On-Premise Medical Information Extraction from German Doctor's Letters under Clinical Constraints

Phillip Richter-Pechański

Department of Computational Linguistics
Heidelberg University

This dissertation is submitted for the degree of
Doctor of Philosophy

Heidelberg University April 2026

Supervisor: Prof. Dr. Anette Frank
Second supervisor: Prof. Dr. Christoph Dieterich

Submission date: 08.12.2025

Acknowledgements

First and foremost, I would like to express my sincere gratitude to Prof. Dr. Anette Frank for the opportunity to work with her and to grow both as a researcher and as a person. As fellow computational linguists working at the intersection of computational linguistics, medicine and medical informatics we faced unfamiliar clinical data, strict regulatory constraints, and evolving clinical requirements. Her high scientific standards, steady guidance, and patience in bridging these disciplines helped me turn challenging practical problems into transparent research questions and study designs. I also learned a great deal from her precise scientific communication, both in style and in substance.

My heartfelt thanks go to my PI and second reviewer, Prof Dr. Christoph Dieterich, for making this PhD possible alongside my work as a research associate at the DieterichLab, and for steadily supporting my development as an independent researcher. His confidence in my scientific abilities, his constructive mentorship, and his patience throughout this project were invaluable. He also enabled access to the time, infrastructure, and data needed for this work. He helped connect me with clinical partners, so that our NLP projects always remained grounded in real-world clinical routine.

I am grateful to the members of Prof Frank's research group for their support and valuable scientific input. My particular thanks go to: Letiţia Pârcălăbescu, Moritz Plenz, Frederick Riemenschneider and Fabian Strobel. Although I worked primarily remotely and could not always participate in the day-to-day life of the group, they made me feel like a full member at all times. Our regular colloquium and reading group enriched me personally and scientifically. A special thanks to Moritz Plenz, who patiently setup remote systems and ensured I could reliably join all meetings and discussions. I also thank all colleagues at the DieterichLab for their generous and valuable feedback throughout my PhD time.

I am deeply grateful to our clinical collaborators: Prof. Dr. Nicolas Geis and his colleagues Dr. Christina Kiriakou and Dr. Dominic M. Schwab. Their clinical expertise and commitment were indispensable. Their insights from clinical routine gave our work both scientific depth and clinical relevance. Our joint meetings were essential for developing NLP use cases that matter for research and routine. Despite demanding schedules, they supported this project with patience and reliable commitment. Without their collaboration, neither

the publication of CARDIO:DE nor the subsequent experimental results would have been possible.

My warmest thanks goes to my wife, Dr. Paulina Richter-Pechanska. She encouraged me to return to academia. With her steady belief in me, I began this journey in computational linguistics, and she stood by me through every step. While I admire her keen scientific mindset and professionalism, I am first of all grateful to have her and our two wonderful children in my life, who patiently endured the many moments of "Daddy is at the computer".

Finally, I thank my family and especially my mother. Sadly, she cannot witness the completion of my PhD. As a single parent in a divided Berlin, she raised four children with love, patience, and steady support. She helped me become who I am today, and for that I am endlessly grateful.

Abstract

A vast amount of German clinical data continues to be stored in unstructured doctor's letters. To make these data available for clinical routine and research, this thesis develops and rigorously evaluates on-premise methods for medical information extraction (MIE) from these letters, converting free text documents into transparent structured data. Automatic extraction systems must be developed and deployed entirely inside the clinical infrastructure and produce trustworthy outputs. At project start, there was no distributable German corpus and only CPU resources available. With the availability of mid-class GPUs, considering best performance-efficiency trade-offs, our approach progressed from supervised encoders to prompt-tuned encoders and finally PEFT-optimized LLMs. Throughout this thesis, we address strict real-world clinical constraints: limited domain expertise, staff time, compute resources, native-language barriers, and strong transparency requirements.

In the first part, we introduce CARDIO:DE, the first distributable German clinical routine corpus containing 500 de-identified cardiology doctor's letters with two high-quality annotation layers (paragraph-level section classes and token-level medication information). The corpus was collected and prepared entirely inside the clinical infrastructure, thus provides a study template for other clinics, and supports transparent and reproducible research in German clinical NLP. We used the corpus along with strong baselines as the foundation data for all experiments in this thesis.

In the second part, as mid-class GPUs became available, we evaluated prompt-tuned encoders for multi-class section classification on CARDIO:DE using pattern-exploiting training (PET). We systematically compare general-domain German BERTs (110M, 340M parameter) with domain/task-adapted and clinical variants. Domain- and task-adapted models consistently outperform general-domain and clinical models in few-shot settings. PET outperforms traditional supervised encoders with only 20 shots. Using a larger encoder and adding context further closes the gap to full-data supervision. We combine PET with efficient prompting and contextualization to reduce domain expertise and staff time demands in a clinical native language environment. Shapley value attributions support training data selection and error analysis, improving transparency. Under clinical constraints, compact encoders are sufficient for most section classes, while larger encoders are supportive for

complex sections. Further-pretraining on local texts is beneficial for general-domain encoders but not clinical ones. Overall, PET is a resource-efficient, interpretable method for native-language section classification.

In the third part, as token-level tasks exceeded capabilities of prompt-tuned encoders and more advanced GPUs became available, we define medication information extraction as a one-step end-to-end task that extracts medication mentions and links each to further attributes (strength, frequency, reason, etc.). We fine-tune open-source Llama models (8b, 70b) with parameter-efficient methods and format-restricting prompts on English and German (CARDIO:DE) corpora and compare against zero-shot and encoder baselines. A feedback LLM supports validation of uncertain predictions. Llama 70b achieves a new state of the art on English and provides the first benchmark for German. Llama 8b offers the best performance-efficiency trade-off. PEFT with format-restriction reduces hallucinations and malformed outputs and simplifies evaluation. Shapley attributions reveal input contributions to structured output. Overall, our approach minimizes expert/staff time demands, keeps compute demand modest, and improves transparency.

Finally, we deploy the pipeline on unseen German data in two clinical applications: (i) detecting the expected guideline-driven shift in oral anticoagulation from vitamin K antagonists (VKA, e.g. phenprocoumon) to direct oral anticoagulants (DOACs, e.g. apixaban) between 2012 and 2021 (DOACs: 16.9% to 59.9%, VKAs 37.7% to 9.9%), (ii) quantifying polypharmacy in longitudinal letters (2008-2016) from a 20-patient cohort, where 75% of letters list > 5 and 44% list > 10 distinct medications and at patient level, 80% have ever exceeded > 10 medications, often for years. Our findings show that our on-premise MIE models generalize to unseen letters and can support downstream clinical analysis.

Under strict on-premise and transparency constraints, we evaluate evolving NLP methods on real-world German and English data and derive a resource-aware guideline for MIE: use prompt-tuned, further-pretrained encoders for native-language section classification and PEFT-optimized, format-restricted LLMs for complex token-level tasks. Combine both with Shapley-based attributions and feedback LLMs to support transparency and evaluation. In a clinical environment, our models generalize to unseen letters, recover guideline-driven anticoagulation shifts and quantify letter- and patient-level polypharmacy, indicating clinical applicability. We expect that the contributions presented in this thesis will foster on-premise, transparent clinical NLP research in a lower-resource setting and support the development of reliable MIE systems.

Table of contents

Chapter 1

Introduction

1.1 Motivation

A vast amount of clinical information is still stored in unstructured documents. Information about patient medication history, longitudinal comorbidities, and physician reasoning remains unexplored for large-scale documentation in free text. A powerful medical information extraction (MIE) system could (1) feed registries and decision-support systems with real-time data, (2) provide researchers with suitable patient cohorts, and (3) relieve physicians from repetitive documentation tasks. Achieving this goal, however, requires methods that can be trained and deployed entirely inside clinical firewalls, work with minimal annotated data, and produce transparent and trustworthy outputs.

In 2017, I began my bachelor's thesis in computational linguistics within the cardiology department at Heidelberg University Hospital. The computational infrastructure was limited to an aging Linux workstation without internet access. All the model weights, Python libraries, and generally all the useful tools I got used to during my studies were not immediately accessible. Instead, they required individual on-demand installation by a local information technology (IT) administrator. Furthermore, a set of unordered doctor's letters were stored on an external hard drive. No gold-standard annotations were available for any fine-tuning or even evaluation experiments. Due to privacy regulations I had no possibility to easily recruit experienced natural language processing (NLP) colleagues from my former institute to conduct annotation projects. Furthermore, the doctor's letters were full of clinical jargon and abbreviations that were initially difficult to understand.

Following the successful completion of a local de-identification project employing conditional random fields (CRF), heuristics and small long short-term memory networks (LSTM), resulting in two publications (Richter-Pechanski et al. 2018; Richter-Pechanski et al. 2019), in 2020 my principal investigator opted to purchase our initial batch of middle-class graphics

processing units (GPU). This enabled me to use recently emerged transformer-based pre-trained bidirectional encoder representations from transformers (BERT) for a first medical information extraction project. To create our own in-house gold standard data, we engaged volunteer assistant physicians from our department and conceptualized and conducted laborious yet educational manual annotation projects that paved the way for all our subsequent NLP projects presented in this thesis (Richter-Pechanski et al. 2021). These experiences revealed two persistent challenges, that are typical for clinical NLP tasks (Richter-Pechanski et al. 2024; Richter-Pechanski et al. 2025):

1. **On-premise resource constraints** All NLP developments must remain inside the hospital firewall.

    1.a. **Domain expertise** Physicians must supply manual annotations and clinical routine expertise.

    1.b. **Staff time** Privacy regulations prohibit outsourcing. All experiments need to be conducted by local clinical staff in a secure infrastructure.

    1.c. **Local compute resources** Privacy regulations prohibit cloud computing. Several parameter-efficient fine-tuning and inference methods have emerged, but local compute restrictions still remain challenging. Most experiments must be performed inside the secure clinical firewall.

    1.d. **Native-language barrier** Doctor's letters are written in native languages. Hence, most language models need careful language adaptation.

2. **Transparency requirements** Clinical routine is a safety-critical domain. Model predictions need to be as transparent and comprehensible as possible.

In recent years, we have also observed several significant methodological shifts in NLP, each offering advancements but introducing new challenges for MIE tasks in a clinical context. At the beginning encoder-based pre-trained language models, such as BERT, achieved state-of-the-art performance in various MIE tasks. However, these models required a substantial amount of annotated data for fine-tuning (cf. challenges 1.a. *Domain expertise* and 1.b. *Staff time*) and, considering the computational resources available at that time, demanded considerably more computing power than earlier methods (cf. challenge 1.c. *Local compute resources*). Furthermore, their pre-training data contained a significant English language bias, requiring careful language adaptations (cf. challenge 1.d. *Native-language barrier*). From 2021 on we observed another shift to a pretrain-then-prompt paradigm, where information extraction tasks are formulated using natural language prompts. However, as outlined by Liu et al. (Liu et al. 2023), although we observed a strong performance improvement

with the use of significantly less annotated data for text classification tasks, more complex token-level MIE tasks, such as named entity recognition (NER) or relation extraction (RE), remained challenging (cf. challenges 1.a. *Domain expertise* and 1.b. *Staff time*). By 2023, large language models (LLM) became the new state-of-the-art for almost all NLP tasks. However, the continually increasing size of the models presented a significant challenge (cf. challenge 1.c. *Local compute resources*). Furthermore, the risk of hallucinations and the need to consistently generate structured outputs crucial for MIE tasks (cf. challenge 2. *Transparency requirements*), along with the emerging difficulties in applying current interpretability techniques (cf. challenge 2. *Transparency requirements*), made the integration of these models into MIE setups a challenging task.

To empirically study this intersection of evolving language model architectures and paradigms, special demands on suitable MIE methods and clinical constraints, we propose to use German as the exemplary lower-resource language and doctor's letters from the cardiology department as the lower-resource document type and domain. By demonstrating how each NLP method: supervised encoders, prompt-tuned encoders and generative LLMs, can be adapted to work entirely inside a clinical firewall while meeting strong transparency requirements, this thesis aims to serve as an empirically grounded, process-oriented guideline for applying MIE projects in lower-resource languages and domains.

1.2 Research Questions

Considering evolving language model architectures and paradigms and strict clinical constraints, such as the need to develop and deploy completely behind a clinical firewall and the demand for transparent predictions, this thesis raises five broad research questions (RQ):

1. **On-premise model adaptation under clinical constraints** How can evolving model architectures and optimization paradigms be applied and deployed entirely inside the clinical firewall, given native-language input, strict data-protection regulations and limited in-house compute resources?

2. **Resource-aware model choice across MIE task complexity** Across various MIE tasks, ranging from text classification to complex structured information extraction (NER+RE), which model architecture and optimization paradigm offers the best balance between performance and resource efficiency in lower-resource clinical setting?

3. **Reducing manual effort during model development and evaluation** To what extent can manual annotation by local clinical staff and evaluation workloads be minimized,

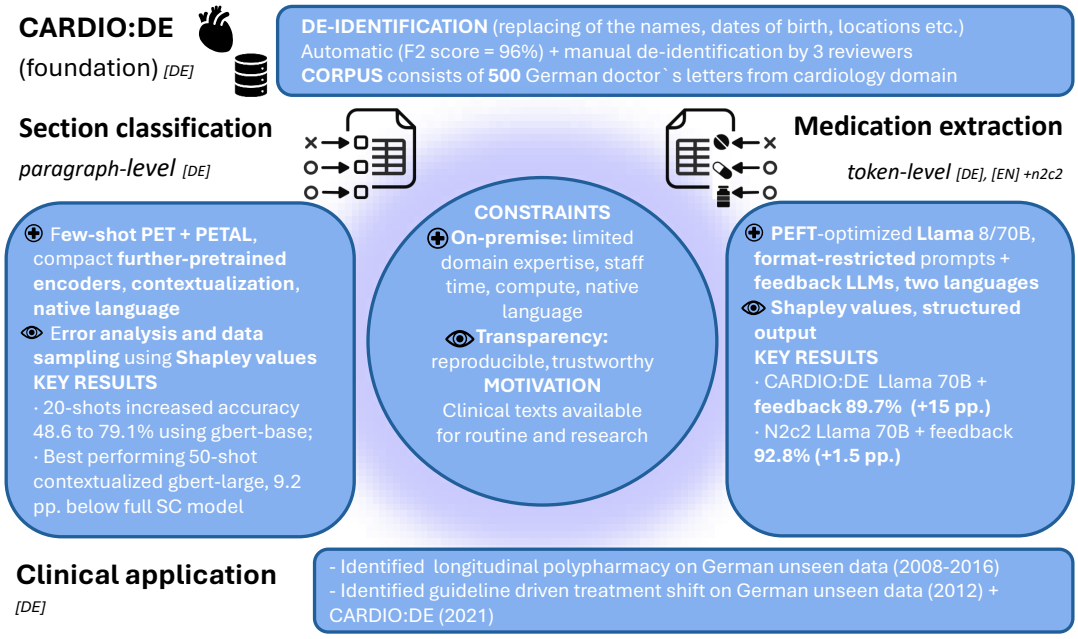


Fig. 1.1 **Thesis overview:** This thesis develops and rigorously evaluates on-premise methods for MIE from German doctor's letters and evaluates their usage under real-world clinical constraints.

moving from zero-shot, few-shot to fully supervised training, by careful prompt engineering, and automatic feedback mechanisms?

4. **Transparency, interpretability and trust** Can interpretability methods reveal model reasoning, prediction errors, and support data sampling and enhance clinical trust, especially for generative LLMs?

5. **Clinical applicability for real-world clinical routine data** How well does our medication information extraction system generalize to unseen German clinical routine doctor's letters and how effectively can it support downstream analyses such as identifying guideline-driven medication treatment shifts and patient-level polypharmacy occurrences?

1.3 Thesis Outline and Contributions

The thesis aims to demonstrate how evolving NLP paradigms can be applied under real-world clinical constraints (cf. graphical abstract Figure 1.1). In Chapter 2 and 3 we discuss the relevant background and related work for this thesis.

In Chapter 4, we outline our foundational work on conceptualizing, developing, and distributing the first German clinical routine corpus (CARDIO:DE) consisting of doctor's letters from the cardiology department (Richter-Pechanski et al. 2023). This corpus enables collaborative and reproducible research in German clinical NLP (cf. challenges 2. *Transparency* and 1.d. Native language). Our prospective study setup complies with current data protection regulations and allowed us to keep the original structure of the documents. To make the data accessible while complying with European data protection regulations, we thoroughly manually de-identified all letters. In close collaboration with physicians, we added two carefully curated annotation layers to CARDIO:DE: (1) medication information and (2) section labels compliant to the clinical document architecture (CDA). Both are directly used in the following chapters and address the challenges of limited domain expertise (cf. 1.a. *Domain expertise*) and restricted staff time for manual annotation (cf. 1.b. *Staff time*).

In Chapter 5 we present a systematic study of prompt-tuned encoders for multi-class section classification in German doctor's letters using annotations of CARDIO:DE (Richter-Pechanski et al. 2024). We compare general-domain German BERT models with medical-adapted variants using task- and domain-adaptation, to reduce the manual demand for domain expertise (cf. 1.a. *Domain expertise* and 1.d. *Native-language barrier*). We apply pattern-exploiting training (PET), at that time the state-of-the-art for prompt-tuning in a few-shot learning setup and evaluate null-prompts and contextualized prompts to further reduce manual annotation and system development time to address challenges 1.a. *Domain expertise* and 1.b. *Staff time*. All experiments used encoders with less than 500 million parameters, emphasizing small models with less than 150 million parameters to reduce the demand of compute resources (cf. 1.c. *Local compute resources*). Finally, we apply Shapley values to optimize few-shot training data selection, prompt strategies and provide more interpretable model predictions (cf. 2. *Transparency requirements*). Altogether, these experiments show that task- and domain-adapted German BERT encoders achieve comparable performance using few-shot learning in comparison to traditional fully-supervised encoders.

While prompt-tuned encoders showed promising results for text classification tasks, they struggled with more complex token-level tasks. Hence, in Chapter 6 we present a systematic study of generative LLMs for end-to-end joint NER+RE medication information extraction for English and German doctor's letters (Richter-Pechanski et al. 2025). We compare a general-purpose foundation Llama model (8 billion and 70 billion parameters) with a domain-specific OpenBioLLM with 8 billion parameters and examine zero-shot prompting versus fine-tuning on gold-standard CARDIO:DE data to reduce the need for in-house domain expertise (cf. challenge 1.a. *Domain expertise* and 1.d. *Native-language*

barrier). Format-restricting prompts and a post-hoc feedback LLM keep predictions in a well-defined javascript object notation (JSON) schema, supporting automatic evaluation to lower manual evaluation efforts (cf. challenge 1.b. *Staff time*). To reduce compute demands we used parameter-efficient low-rank adaptation (LoRA) fine-tuning and quantization in all experiments (cf. challenge 1.c. *Local compute resources*). Moreover, Shapley values, which we previously utilized in our encoder experiments, demonstrated the ability of LLMs to identify entity relationships and capture clinical knowledge, thereby enhancing trust in these models (cf. challenge 2. *Transparency requirements*). Overall, while the parameter-efficiently fine-tuned Llama with 70 billion parameters established a new state-of-the-art performance for complex medication information such as adverse drug events and medication reason, the smaller 8 billion model achieved results that were only marginally lower while remaining highly IT-friendly.

