A Novel way to extract entities from electronic health record in the era of LLMs

Daniele Cecca[1*]

[1*]Computer Science, University Milano Statale, Università Milano Bicocca, Milan,Italy .

Abstract

The aim of this study is to develop a novel system capable of performing Named Entity Recognition (NER) on electronic health records (EHRs).
We explore two distinct approaches:
1. Direct use of Large Language Models (LLMs) to perform NER.
2.A hybrid method, where we fine-tune a **BERT-based** model on a custom NER dataset generated by an LLM. The extracted entities are then used for classification, either by an LLM or a traditional classifier.
We compare these two methodologies to assess whether, in the age of LLMs, alternative approaches still offer practical advantages. This question is especially relevant given the ongoing debate in the literature between end-to-end LLM solutions and more modular, specialized architectures[1][2].

Keywords: NER, LLM, BERT, HER

1 Introduction

In modern healthcare, clinicians often spend a considerable amount of time on administrative tasks, such as transcribing information from patient records (typically in PDF format) into structured formats like Excel. This time-consuming process not only reduces clinical efficiency but also detracts from direct patient care.

To mitigate this issue, we propose an automated system for extracting structured entity data from clinical documents. We explore two complementary approaches.

In the first approach, we leverage the reasoning capabilities of state-of-the-art LLMs to directly perform entity extraction. Given the proliferation of recent LLMs, we also conduct a comparative analysis to identify the most effective model for our use case.

In the second approach, we fine-tune a domain-specific transformer model—**Bio_ClinicalBERT**—which is pre-trained on biomedical texts and tailored for the medical domain. One key challenge in this approach is the scarcity of annotated medical data. To overcome this, we first create a custom NER dataset in IOB format using an LLM. This dataset is then used to fine-tune the model for the NER task.

Through this dual-method analysis, we aim to understand the trade-offs between relying entirely on LLMs and using more traditional, fine-tuned models within the healthcare context.

2 Methodolgies

2.1 Name entity recognition - NER

As written in [3], a named entity is, roughly speaking, anything that can be referred to with a proper name: a person, a location, or an organization. The task of Named Entity Recognition (NER) is to identify spans of text that constitute proper names and to tag the type of the entity. Four entity tags are most common: **PER** (person), **LOC** (location), **ORG** (organization), and **GPE** (geo-political entity).

However, the term *named entity* is commonly extended to include elements that are not entities per se, such as dates, times, temporal expressions, and numerical expressions like prices. The standard approach to sequence labeling for a span-recognition problem like NER is **BIO tagging** (also known as **IOB**). This method allows NER to be treated as a word-by-word sequence labeling task, using tags that capture both the boundary of the entity and its type.

Tag	**Description**
B-XXX	Beginning of an entity of type XXX
I-XXX	Inside (continuation) of an entity of type XXX
O	Outside any named entity

In our case, since the total number of distinct entities was 74 and we lacked the expertise to group them into broader macro-categories, we decided to use a single entity label: **TARGET**. Thus, our tag set consists of the following labels:

- B-TARGET
- I-TARGET
- O

The IOB notation is particularly useful because it enables the extraction of both categorical and numerical values that quantify the different entities.

3 Large Language Model - LLM

Language Model (LM). A language model estimates the probability of a sequence of tokens $w_1, w_2, \ldots, w_n$ by modeling the joint probability:

$$P(w_1, w_2, \ldots, w_n) = \prod_{t=1}^{n} P(w_t \mid w_1, \ldots, w_{t-1})$$

Traditional models include **n-gram** models and **RNNs**. Modern approaches use deep neural architectures like **Transformers** to capture complex language patterns. A large language model is a Transformer-based neural network with billions of parameters, trained on massive text corpora.

A key innovation in LLMs is the **self-attention mechanism**, which allows each token to attend to all others in the sequence. Given token embeddings, self-attention computes:

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

where Q, K, and V are the query, key, and value matrices derived from input embeddings, and d_k is the key dimensionality.

Masking in Attention. is crucial for controlling the flow of information in attention computations. **Autoregressive models** (e.g., GPT) use causal masks to prevent the model from attending to future tokens. This ensures that the prediction of token w_t only depends on $w_1, \ldots, w_{t-1}$. The mask sets attention scores for future tokens to $-\infty$, effectively zeroing them out after the softmax:

$$\text{MaskedAttention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M\right) V$$

where M is a matrix with $-\infty$ in positions corresponding to illegal (future) attention. This ensures that tokens do not have access to information they shouldn't have during generation.

Masking enables autoregressive models to generate text step-by-step without violating causality, ensuring each token is predicted only based on its past context.

3.1 Bidirectional Encoder Representations from Transformers - BERT

BERT is a Transformer-based model designed to capture deep contextual relationships in text by considering both the left and right context of a word. Unlike traditional models such as RNNs, which process text sequentially, BERT utilizes a bidirectional approach, allowing it to access information from both directions simultaneously.

BERT's architecture consists of multiple layers of bidirectional self-attention, enabling it to understand the full context of a word in a sentence. It is pre-trained on a large corpus using two primary objectives:

- **Masked Language Model (MLM)**: Random words in a sentence are masked, and the model is trained to predict them using the surrounding context. This helps BERT learn bidirectional context by forcing it to infer missing words from both the left and right.

- **Next Sentence Prediction (NSP)**: The model is trained to predict whether a given pair of sentences appear consecutively in the text. This helps BERT understand sentence relationships and context across multiple sentences.

The core idea behind BERT's success is its ability to generate highly contextualized word representations by processing input in parallel, rather than sequentially. This parallelization leads to faster training and a more robust understanding of complex language structures.

In application, BERT can be fine-tuned for specific tasks like question answering, named entity recognition (NER), and sentiment analysis by adding task-specific layers on top of the pre-trained model. This approach allows BERT to achieve state-of-the-art performance on a variety of natural language processing (NLP) benchmarks.

4 Dataset

4.1 Text extraction - OCR

The dataset was provided by *Fondazione Alfieri* and consisted of clinical folders from the year 2019. These folders contained scanned documents from patients' electronic health records (EHRs). Since the documents were available only as images, our first task was to extract the textual content. To this end, we chose to convert the extracted content into Markdown format, as large language models (LLMs) generally perform better with structured input.

In this work, we focused primarily on the *discharge letters*. However, in future work, we plan to extend our analysis to include additional document types within the clinical folders, as well as more recent data spanning from 2020 to the present.

To convert the scanned documents into Markdown, we used the `Marker` library. Although Marker supports LLM-based structuring, we opted not to enable this functionality. Our objective was not to achieve perfectly structured output but to extract sufficiently readable content to evaluate our proposed methods. Due to limited computational resources, we also relied on the default OCR model provided by Marker.

We processed a total of 291 documents and identified the relevant discharge letters using keyword matching. Specifically, we searched for the phrase *"RELAZIONE CLINICA ALLA DIMISSIONE – DEFINITIVA"*, following guidance from a medical expert at the foundation.

While the final extracted texts were not fully structured, they proved adequate for evaluating and testing our proposed solutions.

Based on the extracted text, we defined a set of clinical entities to be identified, suggested by the expert. These are shown in the table below.

