UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE BIOLOGIA

Leveraging LLMs for Cardiovascular Disease-Oriented Entity Recognition and Relation Extraction in Electronic Health Records

Diogo Miguel Gomes Mataloto

Mestrado em Bioinformática e Biologia Computacional

Dissertação orientada por:

PhD. Maria Fernandes

Prof. Francisco José Moreira Couto

2025

Acknowledgements

First and foremost, I would like to thank my supervisors, Dr. Maria Fernandes and Professor Francisco Couto, for their invaluable guidance and for all the discussions that greatly enriched this project, as well as for helping me to grow and improve as a researcher. I am especially grateful to Sofia Conceição for always taking the time to help me, and to my colleagues Paulo and Mafalda Moreira for sharing the challenges and joys of this journey. To my dear friends António Schiappa, Nair Moreira, Soraia Amaral, Eduarda Nunes, and Francisco Martins, for their advice, encouragement, and friendship along the way. Finally, I thank my parents, my sister, and my little brother Rodrigo for always being there for me. Each of you, in your own way, made this thesis journey not only possible but also truly rewarding and enjoyable.

Resumo

As Doenças Cardiovasculares e Cerebrovasculares (CCDs) continuam a constituir uma das principais preocupações de saúde pública a nível global. Segundo a Organização Mundial de Saúde (OMS), as doenças cardiovasculares foram responsáveis por aproximadamente 17,9 milhões de mortes em 2019, representando um enorme desafio não apenas para a saúde pública, mas também para a economia. Entre os principais fatores de risco associados encontram-se a diabetes, a hipertensão arterial, a obesidade, a dislipidemia, o tabagismo, a história familiar e o uso de medicamentos para doenças crónicas. Esta tendência tem enfatizado a necessidade urgente de desenvolver mecanismos mais eficazes para prevenção, diagnóstico e tratamento, tornando indispensável a criação de ferramentas computacionais capazes de apoiar o processo clínico.

No contexto da digitalização da saúde, os Registos de Saúde Eletrónicos (*EHRs*, *Electronic Health Records*) assumem um papel fundamental, ao integrarem informação médica tanto estruturada (dados demográficos, códigos diagnósticos, resultados laboratoriais, terapêuticas) como não estruturada (notas clínicas). A literatura atual sublinha que uma parte substancial da informação clínica relevante encontra-se nas notas não estruturadas. Estas capturam descrições detalhadas do estado do doente, raciocínio clínico e decisões terapêuticas. No entanto, a análise manual destas narrativas clínicas não estruturadas é impraticável devido ao seu volume, complexidade e heterogeneidade. Para lidar com este desafio, as técnicas de *Natural Language Processing* (*NLP*) têm ganho destaque, possibilitando a deteção de entidades biomédicas e a extração de relações entre estas.

Apesar dos progressos alcançados, os métodos tradicionais baseados em regras, *machine learning* convencional e redes neuronais iniciais apresentam limitações de generalização e dependem fortemente de conhecimento especializado. Para ultrapassar estas restrições, abordagens híbridas e redes neuronais profundas têm sido exploradas, combinando regras explícitas com *machine learning*. Redes convolucionais (*CNNs*) e recorrentes (*RNNs*), associadas a *embeddings* pré-treinados, melhoraram a deteção de entidades e a extração de relações, capturando padrões contextuais e dependências de longo alcance. Mais recentemente, mecanismos de atenção e arquiteturas *transformer* permitiram modelar dependências complexas e lidar com conceitos sobrepostos ou aninhados em texto clínico, preparando o terreno para a aplicação de modelos de linguagem capazes de generalizar para diferentes tarefas biomédicas.

O avanço dos Modelos de Linguagem de Grande Escala (*LLMs*, *Large Language Models*) veio revolucionar este panorama, permitindo avanços significativos na compreensão de linguagem natural em contexto clínico. Modelos como *GPT-4*, *Mistral*, *TxGemma* e *LLaMA-3* demonstram capacidade para

capturar relações semânticas complexas, lidar com dependências de longo alcance e generalizar para diferentes tarefas clínicas. Estes modelos, graças à sua arquitetura baseada em atenção e pré-treinamento em grandes *corpora*, conseguem inferir relações implícitas, detectar sinônimos médicos, compreender abreviações e adaptar-se a estilos de escrita variados presentes nas notas clínicas, o que é particularmente relevante em contextos multilingues ou em instituições com diferentes protocolos de documentação clínica. Esta evolução oferece novas oportunidades para a extração automática de informação em notas clínicas e, em particular, para o estudo das CCDs.

A presente dissertação tem como hipótese que a informação relativa a CCDs, CCDts e às suas relações pode ser extraída com maior desempenho, superando em precisão e capacidade de generalização os modelos anteriores, através da aplicação de *LLMs*. Para validar esta hipótese, foram definidos dois grandes objetivos: (i) desenvolver um *corpus* supervisionado de CCDs, CCDts e das suas relações, através de uma metodologia híbrida de anotação semi-automática (combinação de anotação automática e correção manual); e (ii) conceber, implementar e avaliar um sistema modular de extração de relações clínicas, integrando módulos de deteção de entidades e de extração de relações, baseado em modelos *LLM* e suportado pelo *corpus* supervisionado criado. Além disso, este estudo considera a importância da reprodutibilidade científica, disponibilizando os *corpora* e modelos de forma aberta, respeitando os termos de privacidade, de modo a permitir que outros investigadores possam replicar os resultados, testar variantes metodológicas e expandir o conhecimento na área de *NLP* clínico.

A metodologia foi estruturada em duas fases principais. Na primeira fase, recorreu-se a 331,794 notas clínicas do *MIMIC-IV-Note*, uma base de dados de referência em investigação biomédica, para a criação de três *corpora* supervisionados: CCDs (anotados com *ICD-10*), CCDts (anotados com *MeSH*) e relações binárias terapêuticas. A anotação seguiu uma abordagem semi-automática: pré-anotação com modelos de última geração (incluindo *BENT* e uma versão inicial do *pipeline LLaMIC*) e subsequente correção manual realizada por um estudante de mestrado em bioinformática. Esta combinação de pré-anotação automática e revisão manual garante um equilíbrio entre eficiência e qualidade, mitigando erros que poderiam ser introduzidos por anotadores humanos ou falhas nos modelos pré-treinados. O resultado foi a construção de três *corpora* supervisionados com a seguinte distribuição: CCDs (8,059 notas de treino, 1,727 de validação e 1,728 de teste), CCDts (4,677 de treino, 584 de validação e 585 de teste) e relações binárias (641 de treino, 78 de validação e 80 de teste).

Na segunda fase, foi desenvolvido o *LLaMIC* (*LLaMA Models Applied to MIMIC*), um *pipeline* modular que integra modelos *LLaMA 3.1–8B* para deteção de entidades e extração de relações. Cada módulo foi treinado nos *corpora* gerados e otimizado para uma subtarefa específica: reconhecimento de entidades (CCDs e CCDts), ligação a terminologias padronizadas (*ICD-10* e *MeSH*), extração dos pares de entidades candidatos a relação e classificação das respetivas relações semânticas, com ênfase nas associações terapêuticas. O *pipeline* foi concebido de forma modular para permitir fácil substituição ou atualização de modelos individuais, possibilitando, por exemplo, a integração de novos *LLMs* ou técnicas de pré-processamento de texto sem a necessidade de *retrain* completo do sistema.

Os resultados obtidos demonstram melhorias significativas em algumas subtarefas em relação aos modelos de referência. No reconhecimento de entidades em modo leniente, o melhor modelo *LLaMIC* atingiu uma precisão de 0,887 para CCDs, superando o *BENT* em 32% e aproximando-se do desempenho

para CCDts (redução de 2% em comparação com o *BENT*). Na ligação de entidades às terminologias padronizadas, observou-se um aumento de 21% para CCDs e uma diminuição de 6% para CCDts. Na extração de relações, o *LLaMIC* superou o *BioLinkBERT* em 4%. Estas métricas demonstram que, neste cenário com notas clínicas extensas, erros ortográficos e abreviações frequentes, a integração de *LLMs* permite melhorar a precisão tanto na identificação de doenças quanto na extração de pares de entidades candidatos a relação.

As contribuições principais desta dissertação são duplas. Em primeiro lugar, a disponibilização de três *corpora* supervisionados para CCDs, CCDts e relações terapêuticas, derivados do *MIMIC-IV* e acessíveis através da plataforma *PhysioNet*. Estes *corpora* fornecem uma base para investigação futura. Em segundo lugar, a apresentação do *LLaMIC*, um *pipeline* modular, flexível e com resultados competitivos, que combina modelos *LLaMA* e que se encontra integralmente disponível em acesso aberto, incluindo código, modelos afinados e documentação.

O *LLaMIC* demonstrou que a integração de modelos de linguagem de grande escala melhora a deteção de entidades e a extração de relações em notas clínicas, mesmo com *datasets* reduzidos, superando modelos de referência e lidando com complexidade textual. Como perspetiva futura, sugere-se expandir e diversificar o *corpus* supervisionado, aumentar o número de anotadores e avaliar a concordância inter-anotador para garantir maior fiabilidade. É também relevante implementar o *pipeline* a outras línguas, testando a robustez do modelo em contextos multilíngues. Além disso, recomenda-se integrar a deteção do estado de assertividade das entidades, identificando, por exemplo, se uma doença é confirmada, negada, hipotética ou incerta, e explorar arquiteturas híbridas para extração de relações, combinando modelos generativos e classificadores discriminativos. A incorporação de conhecimento ontológico ou semântico adicional poderá permitir a inferência mais precisa de relações clínicas que não estão explicitamente presentes nos dados. Para aumentar a eficiência, sugere-se otimizar a paralelização e escalabilidade do processamento e realizar estudos ablatórios mais extensos, testando variantes do *pipeline*, diferentes *LLMs* e estratégias de anotação, bem como avaliando estatisticamente a significância das melhorias observadas.

Palavras-chave: Doenças cardiovasculares e cerebrovasculares, Registos de Saúde Eletrónicos, Reconhecimento de Entidades Nomeadas, Extração de Relações, Grandes Modelos de Linguagem

Abstract

Cardiovascular and cerebrovascular diseases (CCDs) remain among the leading causes of morbidity and mortality worldwide, posing significant challenges to healthcare systems. Electronic Health Records (EHRs) store vast amounts of structured and unstructured clinical data, with critical and detailed information often found in free-text clinical notes. Extracting knowledge from these notes is essential to improve patient care and support clinical research. However, clinical notes present substantial challenges, including considerable document length, highly specialized terminology, ambiguities, and multilingual variation.

This dissertation addresses these challenges through the development of a modular pipeline — LLaMIC (LLaMA Models applied to MIMIC) — designed for entity recognition and relation extraction in clinical notes. The work is organized into two phases. In the first phase, the LLaMIC pipeline is presented, integrating the large-scale language model LLaMA for entity detection, linking these entities to standardized terminologies (International Classification of Diseases and MeSH), and extracting relations, with an emphasis on therapeutic associations. In the second phase, the creation of three supervised corpora for CCDs, therapeutic drugs (CCDt), and their respective relations is described, combining automatic annotation with non-expert manual correction, as well as the implementation of the LLaMIC pipeline on these corpora. Evaluation of LLaMIC demonstrates substantial improvements over baseline models. For entity recognition in a lenient mode, the best LLaMIC model achieved a precision of 0.887 for CCDs, surpassing BENT by 32% and closely matching BENT's performance for CCDt (2 percentage points lower). Relation extraction outperformed BioLinkBERT by 4%. The annotated corpora and optimized models are publicly available.

Keywords: Cardiovascular and Cerebrovascular diseases, Electronic Health Records, Named Entity Recognition, Relation Extraction, Large Language Models

Contents

List of Figures

List of Tables

Nomenclature

CCD	Cardiovascular and Cerebrovascular Disease
CCDt	Therapeutic Drug for Cardiovascular and Cerebrovascular Disease
DICN	Deidentified Free-Text Clinical Note
EHR	Electronic Health Record
LLM	Large Language Model
NEL	Named Entity Linking
NER	Named Entity Recognition
NLP	Natural Language Processing
RE	Relation Extraction

Chapter 1

Introduction

1.1 Motivation

Cardiovascular and Cerebrovascular diseases (CCDs) remain a leading global health concern. According to the World Health Organization (WHO), cardiovascular diseases were responsible for approximately 17.9 million deaths in 2019 [WHO, 2024]. These conditions are closely related to various risk factors, including diabetes, hyperlipidemia, hypertension, smoking, obesity, family history of heart disease, and use of medications for chronic illnesses, as described by the WHO. This trend has posed significant challenges to public health and the economy, highlighting the urgency of spreading knowledge about CCDs prevention, diagnosis, and treatment. Enhancing access to this information could mitigate the impact of CCDs, potentially reducing its associated harm and losses for patients [Fuster et al., 2011].

Electronic Health Records (EHRs) are digital repositories of medical information that integrate both structured data - such as demographic information, laboratory results, medications, diagnostic codes, and procedural information — and unstructured data, including clinical notes. The exponential growth of EHR data, the standardization and authenticity has unlocked new opportunities for the development of automated decision-support systems at the point of care and for advancing clinical and translational research [Woldemariam and Jimma, 2023]. A significant proportion of valuable information within EHRs resides in the unstructured clinical narratives [Sheikhalishahi et al., 2019]. The sheer volume and complexity of this type of data make manual analysis impractical for achieving these advancement[H.Hariri et al., 2019].

To address this challenge, Natural Language Processing (NLP) has emerged an tool, enabling the automatic extraction of meaningful information from EHR clinical notes [Jensen et al., 2012]. Central to these approaches are two critical components: Entity Detection, encompassing Named Entity Recognition and Linking (NER/NEL), which identify biomedical entities and associates them with standardized terminologies, and Relation Extraction (RE), which maps interactions between these entities. While traditional methods, including rule-based systems and conventional machine learning approaches,

have demonstrated utility, they are often limited by their reliance on extensive domain expertise, labor-intensive feature engineering, and constrained generalizability across diverse datasets [Lample et al., 2016; Rajkomar et al., 2018].

Preliminary evidence suggests that large language models (LLMs) may significantly impact the NLP landscape on clinical notes, potentially delivering strong performance across a diverse array of tasks. Models such as GPT-4 [OpenAI, 2024], Mistral [Jiang et al., 2023], TxGemma [Wang et al., 2025], and LLaMA3 [Touvron et al., 2023] utilize advanced deep learning architectures—such as the attention mechanism—and pretraining on large datasets to capture intricate linguistic and semantic relationships. By leveraging these capabilities, LLMs could help overcome many of the challenges associated with processing unstructured clinical notes, including modeling context, handling long-range dependencies, and generalizing across varied tasks [Yang et al., 2022]. Although research in clinical information extraction has made notable progress [Si et al., 2019; Lee et al., 2019], and efforts in CCD-specific information extraction are underway [Chang et al., 2023], the advent of LLMs, which have demonstrated significant breakthroughs in language comprehension, presents new opportunities to enhance performance in this field [Boonstra et al., 2024].

1.2 Problem

As previously noted, clinical notes present significant challenges for tasks such as entity detection and RE due to two primary issues:

1. **Document Length:** Clinical notes are often composed of extensive narratives, with critical information dispersed across lengthy, unstructured text. Some NLP models often fail to capture dispersed contextual interdependencies or suffer from memory degradation when processing long sequences.

2. **Textual Complexity:** The language of clinical notes is marked by highly domain-specific terminology, extensive use of abbreviations, inherent ambiguity (where identical terms may refer to different entities), and frequent misspellings. These characteristics significantly diminish the effectiveness of NLP models.

1.3 Objectives

Current research trends addressing the challenges inherent to clinical notes focus on developing generalizable, interdisciplinary systems for entity recognition and relation extraction tasks. In this context, the primary objectives of this thesis are:

1. **To develop supervised corpora** for cardiovascular and cerebrovascular diseases, therapeutic drugs, and their relations by employing a hybrid annotation methodology that combines automatic annotation and manual correction.

2. **To design, implement, and evaluate a modular clinical relation extraction system**—comprising entity detection and RE modules—capable of identifying and classifying relationships between CCDs and CCDts from clinical text, supported by the supervised corpus.

1.4 Methodology

The overall methods to accomplish the proposed objectives can be divided into two phases, one for each objective. The summarized schematic of this methodology is presented in Figure 1.1. In the first phase is the creation of three supervised corpus (CCDs, CCDts and relations) (Chapter 4). The second phase is design (Chapter 3) and the implementation of a LLM models pipeline for entity detection and RE.

In the first phase, a semi-automated annotation procedure was applied to clinical notes derived from the state-of-the-art MIMIC-IV database. This phase focused on two main entity types and their corresponding standard terminologies: CCDs (ICD-10) and CCDts (MeSH). The semi-automated annotation involved (i) an automatic pre-annotation process using pre-trained transformer models—specifically BERT for entity detection and our unfine-tuned LLaMIC pipeline for both entity detection and relation extraction—and (ii) subsequent manual correction performed by a bioinformatics master's student.

In the second phase, we developed a modular pipeline, referred to as LLaMIC (LLaMA Models Applied to MIMIC), to perform entity detection and relation extraction on clinical notes. The system integrates an open-source large language model (LLaMA 3.1–8B) along with a lexicon-based method for entity detection. Each component of the pipeline is fine-tuned using the previously created supervised corpora and is specialized for a particular subtask: entity recognition modules detect CCDs and CCDts; entity linking modules map recognized spans to standardized terminologies (ICD-10 and MeSH); and the relation extraction component identifies semantic links between entities from a predefined list of possible relations, particularly therapeutic associations.

1.5 Contributions

This work makes two key contributions to clinical natural language processing. First, we introduce an open-access corpus for CCDs, CCDts, and their relations, derived from the MIMIC-IV database with annotations for diseases, drugs, and therapeutic relationships. Second, we present a competitive *LLaMIC*, a modular and flexible pipeline that integrates LLaMA models for entity detection and relation extraction, specifically designed for MIMIC-IV clinical notes, with all code and fine-tuned models publicly available.

1.5.1 Objective 1

Chapter 4 presents a pipeline to generate three weakly supervised corpora for CCDs, CCDts, and their relations. The work for this objective resulted in three restricted-access corpora, which are provided in the *Supplementary Documents* and will be made publicly available after the thesis defense. Each corpus

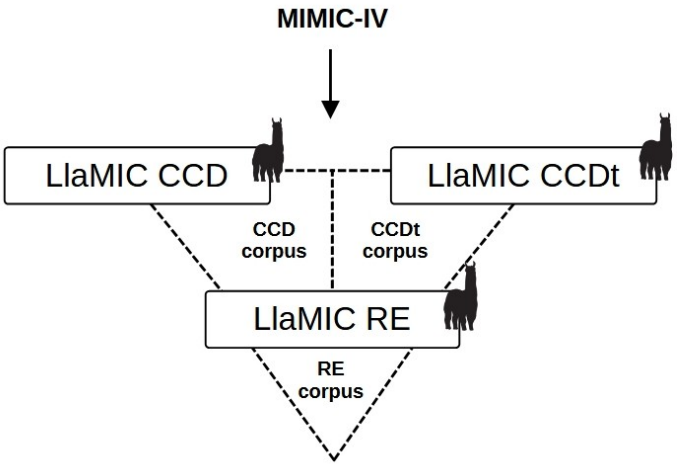

Figure 1.1: Overview of the Methodology: The framework consists of three LLaMIC modules: LLaMIC-CCD for annotating CCDs and linking them to ICD-10, LLaMIC-CCDt for annotating CCDts and linking them to MeSH, and LLaMIC-RE for extracting relationships between the annotated entities. All models were trained on their respective supervised corpora for each task, including supervised CCD, supervised CCDt, and supervised RE corpora. The LLaMIC modules were evaluated not only on the test sets of these supervised corpora but also applied to unsupervised corpora for each task.

includes train/dev/test splits, as well as a prediction corpus generated by direct inference of our best model on an unsupervised MIMIC dataset. The corpora are accessed through PhysioNet under the terms of PhysioNet's License 1.5.0 and Data Use Agreement 1.5.0, following completion of the required CITI training.

1.5.2 Objective 2

Chapter 3 describes the design of our proposed LLaMIC model, which is fully available at `https://github.com/lasigeBioTM/LLAMIC`. The pipeline is highly flexible, allowing the integration of other clinical notes, LLMs, and fine-tuning procedures. Chapter 4 details the implementation and training of the pipeline on the supervised corpora generated in Objective 1, resulting in six fine-tuned LLaMA models that serve as the modules for LLaMIC. Our best LLaMIC model achieved a precision of 0.831 for lenient NER on CCDs, surpassing the BERT baseline by 32%, and performed within 2% of the BERT baseline on CCDts lenient NER. For entity linking, improvements of +21% and -6% were observed for CCDs and CCDts, respectively. In RE, our model outperformed the BERT baseline by 4%. The fine-tuned LLaMA models are included in the *Supplementary Documents* and will be made publicly available after the thesis defense.

1.6 Thesis Structure

This thesis is organized into four main chapters:

- **Chapter 2 – Background and Related Works**: This chapter presents a comprehensive literature review. It begins with an overview of Electronic Health Records (EHRs), highlighting the role of unstructured clinical texts and the application of Natural Language Processing (NLP) techniques in this domain, including entity recognition, entity linking and relation extraction. It then explores the fundamentals of Large Language Models (LLMs), the Transformer architecture, and recent advancements in transfer learning, particularly through QLORA fine-tuning. The chapter concludes with a review of state-of-the-art clinical NLP corpora and evaluation metrics.
- **Chapter 3 – LLaMIC Pipeline: LLM and Lexicon-Based Integration**: This chapter details the proposed LLaMIC pipeline. It describes the modular architecture, which combines LLMs with lexicon-based strategies for robust entity recognition and relation extraction. Adaptability to multiple LLM backbones is also discussed, demonstrating the pipeline's flexibility.
- **Chapter 4 – LLaMIC Implementation and Corpus Construction**: This chapter outlines the practical implementation of the LLaMIC pipeline using the MIMIC-IV dataset. It describes the corpus selection, preprocessing routines, and semi-automatic annotation procedures for both entity and relation annotations. Furthermore, it presents the training and evaluation protocols adopted for LLaMIC and includes a detailed discussion of results and error analysis.
- **Chapter 5 – Conclusion and Future Work**: The final chapter synthesizes the main findings of the thesis, reflects on its limitations, and proposes directions for future research in the domain of clinical NLP with LLMs.

Chapter 2

Background and Related Works

2.1 Electronic Health Records

Electronic Health Records (EHRs) are digital systems designed to automatically compile and manage a patient's health information. These records can consist of both structured and unstructured data. Structured data typically comes from well-defined, standardized sources, such as diagnostic codes, prescription information, or laboratory test results. In contrast, unstructured data includes free-text inputs, such clinical notes.

The adoption of EHRs has accelerated significantly with the digital transformation of healthcare. In the United States, for example, the adoption rate of EHRs rose sharply from 9% in 2008 to 96% in 2023 [AHA and NCHS, 2023]. This rapid increase has created an immense opportunity for research in areas such as "Big Data" and Natural Language Processing (NLP), which are integral to extracting actionable insights from large-scale healthcare data.

2.1.1 Clinical Notes

Structured EHR data alone may not suffice for accurate research, as it can bias findings by underrepresenting incidence and prevalence [Arslan et al., 2022] and low performance of prediction models [Seinen et al., 2023]. In contrast, unstructured sources—like clinical notes—can enrich research results by capturing a more accurate representation of the patient's situation [Wieland-Jorna et al., 2024; Rahal et al., 2021]. However, understanding and extracting meaningful information from clinical notes using NLP techniques remains a significant challenge:

- **Document Length:** Free-text clinical notes often consist of extensive narratives, where important information is dispersed throughout long and frequently unstructured documents.
- **Textual Complexity:** Entities may appear in various forms, including abbreviations, synonyms, and common typographical errors. For example "The patient was started on *asprin* after a suspected

MCA stroke.". Additionally, relationships between entities are often implied by the surrounding context rather than explicitly stated. For instance "Hx of CAD: (...) heparin maintained for prevention."

