Adaption and Evaluation of Generative Large Language Models for German Medical Information Extraction

Sören Spiegel[1], Seid Muhie Yimam[2], Philipp Breitfeld[1], Frank Ückert[1]

[1]Institute for Applied Medical Informatics, University Medical Center Hamburg-Eppendorf,
[2]University of Hamburg

Correspondence: Sören Spiegel s.spiegel@uke.de

Abstract

Analyzing unstructured patient data from electronic health records can improve clinical decision-making. However, the standard approach for medical information extraction (IE) that relies on fine-tuning foundation models for specific tasks faces several challenges, especially in German where training data availability is very scarce. This work investigates the feasibility of large language models (LLMs) equipped with 7 billion parameters for clinical IE in the German medical domain, evaluating their performance on all three publicly available German gold-standard datasets. The results show that LLMs achieve strong performance in drug extraction (F1: 0.71–0.87) but struggle with diagnoses and treatments (F1: 0.44–0.71). Instruction tuning with QLoRA improves performance and reduces hallucinations. While baseline models outperform LLMs on dataset-specific tasks, instruction-tuned LLMs excel on out-of-distribution data, making them a viable option when training data is scarce and detailed accuracy is less critical.

1 Introduction

The widespread adoption of electronic health records has led to the collection of extensive patient data, including family history, symptoms, diagnostic results, and treatments. Effectively analyzing this vast amount of data has the potential to enhance clinical care and support clinical decision-making (Jensen et al., 2012). As most documentation exists in text form, automated information extraction (IE) plays a crucial role in enabling comprehensive analysis of healthcare data (Lentzen et al., 2022).

The potential applications are manifold, one plausible area of application for IE from free written text is premedication in anesthesia. Each patient is examined and assessed preoperatively by an anesthetist to minimize perioperative risk and prepare the patient medically and psychologically for anesthesia (Larsen, 2016). Before the preoperative visit, the anesthetist reviews the patients medical history, focusing on prior diseases, surgeries, allergies, medications, and lab results. This typically involves manually examining physician notes from various sources, such as hospital units or external providers, without any tool to assist in generating premedication reports. This time-consuming process is further complicated by duplicated text in clinical notes (Steinkamp et al., 2022).

Physicians can leverage natural language processing (NLP) methods, such as fine-tuning foundation models, to analyze clinical notes. While current state-of-the-art (SOTA) fine-tuned transformer models achieve remarkable results in clinical named entity recognition (NER), their performance often declines significantly when applied to other datasets (Kühnel and Fluck, 2022). Furthermore, medical IE in Germany faces challenges due to data protection laws, limited annotated datasets, and less research activity in German compared to English (Lentzen et al., 2022; Roller et al., 2022; Richter-Pechanski et al., 2023). Generative large language models (LLMs) offer a promising solution, achieving strong results in various NLP tasks without task-specific fine-tuning (Touvron et al., 2023; Jiang et al., 2024). This paper aims to answer the following research questions: (**RQ1**) Which LLM performs best for medical information extraction in German? (**RQ2**) Which approach yields the highest performance for medical IE: few-shot prompting, instruction fine-tuning or retrieval-augmented generation (RAG)? (**RQ3**) Can generative LLMs with 7 billion parameters compete with smaller task specific fine-tuned models (SLMs) in German clinical IE?

The main **Contributions** of this work include: (1) We present a comprehensive analysis of 7B LLMs for IE in the German medical domain, evaluating their strengths and limitations across multiple datasets. (2) We assess the generalizability of transformer-based NER models by testing models

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 1: Long and Short Papers, pages 33–47, 2025

on several datasets. (3) We establish a benchmark for LLM performance in German clinical IE.

2 Related Works

2.1 Medical Information Extraction

With the rise of electronic health records, extracting structured information from clinical text has become a key research area (Jensen et al., 2012). NER and relation extraction (RE) are central to this task (Landolsi et al., 2023), with early rule-based methods achieving strong results but requiring extensive manual effort. The introduction of transformer-based models like BERT (Devlin et al., 2019) shifted the focus toward machine learning approaches, leading to specialized biomedical models such as BioBERT and ClinicalBERT (Lee et al., 2020; Huang et al., 2020). While these models achieve SOTA performance in shared tasks like CLEF eHealth and n2c2 (Henry et al., 2020), their success relies on well-annotated training data, which is costly to produce and limits generalizability across datasets (Kühnel and Fluck, 2022; Lentzen et al., 2022; Llorca et al., 2023).

German medical IE research lags behind English due to commercialization and data protection constraints (Lentzen et al., 2022; Roller et al., 2022; Richter-Pechanski et al., 2023). Despite these challenges, five annotated German NER datasets have been released: BRONCO (Kittner et al., 2021), CARDIO:DE (Richter-Pechanski et al., 2023), GGPONC2 (Borchert et al., 2022), GERNERMED (Frei and Kramer, 2023) and GPTNERMED (Frei and Kramer, 2023). Note that the latter two are not gold-standard, GERNERMED is an automatically translated subset of n2c2 2018, while GPTNERMED is a synthetic dataset created with GPT NeoX (Black et al., 2022). Studies on token classification models highlight performance disparities across entity types. Early baselines, such as CRF and LSTM on BRONCO (Kittner et al., 2021), showed moderate results (F1: 0.71–0.90), with dictionary-based normalization yielding inconsistent scores. Transformer-based models, including fine-tuned BERT variants on GGPONC2 and CARDIO:DE (Borchert et al., 2022; Richter-Pechanski et al., 2023), consistently outperformed CRF, especially in drug extraction. The work by Frei et al., 2022 demonstrated advancements from CNNs (F1: 0.67) to transformers (F1: 0.91) on GERNERMED, though generalizability remained an issue. Recent efforts, like GPTNERMED (Frei and Frank, 2023), leveraged synthetic data to enhance entity recognition but struggled on external datasets. While progress has been made, the field remains focused on dataset creation and model improvements, with gaps in generalization and broader application.

