Charles University in Prague

Faculty of Mathematics and Physics

MASTER'S THESIS

Goutham V Karunakaran

Automatic Relation Extraction from Clinical Documents: A Study of Fine-Tuned Transformer Models and LLMs

Institute of Formal and Applied Linguistics

Supervisor :	doc. RNDr. Pavel Pecina
Study programme:	Computer Science
Study branch:	Language Technologies and Computational Linguistics

Prague 2024

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

Gth

In Trento date 18.06.2024 signature of the author

Title: Automatic Relation Extraction from Clinical Documents: A Study of Fine-Tuned Transformer Models and LLMs

Author: Goutham V Karunakaran

Department: Institute of Formal and Applied Linguistics

Supervisor: doc. RNDr. Pavel Pecina

Co-Supervisor: Dr. Carlo Strapparava, Alberto Lavelli(FBK)

Abstract: Clinical documents are rich sources of patient information, notably about laboratory tests that inform medical decisions. However, as the volume of such documents grows, there's a pressing need for effective methods to interpret them. This thesis, titled "Automatic Relation Extraction from Clinical Documents: A Study of Fine-Tuned Transformer Models and LLMs", dives into this challenge of pinpointing test results and measurements in clinical documents and associating them with the respective laboratory tests they originate from. We've evaluated several models, such as Multilingual BERT, XLM-RoBERTa, and BioBERT, adapting them for our task. We also explored the potential of advanced large language models like GPT-3.5 and GPT-4 without any fine-tuning. An added dimension to our study is the multilingual nature of the clinical records, spanning Italian, Spanish, and Basque. These languages are often sidelined in research, which mostly centers on English. By focusing on them, we hope to fill a notable research gap. The thesis offers a journey starting with a review of relevant literature, a deep dive into the data and its nuances, a detailed look into our methodology, a discussion on our findings, and ends with insights for future investigations in this sphere.

Keywords: Relation Extraction, Biomedical Text Mining, Transformers, LLM

Acknowledgement

I thank the NLP Research Group in FBK for providing the environment to do research. Thanks are due to Alberto Lavelli and Roberto Zanoli for their guidance and mentorship throughout the course of this thesis. Their insights and discussions shaped my research.

I am grateful to my supervisors: Dr. Carlo Strapparava from the University of Trento and doc. RNDr. Pavel Pecina from Charles University. Their expertise and feedback have been useful in refining my work and thoughts.

Living and studying in both Prague and Trento for two years has enriched my life in countless ways. The vibrant culture of Prague and the serene beauty of Trento have left an indelible mark on my heart. I fondly recall the memories created in these cities and the collaborations with FBK, which have been pivotal in my academic journey.

This two-year odyssey was made achievable by the Erasmus Mundus scholarship from the European Union, allowing me to immerse myself in the Erasmus Mundus European Master Program in Language and Communication Technologies (EM LCT). This thesis stands as a testament to that incredible opportunity.

Lastly, my thanks go to my family, friends, and colleagues in the NLP Research Group.

Contents

Introduction

Automated clinical data mining, encompassing techniques like relation extraction and entity recognition, has emerged as a pivotal tool in modern healthcare informatics. With the exponential growth of electronic health records (EHRs), clinical statements and biomedical literature, there's a pressing need to extract meaningful insights from this vast repository of unstructured data (Jensen et al. [2012]). Relation extraction plays a crucial role in identifying and categorizing relationships between medical entities, such as drugs and their potential side effects or diseases and their associated symptoms (Percha and Altman [2013]). Biomedical entity recognition helps pinpoint specific medical terms or concepts within texts and serves as a foundational task, as a precursor to many downstream applications like relation extraction. Together, these automated techniques enhance clinical decision-making and also foster personalized patient care and advance medical research. Furthermore, by revealing patterns and trends within clinical data, these tools can contribute to epidemiological studies and public health monitoring, enhancing our response to health crises. Their integration into healthcare systems can help clinicians, researchers, and policymakers leverage data more effectively and improve the system overall.

Clinical case

A clinical case is a statement of a clinical practice, detailing the purpose of the patient's visit, description of physical examinations and an evaluation of the patient's condition. They are rich in clinical entities and temporal information.

An example clinical case :

```
The clinical case of an 84-year-old woman, a heavy smoker for
   approximately 30 years, is described. Unfamiliarity with kidney
    disease. Main anamnestic findings: arterial hypertension for
   at least 20 years controlled by pharmacological therapy (
   calcium antagonists, diuretics), cholelithiasis, hypothyroidism
    on hormone replacement therapy. No hearing problems. At the
   age of 70, right breast quadrantectomy for carcinoma. At the
   age of 80, the patient underwent two hospitalizations during
   which multiple, bilateral pulmonary thickenings were
   highlighted and the diagnosis of "extrinsic allergic alveolitis
   " was made. On that occasion, infectious pathology was excluded
    and a steroid was prescribed following which the patient
   showed an improvement in the clinical picture; no data
   available on renal function. In 2005, at the age of 82, the
   patient was hospitalized for rapidly progressive acute renal
   failure (creatinine level 1.5 ->6 mg/dL) and respiratory
   failure. From a systemic point of view, the patient presented
   with widespread osteoarticular pain, low-grade fever and dry
   cough with a single episode of haemoptysis. Her chest x-ray
   revealed bilateral pulmonary enlargements. Chest CT showed
   ground glass appearance of the lung parenchyma, alveolar edema,
    multiple parenchymal thickenings. An echocardiogram was normal
    for his age..... In the hypothesis of a reactivation of the
   autoimmune disease and with the evidence of negative culture
   tests, the patient practiced IV cortisone boluses again but the
    clinical evolution was complicated by K. pneumoniae pneumonia
   which quickly led to the patient’s death.
```

About the Task

This thesis started as a work to establish baselines for a shared task in Relation Extraction organized by NLP research unit of the Fondazione Bruno Kessler in collaboration with HiTZ (Basque Center for Language Technology) at Evalita 2022 (Italian) and IberLEF 2023 (Spanish and Basque) [1]. It expands beyond this foundational work. The primary focus of this thesis is to address the challenge of accurately identifying **test results** and **measurements** within clinical documents and linking them to the corresponding textual mentions of the **laboratory tests** from which they were derived. This involves the identification of both elements in the text, as well as the establishment of the relationship between them.

Eg., in the above clinical case, an example relation is `[1.5 - 6 mg/dL] -->[creatinine level]`

In this thesis, a comparative analysis is conducted to evaluate the performance of various models, including fine-tuned versions of **Multilingual BERT**, **XLM-RoBERTa**, and **BioBERT**. We also investigate how fine-tuning such models with multilingual data can enhance its performance on the clinical statements of each language. Furthermore, given the importance of Large Language Models (LLMs) in the current landscape of NLP research, the thesis also explores the performance of **GPT-3.5** and **GPT-4** in accomplishing the task without any fine-tuning to assess if it can outperform the fine-tuned models.

The clinical statements are in 3 languages, **Italian**, **Spanish** and **Basque** which presents an added layer of complexity, as the majority of research in relation extraction from clinical documents has focused on the English language. This thesis will contribute to the growing body of research on clinical documents in these languages, helping to bridge the gap in the literature. The inclusion of these diverse languages emphasizes the importance of developing robust multilingual models capable of handling the subtleties and variations across different linguistic and medical contexts.

The thesis is organized as follows: Chapter 1 presents a comprehensive review of the related literature, including an overview of relation extraction techniques and their application in the clinical domain. Chapter 2 talks about the datasets and data splits. Chapter 3 details the methodology employed in this research, including the various steps and the implementation of NLP techniques for relation extraction. Chapter 4 discusses the experimental setup and the results with error analysis while Chapter 5 deals with overall discussions and insights gained and future research avenues. This is followed by the conclusion. All code is made available online. [2]

[1] https://e3c.fbk.eu/clinkart#h.one4aqhz1kh, https://e3c.fbk.eu/testlinkiberlef

[2] https://github.com/goutham794/clinical-relation-extraction

1. Background and Related Works

1.1 Overview of Relation Extraction

Relation Extraction (RE) is a critical component of Information Extraction, aiming to identify and classify semantic relationships between entities present within a text. At its core, RE is about connecting the dots; it seeks to understand how two named entities, such as people, organizations, or medical terms, relate to each other in the context of a sentence or a document.

For instance, in a medical text, the sentence "Penicillin is used to treat bacterial infections" establishes a relation between the drug *Penicillin* and the ailment *bacterial infections*. Here, the relationship could be labeled as *treats*. Similarly, in the sentence "Barack Obama was born in Hawaii", the entities *Barack Obama* and *Hawaii* are related by the *born in* relation.

There are a few salient points to note about Relation Extraction:

- **Granularity and Directionality**: Relations can be uni-directional or bi-directional. The direction of the relation is crucial for understanding the context. For example, the *is a parent of* relation is different from the *is a child of* relation.
- **Relation Types**: Depending on the domain, various types of relations can be identified. In biomedical texts, these might revolve around drug-disease interactions or gene-protein associations. In general news or Wikipedia articles, relations could be about geopolitical alliances, familial relationships, or historical events.
- **Challenges**: Relation Extraction is not without its challenges. Ambiguous language, long-distance dependencies between entities, and nested entities can make the task non-trivial. Furthermore, the vast array of potential relations and the need for domain-specific knowledge make the task especially challenging in specialized fields like medicine or law.
- **Applications**: Understanding relations between entities is pivotal for many applications, including building knowledge graphs, question answering systems, and recommendation engines. For instance, a knowledge graph built for medical literature can help researchers identify new drug-disease interactions or side effects by analyzing the relationships extracted from large volumes of text.

1.2 Approaches to Relation Extraction

An early, innovative approach in Information Extraction was the RAPIER (Robust Automated Production of Information Extraction Rules) system developed by Califf and Mooney [1997]. This system automatically generates pattern-match rules by analyzing pairs of text and their corresponding filled templates. RAPIER employs a bottom-up inductive learning approach, beginning with specific rules, progressively generalizing them to capture broader patterns.

Moving to feature-based methods, Kambhatla [2004] discusses a method for extracting relationship between entities using Maximum Entropy models that integrates many lexical, semantic and syntactic features derived from the text. The features used include:

- **Words:** Words of the mentions and words in between.
- **Entity type:** Both entity types. eg. PERSON, LOCATION, ORGANIZATION, etc.
- **Mention Level:** Mention level of both entities. One of NAME, NOMINAL, PRONOUN.
- **Overlap:** Number of words and mentions separating the pair of mentions.
- **Dependency:** Words and syntactic labels from the dependency tree.
- **Parse Tree:** Paths containing mentions in the syntactic parse tree.

Zhao and Grishman [2005] used kernel methods combined with SVM (Support Vector Machine) on the extracted features for the relation detection task. This work combined different levels of syntactic information (sentence tokenization, sentence parsing and deep dependency analysis) using different kernels. They show that each level of information help improve the performance of relation extraction.

Eventually, feature extraction became automated. Nguyen and Grishman [2015] introduces a convolutional neural network (CNN) approach that reduced dependency on manual feature engineering and external linguistic toolkits. Words in sentences were encoded using pre-trained word embeddings (Word2vec by Mikolov et al. [2013]). They used filters with multiple window sizes in the convolutional layer to extract diverse n-gram features. The paper showed that the CNN approach outperformed traditional feature-engineering based methods in their evaluation.

The pipeline approach to RE was outlined in the work by Bassignana and Plank [2022]. Starting from raw text, the first step is to identify entities and possibly assign them a type. They are either nominals or named entities, and hence it is either Named Entity Recognition (NER) or, more broadly, Mention Detection (MD). After entities are identified, the approaches to RE start to diverge as studies have approached RE from different angles. One common approach is to have a subsequent Relation Classification step as shown in Ye et al. [2019].

Wadhwa et al. [2023] discusses using Large Language Models for the task of Relation Extraction by posing it as a text generation task. They show that few-shot learning using GPT-3 can achieve performance comparable to that fully supervised models. The also show that using a prompting technique known as Chain-of-Thought (COT) where they make the model generate intermediate reasoning steps, can yield performance improvement. The work also discusses challenges with evaluating LLM generated text.

1.3 Transformer Models in NLP

Transformer models, as proposed in Vaswani et al. [2017] have revolutionized the landscape of Natural Language Processing (NLP). These models generally outperformed previous state-of-the-art networks, making them the backbone of modern NLP solutions.

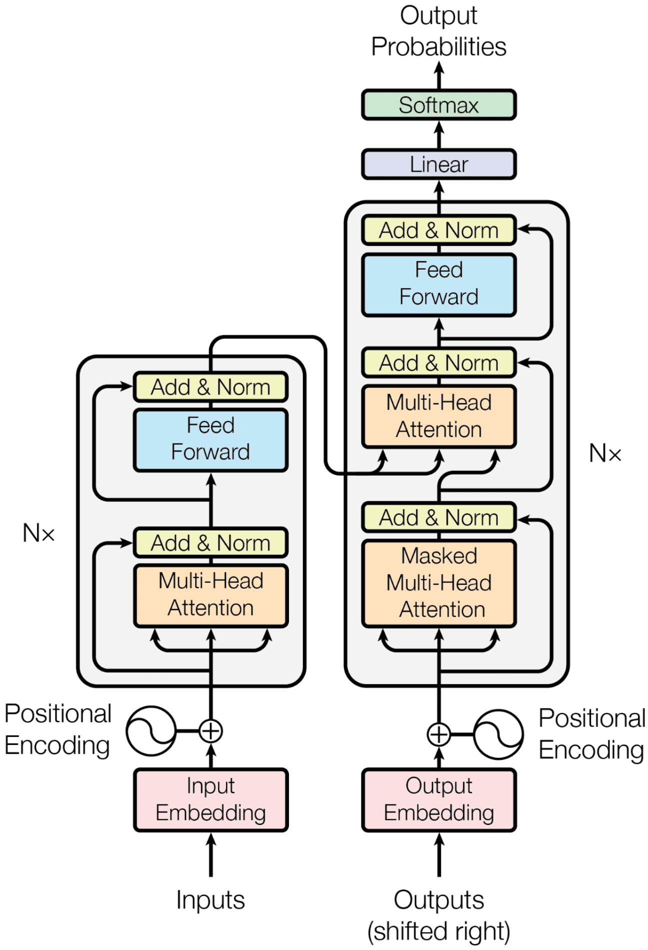

Figure 1.1: Illustration of Transformer model. Figure reprinted from Vaswani et al. [2017]

Basic Structure

The Transformer architecture is based on an *encoder-decoder* structure, which is a common architecture for sequence-to-sequence tasks. The encoder takes an input sequence and converts it into a sequence of hidden representations. The decoder then takes the encoder's output and generates an output sequence, one token at a time. At its core, the Transformer architecture relies on the mechanism of **attention** to draw global dependencies between input and output. Traditional RNNs (Recurrent Neural Networks) and LSTMs (Long Short Term Memory) (Hochreiter and Schmidhuber [1997]) processed sequences word by word, making them inherently sequential. In contrast, the Transformer model processes input data in parallel, ensuring faster computations without compromising the ability to capture sequential information.

Attention

The heart of the Transformer, the attention mechanism, allows the model to weigh the importance of different words in a sentence relative to a given word. For instance, consider the sentence: “The cat was under the car. It then ran away”. When trying to understand the word *it* in the sentence, the attention mechanism can help the model discern whether *it* refers to *the cat* or *the car* mentioned earlier in the text. This capacity to dynamically adjust focus not only enhances the model’s understanding of context within individual sentences but also significantly boosts its ability to handle complex linguistic phenomena such as anaphora and co-reference, which are pivotal for maintaining coherence in longer texts and dialogues.

Working

The working can be better understood by referring to Figure 1.1.

- **Input Embedding**: Both the input sequence (for tasks like translation, this would be the source language) and the output sequence (the target language) are first embedded into a vector representation.

- **Self-Attention Mechanism**: The encoder (and decoder) consists of a stack of self-attention layers. Each self-attention layer learns to attend to different parts of the input sequence.

- **Feed-forward Neural Network**: Each attention output is then passed through a feed-forward neural network (the same one for each position).

- **Stacking Layers**: Several such layers (comprising attention and feed-forward networks) are stacked, which allows the Transformer to learn complex patterns and relationships.

- **Output Sequence** (For tasks like translation): The final layer’s output from the encoder (for the input sequence) is then used as the input to the decoder (for the output sequence). The decoder also has several stacked layers of self-attention and feed-forward networks. In addition, the decoder has cross-attention mechanism that focuses on the encoder’s output. This helps the decoder to know which parts of the input sequence to focus on.

- **Final Linear and Softmax Layer**: The output from the top layer of the decoder is transformed into predicted next-token probabilities using a final linear layer followed by a softmax.

- **Training**: During training, the model is optimized to reduce the difference between its predictions and the actual target output. Commonly used loss functions include the cross-entropy loss.

1.3.1 BERT and its Variants

BERT (Bidirectional Encoder Representations from Transformers) from Devlin et al. [2019] marked a significant shift in NLP by leveraging the Transformer architecture for understanding the context of words in a sentence. It uses a masked language model approach, where random words in a sentence are replaced with a **[MASK]** token, and the model is trained to predict the original word. This mechanism, along with its derivatives like mBERT, RoBERTa, etc has set new benchmarks in various NLP tasks.

With reference to the Transformer architecture as seen in Figure 1.1, BERT is composed of only the **encoder** part, which is the structure on the left. BERT does not need decoder layers because it is not designed to generate text, but rather to produce representations of text that can be used for various downstream tasks, such as question answering, sentiment analysis and named entity recognition.