In Chapter 7 we apply fine-tuned medication information extraction models on a set of unseen German doctor's letters to evaluate LLMs in a real-world clinical scenario, assessing their performance and applicability in practical clinical routine settings. We therefore selected two datasets:

- A temporal cohort (2012 vs. 2020/21) to quantify the guideline-driven shift from Vitamin-K antagonists to direct oral anticoagulants, an evolution documented in the literature but time-consuming to track manually on a larger scale in doctor's letters (Mekaj et al. 2015).

- A longitudinal cohort containing successive doctor's letters per patient to flag polypharmacy whenever more than five distinct active medications are listed, a condition strongly associated with heart-failure readmission (Delara et al. 2022; Unlu et al. 2020).

This chapter demonstrates that our MIE models return robust results on unseen data (Elangovan et al. 2024) to support statistically meaningful analysis of medication trends. Extracting such data is not trivial, as medication information is often stored in free text and the selected letters span a wide time range, thus differ in style and terminology.

We finalize the thesis by discussing and concluding our findings and contributions with respect to research questions 1-5 considering our clinical constraints and the evolving NLP model and optimization paradigms over time. Finally, we outline future work perspectives to further support MIE integration into clinical routine and research. In summary, our main contributions include:

1. We demonstrate end-to-end (e2e), on-premise fine-tuning and deployment of supervised encoders, prompt-tuned encoders and quantized LoRA LLMs in a limited clinical setting (RQ 1: *On-premise model adaptation*).

2. We provide a clear empirical comparison of model architectures and optimization paradigms across different MIE task complexities, showing when lightweight encoders are sufficient and when generative LLMs are essential (RQ 2: *Resource-aware model choice*).

3. We offer comprehensive guidelines to minimize manual efforts for model training and evaluation via few-shot learning, prompt engineering and automatic feedback evaluation LLMs. (RQ 3: *Reducing manual efforts*)

4. We implement and evaluate a set of transparency-enhancing strategies, including format-restricting-prompts, data sampling, deterministic decoding and token-level Shapley value analysis (RQ 4: *Transparency*).

5. We validate real-world utility by applying the best-performing medication information extraction model to unseen clinical routine doctor's letters and quantify guidelines-driven treatment shifts and polypharmacy occurrence on patient level (RQ 5: *Asserting clinical application*).

Unless stated otherwise, novelty claims in this thesis refer to the specific experimental setup of German clinical routine data, cardiology doctor's letters, on-premise clinical constraints, and the respective task settings.

1.4 Own published work

This dissertation is based on the following publications:

- **Richter-Pechanski, P.**, Wiesenbach, P., Schwab, D. M., Kiriakou, C., He, M., Allers, M. M., Tiefenbacher, A. S., Kunz, N., Martynova, A, Spiller, N., Mierisch, J., Borchert, F., Schwind, C., Frey, N., Dieterich, Ch., & Geis, N. A. (2023). A distributable German clinical corpus containing cardiovascular clinical routine doctor's letters. *Scientific Data*, 10(1), pp. 207.

- **Richter-Pechanski, P.**, Wiesenbach, P., Schwab, D. M., Kiriakou, C., Geis, N. A., Dieterich, C., & Frank, A. (2024). Clinical information extraction for lower-resource languages and domains with few-shot learning using pretrained language models and prompting. *Natural Language Processing*, pp. 1-24.

- **Richter-Pechanski, P.**, Seiferling, M., Kiriakou, C., Schwab, D. M., Geis, N. A., Dieterich, C., & Frank, A. (2025). Medication information extraction using local large language models. *Journal of Biomedical Informatics*, pp. 169.

Additional own related works which will be discussed in this thesis:

- **Richter-Pechanski, P.**, Riezler, S., & Dieterich, C. (2018). De-identification of German medical admission notes. In *German Medical Data Sciences: A Learning Healthcare System*, pp. 165-169.

- **Richter-Pechanski, P.**, Amr, A., Katus, H. A., & Dieterich, C. (2019). Deep learning approaches outperform conventional strategies in de-identification of German medical reports. In *German Medical Data Sciences: Shaping Change–Creative Solutions for Innovative Medicine*, pp. 101-109.

- **Richter-Pechanski, P.**, Geis, N. A., Kiriakou, C., Schwab, D. M., & Dieterich, C. (2021). Automatic extraction of 12 cardiovascular concepts from German discharge letters using pre-trained language models. *Digital health*, pp. 1-10.

- Becker, M., Krumscheid, M., Knobelspies, A., Seydel, M., **Richter-Pechanski, P.**, & Karl, A. (2025). Extending CARDIO: DE: Additional annotation guidelines and evaluation of NLP approaches for clinical applications. *International Journal of Medical Informatics*, pp. 1-7.

- Kindermann, A., Tute, E., Benda, S., Löpprich, M., **Richter-Pechanski, P.**, & Dieterich, C. (2021). Preliminary analysis of structured reporting in the HiGHmed use case cardiology: challenges and measures. In *German Medical Data Sciences: Bringing Data to Life*, pp. 187-194.

In this thesis, I use the scientific *we* to report on the work. Chapter 2.6.1 is based on my substantial contributions to de-identification experiments (Richter-Pechanski et al. 2018) and (Richter-Pechanski et al. 2019). Chapter 2.6.2 is based on my work on cardiovascular concept extraction in (Richter-Pechanski et al. 2021).
In Chapter 4 I present my work about distributing a German clinical routine corpus, published in (Richter-Pechanski et al. 2023). My contributions to this research are significant. I substantially conceptualized and administered this project. I was responsible for the project's methodology, validation, machine learning experiments, formal analysis, investigation, data collection and curation and writing the paper.
Chapter 5 is based on my work on section classification presented in (Richter-Pechanski et al. 2024). I contributed to data collection and curation, machine learning experiments, formal

analysis, investigation, project conceptualization and administration. Moreover, I played a key role in documentation and writing the paper.
In Chapter 6 I present our medication information extraction results presented in (Richter-Pechanski et al. 2025). I substantially contributed to data collection and curation, machine learning experiments, formal analysis, investigation, project conceptualization and administration. Moreover, I played a key role in documentation and writing the paper.

Chapter 2

Background

2.1 Clinical NLP: Promises and challenges

The exponential growth of textual data, ranging from doctor's letters to pathology reports in the present healthcare sector, represents an invaluable resource for both, clinical routine and research, which, however, remains mostly unused (Sheikhalishahi et al. 2019). Manual extraction of these data by clinical experts is time-consuming and tedious. The complexity and heterogeneity inherent in this unstructured information present significant obstacles for its automatic extraction (Wang et al. 2018).

However, recent advances in pre-trained language models (PLM) and LLMs showed promising results on various NLP tasks (Alsentzer et al. 2019; Singhal et al. 2023). Thus, development of efficient NLP pipelines has the potential to

- be integrated in an automatic decision support system. After a doctor's letter is stored in a clinical information system, up-to-date diagnosis and medication information could automatically be integrated into clinical registries and trigger medication safety or guideline alerts without delay (Afshar et al. 2023),
- be used to automatically create patient cohorts for clinical research based on free-text clinical documents. This could replace manual screening of these documents and significantly decrease trial recruitment efforts. Furthermore, implementing such a system would add numerous clinical parameters to clinical research, currently only available in free text (Ghosh et al. 2025),
- would help to support documentation efforts from clinicians by offering real-time auto-complete or auto-fill functions during documentation (Perkins et al. 2024).

However, while we have seen an evolving development in model architectures and optimization paradigms on the one hand, and considering the linguistic complexities of clinical texts on the other hand, we identified two systemic challenges for clinical NLP pipelines (Richter-Pechanski et al. 2024; Hahn et al. 2020):

1. **On-premise resource constraints** All NLP development, training and inference must be executed *inside* the clinical firewall. No clinical text can leave the clinical storage.

    a. **Domain expertise** Only local clinicians, specifically physicians, can interpret clinical routine documents and create representative gold-standard annotations. Their time is costly and limited, and each annotation session competes with their obligations in patient care (Tamang et al. 2023). Hence, reducing the amount of annotations for model development is essential.

    b. **Staff time** Privacy regulations prohibit outsourcing of workload to external experts. Hence, every data preparation, model fine-tuning, or error analysis must be performed by local clinical staff within the clinical infrastructure. This limits the staff resource and expertise to a relatively small number of in-house employees. Furthermore, cross-site projects with multiple collaborators can only be conducted after a thorough data protection process involving a careful manual de-identification and a positive ethics vote from each site (Richter-Pechanski et al. 2024). Hence, reducing the time for manual annotation, model development and deployment is crucial.

    c. **Local compute resources** The usage of external high-performance clusters or cloud GPUs from popular commercial vendors are prohibited due to privacy constraints. On-premise compute resources often do not provide any GPU resources at all or only limited capacity shared by multiple projects. Any NLP model architectures must therefore be compute- and memory-efficient (Taylor et al. 2024).

    d. **Native-language barrier** Doctor's letters, outside of the English-speaking countries are typically written in a native language, such as German and contain institution specific abbreviations and terminology. Publicly available open-source models often perform poorly on these data and need careful language and domain adaptation before they become usable (Névéol et al. 2018).

2. **Transparency requirements** Clinical decision-making is a safety-critical domain. Any automatic predictions must be reproducible and transparent. Therefore, NLP models require transparency, deterministic decoding strategies when feasible, and thorough post-hoc quality validation (Hanif et al. 2021).

These challenges, ranging from linguistic complexities of clinical texts, strict data protection regulations, practical limitations of on-premise development and deployment and the demand for model transparency, emphasize the unique and multifaceted challenges that define clinical routine NLP. As a result, advances in this field demand not only sophisticated algorithmic development and a sound adaptation of existing methods to the clinical environment and tasks but also a profound understanding of these real-world constraints to ensure robust and high-performing NLP solutions in the clinical domain.

The remainder of this chapter further discusses core NLP methods and aspects within the fast-changing field of NLP, along with the new challenges each evolution introduces. These concepts are essential to address the constraints highlighted here and provide the foundation for the comparative analysis and thorough evaluation in the subsequent experimental chapters that follow in this thesis, each explicitly aimed at overcoming these constraints. In addition, Subsection 2.6 presents two fundamental preparatory works for the distribution of CARDIO:DE (cf. Chapter 4) and for the experiments in Chapters 5 and 6.

2.2 Characteristics of clinical texts

Clinical routine texts vary significantly from open-domain texts. They contain dense domain-specific terminology (e.g. *Hypertrophe obstruktive Kardiomyopathie (Septumdicke max. 18 mm, TTE v. 01.01.2000)*), non-standardised and locally specific abbreviations (e.g. *Cvrf.* for *cardiovascular risk factors*, *TTE o. B.* for *transthorakale Echokardiographie ohne Befund*), and a fragmented, telegraphic writing style (e.g. *Z. n. NSTEMI — 08/23: PTCA DES RCA Seg3*) (Névéol et al. 2018; Leaman et al. 2015). Furthermore, outside the English-speaking world, clinical documents are written in native languages, limiting the immediate applicability of common publicly available NLP models, typically pre-trained primarily on English data, (Névéol et al. 2018) and intensifying the native-language barrier in clinical NLP (cf. 1.d. *Native languages*).

German cardiology doctor's letters (cf. Chapter 4) follow a standardized but heterogeneous global structure with institutional and author-specific variations in section headings and section ordering (Richter-Pechanski et al. 2023). Frequent sections (with header variants) include: introduction (*Wir berichten über Ihren Patienten . . .*); diagnoses (*Diagnosen, Aktuelle Diagnosen*); cardiovascular risk factors (*Kardiovaskuläre Risikofaktoren, Cvrf, Kvrf*); anamnesis (*Anamnese*); admission medication (*Medikation bei Aufnahme, Aufnahmemedikation*); physical examination (*Körperlicher Untersuchungsbefund, KuB*); diagnostic results (*Echobefund, EKG-Befund, Elektrophysiologischer Befund*); laboratory results (*Labor, Laborwerte*); discharge medication (*Medikation bei Entlassung, Entlassmedikation,*

Therapieempfehlung); and conclusion / hospital course (*Zusammenfassung, Epikrise, Zusammenfassende Beurteilung*).

Content types vary considerably across different sections. Sections such as *anamnesis* or *epicrisis* contain free text, while *diagnosis, results, laboratory values, medication* often contain semi-structured content ranging from comma- or whitespace-separated lists to specific key-value patterns or copied tabular data. Medication descriptions often mix entity and relation information in a one-liner notation (e.g. *ASS 100 mg 0-0-1*) and often contain additional free-text information. Other relation information such as medication reason or adverse drug events frequently appear only in free-text sections (*diagnosis*, *conclusion*). Many doctor's letters are created by copy-pasting from templates or content of different clinical systems and then manually edited, which introduces additional structural inconsistencies.

This structural and semantic heterogeneity makes rule- or heuristic-based document parsing cumbersome and unreliable: section headers vary over time, abbreviations change locally, and structural conventions alter between authors and time periods. Hence we aim (i) to develop a robust section classification model (cf. Chapter 5) to accurately categorize paragraphs in doctor's letters into normalized section types, and (ii) to build an end-to-end medication information extraction model (cf. Chapter 6) to extract medications and their relation information from clinical texts. To support a reproducible development (cf. challenge 2. *Transparency*) considering constraints of domain expertise, staff time and native language (cf. challenges 1.a. *Domain expertise*, 1.b. *Staff time* and 1.d. *Native languages*), we created the German CARDIO:DE corpus (cf. Chapter 4), providing manually de-identified, German cardiology letters including section and medication information annotations used throughout this thesis.

2.3 Core information extraction tasks

Information extraction in NLP aims to extract implicit or explicit information from free text and store the results in a structured format. Common information extraction tasks, relevant for this thesis, cover: text classification, named-entity recognition, and relation extraction (Otter et al. 2021).

2.3.1 Text classification

Text classification or document classification is a fundamental and essential task in NLP (Li et al. 2022a) with the aim to assign a fixed set of classes to a set of documents based on their content. Given a document d and a fixed label set C, in text classification a model is

trained to learn a function $f : d \mapsto C$. Model performance is usually measured by accuracy. Popular text classification tasks encompass sentiment classification, spam detection, or news categorization (Jurafsky et al. 2025).

Early systems were dependent on feature-engineering, bag-of-words approaches and linear classifiers (Joachims 1998). With the rise of deep learning, methods such as convolutional networks (CNN) using pre-trained sentence-level embeddings (Kim 2014) and hierarchical attention networks (Yang et al. 2016) reduced feature engineering and improved performance. Pre-trained encoders such as BERT further increased classification results while reducing manual annotation efforts (Devlin et al. 2019). Prompt-based fine-tuned encoders, reformulate input examples as cloze-style phrases (e.g. PET), allowed similar results in few-shot scenarios (Schick et al. 2021a). Most recent instruction-tuned LLMs like generative pre-trained transformer (GPT)-3 or Llama now reach competitive zero- and few-shot classification by converting the classification task into a generative task, further reducing the demand for manual annotation (Brown et al. 2020; Touvron et al. 2023).

In clinical NLP text classification is used in various aspects, e.g. automatic ICD coding, phenotyping/cohort identification or dividing doctor's letters into semantically consistent sections. This thesis focuses on the latter (Mujtaba et al. 2019). The identification of sections within clinical texts has demonstrated improvements for various medical information extraction tasks (Pomares-Quimbaya et al. 2019). Nonetheless, the progress in this research area has been limited, in part, due to the absence of standardized benchmark datasets (Landolsi et al. 2023). Most studies focus on English clinical data (Denny et al. 2008). Our pioneer experiments described in Chapter 5 are the first to thoroughly investigate the task of section classification on a freely available German clinical corpus (cf. Chapter 4) using prompt-based fine-tuned encoders.

2.3.2 Named entity recognition

A named entity (NE) is anything that can be described with a proper name, e.g. PERSON, LOCATION, ORGANIZATION (Jurafsky et al. 2025). The task of NER involves identifying spans of text that include named entities and categorizing each with its corresponding entity class. NER can be formalized as a sequence labeling task, where entity labels $y_1, \ldots, y_n$ are assigned to a token sequence $x_1, \ldots, x_n$, typically using an inside–outside–beginning (IOB) scheme (Jurafsky et al. 2025). Training is conducted by maximizing the conditional likelihood of the gold label sequence. Performance is usually reported using token-level or entity-level F_β-scores (Li et al. 2022a).

Early automatic systems combined manually constructed gazetteers with CRFs (cf. CoNLL 2003 shared task (Tjong et al. 2003); CRF for de-identification cf. Section 2.6.1

(Richter-Pechanski et al. 2018)). Early deep learning approaches used feed-forward neural networks with sliding windows (Collobert et al. 2011) and later bidirectional LSTM networks with a CRF head (Zhai et al. 2018). Contextual string embeddings (e.g. Flair, ELMo) further improved NER performance in multiple languages (Akbik et al. 2019; Peters et al. 2018) (cf. Section 2.6.1 (Richter-Pechanski et al. 2019)). Pre-trained encoder-based transformers brought further improvement in performance, especially for lower-resource languages and domains (Lee et al. 2020; Alsentzer et al. 2019; Bressem et al. 2024) (cf. Section 2.6.2 (Richter-Pechanski et al. 2021)). Recent developments in instruction-tuned generative LLMs achieve the current state-of-the-art in zero- and few-shot learning setups (Brown et al. 2020; Touvron et al. 2023).

NER is a crucial task in information extraction, serving as the foundation for various downstream NLP tasks such as sentiment analysis or relation extraction (Jurafsky et al. 2025), and is particularly crucial in clinical NLP applications (Bose et al. 2021). While traditionally NER focused on general entities such as persons or locations, it can be adapted to domain-specific demands, particularly in clinical contexts. In Subsection 2.6.2 we define NEs as medical concepts (e.g. angina pectories, dyspnea etc.). For our medication information extraction experiments (cf. Chapter 6) we extend the definition of named entities to all medication names (DRUG) and their relation information, e.g. FREQUENCY or REASON.

2.3.3 Relation extraction

Once NER is completed, RE identifies and classifies the association between entity pairs, such as PERSON $\rightarrow$ LOCATION or DRUG $\rightarrow$ FREQUENCY (Jurafsky et al. 2025). Formally, RE predicts a relation class $r \in \mathscr{R}$ for two entities (e_i, e_j). RE performance is usually evaluated using accuracy or micro-average F_1 score over correctly classified and ordered relation pairs. RE is essential in structuring free text into relational databases or knowledge graphs, such as the Unified Medical Language System (UMLS) (Bodenreider 2004) or drug information databases (Martin et al. 2004).

Early approaches used handcrafted patterns and support vector machines (SVM) (Zelenko et al. 2002; Zhou et al. 2005). Neural methods introduced CNNs and bi-directional LSTMs typically combined with a preliminary NER step (Zhang et al. 2017; Zeng et al. 2014). Transformer-based encoders improved RE results by using pre-trained BERT and incorporating information from target entities (Wu et al. 2019). Universal information extraction frameworks use a unified text-to-structure generation approach, combining NER and RE under a single generative objective to achieved a new state of the art (SOTA) for fully supervised and few-shot scenarios (Lu et al. 2022). Most recently generative LLMs achieved new SOTA results on various relation extraction tasks (Zhou et al. 2024).

In Chapter 6 we follow unified information extraction approaches and adopt an end-to-end NER+RE approach by fine-tuning a generative LLM to output medication entities and their relation information in a single JSON object to tackle both entity recognition and relation extraction in one pass to reduce development time and post-processing efforts.