4.2 Dataset LLM

In the case of the first solution, we use directly the extracted text as dataset so we have 291 inputs.

Table 1 Structured Fields Extracted from Discharge Letters

Campo	Tipo	Note
n_cartella	Number	stringa
data_ingresso_cch	Date	
data_dimissione_cch	Date	
nome	Text	
cognome	Text	
sesso	Categorical (M/F)	categoria ordinata
numero di telefono	Text	
età al momento dell'intervento	Number	continua
data_di_nascita	Date	
Diagnosi	Text	
Anamnesi	Text	
Motivo ricovero	Text	
classe_nyha	Categorical (1–4)	valori 1, 2, 3, 4
angor	Boolean	0/1
STEMI/NSTEMI	Boolean	0/1
scompenso_cardiaco_nei_3_mesi_precedenti	Boolean	0/1
fumo	Categorical (0/1/2)	3 categorie
diabete	Boolean	0/1
ipertensione	Boolean	0/1
dislipidemia	Boolean	0/1
BPCO	Boolean	0/1
stroke_pregresso	Boolean	0/1
TIA_pregresso	Boolean	0/1
vasculopatiaperif	Boolean	0/1
neoplasia_pregressa	Boolean	0/1
irradiazionetoracica	Boolean	0/1
insufficienza_renale_cronica	Boolean	0/1
familiarita_cardiovascolare	Boolean	0/1
limitazione_mobilita	Boolean	0/1
endocardite	Boolean	0/1
ritmo_all_ingresso	Categorical (0/1/2)	3 categorie
fibrillazione_atriale	Categorical (0/1/2)	3 categorie
dialisi	Boolean	0/1
elettivo_urgenza_emergenza	Categorical (0/1/2)	3 categorie
pm	Boolean	0/1
crt	Boolean	0/1
icd	Boolean	0/1
pci_pregressa	Boolean	0/1
REDO	Boolean	0/1
Anno REDO	Date	
Tipo di REDO	Text	
Terapia	Text	
lasix	Boolean	0/1
lasix_dosaggio	Number	continua
nitrati	Boolean	0/1
antiaggregante	Boolean	0/1
dapt	Boolean	0/1
anticoagorali	Boolean	0/1
aceinib	Boolean	0/1
betabloc	Boolean	0/1
sartanici	Boolean	0/1
caantag	Boolean	0/1
esami_all_ingresso	Text	
Decorso_post_operatorio	Text	
IABP/ECMO/IMPELLA	Boolean	0/1
Inotropi	Boolean	0/1
secondo_intervento	Boolean	0/1
Tipo_secondo_intervento	Text	
II_Run	Boolean	0/1
Causa_II_Run_CEC	Text	
LCOS	Boolean	0/1
Impianto_PM_post_intervento	Boolean	0/1
Stroke_TIA_post_op	Boolean	0/1
Necessità_di_trasfusioni	Boolean	0/1
IRA	Boolean	0/1
Insufficienza_respiratoria	Boolean	0/1
FA_di_nuova_insorgenza	Boolean	0/1
Ritmo_alla_dimissione	Categorical (0/1/2)	3 categorie
H_Stay_giorni	Number	continua
Morte	Boolean	0/1
Causa_morte	Text	
data_morte	Date	
esami_alla_dimissione	Text	
terapia_alla_dimissione	Text	

4.3 Dataset BERT

For the creation of the dataset used to train the BERT model, we began by extracting the raw text from the clinical discharge letters. The text was then pre-processed by removing non-informative markdown elements, embedded images, and special symbols deemed irrelevant for the task. Afterward, the text was normalized to ensure consistency in formatting and structure.

Following normalization, the text was segmented into individual phrases using the newline character (‘\n‘) as a delimiter. This resulted in a total of approximately **21,051** phrases.

