German Medical Data Sciences 2024
R. Röhrig et al. (Eds.)
© 2024 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI240861

Zero-Shot LLMs for Named Entity Recognition: Targeting Cardiac Function Indicators in German Clinical Texts

Lucas PLAGWITZ[a,b,1], Philipp NEUHAUS[a], Kemal YILDIRIM[a], Noah LOSCH[a], Julian VARGHESE[a,2], and Antonius BÜSCHER[a,b,c,2]
[a] *Institute of Medical Informatics, University of Münster, Münster, Germany*
[b] *Interdisciplinary Center for Clinical Research (IZKF), University of Münster, Münster, Germany*
[c] *Department for Cardiology II-Electrophysiology, University Hospital Münster, Münster, Germany*
ORCiD-ID: Lucas Plagwitz https://orcid.org/0000-0001-7626-8853

Abstract. Introduction Large Language Models (LLMs) like ChatGPT have become increasingly prevalent. In medicine, many potential areas arise where LLMs may offer added value. Our research focuses on the use of open-source LLM alternatives like Llama 3, Gemma, Mistral, and Mixtral to extract medical parameters from German clinical texts. We concentrate on German due to an observed gap in research for non-English tasks. **Objective** To evaluate the effectiveness of open-source LLMs in extracting medical parameters from German clinical texts, specially focusing on cardiovascular function indicators from cardiac MRI reports. **Methods** We extracted 14 cardiovascular function indicators, including left and right ventricular ejection fraction (LV-EF and RV-EF), from 497 variously formulated cardiac magnetic resonance imaging (MRI) reports. Our systematic analysis involved assessing the performance of Llama 3, Gemma, Mistral, and Mixtral models in terms of right annotation and named entity recognition (NER) accuracy. **Results** The analysis confirms strong performance with up to 95.4% right annotation and 99.8% NER accuracy across different architectures, despite the fact that these models were not explicitly fine-tuned for data extraction and the German language. **Conclusion** The results strongly recommend using open-source LLMs for extracting medical parameters from clinical texts, including those in German, due to their high accuracy and effectiveness even without specific fine-tuning.

Keywords. Large Language Models, Named Entity Recognition, German Clinical Texts, Open-Source Models, Llama 3, Mistral, Gemma

1. Introduction

Since the release of Large Language Models (LLMs) for the public in the form of ChatGPT in 2022, a multitude of application fields for LLMs has emerged. Particularly in medicine, potential use cases seem limitless, as demonstrated by their ability to respond to free-text queries, improve clinical efficiency and increase diagnostic accuracy, despite recognized limitations and the need for regulation [1, 2]. However,

[1] Corresponding Author, Lucas Plagwitz. Address: Institute of Medical Informatics, University of Münster, Albert-Schweitzer-Campus 1, Building A11, 48149 Münster, Germany; E-Mail: lucas.plagwitz@uni-muenster.de

[2] These authors contributed equally to this work.

numerous applications and studies have faced constraints, as processing of personal clinical data through live systems like ChatGPT is not feasible in most countries due to data privacy concerns. The development of locally hostable alternatives such as Llama and Mistral has made it possible to conduct research on closed clinical data.

A significant issue in medical documentation is the lack of standardized and structured collection of data. Although structured data is crucial for research and general data evaluation, there is frequently insufficient flexibility for special cases in data collection. For instance, in many areas, it is still standard practice to evaluate medical documents in human-written format. Methods for converting this data into a structured representation can be found under the term named entity recognition (NER). In recent years, numerous methods and algorithms have been developed to apply NER for various medical databases, ranging from regex-based systems to complex NLP-based neural networks [3]. This issue has gained prominence mainly through a publicly announced challenge known as i2b2 and subsequently i2c2 [4]. Challenges like these are based on different English data sets.

