Improving Few-shot Prompting using Cluster-based Sample Retrieval for Medical NER in Clinical Text

Meethu Mohan C [1], Sneha Shaji Punnan [2], Jeena Kleenankandy [3]
[1,3]Department of Computer Science , Cochin University of Science and Technology, Kerala, India
[2]School of Data Analystics, Mahatma Gandhi University, Kerala, India
[1]cmmeethu@gmail.com, [2]snehapunnan@gmail.com, [3]jeenakk@cusat.ac.in

Abstract

Named Entity Recognition (NER) in the medical domain is crucial for extracting essential information from clinical text. Large Language Models (LLMs) have demonstrated remarkable capabilities in this task, but their performance is highly dependent on the quality of the prompts. Few-shot prompting or prompt-by-example, where the input query to LLM is augmented with one or more sample outputs, is a well-known technique in guiding the LLMs to the expected result. The quality of the sample in the prompt plays an important role in this task. This paper proposes to improve the performance of few-shot prompting for medical NER on clinical text using a cluster-based strategy for sample selection. We employ the concepts from Retrieval Augmented Generation (RAG) and K-means clustering to identify the most similar annotated examples for any given input text. Using these contextually relevant yet divergent training samples as examples, we guide the LLM toward extracting more accurate medical entities. Our experiments using the llama-2 model show that this approach significantly outperforms zero-shot prompting and random sampled few-shot prompting in two data sets chosen for this study, demonstrating the efficacy of cluster-based retrieval in improving few-shot prompting for medical NER tasks.

1 Introduction

Large Language Models (LLM) are deep neural networks trained heavily on massive datasets and are very effective for a wide variety of generative tasks including machine translation (Zhang et al., 2023a), question-answering (Li et al., 2024b), and human-computer conversations (Vemprala et al., 2023). The potential of LLMs in various extraction tasks like Named Entity Recognition (NER) has also been widely explored in the literature (Villena et al., 2024). LLMs are initially pre-trained on language modeling tasks and later fine-tuned to adapt to NER tasks. Their success depends on the quality and quantity of the annotated datasets used for this supervised fine-tuning. Instruction-tuned LLMs used for NER tasks are trained to take the input query as a component of their prompt and respond with the NER annotated text using techniques like slot-filling. Engineering the right prompt for such models is still an active research area. LLM prompts need accurate instructions and appropriate context information to guide them to the correct output for any specific tasks. Few-shot prompting technique (Zhang et al., 2023b) that provides the LLM with sample outputs in their context has been shown to be effective in improving LLM responses.

Medical Named Entity Recognition (NER) is the process of identifying and classifying medical entities from unstructured text. Recognizing medical entities is essential for applications like electronic health record (EHR) analysis, clinical decision support, and biomedical research (Monajatipoor et al., 2024). Medical text contains a wide range of medical terms, including diseases, medications, symptoms, and treatments, which can be expressed in diverse ways. Many studies show that the performance of domain-agnostic LLMs for medical NER is significantly lower than that of models trained specifically for the medical domain. Even domain-specific models require fine-tuning and retraining to adapt to new datasets, and their performance gain is only marginal if not none to traditional extraction techniques. Few-shot prompting in LLMs for medical NER offers opportunities to adapt these models to unseen data with mini-

mal sample data and training effort. Considering the context length limit, identifying the most relevant few-shot samples for the given input text is a crucial step. This challenge has not been much explored in the medical domain.

The main objective of this work is to improve few-shot prompting in pre-trained LLMs for medical NER by proposing a novel approach in sample selection. Our approach makes use of vector-based similarity to select the most appropriate k-shot demonstrations required for model prompting. To ensure relevance and divergence in the retrieved samples, we select samples belonging to k-different clusters whose centroids match the most with the given input query. We experimented with the open-source llama2 model for different numbers of clusters and with multiple values for k, on two clinical datasets. Our findings are reported in the result section which are promising and call for further research in this area.