To generate the dataset in **IOB format** for the Named Entity Recognition (NER) task, we employed a Large Language Model (LLM). Specifically, we utilized the **LLaMA 3-70B** model, which is publicly available via the **Together AI** platform. The model was prompted with a carefully designed instruction to perform IOB tagging, and all 21,051 phrases were processed accordingly. There are already some studies [4] in which datasets were augmented using LLMs by replacing words in sentences with synonyms. However, in our case, we need to create the dataset from scratch, so this approach is not applicable—at least in this phase.

```
promp_base='''Sei un medico e voglio che assegni ad ogni parola una label seguendo il formato IOB (
     Inside-Outside-Beginning),
usato per i task di Named Entity Recognition (NER), alla seguente frase presa da una lettera di
     dimissioni.

L' **unica label** da assegnare è:
**TARGET**

dove le entità target sono le seguenti:

### Mappa delle entità e tipi

| Nome                                        | Descrizione
                                                        |
|-----------------------------------------------------------------------------------|
| n_cartella                                  | Numero identificativo univoco assegnato alla cartella
     clinica del paziente.                             |
| data_ingresso_cch                           | Data in cui il paziente è stato ricoverato presso il
     reparto di Cardiochirurgia.                       |
| data_dimissione_cch                         | Data in cui il paziente è stato dimesso dal reparto di
      Cardiochirurgia.                               |
| nome                                        | Nome proprio del paziente.
                                                       |
| cognome                                     | Cognome del paziente.
                                                       |
| sesso                                       | Sesso biologico del paziente (M = Maschio, F = Femmina
     ).                                                |
| numero di telefono                          | Recapito telefonico del paziente o di un contatto di
     riferimento.                                      |
| età al momento dell'intervento              | Età del paziente calcolata alla data dellintervento
     chirurgico.                                       |
| data_di_nascita                             | Data di nascita del paziente.
                                                       |
| Diagnosi                                    | Diagnosi principale alla base dell'indicazione
     chirurgica.                                             |
| Anamnesi                                    | Anamnesi patologica remota e prossima, utile per la
     valutazione del rischio operatorio.                |
| Motivo ricovero                             | Indicazione clinica per il ricovero in Cardiochirurgia
     .                                                 |
| classe_nyha                                 | Classe funzionale NYHA per scompenso cardiaco (I-IV),
     definisce la gravità dei sintomi.               |
```

```
| angor                                   | Presenza di angina pectoris (dolore toracico di
    origine ischemica).                                 |
| STEMI/NSTEMI                            | Presenza di infarto miocardico acuto con/senza
    sopraslivellamento del tratto ST.                    |
| scompenso_cardiaco_nei_3_mesi_precedenti  | Episodi di scompenso cardiaco documentati nei 3 mesi
    precedenti lintervento.                       |
| fumo                                    | Abitudine al fumo (0 = mai fumato, 1 = ex-fumatore, 2
    = fumatore attivo).                           |
| diabete                                 | Presenza di diabete mellito noto.
                                                  |
| ipertensione                            | Presenza di ipertensione arteriosa.
                                                  |
| dislipidemia                            | Presenza di dislipidemia (colesterolo e/o trigliceridi
     elevati).                                    |
| BPCO                                    | Presenza di broncopneumopatia cronica ostruttiva.
                                                  |
| stroke_pregresso                        | Precedente episodio di ictus cerebrale ischemico o
    emorragico.                                      |
| TIA_pregresso                           | Episodio pregresso di attacco ischemico transitorio (
    TIA).                                          |
| vasculopatiaperif                       | Malattia vascolare periferica documentata (es.
    arteriopatia arti inferiori).                        |
| neoplasia_pregressa                     | Presenza di neoplasie trattate in passato.
                                                  |
| irradiazionetoracica                    | Pregressa radioterapia al torace, rilevante per
    effetti tardivi su cuore e vasi.                    |
| insufficienza_renale_cronica            | Presenza di insufficienza renale cronica diagnosticata
    .                                             |
| familiarita_cardiovascolare             | Familiarità per malattie cardiovascolari premature.
                                                  |
| limitazione_mobilita                    | Presenza di limitazioni significative alla mobilità (
    es. pazienti allettati).                       |
| endocardite                             | Pregressa o attiva endocardite infettiva.
                                                  |
| ritmo_all_ingresso                      | Ritmo cardiaco al momento del ricovero (0 = ritmo
    sinusale, 1 = FA, 2 = altro).                     |
| fibrillazione_atriale                   | Presenza di fibrillazione atriale (0 = mai, 1 =
    parossistica, 2 = permanente/persistente).          |
| dialisi                                 | Paziente in trattamento emodialitico o peritoneale.
                                                  |
| elettivo_urgenza_emergenza              | Tipo di intervento (0 = elettivo, 1 = urgente, 2 =
    emergenza).                                      |
| pm                                      | Presenza di pacemaker.
                                                  |
| crt                                     | Presenza di terapia di resincronizzazione cardiaca (
    CRT).                                           |
| icd                                     | Presenza di defibrillatore impiantabile (ICD).
                                                  |
| pci_pregressa                           | Precedente angioplastica coronarica percutanea (PCI).
                                                  |
| REDO                                    | Intervento cardiochirurgico di revisione (non prima
    chirurgia).                                    |
| Anno REDO                               | Anno in cui è stato eseguito l'intervento REDO
    precedente.                                          |
| Tipo di REDO                            | Descrizione del tipo di intervento REDO eseguito.
                                                  |
| Terapia                                 | Terapia farmacologica in atto al momento del ricovero.
                                                  |
| lasix                                   | Uso documentato di furosemide (Lasix).
                                                  |
| lasix_dosaggio                          | Dosaggio giornaliero di furosemide in mg.
                                                  |
| nitrati                                 | Assunzione di nitrati (vasodilatatori usati per l'
    angina).                                          |
| antiaggregante                          | Presenza di terapia antiaggregante (es. ASA,
    clopidogrel).                                          |
| dapt                                    | Doppia antiaggregazione piastrinica (es. ASA +
    clopidogrel/prasugrel).                              |
```

```
| anticoagorali                            | Terapia anticoagulante in corso (es. warfarin, DOAC).
       |
| aceinib                                  | Uso di ACE-inibitori.
      |
| betabloc                                 | Uso di beta-bloccanti.
      |
| sartanici                                | Uso di sartani (ARBs).
      |
| caantag                                  | Uso di calcio-antagonisti.
      |
| esami_all_ingresso                       | Risultati di laboratorio e strumentali al momento
    dellingresso.                                  |
| Decorso_post_operatorio                  | Descrizione del decorso clinico successivo
    allintervento chirurgico.                             |
| IABP/ECMO/IMPELLA                        | Necessità di supporto meccanico circolatorio (IABP,
    ECMO o Impella).                                 |
| Inotropi                                 | Necessità di farmaci inotropi positivi nel post-
    operatorio.                                          |
| secondo_intervento                       | Esecuzione di un secondo intervento durante la degenza
     attuale.                                 |
| Tipo_secondo_intervento                  | Tipo e motivazione del secondo intervento chirurgico.
       |
| II_Run                                   | Presenza di secondo passaggio in circolazione
    extracorporea (CEC).                                   |
| Causa_II_Run_CEC                         | Motivazione per il secondo utilizzo della CEC.
      |
| LCOS                                     | Sindrome da bassa portata cardiaca (Low Cardiac Output
     Syndrome) post-operatoria.                 |
| Impianto_PM_post_intervento              | Necessità di impianto di pacemaker dopo lintervento.
     |
| Stroke_TIA_post_op                       | Evento neurologico ischemico (TIA/stroke) avvenuto
    dopo lintervento.                                |
| Necessità_di_trasfusioni                 | Necessità di trasfusioni ematiche post-intervento.
      |
| IRA                                      | Insufficienza renale acuta insorta nel post-operatorio
    .                                            |
| Insufficienza_respiratoria               | Insorgenza di insufficienza respiratoria nel post-
    operatorio.                                        |
| FA_di_nuova_insorgenza                   | Fibrillazione atriale di nuova insorgenza nel post-
    operatorio.                                       |
| Ritmo_alla_dimissione                    | Ritmo cardiaco documentato alla dimissione (0 =
    sinusale, 1 = FA, 2 = altro).                          |
| H_Stay_giorni (da intervento a dimissione)| Durata della degenza in giorni, calcolata
    dallintervento alla dimissione.                       |
| Morte                                    | Evento di decesso durante la degenza cardiochirurgica.
      |
| Causa_morte                              | Causa clinica del decesso (es. sepsi, shock cardiogeno
    , ecc.).                                    |
| data_morte                               | Data del decesso, se avvenuto.
      |
| esami_alla_dimissione                    | Risultati di laboratorio e strumentali prima della
    dimissione.                                       |
| terapia_alla_dimissione                  | Terapia farmacologica prescritta alla dimissione.
      |


**ATTENZIONE**:
- Quando stai assegnando la label considera sia sia il nome dell'entità che il suo il valore al'
    interno della stessa entità TARGET
- Non estrarre **nessuna entità TARGET** diversa da quelle elencate.
- Attenzione però i nomi delle entità target che vedi sopra sono in alcuni casi degli acronimi o
    diminutivi delle entità
- Il numero di parole nella frase deve essere **esattamente uguale** al numero di label
    corrispondenti.
- Il risultato deve essere in formato JSON, come una lista di oggetti, ciascuno con le seguenti due
    chiavi:
  - ‘"frase"‘: stringa della frase processata.
  - ‘"label"‘: lista di label IOB corrispondenti alle parole.
```

```
IL NUMERO DI PAROLE NELLA FRASE DEVE ESSERE ESATTAMENTE UGUALE AL NUMERO DI LABEL NELL'ALTRA COLONNA


**NON AGGIUNGERE COMMENTI, NOTE O SPIEGAZIONI**, solo la lista JSON.

---

###Esempi di input

Si dimette in data 02/09/2019
il Sig. BERTOLOTTI FRANCO
Nato il 27/03/1939 telefono 3479927663
ricoverato presso questo ospedale dal 27/08/2019
Numero Cartella 2019034139
Intervento di plastica valvolare mitralica per via percutanea mediante posizionamento di duplice
     dispositivo Mitraclip.
Insufficienza mitralica in status post rivascolarizzazione miocardica chirurgica mediante triplice
     bypass coronarico.
Paziente nega farmacoallergie.
Familiarità positiva per cardiopatia ischemica (padre).
Ex fumatore, stop nel 1990 (1 pack/die).
Diabete mellito in tp ipoglicemizzante orale.
IRC (crea all'ingresso 2,64 mg/dl).


---

###Esempi output(esempio parziale in JSON):

‘‘‘json

  {
    "frase": "Si dimette in data 02/09/2019",
    "label": ["O", "O", "O", "O", "B-TARGET"]
  },
  {
    "frase": "il Sig. BERTOLOTTI FRANCO",
    "label": ["O", "O", "B-TARGET", "B-TARGET"]
  },
  {
    "frase": "Nato il 27/03/1939 telefono 3479927663",
    "label": ["O", "O", "B-TARGET", "O", "B-TARGET"]
  },
  {
    "frase": "ricoverato presso questo ospedale dal 27/08/2019",
    "label": ["O", "O", "O", "O", "O", "B-TARGET"]
  },
  {
    "frase": "Numero Cartella 2019034139",
    "label": ["O", "O", "B-TARGET"]
  },
  {
    "frase": "Intervento di plastica valvolare mitralica per via percutanea mediante posizionamento
     di duplice dispositivo Mitraclip.",
    "label": ["B-TARGET", "I-TARGET", "I-TARGET", "I-TARGET", "I-TARGET", "I-TARGET", "I-TARGET", "I
     -TARGET", "I-TARGET", "I-TARGET", "I-TARGET", "I-TARGET"]
  },
  {
    "frase": "Insufficienza mitralica in status post rivascolarizzazione miocardica chirurgica
     mediante triplice bypass coronarico.",
    "label": ["B-TARGET", "I-TARGET", "I-TARGET", "I-TARGET", "I-TARGET", "I-TARGET", "I-TARGET", "I
     -TARGET", "I-TARGET", "I-TARGET", "I-TARGET"]
  },
  {
    "frase": "Paziente nega farmacoallergie.",
    "label": ["B-TARGET", "I-TARGET", "I-TARGET"]
  },
  {
```

```
    "frase": "Familiarità positiva per cardiopatia ischemica (padre).",
    "label": ["B-TARGET", "I-TARGET", "I-TARGET", "I-TARGET", "I-TARGET", "I-TARGET"]
  },
  {
    "frase": "Ex fumatore, stop nel 1990 (1 pack/die).",
    "label": ["B-TARGET", "I-TARGET", "I-TARGET", "I-TARGET", "I-TARGET", "I-TARGET"]
  },
  {
    "frase": "Diabete mellito in tp ipoglicemizzante orale.",
    "label": ["B-TARGET", "I-TARGET", "I-TARGET", "I-TARGET", "I-TARGET"]
  },
  {
    "frase": "IRC (crea all'ingresso 2,64 mg/dl).",
    "label": ["B-TARGET", "I-TARGET", "I-TARGET", "I-TARGET", "I-TARGET", "I-TARGET"]
  }

  frase da processare:
'''
```