Multilingual BERT, an extension of the original BERT model, has been pre-trained on a large corpus of text in 104 languages. It uses the same architecture. It has 12 stacked layers and embedding dimension of 768. The languages that were part of the pre-training are the ones with the largest Wikipedia data sizes as Wikipedia is the training data for each language.

1.3.2 XLM-RoBERTa

XLM-RoBERTa, multilingual masked language model by Conneau et al. [2020] has emerged as a powerful tool for multilingual natural language processing tasks. Developed to address the challenges of cross-lingual understanding, XLM-RoBERTa is pre-trained on 2.5TB data of 100 languages, making it particularly adept at tasks involving diverse linguistic datasets.

The model's good performance can be attributed to the novel training approach, which combines the strength of masked language modelling and cross-lingual language modelling. By randomly masking 15% of input tokens with a [MASK] token, the model learns to predict the original token, while also leveraging a multilingual corpus to learn cross-lingual alignments. This enables XLM-RoBERTa to capture subtle nuances in language syntax and semantics, as well as develop a good understanding of linguistic variations across different languages. As a result, XLM-RoBERTa in its time achieved state-of-the-art results in a range of multilingual NLP tasks, including cross-lingual natural language inference, named entity recognition, and question answering.

1.3.3 BioBERT

BioBERT has emerged as an important advancement in the realm of biomedical NLP. Specifically tailored for the biomedical domain, BioBERT is a version of BERT pre-trained on vast biomedical corpora, such as PubMed abstracts, to capture domain-specific knowledge and nuances. This domain adaptation has led to significant improvements in various biomedical NLP tasks. For instance, in the context of named entity recognition (NER) within clinical trial eligibility criteria, Li et al. [2022] show that domain-specific transformer models like BioBERT outperform general transformer models, with the embeddings trained from domain-specific corpora playing a crucial role in enhancing performance. Furthermore, probing experiments by Jin et al. [2019] have revealed that embeddings from models like BioBERT and BioELMo intrinsically carry richer entity-type and relational information, which is pivotal for tasks in the biomedical domain. An important point to note is that BioBERT is trained only on English data, and is not inherently multilingual unlike mBERT and XLM-RoBERTa.

The evolution of transformer models, through architectures like BERT and XLM-RoBERTa, has enhanced language processing capabilities and also democratized access to state-of-the-art NLP technologies across various linguistic and domain-specific contexts.

1.4 Transformer Models in Healthcare

Transformer models have applications in healthcare, where they can be used to derive crucial insights from extensive biomedical data. These models have been pivotal in identifying and extracting a diverse array of relational data critical for improving diagnosis, treatment, and overall patient outcomes. Notable relations include **Drug-Disease** interactions, **Symptom-Disease** correlations, **Drug-Dosage** guidelines, **Lab Tests-Result** interpretations, **Anatomical Relations** (Relationships between diseases or symptoms and specific body parts or systems.), **Patient-Intervention outcomes** (Identifying how different interventions (surgical procedures, therapies) affect patient outcomes), and **Comorbidity patterns**. Each of these relational insights plays a vital role in enhancing the precision and effectiveness of medical interventions. In Fraile Navarro et al. [2023], the authors conducted a systematic review of NER and RE tasks using Transformer models in medical texts, highlighting that Transformer models have gained traction steadily but also talk about the need for proper validation of tools and models and suggest requirement of a robust framework for these tasks in order of improve generalizability.

One notable application is the extraction of chemical-protein interactions from scientific literature, a task crucial for drug design and precision medicine. In this context, Weber et al. [2022] utilized pre-trained transformer language models and modeled the task as a relation classification problem. By integrating textual descriptions of chemicals from the Comparative Toxicogenomics Database, the model achieved an impressive F1 score (task - chemical-protein relation extraction), highlighting the power of transformers in biomedical relation extraction.

Scaboro et al. [2023] conducted an extensive evaluation of 19 Transformer-based models for ADE (Adverse Drug Events) extraction, illustrating the broad capabilities of these models to handle informal and colloquial language used in social media. These studies highlight the impact of deep learning technologies in monitoring and analyzing drug safety from online user-generated content.

1.5 Large Language Models in NLP

Large Language Models (LLMs) represent a groundbreaking advancement in the field of Natural Language Processing (NLP). Essentially, they are deep learning models trained on *vast* amounts of text data, enabling them to generate human-like text based on the patterns they recognize from their training. The emergence of LLMs, such as GPT-4, Llama has brought about unparalleled ability to understand and generate complex textual content.

Models like BERT and its variants are also by technical definition, Large Language Models. These models as mentioned earlier are masked language models and are encoder only. Whereas models of the GPT family are *decoders* and are trained as left-to-right language models. In this work, we distinguish the older generation transformer models such as BERT which are generally smaller in size from the newer generation LLMs which are much larger in terms of parameters.

Following Instructions

The base GPT-3 model is trained to predict the next word and does not follow user instructions well. In Ouyang et al. [2022], OpenAI researchers talk about the training of the InstructGPT models which are better at following user intentions. The paper essentially talks about aligning large language models with user intent.

Instruction fine-tuning is a method where the model is fine-tuned by providing it with instructions and examples. The GPT-3 model is first fine-tuned by supervised learning so that it understands the desired model behaviour. Subsequently, a dataset of rankings of model outputs was collected, which was used to further fine-tune the supervised model through **reinforcement learning from human feedback (RLHF)**.

The resulting models were termed **InstructGPT**. In evaluations, the outputs from the 1.3B parameter InstructGPT model were often preferred over those from the 175B GPT-3, even though the former had significantly fewer parameters. The paper underscores the potential of human feedback in aligning language models with human intent, leading to improvements in output quality.

Characteristics of LLMs

This section details the differences between LLMs like the GPTs and the older generation - the BERT Family of models.

- **Architecture and Training**: BERT models, and their various offshoots, are primarily designed for tasks that require understanding the bidirectional context of words in a sentence. They are trained using a masked language model objective. GPT-3.5 and GPT-4, are generative in nature. They are trained using an auto-regressive language modeling objective, where the model predicts the next word in a sequence based on its preceding words. This allows GPT models to generate coherent and contextually relevant sentences or paragraphs.

- **Scale**: GPT models boast hundreds of billions of parameters, storing an astonishing amount of information and dwarfs the BERT family in terms of size.

- **Transfer Learning and Few-Shot Learning**: BERT models ushered in the era of transfer learning in NLP. Once pre-trained on a large corpus, they can be fine-tuned on specific tasks with a relatively small amount of data, thereby transferring the knowledge from the pre-training phase. GPT-3.5

and GPT-4 introduced the concept of few-shot (or even zero-shot) learning. By simply providing a few examples within the prompt, these models can generalize and perform specific tasks without requiring any explicit fine-tuning.

- **Usage and Versatility**: Due to the generative nature, GPT models are much more versatile in terms of tasks it can be used for, which includes text generation, question-answering, translation, summarization.

1.5.1 QLORA Fine-tuning of LLMs

QLoRA fine-tuning (Dettmers et al. [2023]) has emerged as a significant advancement in the training and optimization of large language models (LLMs), usually used as an alternative to full fine-tuning of an LLM when adapting to domain specific task. QLoRA, or Quantized Low-Rank Adaptation, is a method designed to fine-tune pre-trained language models with high efficiency and minimal computational cost.

The precursor to QLORA - LORA (Hu et al. [2021]) basically worked by freezing the pre-trained model and adding a small adapter module which is trained to learn task-specific knowledge. Hence it is efficient, requiring fewer computational resources and less data compared to full fine-tuning.

QLoRA builds upon the foundational principles of LoRA by incorporating quantization into the adaptation process. Quantization involves reducing the precision of the numerical values used in computations, which can significantly decrease the memory and computational resources required.

1.5.2 Retrieval augmented generation (RAG)

Retrieval-Augmented Generation (RAG) (Lewis et al. [2020]) is a technique that combines the capabilities of retrieval-based models and generative models to produce more accurate and contextually relevant responses from an LLM. The system first retrieves relevant documents or data snippets from a large knowledge base or corpus in response to a query. This retrieved information is then used as a contextual reference for the generative model (context is usually appended to the prompt to the LLM), which synthesizes the final output. By leveraging retrieved content, RAG aims to enhance the factual accuracy and depth of the generated text, allowing the model to produce responses that are coherent and grounded on real-world data. This method is particularly useful in tasks requiring detailed knowledge or precise information, such as question answering. This technique could potentially be valuable for improving clinical decision support systems. The retrieved information should be validated and aligned with medical guidelines. Additionally, the generated responses should be interpreted by qualified healthcare professionals.

1.6 Gaps in Current Research

The extraction and linkage of test results with corresponding textual mentions within clinical statements is relatively less explored, especially at the scale that the current data influx demands.

Several gaps and challenges can be identified from the current state of the art:

- **Domain-Specific Complexity**: While much research has been done on relation extraction, the particular challenge of linking test results to their corresponding tests in clinical documents calls for specialized methods. This task, which combines healthcare and natural language processing, requires approaches that are adept in both fields.

- **Language Limitation**: A majority of studies and models in relation extraction have centered on English language datasets. This focus overlooks the vast non-English medical data landscape, which holds immense value. Italian, Spanish and especially Basque are underrepresented in clinical NLP studies, resulting in a significant gap in resources and research for these languages in the clinical domain.

- **Model Comparisons**: Although numerous models have been used for similar tasks, comprehensive comparisons that include both fine-tuned Transformer models such as Multilingual BERT, XLM-RoBERTa, and BioBERT, and Large Language Models (LLMs) are rare. The trade-offs in effectiveness and efficiency between these models, particularly in the medical field, are not well-documented.

- **Exploration of LLMs**: The rise of large language models like GPT-4 has reshaped the NLP landscape, yet their applicability and superiority in specialized tasks like relation extraction in the clinical domain remain an area ripe for exploration.

2. Data

2.1 About the Data

The dataset employed in this study comprises clinical statements obtained from **E3C**, the multilingual European Clinical Case Corpus (Magnini et al. [2020]). The clinical cases were originally obtained from PubMed articles and other corpora.

Annotation

The following annotations were part of the dataset done on the clinical cases from E3C:

- **Laboratory tests and measurement events** - Includes medical procedures and measuring of physical features.
- **Results** - Results of the tests and measurements. They could be texts or values, possibly followed by units, eg. 120 mb/dL.
- **Pertains-to** - Relation connecting the *result* to the *test event.* These relations could be one-to-one, one-to-many and many-to-one.

The data is structured in the PubTator format, a widely recognized standard for biomedical text annotation. This format effectively captures entity and relation annotations.

An example in Italian in the PubTator format with annotations:

```
100509|t|Donna, 87 anni, ipertiroidismo subclinico, artrosi,
    osteoporosi (fratture T10-T11), ipovisus, AH in terapia
    steroidea cronica; ipertensione; scompenso cardiaco diastolico;
     un ricovero per EPA. Recente embolia polmonare, da allora in
    TAO. Recentemente agitazione e dolore resistente a paracetamolo
    . All'ECG RS 66 bpm, deviazione assiale sinistra, BBD
    incompleto. Chest Pain Score e Wells Score bassi. All'
    ecocardiogramma FE 55%. PA 160/90 mmHg. Giordano positivo,
    dolore paravertebrale bilateralmente. Dopo caduta accidentale
    vivo dolore a livello dorsale. Al quadro rx crolli vertebrali
    da T6 a T8 con pregresso crollo di T12. Procrastinata la
    chifoplastica e prescritto un busto, iniziava cauta
    fisioterapia. Videat oculistico e continuazione della terapia
    steroidea. Prescrizione di teriparatide e vitamina D.

100509 REL 309-315 306-308 66 bpm RS

100509 REL 393-398 387-392 bassi Score

100509 REL 393-398 373-378 bassi Score
```

```
100509 REL 423-426 420-422 55% FE

100509 REL 431-442 428-430 160/90 mmHg PA

100509 REL 453-461 444-452 positivo Giordano
```

2.2 Data Statistics

Lang	Split	Num. of Statements	Num. of Tokens	Num. of Relations
Italian	Train	83	28856	658
	Test	80	26437	612
Spanish	Train	81	28815	597
	Test	80	29668	668
Basque	Train	90	34052	1291
	Test	80	12756	345

Table 2.1: Dataset in numbers

Table 2.1 shows the number of clinical cases in the train and test splits for each language. It further shows how many tokens and relations are present in the data. This data split is part of the original dataset.

2.3 Data Splits

15% of the statements were randomly selected from the training set and put into the validation set for each language.

The statements are *sentence tokenized* as part of the dataset. These sentences form the primary individual datapoints in our dataset. We are interested in detecting relations existing in these sentences, and the annotation is such that there are no inter-sentence relations.

The number of datapoints (sentences) in each split is as given in Table 2.2.

Lang	Train	Valid	Test
Italian	972	141	1064
Spanish	1006	128	1232
Basque	2599	527	1083

Table 2.2: Number of sentences in each data split

3. Methodology

This chapter is split into two sections: the methodology for fine-tuning Transformer models and the methodology for few-shot learning with LLMs.

3.1 Transformer Model Fine-tuning: Methodology

3.1.1 Overview

The application of transformer models to the task of relation extraction from clinical documents in this study unfolds in two consecutive stages - **Named Entity Recognition (NER)** and **Relation Classification**.

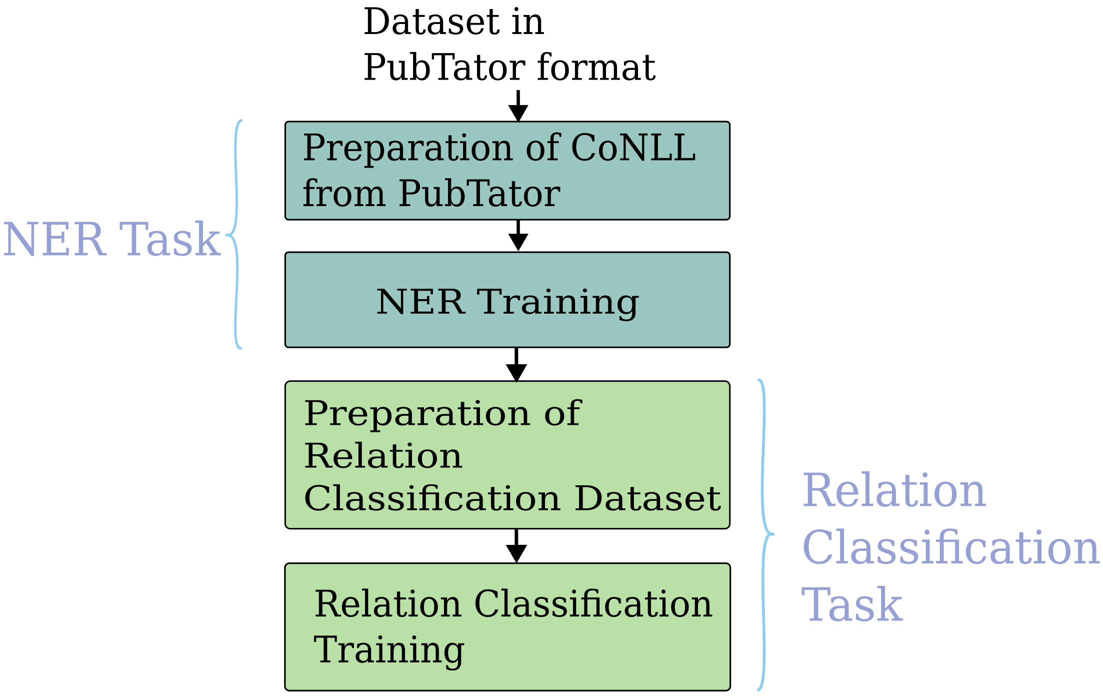
Dataset in PubTator format
Preparation of CoNLL from PubTator
NER Training
NER Task
Preparation of Relation Classification Dataset
Relation Classification Training
Relation Classification Task

Figure 3.1: Flowchart of the Relation Extraction training process

3.1.2 Named Entity Recognition (NER) step

Named Entity Recognition (NER) is the first crucial step in relation extraction, and it entails *identifying* the **test** entities and **result** entities within the data.

Data Processing

The initial data, which is in the PubTator format, needed to be prepared in the CoNLL format. The CoNLL format represents data as sentences with each word on a separate line, and an empty line separating sentences. Each line includes details about the word's features such as its entity type.

The entity types consist of one of five tags - **O**, **B-TST**, **B-RML**, **I-TST**, or **I-RML**. **B** denotes the beginning of an entity, **I** for inside, and **O** for outside. **TST** represents the laboratory test entity, and **RML** signifies the result. A custom script was designed to parse the PubTator annotations and generate data in the CoNLL format, ensuring that no critical information was lost during this transformation process.

Model selection

Choosing an effective model is a critical step. Three models were chosen for this task: Multilingual BERT (**mBERT**), XLM-RoBERTa (**XLM-R**) Conneau et al. [2020] and **BioBERT**. These models have already been discussed in Section 1.3.

mBERT's multilingual capabilities make it well-suited to the task, considering the clinical statements used in this work are in 3 non-English languages. XLM-RoBERTa has a larger architecture and it's promising performance on multilingual NLP tasks made it another good candidate for our NER step. BioBERT, even though not multi-lingual, has been pre-trained on bio-medical data.These models are popular and are widely used in a variety of tasks and applications.

The performance of the NER task does not rely solely on the models. An equally important aspect is the fine-tuning process and the quality of the data fed into these models.

Model Architecture

The Huggingface library (Wolf et al. [2020]) provides a useful "AutoModel" API for the task of Token Classification where the base pre-trained Transformer model is modified with a token classification head.