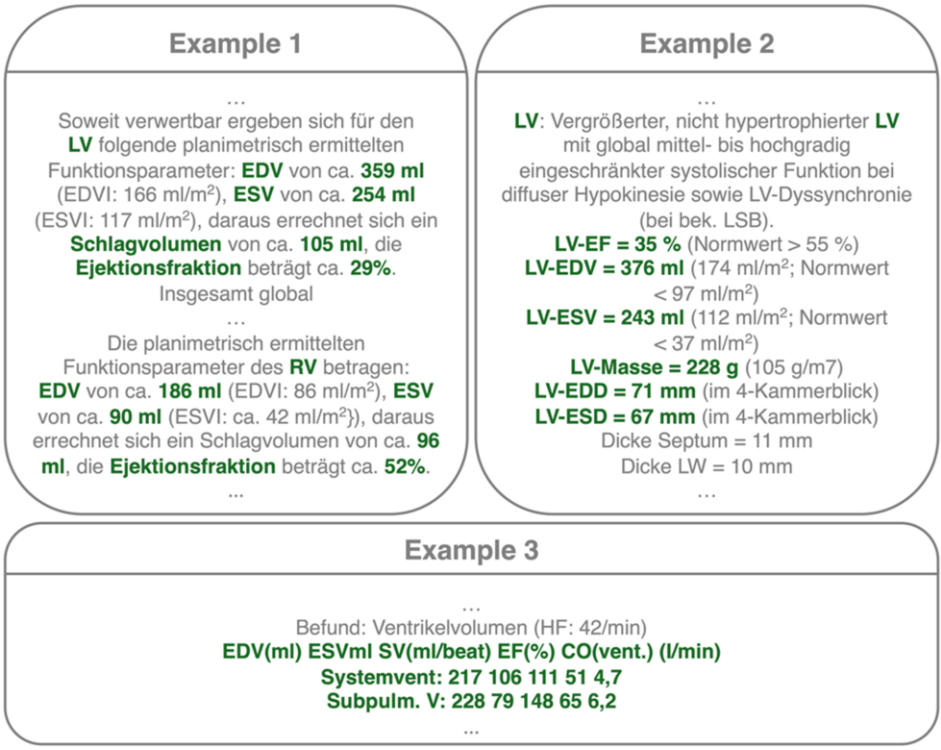

Figure 1. Examples of information representation in different reporting formats. Example 1 shows information within complete sentences, emphasizing the challenge of associating LV/RV with cardiac function and identifying 'Schlagvolumen' as SV and 'Ejektionsfraktion' as EF. Example 2 uses a tabular-like format using key-value pairs. Example 3 features a complete table, highlighting the identification difficulties between LV ("systemvent") and RV ("subpulm. V"). The dataset features a balanced proportion of styles in form of examples 1 and 2, with only a few outliers in the style of example 3.

For our investigation, we consider a dataset of German magnetic resonance imaging (MRI) reports. It contains cardiac function indicators in various formats. Figure 1 shows three examples from different reports. The aim is to correctly assign parameters like

ejection fraction (EF) or volumes separately for the left and right ventricle. In our available dataset, this information is represented in different forms, sometimes in a table-like format and sometimes in complete natural sentences. Each text usually contains only a few parameters. This leads to a combination of NER and Named Entity Normalization (NEN), with particular emphasis on the recognition of key-value pairs.

In this study, we investigate the feasibility of structured data extraction from the MRI reports with various established open-source models. We have set up a simple pipeline using zero-shot prompting. The quality of LLM annotations for the different architectures and model sizes were compared with manual human annotations as benchmark. Our objective is to evaluate the performance of modern and publicly available LLMs at NER and NEN tasks on a German MRI-report dataset.

2. Materials and Methods

2.1. Data set – German clinical text

The dataset used in the present study was gathered from a patient cohort that underwent cardiac MRI for diagnostic workup before or after implantable cardioverter defibrillator (ICD) implantation for primary or secondary prevention of sudden cardiac death at a large tertiary center in Germany. Retrospective analysis of the data was approved by the local Medical Ethics Committee (Ärztekammer Westfalen-Lippe, approval no. 2022-494-f-S) under a waiver of informed consent in accordance with state law for health data privacy (§6 Abs. 2 GDSG NW). 498 written reports of 437 patients were available for analysis. Each report consists of one text of German natural language describing the imaging results including standard cardiac function indicators. These texts follow different structured and unstructured formats (see Figure 1).