2.4 Model architectures and optimization paradigms

This section introduces architectures and optimization paradigms that serve as the basis of the experiments in Section 2.6 and in Chapters 5 and 6, focusing on how each step in the evolution of NLP intersects with our clinical constraints defined in Section 2.1. We assume that the reader understands the basic concept of neural networks and transformers (Vaswani et al. 2017).

2.4.1 Pre-trained encoder models

Pre-trained language models based on transformer encoders such as BERT (Devlin et al. 2019) established a new *pretrain-then-finetune* paradigm. First, a transformer is pre-trained on a large amount of unlabeled text data using a masked-language modeling objective, rather than starting with a random initialization of a deep learning model. This is followed by a task-specific fine-tuning step on a usually significantly smaller amount of labeled data. However, PLMs still demand a considerable amount of annotated data for effective fine-tuning, particularly in lower-resource domains (cf. 1.a. *Domain expertise* and 1.b. *Staff time*). Moreover, because they contain a large number of learnable parameters, they demand substantially more computational resources than earlier NLP approaches. (cf. challenge 1.c. *Local compute resources*). Additionally, as pre-training data is primarily in English, careful language adaptations are crucial (cf. challenge 1.d. *Native-language barrier*). In this thesis we investigate three variants of pre-training scenarios:

- **General-domain PLMs .** While vanilla BERT models achieve strong performance on English-language general domain tasks, they often struggle on specialized clinical and German-language text, motivating domain- and language-adaptation (Gururangan et al. 2020; Sun et al. 2019).
- **Medical PLMs .** Domain-adapted PLMs (e.g. BioBERT (Lee et al. 2020), Clinical-BERT (Alsentzer et al. 2019), or medBERT.de (Bressem et al. 2024)) are either based on (i) models pre-trained entirely from scratch on medical text or (ii) general-domain PLMs further-pretrained on medical text. Chapter 5 demonstrated empirically that,

in lower-resource domains, further-pretraining a general-domain German PLM (e.g. gBERT (Chan et al. 2020)) on clinical routine text achieved higher performance than medical PLMs exclusively pre-trained on clinical text.

- **Further-pretraining strategies.** We investigate three distinguish PLM pre-training strategies on general and medical PLMs (i) domain-adaptation (further-pretraining on large amounts of clinical-domain text), (ii) task-adaptation (further-pretraining on smaller amounts of task-specific text), and (iii) the combination of these approaches (Gururangan et al. 2020). Chapter 5 empirically compares these pre-training variants for our section classification task and puts them into context to our clinical constraints (cf. 1. *On-premise resource constraints*).

2.4.2 Prompt-based fine-tuning of encoders

Fine-tuning PLMs often still requires a significant amount of manually labeled data (Gao et al. 2021) making their usage in the clinical domain challenging due to limited domain expertise, staff time and compute resources (cf. challenges 1. *On-premise resource constraints*).

The *pretrain-then-prompt* paradigm reformulates classification tasks as cloze questions (Shin et al. 2020; Schick et al. 2021a; Gao et al. 2021). The rationale behind this method was to align the pre-training objective with the downstream objective, thereby decreasing the quantity of training data required. While traditional supervised learning trains a model to predict a label y conditioned on x as $P(y|x)$, prompt-based learning is based on language models that model the probability of text directly. The original input x is converted into a textual prompt x' containing an unfilled slot. A language model then predicts the highest probability token to fill the slot, from which the final label can be derived (Liu et al. 2023).

PET was the state-of-the-art for text classification in a few-shot learning scenario. Formally, PET requires a PLM M with vocabulary V, a few-shot dataset with training instances and labels $\mathscr{S} = \{(x_i, y_i)\}$, a pattern function that maps instances to a set of cloze sentences (templates) $P : X \rightarrow V^*$, and a verbalizer $v : Y \rightarrow V$ that maps each label to a single token from the vocabulary of M. The PET workflow contains three basic steps (see Fig. 2.1):

1. **template fine-tuning** applying P to each input instance x_i and fine-tune a model M for each template to obtain the most likely token for the *MASK* token $v(y)$,

2. **semi-supervised labeling** use the ensemble of fine-tuned models M from the previous step and annotate a large unlabeled dataset D with soft labels and

3. **train a classifier** train a final classifier C with a traditional sequence classification head on the labeled dataset D

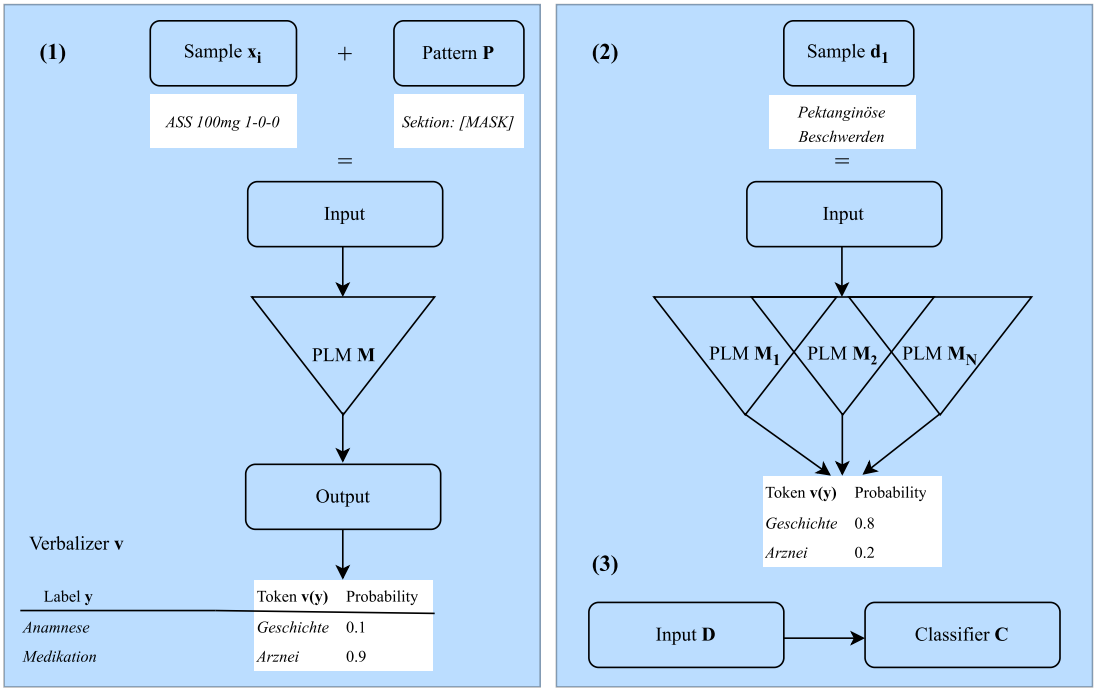


Fig. 2.1 **PET workflow**: Three main steps: (1) Apply pattern function *P(x)* to all few-shot training instances *X*. Fine-tune a PLM *M* using a language model objective on each pattern. The output of the PLM is mapped using a verbalizer function *v(y)*. (2) An ensemble of *M* trained on each pattern is used to annotate an unlabeled dataset *D* with soft labels. (3) A classifier *C* with a classification head is trained on *D*. For a more detailed explanation, see Section 5.3.1. Figure adapted from (Richter-Pechanski et al. 2024).

This three-step approach directly supports RQ 3 and tackles several key clinical constraints: (i) the demand for clinical expertise and staff time is reduced by few-shot learning and semi-supervision while achieving results comparable to fully supervised baselines and (ii) the demand for compute resources is minimized because fine-tuning is conducted with small training data and compact encoders. In Chapter 5 we adopt PET for German section classification, proving how PET meets our clinical constraints while achieving competitive results.

However, as discussed by Liu et al. (Liu et al. 2023), while performance improved in few-shot learning setups for text classification tasks, more complex token-level MIE tasks like NER or RE were difficult to implement (cf. challenges 1.a. *Domain expertise* and 1.b. *Staff time*).

2.4.3 Generative large language models

The latest advancements in NLP are marked by generative LLMs. These models are based on an autoregressive decoder with hundreds of billions of parameters pre-trained on large-scale corpora with a *predict-next-token* language model objective. After pre-training, these models have the capability to perform downstream tasks based on user instructions: The user submits a text prompt and the model then generates the desired answer. The generative abilities of LLMs have considerably broadened the spectrum of applications for NLP, particularly their conversational capabilities allow for more intuitive question answering (QA) interactions between, e.g. physicians and LLMs in the form of a dialogue. However, the continuously increasing size of NLP models posed significant challenges (cf. 1.c. *Local compute resources*). In addition, issues with hallucinations and the need for reliable structured outputs for MIE tasks (cf. 2. *Transparency requirements*), and incompatibilities with well-established interpretability methods (cf. challenge 2. *Transparency requirements*), made the deployment of these models into MIE systems a challenging task.

The Evolution of GPT Models GPT-2, a generative LLM with 1.5 billion parameters illustrated that pre-training a generative model on huge amounts of text data significantly improves their zero- and few-shot capabilities (Radford et al. 2019). GPT-3 was further scaled up in pre-training data and size (175 billion parameters) and significantly improved task-agnostic, few-shot performance, showing even new SOTA on various tasks (Brown et al. 2020) including a broad range of clinical QA benchmarks (Singhal et al. 2023). However, GPT-3 and all successor models are proprietary closed-source systems, unsuitable for local on-premise deployment as required by challenges 1. *On-premise* and lack transparency as required by challenge 2. *Transparency*.

Towards open-source LLMs The release of open-source models such as Llama, Mistral or DeepSeek facilitated advancements in local LLM application development and research (Touvron et al. 2023; Jiang et al. 2023; DeepSeek-AI et al. 2024) due to their applicability to be deployable on-premise. However, their performance in the clinical domain were limited. Quantization and parameter-efficient fine-tuning (PEFT) methods, such as LoRA, enabled further pre-training and fine-tuning of local LLMs on constrained compute infrastructure, resulting in the release of various domain-adapted LLMs (Yang et al. 2023; Huang et al. 2023; Wu et al. 2024). LoRA is a PEFT method which freezes the LLM's weights and injects trainable rank decomposition matrices into each transformer layer. This reduces the number of trainable parameters for fine-tuning and enables efficient task-switching by requiring only small, task-specific parameter updates. This makes LoRA a valuable method to deploy LLMs in resource-constrained environments, such as the clinical domain (Hu et al. 2021). quantized low-rank adapters (QLoRA) further reduces memory usage during fine-tuning and extends LoRA by quantizing precision of the weight parameters of a LLM to 4-bit precision (Dettmers et al. 2023) allowing even a 70b Llama to be fine-tuned on a single middle class GPU, conforming to challenge 1.c. *Compute restrictions*.

Local LLMs in the medical domain The development of PEFT strategies to adapt local LLMs under clinical constraints resulted in the release of several medical-domain-adapted LLMs, including PMC-Llama, Meditron or OpenBioLLM (Chen et al. 2023; Ankit Pal et al. 2024). Several studies showed that fine-tuned, expert LLMs frequently surpass the performance of larger, general LLMs on various medical tasks (Lehman et al. 2023; Lehman 2024; Xu et al. 2024). However, the majority of publications restricted evaluation on English medical QA datasets (cf. MedQA (Jin et al. 2021), MedMCQA (Pal et al. 2022), PubMedQA (Jin et al. 2019)).

LLMs for information extraction (IE) tasks Publications evaluating LLMs on information extraction tasks are limited. One of the main challenges using generative LLMs, producing reliable structured output, is still under intensive research (Liu et al. 2024c; Liu et al. 2024a). Recent studies demonstrated that generative LLMs are capable of information extraction by generating structured output directly from unstructured text. Frameworks like unified information extraction (UIE) treat NER and RE as an end-to-end text-to-structure task, by instructing the model to generate a JSON representation using schema-based prompt mechanisms (Lu et al. 2022).

This end-to-end approach prevents error propagation between pipeline components. In our medication information extraction setup (cf. Chapter 6), we follow a similar approach using

format-restricting prompts by including well-defined PYDANTIC object definitions in the systems prompt.[1] Using PYDANTIC classes simplifies structure definition and maintenance supporting model transparency (challenge 2. *Transparency*) directly contributing to RQ 4 (Richter-Pechanski et al. 2025).

2.4.4 Using feedback LLMs as post-hoc validators in information extraction

Unlike traditional classification models, LLMs generate a sequence of tokens rather than fixed labels. To measure the performance of a model, simply comparing the gold standard and the prediction, often results in artificial false positives/negatives. For instance, representation differences such as list versus concatenated strings or minor differences in abbreviations and units can result in mismatching despite being clinically equivalent. In clinical NLP these mismatches increase manual evaluation efforts (cf. challenges 1.b. *Staff time*, 1.a. *Domain expertise*) and obscures the actual model performance (cf. challenge 2. *Transparency*).

Recent studies address this challenge using various LLMs-as-a-judge paradigms (cf. comprehensive survey (Gu et al. 2024)). Feedback LLMs act as validators over gold and predicted instances to avoid performance penalties coming from harmless formatting or phrasing differences. The feedback LLM is not replacing the primary metrics, such as F_1-score or accuracy, but complements it as a post-hoc validator to better reflect the actual LLM performance. Feedback LLMs helped to speed up the often complex evaluation of LLM outputs in our medication information extraction task (cf. Chapter 6) and helped identify frequent false positives and negatives that were actually correct. Furthermore, they supported our findings that LLMs surpass current SOTA methods.

2.5 Interpretability methods across different model architectures

In safety-critical domains, like clinical routine, it is crucial to understand why a model made a prediction (faithfulness) and evaluate how convincing an explanation is to human experts (plausibility) (Jacovi et al. 2020). Saliency-based local explanations can increase trust by indicating which tokens supported or contradicted a decision and, when errors occur, by helping diagnose their causes.

[1] https://pydantic.dev/articles/llm-intro, accessed 04.12.2025.

Shapley values in NLP

In recent years, Shapley values became a valuable tool in NLP for local text explanations (Attanasio et al. 2023). They offer a systematic approach to attribute the influence of individual textual components (token, token sequences) on a model prediction. Shapley values originate from cooperative game theory, allocating the importance of each input feature by averaging its marginal contribution across all possible feature combinations in predicting an outcome. Formally, for feature i with model f,

$$\phi_i(f) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\left(|N| - |S| - 1\right)!}{|N|!} \Big[f(S \cup \{i\}) - f(S) \Big], \tag{2.1}$$

where N is the set of features and S ranges over subsets of $N \setminus \{i\}$ (Lundberg et al. 2017).

In all our experiments, we use SHapley Additive exPlanations (SHAP) because it offers an optimized algorithm that approximates Shapley values with reduced computational costs (cf. challenge 1.c. *Compute restrictions*), making its application feasible for practical use (Mosca et al. 2022). Furthermore, we conducted experimental explorations and compared several interpretability methods in advance with FERRET, a framework for benchmarking popular explainers on transformers, finding that SHAP was the best-performing method for our setup (Attanasio et al. 2023).

Shapley values for encoders and generative LLMs

Encoders For encoder-based classifiers (e.g., BERT-style sequence or paragraph classification), we compute Shapley values with respect to the target label score probabilities. In our study on prompt-tuned encoders for multi-class section classification in German doctor's letters (cf. Chapter 5), Shapley values were used for two complementary purposes: (i) from a clinical perspective: to make deep learning model predictions more transparent and explainable (cf. challenge 2. *Transparency*) and (ii) from an engineering perspective: to detect biases or errors in the training data and to support choosing the most optimal model architecture (cf. challenge 1.a. *Domain expertise* and 1.b. *Staff time*) (Richter-Pechanski et al. 2024).

Generative LLMs For generative LLMs, predictions are sequences of text (e.g., JSON), hence we do not have a single target class per input. In our medication information extraction experiments (cf. Chapter 6) we use interpretability functionalities that are specifically designed to analyze the behavior of generative LLMs introduced in Captum v. 0.7 (Miglani et al. 2023). This implementation offers various perturbation-based attribution methods and

saliency map visualizations, including Shapley values. The method generates Shapley values of each input token to each generated output token. Finally, we provide further insights in model performance by presenting two use-cases using Shapley values to increase model interpretability (cf. challenge 2. *Transparency*): (1) assessing the contributions of input tokens to relation information output tokens, (2) uncovering implicit knowledge on relation information.

2.6 Foundation work

In the following section, we briefly introduce two foundation works that were essential for processing clinical texts under strict data-protection regulations and to derive important lessons how to extract clinically relevant information: (i) de-identification and (ii) cardiovascular concept extraction from German doctor's letters. Both works provide the basis for the distribution of CARDIO:DE (cf. Chapter 4) and the advanced experiments presented in Chapter 5 and 6.

2.6.1 De-identification of German doctor's letters

Doctor's letters contain highly sensitive personal information, such as patient name or contact information. These documents can only be used for research without legal restrictions if tokens containing personal information, e.g. protected health information (PHI), had been removed by de-identification (Schlünder 2015; Richter-Pechanski et al. 2019). This Subsection is based on our experiments presented in (Richter-Pechanski et al. 2018) and (Richter-Pechanski et al. 2019).

State-of-the-art

While there were numerous publications on the de-identification of English medical texts (Sweeney 1996; Wellner et al. 2007; Uzuner et al. 2008; Liu et al. 2017), due to data protection regulations and administrative challenges only little research was done on de-identification of German texts (Starlinger et al. 2017). At the time of our experiments, CRFs had been the best performing method in major English de-identification competitions, rule-based methods remained widely used and only a few deep learning methods emerged (Yogarajan et al. 2018). This motivated our approach to first leverage well-established CRF and rule-based methods, and then introduce SOTA deep learning models once our compute infrastructure was updated and minimal manual annotation works were conducted.

Practically, at the beginning of this research in clinical NLP, I was faced with an administrative chicken-and-egg problem. Processing of clinical documents was prohibited until all PHI was removed. However, removing the PHI required the documents to be processed in the first place.

Data

To overcome this, we had to perform a de-identification task in accordance with Section 46 Abs.2 Nr.2a (Landeskrankenhausgesetz) and Section 13 Abs.1 Landesdatenschutzgesetz BW. In this context, we had the possibility to use clinical data for the purpose of optimizing internal clinical procedures. Thus, we closely collaborated with physicians who selected and annotated a small set of doctor's letters. As only a single physician could serve as an expert annotator, we could not follow best-practice annotation approaches including redundant annotation, interannotator-agreement scores and an iterative guideline adaptation process (cf. challenges 1.a. *Domain expertise*, 1.b. *Staff time*). The final corpus contained 113 doctor's letters containing 107,229 tokens, among which 5,221 consist of PHI data. We distinguished eight PHI classes: person (PER: *Max Mustermann*), location (LOC: *Musterstadt, Musterstr. 2*), date (DATE: *01. Sep 2000*), phone number (PHONE: *0123 456 34*), organizations (ORG: *Musterklinik*), titles (TITLE: *Dr., Prof.*), salutations (SALUTE: *Herr, Frau*), and postal codes (PLZ: *12345*). This dataset was the prerequisite for training and evaluation of all our de-identification methods.

Objective

Building a de-identification tool under clinical constraints that takes a plain text German doctor's letter as input and returns a de-identified letter, with PHI tokens replaced by their corresponding PHI classes. (e.g. *wir berichten über Ihren Patienten Herrn Mustermann* $\Rightarrow$ *wir berichten über Ihren Patienten SALUTE PER*)

Definitions

To use medical data in research, the consent of the patient is the legal basis preferred by law. In practice, most of the clinical data is stored without obtaining such explicit consent. Thus, most medical texts can only be analyzed without legal restrictions if personal data is completely removed (Starlinger et al. 2017; Schlünder 2015; Richter-Pechanski et al. 2018). There is no consistent terminology for the task of removing PHI data. De-identification, anonymization, and pseudonymization are frequently used in research. All terms are not consistently defined across different data protection regulations. Anonymized data can never

be re-identified, while pseudonymized data can only be re-identified with the use of external information. Hence, we use the broadest term *de-identification* for this thesis, which describes the process of detaching the association between a pre-defined set of identifying data from the data subject (Richter-Pechanski et al. 2019).