2 Related Work

In recent years, large language models (LLMs) have demonstrated immense potential in few-shot and zero-shot clinical information extraction. Agrawal et al., 2022 explored the use of InstructGPT, showing that LLMs can effectively reduce dependency on large annotated datasets and perform well on re-annotated datasets like CASI. This is particularly relevant in healthcare, where annotated clinical datasets are limited. Similarly, techniques using pre-trained language models combined with domain-specific dictionaries and BiLSTM-CRF architectures have enhanced NER in unlabelled medical texts, significantly improving entity recognition rates (Sinha et al., 2024).

The integration of LLMs such as BERT in methods like NESSMa, which uses surrounding sequence matching, has further improved contextual understanding in clinical text mining (Landolsi et al., 2022). However, token-level NER, particularly in rare diseases, remains a challenge. Lu et al., 2024 identified limitations in token-level NER for LLMs like Meditron and UniversalNER but found that Llama2-MedTuned-7b shows promising results, highlighting the potential of open-source models in this field.

Moreover, advanced systems like BERN2 combine traditional rule-based methods with LLMs, enhancing both recognition and normalization tasks (Sung et al., 2022). The use of Retrieval-Augmented Generation (RAG) models has also significantly improved knowledge-intensive NLP tasks, offering better factual accuracy and interpretability (Pichai, 2023). BioMistral, an open-source model for medical domains, further showcases performance improvements through fine-tuning and model merging techniques (Labrak et al., 2024).

The Retrieving and Chain-of-Thought (RT) framework (Li et al., 2024a) enhances few-shot biomedical named entity recognition (NER) by combining retrieval and reasoning. The retrieval module selects relevant examples, providing context for better sentence understanding, while the Chain-of-Thought module applies systematic reasoning to improve entity prediction. Together, they significantly boost model accuracy. A comparative study on BC5CDR and NCBI datasets demonstrated superior performance, while error analysis revealed challenges with long and out-of-vocabulary entities. The RT framework offers a promising approach for advancing NER in biomedical applications.

A prompting method for few-shot Named Entity Recognition (NER) using Large Language Models (LLMs) features a standardized prompt structure with task definition, few-shot demonstrations, and a constrained output format to ensure accurate and structured responses (Cheng et al., 2024). The method also includes error mitigation prompts to correct recognition mistakes and optimizes demonstration selection for better performance. Ablation studies assess the impact of each prompt component. Overall, this approach improves NER accuracy by leveraging LLMs while addressing few-shot learning challenges across datasets.

The Clustered Retrieved Augmented Generation (CRAG) approach (Li et al., 2024a) enhances question-answering systems using Large Language Models (LLMs) by optimizing data processing. It begins with collecting product reviews and generating embeddings using a sentence transformer, followed by clustering these embeddings via K-means. Summarization of each cluster is performed with the Mistral 7B model to reduce redundancy. The cluster summaries are then aggregated into a comprehensive knowledge base. This method reduces token usage while maintaining response quality, improving the system's efficiency.

3 Dataset

Two clinical NER datasets - MTSamples and VAERS used in this paper are sourced from (Hu et al., 2024). MTSamples dataset consists of discharge summaries from MTSamples, which were annotated based on the guidelines from the 2010 i2b2 challenge, focusing on extracting Medical Problems, Treatments, and Tests (Uzuner et al., 2011). VAERS dataset is a collection of publicly accessible safety reports from vaccine adverse event reporting system (VAERS), focused on extracting events related to nervous system disorders (Du et al., 2021).

The MTSamples dataset is completely synthetic, meaning it was generated artificially and does not include any real patient information. The VAERS dataset comes from publicly available post-market safety reports that are anonymized and do not contain any personal data. As a result, no sensitive information was shared, ensuring that this study poses no privacy concerns.

The two datasets were split into training, validation, and test subsets. The training subset is used as the external knowledge base to provide the few-shot samples, whereas the test subset is for evaluating the model's performance and for comparative analysis. Table 1 and Table 2 represent a descriptive summary of the entities found with these datasets.