I chose not to pass the entire text in one go to the model to perform token-level labeling, because I noticed that when the input phrase was too long, the model lost context and began assigning `O` labels.

Another approach I considered (but haven't explored yet) is using the same methodology as GPT-NER[5]. However, since the results from our current pipeline were quite good, we decided to stick with this method for now.

One of the main issues we encountered with this approach was the **mismatch in length** between the list of predicted labels and the list of tokens in the original text. This problem has also been highlighted in the **GPT-NER paper**, where it is noted that using LLMs for token-level classification can result in misalignments between token sequences and predicted labels—especially when tokenization strategies differ or the model generates inconsistent outputs.

To solve the problem we created a function that align the labels by adding `O` or by truncating the labels list. As a result, we achieved **zero mismatches** between tokens and labels.

After alignment, we observed the following distribution of labels:

- `B-TARGET`: 13971
- `I-TARGET`: 42653
- `O`: 70848

Since the number of `O` labels was significantly higher than the number of `TARGET` labels, we decided to **clean the dataset further** by removing sentences that contained only `O` labels and no target entities. This helped improve the quality and balance of the dataset.

To further improve the quality of our dataset, we performed the following heuristic filtering steps:

- Removed **phrases containing only one label**, as they were often too vague to provide meaningful learning signals.
- Discarded **phrases shorter than 15 characters**, which were generally too short to be informative.

In the final version, we had the following label counts:

- `B-TARGET`: 13728
- `I-TARGET`: 42405
- `O`: 56712

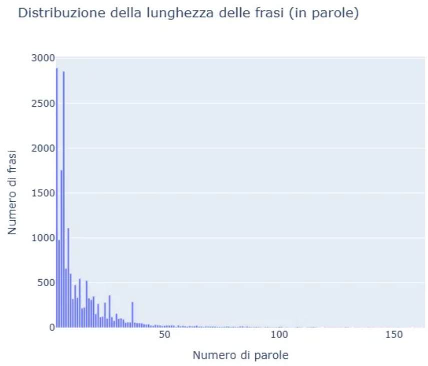

Fig. 1 Sentences lenght distribution

To get better dataset quality and validate it, we also created a Streamlit app to discard wrongly labeled phrases, but for the moment we didn't use it.

Fig. 2 Streamlit app

4.4 Test Data

As test data we used a pdf file of the discharge letter, since in production doctors will load the single pdf files and not the enitire scanned clinical folder.

5 LLM inference for NER

As we already did for the creation of the NER dataset, we use a specific prompt to guide the extraction of correct entities. An effective prompt in this context should include the following components:

- **Goal**: Clearly define the task for the model.
- **Return Format**: Specify the expected structure of the output.
- **Warnings**: Highlight critical constraints or edge cases.
- **Context Dump**: Provide examples or definitions to guide the model.