We focused on the technical aspects of de-identification, using the explicit PHI definitions from the health insurance portability and accountability act (HIPAA). The HIPAA is a quasi-gold standard in de-identification research (Yogarajan et al. 2018) and also used for non-US data due to imprecise definitions in the European general data protection regulation (GDPR) (for further info, see (Richter-Pechanski et al. 2019)). In general, any productive de-identification task needs to be closely supervised by legal advice.

Methods

Three-step approach Due to a lack of sufficient training data and local high performance computing resources at project start in 2018, we implemented a three-step approach for this task. This allowed us to utilize both well-established rule-based de-identification methods and SOTA statistical machine learning methods.[2]

1. *Spelling variant detection* The header of a doctor's letter mostly contains contact information of the patient, the recipients of the note and the clinic. We are using the approach of (Kester et al. 2016), using the minimum edit distance, to compare the tokens in the medical text for spelling variants present in the header.

2. *CRF* Due to the lack of NER CRFs trained on medical texts, we identified Stanford NLP's NER CRF implementation as the best performing NER tool for the recognition of person and location information trained on out-of-domain data (Faruqui et al. 2010; Finkel et al. 2005) (further details: https://github.com/MaviccPRP/ger_ner_evals). A CRF is a statistical machine learning method which, by default, considers context information for each provided input sample unlike other statistical methods such as support vector machines or multilayer perceptrons (for further details about the CRF, see (Lafferty et al. 2001)). To calculate probabilities for each label of a given token the CRF uses a predefined set of feature functions (for further information, see (Richter-Pechanski et al. 2019)).

3. *Regular expressions and gazetteers* The rule-based approach utilizes the Stanford Core NLP RegexNER implementation, a software component for NER inference that uses custom regular expression patterns defined in gazetteers. We combined plain gazetteers

[2]The code is available here: https://github.com/MaviccPRP/Anonymizer, accessed 04.12.2025.

containing lists of names and cities with gazetteers containing German street names extended with regular expressions (Manning et al. 2014).

Deep-learning approach After we successfully presented our three-step approach at a German medical informatics conference (GMDS 2018, (Richter-Pechanski et al. 2018)) we had the opportunity to store clinical texts on a high performance computing central processing unit (CPU) cluster. Furthermore, we finalized our annotated corpus of 113 doctor's letters, enabling us to use the data for model training. Although it remained impractical to fine-tune a SOTA BERT model, it became possible to fine-tune smaller deep learning architectures. Hence, we could include bidirectional LSTM networks. LSTMs are specialized in handling variable length sequence data and belong to the category of recurrent neural networkss (RNN), which take a token per time step as input and keep the information from the previous time steps as context information (Richter-Pechanski et al. 2019). We use German deep contextualized word embeddings from language models (ELMo) and character encoded word embeddings to encode the input text. ELMo, represent an early class of pre-trained language models. They use a bidirectional LSTM architecture, to represent not just the syntactic and semantic properties of the token, but also its context-specific meaning (Peters et al. 2018; Che et al. 2018). All ELMo embeddings were additionally concatenated with character encoded word embeddings to represent token specific properties on character level and to deal with unknown words, often occurring in clinical NLP (Ling et al. 2015). The de-identification model consists of five components, illustrated in Figure 2.2. The ELMo layer maps each token to an ELMo represented embedding vector. The character embedding layer first maps each character to a character embedding and then outputs a single vector to represent the sequence of characters of the given input token (Richter-Pechanski et al. 2019). A dropout layer gets as input a sequence of concatenated vectors of the ELMo and the character embeddings. Next, a bidirectional LSTM layer consisting of a forward and a backward layer takes as input a sequence of output vectors from the drop out layer. Finally, a fully connected layer gets as input the concatenated outputs of the LSTM layers and outputs a softmax vector containing the probabilities of each PHI class (Richter-Pechanski et al. 2019).[3]

Fully supervised CRF Similar to the deep learning model, we trained as well a CRF on our annotated data. For calculating label probabilities for a given token the CRF uses a predefined set of feature functions. A possible feature function for our task could be: *return true if a given token is capitalized.* Equation 2.2 shows how the CRF calculates the conditional

[3]The code is available here: https://github.com/dieterich-lab/clinical_deid, accessed 04.12.2025.

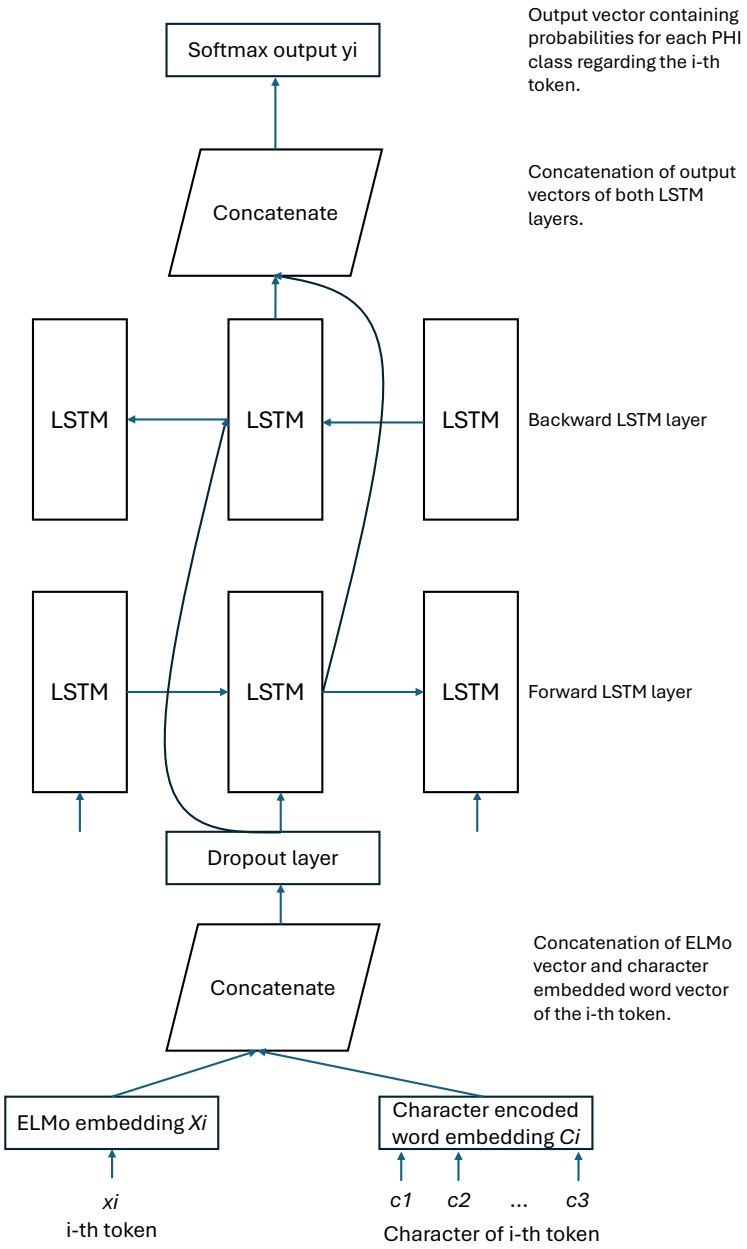


Fig. 2.2 **Neural de-id model:** Bidirectional LSTM architecture using ELMo and character encoded embeddings as input and a softmax classification layer as output.

Model	**Precision [%]**	**Recall [%]**	**F_2-score [%]**
three step approach (3S)	72	69.5	70
CRF	96.5	92	93
neural network (NN)	**97**	**95.5**	**96**

Table 2.1 **Binary de-identification:** Precision, recall and F_2-score for the binary evaluation of the 3S (Richter-Pechanski et al. 2018), CRF and the NN. Best score in bold.

probability of a label `y`, given a text sequence `X`. The first fraction is a normalization term to obtain probabilities. First, we sum over all tokens in the sequence of text and over all feature functions. λ is the weight of feature function `j` to be learned during training. Then we calculate a feature function based on the previous label, the current label the context tokens `X` and the current position of the i-th token. To calculate the most probable output sequence of labels given a text sequence the CRF uses equation 2.3. The CRF is trained using the log likelihood of a given training set. Our CRF is using the following set of features: lower-cased token, previous and following context token, last two and three letters of a token, capitalization, sentence boundaries, digits, part of speech tag of the current token and the first two letters of the part of speech tag. All features are commonly used in other de-identification tasks (Richter-Pechanski et al. 2019).

$$p(y \mid x; \lambda) = \frac{1}{Z(x)} \exp\left(\sum_{i=1}^{n} \sum_{j} \lambda_j f_j(y_{i-1}, y_i, X, i)\right) \tag{2.2}$$

$$\hat{y} = \arg\max_{y} \; p_\lambda(y \mid x), \tag{2.3}$$

Results

Metrics We report median score results using token-level precision, recall and F_2-score using 4-fold cross-validation. In a de-identification task false negatives are more critical than false positives. Hence, the F_2-score places greater weight on recall than on precision. Results for binary classification (PHI recognized vs. PHI not recognized) are reported in Tab. 2.1 and for multiclass classification in Tab. 2.2.

Baselines We compared the deep-learning approach with the training-free (zero-shot) three-step approach of (Richter-Pechanski et al. 2018) and with the fully supervised CRF.

	Precision			Recall			F_2-score		
NE	3S	CRF	NN	3S	CRF	NN	3S	CRF	NN
PER	29.5	98	97.5	81.5	93	98	61	94	**98**
LOC	56	96	96	78	86	92	72	87.5	**92**
DATE	93	98	98.5	91	92	96	91	93	**96.5**
PHONE	100	99	100	57.5	94.5	99	63	95.5	**99**
ORG	6.5	89	91.5	28.5	72.5	81.5	17	75	**82.5**
TITLE	99	94	94	70	97	96	74.5	**96.5**	96
SALUTE	97.5	98	97	67	96.5	98.5	72	99	**99.5**
PLZ	100	98.5	100	89	98	100	90.5	97	**100**

Table 2.2 **Multilcass de-identification:** Precision, recall and F_2-score for the multiclass evaluation of the 3S (Richter-Pechanski et al. 2018), CRF and the NN approach. Best score in bold.

Discussion

Binary evaluation shows that both the CRF and the NN significantly outperform the baseline (Tab. 2.1; *p-value 3S vs. CRF:* $p < 0.001$*, CRF vs. NN:* $p < 0.001$ *(McNemar test)*). CRF increases precision by almost 25 percentage points (pp.) and recall by more than 20 pp. NN reaches the best F_2-score of 95.5%. It shows a more balanced precision and recall score surpassing the CRF particularly in recall by 3.5 pp. (Richter-Pechanski et al. 2019).

In multiclass evaluation (Tab. 2.2) the CRF tops the baseline especially in precision, particularly for PER, LOC, ORG and DATE. For PLZ, SALUTE and PHONE both models have a comparable precision score. Only for the TITLE class the baseline outperforms the CRF by 5 pp. with 99% precision. Regarding recall scores, the CRF outperforms the baseline over all PHI classes (Richter-Pechanski et al. 2019).

Compared to the CRF the NN has similar or slightly better precision results, while it outperforms the CRF in seven of eight PHI classes in the more relevant recall score. From a data protection perspective it is less harming to de-identify a non-PHI token, than not to de-identify a PHI token. Recall of the NN for the ORG class is 9 pp. higher than recall of the CRF. Regarding the F_2-score, the NN tops the baseline and the CRF model over all PHI classes, except the TITLE class. The NN shows the most balanced ratio of precision and recall. The CRF performs slightly worse, while the 3S baseline has the least balanced ratio (Richter-Pechanski et al. 2019).

Deep learning improved performance, while maintaining comparable annotation efforts (cf. challenges 1.a. *Domain expertise*, 1.b. *Staff time*). Taking into account computational resources (cf. challenges 1.c. *Compute restrictions*), the average training time on our CPU

infrastructure per training split fold of the CRF was 13 seconds, while the NN was trained for a minimum of 92 minutes illustrating the significantly higher complexity of deep learning algorithms. In addition, CRFs, being log-linear prediction models, predict outputs that are easier to interpret (cf. challenge 2. *Transparency*). Therefore, it is essential to carefully assess the need for deep learning architectures for specific tasks and to balance performance gains against computational cost and interpretability, to ensure that the added complexity of deep learning is justified by the task requirements.

Conclusion and relevance for the thesis

Our work used SOTA supervised machine learning methods for the first time on German doctor's letters for the task of de-identification. Under clinical constraints regarding corpus creation (cf. challenges 1.a. *Domain expertise* and 1.b. *Staff time*) and supervised machine learning (cf. challenges 1.c. *Compute restrictions* and 2. *Transparency*) we showed that the CRF and the NN approach significantly outperformed the baseline in a binary and multiclass scenario using only a small annotated training corpus. While the CRF achieved strong F_2-scores across all entity classes, the NN achieved the best overall results, particularly in recall and the balance between precision and recall. Most importantly, the results enabled us to conduct further machine learning experiments in the medical domain by providing de-identified doctor's letters on a larger scale, essential in a safety-critical clinical environment.

Our experiments enabled us to conduct on-premise clinical NLP projects (RQ 1), and to reduce manual efforts by including larger automatically de-identified clinical corpora for modeling tasks (RQ 3). Finally, we could establish reproducible de-identification procedures that support transparency of our following experiments (RQ 4) including the distribution of a clinical corpus (cf. Chapter 4).

2.6.2 Cardiovascular concept extraction from German doctor's letters

In 2020, our cluster was upgraded with four NVIDIA RTX6000 GPUs. This addition allowed us, for the first time, to pre-train and fine-tune transformer-based encoders on a significantly larger scale. Hence, we evaluated the performance of these models for the first time on a real-world clinical use case. This Subsection is based on our experiments presented in (Richter-Pechanski et al. 2021).

In this study we evaluated fine-tuning of transformer-based PLMs based on the BERT architecture pre-trained on three different corpus types for the task of cardiovascular concept extraction (CCE) on limited training data. We performed our concept extraction task as a NER-based token-classification task, by assigning each token to a cardiovascular concept

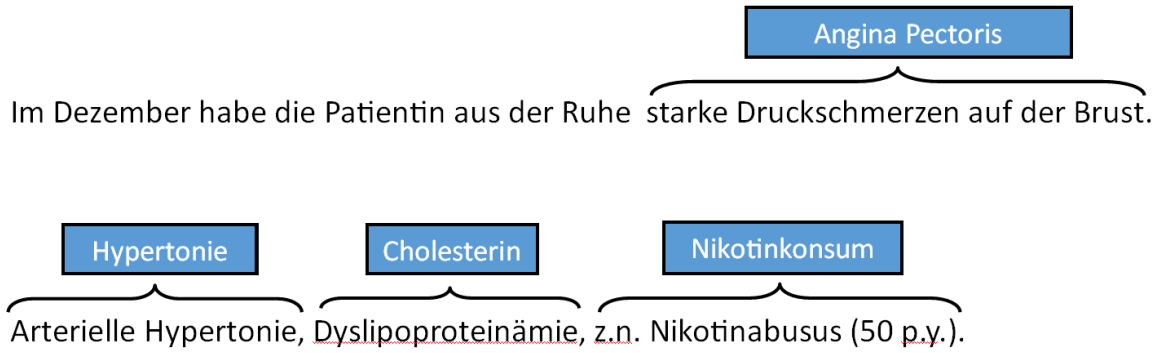


Fig. 2.3 **CCE example**: Doctor's letter snippet annotated with CCs. For example, the sequence *starke Druckschmerzen auf der Brust* is annotated with the concept ANGINA PECTORIS.

(CC) or to a negative class 'O' (cf. Figure 2.3). CCs, such as angina pectoris, dyspnea or edema are crucial in clinical routine but currently require physicians to manually search in doctor's letters. Most of these concepts are stored in unstructured text sections of doctor's letters like anamnesis or conclusion (cf. 2.6.2).

State-of-the-art

Extracting clinical information from unstructured texts was traditionally done via different rule-based and statistical machine learning methods (Long 2005; Turchin et al. 2006; Roller et al. 2018; Bashyam et al. 2005; Zheng et al. 2017; Wang et al. 2018). Later, deep learning methods primarily based on RNNs gained increasing popularity (Jagannatha et al. 2016b; Jagannatha et al. 2016a; Wu et al. 2015; Kittner et al. 2021). Most publications in the cardiovascular domain covered English data (Small et al. 2017; Nath et al. 2016; Patterson et al. 2017; Khalifa et al. 2015; Kaspar et al. 2019; Toepfer et al. 2015; Garvin et al. 2012; Chung et al. 2005; Mykowiecka et al. 2009) only a few used German texts (Kaspar et al. 2019; Toepfer et al. 2015). All publications on cardiovascular data used rule-based approaches.

Due to the lack of annotated clinical gold standard corpora for training, especially in the non-English language, pre-trained encoder-based transformers became more and more popular in clinical information extraction (Li et al. 2019; Beltagy et al. 2019; Si et al. 2019; Scheible et al. 2024; Sänger et al. 2019; Lee et al. 2020; Bressem et al. 2024; Richter-Pechanski et al. 2021).

However, these transformers were typically trained on general-domain data with a strong English bias. Some works further pre-trained these models on English biomedical or clinical text (Gururangan et al. 2020; Beltagy et al. 2019; Lee et al. 2020; Alsentzer et al. 2019). For German, only a handful of general-domain models (e.g., the *dbmdz* BERT family, *GottBERT*,

and *GBERT-large*) were available, and at that time a single model further-pretrained on German medical forum text (*German-MedBERT*) existed (Richter-Pechanski et al. 2021; Sänger et al. 2019).

To the best of our knowledge, there was no study available, evaluating pre-training and fine-tuning language models on an IE task on German doctor's letters from the cardiovascular domain (Richter-Pechanski et al. 2021). This motivated us to systematically compare the performance of three German BERT models pre-trained with different datasets and fine-tuned on manually annotated gold standard data under clinical on-premise constraints for the task of CCE.

Data

For our pre-training experiments we collected a large, in-house clinical corpus (called CardioComplete) containing 200,000 cardiology doctor's letters (2004–2020), containing 218,084,192 tokens automatically de-identified using our previously developed custom de-identification pipeline. For our fine-tuning experiments we selected 204 doctor's letters (called CardioAnno) using stratified sampling. The documents were manually annotated using redundant annotation, interannotator-agreement scores and an iterative guideline adaptation process (Wilbur et al. 2006; Roberts et al. 2009; Lohr et al. 2020). Two annotators (assistant physicians from cardiology) achieved a token-wise inter-annotator agreement using F_1-score of 89.8%. In total, all annotated letters contained 381,628 tokens (36,355 paragraphs) and 1,631 annotated concepts using a set of 12 CCs: ANGINA PECTORIS (AP), DYSPNOE, NYKTURIE, ÖDEME, PALPITATION, SCHWINDEL, SYNKOPE, HYPERTONIE, CHOLESTERIN, DIABETES MELLITUS (DM), FAMILIENANAMNESE (FA) and NIKOTIN.