The base transformer models of mBERT, XLM-RoBERTa and BioBERT modified with a simple token classification head is used. In more detail, the hidden state values of the last layer of the base model are first passed through a dropout layer (Srivastava et al. [2014]) and then through a simple linear classifier with the number of output neurons matching the number of labels in the NER task. The loss is calculated using cross entropy between the logits from the model and the target, which is 1 for the true label and 0 for the others.

Model Training

The training process for NER leverages the capability of transfer learning in our selected Transformer models.

All three models were full fine-tuned on our dataset, which was preprocessed into the CoNLL format. This format is well-suited for NER tasks as it annotates each token with its corresponding entity tag, making it an optimal choice for our use-case.

The models start with their respective pre-trained weights. These weights already encode substantial general language understanding due to the extensive pre-training on a broad multilingual corpus.

The models were trained over multiple epochs. More details on the experiments are in Section 4.2. As mentioned in the previous section, the cross-entropy loss function was used as a measure of dissimilarity between the predicted probability distribution and the actual entity tags.

3.1.3 Relation Classification step

The task of Relation Classification (RC) represents the second stage of this process. It builds upon the results of the Named Entity Recognition (NER) stage. The task is to classify pairs of recognized entities - laboratory test and result (marked using tags : [**TST**] and [**RML**]) - present in a sentence as either positive or negative relations, thus at the end extracting relations from the original text.

Data Processing

This step requires the creation of a new dataset. Each datapoint in this dataset comprises of a sentence with a pair of identified entities: the laboratory test and its associated result. Test entities are highlighted by appending [**TST**] tag as a prefix and suffix and the same is done with result entities using [**RML**] tag.

The methodology to create this dataset differs slightly depending on whether we are preparing data for training or for validation and testing. For the training dataset, we utilize the original test and result entities present in the sentence, creating a separate datapoint for each *test entity* - *result entity* pair. This ensures that the models are trained on the accurate ground-truth entities and relationships. On the other hand, for the validation and testing datasets, we leverage the entities predicted from the first NER stage to create the new dataset. This approach provides a more realistic scenario for evaluation, as it takes into account the potential errors or inaccuracies in the NER stage, reflecting how the system would perform in a real-world, end-to-end setting.

In both cases, each *test-result entity pair* is classified as either having a positive or negative relation, forming the binary class labels for the dataset. Importantly, only sentences that contain at least one *test-result entity pair* pair are included in this dataset.

An example of a positive and negative statement in the relation classification dataset (1=positive, 0=negative) :

```
"L'esame emocromocitometrico ha evidenziato una spiccata
   leucocitosi con [TST]globuli[TST] bianchi pari a [RML]191.000/
   mm3[RML], lieve anemia con Hb 11,1 g/dl e conta piastrinica
   pari a 341.000/mm3.", 1

"L'esame emocromocitometrico ha evidenziato una spiccata
   leucocitosi con globuli bianchi pari a 191.000/mm3, lieve
   anemia con [TST]Hb[TST] 11,1 g/dl e conta piastrinica pari a [
   RML]341.000/mm3[RML].", 0
```

Model selection

For the Relation Classification (RC) stage, the methodology of model selection parallels that of the NER stage. Our focus remained on utilizing Transformer models that were adept at understanding contextual relationships, as the success of the RC stage depends on the model's ability to accurately decipher the association between the identified entities within the context of a sentence.

As such, we again turned to Multilingual BERT, XLM-RoBERTa, BioBERT models, due to their proven track record in similar tasks and their inherently multilingual capabilities and domain knowledge in case of BioBERT.

It is important to note that while the same models are employed in both stages, the focus of their task changes. In the RC stage, these models are used to classify the relation between a pair of identified entities as either positive or negative, based on their context within a given clinical statement. The models, therefore, are not reused across stages but separately fine-tuned and trained for each individual task. This allows the strengths of these models to be leveraged in both entity recognition and relation classification, contributing to a more robust and accurate system overall.

Model Architecture

The "AutoModel" API of Huggingface library for the task of Sequence Classification is used, where the base pre-trained Transformer model is modified with a sequence classification head.

This being a sequence classification task, a pooled output is taken from the base network. For BERT-based models there is a special token ([**CLS**]) added at the start of the text whose embedding represents or captures the entire sequence. The equivalent for XLM-RoBERTa is the <**s**> token. There is a difference in how the pooled output is processed between the BERT-based models and XLM-RoBERTa. For XLM-RoBERTa, the pooled output is passed through dropout, followed by a linear layer that has the same number of output neurons as the input dimension. Then a non-linear activation function (tanh) is applied followed by another dropout and the final classifier that maps the input to the final node that outputs the logits for whether the relation is valid or not.

For the BERT-based models, the architecture is simpler. The pooled output goes through a dropout layer followed by a final linear layer.

Model Training

For the RC task, similar to the NER stage, the models are full fine-tuned on the prepared dataset. The base models that are initially loaded are pre-trained versions of Multilingual BERT, XLM-RoBERTa, and BioBERT. The dataset consists of [test-result] pairs labeled as either positive or negative.

During fine-tuning, each model learns to associate the context of a sentence with the type of relation between the entities. The training is repeated across multiple epochs with the model's performance on the validation dataset monitored.

3.1.4 Multilingual Models

The same methodology as explained above is applied, but on multilingual data *i.e.* all the language datasets are mixed together. This implies training just **one** model (for each of the Transformer models) for the three languages. There are still separate models for the NER Task and the Relation Classification Task. The idea for training Multilingual models is that, when there is limited training data available, such a model can potentially leverage patterns learned from one language on another language, helping the model generalize and ultimately producing a more robust model.

Creating a Multilingual Dataset

A multilingual model can end up biased if there are more dominant languages in the training dataset. For the NER Task, the CoNLL formatted data was taken and **oversampling** was done to ensure equal distribution of all three languages in the dataset. For the RC task, the RC dataset is created as explained in Section 3.1.3. To create the validation split of the RC dataset, the multilingual NER model is used to first predict entities on the validation split of each language, separate RC datasets are created for each language which are then concatenated. So the training and validation splits for both tasks are multilingual. Ultimately, the final prediction on the test set is done separately for each language.

3.2 Few-Shot learning with LLMs: Methodology

3.2.1 Overview

The second part of the methodology focuses on the application of a Large Language Model (LLM), particularly the **GPT-3.5** and **GPT-4** models, for the task of Relation Extraction. Contrary to the typical fine-tuning approach employed in Transformer models, LLMs are exploited for their **few-shot learning** capability, wherein they utilize their extensive pre-training on diverse text data to execute specified tasks based on given prompts.

The LLM-based methodology simplifies the process by encapsulating the tasks of Named Entity Recognition (NER) and Relation Classification into a singular text generation task. Instead of first identifying entities and subsequently classifying their relationships, the GPT model is directly prompted to extract the *test-result* pairs in the clinical text. It can be thought of as **Relation Extraction as Text Generation**.

For efficient utilization of LLMs, precise and well-structured prompts play a critical role. The prompt includes examples of the task at hand to guide the model and help it understand the expected output.

The methodology discussed here provides an in-depth understanding of how the GPT models are employed for the task of Relation Extraction.

3.2.2 Model selection

In the pursuit of implementing an LLM-based approach to the task of Relation Extraction, the models of choice are **GPT-3.5** and **GPT-4**, part of the Generative Pretrained Transformer family developed by OpenAI.

The GPT-3.5 model boasts a remarkable pre-training size of 175 billion parameters, which empowers it to generate coherent and contextually relevant responses.

It has been trained on a diverse range of internet text, equipping it with vast knowledge of different language structures, concepts, and a broad understanding of the world. GPT-4 is considered as a huge improvement over GPT-3.5.

GPT-4 is among the best performing LLMs for a variety of tasks. Facebook's Llama 3 is amongst the best performing open source models but it still lags in terms of multilingual performance, only less than 2% of its pre-training data being non-English languages as mentioned in their release paper (Touvron et al. [2023]).

3.2.3 Few-shot Learning

In contrast to fine-tuning based methodologies, Large Language Models (LLMs) like GPT-3.5, 4 are adapted to our task through a technique known as **few-shot learning**. The premise of this technique is crafting an apt prompt containing a concise task description followed by one or more input-output examples. This approach guides the model towards the desired task without any explicit fine-tuning.

In the context of this study, where we aim to extract relations between tests and results from clinical sentences, the prompt typically starts with a clear explanation of the task. This is followed by examples where each clinical sentence is paired with its corresponding extracted relations.

The input to LLMs and LLM APIs is usually a list of messages. This list of messages simulates a chat history. Each of the messages has two properties, that are, **role** and **content**. The role could be one of "system", "user" or "assistant". The purpose of "system" message is to provide the overall instructions to do the task. Following the system message would be alternate messages from "user" and "assistant" where one can provide the examples that the LLM can learn from, on what an input looks like and what an output should look like. So the first user message would have the clinical statement and the first assistant message would have the extracted relations in the format of our choice, and so on. Ultimately, this chat would be fit into chat template (done on the server side of the API) and converted into a single string which the LLM is prompted with. The other way to prompt an LLM is to fit everything in a single user message which contains instructions and example(s) all in one piece of text. Both of these techniques were experimented with and results vary with both. The following is an example of what a single user message to the LLM for our task could look like:

```
I have a task which is to extract mentions of laboratory tests and
    their results from clinical statements. Here is an example of
   text and output:

Text:
The patient’s blood glucose level is 180 mg/dL.

Output:
blood glucose level | 180 mg/dL

Notice: in the output you first write the result and then the name
    of the test. They are separated by "|".
```

```
Now give me the output for the following text:
{New Statement}
```

The model's task is then to extract the relations from the [New Statement] in the manner shown in the examples. The prompt language and structure can significantly influence the performance of the model, and iterative refinement of the prompt is an important step.

There is no training phase for an LLM model using this approach. Prompt engineering is usually done to test various prompts styles on sample inputs. Prompts are generally made to be informative but concise.

3.2.4 LLM Output Post-Processing

Evaluating the relation extraction task is done with a script that interprets predicted output in the PubTator format, comparing it with the gold standard PubTator. As seen in Section 2.1, the PubTator upholds a certain format, which the script also adheres to. The format necessitates the inclusion of start and end indices of each entity that is part of the relation. This is not part of the LLM output (as we do not require it to) and are appended in the post-processing step. The post-processing step also removes any hallucinations or improperly formatted generations. Model hallucination here refers to when the LLM generates text which is not present in the clinical statement.

Additionally, in line with the annotation guidelines, only the first token of the test entity should be represented in the PubTator. This necessitates the removal of any extra tokens in this post-processing step.

The rationale behind these manual adjustments is to make the job of the LLM as easy as possible. By taking on these more trivial tasks manually, we aim to allow the LLM to focus on its primary function without getting bogged down with these nuances.

3.3 Vocabulary-transfer baseline

This model acts as a simple baseline. In this method, the model identifies an entity (test or result) if the same entity also happened to be in the training split of the data. Additionally, it utilizes regular expressions, which were created from the training data, to detect result entities typically denoted as values accompanied by units (for instance, 100 mg/dL would be detected using an appropriate regex). As for the RC step, any pair of test entity and result entity that co-occur in the same sentence is considered as a positive relation.

3.4 Evaluation and Metrics

For **both** the NER and the Relation Classification tasks, the evaluation is done using these metrics:

- **Precision**: This metric calculates the ratio of correctly predicted positive observations to the total predicted positives. In essence, it answers the question: Of all the entities (or relations) the model labeled as positive,

how many were actually correct?
In the following equations, **TP** - True Positive, **FP** - False Positive, **FN** - False Negative.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.1}$$

- **Recall**: Recall focuses on the actual positives and computes the ratio of correctly predicted positive observations to all the actual positives. It provides insight into how many of the actual positive entities (or relations) our model was able to capture.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.2}$$

- **F1 Score**: Striking a balance between Precision and Recall, the F1 score is the harmonic mean of the two. It ensures that we don't lean excessively towards one metric at the expense of the other, offering a more holistic view of the model's performance.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.3}$$

With the Transformer models, evaluation is done with these metrics for both NER and the final Relation Extraction task. With the LLMs, there is no NER phase, only the final evaluation is done.

Metrics Averaging for NER

In the dataset there exists two entity types, Result entities and Test entities. The metrics mentioned above can be averaged in three different ways to provide an overall summary metric:

- **Micro-Averaging** : Global calculation of metrics by considering the overall true positives, false positives and false negatives, disregarding the entity types. If the goal is to simply look at overall performance, this is a good approach. However if there is minority entity type which is is valued, then micro averaging will not give a good picture of model performance.

- **Macro-Averaging** : Precision, Recall and F1 are calculated separately for each entity type separately and averaged. This is used when we want the model to perform well on all entity type regardless of the frequency.

- **Weighted-Averaging** : This is similar to Macro-averaging except that the averaging is weighted by the frequency of each class.

Macro-averaging was chosen as the averaging method for the NER metrics. In the overall task of Relation Extraction, the NER task forms a preliminary step before the Relation Classification step. In such a case both entity types are equally important. There also exists a class imbalance in the training set, with the ratio of Test entities to Result Entities at 46:54. Macro-averaging would ensure equal consideration for both classes which also aligns with the final task of Relation Extraction.

NER Evaluation Modes

This is regarding the *strictness* of the NER evaluation. Seqeval (Nakayama [2018]) is a popular Python library for evaluation of sequence labelling tasks. It supports a **strict** mode and a **default** or more lenient mode.

The following is a simple example demonstrating how these modes differ, taken from the github repository of the library. Displayed are the macro-averages for the given sequences, although all averaging methods return the same values :

```
true_values = [["B-NP", "I-NP", "O"]]
predicted values = [["I-NP", "I-NP" "O"]]

default
-------
precision = 1.0
recall = 1.0
f1-score = 1.0

strict
------
precision = 0.0
recall = 0.0
f1-score = 0.0
```

In this example, for the first token, the entity type was correct, but tagged with an ‘I’ instead of a ‘B’. The default mode allows for this.

The **default** mode was chosen for the evaluation. It is also important to note that neither of these modes allow for partial matches.

Relation Extraction Evaluation

The same metrics Precision, Recall and F1-Score are used. Only one type of relation is dealt with, hence the concept of averaging metrics is not relevant here. An evaluation script written in Java was used, available on GitHub [1], which was originally written for BioCreative V CDR task (Li et al. [2016]) for chemical disease relation extraction.

The script takes as arguments, PubTator formatted files of the model’s output and the gold results. The output (and the gold results) contain the text spans of the result entities and test entities that form the relation. A relation prediction is considered correct if the start and end indices of both entities are correct and the order of the relation is correct.

[1] https://github.com/JHnlp/BioCreative-V-CDR-Corpus

4. Experiments and Results

4.1 List of Experiments

- **Fine-tuned Transfomers - Experiments**

  1. Fine-tuning of mBERT, XLM-RoBERTa and BioBERT on the three language datasets (separately), on the NER Task followed by the Relation Classification Task.
  2. Fine-tuning **Multilingual** models with mBERT, XLM-RoBERTa and BioBERT, on the NER Task followed by the Relation Classification Task.

- **Experiments with LLMs**

  1. Few-shot learning with GPT-4 and GPT-3.5 with older and newer versions of the models directly on the Relation Extraction task, with prompting experiments.

4.2 Fine-Tuned Transformer Models - Experimental Setup

The data and methodology for the experiments have been detailed in Chapter 2 and 3. The experiments were conducted using a setup featuring an NVIDIA RTX 4000 Ada GPU with 20 GB of VRAM, supported by 50 GB of RAM and 9 virtual CPUs. Simpletransformers library [1] was used for its ease of use, however the library code was modified where necessary to support our experiments. The experiments were run with an initially chosen, fixed random seed. Extensive hyperparameter tuning was done for all the experiments which is detailed in the following section.

4.2.1 Hyperparameter Tuning

Hyperparameters control the learning process of the neural network but are not part of the network itself. Tuning them well can help the model achieve higher metrics. W&B Sweeps of the library Weights & Biases [2] was used for running the model using a grid of hyperparameter values.

The main methods of hyperparameter tuning are **Grid Search**, **Random Search** and **Bayesian Optimization**. Grid search is simply an exhaustive search over the predefined set of hyperparameter values. Random search selects random combinations of hyperparameter values from distributions over the set of parameter values. Bayesian optimization aims to balance between exploitation and exploration. Exploitation being choosing the best based on currently available information and exploration being seeking more information (Yu and Zhu

[1] https://simpletransformers.ai/

[2] https://github.com/wandb/wandb

[2020]. **Bayesian Optimization** was chosen in order to reduce the searching time over the large grid space. **Maximizing the F1** score on the validation set was chosen as the objective.

The Hyperparameter grid

Table 4.1: Hyperparameter Grid

Parameter	**Values or Range**
Learning Rate	$[1e-5, 1e-4]$
Batch Size	8, 16, 32
Number of Epochs	4, 5, 6
Optimizer	AdamW, Adam
Weight Decay	$[0.05, 0.1]$
Warmup Ratio	0.05, 0.1
Max Gradient Norm	1.0, 2.0
Scheduler	constant schedule, polynomial decay schedule with warmup cosine schedule with warmup linear schedule with warmup

Table 4.1 shows all the hyperparameters and the set of values tested with. A brief explanation of all the hyperparameters:

- **Learning Rate**: Learning rate represents the step size that would be taken during gradient descent optimization. It influences the rate at which the model weights would change during training (Goodfellow et al. [2016]). Too high a value can result in the model converging sub-optimally and too low can cause long training times.