Entities	**Train**	**Valid**	**Test**	**Total**
Investigation	148	29	59	236
Nervous adverse event	406	83	162	651
Other adverse event	301	62	167	530
Procedure	338	57	126	521

Table 1: Statistics of the VAERS dataset (Hu et al., 2024)

Entities	Train	Valid	Test	Total
Medical Problem	538	203	199	940
Treatment	149	43	35	227
Test	120	39	50	209

Table 2: Statistics of the MTSamples dataset (Hu et al., 2024)

4 Methodology

The architecture for the proposed Cluster-based Retrieval Augmented Extraction (CRAE) approach is shown in Figure 1. For our experiments, we use the training data split of the respective dataset as the external knowledge store, although it can be any named entity annotated data. These documents are first split into manageable chunks and are embedded using a pre-trained Hugging Face model. The embeddings are stored in FAISS, a vector database designed for efficient retrieval. The nearest k-shots retrieved from this vector store are not only very similar to the input query but also each other. Our experiment showed that augmenting more than one such sample in the prompt did not improve the performance of the model but rather had an inverse effect. To address this issue, we clustered these embeddings into n clusters out of which k clusters were used in prompting. For each input query in the test dataset, we retrieve the most similar example from each cluster by performing a similarity search between the cluster centroids and the input query embedding. These selected examples were then used as context for the query. Both the query and the context are passed to a large language model (LLM) through a pre-defined prompt template to generate informed responses.

The task specific prompt template for MTSamples is shown in Table 3.

Table 4 shows the prompt template used with VAERS dataset.

4.1 Data preprocessing

The BIO (Begin-Inside-Other) format, which is commonly used for classification tasks like Named Entity Recognition(NER) in traditional classification model is not very effective for generative LLM like ChatGPT or llamaChat. The goal of preprocessing is to convert the BIO tagged data into an entity list format devoid of the B and I prefix. Instead of a tagging problem, we model NER as an entity extraction problem where the input is an unstructured text and output is a list of lists, one for each entity type. Any word not classified under the predefined entity types is assigned to the "Other" category. We use a carefully hand-crafted prompt template with entity markup guide, and entity definitions along with the retrieved few-shot examples, as shown in Table 3 and 4. The model is prompted to present a separate list of entities for each type except the "Other" type. The response from the

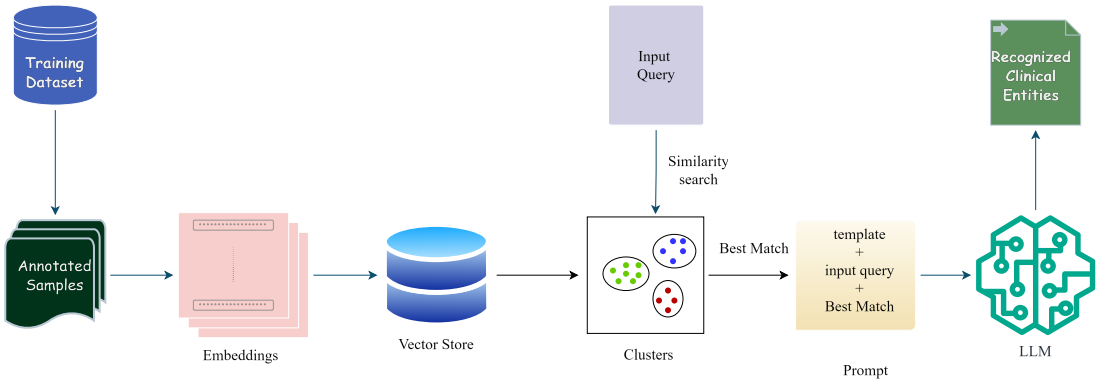

Figure 1: Architecture of Cluster based Retrieval Augmented Extraction (CRAE)

You are a medical named entity recognition model marking up specific entities related to healthcare.

Entity Markup Guide
Use "problem" to denote a Medical Problem.
Use "treatment" to denote a Treatment.
Use "test" to denote a Test.