Beyond pre-trained language models, NLP frameworks like Apache cTakes (Savova et al., 2010) and MetaMap (Aronson, 2001) provide medical IE tools but are optimized for English. Attempts to adapt them to German, such as mapping UMLS concepts from translated clinical notes, resulted in low F1 scores due to linguistic and dataset limitations (Becker and Böckmann, 2016). This highlights a major challenge in clinical NLP: the dominance of English resources, which restricts research in other languages.

2.2 Usage of LLMs for Information Extraction

With the public release of models like LLAMA2 (Touvron et al., 2023), researchers began adapting LLMs for medical tasks, primarily focusing on text summarization and question answering (QA) (Li et al., 2023; Singhal et al., 2023). Notable advancements include Med-PaLM 2, which achieved 86.5% on USMLE-style questions and outperformed physician-written answers on consumer medical questions (Singhal et al., 2023). Similarly, LLMs have shown promise in clinical text summarization and outperforming baseline models in various NLP tasks, such as ICD-9 classification and entity normalization (Van Veen et al., 2023; Gema et al., 2023). Studies also highlight their robustness in low-resource scenarios and generalizability to unseen entities (Wang et al., 2025; Peeters and Bizer, 2024). However, challenges remain in using generative LLMs for token-level tasks like NER, as their architectures are optimized for text generation (Wang et al., 2025).

While initial studies explore LLMs for IE (Agrawal et al., 2022; Wang et al., 2025; Wu et al., 2023), their performance on German medical text remains unexplored. Furthermore, existing research predominantly utilizes LLMs exceeding 70B parameters, making them impractical for users lacking specialized hardware or substantial computational resources.

3 Methods

In this work a mixed method approach is used to explore the applicability of generative LLMs in

the realm of German clinical IE tasks. Traditional NER treats entity recognition as a multiclass classification task, where token classifiers assign labels (e.g., drug, diagnosis, treatment or none often defined as O) to each token. In contrast, generative LLMs are designed to generate text based on a given input, which adds an additional layer of complexity to the task, as the model must not only recognize entities but also generate them correctly in a structured format, ensuring that the output can be processed automatically (Wang et al., 2025). Our study evaluates chat and further instruction-tuned versions of the following LLMs to baseline NER token-classifiers: LLAMA2 7B (Touvron et al., 2023), Mistral 7B (Jiang et al., 2023), Meditron 7B (Chen et al., 2023), Leo hessianai 7B and Leo hessianai Mistral 7B (Plüster, 2023). While Meditron is specifically trained for the medical domain, the latter two are trained for German.

3.1 Few-shot Prompting

When given a few demonstrations, a generative LLM can achieve performance comparable to fine-tuned models specifically trained on one task (Brown et al., 2020). To standardize output and reduce post-processing, we use a few-shot approach, as shown in Figure 1. The LLM is guided to produce structured lists for each entity class. Special instruction tokens *[INST]* within the prompt, used in Mistral 7B Instruct, help distinguish user input from generated content. These tokens vary across models. Following Agrawal et al., 2022, our

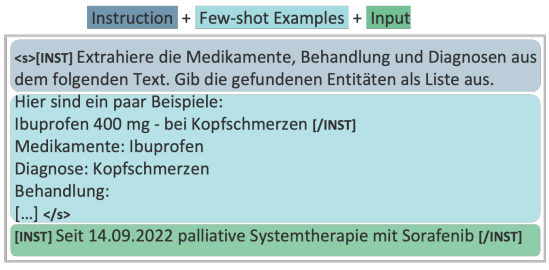

Figure 1: Few-shot approach to extract named entities from input

prompt includes all entity classes, with eight examples covering different entity combinations and a negative case resulting in 8 few-shot examples. To enhance consistency, the inference temperature is set to 0.1. Entity span boundaries are excluded, as even SOTA models like ChatGPT struggle with accurate character counts (OpenAI, 2022).

3.2 QLoRA Instruction Fine-tuning

To enhance LLMs ability to extract instructed medical entities from clinical text, we fine-tune the models on a custom instruction-tuning dataset. The training data comprises four NER datasets described in Section 4. The heterogeneous data formats were harmonized to the Alpaca instruction template (Taori et al., 2023) as illustrated in Figure 2. Each of the 34,096 training records includes an

```
### Instruction:
Extract all medications and diagnoses from the
following text.
Output the found entities as a list.
### Input:
0.4 Diuretics 0.25 1x/week
### Output:
Medication: Diuretics
Diagnosis:
```

Figure 2: Alpaca based schema for instruction fine-tuning.

instruction, input, and output. The instruction specifies dataset-specific entities in either German or English. The input contains the clinical sentence, and the output lists extracted entities by type. Given the LLMs 7B parameters, full fine-tuning is infeasible within this work. Instead, we employ QLoRA (Dettmers et al., 2023), a parameter-efficient fine-tuning (peft) method. The model is first quantized, after which the *LoRA* adapter is added, freezing the base model while enabling fine-tuning. The LoRA configuration follows Dettmers et al., 2023 with (r=64) and (lora_alpha=16). Fine-tuning is conducted using the *Supervised Fine-tuning Trainer* (Wolf et al., 2020), optimizing the model to predict training records.

3.3 RAG Enhanced Information Extraction

Researchers have shown that Retrieval-Augmented Generation (RAG) improves LLM performance by incorporating relevant information from domain-specific knowledge bases (Lewis et al., 2020; Jadhav et al., 2024). Given that BRONCOs diagnosis annotations are based on the German-modified International Statistical Classification Of Diseases (ICD-10-GM) and treatment annotations originate from the German procedure classification (OPS), this approach aims to enrich few-shot prompts with relevant domain knowledge. The retriever model is designed to suggest relevant ICD-10-GM diagnoses for a given input text. To achieve this, a vector database of the ICD-10-GM classification catalog

is constructed using the danielheinz/e5-base-sts-en-de[1] model, which is trained for German semantic textual similarity. Similarly, the model is used to generate a vector database from the OPS procedure catalog to retrieve matching treatments. We chunk the ICD-10-GM data at the category level, including all associated subcategories, ensuring seamless integration of the retriever models output into the few-shot prompt. Notably, all code IDs are removed from the chunks before embedding, enhancing similarity between the vector embeddings and input text, as clinical documents rarely contain ICD-10-GM codes. The Mistral 7B Instruct LLM acts as the generator model, which processes the input sentence with instructions to extract medical entities while incorporating contextual information from the retriever model in the form of ICD-10-GM diagnoses or OPS procedures.