- **Batch Size**: It is the number of training examples processed in one forward and backward pass. According to Goodfellow et al. [2016] larger batch sizes provide better gradient estimates but there is a dimishing return with increasing size. Lower batch sizes provide noisy gradients which can have a regularizing effect (Wilson and Martinez [2004]).

- **Number of Epochs**: An epoch refers to one complete pass through the entire training set. A low number can results in an underfit where the model has not learnt the data well enough and too many epochs can result in overfitting where the model does not generalize well.

- **Optimizer**: Optimizers are the algorithms that determine how the model parameters or weights are updated based on the gradients calculated during the backpropagation phase. **Adam** (Kingma and Ba [2014]) is an algorithm where the learning rate is adapted for each parameter based on its past changes making the learning fast and stable. **AdamW** (Loshchilov and Hutter [2019])is an improved version of Adam, is different on how weight decay is applied.

- **Weight Decay**: Weight decay is a method of regularization by adding the L2 norm of the weights to the loss function. The weight decay value is the parameter that controls the strength of the penalty (Goodfellow et al. [2016]). The range of values is from 0.0 to 0.1 which covers from no regularization to a moderate amount of regularization.

- **Warmup Ratio**: It is the ratio of the total training steps that will be taken for the warm up phase when the learning rate is increased from zero to its peak value. In Izsak et al. [2021], the authors pre-trained BERT and found that values around 0.06 were optimal.

- **Max Gradient Norm**: The gradient norm is clipped to the value provided. It helps in better optimization in regions of sharp loss surface.

- **Scheduler**: Schedulers are for adjusting the learning rate during the training phase of the model. Constant schedule implies a constant learning rate. Linear schedule with warmup - An initial *warmup* phase during which the learning rate increases from zero to the set learning rate, following which it decays linearly. For polynomial decay the the decay follows a polynomial function and for the cosine schedule it follows the cosine function (Wolf et al. [2020]).

  All the optimal hyperparameter values of the models for all experiments are shown in Section 5.6. In total, 24 hyperparameter tuning experiments were conducted across all models and languages.

4.3 Fine-Tuned Transformer Model Results

Training of the models on both tasks was done using the optimal hyperparameters obtained. The validation set performance plots during training for NER and RE are in the Appendix in Sections 5.7 and 5.8 respectively.

In this section, we look into the test set results of both tasks in our methodology. The test set remained untouched during the model development phase, and the performance on this set serves as a final model evaluation. The results are also analyzed to gain more understanding of the performance of the various models.

For each task, there are **six** fine-tuned Transformer models. Hence it is important to clarify the naming convention used:

- The model name on its own indicates a single language model. So just **mBERT** in Italian section indicates the mBERT model fine-tuned only on the Italian language data.

- Use of “Multi” suffix: eg. XLM-RoBERTa-Multi. The “Multi” suffix indicates that the model was fine-tuned on multilingual data, and then evaluated on the particular language data. (mBERT and XLM-RoBERTa are inherently multilingual due to their pre-training data, and here the “Multi” suffix does not refer to that.)

A point to note is that the results from the NER phase implicitly set an upper bound for the Relation Classification results. Given that these tasks are conducted sequentially, any discrepancies or errors in the NER stage directly influence the final Relation Extraction performance.

4.3.1 NER Test Set Results

Tables 4.2, 4.3 and 4.4 show macro-averaged Precision, Recall and F1 for the fine-tuned Transformer models on the task of Named Entity Recognition in Italian, Spanish and Basque respectively. In this section the test entity will be referred to as TST and the result entity as RML.

Italian

Model	Precision	Recall	F1
mBERT	0.69	0.80	0.74
XLM-RoBERTa	0.73	0.75	0.74
BioBERT	0.74	0.72	0.73
mBERT-Multi	0.75	**0.84**	0.79
XLM-RoBERTa-Multi	**0.79**	0.82	**0.80**
BioBERT-Multi	0.78	0.77	0.77

Table 4.2: Macro-averaged NER metrics - Italian

Table 4.2 shows that the multilingual models consistently beat the monolingual models in Italian. Training with all language datasets together proved useful to create a better performing model. Overall the best F1-score is achieved by the largest model, XLM-RoBERTa, trained on multilingual data.

Fine-grained Evaluation: Table 5.9 in the Appendix shows the F1 scores separately for the two entity types. In general, the performance on RML is higher than that for TST. This result can be explained by looking at the distribution of RML entities and TST entities. In the case of RML, there are a few entities that are frequent in the dataset. For example the entity `nella norma` (translated: in the normal range) occurs 23 times in the 503 RML entities in the test set. Other common RML entities are different forms in Italian of the words “positive” and “negative”. The implication of this being, RML should be an easier entity to predict in Italian.

Qualitative Error Analysis: The most common missed TST entities across all six models were `esame` and `esami` which mean “exams” or “tests” in Italian. So this is a scenario where the word “test” is annotated as the TST entity. The following is an example sentence where the TST entity was either wrong or missed by all the models.

> L’esame $_{\texttt{[TST]}}$ obbiettivo capo, collo, cuore, torace e addome
> erano negativi$_{\texttt{[RML]}}$.
> Translation: Physical examination of the head, neck, heart,
> chest and abdomen were negative.

In the above sentence *L’esame* has the label **B-TST** AND *negativi* has the label **B-RML**. The result entity (*negativi*) was correctly predicted by all entities. With regards to the test entity, all models made varied predictions. mBERT predicted *capo* (head), *collo* (neck) and *cuore* (heart) head as the test entities. Other models made similar predictions of the various anatomical areas subjected to the tests.

Concerning RML entities, the monolingual models missed in some cases, the entities which were different forms of Italian for “positive” and “negative”. However these cases reduce with the multilingual models. Long measurement entities were hard for all the models. eg. `pari 0 inferiori a 1.5 mg/dl` (translated: equal to 0 less than 1.5 mg / dl).

Spanish

Model	Precision	Recall	F1
mBERT	0.70	0.80	0.75
XLM-RoBERTa	0.70	0.79	0.75
BioBERT	0.74	0.77	0.76
mBERT-Multi	0.73	0.83	0.77
XLM-RoBERTa-Multi	**0.77**	**0.84**	**0.80**
BioBERT-Multi	0.76	0.78	0.76

Table 4.3: Macro-averaged NER metrics - Spanish

Table 4.3 shows the metrics for the Spanish NER Task, which were again, dominated by XLM-RoBERTa-Multi. The story of multilingual models outperforming continues for Spanish.

Fine-grained Evaluation: The entity specific F1-scores can be seen in Table 5.9. Unlike Italian, some of the models have higher F1-scores for the TST entity. The reason for this could be that, the Spanish statements also have many qualitative RML entities and longer complex measurements. eg. `límites altos de la normalidad` (translated: high limits of normal range). These entities would be harder to detect than simple measurement entities.

Qualitative Error Analysis: The missed TST entities were generally of two categories. One category referring to general terms such as “examination”, “loss” and the other referring to specific medical tests such as “Lowenstein”, “IgM” (Immunoglobulin M). XLM-RoBERTa did not perform well with the general terms which improved marginally with the multilingual model. mBERT struggled less so with the general terms and more with the medical tests and procedures. BioBERT and BioBERT Multilingual had the most diverse range of missed TST entities.

With the RML entities, looking only at the top misses, the performance of the monolingual and multilingual versions of the model are similar. The difference is that monolingual models have a long-tail distribution of missed entities. The commonly missed entities are non-medical terms, eg. `sin alteraciones` (without alterations), `inespecíficas` (non specific). In general, complex measurement entities were hard for all the models eg. `29,5 cm x 27,5 cm x 16 cm`.

Basque

Model	Precision	Recall	F1
mBERT	0.75	0.71	0.71
XLM-RoBERTa	0.74	0.87	0.80
BioBERT	0.73	0.82	0.77
mBERT-Multi	0.80	0.84	0.82
XLM-RoBERTa-Multi	**0.84**	**0.85**	**0.85**
BioBERT-Multi	0.77	0.84	0.80

Table 4.4: Macro-averaged NER Test metrics - Basque

Table 4.4, shows the NER results for Basque. Following the trend set by Italian and Spanish, XLM-RoBERTa Multi achieves the highest F1-score.

Fine-grained Evaluation: Looking at the entity-specific F1-scores (Table 5.9), the RML scores are higher or at-least match the metrics for the TST entities, similar to Italian. The reason, as hypothesized earlier is, the RML entities are usually measurements or words such as “positive” and “negative”.

Qualitative Error Analysis: mBERT had the highest diversity in missed TST entities, showed weakness in recognizing both common and specialized medical terms in Basque. mBERT missed basic terms like `sukarra` (fever) and complex ones like `esplenomegalia` (splenomegaly). XLM-RoBERTa had fewer misses on general terms. Multilingual models generally performed better than their monolingual counterparts.

4.3.2 Relation Extraction Test Set Results

Tables 4.5, 4.6 and 4.7 contain the Precision, Recall and F1 of the final Relation Extraction on the Italian, Spanish and Basque test statements respectively, for the Fine-tuned transformer models. The Vocabulary-transfer baseline is also included in the results.

Italian

Model	Precision	Recall	F1
mBERT	0.51	**0.63**	0.56
XLM-RoBERTa	0.63	0.57	0.60
BioBERT	0.60	0.52	0.56
mBERT-Multi	0.60	**0.63**	**0.61**
XLM-RoBERTa-Multi	0.60	0.59	0.60
BioBERT-Multi	**0.64**	0.51	0.57
Voc. Transfer Baseline	0.30	0.32	0.31

Table 4.5: Relation Extraction metrics - Italian

mBERT-Multi achieves the highest F1, but only 0.01 ahead of XLM-RoBERTa-Multi. The mBERT model gained significantly when trained on multilingual data.

Fine-grained Evaluation: Further error analysis was done looking at the three types of errors - **spurious** Relations, **missed** Relations and **partial-match** Relations. Spurious relations are those that were predicted by the model but are not present in the true relations set. A partial-match is defined as a relation predicted by the model that has at least some overlap for both the RML and the TST entities. The directionality of the relation has to be correct.

Figure 4.1 shows the counts of these errors for each model. mBERT model has the highest number of spurious relations, which concurs with lowest precision as seen in Table 4.5. BioBERT-Multi has the lowest count of partial matches at 25 and XLM-RoBERTa-Multi has the highest at 43. The monolingual versions of these models also have similar counts. BioBERT-Multi has the highest precision, however this model makes the lowest number of total predictions, hence the low recall rate.

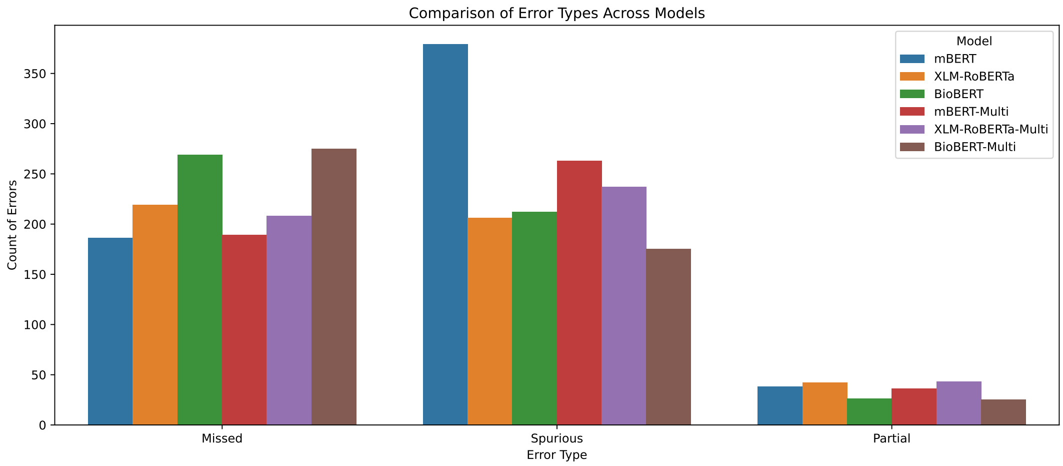

Figure 4.1: Counts of Relation Extraction Error Types for Italian

Qualitative Analysis: The most common **missed** relation across all six models is (`nella norma, esame`)[(in normal range, test)]. In Section 4.3.1, under the analysis for Italian, it was noted that `esame` was the amongst highest missed TST entities. As noted in the referred section, this is a case where the word for "test" is annotated as part of the relation instead of the name of the test or anatomical part subjected to the test. However, in the training set some relations that share the same RML entity are: (nella norma, PCR), (nella norma, PTT), (nella norma, immunoglobuline), where either name of the test or the specific measurement done is part of the relation. This particular pattern of missed relation can be said to stem from the peculiarities of the way the test dataset was annotated.

Looking at the six models together, none of them produced a uniquely missed relation. However, grouping a model and its multilingual version together, there are uniquely missed relations. Both monolingual and multilingual versions of mBERT struggled with relations containing abbreviated test names. Some of the relations missed by them were: (positivo, MTD), (debolmente positivi, tTG), (negativi, LKM).

XLM-RoBERTa models, on the other hand did not do well with complex and longer medical terms eg.(746 μ molL, iperammoniemia). They also struggled with entities expressing ranges, eg. (ai limiti inferiori del range, sideremia) (translation: at the lower limits of the range). BioBERT models having the lowest recall, missed a wider variety of relations, including ones with measurement units and qualitative assessment of results.

Spanish

Model	Precision	Recall	F1
mBERT	0.46	0.63	0.54
XLM-RoBERTa	0.59	0.60	0.60
BioBERT	**0.62**	0.58	0.60
mBERT-Multi	0.61	0.62	0.61
XLM-RoBERTa-Multi	**0.62**	**0.66**	**0.64**
BioBERT-Multi	**0.62**	0.61	0.62
Voc. Transfer Baseline	0.17	0.30	0.22

Table 4.6: Relation Extraction metrics - Spanish

As seen in Table 4.6, XLM-RoBERTa-Multi emerged as the top model for Relation Extraction in Spanish with an F1 score of 0.64. The F1 scores range from 0.54 (mBERT) to 0.64 (XLM-RoBERTa-Multi). Both versions of Bio-BERT tie for the highest precision with XLM-RoBERTa-Multi.

Fine-grained Evaluation: Figure 4.2 shows the counts of the spurious relations, missed relations and partial-match relations for Spanish. As in Italian, mBERT stands out with high count of spurious relations. The benefit of multilingual training is evident in how this error count got reduced in mBERT-Multi. XLM-RoBERTa, Bio-BERT and mBERT-Multi have very close F1 scores but their error profiles are very different. BioBERT has close to 20% more missed relations than mBERT-Multi. mBERT-Multi has the highest count of partial-match errors at 49. This might point to the need for improvement in entity boundary detection in the NER step.

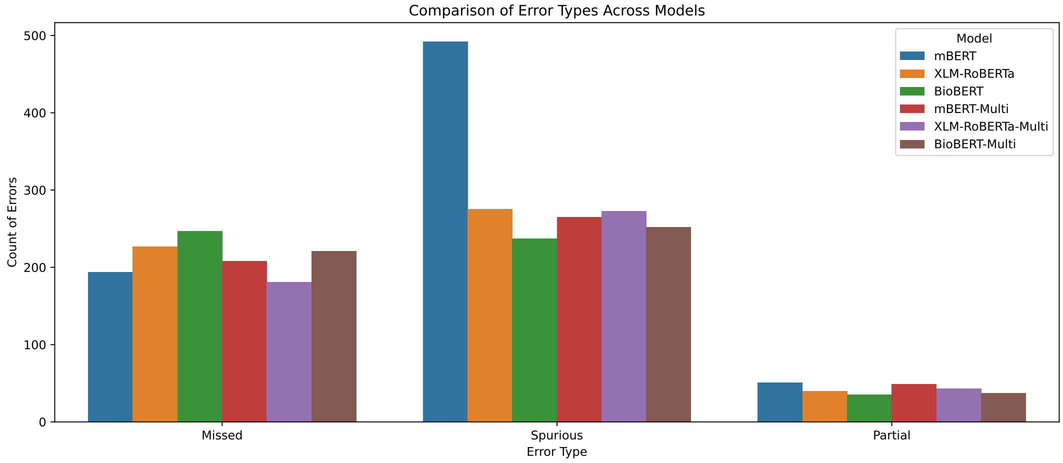

Figure 4.2: Counts of Relation Extraction Error Types for Spanish

Basque

Model	Precision	Recall	F1
mBERT	0.75	0.49	0.59
XLM-RoBERTa	**0.75**	0.77	**0.76**
BioBERT	0.73	0.71	0.72
mBERT-Multi	0.74	0.48	0.58
XLM-RoBERTa-Multi	0.67	**0.79**	0.72
BioBERT-Multi	**0.75**	0.70	0.73
Voc. Transfer Baseline	0.18	0.36	0.24

Table 4.7: Relation Extraction metrics - Basque

As seen in Table 4.7, XLM-RoBERTa achieved the highest F1 score (0.76) but also outperformed its multilingual counterpart by 0.04 points, indicating a stronger performance in the monolingual model for Basque. This finding contrasts with results from Italian and Spanish where multilingual training typically enhanced F1 scores. These discrepancies highlight the unique characteristics of Basque. Notably, the best F1 score for Basque was 0.12 points higher than the highest score observed for Spanish. These results should not be compared as they are different datasets, however a possible reason for this difference is that the number of relations in the Basque test statements is significantly lower (as seen in Table 2.1)

Fine-grained Evaluation: Figure 4.3 illustrates the distribution of the error types - spurious, missed and partial-match relations for Basque. Unlike Italian and Spanish, where mBERT had notably high spurious relation counts, in Basque, both mBERT and mBERT-Multi predominantly show higher counts of missed relations. Conversely, XLM-RoBERTa-Multi exhibits a significantly higher spurious error count.