2.1.2 Natural Language Processing in the Clinical Notes

Language is defined as a system of rules or combined symbols used to convey information. Understanding this information and obtaining meaning from it, particularly from a computer perspective, is a complex task. In unstructured or highly heterogeneous domains, human text involves handling phonology, morphology, semantics, syntax, sentence structure, and pragmatics [Khurana et al., 2022].

In the fields of Artificial Intelligence and Linguistics, this task is referred to as NLP. Despite some progress in NLP, extracting meaningful information remains a formidable challenge, particularly in clinical applications where the language is often domain-specific, ambiguous, and unstructured [Gruber, 1993]. In this context, some of the key tasks researched include Entity Detection, which comprises Named Entity Recognition (NER) and Named Entity Linking (NEL), as well as Relation Extraction (RE) [Goyal and Singh, 2025].

Before talking about these types of complex tasks, it is essential to understand the foundations of NLP. These NPL tasks, broadly categorized into syntactic (low-level) and semantic (high-level, including NER, NEL, and RE) processing, are inherently interconnected, forming the basis for understanding and deriving meaning from text, as illustrated in Figure 2.1 [Nadkarni et al., 2011]

The main low-level tasks include:

- **Tokenization**: Refers to the division of text into basic linguistic units — such as words, numbers, and punctuation. It is a foundational step in NLP, and the complexity of this task can vary depending on the language, domain, or the tokenization strategy used.

- **Part-of-Speech Tagging**: The assignment of grammatical categories (e.g., noun, verb, preposition) to each token in a sentence, based on its syntactic and semantic context. Part-of-Speech (POS) tagging supports syntactic parsing and enhances the interpretation of sentence structure.

- **Parsing**: The analysis of the syntactic structure of sentences according to a formal grammar. Parsing identifies hierarchical relationships between tokens and phrases, enabling deeper linguistic and semantic interpretation. The complexity of parsing depends on the grammar used and the desired level of detail.

2.1.2.1 Entity detection

The NER and NEL tasks involve identifying the boundaries of named entities within text and linking them to a standardized terminologies — such as a classification system or knowledge base — that provides context for the recognized entities [Velupillai et al., 2015].

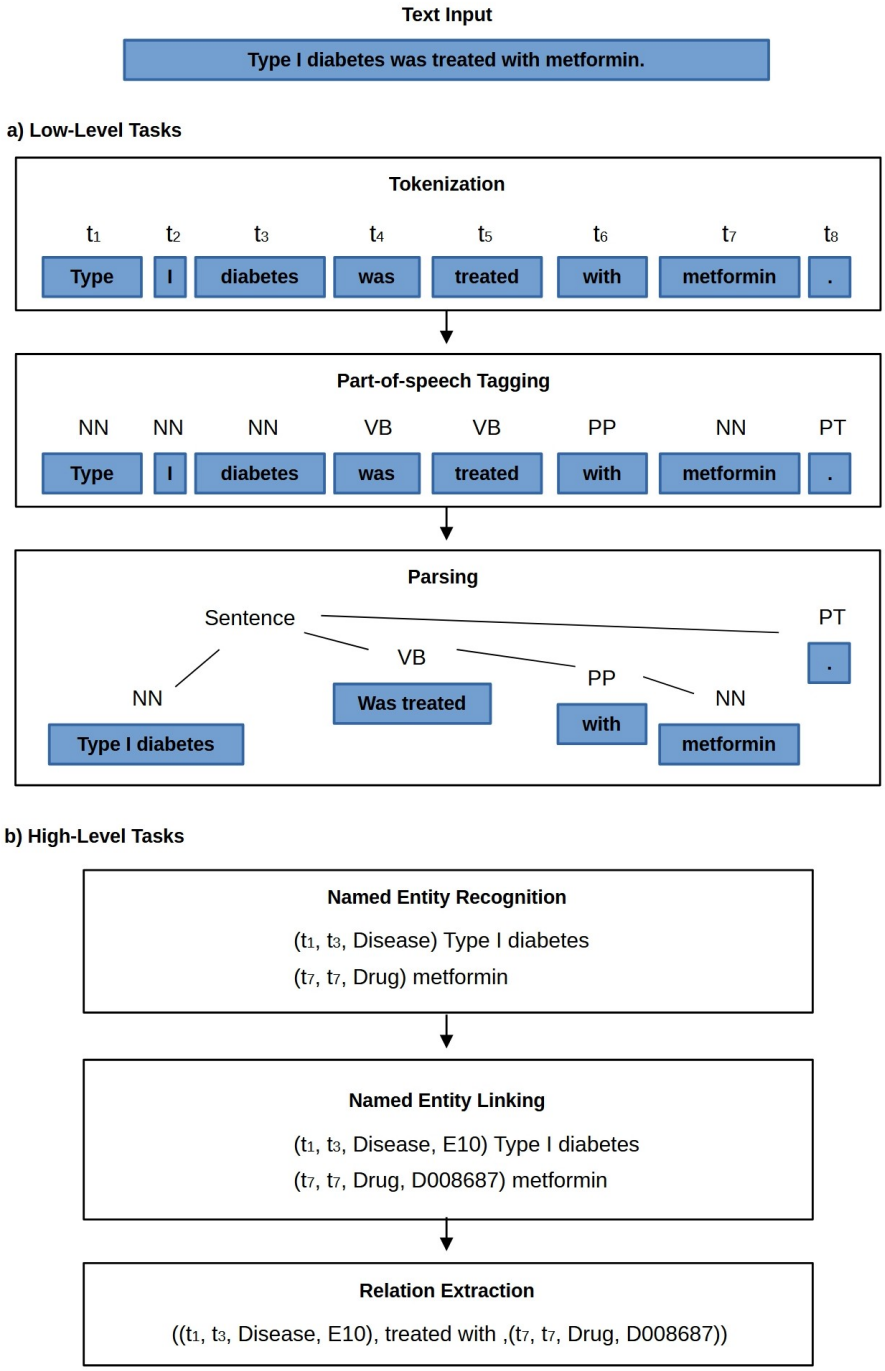

Figure 2.1: Overview of an information extraction pipeline for NLP tasks, broadly categorized into: (a) syntactic (low-level) processing; and (b) semantic (high-level) processing, encompassing tasks such as NER, NEL, and RE. t (token) NN (nouns), VB (verbs), PP (prepositions), and PT (punctuation).

Named Entity Recognition: is typically formulated as a sequence tagging task, building upon earlier techniques such as POS tagging and parsing. Given a sequence of N tokens $T = (t_1, t_2, \ldots, t_N)$, the goal is to identify spans of text that correspond to named entities, represented as tuples (I_s, I_e, l), where I_s and I_e denote the start and end indices of the entity in the token sequence, and l refers to the entity label drawn from a predefined set (e.g., *Disease*, *Drug*). For instance, based on the example in Figure 2.1, in the sentence *"Type 1 diabetes was treated with metformin"*, NER would extract:

- $(t_1, t_3, \textit{Disease})$ for "Type 1 diabetes"
- $(t_7, t_7, \textit{Drug})$ for "metformin"

Named Entity Linking: extends the NER formulation by associating each mention with a unique identifier from a standardized terminologies [Singh et al., 2024]. Standardized terminologies refer to formalized representations of knowledge, such as classification systems or knowledge bases. In the context of EHRs, structured repositories play a critical role in organizing and linking healthcare data, particularly structured data. For example, they enable the creation of tables with ICD-10 diagnosis codes [WHO, 2019] for patients or the organization of clinical interventions and medication prescriptions using standards such as the National Drug Code (NDC). Some commonly used standardized terminologies in biomedical NLP literature include:

- **International Classification of Diseases (ICD):** A globally recognized system maintained by the World Health Organization (WHO) to classify and code diseases, health conditions, and related health problems. ICD-9 and ICD-10 are different versions of the classification, with ICD-9 published in 1977 and ICD-10 in 1990.
- **Disease Ontology (DO):** A standardized ontology developed by the Institute for Genome Sciences at the University of Maryland School of Medicine, designed to provide consistent, structured classification and annotation of human diseases [Schriml et al., 2018].
- **Medical Subject Headings (MeSH):** A comprehensive controlled vocabulary thesaurus maintained by the U.S. National Library of Medicine (NLM) [Lipscomb, 2000]. MeSH terms cover a wide range of biomedical topics, including diseases, chemicals, anatomical terms, and medical procedures.

In NEL, a named entity is now represented as a tuple (I_s, I_e, l, i), where I_s and I_e denote the start and end indices of the mention, l is the entity label, and i is a unique identifier. Continuing with the example from Figure 2.1:

- $(t_1, t_3, \textit{Disease}, \texttt{E10})$ — where `E10` refers to the ICD-10 code for Type 1 diabetes
- $(t_7, t_7, \textit{Drug}, \texttt{D008687})$ — where `D008687` is the MeSH ID for metformin

2.1.2.2 Relation Extraction

RE is the task of identifying and categorizing semantic relationships between entities. In the biomedical domain, RE encompasses both binary relations, i.e., semantic links between pairs of entities such as diseases, drugs, genes, and phenotypes, as well as *n*-ary relations involving three or more entities. The study of *n*-ary relations aims to capture more complex interactions, such as clinical events or biochemical pathways [Peng et al., 2017].

Formally, given a set of recognized entities $\mathcal{E} = \{(I_s^1, I_e^1, l_1), (I_s^2, I_e^2, l_2), \dots\}$ from a token sequence $T = (t_1, t_2, \dots, t_N)$, the goal is to extract semantic relationships between entity pairs. A binary relation is typically represented as a triple (e_h, r, e_t), where e_h and e_t are entity mentions from $\mathcal{E}$, and r is a relation label drawn from a predefined set $\mathcal{R}$ (e.g., *treats*, *causes*, or *associated_with*). Each tuple defines a directed relationship between a head entity e_h and a tail entity e_t mediated by a relation r, for instance, in the sentence *"Type 1 diabetes was treated with metformin"*, the RE system might output:

- (*"Type 1 diabetes"*, *treated_with*, *"metformin"*)

2.1.2.3 From Rules to Recurrent Networks Approaches

Entity detection and RE within clinical notes can be categorized into three primary frameworks: rule-based, machine learning, and deep learning methods [Jauregi Unanue et al., 2017]. Table 2.1 provides an overview of prominent text mining tools exemplifying these approaches, which will be referenced throughout this work.

Model	Architecture	NER	NEL	RE	Reference
MetaMap	Rule-based	x	x		[Aronson and Lang, 2010]
cTAKES	Rule + Machine learning–based	x	x	x	[Savova et al., 2010]
MER	Rule-based	x	x		[Couto and Lamurias, 2018]
DNorm	Machine learning–based	x	x		[Leaman et al., 2013]
CLAMP	Rule + Machine learning–based	x	x	x	[Soysal et al., 2017]
BioBERT	BERT-based	x	x	x	[Ji et al., 2020]
BENT	PubMedBERT - rule - graph-based	x	x		[Ruas and Couto, 2022]
K-RET	Knowledge graphs + BERT-based	x	x	x	[Sousa and Couto, 2023]

Table 2.1: Overview of text mining tools, their architecture, and applicability to NER, NEL, and RE tasks.

Early clinical entity detection and RE systems relied heavily on predefined lexical resources and manually crafted syntactic or semantic rules. **Rule-based approaches** not only utilized regular expressions or pattern matching, but also incorporated additional constraints. For example, NegEx [Chapman et al.,

2001] is a tool designed to identify negated concepts in clinical text (e.g., distinguishing "no evidence of pneumonia"), while ConText [Harkema et al., 2009] extends this by capturing additional contextual cues, such as temporality, certainty, and experiencer (e.g., whether a condition is historical or hypothetical, or pertains to someone other than the patient). Several rule-based annotation tools have been employed in biomedical field, including cTAKES [Savova et al., 2010] and MetaMap [Aronson and Lang, 2010]. Most of these tools can extract various types of named entities and map them to concepts in the UMLS. However, they rely heavily on external dictionaries and are primarily tailored for English-language. Due to this, they present limited adaptability to new datasets and domains.

Machine learning models depended on manually annotated corpora and extensive feature engineering [Zhang et al., 2018]. Among these, Conditional Random Fields (CRFs), kernel-based Support Vector Machines (SVMs) and the generalized model Structured Support Vector Machines emerged as key algorithms for structured prediction tasks in biomedical NLP [Lafferty et al., 2001; Crammer et al., 2002; Tsochantaridis et al., 2005]. For instance, Bundschus et al. [Bundschus et al., 2008] applied CRFs to both NER and RE, focusing on disease–treatment and gene–disease associations from manually annotated PubMed abstracts and the GeneRIF database. Similarly, Rink et al. [Rink et al., 2011] employed SVMs to extract relations between medical records and treatments.

In addition, hybrid methods that combine machine learning with rule-based or feature-engineered components have been proposed to enhance NER and RE performance. Notably, Bruijn et al. [de Bruijn et al., 2011] applied such approaches to extract medical problems, tests, and treatments from discharge summaries and progress notes, while Tang et al. [Tang et al., 2013] explored hybrid strategies for biomedical NER using data from the 2010 i2b2 NLP challenge. In NEL task, DNorm further introduced a pairwise learning-to-rank approach that learns the similarity between entity mentions and concepts in a knowledge base [Leaman et al., 2013]. Despite their contributions, these approaches are inherently limited by the scarcity of labeled data. To address this, in RE **distant supervision** was proposed by Mintz et al. [Mintz et al., 2009], aligning textual mentions with relational facts from a knowledge base.

With the advent of **neural networks** for NLP, the focus shifted from manual feature engineering to the design of network architectures that incorporate inductive biases to support effective feature learning [Zeng et al., 2014]. Neural networks are composed of layers of interconnected nodes or neurons—including input, hidden, and output layers—where each neuron applies a nonlinear activation function to the weighted sum of its inputs. These weights are adjusted during training through algorithms such as backpropagation. The depth of a network, referring to the number of hidden layers, plays a crucial role in enabling the model to capture increasingly abstract and complex patterns. this transition gave rise to deep neural networks such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Another key development was the adoption of word embeddings—replacing traditional bag-of-words models—which enabled richer semantic representations. Furthermore, the availability of large unlabeled corpora facilitated the pre-training of deep neural models, which could then be fine-tuned for specific downstream tasks. This paradigm significantly improved the capacity of models to capture and generalize over complex linguistic structures. In the biomedical domain, CNNs have been applied to

RE tasks at both intra- and inter-sentence levels, such as in the BioCreative-V chemical–disease relation dataset and the DDI corpus for drug–drug interaction extraction [Gu et al., 2017; Liu et al., 2016]. Luo et al. [Luo et al., 2018] introduced a multi-view CNN with a shared multi-task architecture to normalize diagnostic and procedural terms in Chinese discharge summaries by mapping them to standardized concepts.

Bidirectional Long Short-Term Memory (BiLSTM) networks have also been widely explored. For instance, Dong et al. [Dong et al., 2019] proposed a BiLSTM encoder with a segment attention layer and a tensor-based mechanism for NER in Chinese electronic medical records, focusing on problems, treatments, and test entities. Similarly, Weegar et al. [Weegar et al., 2019] applied a BiLSTM-CRF architecture for NER tasks on Swedish general medical corpora—covering body parts, disorders, and clinical findings—as well as Spanish clinical discharge reports for the recognition of diseases and medications.

2.2 Large Language Models

The transformer model, proposed by [Vaswani et al., 2017], abandons the use of recurrent and convolutional layers, described earlier, in favor of a fully attention-based mechanism. This allows the model to focus on the most relevant parts of the input. This architecture gave rise to various transformer families, including pretrained models such as BERT (Bidirectional Encoder Representations from Transformers) and larger-scale models known as large language models (LLMs)[1], which are described in more detail below.

2.2.1 Attention Mechanism

In the attention mechanism, for each token in the input sequence, the model learns three distinct sets of parameters: the query weights W_Q, the key weights W_K, and the value weights W_V.

The attention function can be expressed as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \tag{2.1}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ represent the query, key, and value matrices derived from the input sequence, respectively. The query and key matrices are used to compute a similarity score, while the value matrix holds the actual information to be weighted and passed to the next layer. The softmax operation ensures that the attention weights are normalized, allowing the model to focus on the most relevant tokens in the input sequence. The scaling factor $\sqrt{d_k}$ helps stabilize gradients during training by preventing excessively large dot products in high-dimensional spaces, where d_k is the dimensionality of the key vectors.

[1] In this work, we distinguish earlier transformer-based models, such as BERT, which are generally smaller in scale, from the newer generation of LLMs, which are significantly larger and trained for generative tasks.

This formulation allows the model to dynamically adjust the attention given to different parts of the input sequence based on their relevance to each token.

2.2.2 Transformer Architecture

Briefly, the transformer architecture consists of two main components: an **encoder** and a **decoder**, as illustrated in Figure 2.2. The model receives text input, which is first tokenized into discrete units. During the embedding step, the tokens are converted into numerical vectors, where each token is matched with a pre-trained vector from an embedding table. Positional encoding is then added to capture the order of the sequence.

The **encoder**, depicted on the left-hand side of the same figure, consists of multiple identical layers, each containing the following key components:

- **Multi-Head Self-Attention**: The model performs the attention function in parallel through h distinct heads, as illustrated in Figure 2.3. Each head linearly projects the input embeddings into queries, keys, and values of dimension $d_{\mathbf{K}}$ and $d_{\mathbf{V}}$, applies scaled dot-product attention, and produces an output of dimension $d_{\mathbf{V}}$. The outputs of all heads are then concatenated and linearly transformed. This mechanism allows the model to capture different types of relationships and to assess the importance of each word in the context of the others, regardless of their distance in the sequence.

- **Feed-Forward Network**: Following the self-attention mechanism, a position-wise fully connected feed-forward network processes the attention outputs independently for each position.

The **decoder**, on the right-hand side of the Figure 2.2, mirrors the encoder's structure, with two key distinctions:

- **Cross-Attention**: Cross-attention acts as a bridge between the encoder and decoder. In this sublayer, the decoder's hidden states are used to form the queries, while the keys and values are taken from the encoder's output representations. This multi-head attention mechanism enables the decoder to selectively focus on relevant parts of the encoded input sequence when generating each token.

- **Masked Self-Attention**: In addition to self-attention, the decoder employs a masked self-attention mechanism. This prevents the model from "seeing" future tokens in the sequence during the generation process, thereby maintaining the autoregressive property necessary for generating coherent and contextually appropriate outputs.

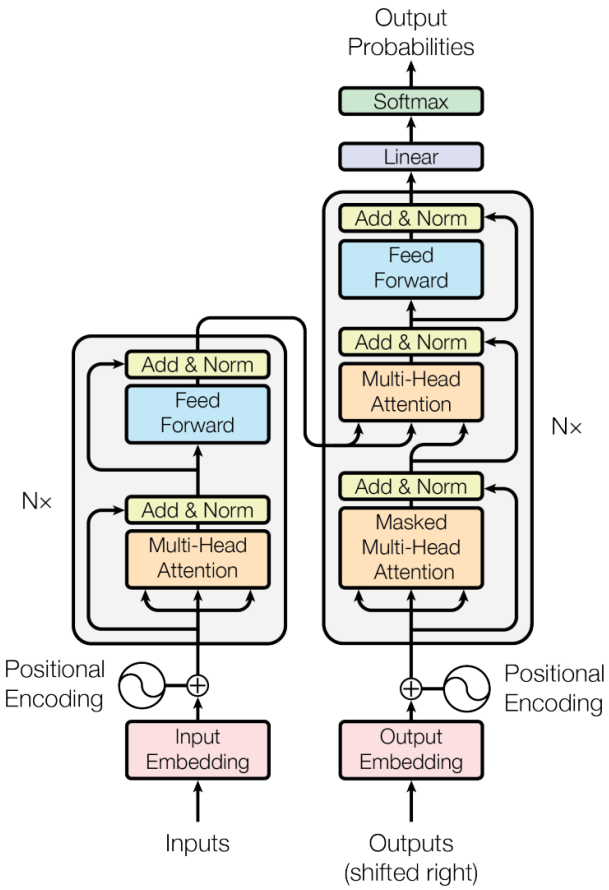

Figure 2.2: The encoder-decoder structure of the transformer [Vaswani et al., 2017]

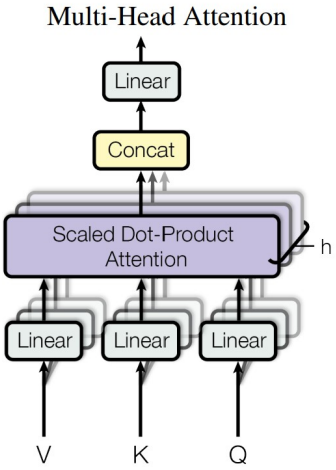

Figure 2.3: Multi-Head Attention [Vaswani et al., 2017]

The generated text is produced one token at a time, with each token being selected based on the output probabilities, which are computed by a linear layer followed by a softmax function.

In contrast to models like BERT, which employ only the encoder component, generative models, such as those based on the GPT architecture, rely solely on the decoder component, focusing on generating text based on a prompt. In decoder-only models, the input text is directly fed into the decoder through masked self-attention layers, without a separate encoder component.

2.3 Large Language Models in Clinical NLP: Beyond BERT

The introduction of self-attention mechanisms established transformer-based architectures as the most used framework in NLP. By replacing recurrence with self-attention, transformers efficiently model long-range dependencies and manage nested or overlapping concepts [Kim et al., 2024]. Their strong transfer learning capabilities, that is, the ability to apply their prior knowledge to new tasks, further support adaptability across NLP tasks [Li et al., 2020; Hu et al., 2024].

Among pre-trained transformer models, encoder-only models, such as BERT which is widely adopted for biomedical NLP [Devlin et al., 2019; Fraile Navarro et al., 2023]. Variants such as BioBERT [Lee et al., 2019], SciBERT [Beltagy et al., 2019], PubMedBERT [Gu et al., 2021], and ClinicalBERT [Huang et al., 2020] —trained on biomedical and clinical corpora—improve performance on domain-specific tasks. A systematic review by [Yang et al., 2020] examined four BERT-based architectures, highlighting their effectiveness in clinical concept extraction.

LLMs, which are decoder-only models, including GPT (Generative Pretrained Transformer), extend the capabilities of transformer architectures by generating coherent and contextually relevant text through training on billions of tokens from diverse corpora. Building on recent biomedical studies, LLMs have demonstrated promising results in clinical NLP tasks such as question answering, named entity recognition, knowledge-guided reasoning, and automated de-identification [Gilson et al., 2023; Hu et al., 2024; Su et al., 2025; Altalla' et al., 2025]. The primary advantages of LLMs for clinical NLP tasks, when compared to encoder-only models, can be outlined as follows:

- **Versatility and Adaptive Learning:** Trained with an auto-regressive objective that predicts each token based on the preceding context, LLMs are inherently generative and highly adaptable. Furthermore, their support for few-shot and zero-shot learning enables them to generalize from a few in-context examples without requiring explicit fine-tuning.

- **Scale and Knowledge Encoding:** LLMs contain hundreds of billions of parameters, enabling them to encode a vast amount of linguistic and domain-specific knowledge. This scale, along with autoregressive training described above and long-range self-attention, allows for efficient processing of complex medical language, including specialized jargon, ambiguous abbreviations, misspellings, and cross-lingual variation.

