Evaluating LLMs for Temporal Entity Extraction from Pediatric Clinical Text in Rare Diseases Context

Judith Jeyafreeda Andrew, Marc Vincent, Anita Burgun, Nicolas Garcelon
Université de Paris, Imagine Institute, Data Science Platform, INSERM UMR 1163
PaRis Artificial Intelligence Research InstitutE (PRAIRIE)
Bd de Montparnasse, Paris, France
{judith.jeyafreeda, marc.vincent, nicolas.garcelon}@institutimagine.org
anita.burgun@aphp.fr

Abstract

The aim of this work is to extract Temporal Entities from patients' EHR from pediatric hospital specialising in Rare Diseases, thus allowing to create a patient timeline relative to diagnosis . We aim to perform an evaluation of NLP tools and Large Language Models (LLM) to test their application in the field of clinical study where data is limited and sensitive. We present a short annotation guideline for temporal entity identification. We then use the tool EDS-NLP, the Language Model CamemBERT-with-Dates and the LLM Vicuna to extract temporal entities. We perform experiments using three different prompting techniques on the LLM Vicuna to evaluate the model thoroughly. We use a small dataset of 50 EHR describing the evolution of rare diseases in patients to perform our experiments. We show that among the different methods to prompt a LLM, using a decomposed structure of prompting method on the LLM Vicuna produces the best results for temporal entity recognition. The LLM learns from examples in the prompt and decomposing one prompt to several prompts allows the model to avoid confusions between the different entity types. Identifying the temporal entities in EHRs helps to build the timeline of a patient and to learn the evolution of a diseases. This is specifically important in the case of rare diseases due to the availability of limited examples. In this paper, we show that this can be made possible with the use of Language Models and LLM in a secure environment, thus preserving the privacy of the patient.

Keywords: Temporal Entities, Vicuna, Prompt-based learning, rare diseases

1. Introduction

Electronic Health Records (EHR) contain several valuable information that help in advancing clinical research. Automatic extraction of information from EHRs has evolved greatly overtime with the development of Machine Learning and Natural Language Processsing (NLP) techniques. In the present article we focus on a sub-task of NLP: Named Entity Recognition (NER) of temporal entities. In particular, we aim at extracting temporal entities from EHRs of patients with Rare Diseases. Identifying the temporal Entities in such texts allows to build the timeline of a patient, allowing for the analysis of patient history, prediction of next steps in the process of diagnosing a disease and the evolution of a patient after a therapeutic decision has been taken. This is a very important application in the field of rare diseases where the data is limited.

There have been several research works for the automatic extraction of information from clinical texts. These works have enabled building several novel methods and models for the extraction of useful information within the clinical texts such as drugs, treatments, diagnosis, symptoms, etc. However, to be able to create a timeline of a patient, the relations between these entities and temporal entities such as date, time, duration etc must be established. For this purpose, the extraction of temporal entities is also essential.

Although there have been considerable efforts in making de-identified EHRs publicly available, accessible after considerable ethical training, the language and format of the EHRs influence greatly the development of Large Language Models for Information Extraction. Models and methods that perform well for the English Language do not necessarily have the same performance on the French language. Also, the format of EHR used in a clinic might not be the same as the format used in another clinic, this also affects the performance of a model. Thus external validation of LLMs with local datasets is essential.

Thus in this work, we use EHR from patients with rare disease particular to the Necker Hospital in Paris [1] [2] for the extraction of temporal entities. Our contributions in this paper are as follows: (i) a short annotation guideline that has been used for a manual annotation. (ii) using existing tools and Large Language models for temporal entity extraction to study their performance and re-usability in a secure environment.(iii) applied to real hospital data.

[1] Hospital located at 149 Rue de Sèvres, 75015 Paris
[2] The dataset is private and cannot be distributed

Proceedings of CL4Health 2024 @LREC-COLING-2024, pages 145–152
20 May, 2024.

