Medical Information Extraction with Large Language Models

Raffaello Fornasiere, **Nicolò Brunello**, **Vincenzo Scotti** and **Mark James Carman**
DEIB, Politecnico di Milano
Via Ponzio 34/5, 20133, Milano (MI), Italy
raffaello.fornasiere@mail.polimi.it nicolo.brunello@polimi.it
vincenzo.scotti@polimi.it mark.carman@polimi.it

Abstract

The increase in clinical text data following the adoption of electronic health records offers benefits for medical practice and introduces challenges in automatic data extraction. Since manual extraction is often inefficient and error-prone, with this work, we explore the use of open, small-scale, Large Language Models (LLMs) to automate and improve the extraction of medication and timeline data. With our experiments, we aim to assess the effectiveness of different prompting strategies –zero-shot, few-shots, and sequential prompting– on LLMs to generate a mixture of structured and unstructured information starting from a reference document. The results show that even a zero-shot learning approach can be sufficient to extract medication information with high precision. The main issues in generating the required information seem to be completeness and redundancy. However, prompt tuning alone seems to be sufficient to achieve good results using these LLMs, even in specific domains like the medical one. Besides medical information extraction, in this work, we address the problem of explainability, introducing a line-number referencing method to enhance transparency and trust in the generated results. Finally, to underscore the viability of applying these LLM-based solutions to medical information extraction, we deployed the developed pipelines within a demo application.

1 Introduction

The rapid integration of digital technologies into healthcare systems has transformed the landscape of patient care and management. *Electronic Health Record* (EHR) systems have become pivotal in modern healthcare environments. However, as a downside, primary care physicians, for example, face a significant burden of documentation. Research indicates that family medicine physicians allocate nearly as much time to interacting with EHR systems as they do to direct patient care (Arndt et al., 2017), leading to reduced clinical efficiency and increased risk of clinician burnout.

To address these issues and automate (or semi-automate) the analysis of these documents and, thus, reduce clinicians burden, we explore the application of *Large Language Models* (LLMs) (Brown et al., 2020; OpenAI, 2023; Anil et al., 2023) as a means to enhance the functionality and efficiency of EHR systems. LLMs are the pivot of the current advancements in *Artificial Intelligence* (AI), present promising solutions for automating routine documentation, extracting information from unstructured data and supporting clinical decision-making through real-time insights from extensive medical databases.

Specifically, with this paper, we explore the application of small-scale openly-available LLMs (Touvron et al., 2023; Jiang et al., 2023; Mesnard et al., 2024) to automate the extraction of medication information and timeline data from clinical text. We evaluate LLMs performance in *zero-shot learning*, *few-shot learning* and *sequential prompting* scenarios. We selected the latter approach to guide the LLM through the multiple steps of information extractions in the cases where the information is not immediately accessible from the raw text. The objective of the evaluations is to assess the accuracy and completeness of the information LLMs extract, such as dosage, frequency, and mode of administration of a drug, as well as LLMs ability to construct patient timelines from clinical narratives. Through this work, we seek not only to deepen our understanding of the capabilities and reliability of LLM-based systems in medical contexts, but also to offer viable strategies for alleviating the documentation burden that detracts from patient-focused healthcare that can serve as possible baselines.

We divide this paper into the following sections. In Section 2, we recap the main results in information extraction. In Section 3, we describe the

pipelines we developed for information extraction. In Section 4, we describe the data sets we used to evaluate our pipelines. In Section 5, we outline the experiments we conducted. In Section 6, we report and comment on the experimental results. Finally, in Section 7, we summarise our work and present possible future directions.

2 Related Works

Information Extraction (IE) is one of the main applications of *Natural Language Processing* (NLP), even outside the medical domain. Traditionally, information extraction encompasses problems like *Named Entities Recognition* (NER), *Relation Extraction* (RE) or *Aspect Classification* (AC) (Jurafsky and Martin, 2024, Chapter 19). NER involves the extraction of named entities like persons and locations, as well as time expressions and even drugs. RE is the task of classifying relations among entities, like the dosage of a specific drug. AC is the classification of events according to their internal temporal structure or temporal contour, for example, identifying whether a patient has been taking a drug before or after hospitalisation.

