Universidade de Brasília – UnB
Campus Gama: Faculdade de Ciências e Tecnologias em Engenharia - FCTE
Programa de Pós-Graduação em Engenharia Biomédica

Detecção de fatores de risco para doenças cardíacas a partir de prontuários eletrônicos, usando técnicas avançadas de processamento de linguagem natural

Taynara de Jesus Carvalho Siqueira

Orientador: Dr. Nilton Correia da Silva

UnB – Universidade de Brasília

FGA – Faculdade Gama

Detecção de fatores de risco para doenças cardíacas a partir de prontuários eletrônicos, usando técnicas avançadas de processamento de linguagem natural

Taynara de Jesus Carvalho Siqueira

Orientador: Dr. Nilton Correia da Silva

Dissertação de Mestrado em
Engenharia Biomédica

Publicação: 207A/2025

Brasília/DF, Maio de 2025

UnB – Universidade de Brasília

FGA – Faculdade Gama

Programa de Pós-Graduação

Detecção de fatores de risco para doenças cardíacas a partir de prontuários eletrônicos, usando técnicas avançadas de processamento de linguagem natural

Taynara de Jesus Carvalho Siqueira

Dissertação de mestrado submetida ao programa de pós-graduação em Engenharia Biomédica da Universidade de Brasília, como parte dos requisitos necessários para a obtenção do grau de mestre em Engenharia Biomédica

Aprovada por:

Dr. Nilton Correia da Silva

(Orientador)

Dr. Cristiano Jacques Miosso

(Examinador interno)

Dr. André Coutinho Castilla

(Examinador externo)

Brasília/DF, Maio de 2025

Ficha Catalográfica

de Jesus Carvalho Siqueira, Taynara
Detecção de fatores de risco para doenças cardíacas a partir de prontuários eletrônicos, usando técnicas avançadas de processamento de linguagem natural
[Distrito Federal], 2024.
42p., (FGA/UnB Gama, Mestrado em Engenharia Biomédica, 2025).
Dissertação de Mestrado em Engenharia Biomédica, Faculdade UnB Gama, Programa de Pós-Graduação em Engenharia Biomédica.

1. Processamento de Linguagem Natural
2. Prontuários Médicos
3. Extração de Informações
4. Doenças Cardiovasculares

I. FGA UnB/UnB.
II. Título (série)

Referência

de Jesus Carvalho Siqueira, Taynara (2025). Detecção de fatores de risco para doenças cardíacas a partir de prontuários eletrônicos, usando técnicas avançadas de processamento de linguagem natural. Dissertação de mestrado em engenharia biomédica, Publicação 207A/2025, Programa de Pós-Graduação, Faculdade UnB Gama, Universidade de Brasília, Brasília, DF, 42p.

Cessão de Direitos

---

tayhcarvalho@gmail.com

Brasília, DF – Brasil

Resumo

As doenças cardiovasculares (DCVs) representam a principal causa de mortalidade em âmbito global, sendo a identificação precoce de fatores de risco fundamental para a redução de sua incidência. No entanto, tais informações frequentemente encontram-se registradas em formato de texto livre nos prontuários eletrônicos, o que dificulta sua extração automatizada.

Esta dissertação propõe e compara duas abordagens para a identificação de fatores de risco cardiovasculares: (1) um *pipeline* tradicional baseado em reconhecimento de entidades nomeadas (*Named Entity Recognition* – NER) com classificação de negação e (2) uma abordagem fundamentada em *Large Language Models* (LLMs), utilizando o modelo DeepSeek-14B com *few-shot prompting*.

Os métodos foram aplicados a um conjunto de prontuários clínicos anotados com cinco fatores de risco: hipertensão, diabetes, obesidade, dislipidemia e tabagismo. O modelo DeepSeek apresentou desempenho superior em termos de *F1-score* geral e sensibilidade, com destaque para a classe "Tabagismo" (F1 = 0,79). Por sua vez, o *pipeline* NER obteve maior precisão em classes como "Obesidade" (0,98) e "Dislipidemia" (0,95).

Os resultados evidenciam o caráter complementar das abordagens avaliadas: enquanto o método tradicional se destaca pela leveza e eficiência computacional, os LLMs demonstram maior robustez na interpretação semântica de textos clínicos. Como perspectivas futuras, propõe-se a aplicação de *instruction tuning* ao modelo DeepSeek, bem como o desenvolvimento de mecanismos automatizados para o cálculo de escores de risco cardiovascular, como o Escore de Framingham.

Palavras-chave: processamento de linguagem natural, extração de informações, fatores de risco, doença cardiovascular, mineração de texto, prontuários médicos.

Abstract

Cardiovascular diseases (CVDs) are the leading cause of mortality worldwide, and the early identification of risk factors is essential to reduce their incidence. However, such information is often recorded as unstructured free text in electronic health records, which hinders automated extraction.

This dissertation proposes and compares two approaches for identifying cardiovascular risk factors: (1) a traditional *pipeline* based on Named Entity Recognition (NER) with negation classification, and (2) a *Large Language Model* (LLM)-based approach using the DeepSeek-14B model with *few-shot prompting*.

The methods were applied to a set of clinical records annotated with five risk factors: hypertension, diabetes, obesity, dyslipidemia, and smoking. The DeepSeek model achieved superior overall *F1-score* and sensitivity, particularly for the "Smoking" class (F1 = 0.79). In contrast, the NER *pipeline* achieved higher precision for classes such as "Obesity" (0.98) and "Dyslipidemia" (0.95).

The results highlight the complementary nature of the evaluated approaches: while the traditional *pipeline* excels in efficiency and computational simplicity, LLMs demonstrate greater robustness in the semantic interpretation of clinical texts. As future work, we propose applying *instruction tuning* to the DeepSeek model and developing automated mechanisms for calculating cardiovascular risk scores, such as the Framingham Risk Score.

Keywords: natural language processing, information extraction, risk factors, cardiovascular disease, text mining, medical records.

Sumário

Lista de Tabelas

Lista de Quadros

Lista de Figuras

Lista de Nomenclaturas e Abreviações

NLP - Natural Language Processing

DCNT - Doenças Crônicas Não Transmissíveis

NER - Named-Entity Recognition

IMC - Índice de Massa Corporal

DM - Diabetes Mellitus

PA - Pressão Arterial

DCV - Doenças cardiovasculares

DRC - Doença Renal Crônica

PEP - Prontuário Eletrônico do Paciente

PEC - Prontuário Eletrônico do Cidadão

CFM - Conselho Federal de Medicina

PLN - Processamento de Linguagem Natural

LLM - Large Language Model

LM - Language Model

IA - Inteligência Artificial

1 Introdução

As doenças cardiovasculares (DCVs) permanecem entre as principais causas de mortalidade e morbidade no mundo, sendo responsáveis por cerca de 17,9 milhões de mortes por ano, segundo a Organização Mundial da Saúde [24]. A maioria desses eventos poderia ser evitada ou mitigada por meio da identificação precoce de fatores de risco, como hipertensão arterial, diabetes mellitus, dislipidemias, obesidade, tabagismo e sedentarismo [29]. A atuação preventiva, baseada em dados clínicos confiáveis e atualizados, é essencial para orientar condutas médicas e políticas públicas eficazes [10].

Com a crescente digitalização dos sistemas de saúde, parte importante das informações clínicas são registradas em prontuários eletrônicos na forma de texto livre, como anotações de evolução, prescrições e laudos médicos. Esse conteúdo textual, por sua natureza não estruturada, é muito variável, subjetivo e carente de padronização terminológica. Além disso, os dados armazenados ainda são frequentemente subutilizados ou se perdem no cotidiano das instituições, dificultando a recuperação de informações e comprometendo a compreensão do fluxo assistencial relacionado aos pacientes [7].

Essa realidade, comum em instituições públicas e privadas, revela um problema estrutural na gestão da informação clínica: embora dados valiosos estejam registrados, eles permanecem ocultos sob a forma de texto livre, fragmentados e redigidos de forma não padronizada. Os Sistemas de Informação em Saúde (SIS), embora representem um avanço na informatização dos serviços, ainda são, em grande parte, voltados ao armazenamento e recuperação pontual de dados, e não à sua interpretação semântica ou análise automatizada [7].

Isso significa que, mesmo com a adoção de prontuários eletrônicos e ferramentas gerenciais, informações críticas como fatores de risco cardiovasculares continuam difíceis de identificar de forma sistemática e em tempo hábil. Como consequência, essas lacunas podem atrasar o diagnóstico, comprometer a estratificação de risco e prejudicar a condução terapêutica, sobretudo em ambientes de alta demanda e escassez de recursos humanos.

Os avanços das técnicas de Processamento de Linguagem Natural (PLN), especialmente com a adoção de modelos baseados em transformers — como o BERT e suas variantes voltadas à área da saúde, como o BioBERT — trouxe ganhos significativos na

interpretação semântica de textos médicos [17]. Iniciativas internacionais, como o banco de dados MIMIC-III [15] e os desafios organizados pelo consórcio i2b2 [35], demonstraram a viabilidade da extração automatizada de informações clínicas relevantes — incluindo fatores de risco, diagnósticos e desfechos — a partir de modelos treinados em conjuntos de dados anotados.

Mais recentemente, modelos de linguagem de grande porte, conhecidos como LLMs (*Large Language Models*), têm ganhado destaque por sua capacidade de compreender e gerar textos de forma muito próxima à linguagem humana. Esses modelos são treinados com bilhões de palavras e conseguem responder perguntas, resumir documentos e até interpretar textos complexos, como os registros médicos [21]. Nesse contexto, Jie Pan et al. [27] propuseram um pipeline que integra LLMs com a expertise de especialistas clínicos para detectar doenças a partir de registros eletrônicos de saúde, sem a necessidade de grandes volumes de dados rotulados. Os resultados demonstraram que o LLM apresentou desempenho superior às abordagens tradicionais, e robustez frente à diversidade textual dos prontuários.

Diante desse cenário, este trabalho propõe e compara duas abordagens para a extração automática de fatores de risco cardiovasculares a partir de prontuários eletrônicos: (1) um *pipeline* tradicional combinando Reconhecimento de Entidades Nomeadas (NER, do inglês *Named Entity Recognition*) com classificação de negação, e (2) uma abordagem inovadora utilizando o LLM DeepSeek-14B [11] via *few-shot prompting* [1]. O estudo foca em soluções *open-source* e computacionalmente eficientes, endereçando três lacunas críticas: (1) falta de modelos clinicamente validados para a língua portuguesa, (2) carência de *benchmarks* públicos para avaliação, e (3) altos custos computacionais de soluções.

2 Fundamentação Teórica

2.1 Doença Cardiovascular (DCV)

Uma das principais causas de morte no Brasil são as doenças cardiovasculares. Segundo dados do Ministério da Saúde, aproximadamente 300 mil pessoas sofrem um Infarto Agudo do Miocárdio (IAM) a cada ano, resultando em óbito em 30% desses casos. Projeções indicam que até 2040, a incidência desses eventos pode aumentar em até 250%. [30]

Uma abordagem efetiva na prevenção do aumento do número de casos das doenças envolve a estratificação dos riscos dos pacientes, seguida pela prescrição de tratamentos específicos. Essas intervenções abrangem desde o estímulo para a melhoria da qualidade de vida com uma dieta equilibrada e maior atividade física até o tratamento das comorbidades existentes no paciente. Os fatores de risco a serem identificados e considerados neste estudo são:

- Obesidade
- Diabetes
- Hipertensão
- Dislipidemia
- Tabagismo
- História familiar

A análise desses fatores permitirá uma abordagem personalizada, visando não apenas a prevenção, mas também a gestão eficaz dos pacientes.

2.2 Obesidade

De acordo com a Organização Mundial de Saúde (OMS), a obesidade e o sobrepeso são definidos como o acúmulo anormal ou excessivo de gordura que apresenta risco à saúde.

A principal causa da obesidade é o desequilíbrio energético entre as calorias consumidas e as calorias gastas. [25]

2.2.1 Diagnóstico

Para a classificação e identificação da obesidade é usado o IMC (Índice de Massa Corporal), que é calculado por meio da divisão da massa em kg pela altura em metros elevada ao quadrado (kg/m2) [28]. O IMC é usado de maneira diferente para crianças e adolescentes. O quadro 2.1 mostra a classificação de obesidade para adultos maiores de 20 anos.