3.4 Information Extraction Evaluation

To compare the NER performance of generative LLMs with SLMs, this study uses three gold-standard annotated German datasets: BRONCO (Kittner et al., 2021), GGPONC2 (Borchert et al., 2022) and CARDIO:DE (Richter-Pechanski et al., 2023). Drug, diagnosis, and treatment extraction is evaluated using SemEval'13 metrics (Segura-Bedmar et al., 2013) on single runs, which also account for partial entity matches, relevant for multi-token entities like diagnoses, where annotation guidelines vary. Following Dror et al., 2018, we assess statistical significance using McNemar's test (McNemar, 1947). The SLMs are fine-tuned versions of medBert.de (Bressem et al., 2024), a SOTA 109M-parameter German medical language model. The different LLMs, in contrast, are tested without prior exposure to the datasets to assess their generalizability. Reference models include GERNERMED by Frei et al., 2022 for drug extraction, and GPTNERMED by Frei and Frank, 2023.

4 Datasets

Instruction Tuning Data. The custom instruction tuning dataset is a collection of the following four NER datasets: GERNERMED (Frei et al., 2022), GPTNERMED (Frei and Frank, 2023), i2b2 (Uzuner et al., 2011), and n2c2 (Henry et al., 2020).

[1] https://huggingface.co/danielheinz/e5-base-sts-en-de

Dataset	Train	Validation	Test
BRONCO	6,279	1,346	1,346
GGPONC2	59,515	12,770	3,000
CARDIO:DE	67,062	14,370	8,969

Table 1: Dataset splits used to train baseline NER tagger and compare LLMs to baseline.

A detailed overview of the datasets and their medical entities can be found in the Appendix A Table 5.

Evaluation Data. We evaluate our approach using the BRONCO, GGPONC2, and CARDIO:DE gold-standard datasets. While BRONCO and CARDIO:DE consist of pseudonymized clinical documents, GGPONC2 is a collection of annotated guidelines, resulting in a larger corpus. As shown in Table 5 in Appendix A, the annotated entity types vary greatly across these datasets. Since the CARDIO:DE annotations do not follow the IOB schema, this test split cannot be evaluated on entity type level but on strict spans only. The dataset splits used for training the baseline model medBert.de are detailed in Table 1, with the test split serving as the evaluation set.

5 Experimentation

All experiments were executed on the same server with the following specifications: NVIDIA RTX A5000 with 24564MiB, Cuda version 11. The source code to reproduce the experiments is available online[2].

5.1 LLM Comparison

Before turning to the LLM comparison to baseline token classifiers, Figure 3 shows the evaluation of instruction-tuned LLMs on the BRONCO test set to answer **RQ1** and explore **RQ2**. All models performed best on medication extraction and worst on identifying treatments, with F1 ranging from 0.88 for medication to 0.14 for treatments. The QLoRA instruction-tuning consistently improved performance, especially for Meditron, while Mistral LLMs outperformed other models across all entity types. We also observed a drastic decline in hallucinations for our instruction fine-tuned LLMs. While the chat-based LLMs generated up to 156 entities that could not be matched to the input text, the instruction-tuned models produced only 0 to 3 hallucinated entities on the BRONCO test sample (Ap-

[2] https://github.com/IAMspiegel/LLMGerMedIE

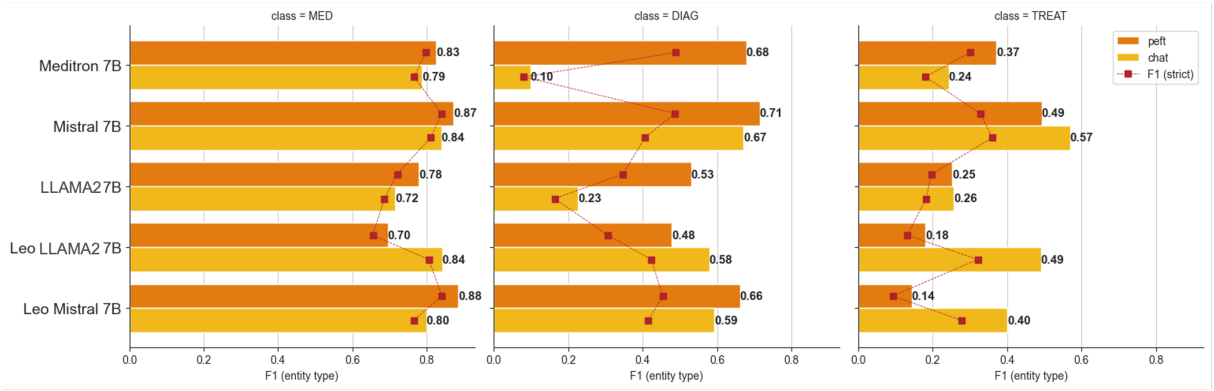

Figure 3: NER results for the BRONCO test sample on entity type level for different versions of LLMs. Compared is the publicly available chat version with instruction tuned (peft) models. Note that the Meditron 7B chat version is not publicly available, thus the base version was used.