In our experiments,we observed that few-shot prompting significantly outperformed zero-shot prompting in terms of accuracy and label consistency.The same result was found in other work like[6], in which are used specif examples for each input.

```
promp_base='''
Sei un medico specializzato in cardiochirurgia. Il tuo compito è estrarre **esclusivamente** le
     seguenti entità dalla **lettera di dimissione** riportata qui sotto.

#####**Entità da estrarre (solo queste):**

### Mappa delle entità e tipi

| Entità                                       | Tipo              | Descrizione

                        |
|-------------------------------------------------------------------------------------|
| n_cartella                                 | Number            | Numero identificativo univoco
     assegnato alla cartella clinica del paziente.
                         |
| data_ingresso_cch                          | Date              | Data in cui il paziente è stato
     ricoverato presso il reparto di Cardiochirurgia.
                      |
| data_dimissione_cch                        | Date              | Data in cui il paziente è stato
     dimesso dal reparto di Cardiochirurgia.
                       |
| nome                                       | Text              | Nome proprio del paziente.

                       |
| cognome                                    | Text              | Cognome del paziente.

                       |
| sesso                                      | Categorical_MF    | Sesso biologico del paziente (M =
      Maschio, F = Femmina).
                       |
| numero di telefono                         | Text              | Recapito telefonico del paziente
     o di un contatto di riferimento.
                      |
| età al momento dell'intervento             | Number            | Età del paziente calcolata alla
     data dellintervento chirurgico.
                      |
| data_di_nascita                            | Date              | Data di nascita del paziente.

                       |
| Diagnosi                                   | Text              | Diagnosi principale alla base
     dell'indicazione chirurgica.
                          |
| Anamnesi                                   | Text              | Anamnesi patologica remota e
     prossima, utile per la valutazione del rischio operatorio.
                          |
| Motivo ricovero                            | Text              | Indicazione clinica per il
     ricovero in Cardiochirurgia.
                             |
| classe_nyha                                | Categorical_1234  | Classe funzionale NYHA per
     scompenso cardiaco (I-IV), definisce la gravità dei sintomi.
                            |
| angor                                      | Boolean           | Presenza di angina pectoris (
     dolore toracico di origine ischemica).
                          |
| STEMI/NSTEMI                               | Boolean           | Presenza di infarto miocardico
     acuto con/senza sopraslivellamento del tratto ST.
                        |
```

```
| scompenso_cardiaco_nei_3_mesi_precedenti   | Boolean           | Episodi di scompenso cardiaco
    documentati nei 3 mesi precedenti lintervento.
                    |
| fumo                                       | Categorical_012   | Abitudine al fumo (0 = mai fumato
    , 1 = ex-fumatore, 2 = fumatore attivo).
                  |
| diabete                                    | Boolean           | Presenza di diabete mellito noto.

                     |
| ipertensione                               | Boolean           | Presenza di ipertensione
    arteriosa.
                             |
| dislipidemia                               | Boolean           | Presenza di dislipidemia (
    colesterolo e/o trigliceridi elevati).
                           |
| BPCO                                       | Boolean           | Presenza di broncopneumopatia
    cronica ostruttiva.
                        |
| stroke_pregresso                           | Boolean           | Precedente episodio di ictus
    cerebrale ischemico o emorragico.
                         |
| TIA_pregresso                              | Boolean           | Episodio pregresso di attacco
    ischemico transitorio (TIA).
                        |
| vasculopatiaperif                          | Boolean           | Malattia vascolare periferica
    documentata (es. arteriopatia arti inferiori).
                       |
| neoplasia_pregressa                        | Boolean           | Presenza di neoplasie trattate in
     passato.
                     |
| irradiazionetoracica                       | Boolean           | Pregressa radioterapia al torace,
     rilevante per effetti tardivi su cuore e vasi.
                  |
| insufficienza_renale_cronica               | Boolean           | Presenza di insufficienza renale
    cronica diagnosticata.
                     |
| familiarita_cardiovascolare                | Boolean           | Familiarità per malattie
    cardiovascolari premature (prima dei 55 anni per uomini, 65 per donne).
                             |
| limitazione_mobilita                       | Boolean           | Presenza di limitazioni
    significative alla mobilità (es. pazienti allettati).
                             |
| endocardite                                | Boolean           | Pregressa o attiva endocardite
    infettiva, rilevante per indicazione chirurgica.
                    |
| ritmo_all_ingresso                         | Categorical_012   | Ritmo cardiaco al momento del
    ricovero (0 = ritmo sinusale, 1 = FA, 2 = altro).
                        |
| fibrillazione_atriale                      | Categorical_012   | Presenza di fibrillazione atriale
     (0 = mai, 1 = parossistica, 2 = permanente/persistente).
                  |
| dialisi                                    | Boolean           | Paziente in trattamento
    emodialitico o peritoneale.
                              |
| elettivo_urgenza_emergenza                 | Categorical_012   | Tipo di intervento (0 = elettivo,
     1 = urgente, 2 = emergenza).
                    |
| pm                                         | Boolean           | Presenza di pacemaker.

                     |
| crt                                        | Boolean           | Presenza di terapia di
    resincronizzazione cardiaca (CRT).
                               |
| icd                                        | Boolean           | Presenza di defibrillatore
    impiantabile (ICD).
                           |
```

```
| pci_pregressa                           | Boolean           | Precedente angioplastica
    coronarica percutanea (PCI).
                               |
| REDO                                    | Boolean           | Intervento cardiochirurgico di
    revisione (non prima chirurgia).
                        |
| Anno REDO                               | Date              | Anno in cui è stato eseguito l'
    intervento REDO precedente.
                         |
| Tipo di REDO                            | Text              | Descrizione del tipo di
    intervento REDO eseguito.
                                |
| Terapia                                 | Text              | Terapia farmacologica in atto al
    momento del ricovero.
                      |
| lasix                                   | Boolean           | Uso documentato di furosemide (
    Lasix).
                         |
| lasix_dosaggio                          | Number            | Dosaggio giornaliero di
    furosemide in mg.
                                |
| nitrati                                 | Boolean           | Assunzione di nitrati (
    vasodilatatori usati per l'angina).
                                 |
| antiaggregante                          | Boolean           | Presenza di terapia
    antiaggregante (es. ASA, clopidogrel).
                                    |
| dapt                                    | Boolean           | Doppia antiaggregazione
    piastrinica (es. ASA + clopidogrel/prasugrel).
                                |
| anticoagorali                           | Boolean           | Terapia anticoagulante in corso (
    es. warfarin, DOAC).
                      |
| aceinib                                 | Boolean           | Uso di ACE-inibitori.

                      |
| betabloc                                | Boolean           | Uso di beta-bloccanti.

                      |
| sartanici                               | Boolean           | Uso di sartani (ARBs).

                      |
| caantag                                 | Boolean           | Uso di calcio-antagonisti.

                      |
| esami_all_ingresso                      | Text              | Risultati di laboratorio e
    strumentali al momento dellingresso.
                           |
| Decorso_post_operatorio                 | Text              | Descrizione del decorso clinico
    successivo allintervento chirurgico.
                     |
| IABP/ECMO/IMPELLA                       | Boolean           | Necessità di supporto meccanico
    circolatorio (IABP, ECMO o Impella).
                      |
| Inotropi                                | Boolean           | Necessità di farmaci inotropi
    positivi nel post-operatorio.
                          |
| secondo_intervento                      | Boolean           | Esecuzione di un secondo
    intervento durante la degenza attuale.
                              |
| Tipo_secondo_intervento                 | Text              | Tipo e motivazione del secondo
    intervento chirurgico.
                         |
| II_Run                                  | Boolean           | Presenza di secondo passaggio in
    circolazione extracorporea (CEC).
                     |
```

```
| Causa_II_Run_CEC                          | Text              | Motivazione per il secondo
     utilizzo della CEC.
                             |
| LCOS                                      | Boolean           | Sindrome da bassa portata
     cardiaca (Low Cardiac Output Syndrome) post-operatoria.
                              |
| Impianto_PM_post_intervento               | Boolean           | Necessità di impianto di
     pacemaker dopo lintervento.
                              |
| Stroke_TIA_post_op                        | Boolean           | Evento neurologico ischemico (TIA
     /stroke) avvenuto dopo lintervento.
                     |
| Necessità_di_trasfusioni                  | Boolean           | Necessità di trasfusioni ematiche
      post-intervento.
                       |
| IRA                                       | Boolean           | Insufficienza renale acuta
     insorta nel post-operatorio.
                             |
| Insufficienza_respiratoria                | Boolean           | Insorgenza di insufficienza
     respiratoria nel post-operatorio.
                            |
| FA_di_nuova_insorgenza                    | Boolean           | Fibrillazione atriale di nuova
     insorgenza nel post-operatorio.
                         |
| Ritmo_alla_dimissione                     | Categorical_012   | Ritmo cardiaco documentato alla
     dimissione (0 = sinusale, 1 = FA, 2 = altro).
                        |
| H_Stay_giorni (da intervento a dimissione) | Number           | Durata della degenza in giorni,
     calcolata dallintervento alla dimissione.
                       |
| Morte                                     | Boolean           | Evento di decesso durante la
     degenza cardiochirurgica.
                           |
| Causa_morte                               | Text              | Causa clinica del decesso (es.
     sepsi, shock cardiogeno, ecc.).
                         |
| data_morte                                | Date              | Data del decesso, se avvenuto.

                       |
| esami_alla_dimissione                     | Text              | Risultati di laboratorio e
     strumentali prima della dimissione.
                             |
| terapia_alla_dimissione                   | Text              | Terapia farmacologica prescritta
     alla dimissione.
                       |



---

### **Istruzioni IMPORTANTI:**

- Ragiona considerando **frase per frase**.
- Non estrarre **nessuna entità** diversa da quelle elencate.
- Se un'entità non è presente nella lettera, **non inventarla** e **non includerla** nel risultato.
- Attenzione però i nomi delle entità che vedi sopra sono in alcuni casi degli acronimi o diminutivi
      delle entità.
- Il formato di output deve essere una lista JSON, dove ogni elemento è un oggetto con **due chiavi
     **:
    - ‘"entità"‘: il nome dell'entità
    - ‘"valore"‘: il valore estratto dell'entità
**NON** aggiungere commenti, spiegazioni, note, intestazioni o altro: **solo** la lista JSON.

---

###Esempio di input(esempio parziale della lettera di dimission)
Si dimette in data 02/09/2019
il Sig. BERTOLOTTI FRANCO
```

```
Nato il 27/03/1939 telefono 3479927663
ricoverato presso questo ospedale dal 27/08/2019
Numero Cartella 2019034139

Diagnosi alla dimissione:
Intervento di plastica valvolare mitralica per via percutanea mediante posizionamento di duplice
     dispositivo Mitraclip.

Motivo del Ricovero:
Insufficienza mitralica in status post rivascolarizzazione miocardica chirurgica mediante triplice
     bypass coronarico.

Cenni Anamnestici:
Paziente nega farmacoallergie.
Familiarità positiva per cardiopatia ischemica (padre).
Ex fumatore, stop nel 1990 (1 pack/die).
Diabete mellito in tp ipoglicemizzante orale.
IRC (crea all'ingresso 2,64 mg/dl).


---

###Esmpio output(esempio parziale in JSON):

‘‘‘json
[
  { "entità": "data_dimissione_cch", "valore": "02/09/2019" },
  { "entità": "nome", "valore": "FRANCO" },
  { "entità": "cognome", "valore": "BERTOLOTTI" },
  { "entità": "data_di_nascita", "valore": "27/03/1939" },
  { "entità": "numero di telefono", "valore": "3479927663" },
  { "entità": "data_ingresso_cch", "valore": "27/08/2019" },
  { "entità": "n_cartella", "valore": "2019034139" },
  { "entità": "Diagnosi text", "valore": "Intervento di plastica valvolare mitralica per via
     percutanea mediante posizionamento di duplice dispositivo Mitraclip." },
  { "entità": "Motivo ricovero", "valore": "Insufficienza mitralica in status post
     rivascolarizzazione miocardica chirurgica mediante triplice bypass coronarico." },
  { "entità": "fumo", "valore": true },
  { "entità": "diabete", "valore": true },
  { "entità": "insufficienza renale cronica", "valore": true },
  { "entità": "familiarita cardiovascolare", "valore": true }
]
'''
```