Quadro 2.1. Classificação internacional da obesidade segundo o índice de massa corporal (IMC) para adultos maiores de 20 anos no Brasil. Fonte: [28]

IMC (KM/M²)	CLASSIFICAÇÃO	OBESIDADE GRAU
<18,5	Abaixo do peso	0
18,5 - 24,9	Normal	0
25 - 29,9	Sobrepeso ou pré-obeso	0
30 - 34,9	Obesidade	I
35.0 - 39,9	Obesidade	II
>= 40	Obesidade Grave	III

Outro critério que pode ser usado juntamente com o IMC para a identificação da obesidade é a medida da circunferência abdominal, a faixa de normalidade é diferente para homens e mulheres. Na América do Sul e América Central, é definida a obesidade para homens se a circunferência abdominal for maior ou igual a 90 cm e as mulheres se for maior ou igual 80 cm. [28]

2.2.2 Tratamentos

A obesidade possui diversos tipos de tratamentos: tratamento não farmacológico, tratamento farmacológico, tratamento dietético, terapia cognitivo-comportamental, terapias heterodoxas e suplementos nutricionais para perda de peso e tratamento cirúrgico. [28]

2.3 Diabetes

Diabetes *mellitus* (DM) é um grupo de doenças metabólicas caracterizadas por hiperglicemia resultante de defeitos na secreção de insulina, na ação da insulina ou em ambos. A hiperglicemia crônica do diabetes está associada a danos a longo prazo, disfunção e falência de vários órgãos, especialmente olhos, rins, nervos, coração e vasos sanguíneos [23].

Segundo o Comitê de Especialistas em Diagnóstico e Classificação do Diabetes Mellitus, os principais tipos de diabetes são [23]:

- Diabetes Tipo 1: (destruição das células β, geralmente levando à deficiência absoluta de insulina);
  - Tipo 1A: Imunomediado;
  - Tipo 1B: Idiopático;
- Diabetes Tipo 2 (pode variar de resistência à insulina com a deficiência de insulina a um defeito predominantemente secretor com resistência à insulina);
- Diabetes Gestacional: Hiperglicemia diagnosticada durante a gestação;
- Defeitos genéticos da função das células β;
- Defeitos genéticos na ação da insulina;
- Doenças do pâncreas exócrino;

2.3.1 Diagnóstico

Na maioria dos casos, a pré-diabetes e a diabetes são assintomáticas, então o diagnóstico deve ser feito por meio de exames laboratoriais que medem o nível de glicose no sangue. Os testes são: o teste de glicose no sangue em jejum, o teste de glicose no sangue após 2 horas de ingestão de 75g de glicose (ou teste de tolerância à glicose), teste de glicose ao acaso e o hemoglobina glicada (A1C ou HbA1c) [23].

O médico também pode considerar fatores como sintomas, histórico familiar e resultados de testes anteriores ao fazer o diagnóstico. O quadro 2.2 mostra o intervalo de valores de cada exame para determinar a pré-diabetes ou a diabetes.

2.3.2 Tratamento

O tratamento para diabetes depende do tipo de diabetes e da gravidade da doença, mas geralmente inclui a combinação de três componentes principais [8]:

- Mudanças no estilo de vida: isso inclui uma alimentação saudável, atividade física regular e controle do peso.
- Medicação: para muitas pessoas com diabetes tipo 2, a medicação oral ou injetável é necessária para controlar os níveis de glicose no sangue.

Quadro 2.2. Critérios laboratoriais para diagnóstico de pré-diabetes e DM. (adaptado de [23]) Fonte: [23]

	Glicose em jejum (mg/dL)	**Glicose 2 horas após sobrecarga com 75 g de glicose (mg/dL)**	**Glicose ao acaso (mg/dL)**	**HbA1c (%)**	**Observações**
Pré-diabetes ou risco aumentado para DM	>= 100 e <126*	>= 140 e <200	-	>= 5,7 e <6,5	Positividade de qualquer dos parâmetros confirma diagnóstico de pré-diabetes.
Diabetes estabelecido	>= 126	>= 200	>= 200 com sintomas inequívocos de hiperglicemia	>= 6,5	Positividade de qualquer dos parâmetros confirma diagnóstico de DM. Método de HbA1c deve ser o padronizado. Na ausência de sintomas de hiperglicemia, é necessário confirmar o diagnóstico pela repetição de testes.

- Monitoramento: as pessoas com diabetes precisam monitorar regularmente seus níveis de glicose no sangue para garantir que estejam dentro de uma faixa saudável.

O tratamento para diabetes é geralmente planejado e supervisionado por um médico, que pode prescrever medicamentos, recomendar mudanças no estilo de vida e ajudar a monitorar a doença. Algumas pessoas com diabetes também precisam de ajuda de um educador em diabetes, um nutricionista ou um psicólogo.

2.4 Hipertensão

A hipertensão arterial é uma Doença Crônica Não Transmissível (DCNT) caracterizada pela elevação persistente da pressão arterial (PA), definida como pressão arterial sistólica (PAS) maior ou igual a 140 mmHg e/ou pressão arterial diastólica (PAD) maior ou igual a 90 mmHg, quando medida com técnica adequada, em pelo menos duas ocasiões distintas e na ausência de tratamento medicamentos. [4]

2.4.1 Diagnóstico

O diagnóstico da hipertensão baseia-se, principalmente, na aferição correta da pressão arterial em consultório e/ou por métodos ambulatoriais, como a Monitorização Ambulatorial da Pressão Arterial (MAPA) e a Monitorização Residencial da Pressão Arterial (MRPA). Essa avaliação deve ser feita com uso de equipamentos validados e calibrados, associados à coleta da história clínica (pessoal e familiar), exame físico e exames

laboratoriais complementares [4].

Os esfigmomanômetros auscultatórios ou oscilométricos são os dispositivos preferidos para a medição da PA.

Quadro 2.3. Classificação da pressão arterial em adultos (adaptado de [4])

Categoria	PAS (mmHg)	PAD (mmHg)
Ótima	< 120	< 80
Normal	120–129	80–84
Pré-hipertensão	130–139	85–89
Hipertensão estágio 1	140–159	90–99
Hipertensão estágio 2	160–179	100–109
Hipertensão estágio 3	≥ 180	≥ 110
Hipertensão sistólica isolada	≥ 140	< 90

2.4.2 Tratamento

O tratamento da hipertensão depende do estágio da doença, presença de comorbidades e risco cardiovascular global. De forma geral, compreende duas abordagens complementares [16]:

- **Mudança no estilo de vida:** Intervenção não medicamentosa que inclui perda de peso, redução da ingestão de sal, aumento do consumo de potássio, prática regular de atividade física, abandono do tabagismo, moderação no consumo de álcool, entre outras.
- **Tratamento medicamentoso:** Uso de medicamentos anti-hipertensivos, individualizado conforme o perfil clínico do paciente.

2.5 Dislipidemia

Dislipidemia é caracterizada pela modificação nos níveis séricos de lipídeos, trazendo alterações no perfil lipídico. Embora o termo descreva uma ampla gama de condições, as formas mais comuns de dislipidemia envolvem [13]:

- Níveis elevados de lipoproteínas de baixa densidade (LDL), ou colesterol ruim;
- Níveis baixos de lipoproteínas de alta densidade (HDL), ou colesterol bom;
- Níveis elevados de triglicerídeos;
- Colesterol alto, o que se refere a níveis elevados de LDL e triglicerídeos.

Os níveis de lipídios no sangue podem variar naturalmente de pessoa para pessoa. Contudo, indivíduos com elevados níveis de LDL e triglicerídeos, ou níveis notavelmente baixos de HDL, apresentam geralmente um maior risco de desenvolver aterosclerose. Esse processo ocorre quando depósitos rígidos e gordurosos, conhecidos como placas, se acumulam nos vasos sanguíneos, prejudicando a fluidez do fluxo sanguíneo.

2.5.1 Diagnóstico

Quadro 2.4. Valores referenciais e de alvo terapêutico do perfil lipídico (adultos > 20 anos)[13]. Fonte: [13]

Lípides	Com jejum (mg/dL)	Sem jejum (mg/dL)	Categoria referencial
Colesterol total	< 190	< 190	Desejável
HDL-c	> 40	> 40	Desejável
Triglicérides	< 150	< 175†	Desejável
Categoria de risco			
LDL-c	< 130	< 130	Baixo
	< 100	< 100	Intermediário
	< 70	< 70	Alto
	< 50	< 50	Muito alto
Não HDL-c	< 160	< 160	Baixo
	< 130	< 130	Intermediário
	< 100	< 100	Alto
	< 80	< 80	Muito alto

* Conforme avaliação de risco cardiovascular estimado pelo médico solicitante;
† colesterol total > 310 mg/dL há probabilidade de hipercolesterolemia familiar;
† Quando os níveis de triglicérides estiverem acima de 440 mg/dL (sem jejum) o médico solicitante faz outra prescrição para a avaliação de triglicérides com jejum de 12 horas e deve ser considerado um novo exame de triglicérides pelo laboratório clínico.

Existem várias formas de diagnosticar a dislipidemia, sendo as mais comuns as seguintes:

- Exames de Sangue Lipídicos: Exames de LDL, HDL, Triglicerídeos e Colesterol Total. O intervalo dos valores estão representados no quadro 2.4
- Testes Genéticos: Em casos de suspeita de hipercolesterolemia familiar (uma condição genética que causa altos níveis de colesterol), testes genéticos podem ser realizados para identificar mutações genéticas específicas.
- Avaliação Clínica: O médico realiza uma avaliação clínica completa, levando em consideração os sintomas do paciente, histórico médico e estilo de vida, para obter uma compreensão holística da saúde cardiovascular.

- Avaliação de Comorbidades: A dislipidemia muitas vezes está associada a outras condições médicas, como diabetes e hipertensão. Portanto, o médico pode realizar exames adicionais para avaliar a presença de comorbidades.

2.5.2 Tratamento

Os principais tratamentos da dislipidemia são: Mudança no estilo de vida, melhorando a alimentação e a prática regular de atividade física, e o uso de medicamentos para controlar os níveis lipídicos no organismo[13].

2.6 Prontuário do Paciente

O conceito de prontuário médico pode ser entendido como um registro que guarda todas as informações de saúde de um paciente, que podem ser necessárias a qualquer momento. No passado, esses documentos eram feitos manualmente, com preenchimento manuscrito, o que levava muito tempo e recursos. Entretanto, com o avanço da tecnologia na medicina, o surgimento do prontuário eletrônico do paciente permitiu a automação dessas rotinas, além de tornar seu acesso e manipulação muito mais simples [33].

2.6.1 Prontuário Eletrônico do Paciente - PEP

Prontuários eletrônicos são registros digitais de informações médicas de pacientes ao longo da vida, usados para armazenar, gerenciar e compartilhar informações de saúde entre os profissionais de saúde. Eles incluem informações como histórico médico, exames, prescrições, resultados de laboratório e outras informações relevantes para o tratamento de um paciente. Eles são considerados mais seguros, eficientes e acessíveis do que os prontuários tradicionais em papel e estão se tornando cada vez mais populares em hospitais e clínicas em todo o mundo [33].

Atualmente, esse prontuário digital deve seguir as definições do Conselho Federal de Medicina para ser válido e legal, além de apresentar diferentes variações de acordo com cada clínica e software.

2.6.2 Prontuário Eletrônico do Cidadão - PEC

O PEC é de um sistema brasileiro desenvolvido de forma exclusiva pelos órgãos nacionais, para ser aplicado nas Unidades Básicas de Saúde de todo o país.

O Prontuário Eletrônico do Cidadão (PEC) é um sistema eletrônico de armazena-

mento de informações de saúde de um indivíduo abrangendo apenas o Sistema Público de Saúde (SUS). A ideia é individualizar o registro, integrar as informações, reduzir o retrabalho na coleta de dados dos pacientes [9].

2.6.3 Estrutura do Prontuário Eletrônico do Paciente

Quanto ao conteúdo, o Conselho Federal de Medicina (CFM) orienta que o prontuário eletrônico pode conter [18]:

- Identificação do paciente.
- Evolução médica, de enfermagem e de outros profissionais.
- Exames laboratoriais, radiológicos, entre outros.
- Raciocínio médico.
- Hipóteses diagnósticas.
- Condutas terapêuticas.
- Prescrições médicas.
- Descrições cirúrgicas e fichas anestésicas.
- Resumos de alta, entre outras informações.