Initially, these problems have been approached with either rule-based systems or classification models combined with *Conditional Random Fields* (CRF) (Jurafsky and Martin, 2024, Chapter 19). Rule-based techniques are known for their precision in entity recognition or relation extraction, particularly when they are meticulously crafted to align with specific data types. These methods typically analyse sentence structures and leverage *Part-of-Speech* (PoS) tags to enhance NER. Both rule-based systems and classification models relied on hand-crafted features and lexical resources to identify medical entities (Landolsi et al., 2024). While they are less flexible and harder to scale, they perform reasonably on well-defined problems.

The advances introduced by *word embeddings* combined with sequence processing techniques based on deep learning like *Recurrent Neural Networks* (Elman, 1990; Hochreiter and Schmidhuber, 1997) and *Transformer Networks* (Vaswani et al., 2017) helped push forward significantly state of the art for IE. In fact, even now, many approaches often favour a combination of *Bi-directional Long Short-Term Memory* (BiLSTM) (an RNN variant) and CRF models or more recent *fine-tuned* Transformers (Symeonidou et al., 2019; Yang et al., 2020; Kafikang and Hendawi, 2023). Bi-directional models (both recurrent and Transformer) excel in capturing high-quality features due to their ability to account for contextual dependencies in both forward and backward directions. Meanwhile, CRF enhances the process by optimising sequence tagging with these features (Çelkmasat et al., 2022). These models exploit a *Begin-Inside-Outside* (BIO) tagging system which allows segmenting an input document into multiple pieces (delineating entity boundaries, for example) while labelling those same pieces (thus, recognising the type of entity, for example). Contextual models like RNN and Transformers play a crucial role in medical information extraction especially when pre-trained on medical texts so that they can incorporate domain knowledge (Lee et al., 2020; Landolsi et al., 2024).

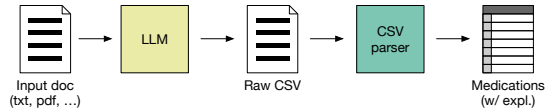
As with many other NLP tasks, LLMs have revolutionised IE as well, offering near state-of-the-art performances out of the box. The in-context learning capabilities of LLMs like *GPT-4* (OpenAI, 2023) or *Gemini* (Anil et al., 2023) have shown promising directions for biomedical NER and RE, especially in scenarios lacking labelled data. Despite these advancements, these LLMs still do not outperform consistently smaller models fine-tuned on task-specific datasets yet (Tian et al., 2023). Additionally, the use of LLMs in IE faces several challenges: For instance, the generative nature of these models may alter the phrasing of recognised entities or predicted relationships, complicating the verification process. Moreover, these LLMs are prone to hallucinations that may lead to the generation of entities and relationships that appear plausible but are not factually accurate. Furthermore, finding suitable prompts for NER and RE tasks can be difficult. These issues underscore the need for further research to explore and develop more effective methods for effectively using LLMs in IE. In this paper, we focus on the medical domain, and we explore solutions for medical IE from patients' records.

3 Methodology

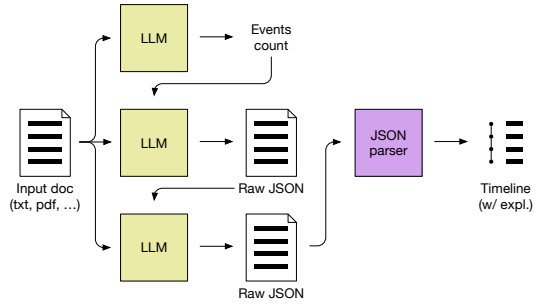
In this section, we describe the two pipelines we propose for medical information extraction from clinical documents. We provide an overview of the pipelines and information extraction approaches in Section 3.1, and then we provide additional details on the explainability in Section 3.2. Finally, we provide practical details on the two tasks of *medication extraction* and *timeline extraction* in

Section 3.3.

3.1 Information extraction overview



(a) Medication.



(b) Timeline.

Figure 1: Information extraction pipelines.