Level	Model (Train Data)	Medication			Diagnosis			Treatment			Micro Avg.
		P	R	F1	P	R	F1	P	R	F1	F1
Entity Type	medBert.de (BRONCO)	.95	.96	**.96**	.84	.94	**.89**	.86	.94	**.90**	**.90**
	medBert.de (GGPONC2)	.75	.85	.80	.63	.84	.72	.44	.88	.59	.69
	medBert.de (CARDIO:DE)	.81	.81	.81	.0	.0	.0	.0	.0	.0	.27
	GermanBERT (GERNERMED)	.39	.92	.55	.0	.0	.0	.0	.0	.0	.28
	GermanMedBERT (GPTNERMED)	.51	.88	.64	.44	.62	.51	.0	.0	.0	.55
	Mistral 7B Instruct (8-shot)	.86	.82	.84	.70	.64	.67	.48	.71	.57	.66
	peft Mistral 7B	.86	.88	.87	.67	.77	.71	.61	.41	.49	.69
Token	medBert.de (BRONCO)	.90	.91	**.91**	.73	.81	**.77**	.75	.82	**.79**	**.80**
	medBert.de (GGPONC2)	.70	.80	.75	.38	.51	.44	.33	.64	.43	.48
	medBert.de (CARDIO:DE)	.81	.81	.81	.0	.0	.0	.0	.0	.0	.27
	GermanBERT (GERNERMED)	.37	.92	.52	.0	.0	.0	.0	.0	.0	.26
	GermanMedBERT (GPTNERMED)	.48	.85	.62	.29	.41	.34	.0	.0	.0	.55
	Mistral 7B Instruct (8-shot)	.83	.80	.81	.43	.39	.41	.30	.45	.36	.46
	peft Mistral 7B	.83	.85	.84	.45	.52	.49	.41	.27	.33	.51

Table 2: Results (precision, recall, F1) of BRONCO test sample. Comparison of best fine-tuned LLM, chat Mistral-7B to the BRONCO baseline model and BERT based models trained on other datasets. The displayed scores are calculated on entity type basis and strict span evaluation (token). The reported scores for medBert.de (GGPONC2) follow the mapping: Clinical_Drug → Medication, Diagnosis_or_Pathology → Diagnosis, Therapeutic and Diagnostic → Treatment. Alternative mappings result in a lower performance.

pendix B.1 Table 6). LLMs tended to over-tag diagnoses, likely due to broader concept recognition beyond ICD-10-GM. In contrast, they struggled with identifying cancer-related treatments, showing high precision but low recall. Overall, LLMs excelled in drug extraction but underperformed in recognizing OPS treatments. An additional analysis of medical abbreviations further confirmed a performance drop for previously unseen diagnosis and treatment phrases, highlighting the models limitations in generalizing to less common or implicitly expressed entities. Since both Mistral versions perform best in their category, we use these LLMs in the following comparison to baseline NER models.

5.2 LLMs Compared to Baseline

5.2.1 BRONCO

The results of the baseline SLMs and top-performing LLMs on the BRONCO test set to answer **RQ2** and **RQ3** are shown in Table 2. As expected, medBert.de fine-tuned on BRONCO achieves the best overall F1 (0.9). Its performance surpasses the baseline of Kittner et al., 2021 but remains below the top-performing model on the BRONCO leaderboard[3]. Notably, GERNERMED and GPTNERMED perform poorly, over-generating medication entities with low precision and failing to capture treatments.

Our fine-tuned peft Mistral 7B model achieves the best F1 for medication and diagnosis second

[3] https://www2.informatik.hu-berlin.de/~leser/bronco/index.html

to medBert.de (BRONCO). However, not significantly better than medBert.de (GGPONC2), which outperforms the LLMs in identifying treatments, although its performance remains significantly below the BRONCO-trained baseline as illustrated in Appendix B.3. Interestingly, Mistral-7B Instruct achieves slightly better results in extracting medications compared to most NER models, reinforcing the strengths of LLM transfer learning over dataset-specific fine-tuning.

5.2.2 GGPONC2

The NER results for the GGPONC2 test sample are presented in Table 3. Its annotation guidelines differ from BRONCO, leading to variations in entity types. The diagnosis class includes pathology related entities, and previous analysis found that mapping GGPONC2s therapeutic class to BRONCOs treatment class yields the highest extraction scores. Since therapeutic entities dominate this category, it serves as the test datasets equivalent to the treatment entity type. As expected, medBert.de (GGPONC2) achieves the highest F1 (0.92) across all entities, slightly surpassing the performance of the BERT based NER model reported by Borchert et al., 2022. While medBert.de (CARDIO:DE) performed reasonably well on BRONCO medications, it struggles on GGPONC2 drugs (F1: 0.57). Both Mistral 7B Instruct and our instruction-tuned peft variant outperform three reference token classifiers across all entities, excelling in the drug class but performing worst in therapeutic. A closer look reveals that medBert.de (BRONCO) and both Mistral LLMs exhibit higher precision than recall for the diagnosis/pathology class, likely because the token classifier was trained for diagnoses only, while the LLMs were prompted for diagnoses but not pathology or symptoms. Our fine-tuned peft Mistral 7B LLM outperforms Mistral-7B Instruct across all entities, with the largest gap in the diagnosis/pathology class. However, it marginally surpasses medBert.de (BRONCO) only in the drug category. While the F1 difference for diagnosis/-pathology is notable, results for therapeutic are comparable.

5.2.3 CARDIO:DE

As the CARDIO:DE annotations distinguish between active ingredients and drug entities but BRONCO and GGPONC2 do not, the entity classes were combined into one single class. The results in Table 4 show that the medBert.de model trained on CARDIO:DE achieves the highest F1 score, with a notable gap between recall and precision. Its performance also exceeds that of the BERT transformer presented in the original CARDIO:DE study by Richter-Pechanski et al., 2023. Other models exhibit comparable trends, while medBert.de (BRONCO) and our peft Mistral 7B have recall scores close to medBert.de (CARDIO:DE), their precision is noticeably lower. GERNERMED and GPTNERMED again show strikingly low precision, suggesting excessive entity tagging. The instruction fine-tuned Mistral 7B achieves the second-best F1 score, outperforming medBert.de models trained on BRONCO and GGPONC2. Notably, Mistral 7B Instruct, with just eight few-shot examples, surpasses medBert.de (BRONCO) in F1, making it the only model with higher precision than recall. Its precision matches the best baseline model, likely because it was not influenced by fine-tuning on datasets where active ingredients were tagged as drugs.