2.7 Sistemas de Apoio à Decisão Clínica

Um dos principais objetivos da utilização de computadores na medicina é ajudar os médicos na tomada de decisões. O diagnóstico médico envolve não apenas determinar a explicação fisiopatológica dos sintomas de um paciente, mas também decidir quais perguntas fazer, quais exames solicitar, e determinar o valor dos resultados em relação aos riscos associados ou custos financeiros. Mesmo quando o diagnóstico é conhecido, há desafiantes decisões de gerenciamento que testam o conhecimento e experiência do médico, como decidir se tratar ou não o paciente e qual tratamento usar [19]. E para lidar com esses desafios, surgem os sistemas de suporte à decisão clínica

Um sistema de suporte à decisão clínica é um programa de computador que ajuda os profissionais de saúde a tomar decisões clínicas. Ele pode ser classificado em três tipos de funções: Ferramentas de Gerenciamento de Informações, Ferramentas para Focar a Atenção e Ferramentas para fornecer recomendações específicas para pacientes.

Com a integração do Processamento de Linguagem Natural (PLN) nos sistemas de apoio à decisão clínica é possível melhorar a capacidade de extrair, interpretar e analisar informações complexas de prontuários eletrônicos, permitindo a identificação precoce de padrões de risco que podem não ser imediatamente evidentes para os profissionais de saúde.

Quando combinado com algoritmos de inteligência artificial que aprendem continuamente com novos dados, o PLN pode oferecer recomendações baseadas em evidências e altamente personalizadas para a gestão de pacientes, otimizando assim as estratégias de prevenção e tratamento de doenças cardiovasculares. Por exemplo, a integração dessas tecnologias em sistemas de suporte à decisão pode auxiliar médicos na escolha de abordagens terapêuticas individualizadas, considerando não apenas as condições clínicas do paciente, mas também suas características sociais e genéticas. Assim, a adoção de PLN e inteligência artificial na prática médica promove uma medicina mais precisa, preventiva, rápida e personalizada.

2.7.1 Ferramentas de Gerenciamento de Informações

Ferramentas de gerenciamento de informações, como sistemas de informação em saúde e sistemas de recuperação de informação, fornecem dados e conhecimentos para os profissionais da saúde. Estações de trabalho especializadas de gerenciamento de conhecimento também estão sendo desenvolvidas para armazenar e recuperar informações clínicas. No entanto, essas ferramentas geralmente não ajudam na aplicação dessas informações para decisões específicas, sendo necessário a interpretação e escolha das informações pelo profissional da saúde [19].

2.7.2 Ferramentas para Focar a Atenção

São ferramentas que tem como o objetivo chamar a atenção do usuário para anormalidades. Esses programas são projetados para lembrar o usuário de diagnósticos ou problemas que poderiam ter sido ignorados. Normalmente, eles usam lógicas simples, exibindo listas ou parágrafos fixos como resposta padrão para uma anormalidade definida ou potencial [19].

2.7.3 Ferramentas para fornecer recomendações específicas para pacientes

Esses programas fornecem avaliações ou conselhos personalizados com base em conjuntos de dados específicos para pacientes. Eles podem seguir lógicas simples (como algoritmos), podem ser baseados na teoria da decisão e na análise custo-benefício, ou po-

dem usar abordagens numéricas apenas como complemento para a resolução de problemas simbólicos [19].

2.8 Reconhecimento de Entidades Nomeadas

Reconhecimento de Entidades Nomeadas ou Named Entity Recognition (NER) é a tarefa de identificar entidades nomeadas como nomes de pessoas, organizações, tempo, locais, entre outros, de um determinado conjunto de dados ou corpus. Entidades nomeadas também incluem outras entidades como entidades de domínio médico, entidades de alimentos e entidades nomeadas definidas pelo usuário no corpus [20].

O NER é muito utilizado na área médica para identificar termos clínicos como, medicamentos, parte do corpo, doenças e etc. Para português existem modelos pré-treinados disponíveis que identificam várias entidades clínicas textos médicos, como o BioBERTpt (Portuguese Clinical and Biomedical BERT) [32]. Porém, não existe um modelo em português que extraia dados sobre patologias específicas, a maioria é de maneira genérica.

2.8.1 Conjunto de dados para treino de reconhecimento de entidades nomeadas

O treinamento de um modelo de Reconhecimento de Entidades Nomeadas (NER) requer um conjunto de dados devidamente rotulado, contendo exemplos de texto que englobem as entidades de interesse, tais como nomes de pessoas, organizações, locais, entre outras. Cada instância de texto deve ser complementada por anotações que indiquem as posições exatas em que as entidades ocorrem, bem como suas categorias correspondentes.

Quadro 2.5. Exemplo de anotação de entidades nomeadas

Texo	Anotação
João	O
é	O
um	O
paciente	O
com	O
diabetes	B-DOENCA
tipo	I-DOENCA
2	I-DOENCA
.	O

No quadro 2.5 é possível ver um exemplo de rótulo identificando a doença, que no caso é como diabetes do tipo 2 (B-DOENCA e I-DOENCA). As outras palavras da frase, que não fazem parte da entidade nomeada, são anotadas como O (outro). Percebe-se que as anotações começam com os prefixo "B"e "I". O prefixo "B-"(de "Beginning") é

usado para marcar a primeira palavra de uma entidade nomeada, ou seja, a palavra que começa a entidade. Já o prefixo "I-"(de "Inside") é usado para marcar todas as palavras subsequentes que fazem parte da mesma entidade.

2.9 Modelos de Linguagem

Um modelo de linguagem é capaz de aprender e gerar texto baseados no dados que foram usados no seu treinamento. Os modelos de linguagem, diferente de outras técnicas de processamento de linguagem natural, aprendem o contexto do que foi passado, fazendo com que ele seja bastante utilizado em várias tarefas de NLP, como geração de texto, tradução e extração de informações.

Para falar de modelos de linguagem, primeiramente é preciso descrever os Transformers. Os Transformers são um modelo arquitetural que é baseado exclusivamente na auto atenção para computar representações de sua entrada e saída sem usar RNNs (Recurrent Neural Network) alinhados sequencialmente ou convolução [37].

Os Transformers ganharam bastante atenção por serem mais rápidos e mais eficientes do que outros tipos de arquiteturas de redes neurais, como por exemplo o LSTM (Long Short-Term Memory) [37].

A Figura 2.1 descreve a arquitetura dos Transformers como de camadas de auto atenção empilhadas e conectadas tanto ao codificador quanto para decodificador. A atenção é uma função que mapeia uma consulta e um conjunto de pares de valores-chave para uma saída, onde a consulta, as chaves, os valores e a saída são todos vetores. O codificador é responsável por receber e coletar as informações e passar adiante e o decodificador é o que finaliza o processo recebendo a informação transformando-a em uma saída.

Surgiu a partir dos Transformers, várias outras variações, como por exemplo o BERT e Roberta.

2.9.1 BERT

O BERT (Bidirectional Encoder Representations from Transformers) é basicamente uma pilha de transformers que foi projetado para pré-treinar representações bidirecionais profundas de texto não rotulado, condicionando conjuntamente os contextos esquerdo e direito em todas as camadas [12]. Os principais problemas que o BERT resolve são: Tradução de textos, sistemas de pergunta e resposta, análise de sentimento, resumos de textos e extração de informações, e como esses problemas necessitam um bom entendimento de linguagem e o BERT é justamente criado para dá um significado e um contexto

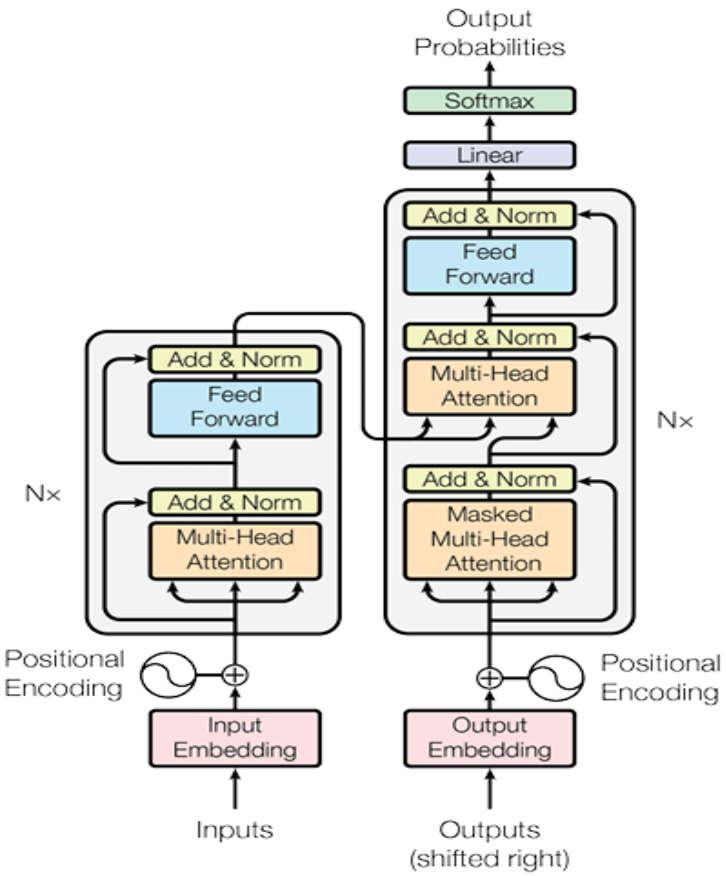

Figura 2.1. Os Transformers - Modelo Arquitetural
Fonte: [37]

para a linguagem. Os procedimentos do BERT são divididos em dois: Pré-treinamento e fine-tune.

O pré-treinamento é quando os parâmetros são utilizados para inicializar o modelo para o entendimento da linguagem e contexto dos textos que estão sendo inseridos e que vão servir para executar tarefas [12]. As tarefas são: Masked Language Model (MLM) e Next Sentence Prediction (NSP). No MLM algumas palavras são mascaradas e substituídas pelo token [MASK] e o modelo tenta prever quais palavras se encaixam melhor naquele token. Por exemplo na frase: "Tinha uma [MASK] no meio do caminho."o modelo irá tentar prever a palavra que se encaixa melhor no token [MASK] de acordo com o contexto, que no caso seria a palavra "pedra". O NSP é parecido com o que se tem no teclado de celulares, quando o usuário vai digitando palavras o sistema tenta "adivinhar"a próxima palavra que o usuário deseja digitar de acordo com o contexto. É dessa forma que o NSP funciona no BERT.

O fine-tune é quando você especializa seu modelo para determinado tipo de dado de entrada. Dependendo do tipo de tarefa que deseja fazer, será necessário ter dados rotulados, como por exemplo, reconhecer entidades nomeadas de textos biomédicos, é

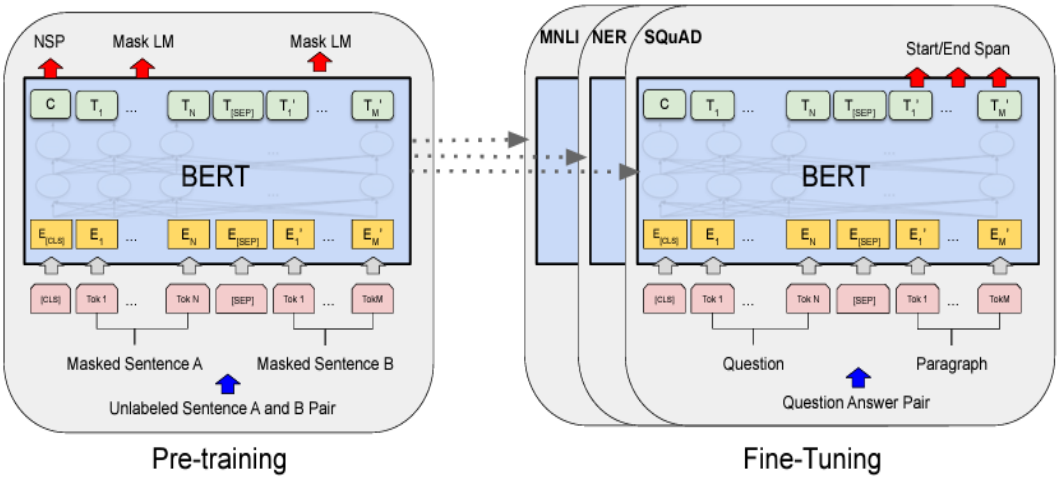

Figura 2.2. Procedimentos de pré-treinamento e fine-tune do BERT
Fonte: [12]

necessário passar para o modelo dados classificados de doenças ou remédios para que ele consiga entender que determinadas frases pertencem ao domínio da biomédica.