The pipeline (depicted in Figure 1) we propose is designed to give a reference medical document (e.g., a discharge letter) as input to an LLM and use the LLM to extract the desired information from the document according to user-provided instructions. As premised, in this work, we focus on medication extraction (i.e., the extraction of information about the drug regime for a patient) and timeline extraction (i.e., the sequence of events characterising a patient's clinical history). The pipeline includes the generation of the required information from the text, the parsing of the generated unstructured text and the rendering of the structured data selected from parsing. An additional passage that enriches our pipeline is that of explainability to justify the generated content.

We considered three different approaches to deal with the information extraction, independently of the actual task:

Zero-shot learning where we provide the LLM with the instructions of the task to complete and a description of the expected output;

Few-shots learning where we provide the same instruction as zero-shot learning, but before asking for the current sample, we append some examples of input and expected output to help guide the generation process;

Sequential prompting where we have the same settings of zero-shot learning, but we break down the task into multiple steps to help the LLM build the solution one piece at a time and keep it aligned with the desired behaviour.

All the aforementioned pieces are deployed as part of a web app demonstrating AI solutions for healthcare (see Figure 2). The demo is agnostic of the underlying LLM. It allows the loading of a reference document and separately generating the table with the medication information, generating the timeline and asking questions to the chatbot about the document. The raw outputs of information extraction are parsed to be converted into structured information and then displayed on the demo web page.

3.2 Explainability

Explainability has become a more and more important step in developing and deploying deep learning-based systems. Explainability helps in understanding where model predictions come from. When it comes to healthcare, the attention to this information is even more crucial.

In our tool, we suggested a simple yet effective solution to explain the results of information extraction. We have a separate pipeline ingesting the clinical document to analyse decorated with the rows numbers. In this way, we can interrogate the LLM automatically asking to point out the number of the row connected to a specific extraction (e.g., where is a specific drug mentioned or where is a specific event mentioned).

This additional explanation can be useful for the clinician. In fact, on one side, having an explanation helps ground the predicted information. On the other side, it helps spot possible errors due to LLM faults, preventing the misinformation of the clinician.

3.3 Medications and Timeline extraction

As premised, in the deployed demo, we approach both medication information extraction and timeline extraction. We approach medication extraction with a zero-shot learning approach and timeline extraction with sequential prompting. We selected these approaches given the results of the experiments we conducted. In both tasks we use the model in *assistant chatbot* format (Scotti et al., 2024), composing a sequence of messages to explain and solve the tasks.