2.2. System infrastructure

To benchmark a broad spectrum of different LLM architectures, we took advantage of the open-source project Ollama in version 0.1.32 [5]. All models were installed in their latest versions as of April 20, 2024. Ollama, integrated through Docker, operated on three Nvidia A40 graphics cards, each with 48 GB of GDDR6 VRAM. Tests have shown that models such as Llama 3 70b and Mixtral 8x7b also operate smoothly on a single A40 graphics card. However, for models like the Mixtral 8x22b, combining multiple graphics cards was necessary.

2.3. Query procedure

For the comparison, we consider different architectures in various sizes, including Mistral (7B), Mixtral (8x7b, 8x22b), Gemma (7b), and Llama 3 (8b, 70B) [6, 7, 8]. For each model, we select the instruct version. We define the system prompt:

```
Only output extractions in form of LV-EF, LV-EDV, LV-ESV, LV-SV,
LV-Masse, LV-EDD, LV-ESD, RV-EF, RV-EDV, RV-ESV, RV-SV, RV-Masse,
RV-EDD, RV-ESD with units if specified. It is very important you
only deliver all 14 values in this specific key notation: LV-EF,
LV-EDV, LV-ESV, LV-SV, LV-Masse, LV-EDD, LV-ESD, RV-EF, RV-EDV, RV-
ESV, RV-SV, RV-Masse, RV-EDD, RV-ESD. EF means 'Ejektionsfraktion'.
```

```
EDV means 'enddiastolisches Volumen'. SV means 'Schlagvolumen' and
ESV means 'Endsystolisches Volumen'. Do not write sentences or
other words except the values. If no matching values are found, use
'-' in the appropriate place. You are solely for extraction.
```

and the user prompt:

```
Extract LV-EF, LV-EDV, LV-ESV, LV-SV, LV-Masse, LV-EDD, LV-ESD, RV-
EF, RV-EDV, RV-ESV, RV-SV, RV-Masse, RV-EDD, RV-ESD from the text
with units if specified. Respond using JSON: \n {MRI report}
```

For each clinical MRI report, we generated one API call. These calls were conducted using the JSON format and zero temperature setting. The Python *eval()* function then converted the JSON response into a Python dictionary. If converting failed due to syntax errors, the entire response was set to "-".

2.4. Evaluation

We structured our evaluation into three components. First, we measured the frequency of parsing errors, specifically cases where converting a JSON string directly to a Python dictionary was unsuccessful. Second, we analyzed the NER accuracy of entries that were successfully converted. This step involved examining individual values, truncating any potential characters following '(' in the data, and removing all whitespace. An entry was marked as correct only if it also included the correct unit. For instance, in the case of an LV-ESV reading "256 ml", both "256ml" and "256 ml (116 ml/m^2)" were labeled as correct. Finally, we combined these two metrics to measure the overall performance. For precision, recall, and F1-score calculations, non-existent entries ("-") were labeled as negatives, while existing entries were marked as positives.

3. Results

A systematic comparison with human extractions revealed mixed results, as shown in Table 1. Different models caused highly varying frequencies of parsing errors. For example, Gemma performed best with an extremely low error rate of 0.4%, whereas Llama 3 8b made parsing errors in 44.6% of cases. Notably, the rate of parsing errors did not seem to be influenced by model size. In contrast, the NER accuracy, which measures the extraction accuracy following successful parsing, seemed to correlate more with model size. Here, larger models such as Llama 3 70b and Mixtral 8x22b showed the best results with 99.8% and 97.5%. However, smaller models like Gemma and Mistral 7b also showed reasonable performance, with NER accuracy of around 94 %. When combining the parsing error with the NER accuracy, the overall performance metrics emerge, with Mixtral 8x7b leading at 95.4% accuracy and 94.6% F1-score.

Table 1. Performances of different open-source instruction models of different sizes. The "Parsing Error" column represents the error rate in the eval function, reflecting the frequency of unsuccessful direct parsing from a JSON string to a Python dictionary. "NER Accuracy" specifically addresses the 14 cardiac function indicators, such as LV-EF, but only for API calls that were successfully parsed into a Python dictionary. "Overall Accuracy", "Overall Precision", "Overall Recall", and "Overall F1-score" capture the total extraction efficacy, including the impact of parsing errors. "Dummy" indicates the performance of a Dummy extractor that returns "-" for every queried indicator. As there are many MRI reports not giving specific numeric measurements for every indicator, the Dummy indicator performance reflects the rate of missing values in the

dataset. The Dummy performance can be considered as a baseline performance when interpreting the results of the different LLMs.