5.3 RAG Information Extraction

To further investigate **RQ2** and to enhance the IE performance of LLMs for diagnosis and treatment entities, we tested a RAG approach, which provides the LLM with additional context, including relevant ICD-10-GM codes or OPS procedures. However, our experiment did not result in an improvement in NER scores on the BRONCO test sample. The embedding model evaluation, shown in Figure 4, reveals that diagnoses were better recognized, with higher similarity scores between the retrieved code categories and input sentences. In contrast, treatment entities did not exhibit the same clear pattern. Despite using optimized prompts, the F1 scores for both diagnosis and treatment entities decreased drastically, as shown in Figure 6 in Appendix B.2, with treatment entities experiencing a more substantial decline due to insufficient context generation and limitations of the OPS knowledge source. Several factors likely contributed to this decline in performance. First, the OPS catalog proved suboptimal for retrieval, often returning context with high lexical similarity but low semantic relevance. Second, the German embedding model may luck sufficient domain coverage, reducing retrieval accuracy. Third, the Mistral 7B chat model struggled to process single-phrase inputs effectively, highlighting limitations in contextual understanding compared to larger models.

Level	Model (Train Data)	Drug			Diagnosis/Pathology			Therapeutic			Micro Avg.
		P	R	F1	P	R	F1	P	R	F1	F1
Entity Type	medBert.de (BRONCO)	.77	.70	.74	.84	.59	.69	.65	.38	.48	.62
	medBert.de (GGPONC2)	.91	.94	**.93**	.91	.94	**.92**	.91	.93	**.92**	**.92**
	medBert.de (CARDIO:DE)	.58	.55	.57	.0	.0	.0	.0	.0	.0	.19
	GermanBERT (GERNERMED)	.45	.70	.55	.0	.0	.0	.0	.0	.0	.12
	GermanMedBERT (GPTNERMED)	.35	.69	.47	.39	.46	.42	.0	.0	.0	.32
	Mistral 7B Instruct (8-shot)	.78	.62	.70	.71	.34	.46	.52	.37	.43	.48
	peft Mistral 7B	.77	.73	.75	.73	.48	.58	.65	.33	.44	.55
Token	medBert.de (BRONCO)	.73	.67	.70	.64	.45	.53	.47	.27	.35	.48
	medBert.de (GGPONC2)	.88	.90	**.89**	.90	.92	**.91**	.90	.92	**.91**	**.91**
	medBert.de (CARDIO:DE)	.58	.55	.57	.0	.0	.0	.0	.0	.0	.19
	GermanBERT (GERNERMED)	.41	.64	.50	.0	.0	.0	.0	.0	.0	.11
	GermanMedBERT (GPTNERMED)	.34	.67	.45	.31	.37	.34	.0	.0	.0	.27
	Mistral 7B Instruct (8-shot)	.73	.58	.65	.58	.28	.37	.42	.30	.35	.40
	peft Mistral 7B	.71	.68	.70	.51	.34	.41	.45	.23	.31	.41

Table 3: Results (precision, recall, F1) of GGPONC2 test sample. Comparison of best fine-tuned LLM, chat Mistral 7B to the baseline model and BERT based models trained on other datasets. The displayed scores are calculated on entity type basis and strict span evaluation (token).

Model (Train Data)	Precision	Recall	F1
medBert.de			
- (BRONCO)	.43	.89	.58
- (GGPONC2)	.61	.83	.70
- (CARDIO:DE)	**.84**	**.93**	**.88**
GermanBERT - (GERNERMED)	.20	.89	.32
GermanMedBERT - (GPTNERMED)	.24	.85	.37
Mistral 7B Instruct (8-shot)	.81	.54	.64
peft Mistral 7B	.59	.88	.71

Table 4: Results of CARDIO:DE test sample. Comparison of best fine-tuned LLM, chat Mistral-7B to the baseline models. The displayed scores are calculated on token level.

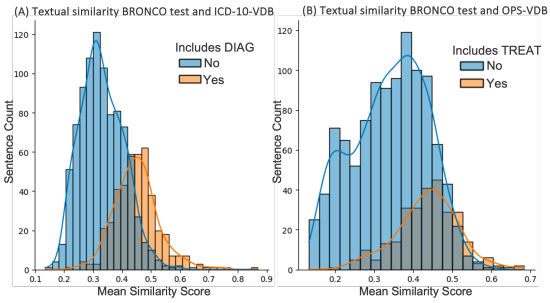

Figure 4: Textual similarity between BRONCO test data and vector databases. (A) Similarity distribution between BRONCO test sentences and the ICD-10-GM vector database (VDB), categorized by whether the sentence contains a diagnosis entity. (B) Similarity distribution between BRONCO test sentences and the OPS VDB, categorized by whether the sentence includes a treatment entity.

5.4 Out-of-Distribution Analysis

To address **RQ3** and assess generalizability, an additional analysis of the datasets reveals that the strong performance of medBert.de (GGPONC2) across all three evaluation datasets, including BRONCO, can be attributed to a high proportion of shared entities between the GGPONC2 training and BRONCO test data. 58% of BRONCOs test treatment entities and 51% of diagnoses appear in GGPONC2, while the overlap with LLM instruction data is remarkably lower (31% for diagnoses, 28% for treatments). Figure 5 further illustrates how models perform on out-of-distribution (OOD) entities. medBert.de (GGPONC2) exhibits a sharp performance drop in the OOD (GGPONC2) dataset, in which all overlapping entities between BRONCO and GGPONC2 were removed. This highlights the models strong reliance training data. Conversely, our instruction fine-tuned peft Mistral 7B maintains more stable performance across OOD datasets, even surpassing medBert.de (GGPONC2) on OOD drugs and diagnoses. Additionally, Mistral 7B Instruct achieves an F1 score for treatment concepts comparable to medBert.de (GGPONC2).

These findings underscore the advantages of domain overlap in training data for token classifiers while demonstrating LLMs greater adaptability to unseen entities.