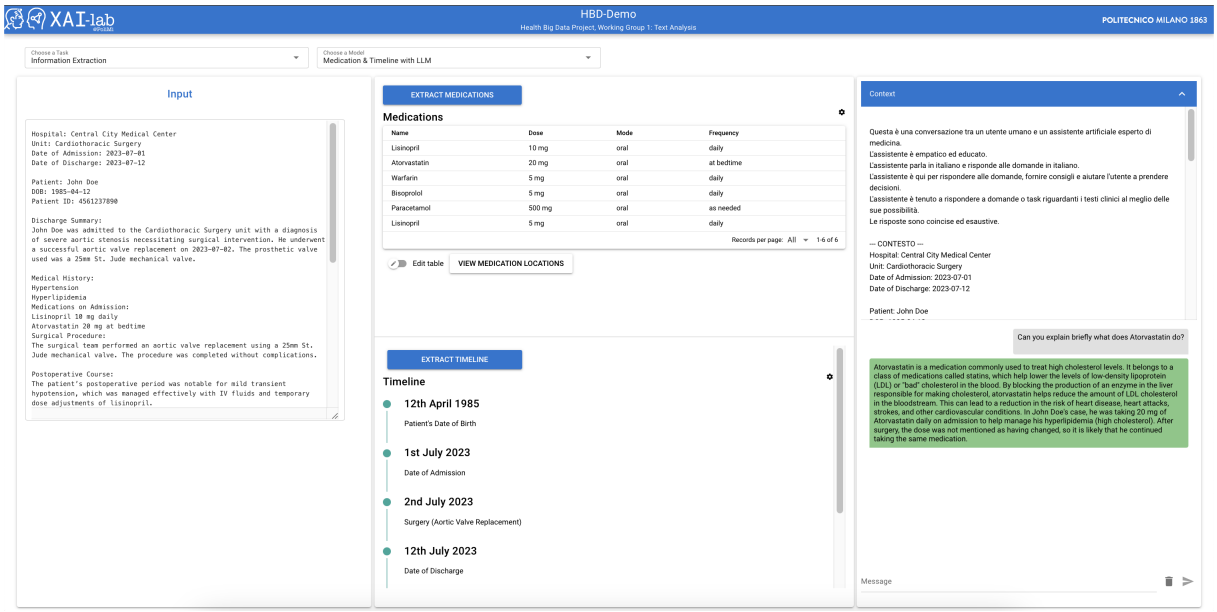


Figure 2: Demo tool using the pipeline to extract medications and medical events from a reference document.

Concerning medication extraction, we task the model to extract all the information at once from a reference document. We provide the reference document to the model as part of the system message, we then append a user message describing the task, and, finally, we force the answer of the assistant to start with the raw text content of the CSV file we want as output. We get the LLM to generate starting from these messages. The CSV table contains the following information about the medication: *name*, *dose* (the specified amount of medication), *mode* (intended as mode of administration), *frequency* (how many times or how often to take the medication), *line* (line in the text where the medication information is mentioned, for explainability).

Concerning the timeline extraction, we follow a sequential prompting approach. We broke down the task into three steps: counting the events mentioned in the document, generating a JSON array with the chronologically ordered events and generating the line number for each event on the array. As for the other task, we provide the reference document as part of the system message and then we alternate user messages with the instructions for the current step and model responses for that step. The elements of the JSON array with the chronological order are *dateValue* (date in the format `"YYYY-MM-DD"`), *dateString* (the string mentioning the date as it appeared in the original document) and *event* (a short description of what the event that occurred at that time point). Each element is decorated in the last step with the line number for explainability. We found empirically these steps to be the most effective to generate the timeline.

4 Data

One of the challenges of working in the healthcare domain is gathering usable data. Given the nature of the task, we focused on finding data sets containing similar samples to what the model would encounter in real-world scenarios. For this project, we resorted to two existing data sets, one for medication extraction and one for timeline extraction (we describe them respectively in Section 4.1 and Section 4.1), and we generated a third additional data set synthetically (we describe this third data set in Section 4.1).

4.1 N2C2

The *National NLP Clinical Challenges* (N2C2) data set (Uzuner et al., 2010) is a collection of 1243 de-identified discharge summaries from *Partners Healthcare*. This data set was released as part of a medical annotation challenge. In the challenge, participants were tasked with extracting medication information from these summaries and collectively provided annotations for 251 documents. The dataset focuses on the identification of medications and medication-related information, including *dosages*, *routes* (i.e., models of admin-

istration), *frequencies*, *durations*, and *reasons* for administration.

Listing 1: Example of the N2C2 input document with numbered lines.

```
41 HASSEL , EDWARDO D. , M.D
42 on order for NEPHROCAPS PO (
       ref 12327843 )
43 POTENTIALLY SERIOUS
       INTERACTION: SIMVASTATIN
       NIACIN ,
44 VIT. B-3 Reason for override:
       home regimen
45 Previous override information:
46 Override added on 4/29/04 by
       GALIPEAU , ENRIQUE R. , M.
       D.
47 DEFINITE ALLERGY ( OR
       SENSITIVITY ) to HMG CoA
       REDUCTASE
48 INHIBITORS Reason for override
       : md aware , home regimen
49 IMDUR ( ISOSORBIDE MONONIT.(
       SR ) ) 30 MG PO QD
50 Food/Drug Interaction
       Instruction
51 Give on an empty stomach (
       give 1hr before or 2hr
       after
```