	Params	Parsing Error	NER Accuracy	Overall Accuracy	Overall Precision	Overall Recall	Overall F1-score
Dummy	-	-	-	57.0 %	-	-	-
Mistral	7b	8.2 %	93.1 %	90.3 %	90 %	87 %	88.5 %
Gemma	7b	**0.4 %**	94.7 %	94.6 %	98.4 %	88.8%	93.4 %
Llama 3	8b	44.6 %	82.8 %	70.3 %	80.1 %	42.6 %	55.6 %
Mixtral	8x7b	2.2 %	96.6 %	**95.4 %**	95.2 %	**94.1 %**	**94.6 %**
Llama 3	70b	9.8 %	**99.8 %**	94.5 %	**99.7 %**	87.4 %	93.2 %
Mixtral	8x22b	5.8 %	97.5 %	94.6 %	99.2 %	88.2 %	93.4 %

4. Discussion

Our research demonstrates robust performance by open-source models in handling the described NER/NEN task with an overall accuracy up to 95.4 % and F1-score up to 94.6%. This adds to the significant progress made in this field over the last years.
Former research projects employed text mining techniques to enhance the accessibility of data for analysis. A survey conducted in 2021 identified a trend towards using machine learning-based approaches and deep learning models for NER [9]. In the same year, Frei et al. (2021) developed the first open natural language processing (NLP) model dedicated to NER in German medical texts [10]. This model successfully labeled words in German medical texts with categories such as Drug, Strength, Frequency, Duration, Form, and Dosage. Building on this foundation, Frei et al. (2022) led to the introduction of a fine-tuned transformer-based German NLP model that outperformed their previous work on entity recognition [11]. More recently, in 2023, advancements in deep learning and NLP for medical text processing resulted in a novel information extraction framework tailored for low-resource languages like German. Utilizing a pre-training strategy with real-world CT head report datasets and subsequent domain-adaptive fine-tuning on various imaging examination reports, their method successfully transferred clinical reporting domain knowledge while maintaining high accuracy even with limited labeled data [12]. In contrast, our study specifically assesses the effectiveness of generalized open-sourced LLMs as extraction tools using targeted zero-shot prompting.

The syntactic quality of the JSON output, measured by parsing errors, was unaffected by model size. Gemma 7b excels in this category with a performance of 0.4%. Similarly, the precision score, which reflects the rate of false positives and hallucinations, also shows independence from model size. For instance, the Llama 3 8b model demonstrates deficiencies with a precision of 80.1%. In contrast, models like Gemma, Llama 3 70b, and Mixtral 8x22b exhibit precision rates exceeding 98%. Despite this, variations in NER accuracy were noted based on model size. Smaller models like Gemma and Mistral 7b achieved NER accuracies around 94% in our tests targeting 14 values. On the contrary, larger models such as Mixtral 8x22b and Llama 70b delivered better results, with performances surpassing 97%. The overall performances, which consider the error rate, further equalizes some models, resulting in Gemma, Llama 3 70b, and the two Mixtral models achieving comparably high accuracies at 95%. The greatest differences in performance can be seen between the smaller models, where Gemma performs on par with the bigger models, while Llama3 8b and to a lesser degree Mistral perform visibly worse. Overall, Gemma produces the fewest parsing errors and is the best performer among the smaller models.

It should be highlighted that there has been minimal optimization within our pipeline. Firstly, we utilize a zero-shot approach, which could be responsible for some flexible responses. Secondly, parsing errors have not been specifically addressed, which results in notably lower performances for models like Llama 3 8b and 70b. Many of these errors could potentially be resolved through various post-processing steps and should be subject of further research.
Parallel to the widespread enthusiasm for large language models, the open-source community is currently experiencing significant growth. Open-source models have been rapidly advancing, closing the gap with proprietary models such as Gemini or GPT-4 from Google and OpenAI. Our results highlight that these models demonstrate high accuracy and efficiency in NER tasks, with notable performance from models like Gemma 8b, Llama 3 70b, and Mixtral.