Objective

Evaluating three pre-training approaches for German PLMs for the task of CCE: (i) fine-tuning a publicly available PLM on CCE gold-standard annotations, (ii) further-pretraining of a publicly available PLM on CardioComplete followed by fine-tuning and (iii) pre-training from scratch on CardioComplete and then fine-tuning. We conducted all experiments entirely on-premise (cf. 1. On-premise) contributing to RQ 1-3.

Methods

We define CCE as a NER task. We compare:

- Traditional baselines: a statistical machine learning method based on a CRF (Lafferty et al. 2001) using linguistic features proposed by Mikhail Korobov and a bidirectional

LSTM model (Hochreiter et al. 1997). For CRF details, see https://sklearn-crfsuite.readthedocs.io/en/latest/tutorial.html#features. For LSTM details, see (Richter-Pechanski et al. 2021).

- Three differently pre-trained BERT models:
  - $\text{BERT}_{\text{base}}$: a publicly available German BERT model (Chan et al. 2020), pre-trained on German Wikipedia, OpenLegalData and various news corpora. (Pre-training data size: $\approx 12GB$)
  - $\text{BERT}_{\text{fine}}$: $\text{BERT}_{\text{base}}$ further-pretrained on CardioComplete. (Pretraining data size: $\approx 12 + 2GB$)
  - $\text{BERT}_{\text{scratch}}$: randomly initialized BERT architecture pre-trained on CardioComplete. (Pretraining data size: $\approx 2GB$)

$\text{BERT}_{\text{fine}}$ and $\text{BERT}_{\text{scratch}}$ were pre-trained using the masked language modeling script based on the HuggingFace transformer library.[4] Training took $\approx$ 20h ($\text{BERT}_{\text{fine}}$) and $\approx 65h$ ($\text{BERT}_{\text{scratch}}$) on 4$\times$ RTX6000 (24GB video random-access memory (VRAM)).

After pre-training, the BERT model is fine-tuned as a supervised downstream task on annotated training data. In our task, we seek and classify phrases containing CCs. Each output vector of the BERT model is used as input to a feed-forward neural network with shared weights and a softmax layer as a final layer to classify each input token into our set of 12 concepts (cf. Figure 2.4) (Richter-Pechanski et al. 2021). Fine-tuning for CCE was performed using the HuggingFace NER script performing 30 epochs with a batch size of 16.[5] Fine-tuning time was $\approx$ 1h per fold on 2$\times$ RTX6000 GPUs (for further methodological details, cf. (Richter-Pechanski et al. 2021)).

Experimental setup and metrics

To evaluate our CC classifiers, including the baseline classifiers, we used identical 4-fold cross-validation splits on the CardioAnno corpus. We calculated token-wise F_1-score (the harmonic mean between precision and recall) per concept and a micro-average F_1-score per classifier. Furthermore, we conducted significance tests using approximate randomization (Koehn 2004; Padó 2006).

[4] https://github.com/huggingface/transformers/blob/v4.0.1-release/examples/language-modeling/run_mlm.py, accessed 10.09.2025.

[5] https://github.com/huggingface/transformers/blob/v4.0.1-release/examples/token-classification/run_ner_old.py, accessed 10.09.2025.

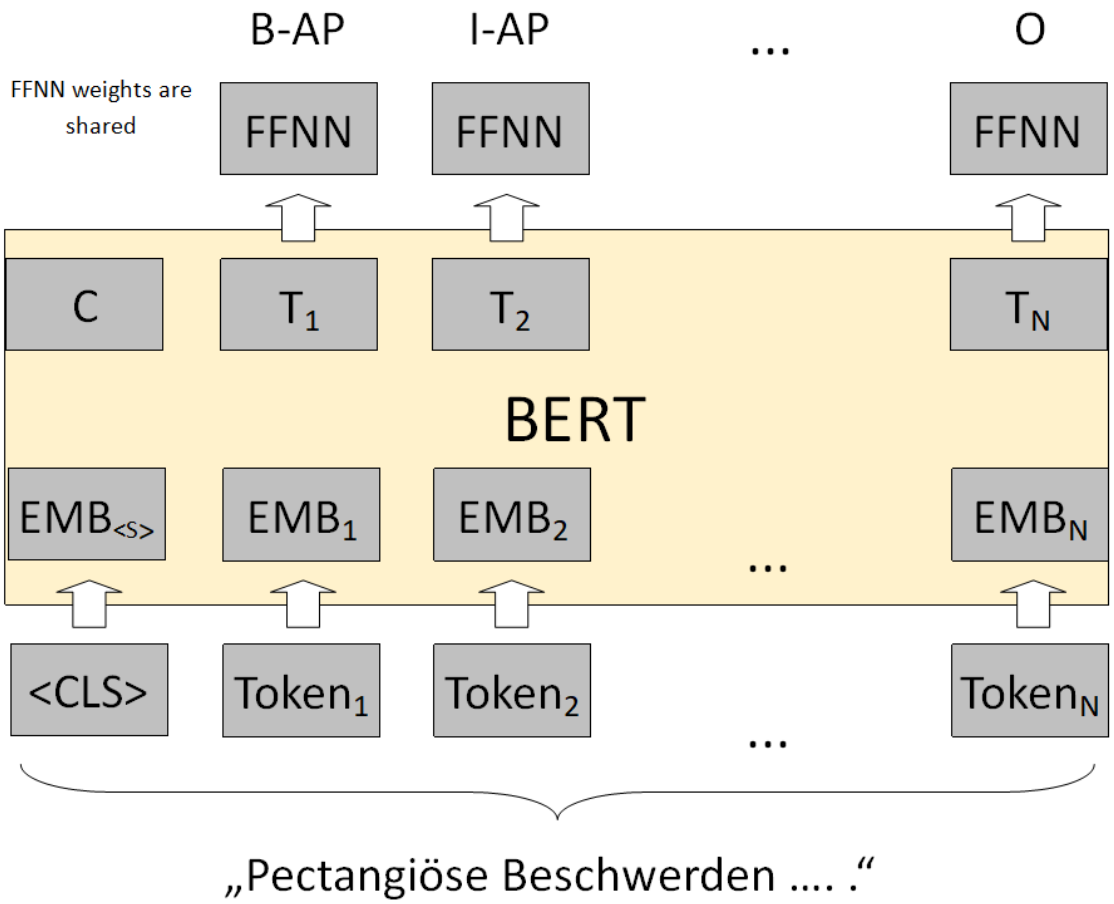


Fig. 2.4 **Fine-tuning BERT for cardiovascular concept extraction:** Input sequence *pectangiöse Beschwerden.*... is tokenized and embedded into a numerical representation. Each output representation T is used as input to a feed forward neural network (FFNN) with a final softmax layer. For example, the token *pectangiöse* is labeled as a B-AP, the token *Beschwerden* is labeled as an I-AP sequence.

CC	CRF	LSTM	$\text{BERT}_{\text{base}}$	$\text{BERT}_{\text{fine}}$	$\text{BERT}_{\text{scratch}}$
AP	69	73	**83**	82	78
DYSPNOE	70	72	**74**	73	70
NYKTURIE	96	92	**97**	91	**97**
ÖDEME	57	79	91	**94**	84
PALPITATION	79	74	**80**	79	77
SCHWINDEL	87	87	95	**98**	92
SYNKOPE	87	85	88	**89**	88
HYPERTONIE	89	90	**93**	87	92
CHOLESTERIN	86	89	**92**	90	89
DM	86	90	90	**91**	**91**
FA	81	77	**82**	74	80
NIKOTIN	86	87	92	90	**94**
Micro average / standard deviation	78/0.83	80/1.87	**86**/1.43	**86**/1.32	83/1.98

Table 2.3 **Results concept extraction:** Mean F_1-score per concept and micro-average F_1-score including standard deviation of the baseline classifiers (CRF and LSTM) and the three pre-trained language models ($\text{BERT}_{\text{base}}$, $\text{BERT}_{\text{fine}}$ and $\text{BERT}_{\text{scratch}}$) in percent. F_1-score is calculated by summing up F_1-scores per fold and dividing it by four. Best score in bold.
CC: cardiovascular concept; CCE: CC extraction; CRF: conditional random field; LSTM: long short-term memory; AP: Angina Pectoris; DM: Diabetes Mellitus; FA: Familial Anamnesis.

Results

$\text{BERT}_{\text{base}}$ and $\text{BERT}_{\text{fine}}$ achieved a token-wise micro-average F_1-score of $\approx 86\%$, significantly outperforming the LSTM (80%) and CRF (78%) baselines. $\text{BERT}_{\text{scratch}}$ only reached 83% F_1-score. The highest F_1-score improvements of these models we observed for ÖDEME (>10 pp.), AP ($\approx$ 10 pp.) and SCHWINDEL (>8 pp.). All BERT models improved recall scores while maintaining a high precision, yielding a better precision/recall balance than the baselines.

Discussion

Our error analysis showed that both baselines were more sensitive to spelling variants (e.g., typos of DYSPNOE: e.g. *Belastungsdysnpoe, Dypnoe, Blstungsdyspnoe, etc.* or inflections or inflected forms of AP: e.g. *pectanginöse/m/n* or *retrosternale/m.*), producing more frequent false negative errors. BERT models, especially $\text{BERT}_{\text{base}}$ and $\text{BERT}_{\text{fine}}$, were more robust to such variance and achieved higher recall scores with only slight precision trade-offs.
Further-pretraining a public BERT model on our large CardioComplete corpus does not

show significant performance gains compared to a plain public BERT model, and both outperformed a BERT model pre-trained from scratch and showed a lower standard deviation. However, both BERT models did not exceed a F_1-score of 86%. Both are important findings for on-premise adaptation with limited data and computational resources (cf. challenges 1.a. *Domain expertise*, 1.b. *Staff time*, 1.c. *Compute restrictions*).

Conclusion and relevance for the thesis

In this study, we performed an in-depth evaluation of fully supervised encoders using language models based on the BERT architecture pre-trained on three different corpora. We fine-tuned them on German doctor's letters from the cardiology domain, which are manually annotated with 12 CCs. We show that pre-trained language models outperform conventional strategies for automatic cardiovascular concept extraction in use-case scenarios where limited training data are available. However, the amount of training data in our experiments can not be considered few-shot.

Our experiments showed that (i) German PLMs can be fine-tuned on limited gold standard annotations entirely within a clinical infrastructure with limited compute resources (cf. RQ 1); (ii) continued further-pretraining ($\text{BERT}_{\text{fine}}$) does not automatically increase performance of general-domain PLMs ($\text{BERT}_{\text{base}}$), but both fully supervised PLMs significantly outperform baselines for our NER task (cf. RQ 2); and (iii) even when using pre-trained language models, creating high-quality gold standard data for fine-tuning remains necessary to improve model performance (cf. RQ 3). Furthermore, the improved recall score and precision/recall balance of all BERT models is relevant for the safety-critical clinical domain, as missing evidence (e.g. anamnesis information, medications) can cause patient harm, whereas the extraction of additional false positives can often be filtered in downstream steps (cf. RQ 4) (Zahl-Holmstad et al. 2023).

Due to data protection regulations, we could not share any data of this project to support reproducibility and transparency. Hence, this work motivated the creation of a distributable German clinical corpus presented in Chapter 4 and the study paved the foundation for two machine learning projects presented later in this thesis: (1) leveraging prompt-tuned encoders for section classification (Chapter 5) to further lower annotation and development costs while staying on-premise; and (2) generative LLMs for an end-to-end medication information extraction task using format-restricted outputs and parameter-efficient fine-tuning methods (Chapter 6), aiming to apply prompting strategies presented in 5 for more complex NER and RE tasks.

Chapter 3

State-of-the-art

In this chapter we position our contributions within the evolving landscape of clinical NLP with a focus on non-English, particularly German language settings. We structure this chapter along four dimensions: (i) clinical corpora and data accessibility; (ii) the evolution of language-model architectures and optimization paradigms under these constraints; (iii) model evaluation under clinical constraints; and (iv) interpretability. We focus on work contemporary to our publications covering CARDIO:DE (2023, cf. Chapter 4), section classification (2024, cf. Chapter 5) and medication information extraction (2025, cf. Chapter 6) and emphasize how methodological shifts motivated our work.

3.1 MIE under clinical on-premise and transparency constraints

MIE in clinical routine is characterized by two main challenges (cf. Section 2.1), strict on-premise constraints (cf. 1. *On-premise*) and transparency requirements (cf. 2. *Transparency*). In summary, our experiments tackle five main limitations remaining in current SOTA: (i) distributing a clinical corpus containing 500 doctor's letters from cardiovascular domain (CARDIO:DE), (ii) evaluating on-premise, few-shot section classification via domain- and task-adapted prompt-tuned encoders, (iii) evaluating joint NER+RE end-to-end medication information extraction using LLMs with reliable structured outputs and parameter-efficient fine-tuning methods, (iv) reducing manual evaluation efforts leveraging external feedback LLMs (v) supporting transparency and trustworthiness applying faithful interpretability methods for encoders and generative LLMs.

3.2 Distributable clinical corpora under strict data protection regulations

Data protection regulations in practice The volume of text data generated in clinical routine is still rapidly increasing. Hence, there is an immense need for high quality, freely accessible clinical routine text corpora derived from this documentation. However, strict data protection regulations are a significant challenge that prevents distributing even limited-size datasets. English clinical corpora in the United States must meet the regulations of the HIPAA.[1] The HIPAA *safe harbor* chapter explicitly lists eighteen PHI identifiers (e.g. person names, dates etc.), which need to be removed from clinical documents to be considered de-identified (Richter-Pechanski et al. 2023).[2]
These explicit regulations enabled the publication of several clinical text corpora in English: e.g. MEDICAL INFORMATION MART FOR INTENSIVE CARE (MIMIC) (2.5 million text documents) (Johnson et al. 2023), and datasets published in context of shared tasks that led to a series of benchmarks and popular clinical NLP frameworks, e.g. (i) corpora distributed by INFORMATICS FOR INTEGRATING BIOLOGY AND THE BEDSIDE (I2B2) AND NATIONAL NLP CLINICAL CHALLENGES (N2C2) (1,748 hospital reports) (Uzuner et al. 2011), (ii) a clinical corpus published by CLINICAL E-SCIENCE FRAMEWORK (CLEF) (150,000 clinical reports) (Roberts et al. 2009) and (iii) by the TEMPORAL HISTORIES OF YOUR MEDICAL EVENTS (THYME) project (1250 discharge letters) (IV et al. 2014; Richter-Pechanski et al. 2023).

In the European Union, the GDPR lacks an equally explicit definition on how to de-identify a clinical document (Regulation (EU) 2016/679), thus non-English corpora are scarce or intentionally falsified (e.g., French MERLOT (Campillos et al. 2018), Spanish IULIA (Marimon et al. 2017)). German resources are largely based on clinical guidelines or synthetic text: GERMAN GUIDELINE PROGRAM IN ONCOLOGY NLP CORPUS (GGPONC) 2.0 (based on oncological guidelines, 2 million token)(Borchert et al. 2022), JSYNCC (synthetic, 867 reports, 313,000 token)(Lohr et al. 2018a) or GRASCCO (synthetic, 60 documents, 43,000 token) (Modersohn et al. 2022). When CARDIO:DE was published, the only German corpus containing clinical routine documents was BRONCO (200 oncological discharge summaries, 89,942 token) from the University Hospitals Berlin (Charité) and Tübingen (Kittner et al. 2021). As the corpus was collected retrospectively and due to the lack of automatic de-identification solutions, each document needed to be carefully manually de-

[1] https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-C/part-164/subpart-E/section-164.502, accessed 01.09.2025.

[2] https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html, accessed 01.09.2025.

identified. Furthermore, all sentences in the corpus were shuffled in order to satisfy the legal requirements of the clinical data protection office and the ethics committee. Consequently, the corpus allows only sentence-level information extraction (Richter-Pechanski et al. 2023).

CARDIO:DE and its relevance for the thesis To the best of our knowledge, CARDIO:DE is the first freely available and distributable German corpus containing coherent and thoroughly de-identified doctor's letters from the cardiovascular clinical routine. The corpus supports on-premise development of clinical NLP pipelines in native language (RQ 1 and 2, cf. 1. *On-premise* challenges) and enables collaborative and reproducible research in German clinical NLP (cf. challenges 1.d. *Native language* and 2. *Transparency*) contributing to RQ 4. Furthermore, we publish two carefully curated annotation layers containing CDA-compliant document section classes and medication information including well-curated annotation guidelines to address the challenges of limited domain expertise (cf. 1.a. *Domain expertise*) and restricted staff time for manual annotation (cf. 1.b. *Staff time*) contributing to RQ 3. The CARDIO:DE gold standard annotations are foundational for Chapters 5 and 6 and support real-world clinical experiments in Chapter 7 contributing to RQ 5.

3.3 Learning paradigms under clinical constraints

In the following sections we discuss SOTA of model architectures and optimization paradigms relevant to (i) supervised encoders (pretrain-then-finetune), (ii) prompt-based encoders in few-shot scenarios, and (iii) generative LLMs for MIE with structured outputs.

3.3.1 Encoders under pretrain-then-finetune paradigm

Limitations of pretrain–then-finetune under clinical constraints Since 2017, most NLP tasks apply a pretrain-then-finetune paradigm: neural models are pretrained with a language modeling objective on large amounts of unlabeled text and then fine-tuned for a downstream task on a smaller amount of annotated data (Liu et al. 2023). However, especially in lower-resource languages and domains, due to the lack of high quality labeled data (Gao et al. 2021) fine-tuning requires a significant amount of manually labeled training data (Liu et al. 2023) which is time-consuming and expensive, especially when physicians are involved (cf. challenges 1.a. *Domain expertise*, 1.b. *Staff time*) (Richter-Pechanski et al. 2024).

Domain adaptation through further-pretraining Further-pretraining (training a PLM further on domain-specific texts using a language model objective) allows domain adaptation

of general-purpose PLMs. Several studies explored further-pretraining on domain-specific data (Zhu et al. 2021), demonstrating that further-pretraining even on limited-sized task-specific data can help to improve results in lower-resource downstream tasks (Gururangan et al. 2020).

PLMs for the medical domain Medical PLMs, pretrained from scratch or further-pretrained on biomedical texts often outperform general-domain PLMs on medical tasks (Sivarajkumar et al. 2023; Taylor et al. 2024). However, clinical routine texts significantly differ from biomedical texts, increasing the complexity of NLP tasks in clinical routine (Leaman et al. 2015; Hahn et al. 2020). Furthermore, most medical PLMs were pretrained primaril on English data (Lee et al. 2020; Li et al. 2023; Bressem et al. 2024).

3.3.2 Prompt-based encoders for section classification

Pretrain-then-prompt Recently scaled-up language models revealed impressive zero-shot capabilities, when NLP tasks shifted to a pretrain-then-prompt paradigm, where tasks are formulated using natural language prompts (Radford et al. 2019; Reynolds et al. 2021; Kojima et al. 2022; Liu et al. 2023). While in many applications at least a few training samples are still required to guide model predictions, prompt-based learning soon matched and even surpassed the performance of fine-tuning in various few-shot learning settings (Liu et al. 2023; Taylor et al. 2024; Richter-Pechanski et al. 2024).