2.10 Grandes Modelos de Linguagem

Os *Large Language Models* (LLMs) são modelos de aprendizado profundo treinados com grandes volumes de dados textuais e compostos por bilhões de parâmetros. Esses modelos são baseados, majoritariamente, na arquitetura *Transformer* [37] e são projetados para prever a próxima palavra em uma sequência, o que lhes permite gerar texto coerente, realizar traduções, responder perguntas, resumir textos, entre outras tarefas de processamento de linguagem natural (PLN).

Durante o treinamento, os LLMs aprendem representações linguísticas ricas que capturam semântica, sintaxe, conhecimento factual e até habilidades de raciocínio. Isso os torna altamente versáteis e adaptáveis a diversas aplicações, mesmo com instruções mínimas (*zero-shot*) ou com exemplos (*few-shot*), como demonstrado com o modelo GPT-3 [6].

Além disso, LLMs podem ser ajustados com técnicas como *instruction tuning* [31] e *reinforcement learning from human feedback* [26], aprimorando sua capacidade de seguir instruções e produzir respostas mais alinhadas às expectativas humanas.

Contudo, eles também apresentam desafios, como alucinação de informações, viés algorítmico e alto custo computacional, o que tem motivado a pesquisa por abordagens mais eficientes e seguras para sua implantação.

2.10.1 DeepSeek-R1

O DeepSeek-R1 é um LLM focado em raciocínio avançado, desenvolvido usando aprendizado por reforço (Reinforcement Learning). Ele faz parte da primeira geração de modelos da iniciativa DeepSeek, voltada para criar LLMs mais inteligentes, com maior capacidade de resolver problemas complexos, fazer inferências lógicas e lidar com múltiplas etapas de pensamento [11].

Os modelos DeepSeek-R1-Zero, DeepSeek-R1 e seis versões densas destiladas — com 1.5B, 7B, 8B, 14B, 32B e 70B parâmetros — foram disponibilizados como código aberto, sendo baseados nos modelos *Qwen* [3] e *LLaMA* [34]. Esses modelos estão acessíveis para uso e pesquisa por parte da comunidade.

Considerando que o foco deste trabalho é comparar diferentes abordagens e modelos de código aberto que possam ser executados localmente com relativa facilidade, optou-se por utilizar o modelo DeepSeek-R1 com 14B de parâmetros.

2.11 Engenharia de Prompts

A engenharia de prompts em modelos de inteligência artificial generativa constitui uma área emergente e estratégica, que influencia diretamente a forma como interagimos com esses sistemas e os resultados que eles são capazes de gerar. Essa prática envolve a elaboração de prompts otimizados para alcançar objetivos específicos, exigindo não apenas familiaridade com o funcionamento interno dos modelos, mas também uma compreensão profunda do contexto de aplicação [1].

Mais do que simplesmente redigir comandos, a engenharia de prompts demanda o domínio de técnicas avançadas, como a utilização de templates dinâmicos baseados em dados e a adaptação do discurso conforme as respostas obtidas. Trata-se de um processo iterativo, que se aproxima do desenvolvimento de software ao incorporar práticas como controle de versão, testes sistemáticos e refinamento contínuo. Essa abordagem permite explorar de forma mais eficaz o potencial dos modelos generativos, promovendo interações mais precisas, eficientes e alinhadas aos objetivos do usuário [1].

2.11.1 O que é uma prompt?

Em modelos de inteligência artificial generativa, o prompt refere-se à entrada textual fornecida pelos usuários com o objetivo de orientar a resposta do modelo. De modo geral, os prompts são compostos por instruções, perguntas, dados de entrada e exemplos. Na prática, para obter uma resposta desejada de um modelo de IA, o prompt deve conter ao

menos uma instrução ou pergunta, sendo os demais elementos opcionais [1].

Prompts básicos em modelos de linguagem podem ser tão simples quanto uma pergunta direta ou uma instrução clara sobre uma tarefa específica. Já os prompts avançados apresentam estruturas mais complexas, como no caso do *chain of thought prompting*, em que o modelo é conduzido a seguir um processo lógico de raciocínio até chegar a uma resposta.

2.11.2 Principais Tipos de Prompt

Diversas estratégias de construção de *prompts* têm sido propostas para maximizar o desempenho de modelos de linguagem, seja na geração de textos mais coerentes, na resolução de tarefas complexas ou na adaptação a contextos específicos. Algumas dessas técnicas são aplicadas durante o *fine-tuning*, enquanto outras exploram a capacidade dos modelos pré-treinados sem a necessidade de ajustes adicionais. Nesta pesquisa, focamos exclusivamente em técnicas de *prompting* aplicáveis diretamente a modelos pré-treinados, sem qualquer readequação de parâmetros.

2.11.2.1 Prompt Simples (Zero-shot Prompting)

O *zero-shot prompting* consiste em fornecer apenas uma instrução ou pergunta ao modelo, sem exemplos adicionais ou contexto de apoio. Essa abordagem avalia a capacidade do modelo de generalizar a partir do conhecimento adquirido durante o pré-treinamento [6].

Exemplo: "Traduza a frase 'Hello, how are you?' para o português."

2.11.2.2 Prompt com Exemplos (Few-shot Prompting)

No *few-shot prompting*, incluem-se exemplos da tarefa desejada dentro do próprio *prompt*, o que auxilia o modelo a identificar padrões e a reproduzir o comportamento esperado com base nas instruções e exemplos fornecidos [6].

Entrada: "O céu está claro." → **Tradução:** "The sky is clear."
Entrada: "Está chovendo." → **Tradução:** [modelo completa]

2.11.2.3 Chain-of-Thought Prompting (Cadeia de Raciocínio)

O *chain-of-thought prompting* estimula o modelo a explicitar seu raciocínio passo a passo antes de apresentar a resposta final. Essa técnica é especialmente eficaz em tarefas

que exigem raciocínio lógico, aritmético ou causal, pois permite que o modelo construa uma sequência de inferências para alcançar uma solução mais confiável [38].

> **Pergunta:** Maria tem 12 balas. Ela dá 3 balas para João e depois compra mais 5 balas. Quantas balas Maria tem agora?
>
> **Resposta:**
> Vamos pensar passo a passo.
> Maria começa com 12 balas.
> Ela dá 3 balas para João, então agora tem $12 - 3 = 9$ balas.
> Depois, ela compra mais 5 balas, totalizando $9 + 5 = 14$ balas.
> **Resposta final:** Maria tem 14 balas.
>
> **Pergunta:** Pedro tem 5 balas. Ele dá 3 balas para Maria e depois compra mais 2 balas. Quantas balas Maria tem agora?
>
> **Resposta:** [modelo completa]

2.12 Argilla

Argilla é uma plataforma de anotação de dados de código aberto para modelos de linguagem. Usando a Argilla, é possível construir modelos de linguagem robustos por meio de uma anotação de dados mais rápida, utilizando tanto feedback tanto humano quanto de máquina [2].

3 Materiais e Métodos

3.1 Prontuários

O conjunto de dados foi disponibilizados pela empresa Neuralmed, e são 6037 prontuários eletrônicos de várias clínicas e hospitais de São Paulo. Inicialmente, juntamente com um especialista na área médica, realizou-se uma análise dos prontuários eletrônicos, onde foram identificados trechos e palavras relacionadas à cada fator de risco e como eles são escritos no texto. Depois a anotação de cada fator de risco é feita manualmente ou utilizando regex, para que se possa ter um dado para o modelo possa ser treinado. As categorias que estão sendo anotadas nos dados são:

3.1.1 Antecedentes e Comorbidades

Condições pré-existentes, como doenças crônicas, lesões ou problemas de saúde que o paciente tenha antes de ser diagnosticado com uma nova condição. E também hábitos que o paciente possa ter como tabagismo e etilismo. É importante registrar esses antecedentes para ajudar no tratamento futuro e na avaliação da saúde geral do paciente. Geralmente aparece nos prontuários como na Figura 3.1

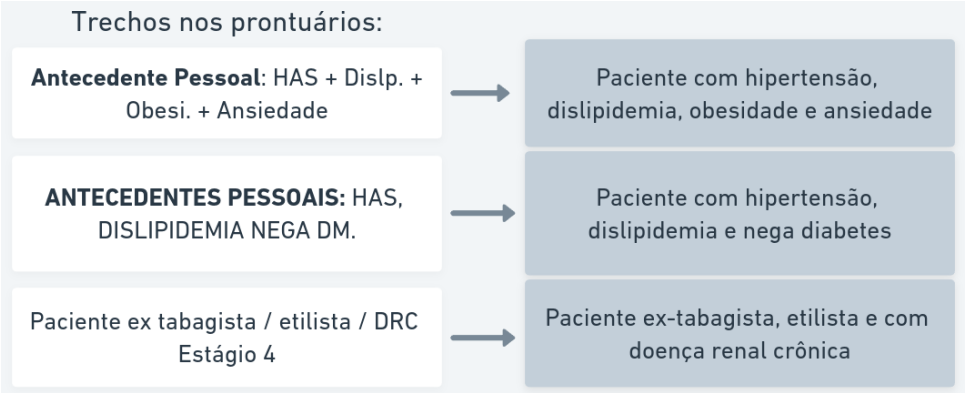

Figura 3.1. Exemplos de antecedentes e comorbidades

A partir desse trecho do prontuário, é possível identificar várias condições que o paciente possui e também quais delas são fatores de riscos.

3.1.2 Antecedentes Familiares

Histórico de doenças ou condições médicas presentes em familiares próximos do paciente, como pais, avós, irmãos ou tios. Esta informação é importantes porque algumas condições médicas podem ser genéticas ou hereditárias e aumentar o risco de desenvolvimento das mesmas.

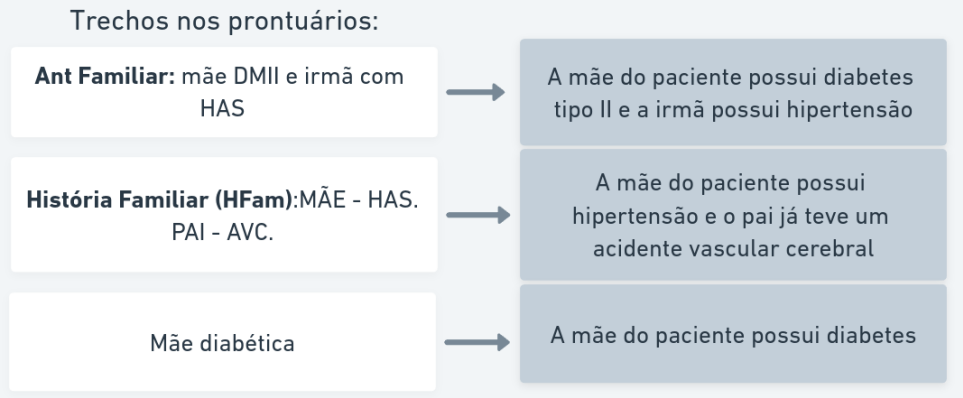

Figura 3.2. Exemplos de antecedentes familiares

3.1.3 Medicamentos

Medicamentos relacionados à diabetes, hipertensão, obesidade e dislipidemia. É importante identificá-los, pois eles podem indicar que o paciente está em tratamento para essas condições.

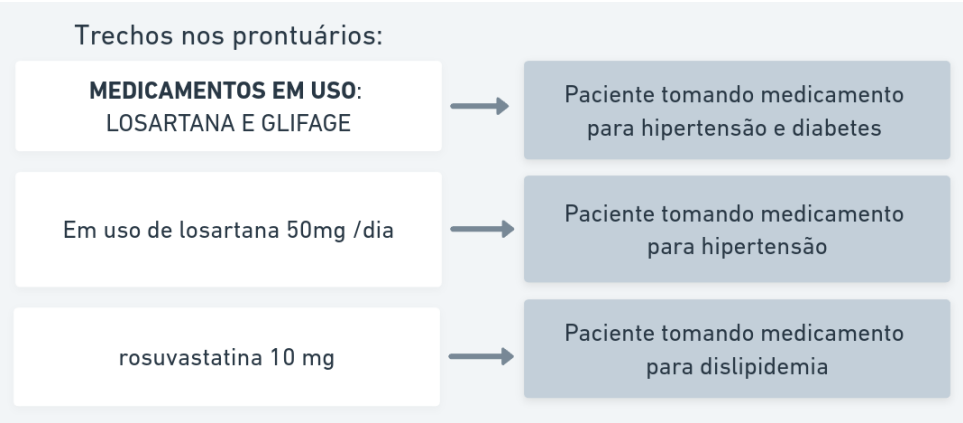

Figura 3.3. Exemplos de medicamentos

3.1.4 Resultados de exames

Os resultados de exames relacionados à diabetes e dislipidemia é uma outra forma de identificar se o paciente possui essas condições, já que os exames são uma das principais

formas de diagnóstico. Os exames são:

- **Hemoglobina Glicada (HbA1c)**: Exame de identificação de altos níveis de glicemia durante períodos prolongados.
- **Glicemia**: O exame de glicemia que descrever o nível de açúcar (glicose) no sangue.
- **HDL**: O exame que descreve o nível de lipoproteínas de alta densidade, ou "colesterol bom".
- **LDL**: O exame que descreve o nível de lipoproteínas de baixa densidade, ou "colesterol ruim".
- **Triglicerídeos**: O exame que descreve o nível de gordura no sangue.
- **Colesterol Total**: O exame que descreve o nível de níveis de colesterol e triglicérides na corrente sanguínea.