2.3.1 Transfer Learning through QLoRA Fine-Tuning of Pretrained LLMs

Fine-tuning has emerged as a key technique for adapting pre-trained LLMs to specialized tasks in domains such as healthcare, where annotated data is often [Zhang et al., 2018]. This process typically involves supervised training on curated datasets—e.g., for clinical entity recognition or diagnosis-treatment relation extraction—where model outputs are aligned with ground truth annotations via backpropagation. Fine-tuning generally includes:

1. **Task-specific input adaptation**: Preprocessing and formatting input data to match the model's expectations, such as tokenization or the inclusion of task-specific prompts.

2. **Model parameter optimization**: Updating the model's weights through gradient-based learning, usually with a low learning rate to retain general knowledge while integrating task-specific patterns.

3. **Custom output mapping**: Adjusting the output layer to match the target task, whether by predicting token classes or generating structured outputs.

To optimize only specific parameters of the models for text-to-text generation, Parameter-Efficient Fine-Tuning addresses the high resource demands of full model finetuning. Methods such as Low-Rank Adaptation (LoRA) and its quantized variant, QLoRA, significantly reduce both the number of trainable parameters and GPU memory usage [Hu et al., 2021; Dettmers et al., 2023]. LoRA introduces trainable low-rank matrices into existing weights, optimizing only these additions. QLoRA further compresses the model by quantizing the low-rank matrices, achieving comparable performance with lower computational cost. These approaches are particularly valuable in low-resource settings where GPU memory is constrained.

The effectiveness of LoRA and QLoRA depends on tuning of key adaptation-related hyperparameters. The rank determines the dimensionality of the low-rank matrices and thus controls the capacity of the adaptation. The alpha scaling factor adjusts the magnitude of the low-rank updates before they are integrated with the base weights, influencing how strongly the model incorporates task-specific knowledge. Additionally, the dropout rate acts as a regularizer by randomly deactivating parameters during training, which helps prevent overfitting.

Beyond these adaptation-specific settings, other model hyperparameters are also critical for optimizing LLM performance. The learning rate sets the step size at each iteration of gradient descent and affects both convergence speed and stability. Batch size, defined as the number of training samples processed per iteration, impacts training stability and memory efficiency. The number of epochs determines how many times the entire training dataset is passed through the model, influencing overall training duration. Finally, weight decay serves as a regularization parameter that penalizes large weights, promoting simpler models and further reducing overfitting.

2.3.2 State-of-Art Open-Source LLMs Models

In the context of biomedical applications, particularly for EHRs, open-source LLMs are increasingly preferred, as they allow local fine-tuning on sensitive data. Below, we highlight some of the most relevant in this biomedical field:

- **mistralai/Ministral-7B-Instruct:** an 7 billion parameter model from the Mistral AI team that outperforms other models of similar size to date [Jiang et al., 2023]. Notably, [Labrak et al., 2024] demonstrates strong performance on the BioMistral multilingual medical evaluation benchmark.
- **meta-llama/Llama-3.1-8B-Instruct:** an 8 billion parameter model from Meta [Touvron et al., 2023]. In the clinical notes domain, various research groups have successfully fine-tuned LLaMA for domain-specific tasks [Wang et al., 2024].
- **google/txgemma-9b-chat:** is a recently developed therapeutic-focused language model fine-tuned from the Gemma 2 base, with 9 billion parameters [Wang et al., 2025]. It is designed to process and understand therapeutic data, including small molecules, proteins, and diseases.

2.4 Clinical NLP Corpora and Benchmarks for CCDs

Over the past two decades, a diverse range of annotated English and non-English clinical corpora for disease and drug entities, with a focus on cardiovascular diseases (Tables 2.2 and 2.3), has become increasingly available, supporting the development of NLP models for entity detection and relation extraction (Table 2.4). Early datasets, such as those from LDS Hospital and the Mayo Clinic, focused on identifying medical problems and diseases (e.g., cardiovascular diseases) in clinical narratives [Meystre and Haug, 2006; Savova et al., 2010]. The i2b2 shared tasks, expanded the annotation scope to include temporal relations, assertions, and complex inter-entity relationships [Uzuner et al., 2010, 2011; Stubbs and Uzuner, 2015; Henry et al., 2020]. These benchmarks have been widely used to evaluate machine learning models, including CRF, SVM, and early deep learning architectures such as LSTM and Transformer-based models [Uzuner et al., 2010, 2011; Stubbs et al., 2015; Houssein et al., 2023; Henry et al., 2020; Belkadi et al., 2023]. Several corpora derived from the MIMIC-III database, including the ShARe corpus and the 2018 n2c2 challenge dataset, have also contributed to this field [Mowery et al., 2013; Pradhan et al., 2014; Elhadad et al., 2015; Henry et al., 2020]. More recently, the CASI Medication Status corpus focused specifically on medication-related entities, achieving high performance in NER tasks using GPT-3 (92% F1). Efforts have also extended beyond English, with annotated corpora in Spanish, including ICD-10 codes [Goeuriot et al., 2020; Suominen et al., 2021; Marimon et al., 2017], French [Campillos et al., 2018], and Chinese, mainly focused on cardiovascular diseases such as the Chinese CVD Risk EMR Corpus and the recent CVDEMRC [Su et al., 2017; Richter-Pechanski et al., 2023]. Additionally, a German corpus has been developed in this domain [Chang et al., 2023]. The JREwBART model applied to the Chinese CVDEMRC corpus achieved 64% F1 in relation extraction tasks [Guo et al., 2024].

ID	Year + Dataset	Annotation / Focus	Size	Data Source
[1]	*2006 – LDS Hospital* [Meystre and Haug, 2006]	NER and Negation of medical problems (e.g., CAD)	160 EHR notes	LDS Hospital Cardiovascular unit
[2]	*2008 – Mayo Clinic EMR* [Savova et al., 2010]	NER of diseases	160 EHR notes	Mayo Clinic
[3]	*2009 – i2b2 Medication Challenge* [Uzuner et al., 2010]	NER and RE of drugs and their attributes (e.g., reason)	251 EHR notes	Partners Healthcare
[4]	*2010 – i2b2 Relations Challenge* [Uzuner et al., 2011]	NER, RE, and Assertions for problems, treatments, and tests	871 reports	Partners Healthcare, Beth Israel Deaconess, University of Pittsburgh
[5]	*2014 – i2b2 De-identification and CAD Risk Factors Challenge* [Stubbs and Uzuner, 2015]	NER (CVD risk factors), RE (temporal), and Assertions	1,304 EHR notes	Diabetic patients at Partners Healthcare
[6]	*2018 – n2c2 Adverse Drug Events and Medication Challenge* [Henry et al., 2020]	NER and RE of medications and ADEs (Adverse Drug Events)	505 EHR notes	MIMIC-III
[7]	*2014 – ShARe Corpus* [Mowery et al., 2013; Pradhan et al., 2014; Elhadad et al., 2015]	NER (medical problems), attributes, and RE (temporal modifiers)	298 clinical reports, +133 for SemEval-2014, +100 for SemEval-2015	MIMIC-II for CLEF/ShARe 2013 and SemEval Tasks
[8]	*2020 – CASI Medication Status [Moon et al., 2014; Agrawal et al., 2022]*	Medication NER + Status	340 drug–status pairs from clinical notes snippets	University of Minnesota-affiliated hospital EHRs

Table 2.2: Overview of clinical NLP corpora including entity, relation, and assertion annotations, with size and source institutions.

ID	Year + Dataset	Annotation / Focus	Size	Data Source
[9]	*2020 – CLEF eHealth Shared Evaluation: Task I* [Goeuriot et al., 2020]	NER and NEL (Named Entity Linking) with CIE10 codes	1,000 clinical cases	CodiEsp (Spanish)
[10]	*2021 – CLEF eHealth Shared Evaluation: SpRadIE* [Suominen et al., 2021]	NER and NEL (findings, anatomical entities, locations, measures, abbreviations)	513 sonography reports	Pediatric hospital in Buenos Aires
[11]	*2017 – IULA Spanish Clinical Record Corpus* [Marimon et al., 2017]	NER (e.g., clinical findings, substances)	3,194 EHR sentences	Hospitals in Barcelona, Spain
[12]	*2018 – MERLOT (Medical Entity and Relation LIMSI Annotated Text corpus)* [Campillos et al., 2018]	NER, Attributes, and RE (e.g., disorders, drugs, disorder-drug relations)	500 EHRs	French healthcare institutions
[13]	*2018 – Chinese CVD Risk EMR Corpus* [Su et al., 2017]	NER (CVD risk factors), Temporal attributes, and Assertions	600 EHRs	Second Affiliated Hospital of Harbin Medical University
[14]	*2023 – Cardiovascular/Stroke Disease EMR Entity and Relation Corpus (CVDEMRC)* [Chang et al., 2023]	NER and RE (CAD and treatments)	7,691 entities; 11,186 relations	Tertiary hospital in Henan Province
[15]	*2023 – CARDIO:DE* [Richter-Pechanski et al., 2023]	NER (Drugs), Attributes (e.g., reasons)	500 clinical letters	Heidelberg University Hospital (Germany)

Table 2.3: Summary of non-English clinical NLP corpora by annotation focus, size, and source institution.

Corpus ID	Models	Metrics (F1 unless stated)
[1]	UMLS MetaMap, NegEx [Meystre and Haug, 2006]	82%
[2]	cTAKES [Savova et al., 2010]	72% (Exact), 82% (Relaxed)
[3]	CRF + SVM + rules [Uzuner et al., 2010]	75% (NER meds), 53% (duration), 46% (reason)
[4]	CRF (NER/RE), SVM (assertion) [Uzuner et al., 2011]	85% (NER), 94% (RE/assertion)
[5]	I2E interface + lexicons + SVM + rules [Stubbs et al., 2015] BERT + Character-BERT (stacked embeddings) [Houssein et al., 2023]	83% (CAD indicators), 93% (risk factors) 94% (risk factors)
[6]	BiLSTM-CRF [Henry et al., 2020] TransformerCRF [Belkadi et al., 2023]	94% (NER), 96% (RE) 97% (NER)
[7]	MetaMap + cTAKES + CRF + SSVM [Tang et al., 2013; Pradhan et al., 2015] CRF + SSVM + MetaMap [Zhang et al., 2014] CRF + SVM + Dictionary (UMLS); SVM (template filling) [Elhadad et al., 2015]	75% (NER), 59% (NEL accuracy) 81% (NER), 74% (NEL accuracy) 75% (NER/NEL), 89% (template accuracy)
[8]	GPT-3 (1-Shot) [Agrawal et al., 2022]	92% (NER)
[10]	BERT + dictionary [Suárez-Paniagua et al., 2021]	69% (relaxed NER), 65% (exact NER)
[13]	BiLSTM-CRF [Su et al., 2017]	Not reported
[14]	JREwBART [Guo et al., 2024]	64% (RE)

Table 2.4: Overview of clinical NLP models with selected performance metrics considered relevant for this study. The corpus ID refers to the IDs of the corpora listed in Tables 2.2 and 2.3 used in the implementation.

2.5 Evaluation Metrics

The evaluation of NLP systems was performed by applying the trained models to a gold standard test set, manually curated or annotated. Performance was measured using precision, recall, and F1-score, defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad \text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Where

Outcome	Description
True Positive (TP)	Correct entity/code or relation triplet
False Positive (FP)	Incorrect entity/code or wrong relation triplet
False Negative (FN)	Missed entity/code or missed relation triplet

These metrics evaluate the model's ability to identify relevant entities or relations. Precision measures the accuracy of positive predictions, recall assesses the model's ability to capture all relevant positives, and the F1-score balances both precision and recall.

Chapter 3

LLaMIC Pipeline: LLM and Lexicon-Based Integration

Clinical notes from EHRs contains a high density and diversity of clinical entities, often exhibiting complex and implicit relationships, as mentioned in Section 2.1.1. Such characteristics present significant challenges for NLP systems, particularly when accurately identifying entities, managing overlapping spans, and handling the presence of multiple entity-relation triplets within a single sentence. To address these challenges, we propose the LLaMIC pipeline — a modular architecture based on LLaMA, a moderately sized open-source LLM, designed primarily for the MIMIC-IV clinical dataset, but adaptable to other clinical notes corpora. The pipeline consists of two main subtasks: entity detection, which comprises NER and NEL, and RE. Specifically, the entity detection component focuses on identifying cardiovascular and cerebrovascular diseases (CCD) with corresponding ICD-10 linking, as well as therapeutic drugs (CCDt) linked to MeSH terms. The full implementation of the pipeline is publicly available at: https://github.com/lasigeBioTM/LLAMIC.

3.1 Entity Detection Module

The entity detection module consists of three sequential steps—entity recognition, linking, and a review—inspired by Su et al. [2025], as shown in Figure 3.1. An example of the expected CSV input format appears in the box below:

idx: 22595853
document: HCV cirrhosis c/b ascites, hiv on ART, h/o IVDU, COPD, bipolar, PTSD, presented from OSH ED with worsening abd distension over past week and confusion. # Ascites - p/w worsening abd distension and discomfort for last week...
entities: {entity: CAD, start: 1303, end: 1306, icd: I25}

In practice, this module takes free-text clinical notes as input and outputs an annotated version of the text, including recognized entities, their character offsets, and corresponding standardized terminology — such as ICD-10 for CCDs and MeSH for CCDts. Each clinical note is provided as a row in a tabular structure containing a unique document identifier and the full raw medical text. When operating in EVAL mode, which evaluates predictions using precision, recall, and F1 metrics, the input must also include an additional column, entities which contains a list of ground-truth entities. This allows for the calculation of evaluation metrics such as precision, recall, and F1-score. The configuration parameters governing this pipeline are summarized in the box below (see Figure 3.2).

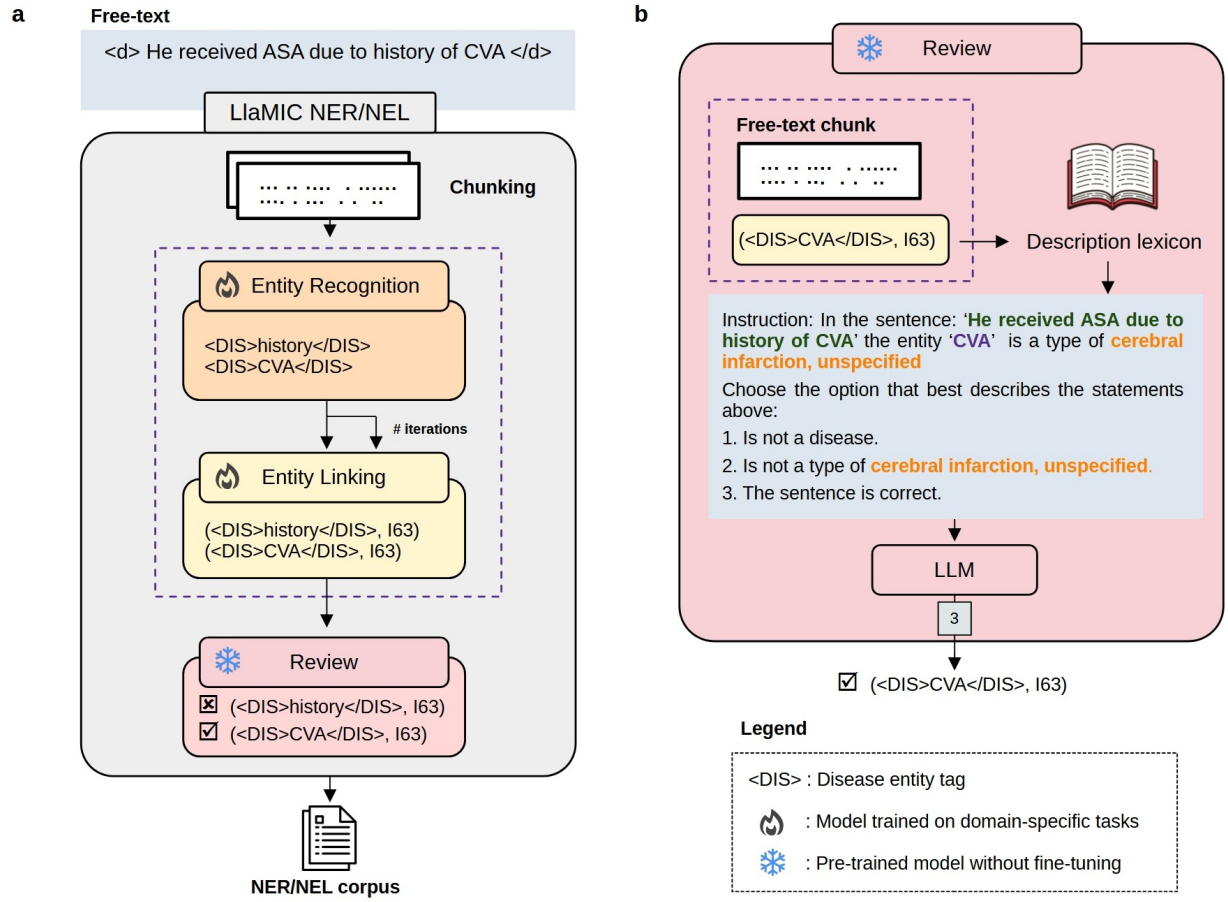

Figure 3.1: Architecture of the LLaMIC module: a) complete pipeline for entity recognition and linking, focused on disease entities; b) detailed view of the Review stage, responsible for filtering false positives from the entity recognition and linking steps.

Given the considerable length and structural variability of clinical texts, the input notes are first partitioned into manageable segments based on character length, controlled by the parameter –window_size

(See Figure 3.2), using the textwrap.wrap function. This segmentation is necessary to accommodate the context window limitations of LLMs. The window size should be selected to balance two factors: (i) preserving sufficient context to ensure accurate entity detection, and (ii) minimizing the risk of hallucinations or incoherent predictions that may occur when the input exceeds the model's effective context length. While this strategy enables the processing of long documents, it is important to note that some contextual loss at segment boundaries may still occur.

`--model_name_or_path_ner`: Path to the NER model (required).
`--model_name_or_path_nel`: Path to the NEL model (required).
`--model_name_or_path_review`: Path to the review/classifier model (required).
`--input_file`: Path to the input JSON file (required).
`--lexicon_path`: Path to the domain-specific lexicon (required).
`--output_dir`: Directory where output will be saved (required).
`--debug`: Enable debug mode for detailed logging (flag).
`--n_iterations`: Number of iterations for lexicon enhancement (default: 1).
`--window_size`: Sliding window size for long document splitting (default: 2636).
`--max_input_tokens`: Maximum input tokens for the model (default: 1024).
`--run_mode`: Execution mode: prediction or evaluation; choices: `PREDICT`, `EVAL` (default: `PREDICT`).
`--save_annotations`: Whether to save annotated outputs; choices: `yes`, `no` (default: `no`).
`--save_chunk_size`: Chunk size for intermediate saves (default: 200).

Figure 3.2: Configuration parameters used by the entity recognition and linking module in the LLaMIC pipeline.

In the first module, the model `--model_name_or_path_ner` performs named entity recognition by generating a flat, linearized list of entities. Formally, given an input chunk x, the model seeks to maximize the conditional probability $p(y \mid C, x)$, where y is the sequence of entity mentions and C is the prompting context, as shown in Appendix A.1.1 for CCDs and Appendix A.1.2 for CCDts. Each predicted entity must correspond to a contiguous substring of x, ensuring all entities are grounded in the input text. To improve recall the model can perform multiple inference passes over the same chunk. This is controlled by a configurable parameter `--n_iterations`. In each pass, entities identified so far are removed and replaced with a blank token (e.g., '___"), the same used in MIMIC de-identification, allowing attention to redistribute over the remaining content. This iterative mechanism offers a trade-off between computational cost and precision versus entity coverage in scenarios with high entity density and lengthy clinical notes. Additionally, a rule-based step ensures that all exact matches of each identified term throughout the note are also annotated. The second module focuses on entity linking, where the model `--model_name_or_path_nel` generates a sequence of $(\text{entity}, \text{ID})$ pairs. Each entity corresponds to a mention extracted in the previous step, and the ID refers to a concept in a standardized terminology. The linking process iterates once through the list of detected mentions, assigning each

one to the most appropriate terminology concept based on the surrounding context. If the list exceeds 12 entities, it is partitioned into batches of up to 12 for processing. The prompt used is the one shown in Appendix A.1.3 for CCDs and Appendix A.1.4 for CCDts. In the final step, review phase, the model `--model_name_or_path_review` reassesses both the accuracy of entity recognition and the correctness of entity linking of each $(\text{entity}, \text{ID})$ pair by revisiting the original text chunk x in combination with a textual description d corresponding to the proposed ID (Figure 3.1b). The model processes not only each unique entity but also evaluates each individual mention when the same entity appears multiple times within the text. The textual description is retrieved from a reference lexicon, `--lexicon_path`, which consists of a dictionary mapping each identifier to its official description. For example, the ICD-10 code `I61` may be associated with the description *"Cerebral infarction, unspecified"* based on ICD-10 terminology. To address potential false positives, the review phase acts as a filtering mechanism, removing incorrect predictions and improving overall precision. This phase is implemented as a multiple-choice task, as described in Appendix A.1.5, where the model outputs a categorical label $R \in \{1, 2, 3\}$ where 1 indicates an invalid entity mention, a 2 reflects a valid entity that is incorrectly linked, and a 3 confirms both a valid mention and a correct link to the identifier. The entire pipeline described operates sequentially, processing one clinical note at a time.

3.2 Relation Extraction Module

The relation extraction task aims to identify semantic relationships between candidate entity pairs within clinical notes. Due to the sparsity of clinically relevant relations in some clinical notes, and the high density of entity mentions, exhaustive all-pairs classification often leads to a large number of false positives, thereby reducing overall precision. To mitigate this, LLaMIC leverages the contextual understanding and generative capabilities of LLMs to filter unlikely pairs during the Pair Generator phase. Subsequently, the Relation Classifier focuses the labeling task on this filtered subset, improving accuracy and efficiency. An overview of the pipeline is presented in Figure 3.3.