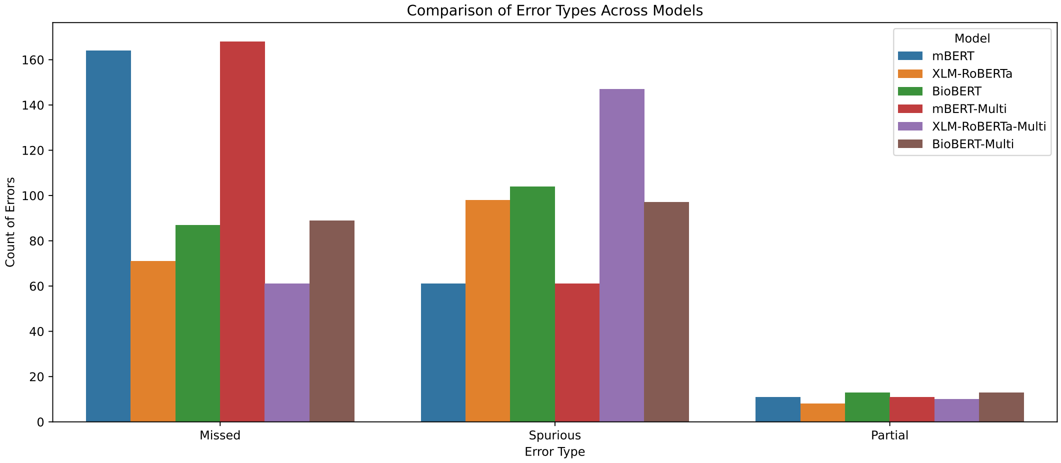

Figure 4.3: Counts of Relation Extraction Error Types for Basque

Qualitative Analysis: The mBERT models exhibited a broad range of missed relations with simple descriptive terms like `normala` (normal) to more complex medical terms and measurements. Both monolingual and multilingual versions demonstrated similar patterns in missed relations. XLM-RoBERTa models effectively handled relations involving ambiguous and generic terms, which are not explicitly medical. This suggests a robust understanding of broader language contexts. These models, however, faced challenges with relations that involved abbreviated medical terms such as `EMG`, `IgE` and `PCR`. This probably stems from inadequate representation in diverse contexts within the training data. BioBERT models were also good at handling relations that included non-medical terms. Multilingual models tended to make the same mistakes as the monolingual ones, which suggests that performance should be improved by addressing the training data deficiencies.

4.4 LLMs Experimental Setup

Few-shot learning using Large Language Models (LLMs) does not involve a training phase, so the experimental setup here is rather different. The experiments were done using the Chat Completions API of the official Open AI API Python library [3].

- **Temperature Setting**: Temperature is a parameter used in the sampling of the model's output distribution. In essence, it influences the randomness of the model's predictions. A high temperature (e.g., close to 1.0) makes the output more random, whereas a lower temperature (e.g., close to 0.0) makes the model's output more deterministic, focusing on the most probable output. In our experiments, we set the temperature to **0.0**. This choice was made to ensure that the LLM provides the most **deterministic**

[3] https://github.com/openai/openai-python

and confident output, minimizing the chances of erratic or random relation predictions, especially given the critical nature of clinical statements.

- **APIs and Model Selection**: We utilized **three** versions of OpenAI's models using both legacy and recently introduced model updates to fully understand their capabilities: Initially, we tested with **gpt-4-0314**.

  In a subsequent phase of testing, we incorporated the newer model updates: **gpt4-turbo-2024-04-09** and and also **gpt3.5-turbo-1106**. This inclusion enabled us to assess advancements in model performance with the enhancements in these newer iterations.

  This approach not only broadens our understanding of the evolution of model capabilities over time but also allows us to contrast the performances of these state-of-the-art models, providing insights into their ongoing development and potential limitations across different versions.

  For clarity and ease of reference, the following naming conventions will be adopted:

  - **gpt3.5-turbo-1106** as **gpt3.5-turbo**
  - **gpt-4-0314** as **gpt4**
  - **gpt4-turbo-2024-04-09** as **gpt4-turbo**

- **Prompt Engineering**: Prompt Engineering is an important step as a well crafted prompt can make a considerable difference in performance. The Chat Completions API accepts a list of messages as input. As discussed in Section 3.2.3, there are two ways of prompting, which would be referred to as follows:

  - **Single-message Prompting**: a single "user" message with instructions and example.
  - **Chat Prompting**: A list of messages, simulating a chat. The system message contains instructions.

  Both of these techniques were experimented with. The other experimentation done was with the language of the prompt. One was to to simply use English as the language, the other was to use the language of the clinical statement. The non-English prompts with Single-message prompting approach are in the Appendix, first part of Section 5.11. The English prompt with Chat-Prompting approach is in Section 5.11.4. The instructions are kept simple and to the point. To be noted: The example included in the prompt is an entire clinical statement and not just one sentence. For Italian, this was 590 tokens long with 43 relations present.

- **Number of examples in prompt**: One example was used in the prompt. The example was chosen to be a long one covering a wide range of relations. Using multiple examples would have made the prompt much longer in terms of token length, making it expensive.

Lang	Model	Precision	Recall	F1
Italian	gpt-3.5-turbo	0.50	0.31	0.39
	gpt-4	0.39	**0.52**	0.45
	gpt-4-turbo	**0.71**	0.38	**0.50**
	Voc. Transfer Baseline	0.30	0.32	0.31
Spanish	gpt-3.5-turbo	0.56	0.29	0.39
	gpt-4	0.43	**0.46**	0.45
	gpt-4-turbo	**0.71**	0.43	**0.54**
	Voc. Transfer Baseline	0.17	0.30	0.22
Basque	gpt-3.5-turbo	0.55	0.26	0.35
	gpt-4	0.48	**0.54**	**0.51**
	gpt-4-turbo	**0.70**	0.36	0.47
	Voc. Transfer Baseline	0.18	0.36	0.24

Table 4.8: Relation Extraction test results for Few-shot learning with LLMs

4.4.1 Relation Extraction Results - LLMs

Table 4.8 presents the performance of Few-shot learning approach with Large Language Models (LLMs) across the three languages. The gpt-4-turbo models generally performed best with non-English prompts with single-message prompting. Basque was the only exception where an English prompt with the chat-prompting performed better, but only by 0.01 points. The older gpt-4 model was only tested with non-English prompt as the model became deprecated at the time of testing. With gpt-3.5-turbo, English prompt with chat-prompting approach generally performed the best. The only exception was Italian where the non-English prompt with single user message outperformed, again only by a small margin (0.02).

In Table 4.8 the scores displayed are with the prompts that generally performed the best for the model. They are as follows:

- **gpt-4-turbo** - Non-English prompt with Single-message Prompting.
- **gpt-4** - Non-English prompt with Single-message Prompting.
- **gpt-3.5-turbo** - English prompt with Chat Prompting

The results of gpt-4-turbo with English prompts and Chat prompting approach are in Table 5.10. The results of gpt-3.5-turbo with non-English prompts and Single-message prompting approach are in Table 5.11. These were the second best performing approaches.

All the models managed to beat the baseline. Gpt-4 models (both turbo and legacy) consistently outperformed gpt-3.5-turbo across all languages. With the exception of Basque, gpt-4-turbo was the best performing model. Gpt-4-turbo achieved the highest precision scores for all languages (0.70 - 0.71). Notably the legacy version, gpt-4 had the highest recall scores. Gpt-4-turbo has much higher Precision at the cost of Recall.

4.4.2 Qualitative Error Analysis

The error analysis was done on the results of gpt4-turbo and gpt-3.5-turbo. Gpt-3.5-turbo showed a broader range of missed relations, including both basic and complex medical terms and measurement entities. In the Italian test statements, there was only one relation which was missed by the best fine-tuned Transformer model but not by gpt-4-turbo. Here is part of a sentence, originally in Italian, illustrating an example error made by gpt-4-turbo. The sentence has been translated to English:

> ... and increase in fat mass and active cell mass with reduction of extracellular water (R/H$_{\texttt{[TST]}}$ + 26.5 ohm/m$_{\texttt{[RML]}}$).

In the above sentence the annotated RML entity is `+ 26,5 ohm/m`, whereas the prediction made by gpt-4-turbo is `26,5 ohm/m`. The fine-tuned Transformer models do not make this mistake as they have learned the peculiarities of the annotations from the training data. This shows one of the limitations of using an LLM for such a task, with only an in-prompt example.

An example of a "spurious" relation predicted by gpt-3.5-turbo in a Spanish statement, translated to English:

> 4 cycles of chemotherapy based on carboplatin and gencitabine were administered, avoiding cisplatin due to cardiac involvement.

In the above sentence gpt-3.5-turbo predicted `[4 cycles]->[chemotherapy]` as a relation. This is an obvious error as the relation does not refer to the result prediction of a medical test. Gpt-4-turbo does not make this error. In this instance, it is the capability of gpt-3.5-turbo as a smaller LLM, that gets exposed.

5. Discussion

This work examined the performance of six fine-tuned Transformer models and three GPT models with Few-shot learning on the task of Relation Extraction in three languages.

The compiled Relation Extraction F1 scores of all the models are plotted in the Figures 5.1, 5.2 and 5.3.

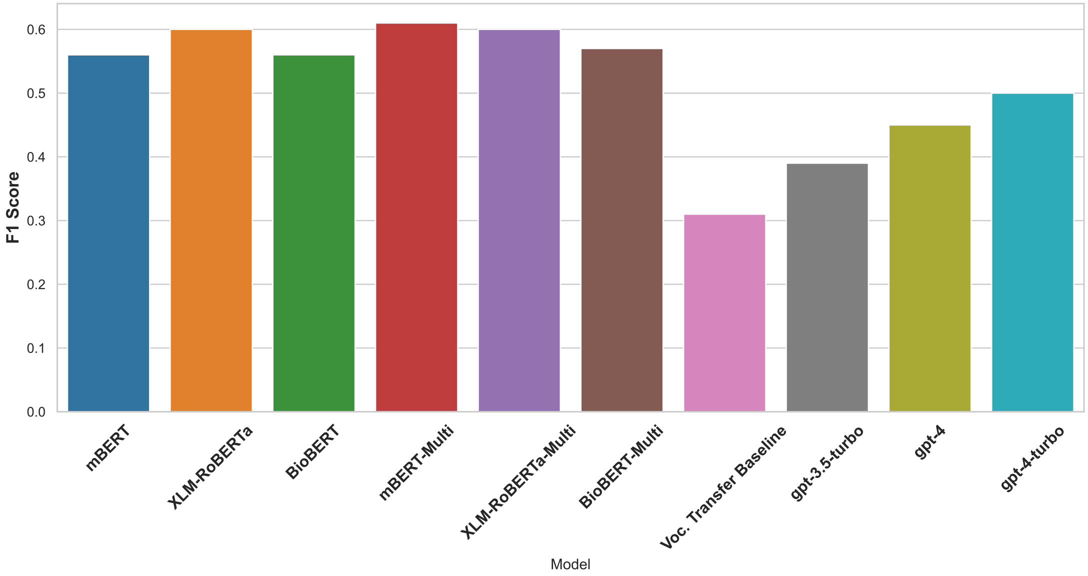

Figure 5.1: RE F1 Scores of all Models - Italian

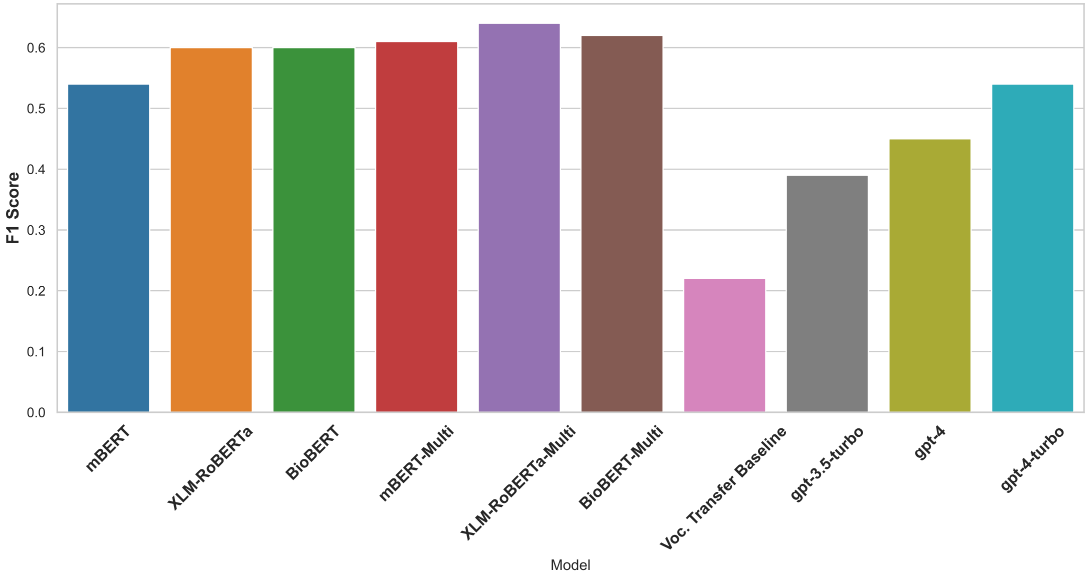

Figure 5.2: RE F1 Scores of all Models - Spanish

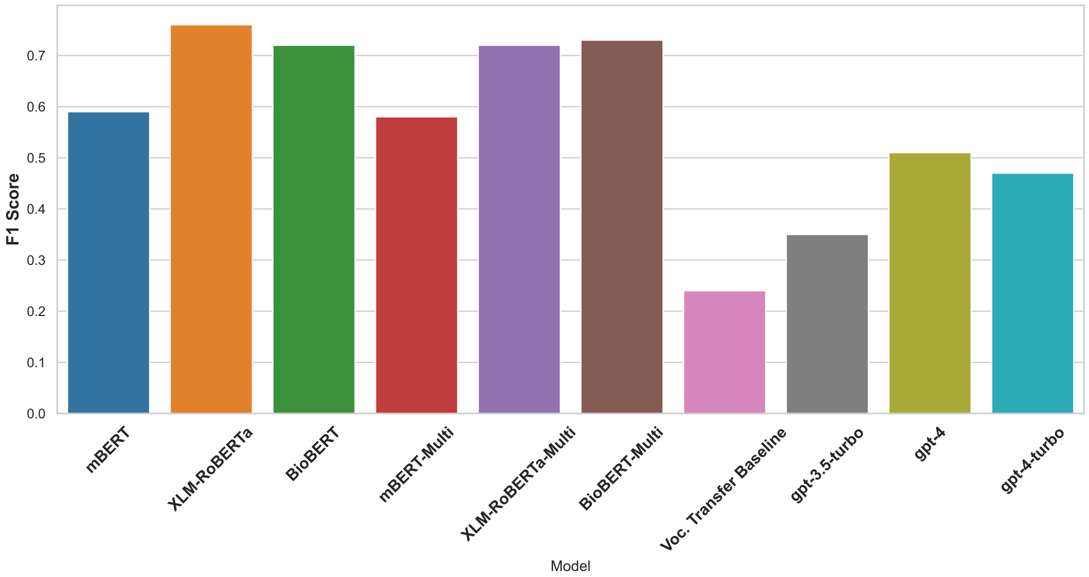

Figure 5.3: RE F1 Scores of all Models - Basque

5.1 Performance in Named Entity Recognition

The results of NER tasks reveal that multilingual models outperform their monolingual counterparts across all three languages. This suggests that training on multilingual data can enhance model performance by cross-lingual knowledge transfer. The XLM-RoBERTa-Multi model, in particular, achieved the highest F1 scores for all three languages, demonstrating the effectiveness of large multilingual models in medical NER tasks. The fine-grained analysis revealed differences in entity type performance. In Italian and Basque, RML (result) entities generally outperformed TST (test) entities, while Spanish showed mixed results. RML entities are usually measurements or often repetitive and contextually straightforward (eg. terms like “positive” or “negative”). Conversely, TST entities could be more varied and context dependent. In general the most commonly missed TST entities were non-medical terms such as “test”, “weighing” and “examination”. However these entities were common in the statements, majority of them were detected. Models often struggled with complex or lengthy RML entities, those involving qualitative descriptions or complex measurements.

5.2 Final Relation Extraction Performance

The performance on the NER task results directly influenced the outcomes of the Relation Extraction task. If an entity is incorrectly tagged or missed in the NER phase, it invariably impacts the subsequent classification of relations involving that entity.

As with NER, multilingual models generally provided a robust performance, with the exception of Basque. With Basque, XLM-RoBERTa performed better as a monolingual model than as a multilingual one. This highlights the unique characteristics of the language. In Italian, the mBERT-Multi obtained the highest F1 score but only 0.01 points ahead of XLM-RoBERTa-Multi. Error analysis revealed a significant number of spurious relations from mBERT in both Italian and Spanish, which improved with multilingual training. Basque revealed a different error profile, with mBERT models exhibiting high missed relations while XLM-RoBERTa-Multi had a high count of spurious relations. As with the NER task, commonly missed relations included general terms which were frequent in the dataset. A limitation observed was that some errors may have stemmed from annotation inconsistencies.