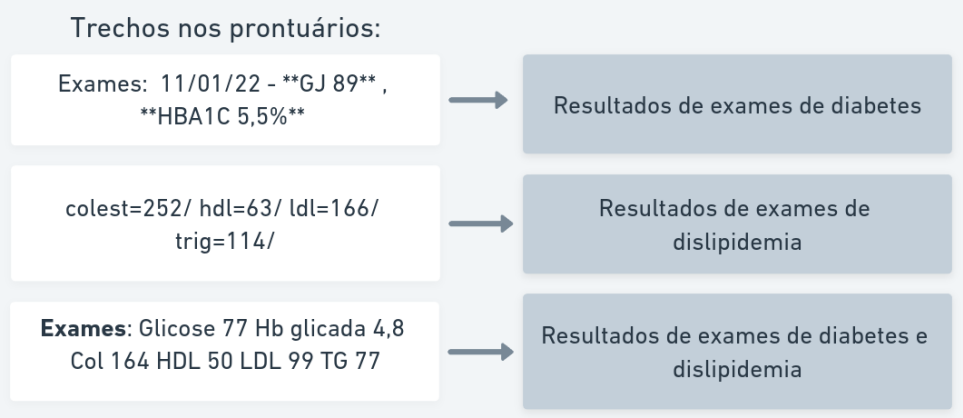

Figura 3.4. Exemplos de resultados de exames

3.1.5 Índice de Massa Corporal - IMC

O IMC é uma medida que serve pra medir os graus de obesidade. Então é um fator importante a ser encontrado. Normalmente ele aparece no texto como no exemplo da Figura 3.5

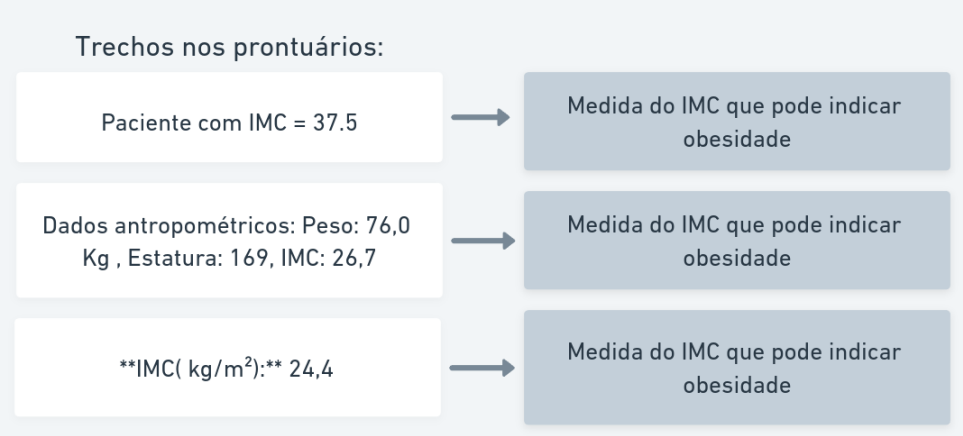

Figura 3.5. Exemplos de IMC

3.1.6 Pressão Arterial - PA

Na pressão arterial temos a medida da pressão que o sangue exerce contra a parede das artérias, e dependendo dos valores, pode indicar hipertensão arterial.

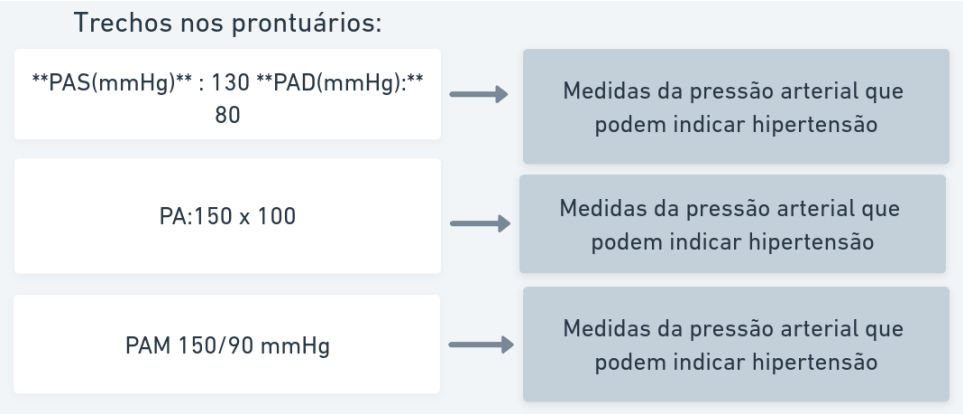

Figura 3.6. Exemplos de pressão arterial

3.1.7 Hipótese Diagnóstica - HD

São as hipóteses que podem descartar ou identificar um diagnóstico.

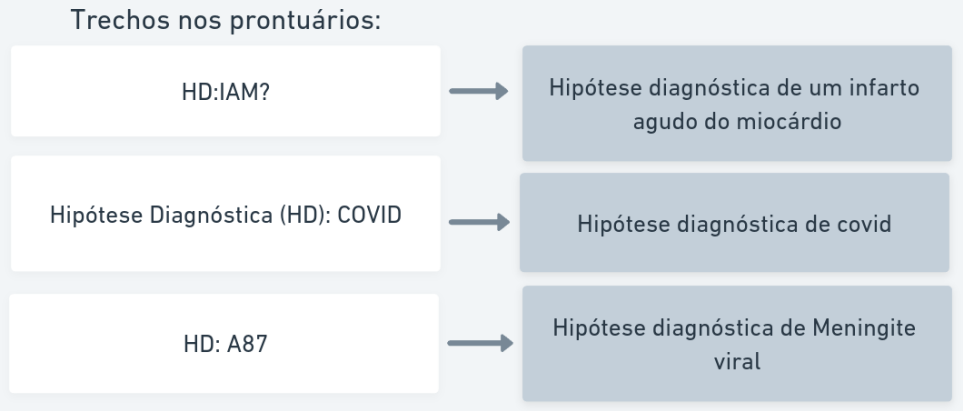

Figura 3.7. Exemplos de hipótese diagnóstica

3.2 Anotações

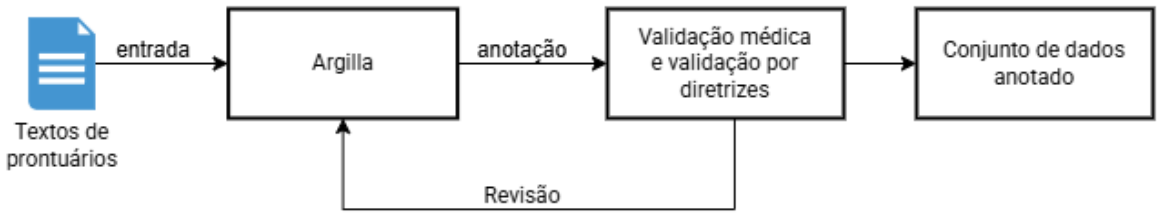

Figura 3.8. Fluxo de fine-tuning do modelo de negação

As anotações dos dados foram feitas na ferramenta Argilla [2], que é open-source e muito utilizada para curadoria de dados de modelo de linguagem. Então, baseada nas categorias analisadas anteriormente, é preciso criar as classes de acordo com o padrão de estrutura para treinamento de modelos NER.

Quadro 3.1. Classes de anotação utilizadas para treinamento do modelo

Classes de anotação	
B-AntecedenteComorbidade	I-AntecedenteComorbidade
B-AntecedenteFamiliar	I-AntecedenteFamiliar
B-ExamesDLP	I-ExamesDLP
B-ExamesDiabetes	I-ExamesDiabetes
B-HipoteseDiagnostica	I-HipoteseDiagnostica
B-IMC	I-IMC
B-MedicamentosDLP	I-MedicamentosDLP
B-MedicamentosDiabetes	I-MedicamentosDiabetes
B-MedicamentosHAS	I-MedicamentosHAS
B-MedicamentosObesidade	I-MedicamentosObesidade
B-PressaoArterial	I-PressaoArterial

Para facilitar a extração das informações, os medicamentos foram separados por condição. Os exames, foram separados por diabetes e dislipidemia. Com as classes definidas, as anotações devem seguir o exemplo do Quadro 3.2.

Quadro 3.2. Exemplo de anotação de exame e medicamento de diabetes

Texto	Anotação
EXAMES:	O
HB	B-ExamesDiabetes
GLICADA:	I-ExamesDiabetes
83	I-ExamesDiabetes
CREATININA:	O
081	O
CONDUTA:	O
GLICLAZIDA	B-DiabetesMedicamento

Com o uso da ferramente, esse tipo de anotação é facilitado por causa da interface, porque é só preciso selecionar a parte do texto específica e colocar a classe desejada. Como é ilustrado na Figura 3.9, onde cada cor representa uma das classes. E no final essa anotação pode ser exportada e é reestruturada como no exemplo do Quadro 3.2.

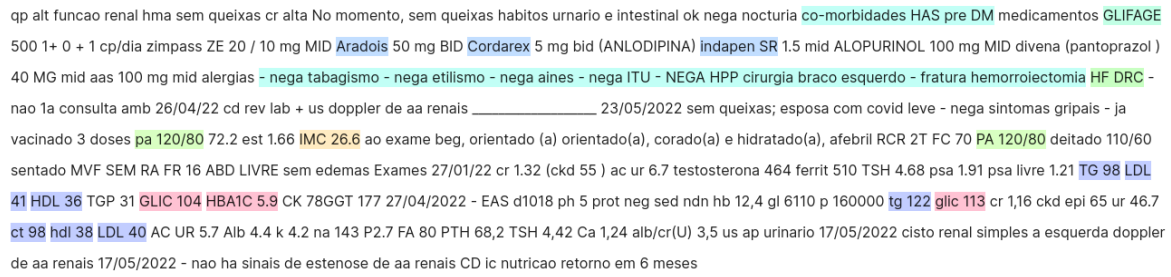

Figura 3.9. Exemplo de anotação no Argilla

O Quadro 3.3 mostra a quantidade de cada classe no conjunto de dados final.

A classe MedicamentosObesidade tem uma quantidade muito pequena, então foi optado por sua remoção pois classes com poucas amostras podem levar a um sobreajuste para estas amostras específicas, reduzindo a capacidade do modelo de generalizar a partir de novos dados. A remoção de classes pouco representativas pode ajudar a balancear o conjunto de dados, melhorando assim a precisão geral do modelo.

Quadro 3.3. Valores das variáveis

Classes	Quantidade
AntecedenteComorbidade	6.031
HipoteseDiagnostica	3.535
PressaoArterial	1.761
MedicamentosHAS	1.171
MedicamentosDiabetes	577
AntecedenteFamiliar	553
ExamesDiabetes	519
MedicamentosDLP	466
ExamesDLP	386
IMC	207
MedicamentosObesidade	12

3.3 Modelo especializados em prontuários

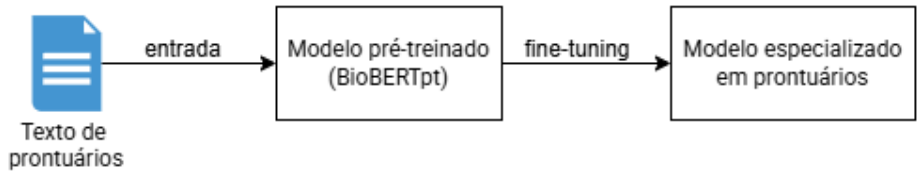

Figura 3.10. Fluxo de fine-tuning do modelo de linguagem especializado em prontuários

Para o desenvolvimento do modelo especializado, foi feito uma *transferência de conheccimento* do modelo ***BioBERTpt - Portuguese Clinical and Biomedical BERT*** [32]. *Tranferência de conhecimento* é uma estratégia essencial neste contexto, porque é possível pegar o conhecimento prévio do *BioBERTpt*, que foi originalmente treinado em uma ampla gama de dados clínicos e biomédicos em língua portuguesa, e adicionar o novo conhecimento específico da tarefa em questão, criando um novo modelo linguagem especializada em prontuários médicos e outros dados médicos.