Input to the model consists of sentences with explicitly annotated entities using indexed markup tags, for example, `<disease1>...</disease1>` and `<drug2>...</drug2>`, where each entity is uniquely identified based on its order of appearance. When operating in EVAL mode, similarly to the Entity Detection module, the pipeline expects a column containing the ground-truth relations, as illustrated below:

idx: `29477116_3`

document: `<disease1>CAD</disease1>: A ___ year old man with suspected <disease2>MI</disease2>, continue <drug1>aspirin</drug1> and <drug2>atorvastatin</drug2>...`

relations: `["<disease1>CAD</disease1>", "indicated", "<drug1>aspirin</drug1>"]`

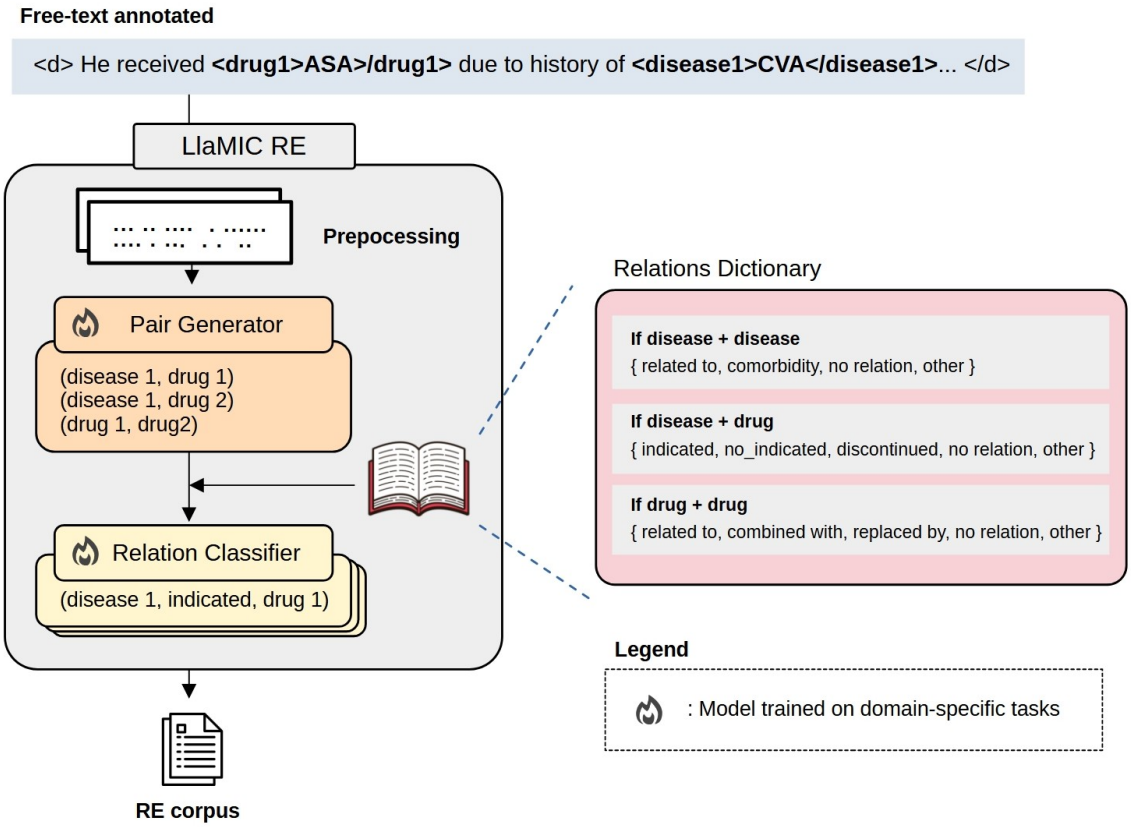

Figure 3.3: Architecture of the LLAMIC module for relation extraction.

Prior to invoking the LLMs, the pipeline automatically generates three distinct variants of each sentence, each emphasizing a specific category of entity pairs: CCD–CCD, CCDt–CCDt, and CCD–CCDt. These variants are processed independently by the model to enable type-specific relation extraction. Due to module processing limitations, inputs exceeding 600 tokens are truncated and not processed in pair generation module, the model `--model_name_or_path_pg` enumerates all candidate pairs $p = (e_h, e_t)$ that potentially exhibit a relation (see Appendix A.1.7, A.1.6, and A.1.8). Given a set of entities $E = \{e_1, e_2, ..., e_n\}$ in sentence n, the candidate pairs $p \in L(n)$ are defined as:

$$L(n) = \begin{cases} \{p_1, p_2, ..., p_k\}, & \text{if } e_h, e_t \in E \\ \emptyset, & \text{otherwise} \end{cases}$$

If no model is provided for this phase, the pipeline adopts a automatic permutation approach, generating all possible permutations of entity pairs within the sentence. Subsequently, in relation classification, the model `--model_name_or_path_rc` processes each candidate pair individually to predict a single relation label r. The labeling function is formalized as:

$$R = f(n, (e_h, e_t)), \quad R \in \{\texttt{relation\_list}\}$$

where `relation_list` denotes a predefined set of possible semantic relations, determined according to the types of entities involved (e.g., diseases and drugs). The prompt used for this task is detailed in Appendix A.1.10. This set of relations can be adjusted depending on the specific domain or application scenario. The configuration parameters are listed in Table 3.4.

`--model_name_or_path_pg`: Path to the Pair Generator model (required).
`--model_name_or_path_rc`: Path to the Relation Classifier model (required).
`--input_file`: Path to the input JSON file (required).
`--output_dir`: Directory where output will be saved (required).
`--debug`: Enable debug mode for detailed logging (flag).
`--run_mode`: Execution mode: prediction or evaluation; choices: `PREDICT`, `EVAL` (default: `PREDICT`).
`--save_annotations`: Whether to save annotated outputs; choices: `yes`, `no` (default: `no`).
`--save_chunk_size`: Chunk size for intermediate saves (default: 200).

Figure 3.4: Configuration parameters used by the relation extraction module in the LLaMIC pipeline.

3.3 LLAMIC integrates with different LLMs

Although the LLaMIC pipeline was initially developed using LLaMA3.1 models, it was designed to be modular and compatible with other base models. As such, it supports any decoder-only causal language model available through the Hugging Face ecosystem or loaded locally. To use the pipeline with another model, the requirements are that the model must support `bitsandbytes` for 4-bit quantization and memory efficiency, be compatible with `AutoModelForCausalLM` from the Hugging Face Transformers library (e.g., Meta-LLaMA 3.1), have a corresponding tokenizer available and compatible with the model, and the environment should provide sufficient GPU memory to load and run the model efficiently.

Chapter 4

LLaMIC Implementation and Corpus Construction

Following the introduction of the LLaMIC pipeline, this chapter describes its implementation in the creation of an annotated corpus through two main phases: first, entity detection, targeting cardiovascular and cerebrovascular diseases (CCD) and therapeutic drugs (CCDt), which were further normalized to ICD-10 and MeSH, respectively; and second, relation extraction, capturing the relationships between these entities from the MIMIC-IV deidentified free-text clinical notes (DICNs). For each task, a corpus was constructed and divided into two complementary subsets. The first is a semi-automatic annotation subset—referred to as the supervised corpus—which was employed to train and validate the LLaMIC pipeline. The second subset consists of the remaining MIMIC-IV DICNs, which were automatically annotated using the trained pipeline, and is referred to as the unsupervised corpus.

4.1 MIMIC-IV Dataset

4.1.1 Dataset Access

The MIMIC-IV v3.1[1] and MIMIC-IV-Note v2.2[2] dataset used in this study was accessed through PhysioNet under the terms of PhysioNet's License 1.5.0 and Data Use Agreement 1.5.0, following completion of the required CITI training [Johnson et al., 2023; Goldberger et al., 2000]. It comprises 26 tables that contain detailed patient information from hospitalizations, including measurements, orders, diagnoses, procedures, treatments, and DICNs. These tables are linked through unique identifiers, typically indicated by the suffix `id` in column names. This work focuses specifically on 331,794 DICNs from 145,915 patients, stored in the `text` column of the `discharges.csv` table from MIMIC-IV-Note, each associated with a hospital admission identified by the `hadm_id`, as illustrated in the bottom part of Figure 4.2a, An example of a DICN section is provided in Figure 4.1. For the preprocessing step, some

[1] https://physionet.org/content/mimiciv/

[2] https://physionet.org/content/mimic-iv-note/2.2/

tables from MIMIC-IV v3.1 will also be used, which are described in more detail below.

4.2 Entity Detection and Annotation Framework

4.2.1 Definition of Entity Types

We define two types of biomedical entities for annotation: (i) CCDs, and (ii) CCDts. The guidelines for identifying and labeling each entity type are detailed below.

I. CCD Entities

Disease is defined as a deviation from normal physiological or psychological functioning, characterized by clinical, pathological, and epidemiological criteria [White, 2020]. This excludes natural biological processes and isolated symptoms. The definition of a disease entity is based primarily on the ICD and the section 'Diseases' (coded C) in the MeSH. In this work, we focus on eight prevalent CCDs, identified by the following ICD-10 codes: coronary heart disease (I20–I25), hemorrhagic stroke (I60–I62), cerebral infarction (I63), and unspecified stroke (I64) [CDC, 2024]. A complete description of each ICD code is provided in the Appendix A.1. Some examples of CCD entities and their respective ICD-10 codes include:

- The patient was placed on heparin drip for stroke prophylaxis (*stroke*, I63)
- The patient was determined to have multiple small acute infarcts in the left centrum semiovale (*multiple small acute infarcts*, I63)
- With prostate cancer, CAD w/multiple stents, recent *E. faecalis* UTI, presenting with hematuria (*CAD w/multiple stents*, I25)

II. CCDt Entities

A therapeutic drug is defined as a pharmacological agent that modulates physiological or pathological processes for therapeutic benefit. Drug entities correspond to concepts in Category D (Chemicals and Drugs) of the MeSH. This study focuses exclusively on CCDts, identified using the "Drug-to-Disease Mapping with ICD Identifiers" dataset from the Therapeutic Target Database (TTD). This dataset enables the retrieval of drugs and targets associated with specific diseases or ICD codes. Only drugs linked to the mentioned CCD ICD codes were retained. A complete list is provided in the Appendix A.2. Some examples of CCDt entities and their respective MeSH code include:

- She was given baby ASA, home metop. (*baby ASA*, D001241 // *metop*, D008790)

4.2.2 Corpus Selection and Preprocessing

The source corpus is derived from DICNs available in MIMIC-IV-Note, previously introduced in Section 4.1. These MIMIC DICNs present two primary challenges for automatic processing: (i) their ex-

```
Name: ___
Unit No: ___
Admission Date: ___
Discharge Date: ___
Date of Birth: ___
Sex: ___

History of Present Illness: The patient presented with complaints
of severe abdominal pain, bloating, and progressive weight loss over
the past month. Symptoms were associated with nausea and occasional
vomiting.

Past Medical History:
1. History of hypertension.
2. Patient denies any history of major surgeries.

Brief Hospital Course: The patient was admitted for evaluation of
chronic abdominal pain and ascites. On admission, vital signs were
stable with no evidence of hemodynamic compromise. Initial laboratory
workup revealed elevated liver enzymes and hyperbilirubinemia.
Imaging with right upper quadrant ultrasound demonstrated coarse and
nodular liver echotexture, consistent with advanced liver disease.
Throughout hospitalization, daily monitoring of liver function tests
and electrolytes was performed, with dose adjustments to medications
as needed.
# Ascites and Cirrhosis: Patient presented with chronic abdominal
pain and clinically significant ascites. Laboratory results showed
elevated liver enzymes and bilirubin levels.
# Hepatic Management: Supportive care was initiated during
hospitalization. Lactulose 15 mL was administered three times daily,
alongside diuretics to reduce fluid overload.

Discharge Medications:
1. Furosemide 40 mg PO DAILY
2. Lactulose 15 mL PO TID

Discharge Diagnosis:
Primary: cirrhosis with ascites
Secondary: hypertension, Type 2 diabetes
```

Figure 4.1: Example of a MIMIC-IV DICN section.

tended length, which may exceed the token limitations of LLAMIC models; and (ii) their frequent lack of explicit references to CCDs or CCDts, which can hinder supervised training. To address challenge (i), following the methodology proposed by Wang et al. [2024], we restricted the input to the "Brief Hospital Course" section of each DICN, which is a succinct summary of an entire hospital encounter (see Figure 4.1). These sections were extracted using regular expressions, and summaries with fewer than 40 characters were excluded to ensure sufficient content for reliable entity detection and future relation extraction, as depicted at the top of the Figure 4.2b.

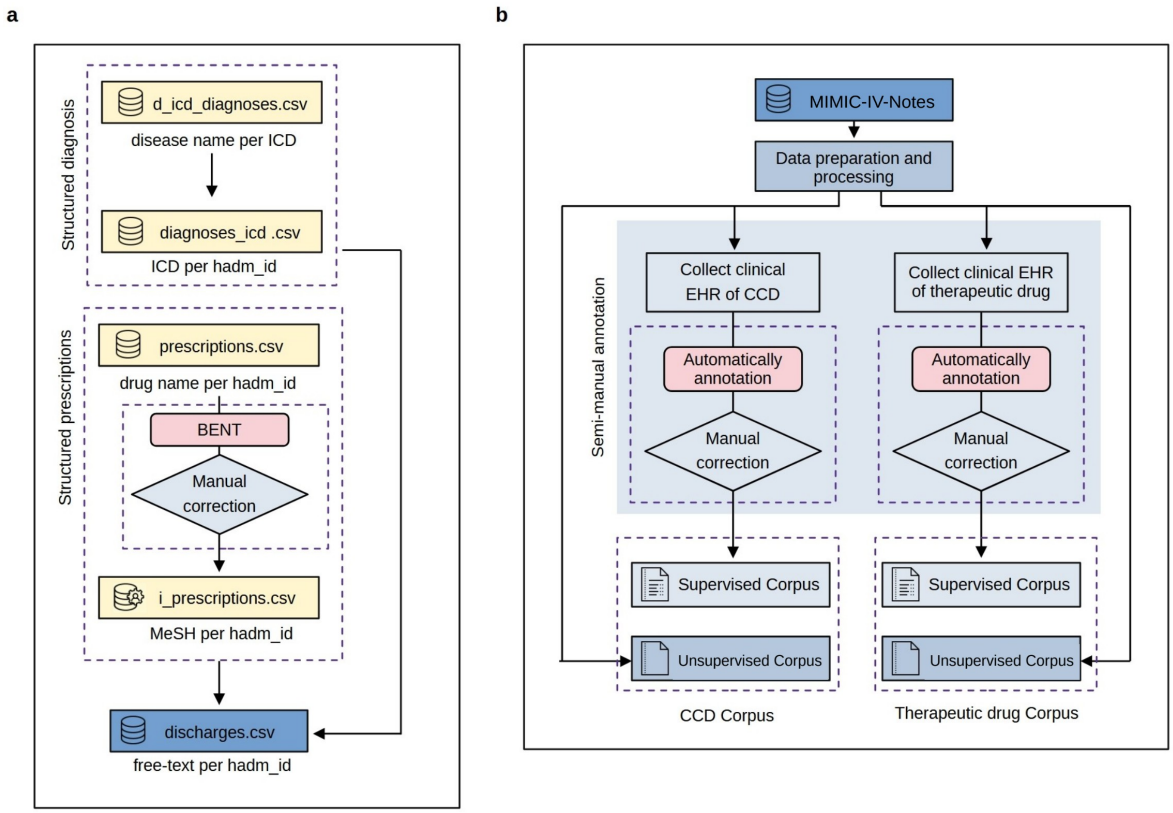

Figure 4.2: Global pipeline for the creation of the LLaMIC DICN Entity Detection Corpus. (a) Extraction of ICD-10 diagnoses and MeSH-based drug prescriptions from structured tables. While diagnosis codes were directly obtained from the diagnosis table, drug mappings required semi-automatic conversion of medication names to MeSH identifiers. (b) Annotation process for creating the supervised and unsupervised LLaMIC corpus, with CCDs on the left and CCDts on the right.

In addition, line breaks, double spaces, and extra spacing were removed to produce a more cohesive and continuous input text. Challenge (ii) is particularly critical when constructing supervised corpus, as the infrequent mention of target entities can introduce class imbalance and adversely affect model performance. To mitigate this, we independently balanced the supervised dataset by selecting an equal

number of DICNs with and without mentions of CCDs, and, in parallel, an equal number with and without mentions of CCDts. This approach may introduce selection bias, because notes that do not explicitly mention an entity might still imply its presence. These two subsets were constructed separately from the entire corpus and were used independently for training and evaluation, as illustrated at the beginning of the ”Semi-manual Annotation” section in the middle of Figure 4.2b. Structured diagnosis records used for the CCD balancing were retrieved from the MIMIC-IV `diagnoses_icd.csv` table, which contains ICD-9 and ICD-10 codes assigned by clinical staff after reviewing finalized patient notes. It is important to note that these codes reflect all diagnoses billed during hospital stays, but do not indicate whether the condition is explicitly mentioned in the corresponding DICN. For the CCDt balancing, medication records were obtained from the `prescriptions.csv` table, which logs all drugs prescribed during each admission. As this file lacks MeSH identifiers, we applied the BENT model to perform NEL to map drug names to their respective MeSH concepts. A manual correction step was performed on BENT's outputs to ensure comprehensive coverage of the target medications and accurate MeSH mappings. Figure 4.2a illustrates this pipeline for extracting such information and its relationship with the DICN records in the `discharges.csv` table. This preprocessing step resulted in two independent semi-automatically annotated subsets: one for CCD entities and another for CCDt entities. These subsets, which will later be semi-manually annotated to obtain their final supervised versions, were further split into training (80%), validation (10%), and test (10%) partitions. The remaining portion of the corpus—consisting of DICNs that were preprocessed but not balanced —was retained separately. This portion serves as an unsupervised, unbalanced dataset to be used for downstream inference once the models have been trained and validated on the supervised data.

4.2.3 Baseline Model

To provide a comparative analysis of the proposed LLaMIC pipeline on MIMIC DICNs, we selected BENT as a baseline model. BENT integrates entity recognition and linking by using PubMedBERT for NER and a graph-based NEL approach leveraging the Personalized PageRank algorithm combined with Information Content metrics for candidate disambiguation [Ruas and Couto, 2022]. This model has demonstrated robust performance in biomedical entity linking, achieving, for example, an F1-score of 0.9401 on the BC5CDR-Disease dataset using the MEDIC ontology. In the context of our work, for disease NER, we used the model trained on disease-specific corpora—referred to as 'disease'—and the 'do' component of BENT relies on the DO as its knowledge base for NEL. To align with our objective of assigning ICD-10 codes as entity identifiers, we extracted corresponding ICD-9 and ICD-10 codes from the DO URIs using the `hasDbXref` property, accessed through the `owlready2` ontology interface. For drug NER, we employed a model fine-tuned on chemical entities— 'chemical'—and utilized MeSH identifiers from the CTD ontology for NEL. To ensure a consistent comparison with our proposed LLaMIC model—which is inherently constrained to target-specific entities—we introduce an extension of the original BENT model, referred to as BENT-Rule. This adaptation applies a rule-based filtering

step post-entity linking, removing all linked entities whose ICD-10 or MeSH codes do not correspond to a predefined set of target entities.

4.2.4 Semi-Automatic Annotation to Build a Supervised Corpus

The annotation process was carried out for the creation of the supervised corpus through a two-stage semi-automatic pipeline: initial automatic pre-annotation of entities (represented by rounded rectangles labeled "Automatic annotation" in Figure 4.2b), followed by manual correction (depicted as a diamond shape labeled "Manual correction" in the same figure).

Automatic pre-annotation: Two models were applied in sequence to perform automatic pre-annotation. First, the BENT-Rule model was used to identify entity mentions. Subsequently, an LLaMIC model based on LLaMA without task-specific fine-tuning was used on the masked version of the input, in which the entities identified by BENT-Rule were temporarily removed. To maximize entity coverage, LLaMIC was executed over three interactions in entity recognition stage (parameters detailed in Section 3.1).

Manual correction: From the pre-annotated corpus, a lexicon was constructed that encompasses all extracted entities found in the DICNs. These entities were grouped based on their textual similarity using TF-IDF vectorization, cosine similarity, and DBSCAN clustering. Manual validation and refinement of the resulting clusters were performed by the first author—a master's student with a background in Human Biology. The corrections followed two main guidelines: first, adhere to the entity definitions outlined in the previous section 4.2.1; second, prioritize the preservation of the highest possible specificity of disease mentions in the notes. Since the entity fragments lacked a surrounding clinical context, ambiguous or unclear mentions were resolved by consulting their original context within the DICN. The manual correction process was inherently constrained by the outputs of the two models used during the automatic pre-annotation phase. To further reduce missed entities in the annotated subset, DICNs with no annotated entities and having structured data indicating their presence were individually reviewed. Importantly, manual correction with preserved offsets was carried out only for the test subset, whereas for the training and validation sets such correction was unnecessary, since the fine-tuning of the LLaMIC modules for Entity Recognition and Entity Linking does not require precise entity localization. This design choice enabled the scalability of the manual correction process. No formal evaluation of the annotation quality was conducted, as the corpus was solely manually corrected by a single researcher. Estimating inter-annotator agreement would require multiple independent annotators to quantify consensus and ensure reproducibility.

4.2.5 Fine-Tuning and Deployment of LLaMIC

Following the generation of the supervised corpus for CCD-related diseases and CCDts, we created a training, validation, and test sets, which were used to fine-tune four LLaMIC models based on the LLaMA 3.1 8B language model. Two models were optimized for NER—one for CCDs and the other for CCDts—and the remaining two for NEL tasks targeting the same entity categories. Fine-tuning was performed

using a learning rate of 5×10^{-5}, with a training batch size of 4 and an evaluation batch size of 2. Training was carried out for 10 epochs using the AdamW optimizer with a weight decay of 0.01. To adapt the model efficiently, we employed LoRa with, with the following configuration: a rank (r) of 8, a LoRA alpha value of 32, and a dropout rate of 0.05. Both models are included in the *Supplementary Documents* and will be made publicly available after the thesis defense. As mentioned in Section 4.2.4, we opted not to train the third model for the Review phase. This model is designed to verify each entity mention within the text, which would require entity offsets in both the training and validation sets. To ensure scalability, we instead relied on the knowledge-grounded baseline capabilities of LLaMA for this task. All LLaMIC models used the default temperature (1.0) with deterministic decoding (do_sample=False) for reproducible generation.

4.2.6 Evaluation Framework of the LLaMIC Pipeline

For NER evaluation, the supervised test corpus was used as ground truth. Evaluating NER introduces challenges due to the generative nature of LLMs models, which may produce variations in entity surface forms. To address this, we adopt both strict matching—where the predicted entity must exactly match the reference—and lenient matching, which allows for semantically equivalent expressions by considering a predicted entity correct if it is either a prefix or suffix of a true entity. The evaluation metrics employed were precision, recall, and F1-score, which are discussed in detail in Section 2.5. The evaluation of NEL was conducted at two levels: (i) NER-tabular — using structured data from MIMIC-IV. This source provides ICD-10/ICD-9 codes (for CCD-related diagnoses) and MeSH identifiers (from prescription records for CCDts) per DICN, but without entity offsets. Consequently, evaluation is limited to verifying whether the predicted set of standardized identifiers matches the reference set of target concepts; and (ii) NER-supervised — the supervised subset, where entity-level annotations are available, allowing a standard evaluation of whether each correctly entity mention is correctly linked to its corresponding ICD or MeSH identifier. Evaluation metrics include precision, recall, and F1-score.

4.2.7 Results and Discussion

4.2.7.1 Preprocessing of MIMIC-IV-Notes

After preprocessing (Section 4.2.2), a total of 310,393 DICNs were retained from the original 331,793 available in MIMIC-IV. These correspond to notes from which a valid “Brief Hospital Course” section could be extracted and which met the minimum content threshold. On average, each DICN contained approximately 2,493 characters and 390 words. Following the label of disease and medication mentions using structured data sources (Figure 4.2a), we observed a notable asymmetry in the presence of target clinical concepts, as summarized in Table 4.1. Specifically, only approximately 25% of DICN were associated with at least one CCD diagnosis, whereas over 80% were linked to CCDt. This discrepancy in coverage may be partially attributed to the broader therapeutic scope of certain medications. Many drugs

used in the treatment of CCDs, such as aspirin, are also commonly prescribed for a wide range of other conditions. In contrast, CCD diagnoses—being more specific by nature—occur less frequently within the corpus, especially considering that MIMIC-IV comprises general ICU data and is not restricted to cardiovascular and cerebrovascular hospitalizations.

Condition	Discharges (number)	Percentage in total (%)
CCDts	252,945	81.49
CCDs	77,677	25.03
Both CCDs and CCDts	71,281	22.96

Table 4.1: Distribution of DICN mentioning target drugs, target diseases, or both.

The frequency of CCDs varies widely on structured data, as shown in Figure A.1a, with chronic ischemic heart disease (I25) being the most frequently recorded diagnosis, appearing in 139,956 hospital admissions. The full list of CCDts is provided in the Appendix A.2.