2. Related Works

(Bose et al., 2021) gives a detailed study on all NER methods and models available in the clinical context until the year 2020. The entities that are often used in the clinical context are drugs, diagnosis, treatment, dosage, family history etc. The methods of NER used include dictionary-based approach, rule-based approach, CRF, Machine Learning based approach, Deep Learning-Based Methods and some hybrid approaches. The authors show the several models that are available in different language, most being in Chinese and some in English. Although this study does not mention extraction of temporal entities, there has been several works done in the field of temporal relation extraction in clinical text in the English Language. (Alfattni et al., 2020) points to the general approach used in Temporal Relation Extraction which include pre-processing, NER of EVENTS and TIMEX entities, TLINK candidate extraction, TLINK classification and post-processing, thus indicating the importance of having an efficient temporal entity recognition method for the task of Temporal Relation Extraction. Within the context of clinical texts in the French language, (Tourille, 2018), has studied various approaches for NER within the clinical context and presented the results on publicly available French corpora. The author uses an LSTM approach with inspiration from sequence labelling for the purpose of NER, while the temporal relation extraction relies solely on LSTM. Lastly, in (Vincent et al., 2022) and (Faviez et al., 2022) the authors use deep learning and hybrid NER methods to perform deep phenotyping on a specialised rare disease dataset, using the resulting models and information extraction to augment the UMLS metathesaurus with specific and previously not included terms.

2.1. Prompt Learning for NER

Prompt learning has gained increasing popularity with the development of LLM and they have been used successfully for several NLP applications (Brown et al., 2020). Prompt learning involves using prompts which are injected to the input into a designed template. This converts the downstream task into a fill-in-the-blank task, then allows the language model to predict the slots in the prompts and eventually deduce the final output. This is often used for text generation and classification tasks. There have been several research works on the several prompting techniques such as discrete and continuous prompt templates (Jiang et al., 2020), (Shin et al., 2020), (Liu et al., 2023),(Li and Liang, 2021), (Lester et al., 2021), (Qin and Eisner, 2021). (Cui et al., 2021) is one of the first attempts in using prompt learning for NER. (Ashok and Lipton, 2023) introduces PromptNER, where a text and a task description is given along with the question for the prediction of entities. This has been tested on the biomedical dataset GENIA (Kim et al., 2003) for NER and outperforms competing models like GPT 3.5. In (Liu et al., 2022), the authors present QaNER, which is a prompt-based learning NER method with Question Answering. The authors of (Ye et al., 2023) propose a decomposed two-stage prompt learning framework for few-shot named entity recognition, which include the entity location and entity typing stages. (Shen et al., 2023) unify entity locating and entity typing in prompt learning for NER with a dual-slot multi-prompt template. (Huang et al., 2022) proposes a few-shot NER approach named COPNER, which combines contrastive learning and prompt guiding, where the prompt is concatenated with the sentence and is then fed to a pre-trained language model.

In this work, we use three different prompts with the "Vicuna" large language model (LLM). The first prompt is a basic question which asks the LLM to identify all the temporal entities in a given clinical text. The second prompt, is a definition based prompt where the entities are defined as part of the prompt which helps the LLM understand the entities that are to be identified. For the third prompt, we decompose the prompt into different prompts (one for each entity).

3. Dataset

As mentioned previously, the language and format of clinical text have a great deal of influence to the performance of large language models. (Youssef A, 2023) has stressed the need for external evaluation in the setting where the LLM models are to be deployed. The selection of testing dataset would depend on the setting of the deployment environment. In this work, we focus on clinical texts in the French language. Our dataset is a collection of patients' EHRS from The Necker pediatric Hospital in Paris, specialised in Rare diseases. (Garcelon et al., 2018) describes Dr Warehouse, which is a database used at the Necker Children's Hospital. The features and capabilities of this database enables efficient use of NLP techniques in a secure environment.