We believe that **the parameters used for prompting the LLM play a crucial role** in achieving accurate and consistent label generation. In particular, since we aim to **avoid hallucination** and ensure that the model **uses the exact words from the input text**, we chose:

- a **very low temperature** of **0.1**,
- and a **high top-p (nucleus sampling)** value, close to **1.0**.

The **temperature** parameter controls the randomness of the model's output. Lower values make the model more **deterministic** and **conservative**, favoring high-probability tokens. Higher values increase randomness and diversity.

Mathematically, temperature modifies the logits before applying *softmax*:

$$P(w_i) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

where:

- z_i is the logit for token i,

- T is the temperature.

As $T \to 0$, the output becomes greedy and predictable.
As $T \to \infty$, the output becomes highly random.
Top-p sampling (or nucleus sampling) controls diversity by limiting the sampling pool to the **smallest set of tokens** whose cumulative probability exceeds a threshold p.
This means:

- With **top-p = 1.0**, the model considers the entire probability distribution (equivalent to not using top-p).
- With **top-p = 0.9**, only the top tokens whose combined probability is $\geq 90\%$ are considered for sampling.

This technique helps filter out low-probability (**potentially noisy or hallucinated**) outputs, while still allowing some controlled variability.

In summary, by setting a **low temperature (0.1)** and **high top-p ($\approx$ 1.0)**, we ensure the LLM remains focused, reduces hallucinations, and adheres closely to the original input text—essential qualities for generating reliable token-level annotations.

5.1 LLMs comparison

We compared three recent state-of-the-art large language models:

- **LLaMA3-70B**,
- **DeepSeek V3**,
- **DeepSeek R1**.

Although several studies have evaluated LLMs for similar tasks, most of them rely on older-generation models [7][1].

After running inference on the dataset, we observed that the output from the reasoning model **DeepSeek R1** was inconsistent—it produced fewer answers than the number of input documents.

For this reason, we excluded DeepSeek R1 from the subsequent analyses.

To compare the performance of the LLaMA and DeepSeek models, we first analyzed the entities extracted by each model, focusing on identifying potential hallucinations or type errors. A **hallucination** occurs when the model generates an entity that is not present in the input text, while a **type error** happens when an entity is extracted but its type or format is incorrect.

To assess this, we developed a scoring function that assigns:

- 1 point for each entity with a wrong type,
- 2 points for each hallucinated (fabricated) entity.

From this analysis, we observed that **no hallucinations were generated**—no new entities were invented by the models. However, we did detect a few type errors. Upon closer inspection, many of these were not true errors, but rather issues of format or representation. For instance, in some cases like “fumo”, the model returned a value such as “Moderato”, which is not part of the expected categorical values, but still reflects a correct interpretation of the text.