4.2.7.2 Supervised Corpus Statistics

The previously described data was divided into an unsupervised and supervised corpus The final dimensions of the supervised corpus—obtained through pre-annotation with BENT-Rule and the LLaMIC framework (using the LLaMA 3.1 8B model), followed by a manual correction phase—and the unsupervised corpus are presented in Table 4.2. During automatic pre-annotation, the BENT-Rule model—being a PubMedBERT variant fine-tuned on well-established and carefully curated biomedical datasets—ensures terminological precision and alignment with standardized biomedical vocabularies, although it lacks the flexibility to specify target diseases. By contrast, LLaMIC leverages domain knowledge to capture more specific entities, enabling focus on CCD diseases and CCDts (which are the scope of our work), as well as semantically complete mentions (e.g., identifying *multiple small acute infarcts* instead of simply *infarcts*), including abbreviated, misspelled, or loosely expressed mentions. To further enhance annotation quality, a manual correction strategy—applied first in a general pass (using clusters) and then in a more specific pass—was implemented. This strategy—allowing scalable processing of large datasets—is particularly important in clinical corpora, where mention variability, especially for CCD entities, is high. Ensuring adequate frequency and diversity of mention types is essential for effective model training and evaluation. Nevertheless, this approach has inherent limitations, such as the potential omission of valid mentions or the introduction of annotation errors due to contextual ambiguity.

Database	Train Set	Validation Set	Test Set	Unsupervised Corpus
CCDs	8,059	1,727	1,728	298,879
CCDts	4,676	584	584	304,546

Table 4.2: Supervised and unsupervised database statistics.

After completing our semi-automatic annotation approach to construct the supervised corpus, a total of 13,521 annotated CCD entities were obtained. The most frequent ICD-10 codes are I25 (36.93%), I21 (25.15%), and I63 (18.50%), while I23 appears least frequently (0.03%) (Table 4.3). At the lexical level, the most common surface forms include “CAD” (19.89%), “NSTEMI” (5.32%), and “stroke” (4.38%) (Supp. Table A.4).

Code	Description	Count	%
ICD-10 Codes (CCDs)			
I25	Chronic ischemic heart disease	4993	36.93
I21	Acute myocardial infarction	3400	25.15
I63	Cerebral infarction	2502	18.50
I20	Angina pectoris	1175	8.69
I61	Nontraumatic intracerebral hemorrhage	896	6.63
I24	Other acute ischemic heart diseases	399	2.95
I60	Nontraumatic subarachnoid hemorrhage	130	0.96
I22	Subsequent myocardial infarction	15	0.11
I62	Other and unspecified nontraumatic intracranial hemorrhage	7	0.05
I23	Certain current complications following STEMI and NSTEMI	4	0.03
Total (ICD)	—	13521	-
MeSH Identifiers (CCDts)			
D001241	Aspirin	1650	51.19
D006493	Heparin	1235	38.32
D005996	Nitroglycerin	153	4.75
D009543	Nifedipine	93	2.89
D017984	Enoxaparin	92	2.85
Total (MeSH)	—	3223	-

Table 4.3: Distribution of annotations in the supervised corpus by ICD-10 codes (CCDs) and MeSH identifiers (CCDts), including counts and percentages.

As many CCDts from the supervised corpus occur only infrequently mentioned in clinical notes, we focused the analysis on the five most frequently mentioned drugs. In total, 3,223 mentions were annotated. The most prevalent MeSH identifiers were D001241 (Aspirin, 51.19%) and D006493 (Heparin, 38.32%), followed by D005996 (Nitroglycerin, 4.75%), D009543 (Nifedipine, 2.89%), and D017984 (Enoxaparin, 2.85%) (Table 4.3). A second balancing of the CCDt supervised corpus was performed, taking into account only the five selected drugs, which explains the smaller size of the supervised corpus compared to that of the CCDts. These corpora are included in the *Supplementary Documents* and will be made publicly available after the thesis defense.

4.2.7.3 Automatic Entity Annotation

We trained LLaMA for the Entity Recognition and Entity Linking modules of the LLaMIC pipeline using the supervised corpus introduced earlier (Table 4.9). The LLaMA for Review module of LLaMIC was not trained. The performance of the BENT-Rule model and the LLaMIC framework—including variants without fine-tuning (LM-NoFT), without review step (LM-NoRV), and with progressively increased numbers of interaction in the Entity Recognition task (LM-Int1 to LM-Int3)—on the test subset are reported for CCD entities in Table 4.4, and for therapeutic drug entities in Table 4.5. For the CCDs, the LLaMIC-Int1 model demonstrates consistent improvements in both F1 score and precision compared to the BENT-Rule baseline across NER and NEL tasks. In the NER task, LLaMIC-Int1 outperforms BENT-Rule in precision by 27% (strict evaluation) and 37% (lenient evaluation). The ablation study shows that fine-tuning significantly enhances performance across all tasks. The optional review step also provides additional gains in precision, although these are not statistically significant. Furthermore, varying the number of interaction steps indicates that increasing the number of interactions improves recall (reaching up to 96% in the lenient setting with three interactions), but at the cost of precision. This is likely due to multiple rounds of entity extraction increasing the risk of false positives.

For the NEL task, LLaMIC continues to outperform BENT-Rule, with gains in precision of 18% under tabular evaluation and 21% under supervised evaluation. Similar to the NER task, fine-tuning and the review step were beneficial. However, while entity linking improved recall, it reduced precision, likely because the higher number of extracted entities in the NER step increased the probability of linking errors.

In contrast, for the CCDts, the BENT-Rule model consistently outperforms all LLaMIC configurations, although the differences are not statistically significant. Specifically, in the NER task, BENT-Rule surpasses LLaMIC-Int3 in precision by 4% (strict) and 2% (lenient). In the NEL task, the gains are also not statistically significant, with 0.2% and 4% higher precision for tabular and supervised evaluation, respectively. In this case, increasing the number of interaction steps in the NER phase was beneficial for both precision and recall. This may be explained by LLaMIC's tendency to under-annotate entities. The issue could stem from the supervised dataset, which contains highly sparse entities and many training cases without any entities (despite attempts to balance this using structured data from prescription.csv).

I. Entity Boundaries

The BENT-Rule model demonstrated limitations in precisely identifying the boundaries of CCDs entities, often resulting in incorrect or partial segmentation of clinical concepts. These challenges can be attributed to the inherent complexity of clinical terminology and the variability with which CCDs are expressed in DICNs. As noted by Stubbs et al. [2015], mentions of CCDs often appear in diverse and context-dependent forms—ranging from simple expressions such as "diffuse SAH" to more complex phrases like "bilateral ___ and ___ cerebellar infarcts." This variability makes it particularly difficult for token level systems to accurately detect and delineate CCD entities.

Metric	BENT-Rule	LM-NoFT	LM-NoRV	LM-Int1	LM-Int2	LM-Int3
NER-strict						
P	0.565	0.435	0.830	**0.831**	0.817	0.808
R	0.551	0.569	0.837	0.837	0.867	**0.869**
F1	0.554	0.463	0.824	0.824	**0.829**	0.823
NER-lenient						
P	0.571	0.445	0.883	**0.887**	0.873	0.865
R	0.554	0.590	0.858	0.905	0.952	**0.960**
F1	0.558	0.475	0.857	0.882	**0.892**	0.889
NEL-tabular						
P	0.515	0.328	0.683	**0.690**	0.679	0.675
R	0.518	0.446	0.729	0.732	**0.741**	**0.741**
F1	0.515	0.352	0.692	**0.698**	0.694	0.691
NEL-supervised						
P	0.610	0.419	0.820	**0.825**	0.8064	0.799
R	0.615	0.522	0.801	0.831	0.8547	**0.859**
F1	0.609	0.443	0.802	**0.819**	0.8182	0.815

LLaMIC-NoFT = without fine-tuning, **LLaMIC-NoRV** = without Review phase; **LLaMIC-Intx** = with x iterations

Table 4.4: Performance of BENT-Rule and LLaMIC models (LLaMA3.1-8B-Instruct) on NER and NEL tasks for CCD entities.

Metric	BENT-Rule	LM-NoFT	LM-NoRV	LM-Int1	LM-Int2	LM-Int3
NER-strict						
P	**0.913**	0.263	0.851	0.860	0.863	0.878
R	**0.898**	0.373	0.824	0.841	0.862	0.882
F1	**0.902**	0.279	0.832	0.845	0.858	0.875
NER-lenient						
P	**0.913**	0.263	0.854	0.878	0.882	0.897
R	0.899	0.378	0.825	0.857	0.881	**0.900**
F1	**0.902**	0.280	0.834	0.861	0.877	0.894
NEL-tabular						
P	0.608	0.401	0.596	0.596	0.600	**0.610**
R	**0.560**	0.401	0.547	0.547	0.552	0.558
F1	**0.574**	0.401	0.562	0.562	0.566	0.573
NEL-supervised						
P	**0.920**	0.571	0.857	0.861	0.863	0.880
R	**0.908**	0.571	0.837	0.841	0.862	0.882
F1	**0.912**	0.571	0.843	0.846	0.858	0.876

LLaMIC-NoFT = without fine-tuning, **LLaMIC-NoRV** = without Review phase; **LLaMIC-Intx** = with x iterations

Table 4.5: Performance of BENT-Rule and LLaMIC models (LLaMA3.1-8B-Instruct) on NER and NEL tasks for CCDt entities.

This boundary recognition challenge is more robustly addressed by the NER module of LLaMIC model, which demonstrates superior performance in recognizing entity boundaries. For example, consider the following clinical note segment, where LLaMIC provides a more complete CCD entity:

Document segment:
`ECG showed signs consistent with demand ischemia, likely secondary to anemia and tachycardia. Troponin levels were mildly elevated.`
BENT-Rule output: *ischemia*
LLaMIC output: *demand ischemia*

Furthermore, in the previously mentioned example of "bilateral ___ and ___ cerebellar infarcts," even with missing tokens, the LLaMIC model leverages contextual clues to accurately infer and capture the full disease entity. Due to the high specificity of disease annotations, this also facilitates a more precise NEL, especially considering that ICD codes provide a highly granular disease classification. Notably, the LLaMIC model demonstrates high recall from the initial extraction, with subsequent iterations—LM-Int2 and LM-Int3—yielding only marginal improvements in entity identification.

In contrast, this limitation is less pronounced in the identification of CCDt. These results suggest that the lexical consistency and domain specificity of CCDt mentions in clinical texts favor a more BERT-like model, such as BENT-Rule, which operates at the token level. On the other hand, LLaMIC models, although showing competitive performance, do not surpass BENT-Rule's precision in this domain. This may be due to previously discussed sample imbalance in the supervised corpus. Nevertheless, iterative variants of LLaMIC (e.g., LM-Int3) achieve precision values that come very close to the baseline.

II. Entity Linking

Regarding CCDs, the NEL module of the LLaMIC model also stood out in the entity linking task. The BENT-Rule model exhibited low specificity in assigning ICD-10 codes, whereas LLaMIC's broader coverage in the NER task enabled it to retrieve more specific ICD codes, reflecting improvements in both recall and precision. However, LLaMIC models demonstrated a greater susceptibility to false positives during entity linking, especially as the number of interaction steps increased (LM-Int2, LM-Int3; see Table 4.4). While additional interaction rounds improved recall—primarily by increasing the number of entities identified during NER—they also introduced spurious mappings, often linking non-disease mentions to inappropriate or unrelated ICD codes. As illustrated in the example below, the term *"cardiac cath"* was incorrectly linked to ICD code I25, despite not representing a pathological condition. This behavior suggests that the model, influenced by semantic proximity in its pretraining data, tends to associate related procedures or anatomical terms with adjacent disease codes.

Document segment:
`The patient was admitted for a cardiac cath to assess cardiac function and guide treatment decisions. The procedure was completed without complications, and the patient tolerated it well.`
LLaMIC: NER output: *cardiac cath*
LLaMIC: NEL output: *I25*

Additionally, in NEL of CDDt, without fine-tuning, the LLaMIC model—relying solely on the general knowledge acquired during LLaMA's pretraining—was unable to effectively perform entity linking to MeSH identifiers for CCDts. However, after fine-tuning on a small annotated training set, the model's NEL performance improved substantially, yielding results nearly comparable to those of the BENT-Rule system, with only a marginal difference of approximately 4% in precision.

III. Review Module: LLaMIC

The LLaMIC architecture incorporates a lexicon-based review mechanism as a validation step. Its impact on precision is limited, with gains of 0.4% and 0.1% in strict and lenient entity recognition, and 0.7% and 0.5% in tabular and supervised linking evaluations, respectively. These results indicate that the review mechanism contributed primarily to the entity linking stage. The Review Module proved effective in rejecting non-pathological mentions- such as clinical findings or procedures- that were erroneously mapped to disease concepts. For example, as show below, the entity "`ST elevations`" was initially linked to `I21` (Acute myocardial infarction), although it represents an electrocardiographic finding:

Document segment:
`ECG revealed ST elevations in the anterior leads, prompting immediate evaluation for myocardial injury.`
LLaMIC: NER output: *ST elevations*
LLaMIC: NEL output: *I21*
Review output: *Is not a disease*

In addition, the module corrected linking errors. For instance, "`critical AS`" (aortic stenosis) was incorrectly mapped to `I61` (Intracerebral hemorrhage). The Review Module identified the mismatch and invalidated the mapping.

Document segment:
`The patient presented with signs of hemodynamic instability and was diagnosed with critical AS requiring urgent intervention.`
LLaMIC: NER output: *critical AS*
LLaMIC: NEL output: *I61*
Review output: *Is not a type of Nontraumatic intracerebral hemorrhage*

IV. Unsupervised Corpus Annotation

Based on the evaluation results, LLaMIC-Int1, which demonstrated the best performance for CCD entities, was used to annotate CCDs, while BENT-Rule, the most effective model for CCDt entities, was employed for their annotation in the unsupervised corpora. These corpora are included in the *Supplementary Documents*.

V. Computational Performance

All experiments were performed on an NVIDIA Tesla T4 GPU (16 GB VRAM). The integration of the Review Module introduced a slight increase in processing time: LLaMIC required on average 14.3 seconds per DICN with the Review Module enabled, compared to 12.8 seconds without it. These values, however, may vary depending on the length of the clinical notes, the density of CCDs and CCDts, and the specific GPU hardware.

4.3 Relation Annotation

4.3.1 Definition of Binary Entity relationships

We define three types of binary relations: (i) between two CCD entities, (ii) between two CCDt entities, and (iii) between a CCD entity and a CCDt entity. A summary of the relation types for each entity pair category is provided in Table 4.6. This list of relation types was derived from a manual generalization of the relations identified by our LLaMIC model in flexible mode (where the model is instructed to infer the relation type)within the supervised MIMIC dataset, and from clustering these relations. A detailed explanation of this process is provided in Section 4.3.4.

The relationships between two CCD entities are defined as follows: *related_to* and *comorbidity*. The *related_to* relation denotes diseases that share a diagnostic or ontological connection, such as belonging to the same pathological spectrum or representing subtypes of a broader condition. The *comorbidity* relation refers to the co-occurrence of two clinically distinct diseases in the same patient, without implying a direct pathological link. Examples are provided below, respectively:

- NSTEMI: Patient a/w left-sided chest pain worse with exertion in setting of known CAD. (*NSTEMI*, *related_to*, *CAD*)
- STEMI: Pt initially presented to ___ with chest pain and EKG showing anterior STEMI. (*STEMI*, *related_to*, *STEMI*)
- Patient with CAD with CABG and CVA. (*CAD*, *comorbidity*, *CVA*)

The relationships between two CCDt entities are defined as follows: *related_to*, *combined_with*, and *replaced_by*. The *related_to* relation refers to drugs that represent the same chemical compound or belong to the same pharmacological class. The *combined_with* relation denotes different drugs that are

co-administered as part of a single therapeutic regimen. The *replaced_by* relation applies when one drug is withdrawn or substituted in favor of another. Examples are provided below, respectively:

- Being on heparin gtt (...) the heparin was held. (*heparin gtt*, *related_to*, *heparin*)
- Ms. ___ was started on a heparin bridge. Also increased aspirin from 81mg qd to 325mg qd. (*heparin*, *combined_with*, *aspirin*)
- She was initially treated with heparin gtt on admission, then transitioned to enoxaparin bridge. (*heparin gtt*, *replaced_by*, *enoxaparin*)

The relationships between a CCD entity and a CCDt entity are defined as follows: *indicated*, *no_indicated*, and *discontinued*. The *indicated* relation applies when a drug is prescribed for the treatment or prevention of a disease. The *no_indicated* relation denotes cases in which the drug is not recommended due to contraindications, inefficacy, or clinical risk. The *discontinued* relation applies when a drug previously prescribed for a disease is no longer administered. Examples are provided below, respectively:

- NSTEMI: Patient started on aspirin, ticagrelor, and atorvastatin. (*NSTEMI*, *indicated*, *aspirin*)
- CAD: Avoided beta-blockers due to history of severe asthma. (*CAD*, *no_indicated*, *beta-blockers*)
- STEMI: Initially on clopidogrel, which was later discontinued after bleeding event. (*STEMI*, *discontinued*, *clopidogrel*)

Entity Pair Type	**Relation Types**
CCD – CCD	*related_to*, *comorbidity*
CCDt – CCDt	*related_to*, *combined_with*, *replaced_by*
CCD – CCDt	*indicated*, *no_indicated*, *discontinued*

Table 4.6: Summary of binary relations defined between entity types.

4.3.2 Corpus Selection and Preprocessing

The dataset used in this task corresponds to the merge of supervised and unsupervised corpus introduced in the previous chapter 4.2.7 which includes DICNs that may or may not contain annotations related to CCDs and CCDts. Since our focus on the extraction of CCD–CCDt relations, and those entities are often sparsely distributed in DICNs, we adopted a co-occurrence method to extract informative text fragments likely to contain meaningful relational candidates. This method, briefly illustrated in Figure 4.3, involves three main steps: (1) filtering notes to retain only those containing at least one CCD and one CCDt annotation, and (2) segmenting the retained notes using the "#" character as a delimiter. This character

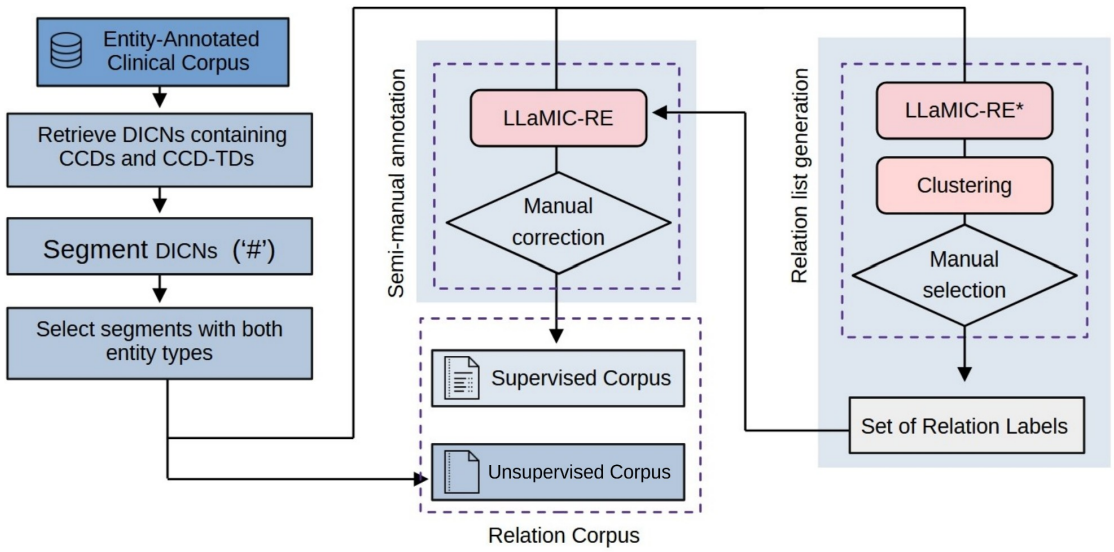

Figure 4.3: Workflow illustrating the annotation process for creating the supervised and unlabeled corpora, for relation extraction of CCDs and CCDts. LLaMIC-RE*: module for freely labeling relations

is frequently used in MIMIC DICNs to separate distinct sections or subcases (e.g., "# Diabetes: ASA 2mg daily… # CAD: Patient reported chest pain…"), allowing the extraction of shorter textual units that maintain clear and focused context (see in 'Brief Hospital Course' section from Figure 4.1) (3) select the segments with both entity types. The resulting segments was divided into two portions: one allocated for manual and semi-automatic annotation, forming the supervised corpus; and the other retained as the unsupervised corpus. The supervised corpus was further partitioned into training, validation, and test sets using an 8:1:1 ratio.

4.3.3 Baseline Models

We used BioLinkBERT for relation extraction, a 340-million-parameter biomedical language model based on LinkBERT and pretrained on PubMed texts with citation links. BioLinkBERT-large outperforms other biomedical language models, such as PubMedBERT [**?**] and BioBERT [Lee et al., 2019], on both NER/NEL tasks (e.g., BC5-chem with 94.04 F1, BC5-disease with 86.39 F1) and relation extraction tasks (e.g., DDI with 83.35 F1 and ChemProt with 79.98 F1) [Yasunaga et al., 2022]. For the model[3], a preprocessing adjustment of the input was required, which involved converting clinical note-level CSV files into pair-level JSON format. In this format, each note is processed to mark entity pairs as `[E1]`/`[E2]` and assign the corresponding relation labels. The hyperparameters used were identical to those reported for the DDI dataset (a corpus of pharmacological substances and drug–drug interactions) in the original

[3] https://github.com/michiyasunaga/LinkBERT/tree/main/src/seqcls

BioLinkBERT paper.

4.3.4 Semi-Automatic Annotation to Build a Supervised Corpus

The construction of the supervised corpus involved three key stages: (i) relation list generation (illustrated on the right side of Figure 4.3), (ii) automatic pre-annotation using the LLaMIC-RE module, and (iii) manual correction of the pre-annotated corpus (both depicted in the central portion of the figure).

Relation List Generation: To derive a meaningful and context-specific set of relations tailored to the DICNs corpus, we leveraged the generative capabilities of the LLaMIC pipeline operating in flexible mode. In this setup, the LLaMIC-RE relation classification module was prompted without any predefined relation list, allowing it to generate relation labels freely based on the contextual semantics of the input segments (Appendix A.1.9). The raw relation candidates obtained were then grouped into semantically coherent clusters by encoding them into dense vector representations using the Sentence-BERT model `all-MiniLM-L6-v2`. This semantic clustering enabled the identification of frequently occurring relational patterns. A summary of the main types of relation per cluster is presented in Figure A.2. From these clusters, a manual selection process was applied to refine the long list of candidate relations. This selection was guided by two main criteria: (1) each relation label should express a meaningful interaction independently of the specific head and tail entity types, and (2) the selected labels should maximize semantic coverage by consolidating similar candidate relations without compromising the granularity. The resulting set of relation types is summarized in Table 4.6.

Automatic Pre-Annotation: The pre-annotation step was performed using the LLaMIC-RE module, which employs a LLaMA-based language model without fine-tuning. The set of relation labels applied in LLaMIC-RE corresponds to those previously defined.

Manual Correction: Given the complexity of types of relations in DICN (as described in Section 3.2), manual corrections were made according to a set of predefined premises proposed in this work, ensuring internal consistency and reproducibility across the corpus. These premises also promote a more cohesive and normalized set of relations, which is essential for training the LLaMA module more efficiently while reducing the risk of erroneous or excessively long outputs. The annotation premises are summarized in Table 4.7. These premises emphasize the requirement for explicit temporal or linguistic cues (Premise P1), enforce inferential consistency through transitivity rules across entities (P2), and guarantee that each entity mention is contextualized within its most recent clinical state (P3).