Listing 2: Example of the N2C2 output labels.

```
m="nephrocaps" 42:3 42:3||do="nm
    "||mo="po" 42:4 42:4||f="nm"||
    du="nm"||r="nm"||ln="list"
m="niacin" 43:5 43:5||do="nm"||mo
    ="nm"||f="nm"||du="nm"||r="nm
    "||ln="list"
m="simvastatin" 43:3 43:3||do="nm
    "||mo="nm"||f="nm"||du="nm"||r
    ="nm"||ln="list"
m="vit.\ b-3" 44:0 44:1||do="nm
    "||mo="nm"||f="nm"||du="nm"||r
    ="nm"||ln="list"
m="imdur ( isosorbide mononit.(
    sr ) )" 49:0 49:6||do="30 mg"
    49:7 49:8||mo="po" 49:9 49:9||
    f="qd" 49:10 49:10||du="nm"||r
    ="nm"||ln="list"
```

The annotations provide the precise location of each piece of information within the discharge summaries, facilitating the development and evaluation of NLP systems for medication information extraction. We report examples of input document (chunk) and corresponding annotations respectively in Listing 1 and Listing 2. As can be evicted by the annotations, the target data contain all the desired details and their reference within the document.

I2B2

The *Informatics for Integrating Biology and the Bedside* (I2B2) data set (Sun et al., 2013), released as part of the homonymous project, consists of 310 discharge summaries annotated with temporal information. This data set was created to facilitate the development and evaluation of NLP systems for temporal reasoning in clinical text.

Listing 3: Example of the N2C2 input document with numbered lines.

```
41 Admission Date :
42 09/29/1993
43 Discharge Date :
44 10/04/1993
45 HISTORY OF PRESENT ILLNESS :
46 The patient is a 28-year-old
       woman who is HIV positive
       for two years .
47 She presented with left upper
       quadrant pain as well as
       nausea and vomiting which
       is a long-standing
       complaint .
48 She was diagnosed in 1991
       during the birth of her
       child .
49 She claims she does not know
       why she is HIV positive .
```

Listing 4: Example of the I2B2 output labels in XML format.

```
<timex3 id="T0" start="18" end
    ="28" text="09/29/1993" type="
    DATE" val="1993-09-29" mod="NA
    " />
<timex3 id="T13" start="2249" end
    ="2271" text="the day of
    discharge ." type="DATE" val
    ="1993-10-04" mod="NA" />
<timex3 id="T3" start="290" end
    ="294" text="1991" type="DATE"
     val="1991" mod="NA" />
```

The annotations focus on three key aspects. *Events*, which include clinical concepts (problems, tests, treatments), clinical departments, evidential information (source of information), and occurrences (e.g., admissions and transfers). Each event is further categorised by type, polarity (positive or negated), and modality (factual, proposed, conditional, or possible). *Temporal expressions*, which include dates, times, durations, and frequencies, normalised to the ISO8601 standard. Each temporal expression is characterised by its type, value, and modifier (exact or approximate). *Temporal Relations* (TLinks), which describe the relationships between events and temporal expressions, indicating whether one occurred before, after, or overlapped with another. We report examples of input document (chunk) and corresponding annotations respectively in Listing 3 and Listing 4. As for the previous data set, the XML entries containing the labels are annotated also with the position of the information withing the source document.

Synthetic Data Set

Given the reduced size of the I2B2 data set, we resorted to ChatGPT to generate some additional data. We refer to this as *synthetic data set* (SD). We prompted ChatGPT 4 to create discharge summaries in both English and Italian[1], along with corresponding annotations for relevant medical information. We report examples of input document (chunk) and corresponding annotations in Listing 5. Differently from the previous two data sets, we have a single entry containing both the input document and the target labels, without explicit annotations of the position of the information within the document (it would have been unrelialable to use ChatGPT annotations for this information, which we can extract searching the matching substrings in the source document)

In general synthesising data offers several advantages, like *personalisation*, *privacy* and *control*. From the personalisation perspective, we have that the content and style of the generated summaries can be tailored to specific requirements, allowing for the creation of diverse and representative samples. Concerning privacy, since the data is synthetic, it inherently avoids privacy concerns associated with real patient data. Finally, about control, we have that the generation process allows for precise control over the types of medical information included, enabling targeted testing of specific extraction challenges. However, it's important to acknowledge that synthetic data may not fully capture the nuances and complexities of real-world clinical documentation. While it serves as a valuable resource for preliminary testing and development, its limitations should be considered when interpreting results and generalising findings to real-world scenarios.