6 Discussion

The presented work was designed to test 7B LLMs for German medical IE on unseen datasets and compare their performance with the baseline SLMs. We found that LLMs exhibit strong generalizability on unseen data due to their extensive pre-training on large corpora and vast parameter size. However, their limited domain-specific knowledge pre-

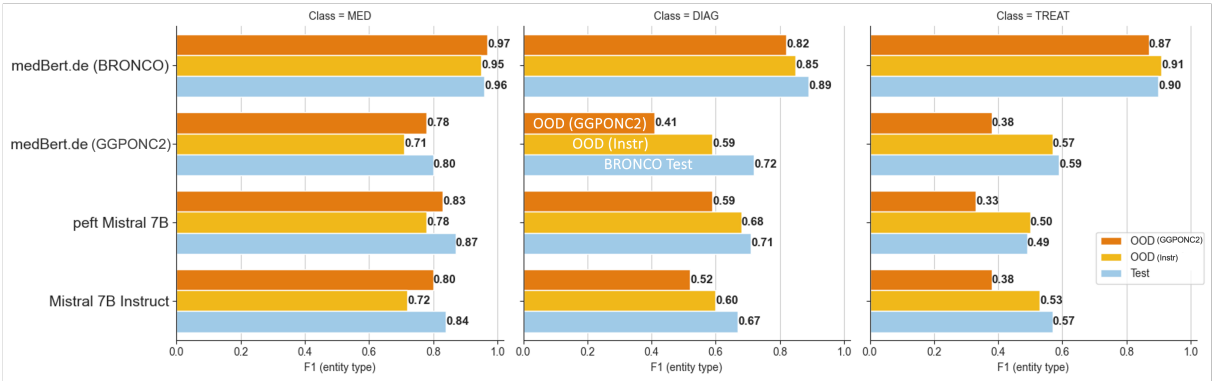

Figure 5: F1 scores for entity type classification on different subsets of the BRONCO test dataset. Each model is evaluated on the full test set (Test) and two OOD subsets: (1) OOD (GGPONC2), where entities present in the GGPONC2 training data are removed, and (2) OOD (Instr), where entities overlapping with the instruction fine-tuning dataset are excluded.

vents them from reaching the F1 scores above 0.90 needed for reliable, production-level applications. While QLoRA instruction tuning enables effective 7B LLM fine-tuning on a single GPU and has demonstrated strong IE performance, this study also identified drawbacks. The trained adapters can reduce generalizability and multitask ability, as noted by Van Veen et al., 2023, who observed worsened text summarization after QLoRA fine-tuning. Similarly, our findings show a considerable drop in treatment extraction performance, likely due to the instruction datasets lack of treatment examples. Despite this, LLMs offer greater flexibility, making them more adaptable to various tasks and domains than SLMs. However, post-processing may be more challenging. As a result, they can be a viable option for extracting medical information from unstructured text when top-tier performance is not essential.

While SOTA commercial LLMs, such as GPT-4, perform well on multilingual clinical tasks (Qiu et al., 2024), we observe that this is not the case for smaller 7B models like Meditron. Despite additional training in the medical domain on English data, Meditron failed to reliably extract German medical terms. Domain-specific terminology and abbreviations may therefore pose a greater challenge for cross-lingual transfer in smaller LLMs.

Computational Resources. LLMs incur higher CPU and GPU costs compared to baseline SLMs, but using the vLLM package (Kwon et al., 2023) significantly reduces inference time, making it comparable to that of baseline SLMs, despite requiring around 24GB of VRAM. Without vLLM, inference times can stretch for hours with just a few thousand samples, underscoring the importance of frameworks like vLLM for optimizing LLM performance and managing resource demands.

Future Work. This paper shows that 7B LLMs lack the medical expertise needed for accurate diagnosis and treatment extraction. To address this, future work could explore larger LLMs, such as the Mixtral 8x7B, which has 47 billion parameters and outperforms 70B models while being manageable during inference. Although our RAG setup did not improve IE performance for several factors, it remains a promising strategy for boosting LLMs capabilities in extracting information (Vithanage et al., 2025; Bartels and Carus, 2025). Another potential approach is integrating medical knowledge via knowledge graphs, as outlined by Pan et al., 2024.

Instead of solely adapting LLMs, future research could focus on disambiguating clinical text to enhance model understanding, as demonstrated by Agrawal et al., 2022.

7 Conclusion

This work explored the ability of 7B LLMs for German medical IE across multiple datasets. Key findings include: **(1)** Mistral 7B outperforms other LLMs, including German trained models, in extracting medications, diagnoses and treatments. **(2)** Instruction tuning via QLoRA improves IE performance but may reduce treatment recognition due to imbalanced training data. **(3)** While LLMs do not surpass SLMs on their domain, they achieve strong medication extraction results (F1: 0.87–0.71) and remain competitive in diagnosis and treatment extraction. **(4)** When comparing LLMs to baseline

SLMs on out-of-domain datasets, the instruction-tuned peft Mistral 7B outperformed four of the five baseline models. Additionally, the Mistral LLMs surpassed the baseline models on out-of-distribution data.

These findings highlight the ongoing challenges in German medical IE and emphasize the need for diverse, high-quality, language-specific datasets for fine-tuning models, particularly in specialized domains like medical NLP. While 7B LLMs are less suited for tasks requiring precise information extraction, they can still be valuable in situations where training data is limited, and the primary objective is to gain a broader understanding of unstructured text rather than achieving excellent accuracy.

8 Limitations

The scarcity of publicly available gold-standard medical datasets in Germany hinders both NER model development and cross-dataset validation. The evaluation datasets BRONCO, GGPONC2, and CARDIO:DE differ in size, origin, document type, and annotation style, making general conclusions about cross-domain performance difficult. BRONCO and GGPONC2 stem from oncology, while CARDIO:DE, from cardiology, lacks annotated diagnoses and treatments. Annotation inconsistencies further impact performance, as BRONCO follows structured classification catalogs (ATC, ICD-10-GM, OPS), whereas GGPONC2 does not. Document-level differences also introduce ambiguity, affecting entity classification. Additionally, BRONCOs shuffled sentence structure disrupts context, potentially impairing LLM performance. While the three German datasets used for evaluation in this study are publicly available only upon request and are stored in CoNLL-style or other structured annotation formats, we cannot rule out the possibility that portions of them were included in the pretraining corpora of the evaluated LLMs. Given that most LLM training data sources are undisclosed, this remains a general limitation in LLM evaluation.