Although model size is a critical factor for prompting (Chowdhery et al. 2023), more compact encoder-based PLMs were also successfully applied in few-shot settings (e.g., AUTOPROMPT(Shin et al. 2020), LM-BFF (Gao et al. 2021), PET (Schick et al. 2021a) and UPT (Wang et al. 2022)). We adopt PET because it fits our clinical setup by combining prompt-based training with a semi-supervised step that enables leveraging internal unlabeled clinical text and, its well-documented public implementation (for more details, cf. Chapter 2.4.2).[3] PET demonstrated a high performance for cloze-style text classification across various tasks. For English topic classification and few-shot size $k = 10$ per task, PET achieved 87.5% accuracy on AG'S NEWS, compared to 25.0% for supervised fine-tuning ($+62,5$ pp.); on YAHOO! ANSWERS PET reached 63.8% vs. 10.1% ($+53.7$ pp.) (Schick et al. 2021a). For cross-lingual stance detection on X-STANCE and $k = 1000$ few-shot samples, PET improved macro-F_1 on German from 43.4% to 66.4 ($+23.1$ pp.) with similar performance gains in French and Italian. On REAL-WORLD ANNOTATED FEW-SHOT TASKS (RAFT), a benchmark dataset of real-word tasks, PET further achieved a new

[3] https://github.com/timoschick/pet, accessed 01.09.2025.

SOTA performance close to non-expert humans in $7/11$ tasks including a biomedical text classification task, where for adverse drug effect classification it surpassed GPT-3 with an F_1-score of 82.2 versus 68.6 (Schick et al. 2022).

PET is particularly attractive under clinical constraints, as it minimizes the need for annotated data, enables incorporating unlabeled data, showed strong performance on German data and runs effectively on small-sized encoders (cf. challenges 1.a. *Domain expertise*, 1.b. *Staff time*, 1.c. *Compute restrictions* and 1.d. *Native language*) (Richter-Pechanski et al. 2024).

Prompting methods in clinical NLP Despite extensive research on medical PLMs, previous research has primarily focused on supervised fine-tuning with large amounts of training data (Wu et al. 2020) with the exception of (Schick et al. 2022) and (Taylor et al. 2024) who both investigated prompting on basic English clinical text classification tasks. Thus, there was a need to thoroughly investigate how prompting influences further-pretrained PLMs in native-language few-shot settings (Richter-Pechanski et al. 2024).

Clinical section classification Identifying sections in clinical texts has been shown to enhance performance on several MIE tasks (Pomares-Quimbaya et al. 2019). However, this research field remains underdeveloped, partly due to the lack of benchmark datasets (cf. comprehensive survey (Landolsi et al. 2023)). Therefore, most studies focus on English clinical texts (Denny et al. 2008; Edinger et al. 2018). In-depth studies focusing on few-shot learning scenarios and prompting are lacking (Ge et al. 2023). Our work is the first to thoroughly investigate these methods on a freely available clinical German benchmark corpus. Furthermore, we extensively explore German PLMs (Bressem et al. 2024) for clinical domains to detect suitable (further-)pretraining methods for prompting and their effect on section classification (Richter-Pechanski et al. 2024).

Relevance for the thesis Due to its suitable attributes under clinical constraints and strong performance on various tasks and domains, in Chapter 5 we apply PET to systematically evaluate prompt-tuned encoders for multiclass section classification in German doctor's letters using annotations of CARDIO:DE (Richter-Pechanski et al. 2024). We compare various prompting strategies and general domain German PLMs with medical-adapted PLMs using task- and domain-adaptation, to reduce the demand for manual annotations, directly addressing challenges 1.a. *Domain expertise*, 1.b. *Staff time* and 1.d. *Native-language barrier*. We use small-sized encoders with less than 500 million parameters, emphasizing small models with less than 150 million parameters to reduce the demand of compute resources (cf. 1.c.

Local compute resources) (Leaman et al. 2015; Hahn et al. 2020; Richter-Pechanski et al. 2024).

We directly contribute to our RQs by demonstrating on-premise feasibility of domain- and task-adaptation of small-sized prompt-based encoders for a section classification task (RQ 1), by providing a resource-aware comparison of general vs. medical-adapted encoder-based PLMs and various fine-tuning strategies (RQ 2) and by reducing annotation and evaluation efforts via pattern-exploiting training and various prompting strategies (RQ 3).

3.3.3 Generative LLMs for MIE

Generative LLMs In recent years generative LLMs showed impressive capabilities in various NLP tasks (Brown et al. 2020; Chowdhery et al. 2023; OpenAI 2023). For example, GPT-4 achieved 86.4% weighted accuracy on the MASSIVE MULTITASK LANGUAGE UNDERSTANDING (MMLU) benchmark, exceeding finetuned encoder-based baselines (ROBERTA) by 58.5 pp. (27.9%) (Hendrycks et al. 2020; OpenAI 2023).

Because most SOTA models are closed-source or contain hundreds of billions of parameters, their usage in the resource-restricted and privacy-sensitive clinical routine remains limited. However, the publication of open source models like Llama and Mistral pushed the development of local LLMs applications and research (Grattafiori et al. 2024; Jiang et al. 2023). Notably Llama 3 70B achieved 86.0% accuracy on the MMLU benchmark, almost closing the gap to GPT-4.

Fine-tuning and domain-expert LLMs Quantization and parameter-efficient fine-tuning (PEFT) methods, such as LoRA and QLoRA, enabled further-pretraining and fine-tuning of LLMs on constrained IT infrastructure across a range of domains. This has resulted in the release of several medical-domain-adapted LLMs, including PMC-Llama, Meditron or OpenBioLLM (Wu et al. 2024; Chen et al. 2023; Ankit Pal et al. 2024). However, additional domain-pretraining did not reliably improve over the source model. For example OpenBioLLM 8B does not consistently outperform its Llama-based source across tasks (Ankit Pal et al. 2024). In contrast, several studies showed that smaller expert LLMs, PEFT fine-tuned on annotated data, frequently outperformed larger general LLMs in various medical tasks (Lehman 2024; Xu et al. 2024; Lehman et al. 2023).

LLMs in the medical domain Google evaluated their large Pathways Language Model (PaLM) model for various medical question-answering (QA) tasks, showing that foundation LLMs encode clinical knowledge, thereby empahiszing the transformative potential of LLMs in healthcare (Singhal et al. 2023; Temsah et al. 2024). While we have seen various studies

evaluating LLMs on medical QA data sets (cf. MedQA (Jin et al. 2021), MedMCQA (Pal et al. 2022), PubMedQA (Jin et al. 2019)), works that apply LLMs for MIE tasks remain limited. Steering LLMs to consistently generate structured output is still under intensive research (Liu et al. 2024c; Liu et al. 2024a). Recently, encoder-decoder models and generative models have been leveraged to solve NER and RE tasks using well-defined JSON strings as output (Dagdelen et al. 2024; Lu et al. 2022). This makes it possible to evaluate these models on existing high-quality benchmark datasets for various IE tasks using well-established metrics, such as precision, recall and F_1-score, and to compare their performance to traditional encoder-based classification models. In our setting, we formulate the MIE task as a one-step end-to-end JSON generation task.

Medication information extraction In the context of shared tasks, three data sets had been published for the task of medication information extraction (i2b2 2009 (Patrick et al. 2010), MADE 1.0 2018 (Jagannatha et al. 2019), n2c2 2018 (Henry et al. 2020)). All data sets contain named entity annotations and relation annotations. As our English data set, we selected the most recent and comprehensive n2c2 2018 track 2 data set. It contains 505 clinical notes extracted from the MIMIC III database (Johnson et al. 2016).
Wei et al, achieved SOTA results for the n2c2 task using a combination of RNNs and CNNs (Wei et al. 2019). El-Allaly et al. used BERT (SciBERT) transformers and established a new SOTA (El-allaly et al. 2021). In this thesis we use both results as our baselines for the English dataset. Most recently some studies used generative LLMs to solve the n2c2 MIE task. (Fornasiere et al. 2024) compared zero-shot and few-shot performance using Mistral 7b and (Hsu et al. 2025) developed a weakly-supervised pipeline using Llama 7b and 13b to label clinical data to fine-tune a downstream BERT model. However, neither study reached new SOTA results, and both presented only micro-averaged F_1-scores per model, excluding class-wise evaluations. (Hu et al. 2024b) used generative LLMs to investigate the suitability of LLMs for MIE tasks. However, they evaluated their models solely for medication extraction using NER and RE separately without any joint evaluation. Furthermore, they only used a small subset of the i2b2 2010 clinical corpus. (Modi et al. 2024), in a recent comprehensive survey, presented all publications using the n2c2 dataset as a benchmark.

Medication information extraction in German For the German language only a few MIE studies are available. Recent publications from (Sharma et al. 2024) and (Roller et al. 2022) used an encoder-based model (BERT) to conduct medication information extraction tasks. However, their data sets are not publicly available as the only distributable German data set with annotated medication information is currently CARDIO:DE (Richter-Pechanski et al.

2023). To our knowledge, we are the first to evaluate SOTA generative LLMs on English and German clinical routine data to extract end-to-end medication information under clinical constraints.

Relevance for the thesis In chapter 6 we evaluate two publicly available LLMs under identical on-premise clinical constraints: (i) Llama, because of its strong performance and availability and (ii) OpenBioLLM, to test weather medical-pretraining helps despite not always surpassing Llama on public benchmarks. We evaluate the feasibility of PEFT-finetuning these models on a German MIE task (addressing RQ 1, cf. 1. *On-Premise* challenges). We develop a joint end-to-end LLM pipeline to conduct a complex NER+RE task and compare performance to SOTA baselines (addressing RQ 2, challenges 1.c. *Compute restrictions* and 1.d. *Native language*). In addition, we want to reduce manual efforts by leveraging clinical knowledge of our selected LLMs in combination with format-restricted prompts (RQ 3, cf. challenges 1.a. *Domain expertise* and 1.b. *Staff time*). Finally, we apply the best performing medication extraction model to two real-world clinical applications, presented in Chapter 7 (addressing RQ 5).

3.4 Evaluation under clinical constraints

Structured outputs are essential in MIE. Encoder models, predict a label per input token, thus evaluation follows a standard token- and entity-wise procedure. Generative LLMs predict a sequence of tokens, thus the output needs to be constrained to predefined schemas (e.g. JSON) to reduce post-processing efforts and to enable automatic evaluation using standard evaluation metrics, such as F_1-score or accuracy (Lu et al. 2022) (cf. Section 3.3.3). Recent studies primarily focused on two methods to steer LLMs to structured output: (i) prompt-based methods using e.g. format-restricting prompts (Dagdelen et al. 2024; Lu et al. 2023) and (ii) decoding-based methods using formal language approaches by guiding the token generation process itself (Geng et al. 2023). We argue that prompt-based methods are lighter-weight and easier portable under clinical on-premise limitations. Hence, we combine format-restricting prompts and supervised fine-tuning to enable even small-sized LLMs to reliable structured output generation (cf. challenge 1.c. *Compute restrictions*).
However, even if the model predicts well-defined structured output, evaluation of IE results via pattern matching often over-penalizes clinically correct predictions (e.g., abbreviations, units, list vs. concatenated strings). Filtering such instances manually is error-prone and time-consuming (Gu et al. 2024). Recent studies used LLM-as-a-judge methods to automatize large-scale LLM output evaluation (Sharif et al. 2025; Gu et al. 2024; Chiang et al. 2023).

Gu et al. formalized the use of an LLM to generate an evaluation E over an input x given a context C (typically a prompt template) as $E \leftarrow P_{\text{LLM}}(x \oplus C)$ (Gu et al. 2024).

Relevance for the thesis In our feedback-LLM scenario, x is the (gold/prediction JSON) combined with an instruction and a task description C. The external feedback LLM returns E as a binary classification output (*SIMILAR/NOT SIMILAR*) label (cf. Chapter 6). However, we keep the primary evaluation scores from pattern matching, and use the feedback LLM only as a post-hoc step to assist in fine-grained output evaluation, to reduce manual evaluation efforts (cf. challenges 1.a. *Domain expertise* and 1.b. *Staff time*). This directly addresses RQ 3 by reducing manual efforts and RQ 4 by supporting reliable detection of prediction errors.

3.5 Interpretability

Given the black-box nature of deep learning architectures, the interpretability of model outputs is challenging and attracts much interest, especially in safety-critical domains such as clinical routine (Fan et al. 2021). Various feature attribution methods have been developed to address these issues, especially for classification models (Ribeiro et al. 2016; Sundararajan et al. 2017; Lundberg et al. 2017), but we still face challenges in assessing their quality (Jacovi et al. 2020; Attanasio et al. 2023).

We follow (Jacovi et al. 2020), by distinguishing between faithfulness (accurately representing the reasoning process behind the model output) and plausibility (how convincing the model interpretation is to humans). Plausibility-based interpretations only assess whether an explanation aligns with human judgment, but they fail to reflect whether they actually represent the inner workings of a model. Hence, they are not sufficient to detect and debug inference errors, which is crucial in the safety-critical clinical domain. Faithful interpretations based on input attributions reflect how input features affect the model output. This supports a transparent and trustworthy application in the clinics. Recent studies suggest that token-level faithfulness may evaluate self-consistency, rather than the true reasoning process of a model. However, we chose Shapley values as our primary interpretation method for both encoders and generative LLM models. We argue that we focus on the direct pathways between the input tokens and the output class/tokens, excluding LLM rationales from our interpretation as discussed in (Parcalabescu et al. 2024). Shapley values provide a theoretically well-founded approach to determine the contribution of individual input features to a model prediction (Shapley 2016; Parcalabescu et al. 2024). Furthermore, we conducted experimental internal explorations and compared several interpretability methods in advance with FERRET, a

framework for benchmarking popular explainers on transformers (Attanasio et al. 2023), finding that Shapley values were the best-performing method for our experiments.
A computationally optimized implementation called SHAP can be applied out-of-the-box on transformer-based models (Lundberg et al. 2017). In this thesis we investigated the use of Shapley values for data and model optimization, since prior work in clinical NLP, to the best of our knowledge, has been limited. Furthermore, application of Shapley value analysis to generative LLMs in a clinical information extraction task remain scarce. We therefore investigate this setting using a recent CAPTUM-based implementation (Miglani et al. 2023).

Relevance for the thesis Our approach supports RQ 4 by integrating Shapley value analyses for transparent error detection, model understanding, training data optimization and trustworthy predictions in a safety-critical setting (cf. challenges 2. *Transparency*). We develop use cases for Shapley values for both: encoder-level attributions (Chapter 5) and generative LLM attributions (Chapter 6).

Chapter 4

CARDIO:DE - Distributing a Clinical Corpus

4.1 Outline and contributions

In this chapter we present CARDIO:DE, a prospective, GDPR-compliant clinical corpus containing 500 German cardiology doctor's letters from the cardiology department at Heidelberg University Hospital. The corpus is enriched with two annotation layers: (1) for token- and relation-level medication information extraction and (2) for paragraph-level section classification.

Section 4.2 motivates the distribution of CARDIO:DE in a clinical NLP context. Section 4.3 describes the characteristics of the corpus, defining the scope, inclusion criteria, and document structures. Section 4.4 addresses ethical considerations, study design and the de-identification and annotation workflow. Section 4.5 presents our baseline results for medication information extraction and section classification using SOTA machine learning (ML) methods. Section 4.6 describes how to access the corpus for research purposes. Finally, Section 4.7 summarizes this chapter in the context of research questions and clinical constraints.

CARDIO:DE is designed to tackle the clinical constraints presented in Section 2.1 and to contribute to our research questions defined in Section 1.2. The prospective study design, a thorough manual de-identification process preserving temporal and structural information of real-world German clinical letters and two well-curated annotation layers allow on-premise NLP experiments and model adaptation inside local clinical IT infrastructures, addressing RQ 1 (cf. challenges 1.a. *Domain expertise*, 1.b. *Staff time*, 1.c. *compute restrictions* and 1.d. *Native language*). Considering RQ 2, two annotation layers covering two task complexities,

paragraph-level and token-/relation-level annotations, serve as a benchmark to compare model architectures and optimization paradigms in Chapter 5 (prompt-tuned encoders for section classification) and Chapter 6 (LLMs for end-to-end medication information extraction) (cf. *compute restrictions* and 1.d. *Native language*). Regarding RQ 3, thoroughly developed annotation guidelines and well-curated annotation layers help to reduce annotation and evaluation efforts (cf. challenges 1.a. *Domain expertise*, 1.b. *Staff time*). RQ 4 is addressed by sharing the first German clinical corpus among research institutions, facilitating transparent and replicable NLP experiments (cf. challenges 1.d. *Native language* and 2. *Transparency*). The corpus serves as the main data source for machine learning experiments in Chapter 5 and 6. In addition, the corpus was used to fine-tune machine learning models, which were then deployed on real-world datasets for evaluation, contributing to RQ 5.
CARDIO:DE400 annotations are released and CARDIO:DE100 is kept as a held-out dataset. Furthermore, we provide two reproducible baselines for both annotation layers.

4.2 Motivation

Despite sustained declines in cardiovascular desease (CVD) mortality in many countries across Europe, CVDs still account for approx. 4.1 million deaths within European Society of Cardiology (ESC) member countries and have remained the most common cause of death within this region (45 and 39% of all deaths in females and males, respectively). Moreover, the prevalence of CVDs across Europe is still high with an estimated 113 million people living with CVD in the 57 ESC member countries, significantly contributing to patient morbidity and hospitalizations (Timmis et al. 2022).

At the same time, in clinical routine, large portions of data like patient anamnesis, cardiovascular risk factors and diagnosis continue to be stored in unstructured form, such as free text in doctor's letters (Starlinger et al. 2017). The predominantly hypothesis-driven strategies used in cardiovascular research should be complemented by computer-assisted methods. Comprehensive analyses of large clinical datasets using automatic information extraction methods will not only significantly expand data sources for clinical care and research, but could also improve clinical decision-making and allow for progress in personalized medicine (Starlinger et al. 2017) (cf. Section 2.1).

The rapid development in the field of NLP in the past 15 years provided powerful tools for automatic text processing (Hahn et al. 2020). A high number of models, based on rule-based, statistical and recently deep learning methods were developed and validated for various tasks. While SOTA generative LLMs achieved impressive zero-shot results, particularly in lower-resource languages and domains, annotated data is still crucial for fine-tuning

(Richter-Pechanski et al. 2024; Richter-Pechanski et al. 2025). Moreover, all methods require annotated data for evaluation and quality control (cf. Chapter 2 and 3).

Therefore, shared corpora in the clinical domain are essential to support transparent and reproducible experiments and foster innovation in the field of clinical NLP (Starlinger et al. 2017; Chapman et al. 2011; Meineke et al. 2023; Lentzen et al. 2022) (cf. Section 3.2 and Chapters 5, 6 and 7).

4.3 Corpus characteristics

CARDIO:DE encompasses 500 cardiovascular doctor's letters covering a broad clinical spectrum of a tertiary care cardiovascular center between 2020 and 2021. Our corpus covers 311 in-patient, 172 outpatient and 17 letters of the cardiac emergency room (chest pain unit). Thus, the included doctor's letters cover both complex multiple-day hospitalizations and brief out-patient presentations. This results in the deployment of a representative collection of clinical documents, covering common doctor's letter sections (e.g. *anamnesis, physical examination, instrumental diagnostics, laboratory results, epicrisis, medication*) in varying degrees and details. Figure 4.1 illustrates an excerpt of a doctor's letter including common section types. The complete corpus contains 993,143 tokens, with approximately 31,952 unique tokens.