[1]We worked with Italian documents to fit the requirements of the project founding this work; for further details, refer to the acknowledgements at the end of this paper.

Listing 5: Example of the SD input dcoments and output annotation in JSON format.

```
{
  "text": "**Discharge Summary
      :**\n\nPatient: Mark Johnson
       \nAge: 38  \nAdmission
      Date: 03-20-2024  \
      nDischarge Date: 03/28/24  \
      n\nPatient History:\nMr.
      Mark Johnson, a 38-year-old
      male, was admitted to our
      facility on March 20, 2024,
      presenting with complaints
      of abdominal pain, nausea,
      and jaundice. He has a past
      medical history ...",
  "annotations": [
    {
      "text": "March 20, 2024",
      "date_value": "2024-03-20"
    },
    {
      "text": "March 24, 2024",
      "date_value": "2024-03-24"
    },
    {
      "text": "eight days",
      "date_value": "2024-03-28"
    },
    ...
  ]
}
```

5 Experiments

In this section, we detail the experiments we run to evaluate our pipelines for medication extraction (Section 5.1) and timeline extraction (Section 5.2). In all the experiments we conducted, we worked with *Mistral 7B* (Jiang et al., 2023), using this LLM as the core of the information extraction system.

5.1 Medication Extraction

In the first set of experiments, we focused on medication extraction from clinical texts. The primary objective is to evaluate the models' ability to extract medication details such as dosage, mode, and frequency from unstructured medical documents, like discharge letters. For this task, we focused on the N2C2 data set. We evaluated the LLM capabilities with different approaches: zero-shot learning, few-shots learning (using 2 examples) and sequential prompting. We conducted the evaluation using the standard metrics: precision, recall, and F_1 score.

Initially, we considered two variants of this task: looking for *full medications* (i.e., we asked the LLM to generate all the medication details: name, dosage, mode, and frequency) or not (i.e., we asked the LLM to generate only the name of the medication). However, as we explain better in Section 6.1, working with full medication yields poor results, as we noticed immediately in the early experiments with a zero-shot learning approach. To measure the metrics in the full medications case, we considered a single string containing all the details, and to measure a match, we standardised the target string and the generated one by removing all spaces and special characters.

Concerning the input and output format, we considered multiple alternatives as well. We explored having as input the whole document to analyse or only a relevant chunk, this approach is helpful with particularly long documents. Moreover, we explored two different output formats: JSON and CSV; in both cases we had the LLM generate directly the raw JSON or CSV strings.

5.2 Timeline extraction

In the second set of experiments, we focused on extracting patient timelines from clinical texts in order to highlight all the relevant events. In this case, we focused only on evaluating the model's capabilities in extracting correctly formatted dates. In fact, from early explorations, we noticed that this task was already challenging as the LLM often deviated from the target format. For this task, we used the I2B2 data set and the synthetic data set. We conducted the evaluation using the standard metrics: precision, recall, and F_1 score.

As for the previous experiment, we evaluated the LLM capabilities with zero-shot learning, few-shots learning (using 4 examples) and sequential prompting approaches. Concerning the input and output format, similar to medication extraction, we considered alternative approaches. As before, we explored using the whole document as input or only a relevant chunk. For the output, we worked only in JSON format and we converted all dates in `"YYYY-MM-DD"` format.

6 Results

In this section, we present and comment on the results of the experiments on medication extraction (Section 6.1) and timeline extraction (Section 6.2). In both cases, we do not compare with the reference baselines coming with the data sets since we approach the evaluations differently and we compute different metrics.

6.1 Medication Extraction

We report the results of this first task of medication extraction in Table 1. Results on precision focusing on the medication name are satisfying, meaning that the model is missing very few medications from the documents. However, the low recall and, subsequent, low F_1 scores hint that the model is often generating information that is not part of the original document. Moreover, results using full medication information are consistently lower, indicating that, as expected, extracting detailed information is harder than simply identifying the medication.