DATE	AGE	DURATION	FREQ	TIME
213	47	12	58	81

Table 1: Number of each entity in the Gold Standard

3.1. Annotation Guidelines

Defining temporal entities within the clinical context can be a difficult task, we build on previous works to do so, most notably the guidelines presented as part of the annotation of the MERLOT corpus (Campillos-Llanos et al., 2018). Broadly, temporal entities can be categorized into the following classes: Dates (including Date of Birth, Date of visit, Date of Report, Date of test, Date of consultation, Date of next scheduled visit), Time, Frequency, Duration and Age. In order to produce reliable and reproducible annotations of the available clinical data, we established the following guidelines, giving precise definitions of each categories as well as informative or borderlines cases that were found by comparing several annotators outputs:

DATE: All dates that are presented within the clinical text. This can be any date including the dates representing the history of the patient, date of birth, date of visit, date of creation of the record, date of identification of a diagnosis, date of commencement of medication etc.

Date mentions can be either complete or incomplete. We consider date mentions to be complete if they mention a year (optionally completed by a month and/or a day), while mentions lacking the mention of a year are considered incomplete (i.e. they require extra information to unambiguously determine the 'absolute' date they refer to). Irrespective of complete or incomplete mentions, these entities are annotated as DATE.

Examples:

- "Craniopharyngiome type decouvert sur des signes d'HTIC en Aout" → **Aout** annotated as DATE
- "Radiotherapie prevue debut Novembre" → **debut Novembre** annotated as DATE
- "Je propose un rendez-vous de consultation le 20 decembre" → **20 decembre** annotated as DATE
- "Dicte le: 02/02/2021" → **02/02/2021** annotated as DATE
- "Paris le 01/07/2000" → **01/07/2000** annotated as DATE

If the Date is written as a range with the year and/month attached to the second part, a fragment with the day,month and year to complete the DATE *Ex:* "Hospitalise(e) du 19 au 29/07/2023": fragment with 19/07/2023 annotated as DATE and another entity 29/07/2023 annotated as DATE (not as DURATION)

If the DATE includes days such as "Lundi 3 Mars 2011", the entire phrase is annotated as DATE, including the day

AGE: This refers to the age of the patient presented in the text, his/her parents or relations, age of a fetus. A fetus's age is usually represented in terms of "SA" or as "Age Gestationnel" *Ex1:* IMG à 33SA + 5jours pour immobilisme foetal, Caryotype normale → **33SA + 5jours** annotated as AGE. *Ex2:* Il a 36 ans → **36 ans** annotated as AGE

DURATION: This entity reference to a continuous duration of time. *Ex:* "depuis le 20/1/2001", "pendant 2 jours", "depuis plus de 25 ans" etc.

FREQUENCY: Any time related quantity repeated at regular intervals. *Ex:* "par jour", "par semaine", "par seconde" "/jour", "/hr" , "/le soir", "/le matin", "tout les matins" etc. FREQUENCY also includes visits to the clinic schedules at specific intervals or tests scheduled at/taken at specific intervals.

- KCL 10ml par jour → **par jour** annotated as FREQUENCY
- Heparine 70 mg dans 48 ml, vitesse 5ml/heure → **/heure** annotated as FREQUENCY

TIME: This entity refers to the any time relative to a date. (i.e) when the date is unclear, it is TIME. *Ex:* "4 semaines", "4 jours", "toujours", "ce moment", "ce jour", "matin", "midi", "soir" etc

- Any specific time to be marked as time. *Ex: "9:28"*
- A "rendez-vous" made after certain amount of time is to be annotated as Time, without a specific date mentioned. *Ex1:* Nouveau controle endoscopique dans 3 mois → 3 mois annotated as TIME. *Ex2:* Prochain RDV dans 1 semaine →*1 semaine* annotated as TIME
- Time indicated as J1, J2 ..etc indicate "Jour 1", "jour 2" etc. Thus these should be annotated as time, since they are relative to the date.

3.2. Annotation Process

For the purpose of testing our experiments, we annotate 50 clinical notes using the annotation guidelines as mention in section 3.1. Three annotators were asked to annotate the same set of clinical notes to be able to establish a gold standard. They were given the same set of the above mentioned annotation guidelines. The methods and models are tested and evaluated on these 50 notes.

A set of 150 EHRs has been annotated by one annotator using the above mentioned guidelines which can be used for training any language model.

4. Experimental Setup

There are indeed several tools that explore temporal entities in the French language. Even if these

tools and models are not particularly tailored for the clinical context, these can be used to identify basic dates and times within the text. In this paper, we perform experiments with 3 existing tools and models on our hospital local dataset. We then evaluate the results to determine how the tools and models perform on our internal dataset.