Premise P1 emphasizes the necessity of explicit temporal or linguistic cues to justify the annotation of a relation between entities. In ambiguous cases where multiple drugs are indicated for the same disease without explicit temporal markers or evidence of combination therapy, no relation is defined between these drugs; instead, each drug-disease pair is independently annotated:

Premise	Description
P1	Relations should only be annotated when supported by temporal or linguistic connectors that indicate continuity, progression, or a direct and unequivocal relation between entities.
P2	Annotated relations admit transitivity. Specifically, if there exist entities $e_1 \xrightarrow{r_x} e_2$ and $e_2 \xrightarrow{r_x} e_3$, then it is possible to infer $e_1 \xrightarrow{r'} e_3$, where $r' = r_x$ if all three entities belong to the same semantic category (e.g., all diseases or all drugs), or $r' =$ related_to if the transitive path involves entities of different types (e.g., drugs and diseases).
P3	The assigned label must reflect the most recent clinical state relevant to the specific mention of the entity. Formally, if an entity e is mentioned multiple times with different clinical states $s_1, s_2, \ldots, s_n$, each mention is associated with the clinical state immediately preceding the subsequent mention of the same entity.

Table 4.7: Annotation Premises Guiding Manual Correction

Document segment:
`ASA (`m_1`) indicated for treatment of MI (`d_1`). Home med includes metoprolol (`m_2`).`

Resolution:
$(m_1) \xrightarrow{\texttt{indicated}} (d_1)$
$(m_2) \xrightarrow{\texttt{indicated}} (d_1)$

Conversely, when a disease mention is followed by a suspected subtype, exacerbation, or progression—linked through temporal or linguistic connectors (e.g., "with possible", "followed by", "progressing to")—the relation is annotated as `related_to`, in accordance with Premise P1. This ensures that the annotation captures meaningful clinical trajectories and reflects the progression or diagnostic uncertainty inherent in such contexts.

Document segment:
`Stable angina (`d_1`) with possible Acute Myocardial Infarction (`d_2`).`

Resolution:
$d_1 \xrightarrow{\texttt{related\_to}} d_2$

In accordance with Premise P2, we adopt a chain-based annotation strategy to systematically omit as many transitive relations as possible in an organized and principled manner. In this strategy, relations are annotated only between consecutive entities in a sequence (e.g., $e_1 \to e_2$, $e_2 \to e_3$, $e_3 \to e_4$). Relations that are transitive under Premise P2 but not explicitly annotated in the LAMIC output are reconstructed via rule-based post-processing, rather than being included directly in the annotated corpus. This strategy helps mitigate the risk of hallucinations in the LAMIC system caused by overly long or dense outputs, while preserving logical continuity in the underlying relational structure.

Document segment:
`Patient diagnosed with CAD (`d_1`), progressing to Ischemic Heart Disease (`d_2`), followed by Acute Myocardial Infarction (`d_3`).`

Resolution:
$d_1 \xrightarrow{\texttt{related\_to}} d_2$
$d_2 \xrightarrow{\texttt{related\_to}} d_3$
$d_1 \xrightarrow{\texttt{related\_to}} d_3$ (automatically inferred)

According to Premise P3, when a drug is mentioned multiple times with varying clinical states over time, each mention should be assigned the label corresponding to its most recent clinical state. This approach ensures that the annotated relation accurately reflects the dynamic therapeutic context associated with the entity.

Document segment:
`Aspirin (`m_1`) started preoperatively for stroke prevention (`d_1`) and discontinued at discharge.`

Resolution:
$m_1 \xrightarrow{\texttt{discontinued}} d_1$

Furthermore, in scenarios where a drug–disease pair exhibits a therapeutic association but the drug is either not prescribed or discontinued due to absence or resolution of the disease, the relation should be annotated as `discontinued`, in accordance with Premise P3.

Document segment:
`Aspirin (`m_1`) was discontinued in the setting of multiple bleeds, and no history of CAD (`d_1`).`

Resolution:
$m_1 \xrightarrow{\texttt{discontinued}} d_1$

In the same way as described for the supervised corpus in Section 4.2.4, the annotated corpus presented here was not subjected to a formal evaluation procedure.

4.3.5 Fine-Tuning and Deployment of LLaMIC

Following the construction of the supervised corpus for RE, we fine-tuned two LLaMIC models based on the LLaMA 3.1 8B language model. The first model was trained to generate candidate entity pairs for the initial module of LLaMIC-RE, while the second model was trained to classify these pairs with the appropriate relation label, based on the predefined set listed in Table 4.6. Both models were fine-tuned using a learning rate of 5×10^{-5}, with a batch size per device of 1 and gradient accumulation steps set

to 8. The pair generation model was trained for 10 epochs, while the relation classification model was trained for 5 epochs. We used the AdamW optimizer with 8-bit quantization (adamw_bnb_8bit), and applied a weight decay of 0.01. LoRA was used with the following configuration: rank $r = 8$, alpha = 32, and dropout = 0.05. The models are included in the *Supplementary Documents*. All LLaMIC models used the default temperature.

4.3.6 Evaluation Framework of the LLaMIC Pipeline

In both evaluation tasks, metrics were computed at the pair level, considering all possible entity pairs across the full test set. In the pair generation task, the model was evaluated based on its ability to correctly identify all relevant entity pairs (head, tail) present in the ground truth. Since the relation labels do not require a specific order between the entities, no directionality is imposed and (head, tail) is treated equivalently to (tail, head)). Evaluation metrics included precision, recall, and F1-score. In the relation classification task, the model was evaluated based on its ability to assign the correct relation label to a given entity pair. The evaluation was carried out at the level of full triplets (head, relation, tail). A prediction was considered correct only if all three components matched the ground truth, regardless of the order of the entities. As in the pair generation task, the evaluation metrics were precision, recall, and F1-score.

4.3.7 Results and Discussion

4.3.7.1 Preprocessing of the MIMIC-IV Annotated Entity Corpus

Out of the original 331,794 DICNs, comprising the merged supervised and unsupervised annotated corpora introduced in Chapter 4.2.7, a total of 45,647 notes segments containing potential interactions between CCDs and CCDts were retained. Table 4.8 presents the distribution of segments according to the presence or absence of the two entity types. We distinguish between the Header Segment (typically a summary of the patient history and hospitalization) and # Segments, which are segments initiated by "#" that separate distinct sections or subcases (e.g., "# Diabetes: ASA 2mg daily… # CAD: Patient reported chest pain…"). Specifically, 11,443 DICNs (4%) included both CCDs and CCDts in their header segment, while 34,204 (2%) did so in the # Segments. Based on these observations, we hypothesize that header segments tend to contain more mentions of CCDs than CCDts. Whereas in # Segments, CCDs and CCDts frequently co-occur. This pattern may explain why the majority of segments do not contain both (90%). In total, 45,647 segments containing both CCDs and CCDts were extracted for further annotation and relation extraction tasks, along with an equal number of segments that contained neither entity type.s<

Group	Both	w/o CCD	w/o CCDt	w/o Both	Total
Head segment	11,443	21,428	46,810	230,712	310,393
	(4%)	(7%)	(15%)	(74%)	
# Segments	34,204	59,354	46,707	1,275,626	1,415,891
	(2%)	(4%)	(3%)	(90%)	
Total	**45,647**	80,782	93,517	1,506,338	1,726,284

w/o = without.

Table 4.8: Distribution of entity co-occurrence across Header Segments and #Segments. Values in parentheses represent proportions relative to the respective row totals.

4.3.7.2 Supervised Corpus Statistics

Out of the 45,647 segments obtained after preprocessing, 799 segments were semi-automatically annotated to construct the supervised corpus. These corpus was subsequently split into training, validation, and test subsets using an 8:1:1 ratio. Table 4.9 summarizes the corpus statistics.

Supervised Corpus	Train Set	Validation Set	Test Set	Unsupervised
Relation Extraction	641	78	80	44,848

Table 4.9: Summary of supervised corpus splits for relation extraction task, including a placeholder for unsupervised data.

Based on the clusters of relation outputs generated by LLaMIC-RE on the these 799 segments (illustrated in Figure A.2), a manually curated list of relation types was created, resulting in the distribution presented in Table 4.10. A total of 1,969 relation instances were annotated across three main entity pair categories: between CCD, between CCDts, and between CCD and CCDt. As shown in Table 4.10, the most frequent relation was `indicated` between CCD and CCDt entities, with 826 occurrences. The next most frequent was `related_to`, observed both among CCD–CCD and CCDt–CCDt pairs. These corpora are included in the *Supplementary Documents*.

4.3.7.3 Automatic Entity Annotation

Using the supervised corpus obtained previously, two models were trained for each task on their corresponding corpora. For the LLaMA Pair Generator task, checkpoint-550 was utilized, while checkpoint-400 was employed for the LLaMA Relation Classifier task.

For the LLaMIC ablation study, we evaluated BioLinkBERT as a baseline, alongside several variants of our LLaMIC pipeline: LLaMIC-NoPG without the Pair Generation phase, relying instead on automatic creation of candidate pairs via permutations) and LLaMIC-NoFT (where the models at each phase

Entity Type	Relation Label	Count	% (Group)
CCD and CCD	related_to	730	(.97)
	comorbidity	23	(.03)
	Subtotal	**753**	
CCDt and CCDt	related_to	189	(.77)
	combined_with	52	(.21)
	replaced_by	5	(.02)
	Subtotal	**246**	
CCD and CCDt	indicated	826	(.85)
	discontinued	103	(.11)
	no_indicated	40	(.04)
	Subtotal	**970**	
Total		**1969**	

Table 4.10: Distribution of annotated relation types in the supervised corpus, grouped by entity pair type, including group totals and percentage breakdowns.

were not fine-tuned on the supervised corpus). For the LLaMA Pair Generator task, the baseline model BioLinkBERT achieved a precision of 0.63 (see Table 4.11). Our proposed LLaMIC model outperformed this baseline, reaching a precision of 0.67, corresponding to an absolute improvement of 0.04. Comparing the baseline approach—which relies on automatic pair permutations—with our LLaMIC-NoPG variant (51% precision) highlights the substantial benefit of LLaMIC's generative pair selection phase. Rather than relying solely on brute-force permutations to generate candidate pairs for relation classification, LLaMIC leverages the model's interpretative and generative capabilities to identify the most probable entity pairs that exhibit a relationship.

Metric	BioLinkBERT	LLaMIC-NoPG	LLaMIC-NoFT	LLaMIC
Precision	0.63	0.51	0.57	**0.67**
Recall	**0.75**	0.49	0.56	0.66
F1 Score	**0.68**	0.45	0.56	0.66

LLaMIC-NoPG = without Pair Generator phase; **LLaMIC-NoFT** = without fine-tuning.

Table 4.11: Evaluation of the pair generation step and ablation study on fine-tuning and pair generation strategies (macro-averaged).

Table 4.12 presents a detailed evaluation of LLaMIC's performance at both relation and note levels.

At the relation level, results are reported for the Pair Generation Phase per class and for the Label Relation Phase per class. At the note level, precision is reported per number of entities in a note and per entity pair type. LLaMIC demonstrates a strong ability to correctly pass non-existent relations (i.e., reducing false positives), achieving a precision of 70.6%, and shows enhanced precision for identifying the "related to" relation (62.9%).

At the note level, precision decreases as the number of entities per note increases, reaching 34.1% for notes containing four or more entities, likely due to the increased complexity of multi-entity relations and the presence of pair types CCD–CCD and CCDt–CCDt, which individually exhibit lower precision (45% and 18%, respectively). The particularly low precision for CCDt–CCDt pairs may result from their limited representation in the supervised training corpus. The supervised corpus was relatively small and lacked diversity in the distribution of relation types, which may explain the observed variability in precision across different relation classes and note types (see Table 4.12). For comparison, previous state-of-the-art results on the CVDEMRC (cardiovascular electronic medical record) corpus reported corpus reported a JREwBART model achieving a relation extraction precision of 0.66 and an F1 score of 0.64 for CAD and treatment relations [Guo et al., 2024]. LLaMIC achieved comparable results, reaching a precision of 0.67, despite being fine-tuned on a substantially smaller and less diverse training and test set.

II. Over-Prediction LLAMIC without Pair Generator phase:

The permutation-based LLaMIC model– LLaMIC-NoPG – tends to over-predict relations, labeling pairs as related when no actual relation exists in the gold standard. This behavior increases false positives, particularly in the *indicated* class. The Pair Generation phase improved precision by 16 percentage points.

Document segment:
`The patient is on chronic <drug1>amlodipine</drug1> for blood pressure. <disease1>Hypertension</disease1> was mentioned in the history, but no change in medication was made during this visit.`

Pair: *<disease1>Hypertension, <drug1>amlodipine*
LLaMIC output: *indicated*
Ground truth: *no relation*

The performance gap between BioLinkBERT and LLaMIC-NoPG may arise from both pretraining and architectural factors. BioLinkBERT's domain-specific pretraining on biomedical literature enhances its understanding of specialized terminology and relationships, improving entity-pair classification. Additionally, as an encoder-based transformer, BioLinkBERT generates bidirectional contextual representations more effectively than the decoder-based LLaMA, which is primarily optimized for autoregressive generative tasks. This architectural difference allows encoder-based models to better capture long-range dependencies and complex interactions between entity pairs, which are critical for accurate relation classification. In contrast, decoder-only models may struggle to incorporate full contextual information across the sequence, leading to reduced performance in structured classification tasksn Brokman and Kavuluru

Metric / Category	LLaMIC	LLaMIC-NoPG	LLaMIC-NoFT
Pair Generation Phase			
P (no relation)	**0.706**	0.698	0.685
P (relation)	**0.645**	0.406	0.497
Label Relation Phase (S $\geq$ 10)			
P (no relation)	**0.706**	0.698	0.685
P (indicated)	**0.505**	0.324	0.417
P (related to)	**0.629**	0.480	0.580
Number of Entities per Note			
P (2 Entities)	**0.486**	0.435	0.309
P (3 Entities)	**0.563**	0.306	0.280
P (>4 Entities)	**0.341**	0.176	0.203
Entity Pair Types			
P (CCD – CCD)	**0.446**	0.316	0.419
P (CCDt – CCDt)	**0.184**	0.203	0.128
P (CCD – CCDt)	**0.569**	0.303	0.309

LLaMIC-NoPG = without Pair Generator phase; **LLaMIC-NoFT** = without fine-tuning;
P = Precision; **S** = Support.

Table 4.12: Precision (P) scores for LLaMIC models across different phases and categories.

[2025].

II. Under-Prediction LLAMIC

As shown in the quantitative evaluation, the fine-tuned (LLaMIC) and unfine-tuned (LLaMIC-NoFT) LLaMIC models tend to under-predict relations, often missing relations that are present in the ground truth, such as the *indicated* relation. This conservative behavior reduces false positives but results in lower recall for certain relation types.

Document segment:
`His NCHCT/MRI showed small hypodensity in the right frontal/temporal white matter c/w <disease2>subacute vs chronic infarct</disease2>. Patient started on <drug1>ASA</drug1> and atorvastatin.`
Pair: *<disease1>subacute vs chronic infarct, <drug1>ASA*
LLaMIC output: *no relation*
Ground truth: *indicated*

Additionally, the model exhibits challenges in handling complex relation scenarios, particularly when

multiple mentions of the same drug appear with differing clinical contexts. In the example below,the model fails to correctly capture the transition from initiation to discontinuation, thereby misclassifying the actual relationship between the disease and the drug state.

Document segment:
`<disease1>CAD</disease1> s/p PCI ___ and CABG ___ (...) Patient was initially started on a <drug1>nitro</drug1> gtt with resolution of his chest pain. The <drug2>nitro</drug2> gtt was weaned without recurrence of his pain.`

Pair: *<disease1>CAD, <drug2>nitro*
LLaMIC output: *indicated*
Ground truth: *discontinued*

IV. Unsupervised Corpus Annotation

Based on the evaluation results, LLaMIC-RE, which demonstrated the highest performance for relation extraction, was used to annotate binary relations in the 44,848 DICN segments of the unsupervised corpora. These corpora are included in the *Supplementary Documents*.

V. Computational Performance

All experiments were conducted on an NVIDIA Tesla T4 GPU (16 GB VRAM). THe processing time of LLaMIC averages 24 seconds per DICN. Processing time varied depending on the length of the notes and the density of annotated entities.

Chapter 5

Conclusion

EHRs contain extensive structured and unstructured clinical data, with critical information often embedded in free-text clinical notes. Extracting knowledge from these notes is particularly relevant in the cardiovascular and CCD domain, where morbidity and mortality remain high. Nevertheless, clinical notes pose significant challenges due to their length, specialized terminology, and inherent ambiguities. Moreover, the scarcity of biomedical datasets addressing advanced tasks such as NER/NEL and RE on free-text clinical notes—especially within the CCD domain—has limited the development of comprehensive NLP approaches. This dissertation addresses these challenges through two main contributions: (i) the creation of a supervised corpus of CCDs, therapeutic drugs, and binary relations derived from real-world clinical notes; and (ii) the development of a state-of-the-art NLP pipeline, *LLaMIC*.

The supervised corpus was constructed using a semi-automatic annotation methodology that combined the LLaMIC pipeline, BENT, and clustering techniques, followed by manual correction (Section 4). The final dataset comprises 13,521 CCD entities and 3,223 CCDt entities, linked to terminologies such as ICD-10 and MeSH, distributed across 11,514 and 5,845 DICNs, respectively. In addition, 1,969 relations were manually annotated across 799 DICNs.

In addition to the corpus, we introduced LLaMIC, a modular and flexible pipeline for CCD and CCDt entity recognition and RE in DICNs (Section 3). This approach integrates lexicon-based methods for entity detection with open-source LLMs, demonstrating strong effectiveness in addressing both document length and textual complexity. By employing an adjustable text-window mechanism, LLaMIC successfully processed extensive DICN notes. Regarding textual complexity, LLaMIC also achieved robust performance in cases involving misspellings, abbreviations, or long multi-word disease mentions. For the RE task, the system achieved competitive results in identifying relation-bearing pairs. While the improvements over baseline systems were not statistically significant, LLaMIC consistently matched or exceeded baseline performance in several metrics. Specifically, in NER, LLaMIC achieved a 37 percentage points precision improvement for CCD entities under lenient evaluation, while it underperformed by 2 percentage points for CCDt entities. In relation extraction, the system reached 67% precision, corre-

sponding to a 4 percentage points gain compared to the baseline. These findings suggest that, although statistical significance was not achieved, there is substantial potential for improvement, particularly in the pair generation stage of relation extraction. Alongside the pipeline, we release all six LLaMIC modules, each fine-tuned for its respective task.

Overall, this work demonstrates not only the feasibility but also the efficiency and accuracy of integrating lexicon-based methods with LLMs for entity and relation extraction in clinical notes. The supervised dataset and the LLaMIC pipeline constitute valuable resources and methodologies that can support future research in clinical NLP for CCDs.

Overall, this work achieves the main objectives set out in this thesis. First, we successfully developed a supervised corpus of CCDs, CCDts, and their relations, employing a hybrid annotation methodology that combined automatic annotation with manual correction. This corpus provides a resource for training and evaluating NLP models in the CCD domain. Second, we designed, implemented, and evaluated LLaMIC, a modular pipeline for entity recognition and relation extraction capable of identifying and classifying relationships between CCDs and CCDts from clinical notes. Beyond these primary objectives, LLaMIC was further adapted to offer flexibility in selecting target entities and in integrating different LLMs, enhancing its applicability to a wider range of clinical NLP tasks.

5.1 Future Work

Several avenues remain open for extending and strengthening the present study. These include:

Entity Expansion

- Extend the annotation schema to cover additional clinical concepts, such as disease findings, drug initiation events, drug dosages, and temporal relations. Existing annotations for diseases and drug mentions can serve as a foundation to identify more complex cases. For instance, disease findings such as “ST elevations” could be annotated as diseases, or information already present in extracted entities, such as dosages (e.g., “5mg ASA”) or status (e.g., “early stroke”), could be corrected and refined.

- Integrate assertion status detection to capture whether a statement is confirmed, negated, hypothetical, resolved, possible, or uncertain, following existing standards and corpora [Mowery et al., 2013]. These annotations are essential for improving relation extraction. Current relations may be ambiguous; for example, in ”CAD with no finding of MI”, the extracted relation is vague (”CAD” → related to → ”MI”). Incorporating assertion status would disambiguate such relations.

- Perform manual correction of existing annotations to address inconsistencies and unmapped entities, thereby improving corpus quality. Given that the corpus was initially curated by a master’s

student, inter-annotator review by domain experts could further enhance accuracy, although access to specialized data may constrain this process.

Relation Extraction

- Explore hybrid architectures, where a generative Pair Generator is complemented by a discriminative classifier (e.g., BERT) to enhance performance. As shown in Section 4.3.7.3, the initial generative phase contributed most to the performance of our LLaMA-based model.
- Incorporate semantic similarity between biomedical entities or integrate ontological knowledge to support inference of relations between entities of the same type, which may not be directly captured in the training data. This approach could improve extraction of relations between entities such as CCDs, which currently rely primarily on the textual context of the clinical note and the prior knowledge encoded in the LLM.

Overall Improvements

- The observed differences in percentage points for NER and relation extraction may not be statistically significant. Future work should include formal statistical tests, such as McNemar's test, to assess whether these improvements are meaningful.
- Expand the supervised dataset and involve more annotators with inter-annotator agreement measures to enable more robust training and evaluation. Also, expand the corpus to increase coverage, particularly for the supervised relation extraction dataset. While entity annotations may cover a sufficiently diverse set of concepts, relation annotations remain relatively limited.
- Apply the LLaMIC pipeline to clinical notes written in languages other than English, enabling cross-lingual evaluation. Given that clinical notes are often written in the local language, this adaptation would evaluate pipeline performance across languages. The pipeline is fully modular and can accept any text, as documented in the GitHub repository[1].
- Investigate strategies for parallelization and scalability, including distributed processing, to efficiently handle larger corpora and reduce latency. In the MIMIC corpus, current processing times were approximately 14s for entity detection and 24s for relation extraction per note, highlighting the need for optimization when processing longer and more complex clinical notes.
- Conduct a more extensive ablation study of LLaMIC, implementing additional compatible LLMs, including models specifically trained for the clinical or biomedical domain. Credentials and configuration details of these models are described in Section 3.3.