Entity Definitions
Definition for Medical Problem, Treatment and Test

Example: {context}

Task
Based on the Example, extract specific entities related to healthcare from the input text . Entities to be identified are of the following categories.
Categories:
-problem
-treatment
-test

NOTE:
1) Output should contain entities which are explicitly mentioned in the input text.
2) Entities should be extracted by strictly following "Entity Markup Guide", "Entity Definitions" and "Annotation Guidelines".
Please provide output in the required format.
Input Text: "text"
Entities:
"""

Table 3: Prompt template for MTSample

LLM is converted to a list of lists for evaluation.

4.2 Models

For our experiments, we utilized the LLaMa-2-7b-chat-hf model. Meta developed and publicly released the Llama 2 family of large language models ranging from 7 billion to 70 billion parameters (Touvron et al., 2023). Llama-2-chat-hf are fine tuned and are optimized for dialogue use cases. This model is used to compare the effectiveness of zero-shot, one-shot, and CRAE few-shot prompting.

4.3 Evaluation

The performance of the models is evaluated based on Accuracy (A), Precision (P), Recall (R), and F1-score at the token level, using both Strict and Relaxed Match. In Token Level Strict Match the boundaries of the entity must perfectly align with the boundaries of the tokens in the text. Relaxed Match allows partial overlap i.e. overlaps between the entity and other tokens are considered partial matches. Consider the example of LLM-generated output shown below.

```
Input Text:
The patient had a persistent cough.
Entities:
- problem: ['cough']
- treatment: []
- test: []
```

Suppose, the ground truth has "a persistent cough" marked as a problem entity, the Strict Match algorithm would consider this output as a mismatch. In Relaxed Match evaluation, the accuracy would be considered as $1/3$ as 1 token out of the 3 have been correctly extracted.

You are a medical named entity recognition model marking up specific entities related to healthcare.

Entity Markup Guide
Use "investigation" to denote an investigation.
Use "nervous_AE" to denote a nervous adverse event.
Use "other_AE" to denote an other adverse event.
Use "procedure" to denote a procedure.

Entity Definitions
Definition for Investigation, Nervous adverse event, Other adverse event and Procedure

Example: {context}

Task
Based on the Example, extract specific entities related to healthcare from the input text . Entities to be identified are of the following categories.
Categories:
-investigation
-nervous_AE
-other_AE
-procedure

NOTE:
1) Output should contain entities which are explicitly mentioned in the input text.
2) Entities should be extracted by strictly following "Entity Markup Guide", "Entity Definitions" and "Annotation Guidelines".
Please provide output in the required format.
Input Text: "text"
Entities:
"""

Table 4: Prompt template for VAERS dataset

5 Result and Discussion

We have experimented the models for different values of n (number of clusters created) and k (number of clusters selected) and those results are discussed. Figure 2 shows the impact of the number of few-shot demonstrations (k) on the performance for MT-Samples with 126 clusters, with k clusters chosen based on the nearest centroids to the input embedding. The evaluation metrics vary as the number of selected clusters changes. The best performance can be seen at k=6. Figure 3 shows the evaluation

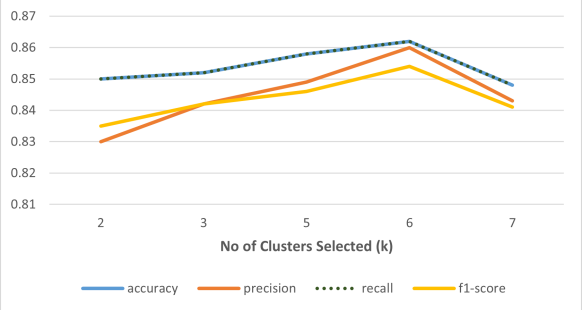

Figure 2: Evaluation results on MTSamples having total of 126 clusters and selecting different number of clusters(k) for best match :Relaxed Match

results on MTSamples using a 2-shot prompt. The results show the performance for different numbers of clusters created, with two clusters being selected based on the centroids closest to the input query embedding. The best performance can be seen at a cluster count of 126.