The experiments are performed using local installations of the tools and models, thus preserving the privacy of patient information.

EDS-NLP: (Wajsburt et al., 2022) is a NLP framework that aims at extracting information from French clinical notes. It is a collection of components or pipes, either rule-based functions or deep learning modules. EDS-NLP has a component (eds.date) for extracting dates in medical reports. In this paper, we apply EDS NLP's *date component* to detect temporal entities in our datatset. This method is able to identify the dates as an entity, however this method fails to differentiate between the temporal entities such as duration and frequency. We use the 50 clinical texts annotated by the 3 annotators to extract the temporal entities. The results are then used to be compared with the manual annotations.

CamemBERT-with-Dates: (Martin et al., 2020) CamemBERT is a state-of-the-art language model for French based on the RoBERTa architecture pretrained on the French subcorpus of the multilingual corpus OSCAR. CamemBERT-with-dates is an extension of french camembert-ner model with an additional tag for dates. This model was trained on an enriched version of wikiner-fr dataset. This model is able to identify the dates as an entity, however this model fails to differentiate between the temporal entities such as duration and frequency, as the model is not trained for these entities. For the first experiment, we extract the temporal entities from the 50 clinical texts annotated by the 3 annotators. The results are then compared with the manual annotations. For the second experiment, we fine-tune the CamemBERT-with-dates model using the 150 clinical texts that has been annotated by one annotator as stated in section 3.2. The fine-tuned model is then tested on the 50 clinical texts (annotated by the 3 annotators). The results from the fine-tuned model is then used to be compared with the manual annotations.

Large Language Model: In this work, we use the Vicuna model (Chiang et al., 2023) for testing the prompt based approach on the dataset. Vicuna is an open-source large Language Model (LLM) with 13 billion parameters. There are several versions of Vicuna available. For experimentation, we use Vicuna v1.5. This model is fine-tuned from Llama2 with supervised instruction fine-tuning and linear RoPE scaling. The training data is around 125K conversations collected from ShareGPT.com. These conversations are packed into sequences that contain 16K tokens each.

In this work, we setup a local version of the model that is used for experimentation, so as to preserve the privacy of the dataset. This model is prompted with three different kinds of prompts to identify the temporal entities.

We use prompt based methods to query the LLM for the purpose of identifying temporal entities. As mentioned in section 2.1, there have been several works on using various types of templates for prompting LLMs. In this work, we experiment with 3 different prompts to extract temporal entities using the Vicuna LLM. They are as follows:

- Posing a general question to the LLM (Vicuna) to identify the temporal entities (i.e What are the temporal entities in the text "..."?).

- Defining the temporal entities to the LLM before posing the question to the LLM. For example: We define all entities together such as "date: date written in any format. time: time of the day or any time without mention of date. age is the age of the patient or fetus. frequency: time related quantity repeated at regular intervals. Ex: “par jour”, “par semaine”, “par seconde” “/jour”, “/hr” , “/le soir”, “/le matin”, “tout les matins” etc Duration: a continuous duration of time. Ex: “depuis le 20/1/2001”, “pendant 2 jours”, “depuis plus de 25 ans” etc.” and then ask Vicuna to identify all temporal entities defined above

- Decomposing the prompt into several parts. In this part, we split the prompt into 5 different prompts (one for each entity). Each of the prompt has a definition of the entity with examples and a question asking the LLM to identify that particular entity. For example: "time is defined as any time of the day like "matin", "soir", "midi" or any time without mention of date like "ce jour", "ce moment", "aujourd'hui" or time indicated as number of says like "Jour 1", "Jour 2" etc or "J1", "J2" etc. Identify all the mentions of TIME entities in the following text: ..."

For the purpose of evaluation, a certain amount of post-processing is required as comparison to the gold standard annotation requires the outputs from the tools and models to have span (start and end indices) of the entities. As mentioned in (Ashok and Lipton, 2023), one of the limitations of prompting LLMs is the preservation of spans for the entities. As the testing data is small (50 EHR), the post processing of matching the entity with the span was done manually.