5. Conclusion

The use of LLMs for NER tasks is straightforward and effective. Intuitive prompting allows for effective parameter extraction from medical texts, where alternatives like regex fall short. Additionally, even smaller models that can operate locally on standard-performance computers, such as Mistral 7b or Gemma, prove to be beneficial. However, our results suggest employing models like Llama 3 70b and Mixtral 8x22b, together with an optimized JSON-Dict-Parsing process, leads to the highest quality annotations. The presented models could be used in a framework supporting medical personal in transferring information from unstructured text into structured data.
Moving forward, future work could explore the impact of differently phrased prompts and the effects of fine-tuning models to extract specific parameters. Particularly, fine-tuning is expected to reduce parsing errors and enhance NER accuracy. This continued research will further refine and enhance the capability of LLMs in medical text processing, making them even more valuable in practical applications.

Declarations

Conflict of Interest: The authors declare that there is no conflict of interest.

Author contributions: LP and PN executed the computations; AB and JV supervised computational side of the work; AB supervised the medical side; LP, NL drafted the manuscript. KY was responsible for data accessibility and management. All authors approved the manuscript in the submitted version and take responsibility for the scientific integrity of the work.

Ethics: Retrospective analysis of the data was approved by the local Medical Ethics Committee (Ärztekammer Westfalen-Lippe, approval no. 2022-494-f-S) under a waiver of informed consent in accordance with state law for health data privacy (§6 Abs. 2 GDSG NW).

References

[1] Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW, Large language models in medicine, Nature Medicine, 29, pp. 1930-1940, 2023, doi: 10.1038/s41591-023-02448-8Petitti DB, Crooks VC, Buckwalter JG, Chiu V. Blood pressure levels before dementia. Arch Neurol. 2005 Jan;62(1):112-6. DOI: 1111/1234X

[2] Varghese S, Riepenhausen S, Plagwitz L, Varghese J, Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks, Nature Communications, 2050, 15, 2024, doi: 10.1038/s41467-024-46411-8.

[3] Keraghel I, Stanislas M, and Mohamed N. A survey on recent advances in named entity recognition. *arXiv preprint arXiv:2401.10825* (2024), doi: 10.48550/arXiv.2401.10825

[4] Henry S, Buchan K, Filannino M, Stubbs A, and Uzuner O, 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records, J. Am. Med. Inform. Assoc., vol. 27, no. 1, pp. 3–12, Jan. 2020, doi: 10.1093/jamia/ocz166

[5] Morgan J, jmorganca/ollama. Jan. 18, 2024. Accessed: Jan. 18, 2024. [Online]. Available: https://github.com/jmorganca/ollama

[6] Touvron H *et al.*, 'LLaMA: Open and Efficient Foundation Language Models. *arXiv*, Feb. 27, 2023. doi: 10.48550/arXiv.2302.13971.

[7] Jiang AQ *et al.*, Mixtral of Experts. *arXiv*, Jan. 08, 2024. doi: 10.48550/arXiv.2401.04088.

[8] Mesnard T et al. Gemma: Open Models Based on Gemini Research and Technology, *arXiv*, Mar. 13, 2024, doi: 10.48550/arXiv.2403.08295

[9] Bose P, Srinivasan S, Sleeman WC IV, Palta J, Kapoor R, Ghosh P. A Survey on Recent Named Entity Recognition and Relationship Extraction Techniques on Clinical Texts. *Applied Sciences*. 2021; 11(18):8319, doi: 10.3390/app11188319

[10] Frei J, Kramer F. GERNERMED: An open German medical NER model. Software Impacts, arXiv, Dec. 10, 2021, doi: 10.48550/arXiv.2109.12104

[11] Frei J, Frei-Stuber L, Kramer F. GERNERMED++: Transfer Learning in German Medical NLP. arXiv, Oct. 9, 2022, doi: 10.48550/arXiv.2206.14504

[12] Jantscher M, Gunzer F, Kern R *et al.* Information extraction from German radiological reports for general clinical text and language understanding. Sci Rep 13, 2353 (2023), doi: 10.1038/s41598-023-29323-3