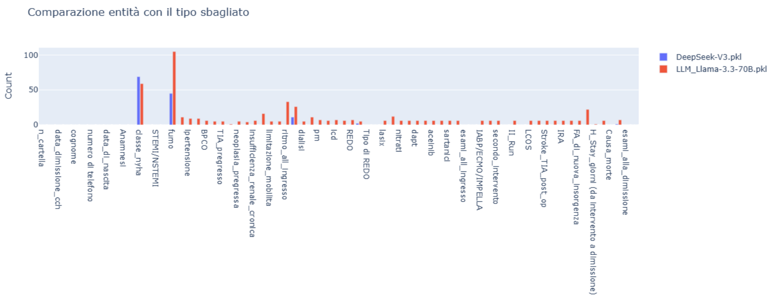

Fig. 3 Comparison wrong type

To further compare the behavior and stability of the models, we plotted the distribution of the number of extracted entities per patient. The results show that:

- The **LLaMA model** tends to extract **fewer entities on average** and has a **lower standard deviation**.
- The **DeepSeek model** extracts **more entities**, and while its distribution is more dispersed, this could indicate richer coverage.

However, since the distribution is not symmetric, standard deviation alone may not be a meaningful indicator of performance. In general, we found that **DeepSeek captures a wider range of entities**, especially boolean-type entities that are implicitly present in the text. For example, even when the presence of an entity is not explicitly mentioned, DeepSeek was often able to infer and extract it correctly.

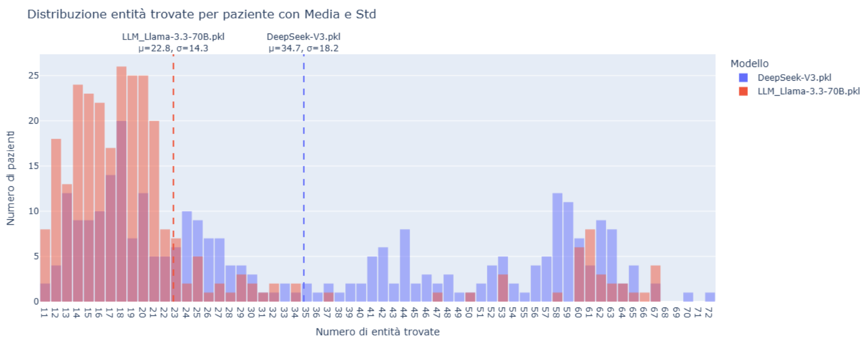

Fig. 4 Frequency distribution

Moreover, when comparing type errors across the two models, DeepSeek showed significantly fewer errors, reinforcing the idea that it is more reliable for this task. That said, neither model extracted every possible entity—but this is likely due to the absence of those entities in the original documents, rather than a failure of the models themselves

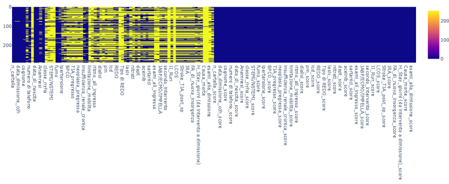

Fig. 5 Llama3 entities

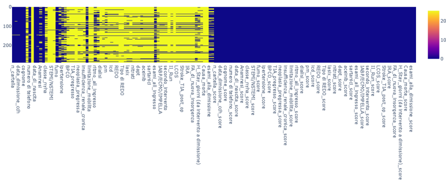

Fig. 6 Deep seekV3 entities

5.1.1 Text-type entities comparison

We also conducted a specific comparison of **text-type entities** (i.e., entities represented as free-text values rather than categorical or boolean values).

To evaluate their similarity, we:

1. Extracted all text-type entities from both models.
2. Used a BERT-based model to embed the entity texts.
3. Calculated centroids for each entity type (in this way we "normalize" them because we don't consider the frequency).
4. Measured the **cosine similarity** between corresponding centroids.

The results showed that the embeddings of the entities extracted by the two models were **highly similar**, suggesting that both models are consistent and that hallucinations are unlikely in this category as well.

To visualize this more clearly, we used t-SNE to project all entity embeddings (not just centroids) into 2D space. The resulting plot confirmed our observations: most of the entities from both models overlapped significantly. The only noticeable outliers were entities identified by DeepSeek but not by LAMA, further highlighting the former's broader recall capability.

Entity	Similarity Score
nome	1.0000
cognome	0.9999
numero di telefono	0.9999
Diagnosi	0.9971
Anamnesi	0.9955
Motivo ricovero	0.9978
Tipo di REDO	0.9980
Terapia	0.9118
esami_all_ingresso	0.8472
Decorso_post_operatorio	0.9979
Tipo_secondo_intervento	1.0000
Causa_II_Run_CEC	0.9999
Causa_morte	0.9999
esami_alla_dimissione	0.7914
terapia_alla_dimissione	0.9174

Table 2 Cosine similarity between centroid embeddings of text-type entities from the two models.

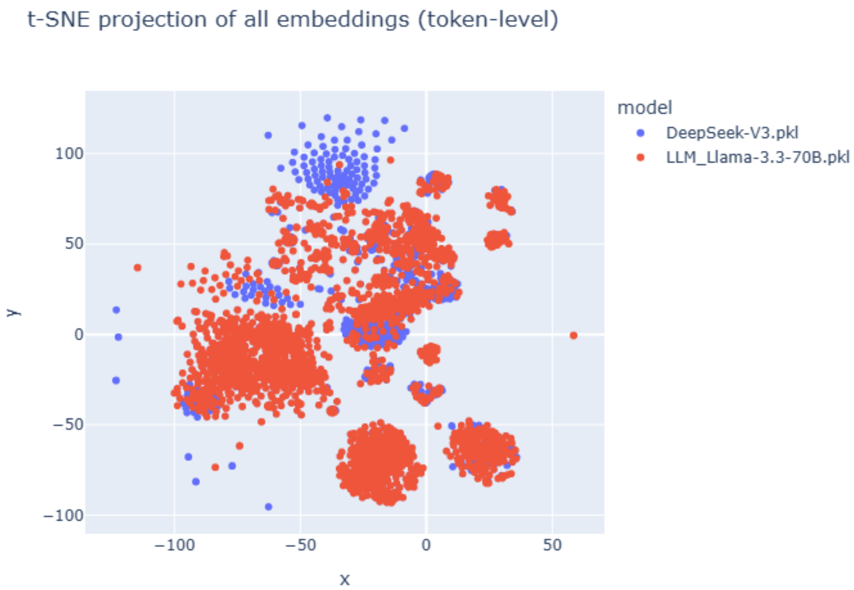

Fig. 7 projection tsne

These findings support the conclusion that **DeepSeek not only extracts more entities, but it does so without compromising the quality or semantic alignment of the output**.

6 BIOClinical BERT

After creating our dataset for Named Entity Recognition in the clinical domain, we fine-tuned **Bio_ClinicalBERT**[8], a domain-specific language model designed to handle medical and clinical text with greater precision than general-purpose BERT models. Bio_ClinicalBERT is a specialized version of BERT that builds on the strengths of BioBERT[9], which was originally trained on a large corpus of biomedical literature from PubMed and PMC. To make it more effective for clinical tasks, Bio_ClinicalBERT was further pretrained on real clinical notes from the MIMIC-III database—an extensive collection of de-identified ICU patient records from Beth Israel Deaconess Medical Center in Boston.