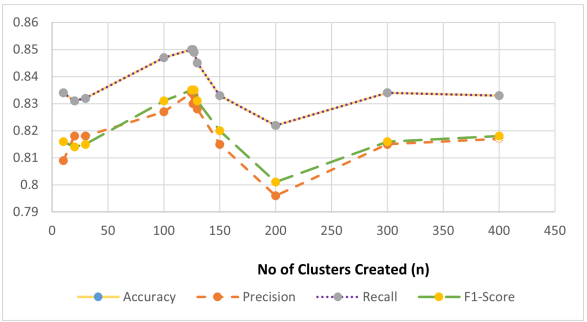

Figure 3: Evaluation results on MTSamples for 2-shot prompt with different clusters :Relaxed Match

The Table 5 shows the performance of zero-shot and CRAE-k shots using Llama-2-7b-chat-hf model, evaluated under both Token Level-Strict and Relaxed matching criteria on MTSamples. Overall, the CRAE-k model achieved the highest F1-score 0.831 and 0.854 for both Token Level-Strict and Relaxed matching respectively, indicating the best overall performance at $k = 6$.

	Token Level -Strict Match				Relaxed Match			
	A	**P**	**R**	**F1**	**A**	**P**	**R**	**F1**
zero-shot	0.791	0.754	0.791	0.758	0.805	0.777	0.805	0.776
one-shot	0.798	0.77	0.798	0.777	0.815	0.795	0.815	0.798
CRAE-1	0.84	0.828	0.84	0.826	0.8592	0.847	0.859	0.847
CRAE-k	0.84	0.833	0.84	0.831	0.862	0.86	0.862	0.854

Table 5: Evaluation results on MTSamples for zero-shot, one-shot, CRAE-1 and CRAE -k, where k = 6

	Token Level -Strict Match				Relaxed Match			
	A	**P**	**R**	**F1**	**A**	**P**	**R**	**F1**
zero-shot	0.728	0.679	0.728	0.636	0.728	0.678	0.728	0.637
one-shot	0.75	0.688	0.75	0.703	0.757	0.7	0.757	0.711
CRAE-1	0.75	0.73	0.75	0.703	0.763	0.735	0.763	0.709
CRAE-k	0.747	0.716	0.747	0.677	0.75	0.72	0.75	0.68

Table 6: Evaluation results on VAERS dataset for zero-shot, one-shot , CRAE-1 and CRAE -k where k= 15

The Table 6 shows the performance on VAERS dataset. Overall, the CRAE-1 achieved the highest values for evaluation metrices like accuracy 0.758 and 0.763 for both Token Level-Strict and Relaxed matching respectively, indicating the best overall performance.

6 Conclusion and Futureworks

This paper presents a novel approach to enhance few-shot prompting for Named Entity Recognition (NER) in the medical domain. By leveraging K-means clustering and embedding similarity, we propose a cluster-based sample selection strategy to identify contextually relevant yet diverse training examples in few-shot prompting. Our experiments demonstrate that this approach significantly outperforms zero-shot prompting and random sampled few-shot prompting on two medical datasets. However, further research is needed to explore this approach using more open-source domain-agnostic models. A thorough qualitative analysis of the results could give more insights to improve the results. Our study was limited by the availability of computing power to run the model. The usage of more advanced embedding models or fine-tuning the current models can make the similarity search more accurate. This would help retrieve better and more relevant context. Exploring the use of hybrid search methods that combine semantic and keyword-based approaches might further enhance retrieval performance. This work could potentially lead to even more robust and generalizable solutions for NER few-shot prompting across various domains.

7 Limitations

There are several limitations that affect the overall performance and scalability of the approach. The availability of models can restrict our approach, as not all models are accessible for research purposes. Additionally, some of the most effective models are expensive, which can pose financial constraints for researchers and institutions. Processing time is another concern, as more complex models may require significant computational resources and time to generate results .The availability of annotated datasets is very important for training and evaluating models, but such datasets are often limited or difficult to obtain, especially in specialized domains. These factors collectively impact the scalability and applicability of our findings.