5.2.1 Relation Extraction with LLMs and Few-shot learning

The models tested included a legacy version, **gpt4** and newer versions: **gpt-4-turbo** and **gpt-3.5-turbo**. Overall, the best performing model is the newest and largest model, gpt-4-turbo. Basque is the only exception again, with the legacy gpt-4 achieving the highest F1. However, the best performing LLM is 0.11, 0.10 and 0.25 points less in Italian, Spanish and Basque respectively compared to the best fine-tuned Transformer model.

So, one-shot learning using LLMs still lags behind fine-tuning Transformer models. A single example is not enough for the LLM to learn the quirks in the annotation. An approach to imrove could be be to feed it well-written guidelines in the prompt. The clinical statement fed to the LLM is also a long piece of text and the model might not be able to retain all the key information in it to generate the required output. If the task was done at sentence level (as done for the Transformer models) it is plausible that the LLMs would fare much better. This would also be an expensive experiment.

The gpt-4 models give a good performance jump over gpt-3.5 across languages. Gpt-4 models perform better with non-English prompts, although not by a big margin. Gpt-3.5-turbo, however, performed much better with English prompts. Also, non-English prompts performed better with the Single-message prompting approach, whereas English prompts worked better with the Chat Prompting approach. A possible explanation for this occurrence could be that the "system message" might be better understood by the model in English.

It is reasonable to expect that adding more examples to the prompt would help with the performance, albeit at a higher usage cost. Another technique which can possibly improve scores: for each clinical statement, select best example(s) to use in the prompt from a group of n examples by measuring semantic similarity with the input statement.

A notable observation with the newer model version is the significant increase in Precision coinciding with a drop in Recall. The newer gpt-4 appears to have a conservative prediction strategy, possibly a change introduced by OpenAI to be maximally accurate with generations.

5.3 Challenges and Limitations

Data Limitations

- **Size of the Dataset** A larger dataset often paves the way for models, especially deep learning ones, to better generalize and learn the intricacies of the task. In this case, the number of clinical statements available for

training and validation purposes is not vast. This limitation influences the models' performance, given that deep learning architectures such as transformer models significantly benefit from vast amounts of data.
A reason for this is that annotation of these statements is quite an expensive and time consuming process, requiring specific expertise.

- **Representativeness** With the dataset not being expansive comes the concern about its representativeness. The dataset might not capture complete variability and range of clinical scenarios, terminologies or linguistic constructs that exist in the domain of clinical documentation.

LLM - Cost Implications

A drawback of working with API based implementations like GPT, is the associated cost. With clinical statements, the prompts required for the task often tend to be lengthy. The usage is billed based on the number of tokens in the prompt and given the verbose nature of the statements, the expenses can quickly escalate. This factor becomes prohibitive for extensive datasets and iterative testing of prompts.

5.4 Implications for Clinical Practice

Advancements in relation extraction from clinical documents can help in enhancing clinical decision making. Having access to structured and contextualized patient information, healthcare professionals can make more informed judgments, improving patient outcomes. The transformation of unstructured clinical narratives into structured data can considerably simplify the storage, retrieval, and sharing of patient information. Automating the process of relation extraction reduces the need for manual chart reviews and data entry, which can be labor-intensive and prone to errors.

With the growing interest in monitoring the health status of populations, structured clinical data can empower public health agencies to detect patterns, track disease outbreaks, and respond swiftly to health emergencies.

5.5 Future Research Avenues

- **Fine-tuning LLMs** LLMs require a lot of computational resources and memory to fine-tune. As explained in Section 1.5.1 **QLORA** technique can make fine-tuning faster and cheaper. Fine-tuning is typically performed using open-source LLMs, as they provide the flexibility and transparency needed for customization. However, some providers, such as OpenAI, offer fine-tuning APIs without disclosing the underlying details of the fine-tuning process. While this can be convenient for users, it may limit the understanding and control over the fine-tuning procedure. Nevertheless, these APIs can still be useful for those who prioritize ease of use over customization.

  One significant limitation of many well-performing open-source LLMs is their limited support for non-English languages. These models are often pre-trained on large English text corpora, resulting in suboptimal performance when applied to tasks in other languages. Fine-tuning LLMs on non-English tasks typically requires a substantial amount of data to compensate for the lack of exposure to the target language during pre-training. In such cases, fine-tuning data can be sourced from resources such as Wikipedia or other large text repositories in the desired language. However, the quality and quantity of available data in non-English languages may vary, potentially impacting the effectiveness of fine-tuning.

- **Medical LLMs** LLMs like GPT-3.5 and GPT-4 have shown impressive capabilities in general domain, but there is an evident gap when it comes to specialized multilingual medical contexts. Large Language Models aligned to the medical domain can solve this problem. Google's Med-Palm 2 (Singhal et al. [2023]) was trained on a variety of biomedical datasets and different tasks.

  However, it is important to note that the development of medical LLMs also comes with challenges and considerations. One major challenge is the availability and accessibility of large-scale, high-quality medical datasets for training these models. Medical data often contains sensitive and confidential information, and strict regulations and privacy concerns may limit the sharing and use of such data. Efforts to create de-identified and anonymized medical datasets, as well as collaborative initiatives among healthcare institutions and researchers, can help address this challenge.

  Hallucination in LLMs which refers to generating plausible but incorrect or unverifiable information is a serious problem to be tackled especially with sensitive applications like healthcare to be a with LLMs. In Pal et al. [2023] the authors address the issue of hallucinations in LLMs in the medical domain. It introduces a new benchmark and dataset, Med-HALT, to evaluate and mitigate hallucinations in LLMs.

- **Larger and more representative datasets** A broader dataset would encapsulate the extensive variability in clinical terminologies, linguistic structures, and clinical scenarios that these models are expected to handle in real-world settings. As the field evolves, it will be crucial to investigate how models trained on extensive clinical datasets perform in related domains, like biomedical research papers. This cross-domain adaptation can further test the models' robustness and adaptability. **Privacy-enabled AI** can play a pivotal role in accessing and utilizing more useful data, especially in domains such as healthcare where data sensitivity is a paramount concern. It can help with compliance regulation and unlocking siloed data.

Conclusion

In the area of clinical document analysis, this work was an effort to understand the efficacy of contemporary models with a focus on relation extraction from multilingual clinical records.

The work underscores the effectiveness of fine-tuning with multilingual data and large encoder models for the tasks of Named Entity Recognition and Relation Extraction. Basque stood out as an exception, with the monolingual XLM-RoBERTa outperforming other models in the Relation Extraction task. Italian and Spanish generally exhibited similar error profiles with the Transformer models. The models struggled with complex medical terms and test measurements. Large representative datasets with medical entities present in diverse contexts are important to enhance performance on tasks with medical texts.

Our experimentation with Large Language Models, specifically the GPT-3.5 and GPT-4, offered intriguing takeaways. While one-shot learning demonstrated potential, it falls short of the performance of fine-tuned Transformers. It underlines the challenges with domain specific tasks especially one such as Relation Extraction where there might exist specific guidelines for the entities and relations. With limited examples and basic instructions the current LLMs fall short of traditional fine-tuning approaches. Smaller LLMs also suffer from limited capability of understanding the text. The tangible improvement from GPT-3.5 to GPT-4 illuminates the rapid strides being made in this domain. The newer version of the GPT-4 model tends to be more conservative and precise in its generation, resulting in poorer recall. The enhancements have not translated into substantial improvements in clinical relation extraction, highlighting the challenges in balancing model sophistication with practical performance gains.

Looking ahead, the frontier of clinical document analysis is ripe for advancements. Fine-tuning LLMs, especially with tools like QLORA is quite promising. Medical LLMs, fine-tuned for multilingual contexts, seems not only feasible but also imperative. Such specialized models, like the aforementioned Google's Med-Palm 2, showcase the exciting direction in which the broader field of Biomedical text mining is headed.

In summation, this thesis demonstrates the evolving capabilities of Transformer models and LLMs in the context of clinical document interpretation. Future endeavors in this domain promise tangible enhancements in medical data interpretation, which can possibly improve healthcare outcomes globally.

Bibliography

Elisa Bassignana and Barbara Plank. What do you mean by relation extraction? a survey on datasets and study on scientific relation classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 67–83, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-srw.7. URL `https://aclanthology.org/2022.acl-srw.7`.

Mary Elaine Califf and Raymond J. Mooney. Relational learning of pattern-match rules for information extraction. In *CoNLL97: Computational Natural Language Learning*, 1997. URL `https://aclanthology.org/W97-1002`.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL `https://aclanthology.org/2020.acl-main.747`.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

David Fraile Navarro, Kiran Ijaz, Dana Rezazadegan, Hania Rahimi-Ardabili, Mark Dras, Enrico Coiera, and Shlomo Berkovsky. 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, 2023. ISSN 1386-5056. doi: https://doi.org/10.1016/j.ijmedinf.2023.105122. URL `https://www.sciencedirect.com/science/article/pii/S1386505623001405`.

Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning.* MIT Press, Cambridge, MA, USA, 2016. `http://www.deeplearningbook.org`.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997. doi: 10.1162/neco.1997.9.8.1735.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

Peter Izsak, Moshe Berchansky, and Omer Levy. How to train BERT with an academic budget. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10644–10652, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.831. URL `https://aclanthology.org/2021.emnlp-main.831`.

Peter B Jensen, Lars J Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, June 2012.

Qiao Jin, Bhuwan Dhingra, William Cohen, and Xinghua Lu. Probing biomedical embeddings from language models. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 82–89, Minneapolis, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-2011. URL `https://aclanthology.org/W19-2011`.

Nanda Kambhatla. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*, ACLdemo '04, page 22–es, USA, 2004. Association for Computational Linguistics. doi: 10.3115/1219044.1219066. URL `https://doi.org/10.3115/1219044.1219066`.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL `https://api.semanticscholar.org/CorpusID:6628106`.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

J. Li, Q. Wei, O. Ghiasvand, M. Chen, V. Lobanov, C. Weng, and H. Xu. A comparative study of pre-trained language models for named entity recognition in clinical trial eligibility criteria from multiple corpora. *BMC Med Inform Decis Mak*, 22(Suppl 3):235, Sep 2022.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068, 05 2016. ISSN 1758-0463. doi: 10.1093/database/baw068. URL `https://doi.org/10.1093/database/baw068`.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

Bernardo Magnini, Begoña Altuna, Alberto Lavelli, Manuela Speranza, and Roberto Zanoli. The E3C project: Collection and annotation of a multilingual corpus of clinical cases. In *Proceedings of the Seventh Italian Conference* *on Computational Linguistics CLiC-it 2020*, pages 258–264. Accademia University Press, 2020.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013. URL `https://arxiv.org/abs/1301.3781`.

Hiroki Nakayama. seqeval: A python framework for sequence labeling evaluation, 2018. URL `https://github.com/chakki-works/seqeval`. Software available from https://github.com/chakki-works/seqeval.

Thien Huu Nguyen and Ralph Grishman. Relation extraction: Perspective from convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 39–48, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.3115/v1/W15-1506. URL `https://aclanthology.org/W15-1506`.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

Ankit Pal, Logesh Umapathi, and Malaikannan Sankarasubbu. Med-halt: Medical domain hallucination test for large language models. pages 314–334, 01 2023. doi: 10.18653/v1/2023.conll-1.21.

Bethany Percha and Russ B Altman. Informatics confronts drug-drug interactions. *Trends Pharmacol Sci*, 34(3):178–184, February 2013.

Simone Scaboro, Beatrice Portelli, Emmanuele Chersoni, Enrico Santus, and Giuseppe Serra. Extensive evaluation of transformer-based architectures for adverse drug events extraction. *Knowledge-Based Systems*, 275: 110675, 2023. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2023.110675. URL `https://www.sciencedirect.com/science/article/pii/S0950705123004252`.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera Y Arcas, Dale Webster, Greg S Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620 (7972):172–180, July 2023.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, jan 2014. ISSN 1532-4435.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Somin Wadhwa, Silvio Amir, and Byron C. Wallace. Revisiting relation extraction in the era of large language models. *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2023:15566–15589, 2023. URL `https://api.semanticscholar.org/CorpusID:258564662`.

Leon Weber, Mario Sänger, Samuele Garda, Fabio Barth, Christoph Alt, and Ulf Leser. Chemical–protein relation extraction with ensembles of carefully tuned pretrained language models. *Database*, 2022:baac098, 11 2022. ISSN 1758-0463. doi: 10.1093/database/baac098. URL `https://doi.org/10.1093/database/baac098`.

D. Wilson and Tony Martinez. The general inefficiency of batch training for gradient descent learning. *Neural networks : the official journal of the International Neural Network Society*, 16:1429–51, 01 2004. doi: 10.1016/S0893-6080(03)00138-2.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Perric Cistac, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-Art Natural Language Processing. pages 38–45. Association for Computational Linguistics, October 2020. URL `https://www.aclweb.org/anthology/2020.emnlp-demos.6`.

Wei Ye, Bo Li, Rui Xie, Zhonghao Sheng, Long Chen, and Shikun Zhang. Exploiting entity BIO tag embeddings and multi-task learning for relation extraction with imbalanced data. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1351–1360, Florence, Italy, July

2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1130. URL https://aclanthology.org/P19-1130.

Tong Yu and Hong Zhu. Hyper-parameter optimization: A review of algorithms and applications. *ArXiv*, abs/2003.05689, 2020. URL https://api.semanticscholar.org/CorpusID:212675087.

Shubin Zhao and Ralph Grishman. Extracting relations with integrated information using kernel methods. In Kevin Knight, Hwee Tou Ng, and Kemal Oflazer, editors, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 419–426, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219892. URL https://aclanthology.org/P05-1052.

List of Figures

List of Tables

List of Abbreviations

BERT - Bidirectional Encoder Representations from Transformers

GPT - Generative Pre-trained Transformer

LLM - Large Language Model

NER - Named Entity Recognition

RC - Relation Classification

RE - Relation Extraction

RLHF - Reinforcement Learning from Human Feedback

Appendices

5.6 Optimal Hyperparameters

NER Task

mBERT

Hyperparameter	Optimal value		
	it	**es**	**eu**
Learning Rate	8e-5	4e-5	8e-5
Batch Size	16	16	32
Epochs	6	6	6
Optimizer	AdamW	AdamW	AdamW
Weight Decay	0.06	0.05	0.05
Warmup Ratio	0.1	0.05	0.1
Max Gradient Norm	2.0	1.0	2.0
Scheduler	polyn.	polyn.	linear sched.

Table 5.1: Optimal hyperparameter values for mBERT and NER Task across languages

XLM-RoBERTa

Hyperparameter	Optimal value		
	it	**es**	**eu**
Learning Rate	4e-5	1e-4	8e-5
Batch Size	32	32	32
Epochs	5	6	6
Optimizer	AdamW	AdamW	AdamW
Weight Decay	0.1	0.07	0.06
Warmup Ratio	0.1	0.1	0.1
Max Gradient Norm	2.0	2.0	2.0
Scheduler	cosine sched.	polyn.	polyn.

Table 5.2: Optimal hyperparameter values for XLM-RoBERTa and NER Task across languages

BioBERT

Hyperparameter	**Optimal value**		
	it	**es**	**eu**
Learning Rate	8e-5	8e-5	9e-5
Batch Size	16	16	32
Epochs	6	5	6
Optimizer	AdamW	AdamW	AdamW
Weight Decay	0.08	0.08	0.07
Warmup Ratio	0.1	0.1	0.05
Max Gradient Norm	2.0	1.0	2.0
Scheduler	linear sched.	constant sched.	cosine sched.

Table 5.3: Optimal hyperparameter values for BioBERT and NER Task across languages

Joint Multilingual model

Hyperparameter	**Optimal value**		
	mBERT	**XLM-R**	**BioBERT**
Learning Rate	4e-5	9e-5	5e-5
Batch Size	16	32	32
Epochs	5	6	6
Optimizer	AdamW	AdamW	AdamW
Weight Decay	0.08	0.06	0.05
Warmup Ratio	0.1	0.05	0.1
Max Gradient Norm	2.0	2.0	2.0
Scheduler	polyn.	polyn.	polyn.

Table 5.4: Optimal hyperparameter values for Multilingual models and NER Task

Relation Classification Task

mBERT

Hyperparameter	**Optimal value**		
	it	**es**	**eu**
Learning Rate	6e-5	4e-5	2e-5
Batch Size	16	16	8
Epochs	6	6	6
Optimizer	AdamW	AdamW	AdamW
Weight Decay	0.05	0.06	0.08
Warmup Ratio	0.05	0.1	0.05
Max Gradient Norm	2.0	2.0	2.0
Scheduler	cosine sched.	polyn.	const sched.

Table 5.5: Optimal hyperparameter values for mBERT and RC Task across languages

XLM-RoBERTa

Hyperparameter	**Optimal value**		
	it	**es**	**eu**
Learning Rate	8e-5	4e-5	3e-5
Batch Size	32	32	32
Epochs	6	6	6
Optimizer	AdamW	AdamW	AdamW
Weight Decay	0.09	0.1	0.05
Warmup Ratio	0.1	0.1	0.05
Max Gradient Norm	2.0	2.0	1.0
Scheduler	polyn.	constant sched.	constant sched.

Table 5.6: Optimal hyperparameter values for XLM-RoBERTa and NER Task across languages

BioBERT

Hyperparameter	**Optimal value**		
	it	**es**	**eu**
Learning Rate	2e-5	3e-5	9e-5
Batch Size	32	32	32
Epochs	6	6	6
Optimizer	AdamW	AdamW	AdamW
Weight Decay	0.06	0.05	0.1
Warmup Ratio	0.1	0.1	0.1
Max Gradient Norm	2.0	2.0	1.0
Scheduler	linear sched.	polyn.	cosine sched.

