymization — A Data Anonymization Reference Classification Merging Legal and Technical Considerations. *Journal of Law, Medicine and Ethics*, v. 48, n. 1, p. 228–231, 2020. DOI: <https://doi.org/10.1177/1073110520917025>.

WEI, Jason *et al.* Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, p. 24824-24837, 2022.

WUEST, Isabella Catharina *et al.* LLM-Alx: An open-source pipeline for Information Extraction from unstructured medical text based on privacy-preserving Large Language Models. *medRxiv*, 2024. DOI: <https://doi.org/10.1101/2024.09.02.24312917>

YANG, Xi et al. A large language model for electronic health records. *npj Digital Medicine*, v. 5, n. 194, 2022. Disponível em: <https://www.nature.com/articles/s41746-022-00742-2>.