We randomly split our corpus into two parts, similar to (Kittner et al. 2021). CARDIO:DE400 contains 400 documents, 805,617 tokens and 114,348 annotations. CARDIO:DE100 contains 100 documents, 187,526 tokens and 26,784 annotations. Both corpora are published for scientific research purposes. Scientific research excludes processing the data for marketing purposes. We only published annotations of CARDIO:DE400, annotations of CARDIO:DE100 are kept in-house as held-out data for a shared task on various MIE tasks, which we want to organize in the future.

Table 4.1 shows a quantitative analysis per CARDIO:DE splits. In Table 4.2 we present the most common 50 whitespace separated token in CARDIO:DE including token count.

We published cardiovascular doctor's letters as close as possible to clinical routine documents. To achieve this, CARDIO:DE is based on a prospective study design with patient consent, which enabled us to keep the original document structure of clinical routine doctor's letters (Figure 4.2). Although collecting patient consents can be time consuming and tedious, this procedure ensures that we comply the best with current data protection regulations in Germany. Moreover, similarly to recent corpus distribution projects (Johnson et al. 2023; Kittner et al. 2021; Lohr et al. 2024) we preserved the information on patient's

Universitätsklinikum Musterstadt Station Sowieso | Beispielstr. 12 | 12345 Musterstadt

Frau
Dr. med. Paul Beispiel
Musterplatz 1
56789 Beispielstadt

Test-Klinik
Zentrum für Kardiologie

Klinik für Kardiologie
Station II

Dr. med. Muster
Ärztlicher Direktor
Station II

Station Sowieso
Besipiestr. 123
12345 Musterstadt
Tel +123 23 45 67
Fax +123 23 45 66

01.01.2010

Nachrichtlich:
Herrn Max Mustermann, Beispielplatz 1, 12345 Musterstadt

Sehr geehrter Herr Kollege Muster,

wir berichten über Ihre Patientin Frau Maxima Musterfrau geboren am 01.01.1970, wohnhaft in 12345 Musterstadt, Beispielstr. 1, die sich vom bis in unserer stationären Behandlung befand.

Diagnosen:

Schwerer Infarkt der … am 01.02

Cvrf: **Hyperlipidämie, Nikotinkonsum seit 01.01.1980, 30 py.**

Allergien: Hausstaub

Anamnese:

Die stationäre Übernahme von Frau Musterfrau erfolgte über die Chirurgie. Die Patientin klagt über **Tachykardien**. Auf gezielte Nachfrage eingeschränkte Belastbarkeit, **belastungsabhängiges thorakales Druck- und Engegefühl** außerdem **progrediente Belastungsdyspnoe**. Es bestehen **Ödeme bds.**, **kein Schwindelgefühl**, **keine Synkopen**.

Wir danken für die vertrauensvolle Zusammenarbeit und stehen bei Rückfragen selbstverständlich jederzeit gerne zur Verfügung.

Labor:

Bezeichnung	Wert	Datum
Abc	123	01.01.2010

Medikation:

ASS 50mg 1-0-0
Clexane 12mg 0-1-1 bis Mai 2011

Mit freundlichen Grüßen

Dr. med. Muster
Ärztl. Direktor

Dr. Platzhalter
Oberarzt

Fig. 4.1 **Structure of a doctor's letter:** German dummy doctor's letter from cardiology domain used in CARDIO:DE corpus. The letters are semi-structured binary texts MS DOC files. Most of the letters contain at least a header with contact information, a salutation, a diagnosis section, an anamnesis, laboratory values, medication plan and a conclusion/epicrisis.

Statistic	CARDIO:DE	CARDIO:DE400	CARDIO:DE100
Total	993,143	805,617	187,526
Mean	1,986	2,014	1,875
Min	588	588	597
25%	1,064	1,082	992
Median	1,704	1,764	1,448
75%	2,638	2,647	2,562
Max	6,644	6,644	5,322

Table 4.1 **Corpus token statistics**: Total token count and quantitative analysis of token count per doctor's letter per CARDIO:DE split.

Token	Count	Token	Count
Parameter	3,404	/nl	936
Datum *(date)*	3,388	links *(left)*	919
Wert *(value)*	3,381	Vorstellung *(presentation)*	913
Normb *(norm range)*	3,365	Beschwerden *(complaints)*	906
Dimension	3,365	Nachweis *(proof)*	899
min	1,999	rechts *(right)*	881
mmHg	1,843	Befund *(finding)*	872
empfehlen *(recommend)*	1,784	guter *(good)*	871
erfolgte *(took place)*	1,619	Ruhe-EKG *(resting ECG)*	858
Verlauf *(course)*	1,488	gute *(good)*	850
Patienten *(male patient)*	1,356	Medikation *(medication)*	846
wurde	1,336	Hinweis *(evidence)*	831
Mmol	1,285	Kontrolle *(check-up)*	828
Pumpfunktion *(pumping function)*	1,262	QRS	796
zeigte *(showed)*	1,175	Ödeme *(edema)*	788
Patientin *(female patient)*	1,139	QTc	786
Therapie *(therapy)*	1,120	gut *(good)*	773
Aufnahme *(admission)*	1,120	bitten *(asking for)*	756
gerne *(preferred)*	1,108	Funktion *(function)*	755
ca.	1,057	leicht *(mild)*	754
Z.n *(S/P)*	1,054	regelrecht *(correct)*	724
normal	1,026	LAD	723
unserer *(our)*	1,018	regelmäßige *(regularly)*	720
Sinusrhythmus *(sinus rhythm)*	973	ASS	714
Risikofaktoren *(risk factors)*	937	Echokardiographie *(echocardiography)*	709

Table 4.2 **Most common CARDIO:DE token:** 50 most common whitespace separated tokens in CARDIO:DE including count per token (selected English terms in brackets).

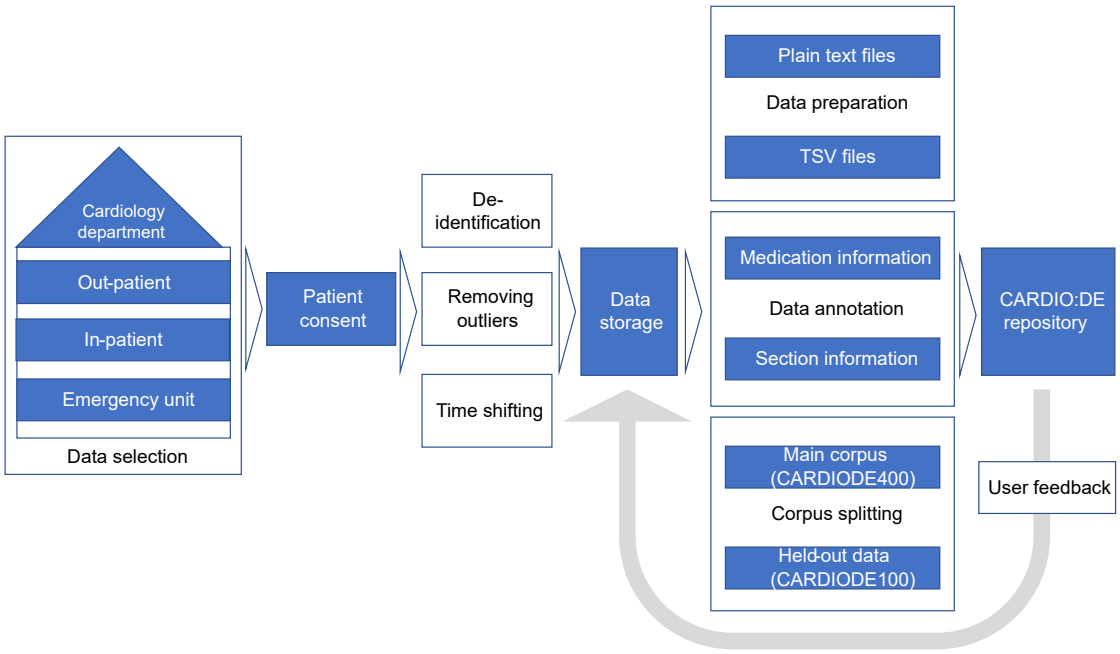


Fig. 4.2 **CARDIO:DE study design:** Visualization of the development process of CARDIO:DE. (1) Data selection, (2) data collection and storage, (3) data preparation, annotation and corpus splitting, (4) CARDIO:DE repository.

age and time/date in the documents. Thus, the corpus can be used for various information extraction tasks at the document level in the cardiovascular domain.

4.4 Methods

4.4.1 Ethics declaration

This study has been approved by the ethics committee of the Heidelberg University Hospital (S-498/2020) and has therefore been performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki. All persons gave their informed consent prior to their inclusion in the study.

4.4.2 Data selection and collection

Our study was designed monocentric, non-interventional, non-randomized, and prospective by collecting 500 patient consents between 2020 and 2021 in the cardiology department at the Heidelberg University Hospital. One doctor's letter per patient was included into the CARDIO:DE corpus. Inclusion criteria were as follows: (1) age of at least 18 years,

(2) written consent signed by the patient, and (3) a diagnosis with a cardiovascular system disease. Patients were included after revision of the criteria by the recruiting study assistant, adequate information and subsequent signing of the CARDIO:DE consent form. By signing, the patient's next generated doctor's letter was included into the corpus. No study-specific additional examinations or further measures are performed within the scope of the project; thus the patient was not exposed to any additional risk. We then exported each document and converted it from binary MS DOC to MS DOCX to a plain text format keeping the paragraph sections consistent, highlighted with the "¶" symbol in the original MS DOC.[1] We split each document by paragraph and tokenized each document using SPACY (v.3.2.1, language pipeline: `de_dep_news_trf`) (Honnibal et al. 2019).

4.4.3 De-identification

All documents were initially de-identified using our deep learning model trained on manually annotated in-house data (further details, cf. Chapter 2.6.1) (Richter-Pechanski et al. 2019). In a second step, clinical experts manually reviewed the automatic de-identified documents and replaced remaining un-deidentified PHI token with appropriate semantic placeholders. To keep the chronological order in the documents, we followed best-practice procedures by shifting all dates by a random number per document (Johnson et al. 2023). Information about weekdays, time of day and seasons were kept. We also kept the information about patient age in the documents. If the patient was older than 80 years, we followed best-practice approaches, by shifting age by a random number larger than 300 (Johnson et al. 2023).

We added three initial lines to each document containing pseudonymized meta information about: (1) admission date, (2) date of birth and (3) patient age. To further ensure anonymity, we removed outliers in laboratory values of each document, including patient height and weight. To identify outliers, we used a z-score approach (Rousseeuw et al. 2011). Finally, we stored each doctor's letter in our clinical data storage to save the corpus for further data preparation and annotation.

4.4.4 Data annotation

We created two annotation layers for CARDIO:DE: (1) paragraph-level section classes and (2) token- and relation-level medication information. We used the annotation tool INCEPTION (v. 22.3) optimized for span annotations, including a monitoring and curation

[1] using LIBREOFFICE 6.1.5.2, cf. https://wiki.documentfoundation.org/Faq/General/150, accessed: 08.09.2025.

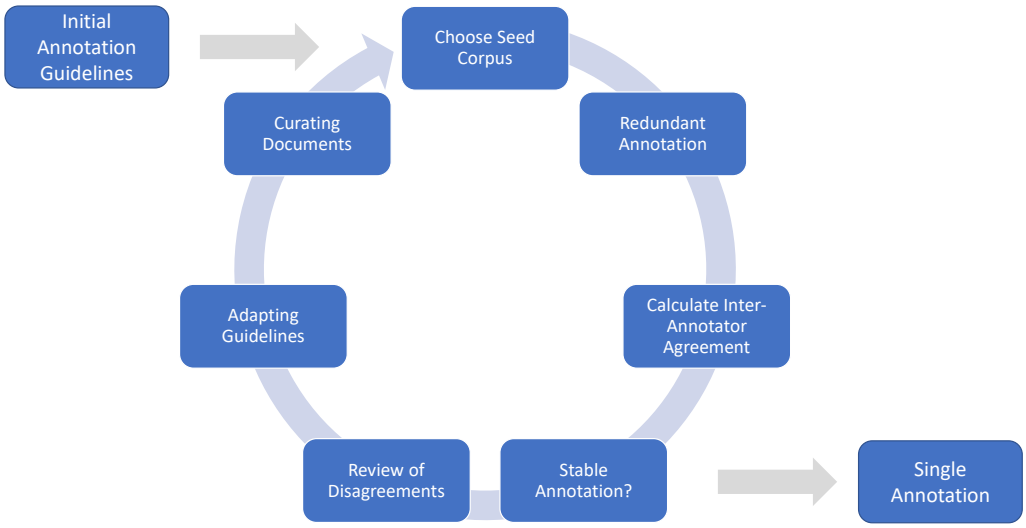


Fig. 4.3 **Annotation workflow:** Iterative guideline adaptation process used for CARDIO:DE annotation layers. Redundant iterations helped synchronize the annotations with the guidelines. After meeting the IAA threshold, a final batch was assigned to each annotator.

tool (Klie et al. 2018). INCEPTION was installed as a web service in the clinical network to facilitate its use and data access for our annotation team within the clinical infrastructure.

We used well-established annotation methods (Gurulingappa et al. 2012; Lohr et al. 2020; Roberts et al. 2009; Wilbur et al. 2006), including a guideline adaptation process by redundantly annotating documents involving an inter-annotator agreement score (IAA) in an iterative approach (Figure 4.3).

After drafting initial guidelines with domain experts, a subset of documents was sampled from the main corpus for redundant annotation by all annotators. After each iteration the annotation master reviewed all annotations and documented all disagreements. To measure annotation quality an IAA score was calculated. During the following review meeting with all annotating participants including the annotation master all disagreements were discussed and a joint solution was defined. If necessary, the guidelines were adapted and a new iteration round was initialized until a pre-defined IAA threshold was met, depending on the annotation task. The annotation master curated all documents of each iteration based on the adapted guidelines. After redundant annotation was completed each annotator was assigned to a distinct subset of the remaining documents. To ensure high annotation quality nevertheless, the annotation master carefully reviewed all single annotated documents in compliance with the final guidelines.

The annotation guidelines used for the initial corpus release and the later revised versions are provided as supplementary material on heiDATA (cf. https://doi.org/10.11588/DATA/USQLMB). See Appendix A.1 for further guideline information.

Medication information annotation

Our medication information annotation scheme is based on (Uzuner et al. 2010a), and was adapted to the specific structure of our CARDIO:DE data (guidelines, cf. (Richter-Pechanski et al. 2023)). While annotating token containing medical entities, many of them were not annotated with a class type, thus we used F_1-score (harmonic mean between precision and recall) as IAA to reflect the proportion of true negatives. We performed three iterations of redundant annotation containing 15, 15 and 10 documents. Our annotator team included four medical informatics master's students, of whom three had clinical experience and two medical students in their seventh semester (third clinical) with clinical routine experience. The entire project lifetime, including preparation, annotation and evaluation, was three months. Approximate annotation time per document was 5–10 min.

Most of medication information in a doctor's letter is listed in a separate semi-structured section *(e.g. Therapieempfehlung, Medikation bei Aufnahme, etc.*). In addition, we annotated medication information in narrative text sections. For all annotated medications in a doctor's letter, the patient had to be the experiencer. We neither made any assumptions nor considered longitudinal information from external sources about a patient.

Our annotation objective was to identify a relevant drug (DRUG) or active ingredient (ACTIVEING) and its relation information (DOSAGE, ROUTE, FREQUENCY, DURATION, STRENGTH, REASON AND FORM). Moreover, we added a binary attribute (INNARRATIVE) to each DRUG/ACTIVEING, to mark whether the medication information is in a semi-structured or in a plain text section (for an example annotation, see Figure 4.4). We did not add entity normalization to the medication information layer.

Annotation quality Figure 4.5 shows all token level median IAA scores of all annotator combinations per iteration per medication information class of our initial corpus version 1.0. Detailed information of IAA scores including standard deviation, see Table 4.3.

IAA could be improved consistently for three classes (DRUG, FORM and FREQUENCY) over all iterations. For classes DURATION, STRENGTH and REASON IAA could be increased in second iteration and slightly decreased in third iteration. The complex REASON class still achieved a relatively low IAA (0.41) in iteration 3. For classes ROUTE and ACTIVEING IAA continuously decreased over all iterations. Standard deviation decreased in iteration 3 for

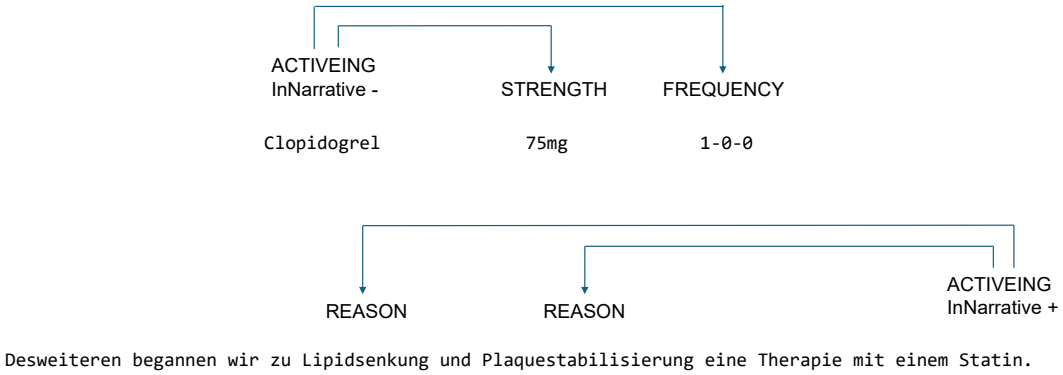


Fig. 4.4 **Example annotations for medication information:** Two annotated text snippets including medication information annotations and relations to other tokens. (Top) The ACTIVEING entity contains an attribute INNARRATIVE to specify if the entity is inside a semi-structured section (-) or plain text section (+) of a doctor's letter. In this example, the ACTIVEING entity is inside a semi-structured section and related to a STRENGTh and a FREQUENCY attribute. (Bottom) The ACTIVEING entity, inside a plain text section (positive INNARRATIVE), is related to two REASON attributes.

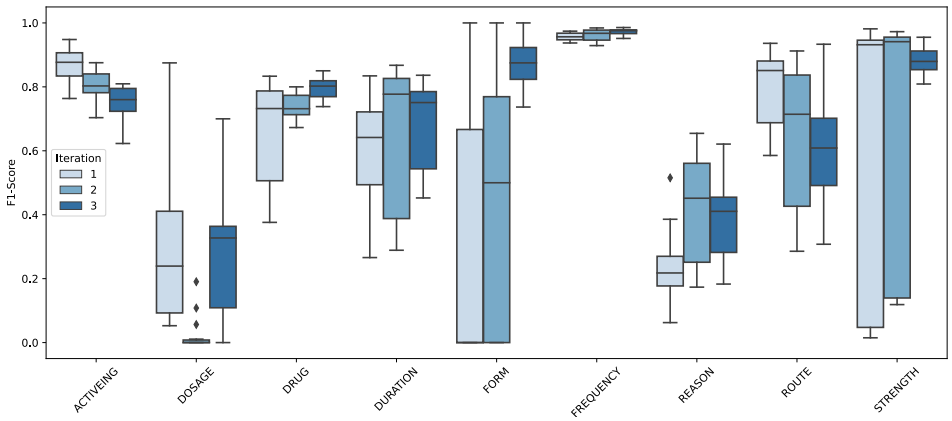


Fig. 4.5 **IAA medication information:** Boxplot to illustrate the development of token level median IAA scores of all annotator combinations per iteration per medication information class (entity). X-axis: medication information class, y-axis: IAA F_1-scores per iteration.