A etapa de treinar um novo modelo de linguagem especializado pode ser feito diversas vezes e com diferentes modelos de base, como mostra a Figura 3.10. Além disso, é possível mudar a arquitetura do modelo e fazer diversos testes para qual estratégia se tem uma melhor performance.

3.4 Modelo NER

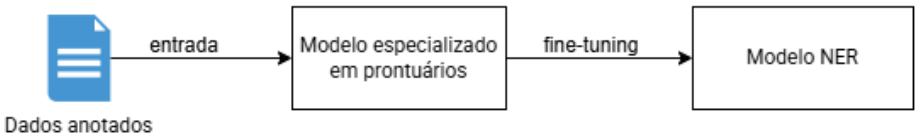

Figura 3.11. Fluxo de fine-tuning do modelo NER

Para o desenvolvimento do modelo foi usado o fine-tuning do modelo especializado. Fine-tuning em modelos de linguagem é o processo de usar um modelo pré-treinado em grandes quantidades de dados e adaptá-lo a um conjunto de dados de tarefa específico. O objetivo é aproveitar a capacidade do modelo de aprender padrões gerais de linguagem durante o treinamento com grandes quantidades de dados, ao mesmo tempo em que se permite que o modelo se adapte aos dados de tarefa específicas.

A tarefa utilizada foi a de classificação de tokens, com o objetivo de criar modelos NER. O treinamento de NER com o Transformer envolve a seguinte sequência de etapas:

- Tokenização: Converter o texto em uma sequência de tokens (palavras ou subpalavras).
- Conversão de tokens em vetores: Codificar cada token como um vetor numérico para que possa ser utilizado como entrada na rede neural.
- Treinamento do transformer: Treinar a Transformer com os dados de treinamento etiquetados e obter os pesos das camadas da rede.
- Validação: Avaliar a precisão da rede neural com um conjunto de dados de validação anotados.
- Previsão: Utilizar o modelo treinado para prever as entidades nomeadas em textos desconhecidos.

3.5 Normalizacão dos dados

A normalizacão dos dados de predição do modelo é essencial, porque como os prontuários são escritos de maneira livre, em linguagem natural, e não padronizadas, é preciso definir e mapear o que cada coisa significa. Por exemplo, a condição "diabetes"pode ser escrita de diversas formas, usando siglas, ou nomes mais formais. Existem vários sistemas de vocabulários, como por exemplo, o UMLS (Unified Medical Language System), porém a maioria é em inglês.

Nesse estudo, o vocabulário foi feito manualmente por meio de análise dos textos do prontuários e criando assim o próprio banco de vocabulário para prontuários. Essa normalização é presentada na Figura 3.12, onde se tem uma predição de antecedentes do paciente e após a normalização temos os nomes das condições padronizadas.

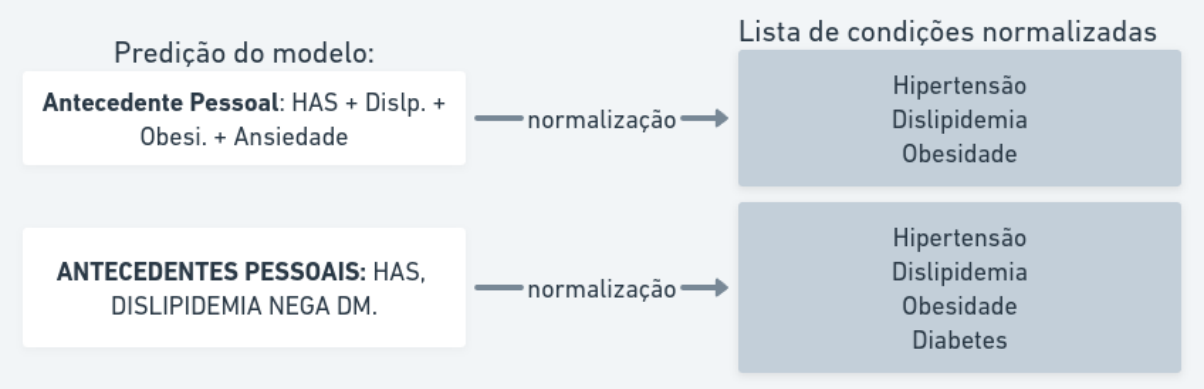

Figura 3.12. Exemplo de normalização de antecedentes

Além dos trechos de antecedentes que precisam ser normalizados, é preciso identificar as condições por meio dos exames. De acordo com as formas de diagnósticos descritos nos Quadros 2.4 e 2.2, um algoritmo foi feito, usando o intervalo dos exames e identificando se o paciente tem ou não diabetes ou dislipidemia.

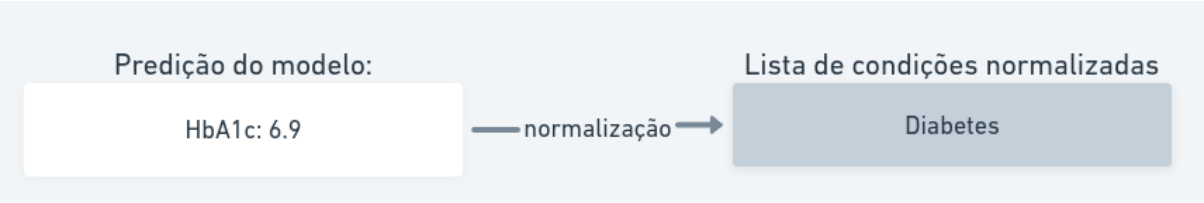

Figura 3.13. Exemplo de normalização de exames

A Figura 3.13 mostra a predição de um exame de hemoglobina glicada com o valor de 6.9, que de acordo com as diretrizes, o paciente possivelmente tem diabetes. Lembrando que o intuito do estudo é sempre o auxílio na procura de pacientes com fatores de risco e não dar um diagnóstico, por isso é preciso sempre validar as informações que são dadas. Os medicamentos também podem passar por um fluxo de normalização parecido, como eles já são especificados com as classes específicas de cada condição, esse trabalho é mais facilitado.

3.6 Modelo de Negação

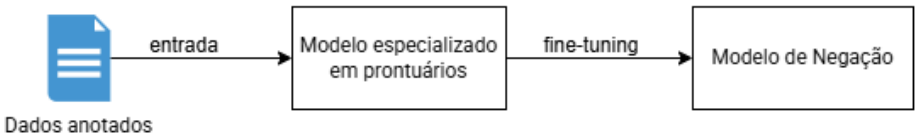

Figura 3.14. Fluxo de fine-tuning do modelo de negação

É muito comum que os prontuários médicos contenham a negação de determinadas condições, especialmente quando o médico ou o profissional da saúde descreve os antecedentes do paciente. Por isso, é importante ter uma forma eficaz para identificar com precisão o que está sendo negado, minimizando assim a ocorrência de falsos positivos durante a extração de informações. Uma investigação aprofundada foi feita em busca de modelos pré-existentes treinados especificamente para tarefas de classificação de negação na área médica.Essa análise revelou uma lacuna significativa: o único modelo disponível para tarefas relacionadas à negação clínica era era o Clinical Assertion / Negation Classification BERT [36], desenvolvido exclusivamente para textos em língua inglesa. Esse cenário evidenciou a necessidade urgente de um modelo personalizado e adaptado à realidade da língua portuguesa, especialmente para o processamento de documentos médicos.

Para treinar um modelo parecido com o que foi encontrado em inglês para português, é preciso um conjunto de dados com as classificações de cada frase, então foi preciso anotar manualmente os dados, juntamente com o auxílio de regex e de outros LLMs. Um segundo conjunto de dados foi criado com o propósito de classificar condições médicas nas categorias POSSIBLE, ABSENT ou PRESENT.

Quadro 3.4. Exemplos de condições médicas classificadas

Texto	Classe
NEGA [entity] HAS [entity] E OU DIABETES	ABSENT
NEGA HAS E OU [entity] DIABETES [entity]	ABSENT
Portadora de [entity] HAS [entity]	PRESENT
Paciente em investigação para [entity] diabetes [entity]	POSSIBLE

Inicialmente, as anotações foram realizadas para indicar claramente quais condições estavam PRESENT, ABSENT ou POSSIBLE dentro das sentenças. Em seguida, essas condições foram envolvidas pela tag [entity], conforme ilustrado no Quadro 3.4. Essa marcação contribui para uma melhor contextualização por parte do modelo, auxiliando na identificação precisa de condições mencionadas em contextos de negação. O conjunto final anotado é composto por 11424 amostras, sendo 7039 de ABSENT, 3931 PRESENT e 454 POSSIBLE como demonstrado na Figura 3.15.

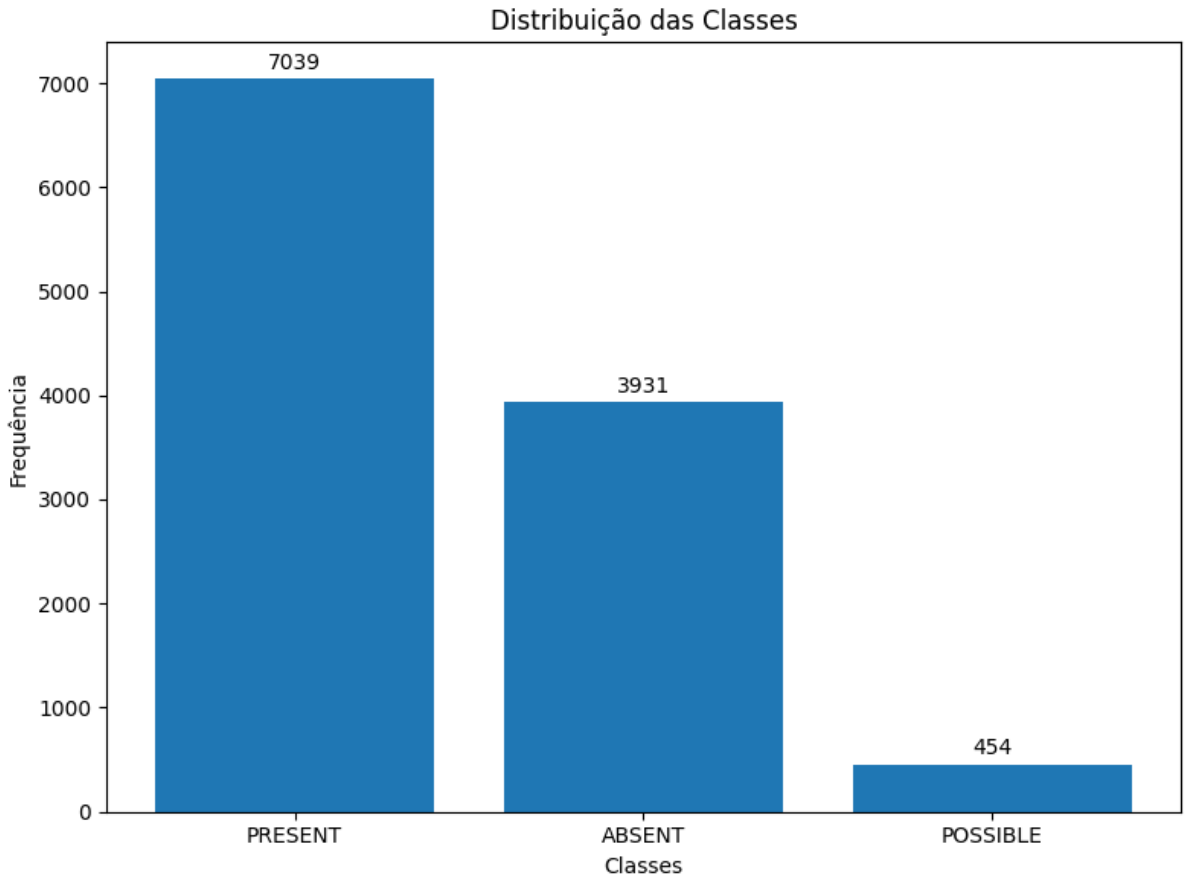

Figura 3.15. Distribuição das classes do conjunto de dados de negação

É possível notar que temos um desbalanceamento entre as classes, a classe POSSIBLE tem apenas 454, isso pode prejudicar a performance do modelo. Por isso, foi dado pesos para cada classe e utilizados no treinamento do modelo.