Ethics Statement

This work adheres to the ACL Ethics Policy and considers the broader impacts of our research. We acknowledge that language models, including those utilized in this study, can inadvertently perpetuate biases present in training data, leading to potential ethical concerns in applications. Our approach emphasizes transparency, aiming to mitigate these biases by employing rigorous evaluation techniques and promoting fair usage. Additionally, we recognize the importance of data privacy and confidentiality; all datasets used in our research comply with relevant regulations and ethical standards. We encourage further discussions on the implications

of our findings and their responsible application in real-world scenarios. Through this work, we aim to contribute positively to the field while fostering ethical research practices.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qi Cheng, Liqiong Chen, Zhixing Hu, Juan Tang, Qiang Xu, and Binbin Ning. 2024. A novel prompting method for few-shot ner via llms. *Natural Language Processing Journal*, 8:100099.

Jingcheng Du, Yang Xiang, Madhuri Sankaranarayanapillai, Meng Zhang, Jingqi Wang, Yuqi Si, Huy Pham, Wang Qi, Yong Chen, and Cui Tao. 2021. Extracting postmarketing adverse events from safety reports in the vaccine adverse event reporting system (vaers) using deep learning. *Journal of the American Medical Informatics Association*, 28.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, Kirk Roberts, and Hua Xu. 2024. Improving large language models for clinical named entity recognition via prompt engineering.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains.

Mohamed Yassine Landolsi, Lotfi Ben Romdhane, and Lobna Hlaoua. 2022. Medical named entity recognition using surrounding sequences matching. *Procedia Computer Science*, 207:674–683. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 26th International Conference KES2022.

M. Li, H. Zhou, H. Yang, and R. Zhang. 2024a. Rt: a retrieving and chain-of-thought framework for few-shot medical named entity recognition. *Journal of the American Medical Informatics Association*, 31(9):1929–1938.

Zhenyu Li, Sunqi Fan, Yu Gu, Xiuxing Li, Zhichao Duan, Bowen Dong, Ning Liu, and Jianyong Wang. 2024b. Flexkbqa: A flexible llm-powered framework for few-shot knowledge base question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18608–18616.

Qiuhao Lu, Rui Li, Andrew Wen, Jinlian Wang, Liwei Wang, and Hongfang Liu. 2024. Large language models struggle in token-level clinical named entity recognition.

Masoud Monajatipoor, Jiaxin Yang, Joel Stremmel, Melika Emami, Fazlolah Mohaghegh, Mozhdeh Rouhsedaghat, and Kai-Wei Chang. 2024. Llms in biomedicine: A study on clinical named entity recognition.

Kieran Pichai. 2023. A retrieval-augmented generation based large language model benchmarked on a novel dataset. *Journal of Student Research*, 12(4).

Shweta Sinha, Tushar Agarwal, and Pratiyush Pandey. 2024. Artificial intelligence in healthcare: medical named entity recognition based audio prescription generator. In *Proceedings of the 5th International Conference on Information Management & Machine Intelligence*, ICIMMI '23, New York, NY, USA. Association for Computing Machinery.

Mujeen Sung, Minbyul Jeong, Yonghwa Choi, Donghyeon Kim, Jinhyuk Lee, and Jaewoo Kang. 2022. Bern2: an advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics*, 38(20):4837–4839.

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. 2023. Llama 2: Open foundation and fine-tuned chat models.

Ozlem Uzuner, Brett South, Shuying Shen, and Scott DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 18:552–6.

Sai Vemprala, Rogerio Bonatti, Arthur Bucker, and Ashish Kapoor. 2023. Chatgpt for robotics: Design principles and model abilities.

Fabián Villena, Luis Miranda, and Claudio Aracena. 2024. llmner: (zero|few)-shot named entity recognition, exploiting the power of large language models.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023b. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481, Singapore. Association for Computational Linguistics.