The experiments with zero-shot approach showed that the LLM is not capable of extracting all the medication information just from the instructions. Looking at the generated output, we noticed that sticking to the target output format was difficult, and even output post-processing and string normalisation were not sufficient to match the target and predicted output. CSV format seems to be harder to get to work independently of the target being name only or full medication information.

From the results of the few-shots approach and sequential approach, there seems to be no clear solution for the output format. In fact, depending on the approach, generating CSV or JSON output seems to yield the best results. Concerning the difference between the approaches, there is not clear difference between zero-shot and sequential approaches. Few-shots approach does not improve significantly over the other approaches over precision, but improves the recall and, thus, the F_1.

Approach	Format	Chunked docs	Full medication	Precision	Recall	$\mathbf{F}_1$
zero-shot	JSON	✗	✗	0.964	0.392	0.513
		✗	✓	0.446	0.115	0.181
	CSV	✗	✗	0.557	0.453	0.498
		✗	✓	0.418	0.217	0.278
few-shots	JSON	✗	✗	0.885	0.479	0.606
		✗	✓	0.364	0.109	0.166
		✓	✗	**0.965**	0.547	**0.683**
		✓	✓	0.616	0.243	0.342
	CSV	✗	✗	0.857	0.546	0.660
		✗	✓	0.366	0.136	0.198
		✓	✗	0.837	0.526	0.636
		✓	✓	0.380	0.160	0.224
sequential	JSON	✗	✗	0.961	0.358	0.512
		✗	✓	0.597	0.134	0.217
	CSV	✗	✗	0.808	0.318	0.442
		✗	✓	0.288	**0.550**	0.378

Table 1: Results on N2C2 for medication extraction (bold values correspond the best score).

Dataset	Approach	Chunked docs	Precision	Recall	$\mathbf{F}_1$
I2B2	zero-shot	✗	0.811	0.589	0.651
	few-shots	✗	0.803	**0.794**	**0.790**
		✓	**0.954**	0.592	0.701
	sequential	✗	0.757	0.644	0.660
SD	zero-shot	✗	0.949	0.806	0.861
	few-shots	✗	0.926	**0.917**	0.916
		✓	**0.975**	0.898	**0.931**
	sequential	✗	0.966	0.898	0.926

Table 2: Results on I2B2 and Synthetic Data (SD) for timeline extraction (bold values correspond the best score for each data set).

6.2 Timeline extraction

We report the results on this second task of timeline extraction in Table 2. As can be seen, the results are good, yet there is a lot of space for improvement. Results on the synthetic data are always better than those on the I2B2 data set.

Comparing the results of zero-shot and few-shots learning, we can see that in most cases, using the few-shots approach helped significantly improve the results on recall and, thus, F_1. The higher results on chunked documents seem to indicate that, in this case, using longer documents negatively affects the ability to extract the time information.

Both sequential prompting and zero-shot work without reference examples, yet sequential prompting performed in terms of recall and F_1, and performed comparably to the few-shots approach. This hints that the sequential approach helped the LLM capture better the target task and output format.

7 Conclusion

In this paper, we showed how we approached the problem of medical information and events extraction using LLMs. The results of the conducted experiments highlight the potential of these LLMs for

automating the extraction of this information from clinical texts. The performance of these models resulted sufficiently robust for practical application in real-world settings, though there is still room for further improvements. To complete the proposed pipelines and make them more reliable, we provided also an explanation tool.

Concerning the evaluations, the LLMs exhibited significantly better performance in few-shot learning settings when compared to zero-shot learning ones, achieving, as expected, higher precision, recall, and F1 scores. However, it is important to point out that the effectiveness of the LLM varied significantly depending on factors such as the chosen output format (JSON vs CSV). For instance, although the models are capable of adapting to the requested output format, it remains unclear which format yields the most effective results. While, in some cases, the performance we achieved is suitable for practical application, these fluctuations pinpoint a challenge taht highlights the need for better models before moving to real-world applications of the LLM technology for healthcare.