The model used in our work was initialized from BioBERT-Base v1.0 and then trained on all available notes from MIMIC-III, including discharge summaries, nursing notes, radiology reports, and more. In total, this corpus consists of roughly 880 million words. Each clinical note was carefully preprocessed: first split into logical sections (e.g., *History of Present Illness* or *Family History*) using rule-based heuristics, and then further broken into sentences using the SciSpacy tokenizer trained on scientific text.

This extended pretraining allowed the model to better understand the language, structure, and patterns specific to hospital documentation. For instance, abbreviations, shorthand notations, and medical jargon are prevalent in clinical records, and Bio_ClinicalBERT is well-equipped to handle these thanks to its exposure to both biomedical literature and real-world hospital data.

The training was performed using the original BERT pretraining approach. It included masked language modeling with a 15% token masking probability, a batch size of 32, and a learning rate of 5×10^{-5}. The model was trained for 150,000 steps, with a duplication factor of 5 to introduce variation in masking patterns. This setup helps the model generalize better and learn contextual information more effectively.

Thanks to this rigorous pretraining on relevant data, Bio_ClinicalBERT performs significantly better than standard BERT or even BioBERT on clinical tasks such as NER.

6.1 Finetuning

To fine-tune the **Bio_ClinicalBERT** model for our Named Entity Recognition (NER) task, we configured a set of training hyperparameters and evaluation strategies using Hugging Face's `TrainingArguments`. The following table summarizes all the key training parameters, optimizer, and loss function used in our fine-tuning process:

7 Test

To evaluate the two proposed solutions, we tested them using a discharge letter, as previously mentioned in the dataset testing section.

For the first approach—using an LLM—we applied the same prompt used in our earlier comparison of different LLMs. For the second approach—based on the fine-tuned BERT model—we performed inference on each individual sentence of the

Parameter	Value	Description
`evaluation_strategy`	"epoch"	Evaluation is run at the end of each training epoch.
`learning_rate`	5e-5	Learning rate used during training.
`per_device_train_batch_size`	32	Batch size for training per GPU.
`per_device_eval_batch_size`	32	Batch size for evaluation per GPU.
`num_train_epochs`	20	Total number of training epochs.
`weight_decay`	0.01	Weight decay used for regularization.
`save_strategy`	"epoch"	Model is saved at the end of each epoch.
`load_best_model_at_end`	True	After training, the best model (based on metric) is reloaded.
`metric_for_best_model`	"f1"	Metric used to evaluate and select the best model.
`report_to`	"wandb"	Logging and tracking is done via Weights & Biases.
Optimizer	AdamW	Common optimizer used for Transformer models (default in Hugging Face Trainer).
Loss Function	Cross Entropy Loss	Standard loss used for token classification tasks like NER.

Table 3 Training parameters, optimizer, and loss function used for fine-tuning Bio_ClinicalBERT.

discharge letter. The model identified entities sentence by sentence, and the collected outputs were then passed to an LLM along with a prompt for classification. The prompt used was similar in structure to the one used in the direct LLM approach.

In the future, this classification step could be replaced by a dedicated model trained specifically for classification tasks, enhancing both accuracy and speed.

Upon comparing the results of both pipelines, we observed that the extracted entities were largely the same. However, in the BERT + LLM pipeline, two entities were hallucinated—that is, they were not part of the expected list of entities defined in the prompt. Despite this, the values extracted for the correctly identified entities were consistent across both approaches, even for complex text-based entities.

BERT+LLM

Entità	Valore
data_dimissione_cch	27/01/2025
nome	MASSIMO
cognome	RICCA
data_di_nascita	17/02/1966
numero di telefono	3287351755
n_cartella	2025003002
data_ingresso_cch	20/01/2025
Diagnosi	In data 21/01/2025 Sostituzione valvolare aort...
Motivo ricovero	Stenosi aortica
familiarita cardiovascolare	True
fumo	False
classe_nyha	IIb
angor	False
dispnea	*False*
sincopi	*False*

LLM

Entità	Valore
data_dimissione_cch	27/01/2025
nome	MASSIMO
cognome	RICCA
data_di_nascita	17/02/1966
numero di telefono	3287351755
n_cartella	2025003002
data_ingresso_cch	20/01/2025
Diagnosi	Sostituzione valvolare aortica con protesi bio...
Motivo ricovero	Stenosi aortica
familiarita cardiovascolare	True
fumo	False
classe_nyha	IIb
angor	False

8 Conclusion

Based on the test, we conclude that directly using an LLM for this task is more effective and scalable. LLMs demonstrate strong performance out-of-the-box, and they can be easily adapted to new use cases by simply modifying the prompt to add or remove entities. Moreover, the cost of using LLM APIs—especially from providers like Together AI—is relatively low, while still ensuring compliance with privacy regulations.

Finally, should even higher performance be required in the future, further improvements can be achieved by fine-tuning the LLM using lightweight techniques such as LoRA, once the system is in production and real-world data has been collected.

9 AI Usage Disclaimer

Parts of this projects have been developed with the assistance of OpenAI's ChatGPT (GPT-4). The AI was used to support the development of project ideas, the structuring of methodological workflows, the drafting of descriptive texts, and the identification of relevant datasets and references. All content produced with AI assistance has been carefully reviewed, edited, and validated by me. I take full responsibility for the final content and its accuracy, relevance, and academic integrity

.

References

[1] Hu, Y., Chen, Q., Du, J., Peng, X., Keloth, V.K., Zuo, X., Zhou, Y., Li, Z., Jiang, X., Lu, Z., Roberts, K., Xu, H.: Improving Large Language Models for Clinical Named Entity Recognition via Prompt Engineering (2024). https://arxiv.org/abs/2303.16416

[2] Ntinopoulos, V., Rodriguez Cetina Biefer, H., Tudorache, I., Papadopoulos, N., Odavic, D., Risteski, P., Haeussler, A., Dzemali, O.: Large language models for data extraction from unstructured and semi-structured electronic health records: a multiple model performance evaluation. BMJ Health Care Inform **32**(1), 101139 (2025) https://doi.org/10.1136/bmjhci-2024-101139

[3] Jurafsky, D., Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edn. (2025). Online manuscript released January 12, 2025. https://web.stanford.edu/~jurafsky/slp3/

[4] Ye, J., Xu, N., Wang, Y., Zhou, J., Zhang, Q., Gui, T., Huang, X.: LLM-DA: Data Augmentation via Large Language Models for Few-Shot Named Entity Recognition (2024). https://arxiv.org/abs/2402.14568

[5] Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., Wang, G.: GPT-NER: Named Entity Recognition via Large Language Models (2023). https://arxiv.org/abs/2304.10428

[6] Jiang, G., Ding, Z., Shi, Y., Yang, D.: P-ICL: Point In-Context Learning for Named Entity Recognition with Large Language Models (2024). https://arxiv.org/abs/2405.04960

[7] Gu, B., Shao, V., Liao, Z., *et al.*: Scalable information extraction from free text electronic health records using large language models. BMC Medical Research Methodology **25**, 23 (2025) https://doi.org/10.1186/s12874-025-02470-z

[8] Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., McDermott, M.B.A.: Publicly Available Clinical BERT Embeddings (2019). https://arxiv.org/abs/1904.03323

[9] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2019) https://doi.org/10.1093/bioinformatics/btz682