5. Results and Discussion

The results from our experiments (as mentioned in section 4) are presented in Tables 2 and 3. Table 2 gives the F1 scores of the entities, while table 3 provides a token level evaluation that counts partial token matches of multi-tokens terms as positives.

EDS-NLP and CamemBERT: Both EDS-NLP and CamemBERT, do not differentiate dates with frequency, duration, time or age. That is, every temporal entity is labelled as DATE. For example: In the text: *"Depuis Juin 2008, la creatininemie augmente"*, the entity *"Juin 2008"*, is marked as DATE by both EDS-NLP and CamemBERT, while according the Gold Standard annotations they should be marked as DURATION. Phrases such as *"Il y a 5 mois"*, *"par semaine"* are also marked as DATE by both EDS-NLP and CamemBERT, while according the Gold Standard annotations they should be marked as TIME and FREQUENCY respectively. Thus, to have a fair evaluation of these tools, we mark all temporal entities as DATE in the Gold Standard as well (i.e), all the other entities (AGE, DURATION, FREQUENCY and TIME) are renamed as DATE for the purpose of evaluating EDS-NLP and CamemBERT with our test dataset.

It has to be noted that EDS-NLP has been developed for French Clinical texts, while CamemBERT-with-dates has been trained for the French language but not particularly for clinical texts.

CamemBERT Finetuned: For the purpose of fine-tuning a language model, we use the 150 documents annotated by one annotator. All temporal entities in these 150 texts are DATE, (i.e), all the other entities (AGE, DURATION, FREQUENCY and TIME) are renamed as DATE. This will help to fine-tune the CamemBERT-with-Dates model more efficiently as DATE is already a supported entity by the model. The fine-tuned model (dubbed CamemBERT-ft in the results tables) is then tested on the 50 EHRs (annotated by 3 annotators). As seen from Tables 2 and 3, there is definitely improvements in the results when a fine-tuned model is used. However, table 2 shows very low F1 score (0.047) for the DATE entity. This is because of variations in the tokenization used by the model. For example: the text "16.04.1968" is marked as a whole as DATE, however, the model splits the tokens into three different tokens as "16","04","1968" and each of them are labelled as DATE. This is evident from Table 3 where the token level evaluation is presented. This shows a F1 score of 0.758 for the fine-tuned CamemBERT-with-Dates model. It is to be noted that only 150 documents were used to fine-tune the model. The number of Epochs used for fine tuning is 25. Given the improvement in result of a fine-tuned model when compared to the raw model, even while using such a small amount of data for fine-tuning, it can be envisioned that using a bigger amount of data for fine-tuning could result in a more competitive model.

LLM - Vicuna: We have used three different prompts with Vicuna to extract the temporal entities in the text. It has to be noted that Vicuna is not particularly trained for the French Language, nor particularly for clinical texts but positive results on early experiments prompted us to continue testing it.

The first prompt, being a very general prompt demanding the LLM to identify all temporal entities, while performing well for the identification of DATE, AGE and Duration entities, does not perform well for FREQUENCY and TIME (Tables 2 and 3). It has a poor performance specifically for the FREQUENCY entity as the LLM is not able to understand our definition of FREQUENCY. For example: In the text, *KCL 10 ml par jour*, the entity "par jour" is not marked at all, while it has to be marked as frequency. this is because a general question to the LLM demanding the identification of temporal entities is not well understood by the model.

The second prompt, where the definitions of all the entities are given to the LLM before posing a question asking for the identification of the defined entities, the results (Tables 2 and 3) are better. The results for the entity FREQUENCY has improved a lot as the model is now able to understand each entity. The definition of the FREQUENCY and DURATION also includes examples for each entity, thus helping Vicuna to learn from example. For the TIME entity, there seem to be several TIME entities misclassified as DATE like "ce jour", "ce semaine" etc.

The third prompt, where a prompt is generated for each entity with examples before posing questions to the LLM, performs the best. In particular, the TIME entity improves in performance drastically. Not only does the model learn from examples but by giving individual prompts for each entities, the confusion between DATE and TIME is avoided. Thus entities like "ce jour", "ce matin", "aujourd'hui" etc which are classified as DATE while using the second prompt is correctly classified while using the third prompt.