[1] https://github.com/lasigeBioTM/LLAMIC

References

Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., and Sontag, D. (2022). Large language models are few-shot clinical information extractors. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. 19, 21

AHA and NCHS (2023). Annual survey information technology supplement and national ambulatory care survey / national electronic health record survey. https://www.cdc.gov/nchs/nehrs/index.html. Data covering 2008–present. Accessed: 2025-08-10. 7

Altalla', B., Abdalla, S., Altamimi, A., Bitar, L., Omari, A. A., Kardan, R., and Sultan, I. (2025). Evaluating gpt models for clinical note de-identification. *Scientific Reports*, 15(1):3852. 16

Aronson, A. and Lang, F. (2010). An overview of metamap: historical perspective and recent advances. *J Am Med Inform Assoc*, 17(3):229–236. 11, 12

Arslan, I. G., Damen, J., de Wilde, M., van den Driest, J. J., Bindels, P. J. E., van der Lei, J., Schiphof, D., and Bierma-Zeinstra, S. M. A. (2022). Incidence and prevalence of knee osteoarthritis using codified and narrative data from electronic health records: A population-based study. *Arthritis Care & Research*, 74(6):937–944. 7

Belkadi, S., Han, L., Wu, Y., and Nenadic, G. (2023). Exploring the value of pre-trained language models for clinical named entity recognition. In *2023 IEEE International Conference on Big Data (BigData)*, pages 3660–3669. 18, 21

Beltagy, I., Lo, K., and Cohan, A. (2019). Scibert: A pretrained language model for scientific text. 16

Boonstra, M. J., Weissenbacher, D., Moore, J. H., Gonzalez-Hernandez, G., and Asselbergs, F. W. (2024). Artificial intelligence: revolutionizing cardiology with large language models. *European Heart Journal*, 45(5):332–345. 2

Brokman, A. and Kavuluru, R. (2025). How important is domain-specific language model pretraining and instruction finetuning for biomedical relation extraction? In Ichise, R., editor, *Natural Language Processing and Information Systems*, pages 80–94, Cham. Springer Nature Switzerland. 53

Bundschus, M., Dejori, M., Stetter, M., Tresp, V., and Kriegel, H.-P. (2008). Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinformatics*, 9:207 – 207. 12

Campillos, L., Deléger, L., Grouin, C., Hamon, T., and Névéol, A. (2018). A french clinical corpus with comprehensive semantic annotations: development of the medical entity and relation limsi annotated text corpus (merlot). *Language Resources and Evaluation*, 52(2):571–601. 18, 20

CDC (2024). Cardiovascular disease - chronic disease indicators. Accessed: December 12, 2024. 32

Chang, H., Zan, H., Zhang, S., Zhao, B., and Zhang, K. (2023). Construction of cardiovascular information extraction corpus based on electronic medical records. *Mathematical Biosciences and Engineering*, 20(7):13379–13397. 2, 18, 20

Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., and Buchanan, B. G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5):301–310. 11

Couto, F. and Lamurias, A. (2018). Mer: a shell script and annotation server for minimal named entity recognition and linking. *J Cheminform*, 10:58. Received: 24 July 2018; Accepted: 30 November 2018; Published: 05 December 2018. 11

Crammer, K., Singer, Y., Cristianini, N., Shawe-Taylor, J., and Williamson, B. (2002). On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res*, 2. 12

de Bruijn, B., Cherry, C., Kiritchenko, S., Martin, J., and Zhu, X. (2011). Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *Journal of the American Medical Informatics Association*, 18(5):557–562. 12

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. 17

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. 16

Dong, X., Chowdhury, S., Qian, L., Li, X., Guan, Y., Yang, J., and Yu, Q. (2019). Deep learning for named entity recognition on chinese electronic medical records: Combining deep transfer learning with multitask bi-directional lstm rnn. *PLOS ONE*, 14(5):1–15. 13

Elhadad, N., Pradhan, S., Gorman, S., Manandhar, S., Chapman, W., and Savova, G. (2015). SemEval-2015 task 14: Analysis of clinical text. In Nakov, P., Zesch, T., Cer, D., and Jurgens, D., editors, *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 303–310, Denver, Colorado. Association for Computational Linguistics. 18, 19, 21

Fraile Navarro, D., Ijaz, K., Rezazadegan, D., Rahimi-Ardabili, H., Dras, M., Coiera, E., and Berkovsky, S. (2023). Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review. *International Journal of Medical Informatics*, 177:105122. 16

Fuster, V., Kelly, B. B., and Vedanthan, R. (2011). Global cardiovascular health. *JACC*, 58(12):1208–1210. 1

Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A., and et al. (2023). How does chatgpt perform on the united states medical licensing examination? the implications of large language models for medical education and knowledge assessment. *JMIR Medical Education*, 9(1):e45312. 16

Goeuriot, L., Suominen, H., Kelly, L., Miranda-Escalada, A., Krallinger, M., Liu, Z., Pasi, G., Gonzalez Saez, G., Viviani, M., and Xu, C. (2020). Overview of the clef ehealth evaluation lab 2020. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings*, page 255–271, Berlin, Heidelberg. Springer-Verlag. 18, 20

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., and Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220. Online publication. 31

Goyal, N. and Singh, N. (2025). Named entity recognition and relationship extraction for biomedical text: A comprehensive survey, recent advancements, and future research directions. *Neurocomputing*, 618:129171. 8

Gruber, T. (1993). A translational approach to portable ontologies. *Knowledge Acquisition*, 5:199–220. 8

Gu, J., Sun, F., Qian, L., and Zhou, G. (2017). Chemical-induced disease relation extraction via convolutional neural network. *Database*, 2017(1):bax024. 13

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1). 16

Guo, Y., Zan, H., Chang, H., Zhou, L., and Zhang, K. (2024). *A BART-Based Study of Entity-Relationship Extraction for Electronic Medical Records of Cardiovascular Diseases*, pages 82–97. 18, 21, 53

Harkema, H., Dowling, J. N., Thornblade, T., and Chapman, W. W. (2009). Context: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics*, 42(5):839–851. Biomedical Natural Language Processing. 12

Henry, S., Buchan, K., Filannino, M., Stubbs, A., and Uzuner, O. (2020). 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *J Am Med Inform Assoc*, 27(1):3–12. 18, 19, 21

H.Hariri, R., Fredericks, E., and Bowers, K. (2019). Uncertainty in big data analytics: survey, opportunities, and challenges. *Journal of Big Data*, 6:44. 1

Houssein, E. H., Mohamed, R. E., and Ali, A. A. (2023). Heart disease risk factors detection from electronic health records using advanced nlp and deep learning techniques. *Scientific Reports*, 13(1):7173. 18, 21

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. 17

Hu, Y., Chen, Q., Du, J., Peng, X., Keloth, V. K., Zuo, X., Zhou, Y., Li, Z., Jiang, X., Lu, Z., Roberts, K., and Xu, H. (2024). Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 31(9):1812–1820. 16

Huang, K., Altosaar, J., and Ranganath, R. (2020). Clinicalbert: Modeling clinical notes and predicting hospital readmission. 16

Jauregi Unanue, I., Zare Borzeshi, E., and Piccardi, M. (2017). Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. *J Biomed Inform*, 76:102–109. 11

Jensen, P. B., Jensen, L. J., and Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nat. Rev. Genet.*, 13(6):395–405. 1

Ji, Z., Wei, Q., and Xu, H. (2020). Bert-based ranking for biomedical entity normalization. In *AMIA Joint Summits on Translational Science proceedings*, pages 269–277. 11

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023). Mistral 7b. 2, 18

Johnson, A., Pollard, T., Horng, S., Celi, L. A., and Mark, R. (2023). Mimic-iv-note: Deidentified free-text clinical notes (version 2.2). Accessed: 2024-12-11. 31

Khurana, D., Koli, A., Khatter, K., and Singh, S. (2022). Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, 82(3):3713–3744. 8

Kim, H., Kim, J.-E., and Kim, H. (2024). Exploring nested named entity recognition with large language models: Methods, challenges, and insights. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors,

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 8653–8670, Miami, Florida, USA. Association for Computational Linguistics. 16

Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.-A., Rouvier, M., and Dufour, R. (2024). Biomistral: A collection of open-source pretrained large language models for medical domains. 18

Lafferty, J. D., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*. 12

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. 2

Leaman, R., Islamaj Doğan, R., and Lu, Z. (2013). Dnorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917. 11, 12

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240. 2, 16, 46

Li, X., Feng, J., Meng, Y., et al. (2020). A unified mrc framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859. 16

Lipscomb, C. E. (2000). Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265–266. 10

Liu, S., Tang, B., Chen, Q., and Wang, X. (2016). Drug-drug interaction extraction via convolutional neural networks. *Computational and Mathematical Methods in Medicine*, 2016(1):6918381. 13

Luo, Y., Song, G., Li, P., and Qi, Z. (2018). Multi-task medical concept normalization using multi-view convolutional neural network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1). 13

Marimon, M., Vivaldi, J., and Bel, N. (2017). Annotation of negation in the IULA Spanish clinical record corpus. In Blanco, E., Morante, R., and Saurí, R., editors, *Proceedings of the Workshop Computational Semantics Beyond Events and Roles*, pages 43–52, Valencia, Spain. Association for Computational Linguistics. 18, 20

Meystre, S. and Haug, P. J. (2006). Natural language processing to extract medical problems from electronic clinical documents: Performance evaluation. *Journal of Biomedical Informatics*, 39(6):589–599. 18, 19, 21

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In Su, K.-Y., Su, J., Wiebe, J., and Li, H., editors, *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics. 12

Moon, S., Pakhomov, S., Liu, N., Ryan, J. O., and Melton, G. B. (2014). A sense inventory for clinical abbreviations and acronyms created using clinical notes and medical dictionary resources. *Journal of the American Medical Informatics Association*, 21(2):299–307. 19

Mowery, D. L., Jordan, P., Wiebe, J., Harkema, H., Dowling, J., and Chapman, W. W. (2013). Semantic annotation of clinical events for generating a problem list. In *Proceedings of the AMIA Annual Symposium*, pages 1032–1041. AMIA. 18, 19, 58

Nadkarni, P., Ohno-Machado, L., and Chapman, W. (2011). Natural language processing: An introduction. *Journal of the American Medical Informatics Association : JAMIA*, 18:544–51. 8

OpenAI (2024). Openai gpt-4 technical report. 2

Peng, N., Poon, H., Quirk, C., Toutanova, K., and Yih, W.-t. (2017). Cross-sentence n-ary relation extraction with graph LSTMs. *Transactions of the Association for Computational Linguistics*, 5:101–115. 11

Pradhan, S., Elhadad, N., Chapman, W., Manandhar, S., and Savova, G. (2014). SemEval-2014 task 7: Analysis of clinical text. In Nakov, P. and Zesch, T., editors, *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 54–62, Dublin, Ireland. Association for Computational Linguistics. 18, 19

Pradhan, S., Elhadad, N., South, B. R., Martinez, D., Christensen, L., Vogel, A., Suominen, H., Chapman, W. W., and Savova, G. (2015). Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *Journal of the American Medical Informatics Association*, 22(1):143–154. 21

Rahal, R. M., Mercer, J., Kuziemsky, C., and Yaya, S. (2021). Factors affecting the mature use of electronic medical records by primary care physicians: a systematic review. *BMC Medical Informatics and Decision Making*, 21(1):67. 7

Rajkomar, A., Oren, E., Chen, K., Dai, A., Hajaj, N., Liu, P., Liu, X., Sun, M., Sundberg, P., Yee, H., Zhang, K., Duggan, G., Flores, G., Hardt, M., Irvine, J., Le, Q., Litsch, K., Marcus, J., Mossin, A., and Dean, J. (2018). Scalable and accurate deep learning for electronic health records. *npj Digital Medicine*, 1. 2

Richter-Pechanski, P., Wiesenbach, P., Schwab, D. M., Kiriakou, C., He, M., Allers, M. M., Tiefenbacher, A. S., Kunz, N., Martynova, A., Spiller, N., Mierisch, J., Borchert, F., Schwind, C., Frey, N., Dieterich, C., and Geis, N. A. (2023). A distributable german clinical corpus containing cardiovascular clinical routine doctor's letters. *Scientific Data*, 10(1):207. 18, 20

Rink, B., Harabagiu, S., and Roberts, K. (2011). Automatic extraction of relations between medical concepts in clinical texts. *Journal of the American Medical Informatics Association*, 18(5):594–600. 12

Ruas, P. and Couto, F. M. (2022). Nilinker: Attention-based approach to nil entity linking. *Journal of Biomedical Informatics*, 132:104137. 11, 35

Savova, G. K., Masanz, J. J., Ogren, P. V., et al. (2010). Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513. 11, 12, 18, 19, 21

Schriml, L. M., Mitraka, E., Munro, J., Tauber, B., Schor, M., Nickle, L., and Greene, C. (2018). Human disease ontology 2018 update: classification, content and workflow expansion. *Nucleic acids research*, 47(D1):D955–D962. 10

Seinen, T. M., Kors, J. A., van Mulligen, E. M., Fridgeirsson, E., and Rijnbeek, P. R. (2023). The added value of text from dutch general practitioner notes in predictive modeling. *Journal of the American Medical Informatics Association*, 30(12):1973–1984. 7

Sheikhalishahi, S., Miotto, R., Dudley, J. T., Lavelli, A., Rinaldi, F., and Osmani, V. (2019). Natural language processing of clinical notes on chronic diseases: Systematic review. *JMIR Medical Informatics*, 7(2). Cited by: 322; All Open Access, Gold Open Access, Green Open Access. 1

Si, Y., Wang, J., Xu, H., and Roberts, K. (2019). Enhancing Clinical Concept Extraction with Contextual Embedding. *Journal of the American Medical Informatics Association*. 2

Singh, A., Krishnamoorthy, S., and Ortega, J. E. (2024). Neighbert: Medical entity linking using relation-induced dense retrieval. *Journal of Healthcare Informatics Research*, 8(2):353–369. 10

Sousa, D. F. and Couto, F. M. (2023). K-ret: knowledgeable biomedical relation extraction system. *Bioinformatics*, 39(4):btad174. 11

Soysal, E., Wang, J., Jiang, M., Wu, Y., Pakhomov, S., Liu, H., and Xu, H. (2017). Clamp – a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*, 25(3):331–336. 11

Stubbs, A., Kotfila, C., Xu, H., and Uzuner, Ö. (2015). Identifying risk factors for heart disease over time: Overview of 2014 i2b2/uthealth shared task track 2. *Journal of Biomedical Informatics*, 58 Suppl:S67–S77. 18, 21, 40

Stubbs, A. and Uzuner, □. (2015). Annotating risk factors for heart disease in clinical narratives for diabetic patients. *Journal of Biomedical Informatics*, 58 Suppl:S78–S91. Epub 2015 May 21. 18, 19

Su, J., He, B., Guan, Y., Jiang, J., and Yang, J. (2017). Developing a cardiovascular disease risk factor annotated corpus of chinese electronic medical records. *BMC Medical Informatics and Decision Making*, 17(1):117. 18, 20, 21

Su, X., Wang, Y., Gao, S., Liu, X., Giunchiglia, V., Clevert, D.-A., and Zitnik, M. (2025). Kgarevion: An ai agent for knowledge-intensive biomedical qa. 16, 24

Suárez-Paniagua, V., Dong, H., and Casey, A. (2021). A multi-bert hybrid system for named entity recognition in spanish radiology reports. In *CLEF eHealth 2021, CLEF 2021 Evaluation Labs and Workshop: Online Working Notes*. CEUR-WS. Online Working Notes. 21

Suominen, H. et al. (2021). Overview of the clef ehealth evaluation lab 2021. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 308–323. Springer. 18, 20

Tang, B., Cao, H., Wu, Y., et al. (2013). Recognizing clinical entities in hospital discharge summaries using structural support vector machines with word representation features. *BMC Med Inform Decis Mak*, 13:S1. 12, 21

Touvron, H. et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*. 2, 18

Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y. (2005). Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484. 12

Uzuner, O., Solti, I., and Cadag, E. (2010). Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518. 18, 19, 21

Uzuner, Ö., South, B. R., Shen, S., and DuVall, S. L. (2011). 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556. 18, 19, 21

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Łukasz Kaiser, and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30*. NeurIPS. 13, 15

Velupillai, S., Mowery, D., South, B. R., Kvist, M., and Dalianis, H. (2015). Recent advances in clinical natural language processing in support of semantic analysis. *Yearb Med Inform*, 10(1):183–193. 8

Wang, E., Schmidgall, S., Jaeger, P. F., Zhang, F., Pilgrim, R., Matias, Y., Barral, J., Fleet, D., and Azizi, S. (2025). Txgemma: Efficient and agentic llms for therapeutics. 2, 18

Wang, H., Gao, C., Dantona, C., Hull, B., and Sun, J. (2024). Drg-llama: Tuning llama model to predict diagnosis-related group for hospitalized patients. *NPJ Digital Medicine*, 7(1):16. 18, 34

Weegar, R., Pérez, A., Casillas, A., and Oronoz, M. (2019). Recent advances in swedish and spanish medical entity recognition in clinical texts using deep neural approaches. *BMC Medical Informatics and Decision Making*, 19(7):274. 13

White, F. (2020). Application of disease etiology and natural history to prevention in primary health care: A discourse. *Medical Principles and Practice*, 29(6):501–513. Epub 2020 May 18. 32

WHO (2019). *International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10)*, 10th edition. Accessed: 2025-09-20. 10

WHO (2024). Health topics: Cardiovascular diseases. Accessed: 27 November 2024. 1

Wieland-Jorna, Y., van Kooten, D., Verheij, R. A., de Man, Y., Francke, A. L., and Oosterveld-Vlug, M. G. (2024). Natural language processing systems for extracting information from electronic health records about activities of daily living. a systematic review. *JAMIA Open*, 7(2):ooae044. 7

Woldemariam, M. and Jimma, W. (2023). Adoption of electronic health record systems to enhance the quality of healthcare in low-income countries: a systematic review. *BMJ Health Care Inform*, 30(1):e100704. 1

Yang, X., Bian, J., Hogan, W. R., and Wu, Y. (2020). Clinical concept extraction using transformers. *Journal of the American Medical Informatics Association*, 27(12):1935–1942. 16

Yang, X., Chen, A., PourNejatian, N., Shin, H. C., Smith, K. E., Parisien, C., Compas, C., Martin, C., Costa, A. B., Flores, M. G., Zhang, Y., Magoc, T., Harle, C. A., Lipori, G., Mitchell, D. A., Hogan, W. R., Shenkman, E. A., Bian, J., and Wu, Y. (2022). A large language model for electronic health records. *npj Digital Medicine*, 5(1):194. 2

Yasunaga, M., Leskovec, J., and Liang, P. (2022). Linkbert: Pretraining language models with document links. 46

Zeng, D., Liu, K., Lai, S., Zhou, G., and Zhao, J. (2014). Relation classification via convolutional deep neural network. In Tsujii, J. and Hajic, J., editors, *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, Dublin, Ireland. Dublin City University and Association for Computational Linguistics. 12

Zhang, J., Li, J., Wang, S., Zhang, Y., Cao, Y., Hou, L., and Li, X.-L. (2018). Category multi-representation: A unified solution for named entity recognition in clinical texts. In *Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC,*

Australia, June 3-6, 2018, Proceedings, Part II, page 275–287, Berlin, Heidelberg. Springer-Verlag. 12, 17

Zhang, Y., Wang, J., Tang, B., Wu, Y., Jiang, M., Chen, Y., and Xu, H. (2014). UTH_CCB: A report for SemEval 2014 – task 7 analysis of clinical text. In Nakov, P. and Zesch, T., editors, *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 802–806, Dublin, Ireland. Association for Computational Linguistics. 21

Appendix A

Extra Information

A.1 Prompt Templates

A.1.1 Prompt for CCD NER

```
Given the following text, identify and extract all medically relevant ischemic heart
diseases and strokes explicitly or implicitly contained.
Explicit mentions refer to diseases named directly within the text, preserving exact
wording, formatting, and order. Implicit mentions must only be included when there
is clear evidence linking the described condition to an ischemic heart disease or
stroke.
Do not include mentions of medication, symptoms, medical devices, treatments, or
medical procedures, even if they are associated with ischemic heart diseases or
strokes. Focus exclusively on identifying disease names.
Exclude conditions that are not explicitly or implicitly ischemic heart diseases or
strokes, such as hypertension, even if they suggest cardiovascular risk.
Return the output in JSON format with the key "cvd_terminologies", and the value
as a list of extracted ischemic heart diseases and strokes, maintaining the exact
original wording and order.
If no ischemic heart diseases or strokes are found in a particular part, return an
empty list in the JSON.
Text: {{text}}
The final JSON is:
```

A.1.2 Prompt for CCDt NER

```
Given the following text, identify and extract all medically relevant drugs related
with coronary diseases and strokes related with coronary diseases and strokes
explicitly contained.
Explicit mentions refer to drugs related with coronary diseases and strokes related
with coronary diseases and strokes named directly within the text, preserving exact
wording, formatting, and order.
Avoid mentions of doses, symptoms, medical devices, or medical procedures, even if
they are associated with the drug. Focus exclusively on identifying drugs related
with coronary diseases and strokes related with coronary diseases and strokes names.
Return the output in JSON format with the key "drugs_terminologies", and the value
as a list of extracted drugs related with coronary diseases and strokes related with
coronary diseases and strokes, maintaining the exact original wording and order.
If no drugs related with coronary diseases and strokes are found in a particular
part, return an empty list in the JSON.
Text: {{text}}
The final JSON is:
```

A.1.3 Prompt for CCD NEL

```
Given the following text and a list of diseases present in the text, generate a
set of related pairs. Each pair should consist of a disease name (identical to the
one provided in the list) and its respective 3 digits ICD-10 code. The ICD code
must best represent the disease based on the context provided in the text and your
knowledge of ICD.
The ICD code must have 3 digits only, like the following : ["I20", "N93"]
Ensure that the disease name in the output pairs matches exactly as given in the
input List_Diseases.
Each pair should be formatted as: (Disease Name, 3 digit ICD-10).
Return the generated pairs in a structured JSON format where the key is "Pairs" and
the value is a list of pairs (with the format: [(Disease Name, ICD), (Disease Name,
ICD)]).
Input:
Text: {{text}}
List_Diseases: {{diseases_list}}
The final JSON is:
```

A.1.4 Prompt for CCDt NEL

```
Given the following a list of drugs related with coronary diseases and strokes
generate a set of related pairs. Each pair should consist of a drug name (identical
to the one provided in the list) and its respective MeSH ID. The MeSH ID must best
represent the drug based on your knowledge of MeSH.
The MeSH ID must have MeSH + unique number.
Ensure that the drug name in the output pairs matches exactly as given in the input
List_drugs related with coronary diseases and strokes.
Each pair should be formatted as: (Drug Name, MeSH ID).
Return the generated pairs in a structured JSON format where the key is "Pairs" and
the value is a list of pairs (with the format: [(Drug Name, MeSH ID), (Drug Name,
MeSH ID)]).
Input: Text: {{text}}
List_drugs related with coronary diseases and strokes: {{drug_list}}
The final JSON is:
```

A.1.5 Prompt for Review

```
Given the following sentence: {{text}}, identify whether the entity {{entity}}
represents a disease and if it correctly matches the category {{description}}.
Respond based on the following guidelines:
1. If the entity {{entity}} is not a disease, select option "1".
2. If the entity {{entity}} is a disease, but does not fit the category
{{description}}, select option "2".
3. If the entity {{entity}} is a disease and matches the category {{description}},
select option "3".
Return the response in a structured JSON format where the key is Result and the
value is the selected option as an integer.
Input:
Entity: {{entity}}
Category: {{description}}
Response:
```

A.1.6 Prompt for CCD–CCD Pair Generation

```
Given the following sentence, extract all relevant relations between diseases ONLY.
IMPORTANT RULES:
- Consider ONLY entities explicitly marked with <diseaseidx>.
- DO NOT create new entities.
- Extract ONLY relations between diseases.
- Return a structured JSON response with the key "Pairs" and a list of list of pairs
of entities in the form:
- JUST A EXAMPLE: ['<disease4>', '<disease5>'], ['<disease10>', '<disease11>']
- ONLY MENTION THE DISEASE BY THE TAG <diseaseidx>, where the idx MUST BE THE SAME
AS IN THE INPUT.
Sentence to process:
sentence: {{sentence}}
List of diseases ONLY to consider:list_diseases
JSON Response:
```

A.1.7 Prompt for CCDt–CCDt Pair Generation

```
Given the following sentence, extract all relevant relations between drugs ONLY.
IMPORTANT RULES:
- Consider ONLY entities explicitly marked with <drugidx>.
- DO NOT create new entities.
- Extract ONLY relations between drugs.
- Return a structured JSON response with the key "Pairs" and a list of list of pairs
of entities in the form:
- JUST A EXAMPLE: ['<drug4>', '<drug5>'], ['<drug10>', '<drug11>']
- ONLY MENTION THE DRUG BY THE TAG <drugidx>, where the idx MUST BE THE SAME AS IN
THE INPUT.
Sentence to process:
sentence: {{sentence}}
List of drugs ONLY to consider: {{list_drugs}}
JSON Response:
```