To improve the overall pipeline robustness and utility, we will be working on minimising the LLM's sensitivity to minor variations in prompts, for example, working on our own fine-tuning for chatbot assistant or instruction following rather than resorting to existing solutions. Similarly, we are interested in exploring alternative evaluation metrics that assess the semantic accuracy of the extracted information, rather than relying solely on string matching. We expect that advancements in these two directions will better gauge the practical applicability and effectiveness of LLMs in processing clinical texts. At the same time, to expand the tool capabilities, we are interested in exploring more complex scenarios, where the information to extract is scattered across multiple documents, which represent a more challenging task also from the explainability perspective.

Limitations

In this paper, we mainly focused on the development and deployment of the pipeline, rather than exhaustive experiments. The first limitation is in the choice of the LLM: as for now, we evaluated the results using only Mistral 7B. A proper evaluation would require exploring other openly accessible models of the same and different sizes and closed-access models to have a reference for the comparisons. The second limitation is the size of the available data sets, which we can consider small if compared with data sets for other information extraction tasks, thus the results may be subject to high variance.

Ethics Statement

The authors do not foresee any considerable risks associated with the work presented in this paper. In principle, the presented framework is intended for information extraction from medical documents as it is thought to be used by clinicians (or other similar experts) who have authorised access to the target documents. The authors pledge to make the source code publicly available to ensure the reproducibility of the experiments.

Acknowledgements

This work has been supported by the Health Big Data project, funded by the Italian Ministry of Economy and Finance and coordinated by the Italian Ministry of Health, and results from the interaction with numerous knowledgeable researchers, clinicians, and other field experts from the Politecnico di Milano and the IRCCSs (Istituto di ricovero e cura a carattere scientifico). The work was also partially supported by the FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, Investment 1.3, Line on Artificial Intelligence).

References

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Brian G. Arndt et al. 2017. Tethered to the ehr: Primary care physician workload assessment using ehr event

log data and time-motion observations. *Annals of Family Medicine*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Gökberk Çelkmasat, Muhammed Enes Aktürk, Yunus Emre Ertunç, Abdul Majeed Issifu, and Murat Can Ganiz. 2022. Biomedical named entity recognition using transformers with bilstm + CRF and graph convolutional neural networks. In *International Conference on INnovations in Intelligent SysTems and Applications, INISTA 2022, Biarritz, France, August 8-12, 2022*, pages 1–6. IEEE.

Jeffrey L. Elman. 1990. Finding structure in time. *Cogn. Sci.*, 14(2):179–211.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Dan Jurafsky and James H. Martin. 2024. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition (3rd edition). Draft.

Maryam Kafikang and Abdeltawab M. Hendawi. 2023. Drug-drug interaction extraction from biomedical text using relation biobert with BLSTM. *Mach. Learn. Knowl. Extr.*, 5(2):669–683.

Mohamed Yassine Landolsi, Lobna Hlaoua, and Lotfi Ben Romdhane. 2024. Extracting and structuring information from the electronic medical text: state of the art and trendy directions. *Multim. Tools Appl.*, 83(7):21229–21280.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Cristian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, and et al. 2024. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Vincenzo Scotti, Licia Sbattella, and Roberto Tedesco. 2024. A primer on seq2seq models for generative chatbots. *ACM Comput. Surv.*, 56(3):75:1–75:58.

Weiyi Sun, Anna Rumshisky, and Özlem Uzuner. 2013. Annotating temporal information in clinical narratives. *J. Biomed. Informatics*, 46(6):S5–S12.

Anthi Symeonidou, Viachaslau Sazonau, and Paul Groth. 2019. Transfer learning for biomedical named entity recognition with biobert. In *Proceedings of the Posters and Demo Track of the 15th International Conference on Semantic Systems co-located with 15th International Conference on Semantic Systems (SEMANTiCS 2019), Karlsruhe, Germany, September 9th - to - 12th, 2019*, volume 2451 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C. Comeau, Rezarta Islamaj, Aadit Kapoor, Xin Gao, and Zhiyong Lu. 2023. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *CoRR*, abs/2306.10070.

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, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *J. Am. Medical Informatics Assoc.*, 17(5):514–518.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Xi Yang, Jiang Bian, William R. Hogan, and Yonghui Wu. 2020. Clinical concept extraction using transformers. *J. Am. Medical Informatics Assoc.*, 27(12):1935–1942.