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. Association for Computational Linguistics.

Alan R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

Stefan Bartels and Jasmin Carus. 2025. From text to data: Open-source large language models in extracting cancer related medical attributes from german pathology reports. *International Journal of Medical Informatics*, 203:106022.

Matthias Becker and Britta Böckmann. 2016. Extraction of umls® concepts using apache ctakes™ for german language. In *Health Informatics Meets Ehealth*, pages 71–76. IOS Press.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20b: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136. Association for Computational Linguistics.

Florian Borchert, Christina Lohr, Luise Modersohn, Jonas Witt, Thomas Langer, Markus Follmann, Matthias Gietzelt, Bert Arnrich, Udo Hahn, and Matthieu-P. Schapranow. 2022. GGPONC 2.0 - the german clinical guideline corpus for oncology: Curation workflow, annotation policy, baseline NER taggers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3650–3660. European Language Resources Association.

Keno K. Bressem, Jens-Michalis Papaioannou, Paul Grundmann, Florian Borchert, Lisa C. Adams, Leonhard Liu, Felix Busch, Lina Xu, Jan P. Loyen, Stefan M. Niehues, Moritz Augustin, Lennart Grosser, Marcus R. Makowski, Hugo J. W. L. Aerts, and Alexander Löser. 2024. medBERT.de: A comprehensive german BERT model for the medical domain. *Expert Systems with Applications*, 237:121598.

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. *Preprint*, arxiv:2005.14165 [cs].

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. MEDITRON-70b: Scaling medical pretraining for large language models. *Preprint*, arxiv:2311.16079 [cs].

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. *Preprint*, arxiv:2305.14314 [cs].

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392. Association for Computational Linguistics.

Johann Frei and Kramer Frank. 2023. Annotated dataset creation through large language models for non-english medical NLP. *Journal of Biomedical Informatics*, 145:104478.

Johann Frei, Ludwig Frei-Stuber, and Frank Kramer. 2022. GERNERMED++: Transfer learning in german medical NLP. *Preprint*, arxiv:2206.14504 [cs].

Johann Frei and Frank Kramer. 2023. German medical named entity recognition model and data set creation using machine translation and word alignment: Algorithm development and validation. *JMIR Formative Research*, 7(1):e39077.

Aryo Pradipta Gema, Luke Daines, Pasquale Minervini, and Beatrice Alex. 2023. Parameter-efficient finetuning of LLaMA for the clinical domain. *Preprint*, arxiv:2307.03042 [cs].

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. *Preprint*, arxiv:1904.05342 [cs].

Suramya Jadhav, Abhay Shanbhag, Sumedh Joshi, Atharva Date, and Sheetal Sonawane. 2024. Maven at MEDIQA-CORR 2024: Leveraging RAG and medical LLM for error detection and correction in medical notes. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 374–381. Association for Computational Linguistics.

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

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. *Preprint*, arxiv:2310.06825 [cs].

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arxiv:2401.04088 [cs].

Madeleine Kittner, Mario Lamping, Damian T Rieke, Julian Götze, Bariya Bajwa, Ivan Jelas, Gina Rüter, Hanjo Hautow, Mario Sänger, Maryam Habibi, Marit Zettwitz, Till de Bortoli, Leonie Ostermann, Jurica Ševa, Johannes Starlinger, Oliver Kohlbacher, Nisar P Malek, Ulrich Keilholz, and Ulf Leser. 2021. Annotation and initial evaluation of a large annotated german oncological corpus. *JAMIA Open*, 4(2):ooab025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. *Preprint*, arxiv:2309.06180 [cs].

Lisa Kühnel and Juliane Fluck. 2022. We are not ready yet: limitations of state-of-the-art disease named entity recognizers. *Journal of Biomedical Semantics*, 13(1):26.

Mohamed Yassine Landolsi, Lobna Hlaoua, and Lotfi Ben Romdhane. 2023. Information extraction from electronic medical documents: state of the art and future research directions. *Knowledge and Information Systems*, 65(2):463–516.

Reinhard Larsen. 2016. Präoperative Einschätzung und Prämedikation. *Anästhesie und Intensivmedizin für die Fachpflege*, pages 26–35.

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. *Bioinformatics*, 36(4):1234–1240.

Manuel Lentzen, Sumit Madan, Vanessa Lage-Rupprecht, Lisa Kühnel, Juliane Fluck, Marc Jacobs, Mirja Mittermaier, Martin Witzenrath, Peter Brunecker, Martin Hofmann-Apitius, Joachim Weber, and Holger Fröhlich. 2022. Critical assessment of transformer-based AI models for german clinical notes. *JAMIA Open*, 5(4):ooac087.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. ChatDoctor: A medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge. *Preprint*, arxiv:2303.14070 [cs].

Ignacio Llorca, Florian Borchert, and Matthieu-P. Schapranow. 2023. A meta-dataset of german medical corpora: Harmonization of annotations and cross-corpus NER evaluation. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 171–181. Association for Computational Linguistics.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

OpenAI. 2022. Introducing ChatGPT. https://openai.com/index/chatgpt/.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599.

Ralph Peeters and Christian Bizer. 2024. Entity matching using large language models. *Preprint*, arxiv:2310.11244 [cs].

Björn Plüster. 2023. LeoLM: Igniting german-language LLM research | LAION. https://laion.ai/blog/leo-lm.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *Nature Communications*, 15(1):8384. Publisher: Nature Publishing Group.