A.1.8 Prompt for CDD–CDDt Pair Generation

```
Given the following sentence, extract all relevant relations between diseases and
drugs ONLY.
IMPORTANT RULES:
- Consider ONLY entities explicitly marked with <diseaseidx> and <drugidx>.
- DO NOT create new entities.
- Extract ONLY relations between diseases and drugs.
- Return a structured JSON response with the key "Pairs" and a list of list of pairs
of entities in the form:
- JUST A EXAMPLE: ['<disease4>', '<drug5>'], ['<drug10>', '<drug11>']
- ONLY MENTION THE DRUG BY THE TAG <drugidx> AND THE DISEASE BY THE TAG
<diseaseidx>, where the idx MUST BE THE SAME AS IN THE INPUT.
Sentence to process:
sentence: {{sentence}}
List of diseases ONLY to consider: {{list_diseases}}
List of drugs ONLY to consider: {{list_drugs}}
JSON Response:
```

A.1.9 Flexible Prompt for Relation Classification

```
Given the following context and a pair of medical entities, determine the most
accurate relationship between them.
Instructions:
- The relation label must be as precise as possible, using a maximum of five words.
- The relation label should best describe the relationship between the two entities.
- Only consider interactions relevant to the provided context.
- Ensure that the extracted relationship is meaningful in a medical or biomedical
setting.
- The entities in the triplet MUST be exactly as provided in the pair, including
their respective tags with the correct numbering, as presented in the context and
entity pair.
- The response MUST be in a structured JSON format, with the key "Label"
and the value as a triplet: ["<tagidx>Entity1</tagidx>", "relation",
"<tagidx>Entity2</tagidx>"].
IMPORTANT RULE:
- When mentioning the disease or drug, use the tag <diseaseidx> or <drugidx>
respectively, where the idx MUST BE THE SAME NUMBER AS IN THE INPUT.
Input Data:
- Context: {{new_query}}
- Entity Pair: {{entity1_name}}, {{entity2_name}}
JSON Response:
```

A.1.10 Prompt for Relation Classification

```
Given the following context and a pair of medical entities, determine the most
accurate relationship between them.
Instructions:
- The relation label must be as precise as possible based on the context.
- The relation label should be one of the following: {{relation_list}}.
- The entities in the triplet MUST be exactly as provided in the pair, including
their respective tags with the correct numbering, as presented in the context and
entity pair.
Relation Descriptions:
- '{{label1}}': {{description1}}
- '{{label2}}': {{description2}}
...
- '{{labelN}}': {{descriptionN}}
IMPORTANT RULE:
- The response MUST be in a structured JSON format, with the key "Label" and the
value as a triplet: ["<tag>Entity1</tag>", "relation", "<tag>Entity2</tag>"].
- When mentioning the disease or drug, use the tag <diseaseidx> or <drugidx>
respectively, where the idx MUST BE THE SAME NUMBER AS IN THE INPUT.
Input Data:
- Context: {{context}}
- Entity Pair: {{entity1}}, {{entity2}}
JSON Response with the key "Label" and the value as a triplet:
```

A.1.11 LLaMIC Implementation on MIMIC-IV

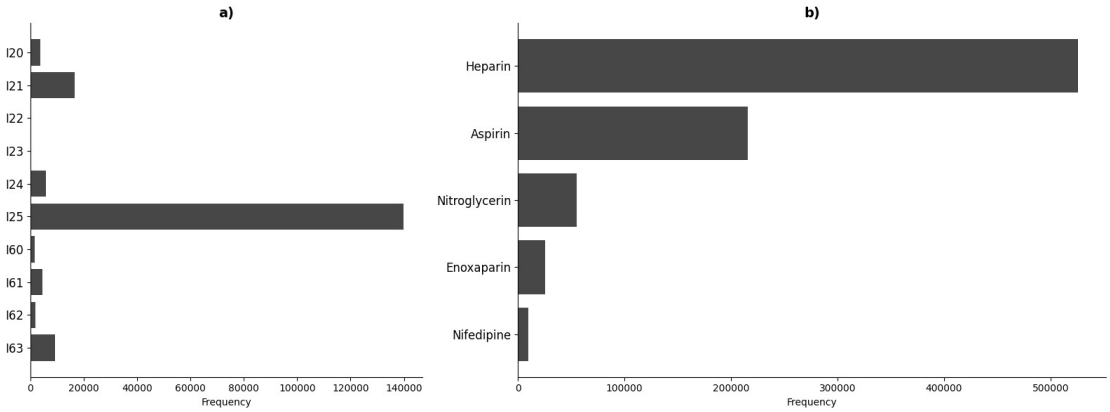

Figure A.1: Distribution of CCDs (a) and top five CCDts (b) from structured data.

ICD-10	Description
I20–I25	Ischaemic heart diseases
I20	Angina pectoris
I21	Acute myocardial infarction
I22	Subsequent myocardial infarction
I23	Certain current complications following acute myocardial infarction
I24	Other acute ischaemic heart diseases
I25	Chronic ischaemic heart disease
I60–I69	Cerebrovascular diseases
I60	Subarachnoid haemorrhage
I61	Intracerebral haemorrhage
I62	Other nontraumatic intracranial haemorrhage
I63	Cerebral infarction
I64	Stroke, not specified as haemorrhage or infarction

Table A.1: List of ICD-10 codes and their corresponding disease descriptions.

MeSH	Description
D001241	Aspirin
D010869	Pindolol
D000077550	Ivabradine
D017984	Enoxaparin
D000241	Adenosine
D000077466	Tirofiban
D005437	Flucytosine
D000077542	Eptifibatide
D009543	Nifedipine
D000255	Adenosine Triphosphate
D007548	Isosorbide Dinitrate
D005996	Nitroglycerin
D012977	Sodium Nitrite
D000077785	Tenecteplase
D000077764	Dronedarone
D000077486	Ticagrelor
D000077425	Fondaparinux
D000068799	Prasugrel Hydrochloride
D006493	Heparin
D000069458	Ranolazine
D005297	Ferrozine

Table A.2: List of therapeutic drugs annotated in the corpus with corresponding MeSH identifiers.

Name	Count	%
CAD	2613	19.88
NSTEMI	699	5.32
stroke	575	4.37
MI	514	3.91
ACS	389	2.96
coronary artery disease	327	2.49
CVA	322	2.45
ischemia	296	2.25
STEMI	222	1.69
demand ischemia	221	1.68
Total	6178	47.00

Table A.3: Top 10 most frequent CCD names in the supervised corpus.

Name	Count	%
heparin	1318	22.84
aspirin	1073	18.59
ASA	938	16.25
Aspirin	311	5.39
Heparin	218	3.78
nifedipine	119	2.06
enoxaparin	113	1.96
Lovenox	89	1.54
heparin gtt	87	1.51
Total	4572	79.22

Table A.4: Top 10 most frequent CCDts name mentions in the supervised corpus

Figure A.2: Semantic clustering of triplet relations among CCD and CCDt.

Appendix B

Supplementary Documents

The following supplementary materials will be made available after the thesis defense.

Return to project overview

The project has passed all automatic checks.

Model | Credentialed Access

[Preview]: LLaMIC Modules: Fine-Tuned Large Language Models for Clinical Entity Detection and Relation Extraction

Diogo Mataloto , **Maria Fernandes** , **Francisco Couto**

Published: [dd/mm/yyyy] - Version: 1.0.0

When using this resource, please cite: (show more options)
Mataloto, D., Fernandes, M., & Couto, F. (2025). LLaMIC Modules: Fine-Tuned Large Language Models for Clinical Entity Detection and Relation Extraction (version 1.0.0). *PhysioNet*. RRID:SCR_007345. https://doi.org/10.13026/*****

Please include the standard citation for PhysioNet: (show more options)
Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220. RRID:SCR_007345.

Abstract

With the digitalization of healthcare, the increasing availability of electronic health records (EHRs) has enabled the development of robust algorithms capable of inferring patient diagnoses and treatments from unstructured clinical notes.

We present LLaMIC modules, a suite of LLaMA-based large language models fine-tuned on a manually and semi-automatically annotated corpus of cardiovascular and cerebrovascular diseases (CCDs), their associated therapeutic drugs (CCDts) and binary relations between these entities derived from MIMIC-IV clinical notes. The models were trained to recognize entities and extract binary relations between diseases and drugs, leveraging annotations standardized with International Classification of Diseases 10 (ICD-10) and MeSH identifiers. These fine-tuned checkpoints are designed to be directly implemented within the LLaMIC pipeline (https://github.com/lasigeBioTM/LLAMIC).

Background

Building on recent studies [1,2,3,4], large language models (LLMs) have shown strong performance in clinical NLP tasks such as question answering, named entity recognition, knowledge-guided reasoning, and automated de-identification. However, there is still limited literature on the use of open-source LLMs applied to real-world medical data for information extraction tasks, particularly in identifying CCDs and their corresponding therapeutic drugs.

Central to these approaches are two critical components: Named Entity Recognition and Linking (NER/NEL), which identify biomedical entities and associate them with standardized knowledge bases, and Relation Extraction (RE), which maps interactions between these entities.

Model Description

The LLaMIC suite comprises six fine-tuned LLaMA3.1 8B foundation models, each dedicated to a specific subtask in the LLaMIC pipeline. For entity detection, four models are used: two for NER — one specialized in detecting disease mentions and another for drug mentions, each trained to extract as many target entities as possible from a clinical note — and two for NEL, mapping detected entities to standardized terminologies (ICD-10 for CCDs and MeSH for CCDts).

The remaining two models address binary relation extraction: a Pair Generation model, optimized to identify candidate pairs with potential relations, and a Relation Labeling model, trained to assign a specific relation label to each pair.

All models were fine-tuned on a semi-automatically annotated subset of MIMIC-IV clinical notes (https://physionet.org/content/), split into training, validation, and test sets using an 8:1:1 ratio.

Technical Implementation

For NER, fine-tuning was performed using a learning rate of 5×10^{-5}, with a training batch size of 4 and an evaluation batch size of 2. Training was carried out for 10 epochs using the AdamW 8-bit optimizer with a weight decay of 0.01. To adapt the model efficiently, LoRA was employed, with the following configuration: a rank (r) of 8, a LoRA alpha value of 32, and a dropout rate of 0.05.

For RE, the same hyperparameters and LoRA configuration were applied, except for the following differences: the training and evaluation batch sizes were reduced to 1, combined with gradient accumulation over 8 steps; mixed-precision training (fp16) and gradient checkpointing were enabled to improve efficiency; and a cosine learning rate scheduler was employed, with a warmup ratio of 0.05 and a maximum gradient norm of 0.3.

Model training and monitoring were conducted using Weights & Biases (W&B) [5].

Installation and Requirements

Download the LLaMIC models to the corresponding models directory in the LLaMIC script (https://github.com/lasigeBioTM/LLAMIC).

Usage Notes

Given LLaMIC models are large language model architecture it's important to consider the computational resources required. We advise using a GPU with at least 16GB of memory.

Ethics

All authors completed CITI training, were credentialed by PhysioNet, and signed the Data Use Agreement to access the de-identified MIMIC-IV notes.

Large Language Models have been shown to be susceptible to the leakage of sensitive information. Users must not attempt to use these models to extract or reproduce any confidential data. All models were trained exclusively on MIMIC-IV data.

The LLaMIC modules comply with the terms and conditions outlined in the META LLAMA 3 Community License Agreement.

Conflicts of Interest

The authors have no conflicts of interest to declare.

References

1. Gilson, Aidan, et al. “How does CHATGPT perform on the United States Medical Licensing Examination (USMLE)? the implications of large language models for medical education and knowledge assessment.” JMIR Medical Education, vol. 9, 8 Feb. 2023, https://doi.org/10.2196/45312.
2. Hu, Yan, et al. “Improving large language models for clinical named entity recognition via prompt engineering.” Journal of the American Medical Informatics Association, vol. 31, no. 9, 27 Jan. 2024, pp. 1812–1820, https://doi.org/10.1093/jamia/ocad259.
3. Su, X., Wang, Y., Gao, S., Liu, X., Giunchiglia, V., Clevert, D. A., & Zitnik, M. (2024). KGARevion: an AI agent for knowledge-intensive biomedical QA. arXiv preprint arXiv:2410.04660.
4. Altalla', B., Abdalla, S., Altamimi, A., Bitar, L., Al Omari, A., Kardan, R., & Sultan, I. (2025). Evaluating GPT models for clinical note de-identification. Scientific reports, 15(1), 3852. https://doi.org/10.1038/s41598-025-86890-3
5. Biewald, L. (2020). Experiment tracking with Weights and Biases [Software]. Weights & Biases. https://www.wandb.com/

Parent Projects

LLaMIC Modules: Fine-Tuned Large Language Models for Clinical Entity Detection and Relation Extraction was derived from:

- MIMIC-IV v3.1
- MIMIC-IV-Note: Deidentified free-text clinical notes v2.2

Please cite them when using this project.

Access

Access Policy:
Only credentialed users who sign the DUA can access the files.

License (for files):
PhysioNet Credentialed Health Data License 1.5.0

Data Use Agreement:
PhysioNet Credentialed Health Data Use Agreement 1.5.0

Required training:
CITI Data or Specimens Only Research

Discovery

DOI (version 1.0.0):
https://doi.org/10.13026/*****

DOI (latest version):
https://doi.org/10.13026/*****

Corresponding Author

Diogo Mataloto, LASIGE Computer Science and Engineering Research Centre, fc62747@alunos.fc.ul.pt

Files

Folder Navigation: <base>

Name	Size	Modified
LLaMIC_Entities		
LLaMIC_RE		

MIT Laboratory for Computational Physiology

National Institute of Biomedical Imaging and Bioengineering (NIBIB) under NIH grant number R01EB030362

Navigation

Discover Data

The project has passed all automatic checks.

Database Credentialed Access

[Preview]: LLaMIC Corpus: Annotated Cardiovascular and Cerebrovascular Entities, Therapeutic Drugs and Binary Relations from MIMIC-IV Clinical Notes

Diogo Mataloto , **Maria Fernandes** , **Francisco Couto**

Published: [dd/mm/yyyy] - Version: 1.0.0

When using this resource, please cite: (show more options)
Mataloto, D., Fernandes, M., & Couto, F. (2025). LLaMIC Corpus: Annotated Cardiovascular and Cerebrovascular Entities, Therapeutic Drugs and Binary Relations from MIMIC-IV Clinical Notes (version 1.0.0). *PhysioNet*. RRID:SCR_007345. https://doi.org/10.13026/*****

Please include the standard citation for PhysioNet: (show more options)
Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220. RRID:SCR_007345.

Abstract

With the digitalization of healthcare, the increasing availability of electronic health records (EHRs) has enabled the development of robust algorithms capable of inferring patient diagnoses and treatments from unstructured clinical notes, as well as constructing knowledge graphs from this information.

This project introduces the LLaMIC corpus, a manually and semi-automatically annotated dataset of cardiovascular and cerebrovascular diseases (CCDs) their associated therapeutic drugs (CCDt) and binary relations between these entities, derived from the MIMIC-IV clinical notes. The annotation pipeline combines a Biomedical Entity Annotator (BENT)[1], a large language model (LLaMA) and manual correction (conducted by a master's student in bioinformatics and computational biology) to identify and standardize eight prevalent CCD entities using ICD-10 codes, alongside five drug entities mapped to MeSH identifiers. This corpus was developed as part of a master's thesis to train and evaluate the LLaMIC framework (LLaMIC: LLaMA Model Applied to MIMIC, https://github.com/lasigeBioTM/LLAMIC).

Background

Cardiovascular and Cerebrovascular Diseases (CCDs) remain a leading global health concern. According to the World Health Organization, an estimated 17.90 million people died from cardiovascular diseases in 2019 [2]. These conditions are closely associated with multiple risk factors, including diabetes, hyperlipidemia, hypertension, smoking, obesity, family history of heart disease, and chronic medication use, as outlined by the WHO. This trend poses significant public health and economic challenges, underscoring the urgency of disseminating knowledge about CCD prevention, diagnosis, and treatment. Enhanced access to such information may mitigate the impact of CCDs, thereby reducing patient morbidity and mortality [3].

Electronic Health Records (EHRs) represent digital repositories of patient medical information that integrate both structured data—such as demographics, laboratory results, medications, diagnostic codes, and procedural information—and unstructured data, including clinical notes. The exponential growth, standardization, and authenticity of EHR data have unlocked new

opportunities for developing automated decision-support systems at the point of care, as well as for advancing clinical and translational research [4]. A significant proportion of valuable information within EHRs resides in unstructured clinical notes [5]. The vast volume and complexity of such data render manual analysis impractical for achieving these advancements [6].

To address this challenge, Natural Language Processing has emerged as an indispensable tool, enabling the automatic extraction of meaningful information from unstructured EHR narratives [7]. Central to these approaches are two critical components: Named Entity Recognition and Linking (NER/NEL), which identify biomedical entities and associate them with standardized knowledge bases, and Relation Extraction (RE), which maps interactions between these entities. Having supervised corpora for each of these tasks is essential for the development of these NLP systems.

Methods

This corpus targets eight prevalent cardiovascular and cerebrovascular diseases (CCDs), identified by the following ICD-10 codes: coronary heart disease (I20–I25), hemorrhagic stroke (I60–I62), cerebral infarction (I63), and unspecified stroke (I64) [8]. For the therapeutic drug targets associated with these CCDs, we selected the six most frequently mentioned drugs in MIMIC-IV clinical notes that exhibit a therapeutic link with the ICD-10 targets, based on the 'Drug-to-Disease Mapping with ICD Identifiers' dataset from the Therapeutic Target Database (TTD). In addition, a supervised corpus of binary relations between CCDs and the corresponding therapeutic drugs was created.

The annotation process of the supervised corpus was conducted on a subset of the MIMIC-IV-Note corpus using a semi-automatic pipeline: an initial automatic pre-annotation of entities (BENT model and the LLaMIC model—based on LLaMA without task-specific fine-tuning), followed by weak manual correction performed by a master's student in Bioinformatics and Computational Biology.

Data Description

We created a dataset of clinical notes comprising two subsets: (1) supervised, generated through a semi-automatic pipeline, and (2) unsupervised, produced by our LLaMIC model after applying an LLM fine-tuned on the supervised corpus. The only exception is the unsupervised CCDt corpus, for which the BENT model was employed, as it outperformed the LLaMIC-based approach. The statistics for the supervised and the unsupervised dataset are summarized in the table below:

Entity Type	Train Set	Validation Set	Test Set	Predicted Set
CCDs	8,059	1,727	1,728	298,879
CCDts	4,676	584	584	304,546
RE	641	78	80	44,848

In addition, a supervised corpus of binary relations between CCDs and the corresponding therapeutic drugs was created to support relation extraction tasks. field "id" (integer) corresponds to the "HADM_ID" (integer) from the NOTEEVENTS table in MIMIC-III v1.4. The field "documents" (string) corresponds to the "TEXT" (string) from NOTEEVENTS, preprocessed by removing extra spaces. In the supervised corpus, the field "entities" (list) contains the entities identified in the document, where each entity record includes: "entity" (the mention in text), "icd" or "mesh" (the associated code), and "start" and "end" (character offsets in the preprocessed text). Note that the training and validation sets do not include the "start" and "end" offsets, as these are not required for training the LLaMIC model

Supervised Output:

For entities corpus:

```
id,documents,entities
23434534," year old woman with a history of CAD, HTN, HLD, afib on dispyramide status post ___ ___ ...",
         "[{'entity': 'CAD', 'icd': 'I25', 'start': 34, 'end': 37}]"
(...)
```

For relations corpus:

The index at the end of the IDs corresponds to the segment number of the clinical note resulting from the https://github.com/lasigeBioTM/LLAMIC/blob/main/mimic-implementation/semi_automatic_re_correction.ipynb processing. An index of zero indicates that it is the first segment of the discharge summary.

```
id,documents,entities
23434534_0,"Patient with <disease1>CAD</disease1> treated with <drug1>nitroglycerin</drug1> ...",
         "[["<disease1>CAD</disease1>", "indicated", "<drug1>nitro</drug1>"]]"
(...)
```

Unsupervised Output:

For entities corpus:

For storage efficiency, the text of the clinical notes in the unsupervised output is stored in a CSV file generated by mimic_preprocessing.py. The id column serves as the key linking the two tables, while the clinical note text is contained in the documents column. The preprocessing script is available at the https://github.com/lasigeBioTM/LLAMIC/blob/main/mimic-implementation/mimic_preprocessing.py

```
{
    "results": [
        {
            "id": "28253766",
            "entities": [{'entity': 'aspirin', 'mesh': 'D001241', 'start': 1405, 'end': 1412}]
        },
        (...)
    ]
}
```

For relations corpus:

In contrast to the entities corpus, the CSV file containing the clinical notes for the relations corpus is already provided as unsupervised.csv, with the columns id and document—similar to the supervised tables—but without the entities column.

Usage Notes

Since this corpus of annotated deidentified clinical notes contains original healthcare data that includes protected health information (PHI) under the Health Insurance Portability and Accountability Act of 1996 (HIPAA) and can be linked to the MIMIC-IV database, access is restricted to individuals who have met all requirements for accessing MIMIC-IV data.

Ethics

All authors completed CITI training, were credentialed by PhysioNet, and signed the Data Use Agreement to access the deidentified MIMIC-IV notes.

Conflicts of Interest

The authors have no conflicts of interest to declare

References

1. Ruas, P., & Couto, F. M. (2022). NILINKER: Attention-based approach to NIL entity linking. Journal of Biomedical Informatics, 132, 104137. https://doi.org/10.1016/j.jbi.2022.104137
2. World Health Organization. (2024). Health topics: Cardiovascular diseases. Retrieved November 27, 2024, from https://www.who.int/health-topics/cardiovascular-diseases
3. Fuster, V., Kelly, B. B., & Vedanthan, R. (2011). Global cardiovascular health. Journal of the American College of Cardiology, 58(12), 1208-1210
4. Woldemariam, M. T., & Jimma, W. (2023). Adoption of electronic health record systems to enhance the quality of healthcare in low-income countries: A systematic review. BMJ Health & Care Informatics, 30(1), e100704. https://doi.org/10.1136/bmjhci-2022-100704
5. Sheikhalishahi, S., Miotto, R., Dudley, J. T., Lavelli, A., Rinaldi, F., & Osmani, V. (2019). Natural language processing of clinical notes on chronic diseases: Systematic review. JMIR Medical Informatics, 7(2), e12239
6. Hariri, R. H., Fredericks, E. M., & Bowers, K. M. (2019). Uncertainty in big data analytics: Survey, opportunities, and challenges. Journal of Big Data, 6, 44. https://doi.org/10.1186/s40537-019-0206-3
7. Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: Towards better research applications and clinical care. Nature Reviews Genetics, 13(6), 395-405. https://doi.org/10.1038/nrg3208
8. Centers for Disease Control and Prevention. (2024). Cardiovascular disease — Chronic Disease Indicators. Retrieved December 12, 2024, from https://www.cdc.gov/cdi/indicator-definitions/cardiovascular-disease.html

Parent Projects

LLaMIC Corpus: Annotated Cardiovascular and Cerebrovascular Entities, Therapeutic Drugs and Binary Relations from MIMIC-IV Clinical Notes was derived from:

- MIMIC-IV v3.1
- MIMIC-IV-Note: Deidentified free-text clinical notes v2.2

Please cite them when using this project.

Access

Access Policy:
Only credentialed users who sign the DUA can access the files.

License (for files):
PhysioNet Credentialed Health Data License 1.5.0

Data Use Agreement:
PhysioNet Credentialed Health Data Use Agreement 1.5.0

Required training:
CITI Data or Specimens Only Research

Discovery

DOI (version 1.0.0):
https://doi.org/10.13026/*****

DOI (latest version):
https://doi.org/10.13026/*****

Corresponding Author

Diogo Mataloto, LASIGE Computer Science and Engineering Research Centre, fc62747@alunos.fc.ul.pt

Files

Folder Navigation: <base>

Name	Size	Modified
LLaMIC_Entities		
LLaMIC_RE		