3.7 Processo de extração

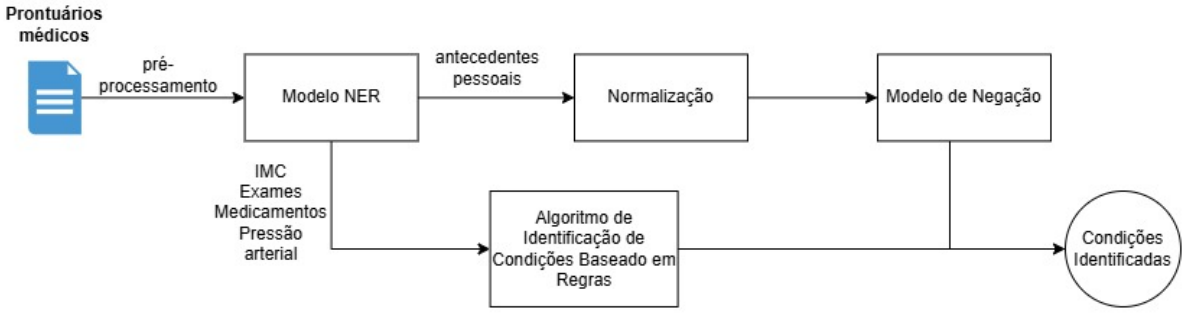

Figura 3.16. Diagrama do processo de extração das informações

Para extrair os fatores de risco para doenças cardiovasculares, os modelos treinados são utilizados em uma pipeline, conforme demonstrado na Figura 3.16. Nesta pipeline, os dados passam por um pré-processamento e são inseridos em um modelo de Reconhecimento de Entidades Nomeadas (Named Entity Recognition — NER), que extrai informações essenciais, como antecedentes pessoais, exames, IMC, pressão arterial e medicamentos. Em seguida, os antecedentes pessoais passam por uma normalização de vocabulário e são processadas por meio de um modelo de negação para confirmar a presença das condições. Simultaneamente, um algoritmo baseado em regras para identificação de condições analisa os demais dados extraídos. Ambos os processos convergem para identificar as condições clínicas finais.

Foi necessário criar um vocabulário com todas as variações dos fatores de risco identificados. Em português, é raro encontrar algo semelhante ao SNOMED-CT [14] e ao UMLS [5], que oferecem mapeamentos para esses termos em inglês. Para este estudo, o mapeamento foi realizado manualmente, por meio da análise dos termos referentes a cada risco identificado nos prontuários médicos.

O Algoritmo de Identificação de Condições Baseado em Regras é responsável por definir condições médicas com base em regras específicas, principalmente para fins de identificação de doenças. Por exemplo, se um exame de HbA1c apresentar nível de 6,7%, isso indica diabetes, uma vez que níveis de HbA1c acima de 6,5% são diagnósticos para essa condição [23]. Da mesma forma, se uma leitura de pressão arterial exceder consistentemente 140/90 mmHg, isso indica hipertensão [4]. Quando um medicamento associado a um fator de risco, como insulina para diabetes ou anti-hipertensivos para pressão alta, é identificado, o algoritmo marca o paciente como portador da condição. Além disso, o algoritmo pode identificar condições como hiperlipidemia se um painel lipídico revelar níveis elevados de colesterol, ou obesidade se o IMC de um paciente for superior a 30 [28]. Essa abordagem garante uma identificação abrangente e precisa de diversas condições de saúde com base em critérios médicos estabelecidos.

3.8 DeepSeek 14b + Few-Shot Prompting

Para a tarefa de extração de informações clínicas a partir de prontuários médicos em linguagem natural, utilizamos o LLM *DeepSeek 14b*, executado localmente via a plataforma *Ollama* [22]. O Ollama é uma interface de execução e gerenciamento local de modelos de linguagem de grande porte, que permite acesso rápido e seguro por meio de API HTTP, sem a necessidade de envio de dados sensíveis para servidores externos.

O modelo é guiado por um *prompt* do tipo *few-shot prompting*, conforme descrito por Brown et al. [6], combinando instruções explícitas com exemplos anotados. O objetivo do

prompt é identificar automaticamente menções a cinco condições clínicas: hipertensão, diabetes, dislipidemia, obesidade e tabagismo.

A saída esperada do modelo é um objeto JSON contendo os seguintes campos estruturados:

- `antecedentes_clinicos`: lista de condições clínicas mencionadas;
- `exames_laboratoriais`: lista de exames relevantes identificados;
- `historico_familiar`: condições mencionadas como familiares;
- `pressao_arterial`: valores de PA, quando disponíveis;
- Um campo para cada condição clínica (ex.: `hipertensao`, `diabetes`, etc.), com os atributos:
  - `present`: valor booleano indicando presença da condição;
  - `texto`: trechos do prontuário que justificam a identificação.

Após a resposta da API, o JSON retornado é extraído e validado quanto à presença de todos os campos esperados. Caso algum campo esteja ausente, uma nova tentativa é realizada automaticamente. Se a resposta continuar incompleta, os campos faltantes são preenchidos com valores padrão (listas vazias ou objetos com `present = false` e texto vazio), garantindo robustez na análise automatizada.

Essa abordagem permite utilizar modelos de linguagem pré-treinados sem necessidade de *fine-tuning*, tornando a técnica eficiente e reprodutível para a tarefa de extração de informações clínicas.

4 Resultados e Discussões

4.1 Métricas dos modelos

4.1.1 Modelo de linguagem especializado em prontuários

O modelo especializado, treinado com o texto dos prontuários e utilizado para o fine-tuning do modelo NER, atingiu um *loss* de treinamento de 4.0662 ao final do processo. É perceptível que o *loss*, tanto de treinamento quanto de validação, diminui à medida que o modelo converge. Isso sugere que o modelo está aprimorando sua capacidade de generalização e de aprendizado conforme o contexto em que está sendo treinado.

Quadro 4.1. Resultados de treinamento e validação ao longo das épocas.

Epoch	Training Loss	Validation Loss
1	5.0772	4.9537
2	5.0240	4.8402
3	4.8405	4.7441
4	4.7057	4.6670
5	4.5797	4.6074
6	4.5069	4.5579
7	4.4158	4.5190
8	4.3402	4.4838
9	4.2821	4.4521
10	4.2321	4.4353
11	4.1994	4.4140
12	4.1367	4.4013
13	4.1229	4.3925
14	4.0957	4.3843
15	4.0662	4.3834

4.1.2 Modelo de Reconhecimento de Entidades Nomeadas

As métricas do modelo estão descritas no Quadro 4.3 e 4.2. A maioria as classes conseguiram um F1 Score acima de 80%. O F1 Score é uma métrica de avaliação de modelos de classificação que combina precisão e sensibilidade em uma única medida. Essa

combinação é particularmente útil quando se deseja um equilíbrio entre a capacidade de identificar corretamente as instâncias positivas (verdadeiros positivos) e a importância de minimizar os falsos positivos. Um F1 Score de 80% indica um equilíbrio entre precisão e sensibilidade, significando que o modelo tem a capacidade de identificar corretamente os verdadeiros positivos (alta precisão), evitando falsos positivos e capturar uma grande proporção dessas amostras que de fato são positivas (alto recall).

Quadro 4.2. Resultados das métricas de desempenho.

Métrica	**Valor**
Precisão geral	0.8735
Recall geral	0.8869
F1 geral	0.8802
Acurácia geral	0.9845

Quadro 4.3. Métricas de desempenho classe

Classes	Precisão	Sensibilidade	F1 Score	Número
AntecedenteComorbidade	0.8919	0.8950	0.8934	4553
AntecedenteFamiliar	0.9085	0.8862	0.8972	325
ExamesDLP	0.8095	0.8395	0.8242	243
ExamesDiabetes	0.6547	0.6594	0.6570	138
HipoteseDiagnostica	0.8769	0.8935	0.8851	3540
IMC	0.7727	0.9623	0.8571	53
MedicamentosDLP	0.9143	0.9014	0.9078	71
MedicamentosDiabetes	0.8511	0.8247	0.8377	97
MedicamentosHAS	0.8390	0.9069	0.8716	247
PressaoArterial	0.7761	0.8387	0.8062	434

Ao interpretar os dados do Quadro 4.3, é possível observar que a precisão, recall e F1 Score variam entre as diferentes classes. Por exemplo, classes como IMC e ExamesDiabetes apresentam F1 Score relativamente baixos, sugerindo um desempenho menos eficaz do modelo nessas categorias específicas. Por outro lado, classes como AntecedenteComorbidade e MedicamentosDLP exibem pontuações mais elevadas nessas métricas, indicando uma melhor capacidade do modelo em classificar corretamente essas informações.

Essas métricas trazem informações valiosas, por meio delas é possível identificar quais campos específicos estão propensos a uma performance inferior por parte do modelo. Isso é fundamental para direcionar esforços de melhoria e refinamento do modelo e do conjunto de dados, concentrando recursos onde são mais necessários.

4.1.3 Modelo de Negação

O modelo demonstrou uma acurácia geral de 94%, com uma pontuação F1 macro-média de 93% e uma pontuação F1 média ponderada de 94%. Os resultados específicos

para cada categoria—ABSENT, PRESENT e POSSIBLE—são apresentados em termos de *Precision*, *Recall* e *F1 Score* no Quadro 4.4.

Quadro 4.4. Resultados experimentais do modelo de asserção clínica

Classes	Precisão	Sensibilidade	F1 Score	Número
ABSENT	0.94	0.89	0.91	904
PRESENT	0.93	0.97	0.95	1618
POSSIBLE	0.99	0.89	0.94	114

Apesar de o modelo ter alcançado uma acurácia geral de 94%, é importante observar que ele foi treinado em um conjunto de dados homogêneo. Embora contextualize com sucesso as informações dentro desses parâmetros, podem existir limitações ao processar textos que se desviam significativamente dos dados de treinamento.

Durante os testes, foi observado que o modelo tende a apresentar desempenho inferior ao lidar com seções de comorbidades apresentadas de forma estruturada, como listas de verificação ou formulários. Além disso, o desempenho do modelo é significativamente comprometido quando se depara com erros semânticos ou erros tipográficos no texto. Tais imprecisões podem levar a classificações incorretas, refletindo a sensibilidade do modelo à qualidade dos dados de entrada. Por exemplo, textos como "HAS NEGA CA DE MAMA" contêm abreviações ou informações fragmentadas que podem ser desafiadoras até mesmo para interpretação humana. Esse tipo de erro ou abreviação informal geralmente carece de contexto suficiente, dificultando o processamento e a classificação precisa por parte do modelo.

4.1.4 Comparação entre modelo NER + pipeline e DeepSeek com few-shot prompting

Os Quadros 4.5 e 4.6 apresentam um comparativo de desempenho entre duas abordagens diferentes para extração de riscos cardiovasculares de prontuários médicos: uma pipeline baseada em Named Entity Recognition (NER) com um modelo de classificação de negação, conforme ilustrado na Figura 3.16, e o modelo DeepSeek 14b (14 bilhões de parâmetros) utilizado em uma configuração few-shot, na qual exemplos de anotações foram incluídos diretamente no *prompt*.

A avaliação foi conduzida utilizando um conjunto composto por 760 amostras reais e aleatórias de prontuários eletrônicos, contendo registros clínicos variados. Cada amostra possui uma anotação de referência manual e os resultados das abordagens, representando respectivamente as saídas do pipeline tradicional e do modelo DeepSeek. O conteúdo dos textos inclui anotações de enfermagem, anamneses e evoluções clínicas, abrangendo diferentes estilos e níveis de estruturação textual.