6. Conclusion

In this paper, we performed an external validation for extraction of temporal entities using the NER tool (EDS-NLP), Language model (CamemBERT-with-Dates) and Large Language Model (Vicuna). There are several other LLM, such as described in (Touvron et al., 2023) with Llama models ranging from 7B to 70B parameters. There are also newer models such as Mistral-7B-v0.1 (Jiang et al., 2023), which is a small (7-billion parameters) but powerful

Method	DATE	AGE	DURATION	FREQ.	TIME
EDS-NLP	0.560	NA	NA	NA	NA
CamemBERT	0.024	NA	NA	NA	NA
CamemBERT-ft	0.047	NA	NA	NA	NA
Vicuna Prompt1	0.842	0.84	0.857	0.067	0.527
Vicuna Prompt2	0.853	0.854	0.957	0.840	0.615
Vicuna Prompt3	0.862	0.860	0.960	0.848	0.860

Table 2: F1 scores for entities

Method	DATE	AGE	DURATION	FREQ.	TIME
EDS-NLP	0.779	NA	NA	NA	NA
CamemBERT	0.543	NA	NA	NA	NA
CamemBERT-ft	0.758	NA	NA	NA	NA
Vicuna Prompt1	0.830	0.861	0.822	0.097	0.577
Vicuna Prompt2	0.867	0.840	0.938	0.852	0.667
Vicuna Prompt3	0.912	0.881	0.938	0.867	0.90

Table 3: Token wise F1 evaluation

language model adaptable to several down-stream tasks and shown to perform better than Llama 2 13B on all tested benchmarks. (Jiang et al., 2023). We made a choice to use Vicuna for our experiments as we had the computing power and memory to store a Vicuna model (13 billion parameters), and it displayed good performances (Zheng et al.) that our early experiments confirmed. As the set of available LLMs changes rapidly we intend to test further models such as Mistral-7B-v0.1, keeping in mind performance to cost ratio. Indeed, deploying a Language Model (Large or small) locally in a clinic can be difficult as it requires a significantly higher amount of storage space and computing power than smaller deep learning models, proportional to the increase in the number of parameters (assuming comparable implementations - other factors coming into play such as quantization, method for underlying attention, etc...).

Fine-tuning and storing any Language Model locally is expensive, thus the efficiency of the model is an important factor to be considered. We have selected other tools and models to perform a comparison study between tools tailored for clinical texts, models trained for French (not for clinical texts in particular) and an entirely different model without any context to the french language or for clinical text. This gives us a variety of options to consider before deployment.

From Tables 2 and 3, it can be seen that prompting a LLM with question for NER performs better than EDS-NLP and CamemBERT-with-Dates, even-though Vicuna is not specifically trained for French clinical texts. It is important to note that the dataset used for testing is small (50 EHR). This is a small sample size to generalize the results globally, however locally (within the clinic) it is a good amount to be able to understand the requirements for good performance.

Language Models such as CamemBERT, though trained on fewer parameters, are easier to fine-tune for downstream tasks. While LLMs such as Vicuna, can have a good performance without any fine-tuning which can make them very useful in a context where data is not readily available and costly to produce. Thus choosing a model for extraction of information depends greatly on the local requirements.

The tools and models have been tested for temporal entities in EHRs of patients with rare diseases, however, this could be easily extended to other entities in any type of clinical text. Thus this presents a feasible method for analysing a patient's history, prediction of next steps and the evaluation of decisions taken.

7. Acknowledgement

This work was supported by state funding by The French National Research Agency (ANR) under the C'IL-LICO project (ANR-17-RHUS-0002) and as part of the "Investissements d'avenir" program (ANR-19-P3IA-0001) (PRAIRIE 3IA Institute). The authors acknowledge URC-CIC Paris Centre for the implementation of the study.

8. Bibliographical References

Ghada Alfattni, Niels Peek, and Goran Nenadic. 2020. Extraction of temporal relations from clinical free text: A systematic review of current ap-

proaches. *Journal of Biomedical Informatics*, 108:103488.