Phillip Richter-Pechanski, Philipp Wiesenbach, Dominic M. Schwab, Christina Kiriakou, Mingyang He, Michael M. Allers, Anna S. Tiefenbacher, Nicola Kunz, Anna Martynova, Noemie Spiller, Julian Mierisch, Florian Borchert, Charlotte Schwind, Norbert Frey, Christoph Dieterich, and Nicolas A. Geis. 2023. A distributable german clinical corpus containing cardiovascular clinical routine doctor's letters. *Scientific Data*, 10(1):207.

Roland Roller, Laura Seiffe, Ammer Ayach, Sebastian Möller, Oliver Marten, Michael Mikhailov, Christoph Alt, Danilo Schmidt, Fabian Halleck, Marcel Naik, Wiebke Duettmann, and Klemens Budde. 2022. A medical information extraction workbench to process german clinical text. *Preprint*, arxiv:2207.03885 [cs].

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350. Association for Computational Linguistics.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models. *Preprint*, arxiv:2305.09617 [cs].

Jackson Steinkamp, Jacob J. Kantrowitz, and Subha Airan-Javia. 2022. Prevalence and sources of duplicate information in the electronic medical record. *JAMA network open*, 5(9):e2233348.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

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. *Preprint*, arxiv:2307.09288 [cs].

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2023. Clinical text summarization: Adapting large language models can outperform human experts. *Preprint*, arxiv:2309.07430 [cs].

Dinithi Vithanage, Chao Deng, Lei Wang, Mengyang Yin, Mohammad Alkhalaf, Zhenyu Zhang, Yunshu Zhu, and Ping Yu. 2025. Adapting generative large language models for information extraction from unstructured electronic health records in residential aged care: A comparative analysis of training approaches. *Journal of Healthcare Informatics Research*, 9(2):191–219.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. GPT-NER: Named entity recognition via large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. PMC-LLaMA: Towards building open-source language models for medicine. *Preprint*, arxiv:2304.14454 [cs].

A Dataset Descriptions

	Instruction Data			
	GPTNERMED (Frei and Frank, 2023)	**GERNERMED** (Frei et al., 2022)	**i2b2 2010** (Uzuner et al., 2011)	**n2c2 2018** (Henry et al., 2020)
Textual Elements				
Language	GER	GER	EN	EN
Segments	18,683	24,091	16,701	19,276
Tokens	76,347	109,011	133,799	516,723
Entities				
Diagnosis	4,756	-	-	-
ADE	-	-	-	867
Drug	7,889	6,633	-	14,617
Problem	-	-	6,327	-
Reason for Drug	-	-	-	3,457
Treatment	-	-	4,368	-

	Evaluation Data		
	BRONCO (Kittner et al., 2021)	**GGPONC2** (Borchert et al., 2022)	**CARDIO:DE** (Richter-Pechanski et al., 2023)
Document Type	Discharge Summaries	Guidelines	Doctor Letters
Textual Elements			
Documents	200	30	500
Sentences	11,434	78,090	
Tokens	89,942	1877,100	993,143
Entities			
ActiveIng	-	-	7,580
Diagnosis	5,245	-	-
Diagnostic	-	27,829	-
Diagnosis|Pathology	-	81,380	-
Drug	2,013	19,478	2,093
Other Finding	-	51,376	-
Therapeutic	-	61,034	-
Treatment	2,013	-	-

Table 5: Overview of datasets and entities used for instruction fine-tuning, training baseline models and evaluation. ADE is short for Adverse Drug Events.

B Additional Analysis

B.1 Hallucinations

Model	Match	Fuzzy Match	Hallucination
peft Mistral 7B	1178	2	0
peft LLAMA2 7B	613	6	0
peft Meditron 7B	1248	6	1
peft Leo LLAMA2 7B	1120	5	3
Leo LLAMA2 7B chat	846	16	36
Mistral 7B Instruct	1186	27	76
LLAMA2 7B chat	596	43	151

Table 6: Count of entities generated by the LLMs when prompted to extract diagnoses, medications and treatments for the BRONCO test samples. An entity is categorized as a match if it is identical to a phrase in the input. If no exact match is found but the entity has a low Levenshtein distance to an input phrase, it is counted as a fuzzy match. Entities that cannot be matched to any input phrase are categorized as hallucinations

B.2 RAG Information Extraction

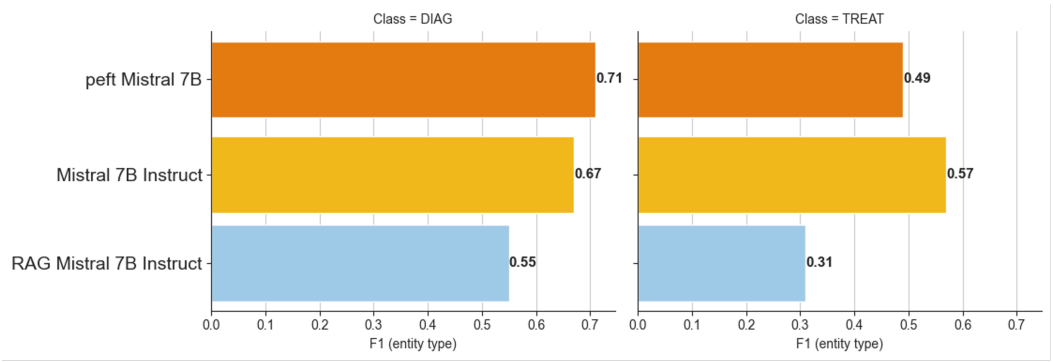

Figure 6: Results on the test BRONCO dataset for different versions of the Mistral 7B LLM. The chat version Mistral 7B Instruct and the fine-tuned peft Mistral 7B are compared to the retrieval-augmented generation approach RAG Mistral 7B Instruct. The F1 score is presented on entity type level.

B.3 Significance Tests

Figure 7: Pairwise McNemar significance heatmaps for entity-level F1 scores across all test datasets and entities, as well as micro-average. Each cell shows the F1-score difference between model i (row) and model j (column). Colors indicate statistical significance based on McNemar's test ($p < 0.05$) based on span-level entity comparisons using strict matching criteria. Green: model i significantly better than j, Red: model j significantly better than i, Gray: no significant difference.