Quadro 4.5. Desempenho do modelo NER + negação

Classes	Precisão	Sensibilidade	F1 Score	Número
Hipertensão	0.92	0.83	0.87	270
Diabetes	0.74	0.76	0.75	168
Obesidade	0.98	0.80	0.88	90
Dislipidemia	0.95	0.87	0.91	137
Tabagismo	0.57	0.46	0.51	95
Micro média	0.85	0.77	0.81	760
Macro média	0.83	0.74	0.78	760
Média ponderada	0.85	0.77	0.81	760
Média amostral	0.33	0.31	0.31	760

Quadro 4.6. Desempenho do modelo DeepSeek com prompting few-shot

Classes	Precisão	Sensibilidade	F1 Score	Número
Hipertensão	0.84	0.97	0.90	270
Diabetes	0.78	0.95	0.86	168
Obesidade	0.79	0.92	0.85	90
Dislipidemia	0.76	0.95	0.84	137
Tabagismo	0.68	0.94	0.79	95
Micro média	0.78	0.96	0.86	760
Macro média	0.77	0.95	0.85	760
Média ponderada	0.79	0.96	0.86	760
Média amostral	0.37	0.40	0.38	760

Ao observar os resultados, nota-se uma diferença significativa entre as abordagens, especialmente em termos de *sensibilidade*. O modelo DeepSeek apresenta valores elevados em todas as classes (variando entre 0.92 e 0.97), enquanto o pipeline NER + negação demonstra desempenho mais modesto, com *sensibilidade* variando de 0.46 (Tabagismo) a 0.87 (Dislipidemia). Essa diferença indica que o DeepSeek é mais eficaz em capturar menções relevantes às condições clínicas, sendo especialmente robusto em casos como "Hipertensão"e "Diabetes", que são frequentemente referenciados de maneira ambígua ou indireta nos textos médicos.

No que diz respeito à precisão, o modelo baseado em NER obtém valores mais altos em algumas classes, como "Obesidade"(0.98) e "Dislipidemia"(0.95), sugerindo maior especificidade na identificação correta dos termos. Por outro lado, o DeepSeek mantém uma precisão consistente, acima de 0.75 em todas as categorias.

O F1-score, que equilibra precisão e *sensibilidade*, mostra um desempenho geral superior do DeepSeek na maioria das classes. A média ponderada do F1-score do DeepSeek (0.86) supera a do modelo NER + negação (0.81), indicando melhor desempenho global. Vale destacar a performance em "Tabagismo", onde o F1-score sobe de 0.51 no modelo tradicional para 0.79 com o DeepSeek, demonstrando a vantagem das LLMs em contextos

com linguagem ambígua ou referências implícitas.

A média amostral (*samples avg*) também reflete essa diferença de capacidade de generalização, com o DeepSeek alcançando 0.38 no F1-score versus 0.31 do modelo tradicional.

Em resumo, os resultados apontam que, apesar do modelo NER + pipeline apresentar boa precisão em alguns casos, o DeepSeek, mesmo utilizando apenas poucos exemplos no prompt (few-shot), demonstrou ser substancialmente mais eficaz na cobertura e identificação das entidades clínicas, especialmente em situações com linguagem menos estruturada. Isso evidencia o potencial das LLMs como alternativa mais flexível e escalável para tarefas de extração de informação em domínios médicos complexos.

5 Conclusão

Este estudo demonstrou o potencial e os desafios do Processamento de Linguagem Natural (PLN) na análise de prontuários eletrônicos para a extração de fatores de risco cardiovascular. Por meio de uma abordagem multidisciplinar, foram desenvolvidos e avaliados modelos capazes de identificar condições clínicas relevantes, como hipertensão, diabetes, obesidade, dislipidemia e tabagismo, em textos médicos não estruturados.

Duas abordagens distintas foram comparadas: um pipeline tradicional baseado em reconhecimento de entidades nomeadas (NER) com um classificador de negação, e uma abordagem baseada em modelos de linguagem de grande porte, utilizando o DeepSeek 14b com prompting do tipo few-shot. A análise quantitativa revelou que o modelo DeepSeek apresentou desempenho superior em termos de *sensibilidade* e F1-score em praticamente todas as classes, indicando maior sensibilidade e capacidade de captar menções clínicas, inclusive quando expressas de forma implícita ou ambígua. Destaque especial foi observado na classe "Tabagismo", cujo F1-score subiu de 0.51 (pipeline NER) para 0.79 com o DeepSeek.

Por outro lado, o pipeline NER + negação demonstrou maior precisão em algumas categorias, como "Obesidade"(0.98) e "Dislipidemia"(0.95), sugerindo uma menor taxa de falsos positivos. Além disso, essa abordagem apresenta menor custo computacional e maior velocidade de inferência, sendo mais adequada para ambientes com recursos limitados ou que exijam respostas em tempo real. No entanto, sua limitação em lidar com variações linguísticas e construções mais complexas compromete sua capacidade de generalização.

Já o modelo DeepSeek, embora mais robusto e eficaz na cobertura dos fatores de risco, requer maior poder computacional, tempo de processamento mais elevado e cuidado especial na engenharia de prompts. Ainda assim, sua flexibilidade e alta performance mesmo sem *fine-tuning* o tornam uma solução promissora para aplicações clínicas que demandem maior profundidade na análise textual.

Em resumo, os resultados indicam que ambas as abordagens têm valor complementar. O pipeline NER pode ser útil em aplicações mais restritas e de baixo custo, enquanto os modelos baseados em LLMs, como o DeepSeek, oferecem maior abrangência e capacidade

de interpretação semântica, sendo ideais para sistemas de apoio à decisão clínica mais sofisticados.

Este trabalho destaca, portanto, o papel crescente da inteligência artificial na medicina preventiva, ao permitir a automação da identificação de riscos cardiovasculares em larga escala. Para estudos futuros, recomenda-se a expansão dos conjuntos de dados, bem como a realização de um fine-tuning baseado em *instruction tuning* no modelo DeepSeek, com o objetivo de adaptá-lo de forma mais precisa à tarefa de extração de informações clínicas relacionadas a fatores de risco cardiovascular. Além disso, a integração de validação clínica com profissionais da saúde pode contribuir significativamente para o desenvolvimento de soluções mais confiáveis e aplicáveis na prática médica real. Conclui-se que as abordagens avaliadas apresentam caráter complementar: enquanto o *pipeline* tradicional se destaca pela leveza e velocidade de execução, os LLMs demonstram maior robustez na interpretação semântica de textos clínicos. Como desdobramento, propõe-se também o desenvolvimento de cálculos automatizados para escores de risco cardiovascular, como o Escore de Framingham.

Agradecimentos

Agradeço à Neuralmed pelo fornecimento dos dados de prontuários médicos e ao Dr. Antônio Cesar pelo apoio nas validações dos dados nas contribuições clínicas.

Lista de Referências

[1] Xavier Amatriain. Prompt design and engineering: Introduction and advanced methods, 2024.

[2] Argilla. Argilla, 2024. Available at: https://argilla.io/.

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Han, et al. Qwen technical report, 2023.

[4] Weimar Kunz Sebba Barroso, Cibele Isaac Saad Rodrigues, Luiz Aparecido Bortolotto, Marco Antônio Mota-Gomes, Andréa Araujo Brandão, Audes Diógenes de Magalhães Feitosa, et al. Diretrizes Brasileiras de Hipertensão Arterial. https://abccardiol.org/article/diretrizes-brasileiras-de-hipertensao-arterial-2020, 2020. Accessed: 2022-12-31.

[5] O. Bodenreider. Unified medical language system (UMLS), 2021. Available at: https://www.nlm.nih.gov/research/umls/.

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners, 2020.

[7] Ricardo Bezerra Cavalcante, Marina Nagata Ferreira, e Poliana Cavalcante Silva. Sistemas de informação em saúde: possibilidades e desafios. *Revista de Enfermagem da UFSM*, 1:290, 2011.

[8] Ministério da Saúde. *Caderno de Atenção Básica*. Brasília - DF, 16ª edição, 2006.

[9] Ministério da Saúde. MANUAL DE USO DO SISTEMA COM PRONTUÁRIO ELETRÔNICO DO CIDADÃO – PEC, 2018.

[10] Ministério da Saúde. Plano de ações estratégicas para o enfrentamento das doenças crônicas e agravos não transmissíveis no Brasil. 2021.

[11] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, e Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2019.

[13] Izar M. C. O. Saraiva J. F. K. Chacra A. P. M. Bianco H. T. Afiune A. Faludi, A. A. et al. Atualização da diretriz brasileira de dislipidemias e prevenção da aterosclerose – 2017. *Arquivos brasileiros de cardiologia*, 2017.

[14] SNOMED International. SNOMED CT: Systematized nomenclature of medicine – clinical terms, 2021. Available at: https://www.snomed.org/.

[15] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li Wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, e Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3, 5 2016.

[16] Osvaldo Kohlmann Jr., Armênio Costa Guimarães, Maria Helena C. Carvalho, Hilton de Castro Chaves Jr., Carlos Alberto Machado, José Nery Praxedes, José Luiz Santello, et al. III Consenso Brasileiro de Hipertensão Arterial. *Arquivos Brasileiros de Endocrinologia Metabologia*, 43:247–249, 1999.

[17] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, e Jaewoo Kang. Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234–1240, 2020.

[18] CONSELHO FEDERAL DE MEDICINA. Código de ética médica. Resolução nº 1.638/2002, 2002.

[19] Mark A. Musen, Yuval Shahar, Edward H. Shortliffe, e James J. Cimino (Eds.). *Biomedical Informatics - Computer Applications in Health Care and Biomedicine*. 3 edição, 2006. p. 698-736.

[20] Salman Naseer, Muhammad Ghafoor, Sohaib, Sohaib Khalid Alvi, Anam Kiran, Shafique Ur Rehman, Ghulam Murtaza, Jehlum Campus, e Pakistan Jehlum. Named entity recognition (NER) in NLP techniques, tools accuracy and performance. page 2, 01 2022.

[21] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, e Ajmal Mian. A comprehensive overview of large language models. 2023.

[22] Ollama. Ollama: Run LLMs locally. https://ollama.com, 2024. Acesso em: mar. 2025.

[23] The Expert Committee on the Diagnosis e Classification of Diabetes Mellitus. Report of the expert committee on the diagnosis and classification of diabetes mellitus. *Diabetes Care*, 20(7):1183–1197, 1997.

[24] World Health Organization. Cardiovascular diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Accessed: 2025-05-04.

[25] World Health Organization. WHO Obesity and overweight. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight. Accessed: 2022-12-18.

[26] Long Ouyang, Jeff Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. Technical report, 2022.

[27] Jie Pan, Seungwon Lee, Elliot A Martin, Kiarash Riazi, Hude Quan, e Na Li. Integrating large language models with human expertise for disease detection in electronic health records. 2025.

[28] Associação Brasileira para o Estudo da Obesidade e da Síndrome Metabólica. *Diretrizes Brasileiras de Obesidade*. São Paulo, 4ª edição, 2016.

[29] Jaqueline L Pereira, Michelle A de Castro, Jean M R S Leite, Marcelo M Rogero, Flavia M Sarti, Chester Luís Galvão César, Moisés Goldbaum, e Regina M Fisberg. Overview of cardiovascular disease risk factors in adults in são paulo, brazil: Prevalence and associated factors in 2008 and 2015. *International Journal of Cardiovascular Sciences*, 35:230–242, 12 2021.

[30] Dalton Bertolim Precoma, Gláucia Maria Moraes de Oliveira, Antonio Felipe Simão, et al. Updated cardiovascular prevention guideline of the brazilian society of cardiology - 2019. *Arq Bras Cardiol*, 113(4), 2019.

[31] Victor Sanh, Albert Webson, Colin Raffel, et al. Multitask prompted training enables zero-shot task generalization. 2021.

[32] Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Julien Knafou, Jenny Copara, Lucas E S Oliveira, Yohan B Gumiel, Lucas F A de Oliveira, Douglas Teodoro, Emerson Cabrera Paraiso, e Claudia Moro. BioBERTpt - A Portuguese Neural Language Model for Clinical Named Entity Recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, páginas 65–72, Online, November 2020. Association for Computational Linguistics.

[33] PAUL C. TANG e CLEMENT J. MCDONALD. *Electronic health record systems*. New York: Springer, 2006. p. 447-475.

[34] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, e Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[35] Ozlem Uzuner, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, e Brett South. Evaluating the state of the art in coreference resolution for electronic medical records, 2012.

[36] Betty van Aken, Ivana Trajanovska, Amy Siu, Manuel Mayrdorfer, Klemens Budde, e Alexander Loeser. Assertion detection in clinical notes: Medical language models to the rescue? In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*. Association for Computational Linguistics, 2021.

[37] Ashish Vaswani, NoamShazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, e Illia Polosukhin. Attention is all you need. 2017.

[38] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, e Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.