Dhananjay Ashok and Zachary C. Lipton. 2023. Promptner: Prompting for named entity recognition.

Priyankar Bose, Sriram Srinivasan, William C. Sleeman, Jatinder Palta, Rishabh Kapoor, and Preetam Ghosh. 2021. A survey on recent named entity recognition and relationship extraction techniques on clinical texts. *Applied Sciences*, 11(18).

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Leonardo Campillos-Llanos, Louise Deléger, Cyril Grouin, Thierry Hamon, Anne-Laure Ligozat, and Aurélie Névéol. 2018. A french clinical corpus with comprehensive semantic annotations: development of the medical entity and relation limsi annotated text corpus (merlot). *Language Resources and Evaluation*, 52.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-based named entity recognition using BART. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1835–1845, Online. Association for Computational Linguistics.

Carole Faviez, Marc Vincent, Nicolas Garcelon, Caroline Michot, Genevieve Baujat, Valerie Cormier-Daire, Sophie Saunier, Xiaoyi Chen, and Anita Burgun. 2022. Enriching umls-based phenotyping of rare diseases using deep-learning: Evaluation on jeune syndrome. In *Challenges of Trustable AI and Added-Value on Health*, pages 844–848. IOS Press.

Nicolas Garcelon, Antoine Neuraz, Rémi Salomon, Hassan Faour, Vincent Benoit, Arthur Delapalme, Arnold Munnich, Anita Burgun, and Bastien Rance. 2018. A clinician friendly data warehouse oriented toward narrative reports: Dr. warehouse. *Journal of Biomedical Informatics*, 80:52–63.

Yucheng Huang, Kai He, Yige Wang, Xianli Zhang, Tieliang Gong, Rui Mao, and Chen Li. 2022. Copner: Contrastive learning with prompt guiding for few-shot named entity recognition. In *Proceedings of the 29th International conference on computational linguistics*, pages 2515–2527.

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.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Andy T. Liu, Wei Xiao, Henghui Zhu, Dejiao Zhang, Shang-Wen Li, and Andrew Arnold. 2022. Qaner: Prompting question answering models for few-shot named entity recognition.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2023. Gpt understands, too. *AI Open*.

Zengjian Liu, Ming Yang, Xiaolong Wang, Qingcai Chen, Buzhou Tang, Zhe Wang, and Hua Xu. 2017. Entity recognition from clinical texts via recurrent neural network. *BMC medical informatics and decision making*, 17:53–61.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying lms with mixtures of soft prompts. *arXiv preprint arXiv:2104.06599*.

Yongliang Shen, Zeqi Tan, Shuhui Wu, Wenqi Zhang, Rongsheng Zhang, Yadong Xi, Weiming Lu, and Yueting Zhuang. 2023. Promptner: Prompt locating and typing for named entity recognition. *arXiv preprint arXiv:2305.17104*.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Julien Tourille. 2018. *Extracting Clinical Event Timelines : Temporal Information Extraction and Coreference Resolution in Electronic Health Records*. Ph.D. thesis.

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

Marc Vincent, Maxime Douillet, Ivan Lerner, Antoine Neuraz, Anita Burgun, and Nicolas Garcelon. 2022. Using deep learning to improve phenotyping from clinical reports. *Stud Health Technol Inform*, 290:282–6.

Perceval Wajsburt, Thomas Petit-Jean, Basile Dura, Ariel Cohen, Charline Jean, and Romain Bey. 2022. Eds-nlp: efficient information extraction from french clinical notes.

Website. 2023. Temporal entity definition.

Feiyang Ye, Liang Huang, Senjie Liang, and KaiKai Chi. 2023. Decomposed two-stage prompt learning for few-shot named entity recognition. *Information*, 14(5).

Thakur A Zhu T Clifton D Shah NH Youssef A, Pencina M. 2023. External validation of ai models in health should be replaced with recurring local validation.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. corr, abs/2306.05685, 2023. doi: 10.48550. *arXiv preprint arXiv.2306.05685*.