Table 5.7: Optimal hyperparameter values for BioBERT and RC Task across languages

Joint Multilingual model

Hyperparameter	**Optimal value**		
	mBERT	**XLM-R**	**BioBERT**
Learning Rate	9e-5	4e-5	1e-4
Batch Size	32	32	32
Epochs	4	5	6
Optimizer	AdamW	AdamW	AdamW
Weight Decay	0.05	0.09	0.06
Warmup Ratio	0.1	0.05	0.1
Max Gradient Norm	2.0	2.0	2.0
Scheduler	linear sched.	polyn.	polyn.

Table 5.8: Optimal hyperparameter values for Multilingual models and RC Task

5.7 NER Validation set metrics

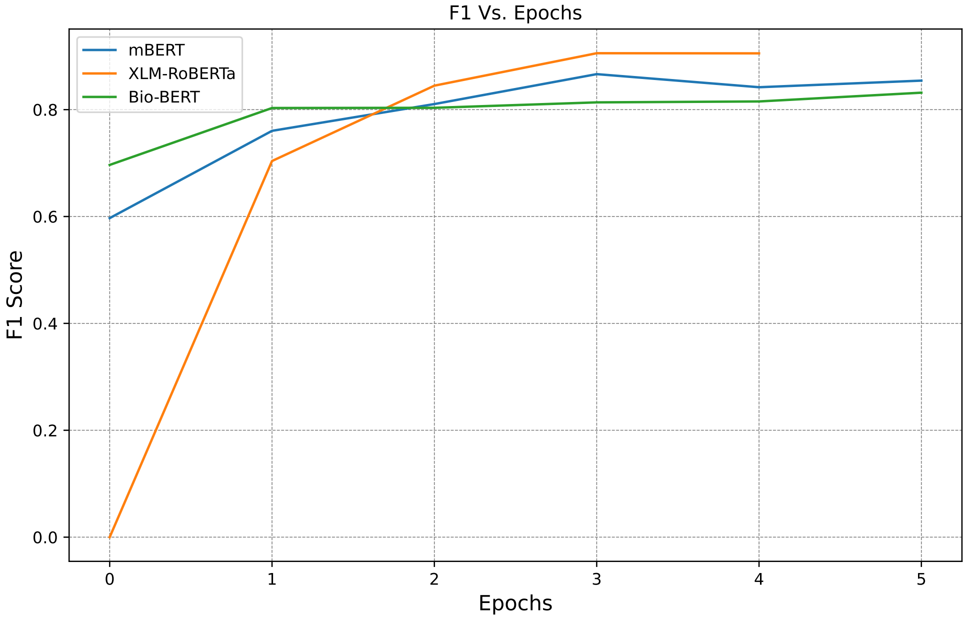

Figure 5.4: Validation set F1 Scores for Monolingual models - Italian

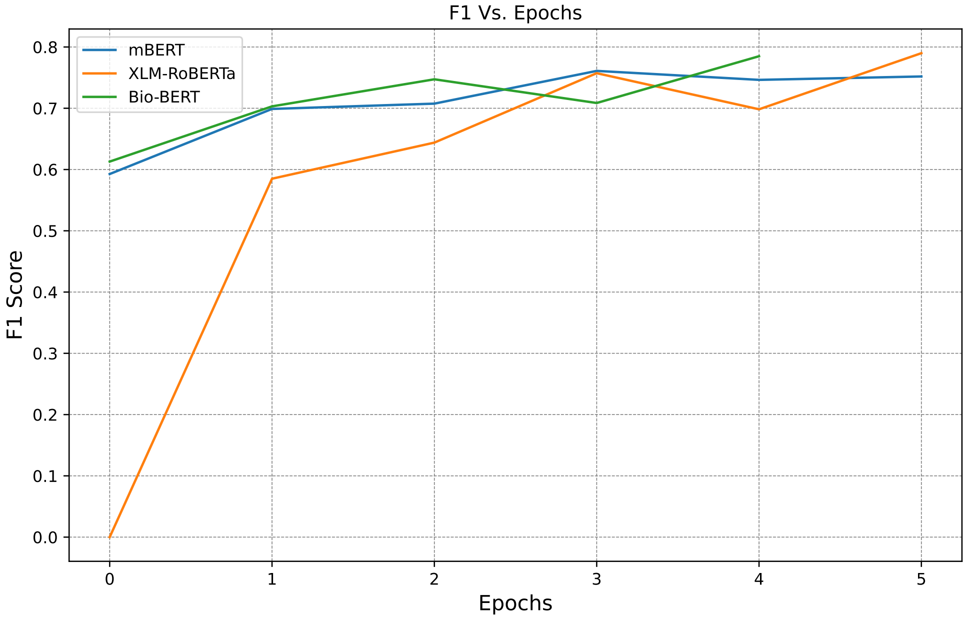

Figure 5.5: Validation set F1 Scores for Monolingual models - Spanish

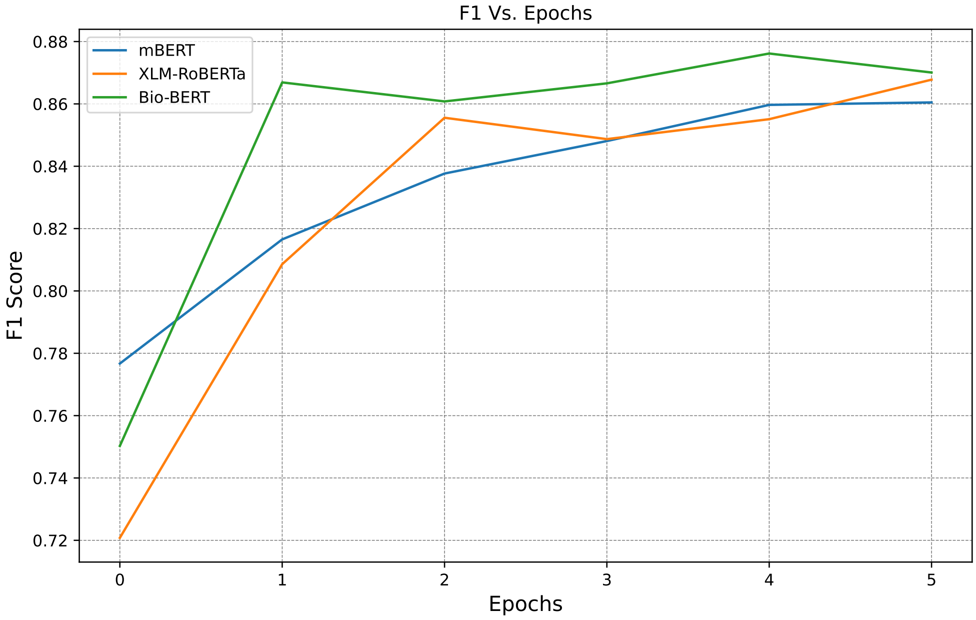

Figure 5.6: Validation set F1 Scores for Monolingual models - Basque

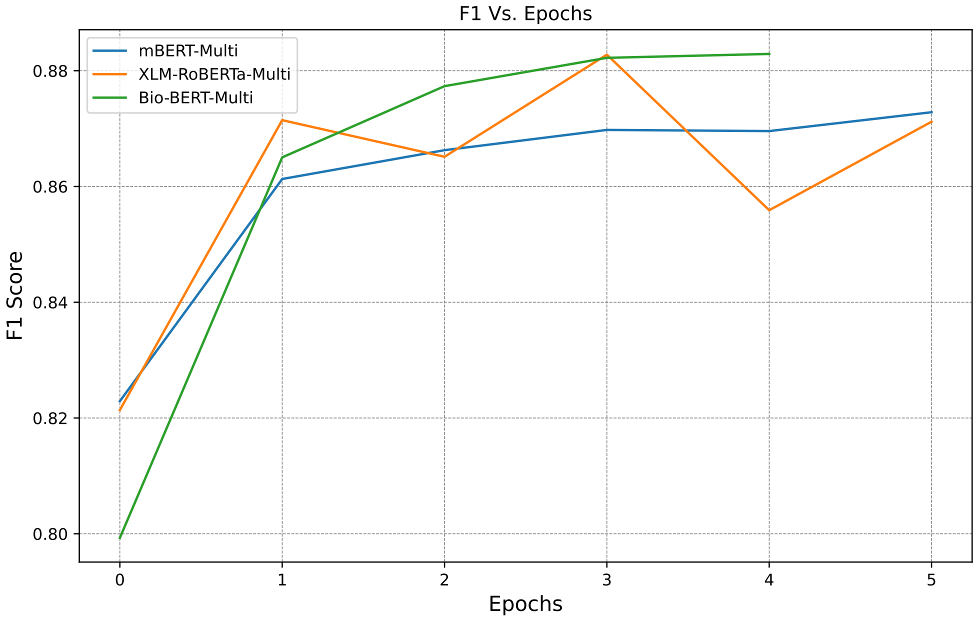

Figure 5.7: Validation set F1 Scores for Multilingual models

5.8 RE Validation set metrics

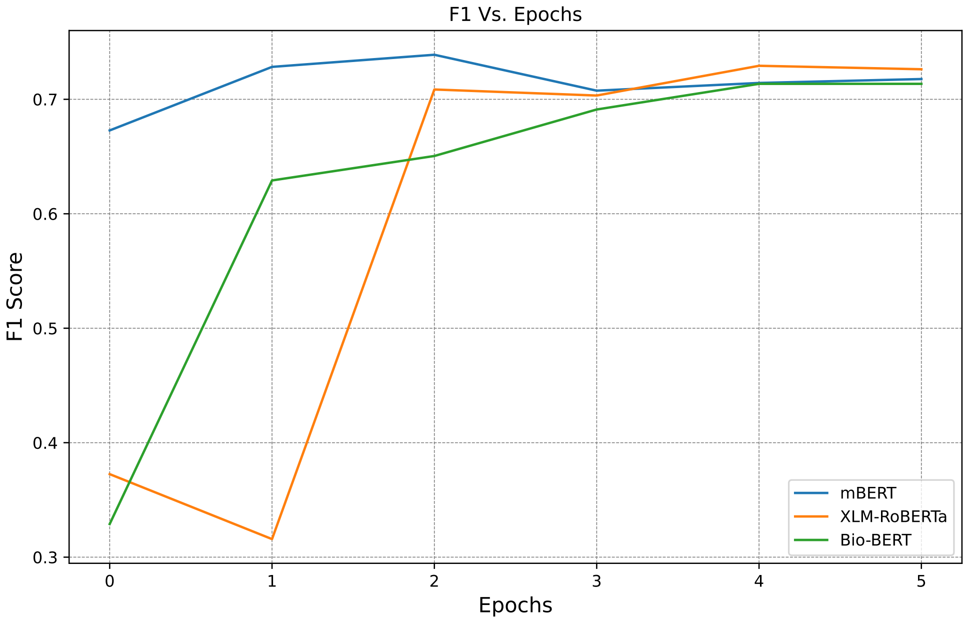

Figure 5.8: Validation set F1 Scores for Monolingual models - Italian

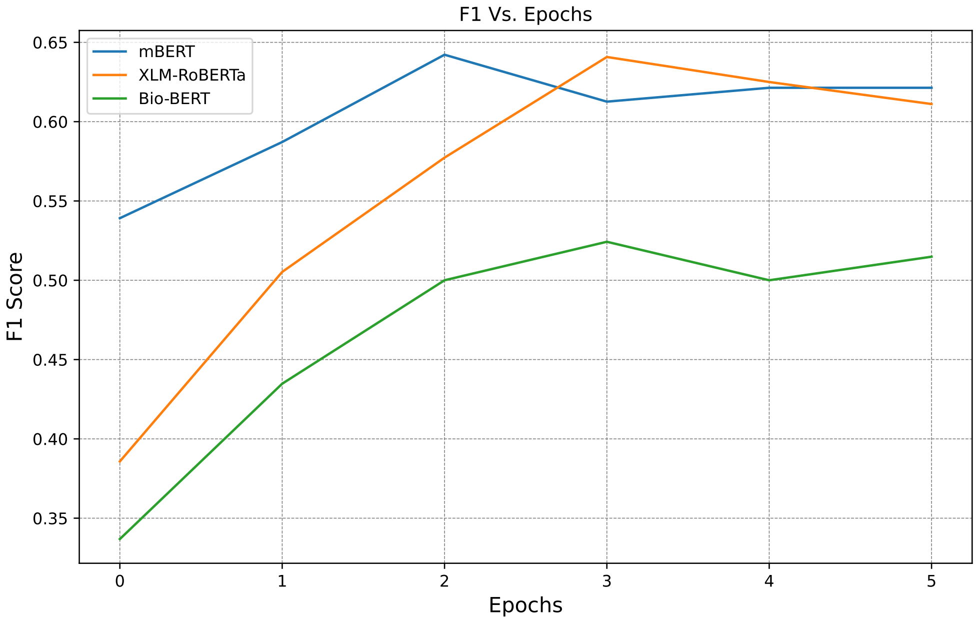

Figure 5.9: Validation set F1 Scores for Monolingual models - Spanish

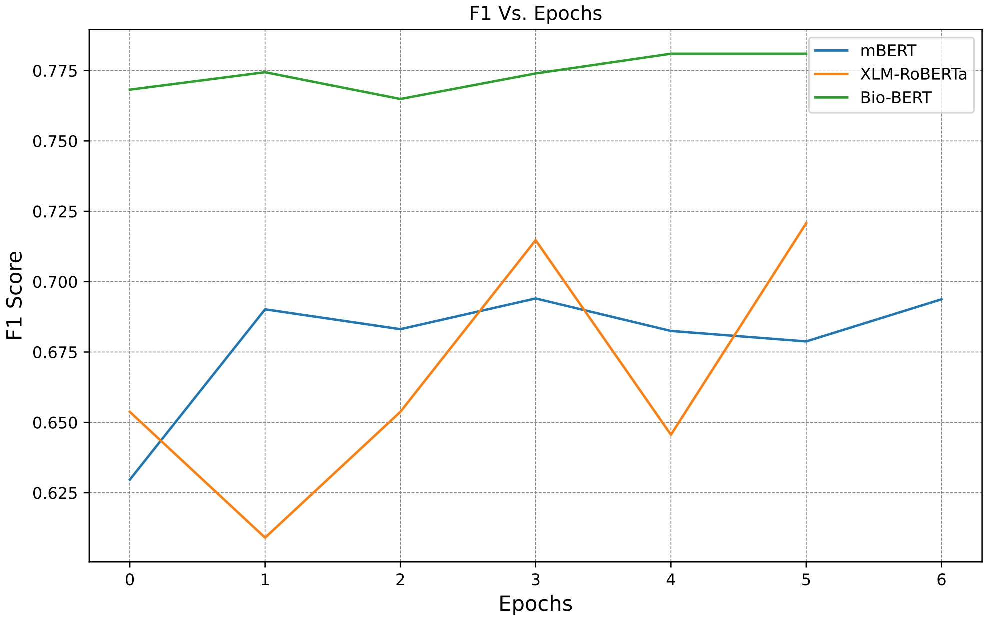

Figure 5.10: Validation set F1 Scores for Monolingual models - Basque

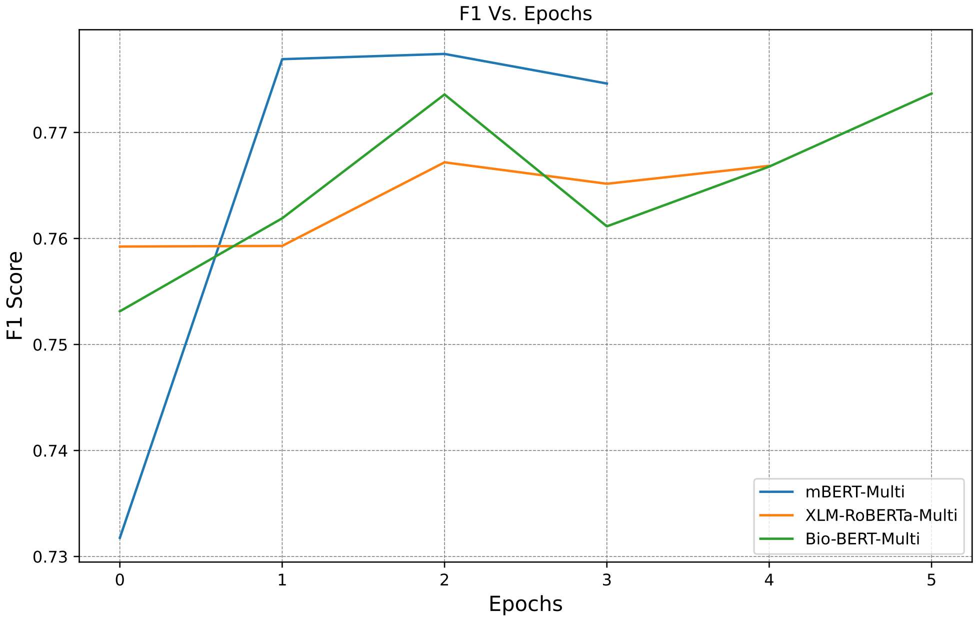

Figure 5.11: Validation set F1 Scores for Multilingual models

5.9 Fine-grained NER results

Lang	Model	RML	TST
Italian	mBERT	0.76	0.72
	XLM-RoBERTa	0.74	0.73
	BioBERT	0.78	0.68
	mBERT-Multi	0.79	0.79
	XLM-R-Multi	0.81	0.79
	BioBERT-Multi	0.79	0.75
Spanish	mBERT	0.74	0.76
	XLM-RoBERTa	0.72	0.77
	BioBERT	0.78	0.74
	mBERT-Multi	0.77	0.78
	XLM-R-Multi	0.80	0.80
	BioBERT-Multi	0.77	0.76
Basque	mBERT	0.75	0.67
	XLM-RoBERTa	0.81	0.78
	BioBERT	0.77	0.77
	mBERT-Multi	0.82	0.82
	XLM-R-Multi	0.85	0.84
	BioBERT-Multi	0.82	0.79

Table 5.9: Entity specific NER F-1 Scores

5.10 Other prompt results

Lang	Precision	Recall	F1
Italian	0.73	0.35	0.47
Spanish	0.79	0.37	0.50
Basque	0.82	0.39	0.52

Table 5.10: Metrics for English prompt (chat-prompting) with gpt-4-turbo

Lang	Precision	Recall	F1
Italian	0.57	0.31	0.40
Spanish	0.44	0.19	0.27
Basque	-	-	-

Table 5.11: Metrics for non-English prompt (single-message) with gpt-3.5-turbo

5.11 LLM Prompts

5.11.1 Prompt for Italian

Ho un compito che è quello di estrarre menzioni di test di laboratorio e dei loro risultati da dichiarazioni cliniche. Ecco un esempio di testo e output:

Testo :
Giunge alla nostra osservazione in Pronto Soccorso con mezzo proprio una donna di 58 anni, riferendo di aver accidentalmente ingerito una piccola quantità di un bicchiere di acido borico diluito, usato per la disinfezione di lesioni psoriasiche. Al triage vengono rilevati i seguenti parametri vitali: PA 140/70 mmHg, F.C. 100/min, ritmica, SpO2 98% in a.a., GCS 15/15. Viene assegnato un codice giallo e inviata la paziente in sala visita. Dall'anamnesi emerge storia di psoriasi e pregressa nefrectomia con calcolosi a stampo. La paziente è in terapia con antisecretivi gastrici. Nega allergie. All'esame obiettivo non si evidenziano lesioni del cavo orale e faringe, sebbene la paziente riferisca faringodinia. L'obiettività cardio-respiratoria è nella norma. Le lesioni cutanee psoriasiche non appaiono in fase florida. La paziente risulta pertanto stabile e si prosegue con la stima del tossico ingerito: viene riferita assunzione di meno di un bicchiere di una soluzione al 3% ottenuta diluendo una bustina da 30g di acido borico in un litro d'acqua. Pertanto, considerando la capacità di un bicchiere di circa 200 ml, è stato assunto 1/5 dei 30g diluiti, pari a 6 g. Si contatta il Centro Antiveleni di Pavia che consiglia di posizionare SNG, aspirare e somministrare carbone attivo. Poiché la sostanza è gravata da nefrotossicità, riferisce inoltre di idratare la paziente con un target pari a 4000 ml nelle 24 ore, con monitoraggio ogni 6 ore della funzionalità renale. Considerata la necessità di un attento monitoraggio, si decide di ricoverare la paziente presso l'Unità di Osservazione Breve Intensiva (OBI) del nostro Pronto Soccorso. Agli esami di laboratorio effettuati all'ingresso, la funzionalità renale è nella norma, ma si evidenzia un lieve rialzo della bilirubina diretta (1,25 mg/dl). Al secondo controllo gli esami di laboratorio risultano sovrapponibili ai precedenti. Sono trascorse 12 ore dall'ingresso della paziente e il CAV di Pavia, nonostante la normalità degli esami di laboratorio, indica l'opportunità di effettuare comunque una seduta dialitica. Dopo la seduta la paziente esegue esami ematochimici di controllo in cui compare brusco incremento delle transaminasi e della bilirubina; in particolare: bil. tot. da 1,71 mg/dl a 2 (diretta da 1,28 a 1,2), AST da 49 U/L a 207, ALT da 49 U/L a 237, GGT da 36 U/L a 186. Tali valori tendono poi a calare già a partire dal controllo effettuato 6 ore dopo. Si esegue in ogni caso nuova seduta emodialitica il giorno successivo, ed il CAV di Pavia, informato dell'evento, ci informa (ribadendolo poi ancora successivamente) che non esiste alcuna correlazione tra l'intossicazione e l'incremento degli indici di citonecrosi epatica. A questo punto la paziente introduce un elemento anamnestico fino a quel momento taciuto: riferisce di essere affetta da sindrome di Dubin-Johnson, con eventi di rialzo delle transaminasi in occasione di gastroenteriti. La paziente sospende le sedute dialitiche e continua con idratazione, senza ulteriori complicazioni fino al 6° giorno di degenza in OBI, quando compaiono astenia, malessere generale e febbre, che nel pomeriggio raggiunge il picco di 38.5 °C. Agli esami di laboratorio si evidenzia nuovo incremento delle transaminasi e bilirubina, che restano però su valori lievemente più bassi rispetto al primo picco (bil. tot. da 1,9 mg/dl a 2,3 (diretta da 1,17 a 1,38), AST da 32 U/L a 141, ALT da 93 U/L a 191, GGT da 74 U/L a 121), e successivo calo dopo 6 ore.

Output :
140/70 mmHg |PA
100/min |F.C
9815/15 |GCS
nella norma |L'obiettività
nella norma |funzionalità
1,25 mg/dl |bilirubina
la normalità |esami

```
1,71 mg/dl |bil
2 |bil
1,28 |diretta
1,2 |diretta
49 U/L |AST
207 |AST
49 U/L |ALT
237 |ALT
36 U/L |GGT
186 |GGT
38.5 °C |febbre
38.5 °C |picco
1,9 mg/dl |bil
2,3 |bil
1,17 |diretta
1,38 |diretta
32 U/L |AST
141 |AST
93 U/L |ALT
191 |ALT
74 U/L |GGT
121 |GGT
```

Nota: nell'output viene scritto prima il risultato e poi il nome del test. Sono separati da '|'.

Ora dammi l'output per il seguente testo:

{New Statement}

Stampa solo l'output se presente e nient'altro.

5.11.2 Prompt for Spanish

Tengo una tarea que consiste en extraer menciones de pruebas de laboratorio y sus resultados de declaraciones clínicas. Aquí hay un ejemplo de texto y salida:
Texto :
Mujer de 27 años, politoxicómana activa, que 15 días antes de ingresar comienza con tos seca, disnea de intensidad progresiva, fiebre, astenia, anorexia, náuseas y vómitos. Tenía antecedentes de infección por VIH de reciente diagnóstico, desconociendo sus parámetros inmunológicos y virológicos. La exploración al ingreso mostraba caquexia, 40 respiraciones/ min., 120 latidos/min. TA:110/60 mm Hg, muguet, acropaquias, adenopatías cervicales bilaterales menores de 1 cm y supraclavicular derecha de 2 cm. Disminución generalizada del murmullo vesicular y crepitantes bibasales, soplo sistólico II/VI polifocal, hepatomegalia a 2 cm del borde costal y condilomas vulvares. El resto de la exploración somática y neurológica fueron normales. Se realizaron al ingreso las siguientes pruebas complementarias: Hb 11,3 g/dL, 6.900 leucocitos/µL (79% seg-

mentados, 14% linfocitos), 682.000 plaquetas/µL; sodio 119 mmol/L, LDH 603 UI/L, siendo normales el resto de los valores del autoanalizador. Gasometría arterial basal al ingreso: pH 7,54, pCO2 34,8 mm Hg, pO2 73,4 mmHg. En la placa de tórax se observa una masa mediastínica anterior derecha que comprime y desplaza la tráquea y un infiltrado alveolo-intersticial bilateral. Posteriormente se realiza ecografía abdominal y TAC torácico, abdominal y pélvico en los que se aprecian adenopatías axilares bilaterales menores de un cm, una gran masa mediastínica que comprime tráquea y vena cava superior, patrón intersticial pulmonar en resolución y engrosamiento de pared de asas intestinales con pequeña cantidad de líquido libre entre ellas. Se realizó ecocardiograma transtorácico en el que no se encontraron alteraciones. Se inicia tratamiento con cotrimoxazol, levofloxacino y amfotericina-B, a pesar de lo cual desarrolla insuficiencia respiratoria progresiva, precisando ventilación mecánica. Durante los días siguientes se observa mejoría de la insuficiencia respiratoria, posibilitando la retirada de la ventilación mecánica. Posteriormente presenta crisis tónico-clónica generalizada seguida de hemiparesia izquierda; se realiza TAC y RMN craneal, observándose una masa de pared gruesa y centro hipodenso de 4 cm de diámetro en encrucijada témporo-parieto-occipital derecha. Se añade al tratamiento sulfadiacina, pirimetamina, ácido fólico y fenitoína, que se retiran tres semanas más tarde por ausencia de mejoría clínica y de disminución de la lesión intracraneal. Se realizó PAAF y biopsia de adenopatía supraclavicular que se informa como metástasis de tumor maligno indiferenciado de origen epitelial. Más tarde se reciben los resultados siguientes: linfocitos CD4+: 440/mL, carga viral VIH: 25.000 copias/ mL, alfafetoproteina normal y b-HCG: 650 mUI/mL (normal ¡ 5). Los estudios microbiológicos en sangre, orina, esputo, brocoaspirado y punción ganglionar fueron negativos para bacterias, micobacterias y hongos. Las tinciones de muestras respiratorias fueron negativas para P. carinii (P. jiroveci). Las serologías para Toxoplasma, virus hepatotropos y RPR fueron negativas; el antígeno criptocócico fue negativo. Diagnóstico y evolución: Inicialmente se realizó el diagnóstico de neumonía difusa presuntivamente por P. carinii (P. jiroveci) y probable linfoma diseminado con afectación mediastínica, cerebral e intestinal. Tras recibir los resultados de la biopsia ganglionar y de los marcadores tumorales (alfafetoproteina y ß-HCG), se realiza el diagnóstico definitivo: tumor de células germinales extragonadal con afectación ganglionar supraclavicular y mediastínica, y probablemente cerebral. Se propone a la paciente realizar estudio de extensión tumoral e iniciar tratamiento con poliquimioterapia, que rechaza, a pesar de explicársele que la elevada quimiosensibilidad del tumor lo hacía potencialmente curable. La paciente falleció en otro centro varias semanas después del alta.

Output :

```
120 latidos/min |exploración
110/60 mm Hg |TA
menores de 1 cm |adenopatías
2 cm |adenopatías
normales |resto
11,3 g/dL |Hb
6.900 leucocitos/µL |pruebas
79% |segmentados
14% |linfocitos
682.000 plaquetas/µL |pruebas
119 mmol/L |sodio
603 UI/L |LDH
```

```
normales |pruebas
7,54 |pH
34,8 mm Hg |pCO2
73,4 mmHg |pO2
menores de un cm |adenopatías
4 cm de diámetro |masa
440/mL |linfocitos
25.000 copias/ mL |carga
normal |alfafetoproteina
650 mUI/mL |b-HCG
negativos |micobacterias
negativos |estudios
negativos |punción
negativos |hongos
negativos |brocoaspirado
negativos |bacterias
negativas |tinciones
negativas |P
negativas |RPR
negativas |virus
negativas |Toxoplasma
negativo |antígeno
```

Nota: El resultado se escribe primero y luego el nombre de la prueba en la salida. Están separados por '|'

Ahora dame la salida para el siguiente texto:

{New Statement}

Imprime solo la salida y si no la hay, no imprimas nada.

5.11.3 Prompt for Basque

Laborategiko proben aipamenak eta haiei dagozkien emaitzak adierazpen klinikoetatik ateratzeko zeregina daukat. Hona hemen testuaren eta irteeraren adibide bat:
Testua :
ENDOKRINOLOGIA ETA NUTRIZIOA OSPITALERATZEKO/KONTSULTARAKO ARRAZOIA. Hipergluzemia duen 24 urteko emakumea. AURREKARIAK Ez du alergia ezagunik. Hipertentsio arteriala du eta erretzailea da. Ez du ohiko tratamendurik. EGUNGO HISTORIA Hilabetez geroztik egarria eta pixa askotan egiteko gogoa sentitu du. Aldi berean, argaldu egin da (5 kg gutxiago azken 3-4 asteetan) apetitua galdu gabe. Ez du sukarrik izan ez eta infekzio-zeinurik ere. Hasieran ez zen medikuarengana joan, baina gaur goizean goragaleak izan ditu, eta, arnasa hartzeko zailtasun txikia sentitu duenez, osasun-zentrora gerturatu da. Han gluzemia kapilarra neurtu eta oso altua zuela-eta, ospitaleko Larrialdietara bidali dute, diabetesaren susmoa egiaztatzeko eta, hala badagokio, kontrol metabolikoa ezarri eta hezkuntza diabetologikoa

hasteko. AZTERKETA FISIKOA. Tentsio arteriala: 120/50; bihotz-maiztasuna (BM) 98 tau/min; arnas maiztasuna 20 arnas/min; O2Sat,% 98. Kontziente, orientatua eta nahiko ondo hidratatua dago, baina takipnea pixka bat du. Fetor zetosikoa (+). Gluzemia kapilarra, 365 mg/dl. Zetonuria (+ + +) Burua eta lepoa. Ezer berezirik ez. Biriken auskultazioa. Biriketako murmurio normala. Bihotzaren auskultazioa. Erritmikoa. Sabelaldea. Ez dago aurkikuntza patologikorik. Gorputz-adarrak. Ez du edemarik; pultsu periferikoak normalak dira eta perfusioa egokia da. PROBA OSAGARRIAK ELEKTROKARDIOGRAMA. Erritmo sinusala; bihotz-maiztasuna (BM), 95 tau/min; ST segmentuaren aldaketa ez-espezifikoak. BULARREKO ERRADIOGRAFIA. Normala. ANALITIKA. Kreatinina 1,65 mg/dl; urea 62 mg/dl; Cl, 94 meq/l; Na, 134 meq/l; K, 5,2 meq/l; glukosa, 427 mg/dl, T troponina ultrasentsiblea, CPK eta CKMB normalak; GPT, 26; GOT, 21; GGT, 31; amilasa, 26; hemoglobina, 14,4 g/dl; hematokritoa,% 41,7; plaketak, 162.000/ml; leukozitoak, 16.580/ml; INR, 1,07; APTT, 31,1 seg. GASOMETRIA. pH, 7,15; HCO3, 10; BG (base gehiegi), -16,5; pO2, 112; pCO2, 16. GERNU-JALKINA. Normala. Glukosuria 1.000 mg/dl. Zetonuria 50 mg/dl. BULARREKO ERRADIOLOGIA. Normala. EBOLUZIOA Larrialdietako gelan 1 motako diabetesaren desoreka hipergluzemiko zetoazidosikoa diagnostikatu zioten. Zainketa Intentsiboetako Unitatean ospitaleratzea baztertu da, haren egoera klinikoa egonkorra delako. Egoera metabolikoa pixka bat hobetutakoan, Endokrinologiako Zerbitzuan ospitaleratzea erabaki da, intsulinaren bena barneko infusioa eta fluidoterapia hasteko. DIAGNOSTIKOA 1 motako diabetesaren debuta. Desoreka hipergluzemiko zetoazidosikoa. Ez zegoen faktore azkartzaile berezirik. TRATAMENDUA Intsulinaren infusioa bena barnetik 6 IU/h (intsulinaren 50 IU 500 ml-ko serumean, 0.1 IU/ml lortuta, hau da, hasteko, 60 ml/h jarriz) Gluzemia kapilarraren kontrola, bi ordutik behin. Zetonemiaren kontrola, bi ordutik behin. Serum fisiologikoa, 500 ml/8 ordutik behin, eta% 5eko serum glukosatua, 500 ml/8 orduz behin, biak batera " Y eran " jarriak. Tratamendu honen lehenengo 4 orduak ClK jarri gabe, ondoren hasi ClK serumean diluitua jartzen (20 meq 500 ml serumeko).

Irteera :
oso altua |gluzemia
120/50 |Tentsio
98 tau/min |BM
98 tau/min |bihotz-maiztasuna
20 arnas/min |maiztasuna
+ |Fetor
365 mg/dl |Gluzemia
+ + + |Zetonuria
normalak |pultsu
95 tau/min |BM
95 tau/min |bihotz-maiztasuna
1,65 mg/dl |Kreatinina
62 mg/dl |urea
94 meq/l |Cl
134 meq/l |Na
5,2 meq/l |K
427 mg/dl |glukosa
normalak |troponina
normalak |CPK
normalak |CKMB

26 |GPT
21 |GOT
31 |GGT
26 |amilasa
14,4 g/dl |hemoglobina
162.000/ml |plaketak
16.580/ml |leukozitoak
1,07 |INR
31,1 seg |APTT
7,15 |pH
10 |HCO3
-16,5 |BG
-16,5 |base
112 |pO2
16 |pCO2
Normala |GERNU-JALKINA
1.000 mg/dl |Glukosuria
50 mg/dl |Zetonuria

Oharra: emaitza idazten da lehenik eta gero probaren izena irteeran. ’|’z bereizten dira.

Orain eman iezadazu testu honen irteera:

{New Statement}

Inprimatu bakarrik irteera existitzen bada eta kito.

5.11.4 English prompt

```
system
You are provided with a clinical statement in [[language]].
Your task is Relation Extraction. Extract mentions of laboratory tests
and their results from the statement. Note: the result is written
first in the output and then the name of the test. They are separated
with a ‘|’. Print only the relations if any and nothing else.

user
[[ example statement ]]

assistant
[[ example relations ]]

user
[[ new statement ]]
```