Label	Iteration 1	Stddev	Iteration 2	Stddev	Iteration 3	Stddev
ACTIVEING	0.87	0.05	0.80	0.05	0.76	0.05
DOSAGE	0.24	0.28	0.00	0.05	0.33	0.19
DRUG	0.73	0.16	0.73	0.04	0.80	0.03
DURATION	0.64	0.16	0.78	0.22	0.75	0.13
FORM	0.00	0.37	0.50	0.38	0.88	0.07
FREQUENCY	0.96	0.01	0.97	0.02	0.98	0.01
REASON	0.21	0.11	0.45	0.17	0.41	0.12
ROUTE	0.85	0.11	0.71	0.22	0.61	0.16
STRENGTH	0.93	0.44	0.94	0.40	0.89	0.04

Table 4.3 **Median F_1-scores per medication information class:** IAA for all three redundant annotation iterations including standard deviation (stddev).

Metric	Iteration 1	Iteration 2	Iteration 3
Token-wise	0.85	0.89	0.85
Entity-wise	0.84	0.79	0.81
Relation annotation	0.37	0.62	0.61

Table 4.4 **Median micro-average F_1 score:** IAA for medication information (token-wise, entity-wise) and median micro-average F_1-scores for medication relation annotations.

all but the DOSAGE class, which showed the overall lowest IAA scores with a maximum of 0.33 in iteration 3.

Considering median micro-average F_1-score, token-wise F_1-score could be improved in second iteration from 0.85 to 0.89, but only leveled at 0.85 in third iteration. Entity-wise median F_1-score decreased in second iteration from 0.84 to 0.79, but increased to 0.81 in third iteration, which is 3 percentage points below IAA of first iteration (Table 4.4).

In addition to calculating IAA for token-level medication information, we computed micro-average F_1-scores to assess the annotation quality of the annotations of medication information relation. We could improve IAA in second iteration by 0.25. In the last iteration IAA decreased to 0.61 (Table 4.4). As also reported by other publications, the IAA scores for relation annotation were generally lower, than for entity annotations (Campillos et al. 2018; Roberts et al. 2009).

Overall, the synchronization of the annotation for medication information was very challenging. While annotation quality of medication information in the structured section of doctor's letters quickly improved, annotation of more complex medication information samples in plain text remained difficult. Due to time restrictions, we stopped redundant annotations after the third iteration (cf. challenges 1.a. *Domain expertise* and 1.b. *Staff*

time). Moreover, although some IAA scores could not be increased or even decreased, IAA scores of the most frequent classes (DRUG, ACTIVEING, STRENGTH, FORM, FREQUENCY, DURATION) achieved sufficient quality ($0.75 - 0.98$).

One reason for the challenging synchronization could be rooted in the different educational backgrounds of the annotators. Generally, medical students and medical informatics students with experiences in clinical routine shared higher IAA scores. This was apparent for the REASON class, as this information demanded a more profound clinical knowledge. The DOSAGE class achieved lowest IAA scores as DOSAGE entities were easily confused as STRENGTH entities (e.g. *max Zufuhr von 4 IE Insulin/h: 4 IE* repeatedly annotated as DOSAGE; *Torasemid 5 mg 1-1-0: 5 mg* repeatedly annotated as *Dosage*). In addition, STRENGTH was occasionally annotated as DOSAGE in the structured medication sections. In general, DOSAGE entities are rarely represented in the corpus; thus, these results are not fully representative.

At the end of our annotation iterations, we had to face another challenge. To increase heterogeneity of the doctor's letters in each iteration, we did not just randomly select letters from the complete corpus but restricted each sampling for each iteration to a specific time period of patient recruitment. For example; iteration 1 contained the first ten letters from the beginning of the projects recruiting phase, while iteration 2 contained letters from the middle of this phase. This resulted in a bias, as the patient recruitment was sometimes dominated by in-patients, out-patients or patients from the chest pain unit. The structure of these letters can vary significantly, therefore the medication information in each iteration batch varied in its notational form. For future annotation projects, we recommend a more balanced distribution of such letters in each iteration in order to improve agreement between different iterations. Due to the non-satisfying IAA scores of REASON, ROUTE and DOSAGE, shortly after publishing the initial corpus version 1.0 we conducted a revision annotation project. All medication annotations were thoroughly reviewed by two experienced physicians working in the cardiology department at Heidelberg University Hospital. To measure IAA we selected 150 discharge letters in two iterations of 100 and 50 letters. We could substantially improve IAA scores for all medication information classes, but the FORM class and published CARDIO:DE v.1.1 in 2025 (cf. Table 4.5).

The most recent CARDIO:DE corpus version 1.1 contains in total 27,155 annotated medication information entities (Table 4.6) with 19,336 medication relations. The most frequent medication information classes are related to ACTIVEING/DRUG and their relations FREQUENCY and STRENGTH. Figure 4.6 illustrates the most frequently annotated Drug entities in CARDIO:DE.

Class	F_1-score (v. 1.0)	F_1-score (v. 1.1)
ACTIVEING	0.76	**0.93**
DOSAGE	0.33	**0.73**
DRUG	0.80	**0.84**
DURATION	0.75	**0.87**
FORM	**0.88**	0.82
FREQUENCY	0.98	**0.98**
REASON	0.41	**0.74**
ROUTE	0.61	**0.92**
STRENGTH	0.89	**0.98**

Table 4.5 **IAA medication information of CARDIO:DE v. 1.0 versus v. 1.1:** Median F_1-scores per medication information class of CARDIO:DE v. 1.0 and, after revision v. 1.1.

Type	CARDIO:DE400	CARDIO:DE100
Medication counts		
ACTIVEING	6,032	1,508
DOSAGE	127	16
DRUG	1,747	448
DURATION	1,284	312
FORM	152	29
FREQUENCY	5,135	1,366
REASON	1,631	359
ROUTE	452	115
STRENGTH	5,071	1,371
Total	**21,631**	**5,524**
Relation counts		
DOSAGE	130	16
DURATION	1,487	363
FORM	155	30
FREQUENCY	5,231	1,427
REASON	2,116	486
ROUTE	1,058	327
STRENGTH	5,093	1,417
Total	**15,270**	**4,066**

Table 4.6 **Medication information statistics CARDIO:DE v. 1.1.:** Entity and relation counts per CARDIO:DE v. 1.1. split.

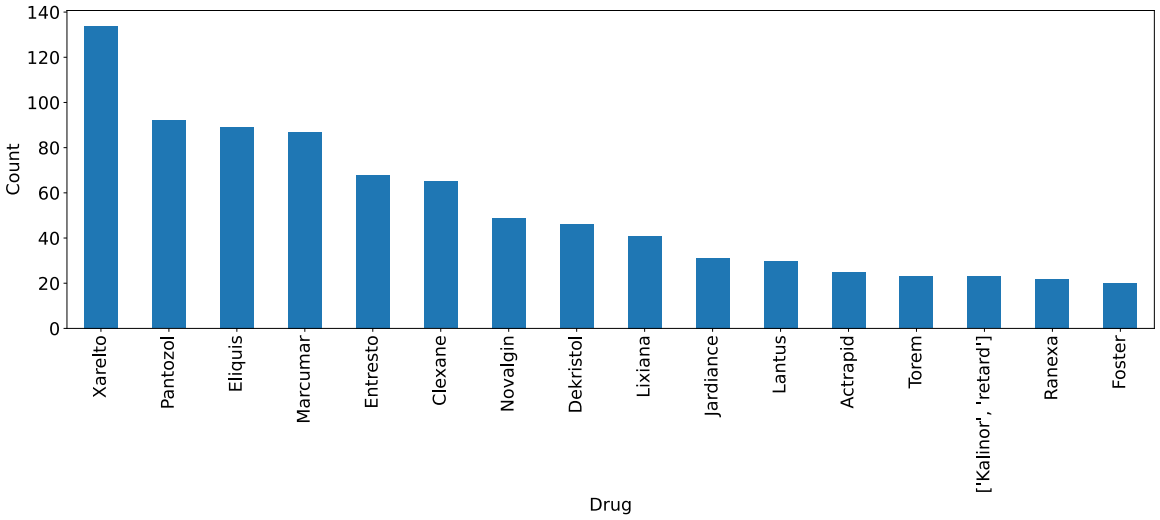


Fig. 4.6 **Drug entity counts:** Counts of Drug entity annotations in CARDIO:DE v. 1.1 (complete) corpus, if count ≥ 20 entities.

Section type annotation

Our section type annotation scheme is based on (Lohr et al. 2018b), but is more coarse-grained and carried out on paragraph-level (guidelines cf. (Richter-Pechanski et al. 2023)). To measure the quality of annotations we calculated an IAA using Krippendorff's alpha. Krippendorff's alpha is a chance corrected IAA and can be used for any number of annotators and class labels (Krippendorff 2004). We performed three iterations for redundant annotation with three annotators (two clinical data scientists researching on clinical routine documents at the cardiology department, and one research student assistant studying computational linguistics (B.A.) in sixth semester) containing 35, 30 and 20 documents. The project lifetime, including preparation, annotation, and evaluation, was two months. The approximate annotation time per document was 3–8 min.

We annotated fourteen section types. Nine section types are mapped to HL7 CDA elements.[2] Sections related to diagnosis are not mapped to CDA elements. The CDA standard separates diagnosis sections into ENTLASSUNGSDIAGNOSE (discharge diagnosis) and AUFNAHMEDIAGNOSE (admission diagnosis). Neither of them is explicitly part of doctor's letters in CARDIO:DE. After consultation with cardiologists, we decided to use the most representative heading names in the original doctor's letters as section class labels. There are typically two section headings related to diagnosis: (1) AKTUELLDIAGNOSEN: This section contains discharge diagnosis information and is part of most of the letters.

[2] ARZTBRIEF PLUS, v. 3.15, https://wiki.hl7.de/index.php?title=IG:Arztbrief_Plus, accessed 08.09.2025.

German Section Header	CDA ID
ANREDE *(Salutation)*	1.2.276.0.76.10.3001
AKTUELLDIAGNOSEN *(Current diagnosis)*	—
DIAGNOSEN *(Diagnosis)*	—
ALLERGIENUNVERTRÄGLICHKEITENRISIKEN *(Allergies intolerances risks)*	1.2.276.0.76.10.3028
ANAMNESE *(History of present illness)*	1.2.276.0.76.10.3022
AUFNAHMEMEDIKATION *(Admission medication)*	1.2.276.0.76.10.3029
KUBEFUNDE *(Physical examination)*	—
BEFUNDE *(Results)*	1.2.276.0.76.10.3100
ECHOBEFUNDE *(Echocardiographic findings)*	—
LABOR *(LaboratoryResultObservation)*	1.2.276.0.76.10.4254
ZUSAMMENFASSUNG *(Hospital course)*	1.2.276.0.76.10.3021
MIX *(Other)*	—
ENTLASSMEDIKATION *(Discharge medication)*	1.2.276.0.76.10.3031
ABSCHLUSS *(Final remarks)*	1.2.276.0.76.10.3034

Table 4.7 **CARDIO:DE section types:** CDA code for each section type, if available (English translation in brackets).

Iteration	Krippendorff's α	Std. dev.
1	0.91	0.06
2	0.89	0.17
3	0.96	0.07

Table 4.8 **CARDIO:DE section type IAA:** Median Krippendorff Alpha IAA score per iteration for the section annotation layer including standard deviation.

(2) DIAGNOSEN: This section type contains admission or discharge diagnosis information. In original documents in MS DOC format, important diagnosis information is commonly written as bold type. This information is not part of the documents in CARDIO:DE. After consultation with physicians, in addition to CDA section type BEFUNDE, we annotated section types KUBEFUNDE and ECHOBEFUNDE. Both appear frequently in CARDIO:DE letters and are considered relevant for cardiovascular clinical routine and research. Sections and paragraphs which cannot be mapped to one of the thirteen section types listed in Table 4.7 are annotated with the generic section type MIX.

Annotation quality During iterative redunant annotation we could increase IAA from a Krippendorff's alpha score of 0.91 to 0.96 (Table 4.8).

Section class	CARDIO:DE400	CARDIO:DE100	CARDIO:DE
ABSCHLUSS	2,802	694	3,496
AKTUELLDIAGNOSEN	3,250	694	3,944
ALLERGIEN	1,033	236	1,269
ANAMNESE	1,188	281	1,469
ANREDE	405	100	505
AUFNAHMEMEDIKATION	2,056	593	2,649
BEFUNDE	9,635	2,519	12,154
DIAGNOSEN	4,724	1,044	5,768
ECHOBEFUNDE	1,299	290	1,589
ENTLASSMEDIKATION	4,138	1,034	5,172
KUBEFUNDE	4,199	1,105	5,304
LABOR	55,684	12,220	67,904
MIX	945	242	1,187
ZUSAMMENFASSUNG	3,644	843	4,487
Total	**95,002**	**21,895**	**116,897**

Table 4.9 **CARDIO:DE section statistics:** Section counts (total and per section type) per CARDIO:DE split. Note: ALLERGIEN abbreviates the class ALLERGIENUNVERTRÄGLICHKEITENRISIKEN.

The final corpus contains in total 116,897 annotated paragraphs with section classes. The most frequent section class was LABOR and BEFUNDE. BEFUNDE is a meta class, containing all kinds of findings, excluding KUBEFUNDE and ECHOBEFUNDE. LABOR contains laboratory information in a flattened tabular format. The least annotations are related to the class ANREDE. This includes typically a single introductory sentence at the top of a doctor's letter, containing information of the patient and the receiving department (Table 4.9).

Section annotation was performed at paragraph level. In the final annotation schema, annotators marked only the first paragraph of each new section type. The beginning of the next section type implicitly defined the end of the previous section. Accordingly, all following paragraphs were assigned to that section class until the next section start was annotated. This procedure did not reduce the relevant annotation information, but substantially simplified the annotation process. Initially, we annotated every paragraph belonging to a section span. However, this resulted in slow performance of the INCEPTION tool and made the annotation process unnecessarily time-consuming and laborious (cf. challenges 1.a. *Domain expertise*, 1.b. *Staff time* and 1.c. *Compute restrictions*). We therefor changed the schema and continued with annotating only the first paragraph of each new section. This change significantly increased annotation speed, but in iteration 2 it also introduced a new

annotation error: annotators occasionally missed the beginning of a new section, especially for section classes embedded in BEFUNDE. In particular, the end of ECHOBEFUNDE or ALLERGIENUNVERTRÄGLICHKEITENRISIKEN sections was sometimes not annotated correctly, causing these sections to extend over too many paragraphs. After detailed review meetings, this problem was substantially reduced in iteration 3.

Other issues involved how to define the section type ENTLASSMEDIKATION. Some doctor's letters contained two explicit medication sections: section type AUFNAHMEMEDIKATION at the beginning and section type ENTLASSMEDIKATION at the end. ENTLASSMEDIKATION was frequently introduced by the header *Aktuelle Medikation*. But a couple of letters only contained a single medication section at the beginning of a document. Therefore, some annotators interpreted these sections as AUFNAHMEMEDIKATION, even though they are introduced with the header *Aktuelle Medikation*. After review meetings and consultations with physicians, we updated the definition of the initial medication section with this header as ENTLASSMEDIKATION within the guidelines.

Summary

Our IAA scores for medication information and section classes are comparable to previously published similar annotation projects: e.g. IAA from I2B2 corpus for medication information on token-level F_1: $0.82-0.88\%$ (Uzuner et al. 2010b), IAA for section types, median Krippendorff alpha: for seven classes: 0.85%, $11-21$ classes: $0.70-0.84\%$ (Lohr et al. 2018b). However, these IAAs are not completely comparable, due to different pre-processing steps and measurement procedures. Since the corpus is freely available for research purposes, we encourage the community to improve existing annotations and add new annotation layers to CARDIO:DE. Our internal team has already upgraded CARDIO:DE from version 1.0 to 1.1 by thoroughly revising medication information annotations. Moreover, in 2025, through the generous contribution of colleagues of the *Bavarian Center for Digital Health and Social Care* the corpus annotation schema was extended by standardized token-level labels such as DIAGNOSIS, THERAPY, and MEDICAL FINDINGS (Becker et al. 2025).

4.5 Baselines and technical validation

We used baseline classifiers to demonstrate what well-established and publicly available machine learning models (supervised statistical models and encoder-based BERT models) could achieve out-of-the-box on CARDIO:DE v.1.0 annotations. We trained our classifiers on annotations of CARDIO:DE400 and assessed their performance on annotations of CARDIO:DE100. For these models we performed neither hyperparameter tuning, nor

architecture optimization. The results were intended to give a first impression of possible applications and how to train MIE models on CARDIO:DE annotations for both tasks.

4.5.1 Medication information extraction

As an example use case, we evaluated a statistical and a neural encoder-based model for medication information extraction on CARDIO:DE v. 1.0 (for a detailed entity and token-count statistics of CARDIO:DE v. 1.0, cf. (Richter-Pechanski et al. 2023)). Our statistical model is based on a CRF (Lafferty et al. 2001). The CRF uses basic linguistic features (similar to CCE exraction in Subsection 2.6.2). Our neural model is based on a well-documented Hugging Face BERT language model for NER, pre-trained on different publicly available German language corpora (Devlin et al. 2019; Chan et al. 2020) (further hyperparameters, see Section A.2).[3]

In addition, we evaluated a freely available German GGPONC NER classifier[4], trained on SYSTEMATIZED NOMENCLATURE OF MEDICINE (SNOMED CT) annotations in GGPONC version 2.0 (Borchert et al. 2022).[5] The fine-grained schema includes the entity type CLINICAL DRUG. For a detailed description of this class, see the GGPONC 2.0 guidelines.[6] CLINICAL DRUG describes a pharmaceutical product produced for diagnostic or therapeutic purposes. While the guidelines do not exactly match our definition of DRUG/ACTIVEING we follow two mappings during evaluation, (1) mapping CLINICAL DRUG to our DRUG/ACTIVEING (short) and (2) to DRUG/ACTIVEING/FREQUENCY/STRENGTH (long).

The objective of this task is to conduct NER by assigning a set of six medication information classes (ACTIVEING, DRUG, DURATION, FORM, FREQUENCY, STRENGTH) to each input token. Medication information can consist of one single token or a sequence of tokens. Table 4.10 shows a tokenized input snippet, containing 20 token and their assigned medication information classes.

We evaluated this task using the F_1-score, the harmonic mean between precision and recall per class, and the micro-average F_1-score per classifier (Table 4.11). We evaluated token-wise and entity-wise (for further details on NER and the IOB schema, cf. Section 2.3.2). For token-wise we removed all *B*- and *I*- substrings from the labels before calculating

[3] `deepset/gbert-base`.

[4] `04_ggponc_fine_long`.

[5] https://confluence.ihtsdotools.org/display/DOCSTART/6.+SNOMED+CT+Concept+Model, accessed 09.09.2025.

[6] https://github.com/hpi-dhc/ggponc_annotation/blob/master/annotation_guide/anno_guide.pdf, accessed 09.09.2025.

Token	**Tag**
Medikation	O
bei	O
Aufnahme	O
:	O
Ramipril	B-ActiveIng
10	B-Strength
mg	I-Strength
1	B-Frequency
-	I-Frequency
0	I-Frequency
-	I-Frequency
0	I-Frequency
,	O
HCT	B-ActiveIng
25	B-Strength
mg	I-Strength
1	B-Frequency
-	I-Frequency
0	I-Frequency
-	I-Frequency
0	I-Frequency

Table 4.10 **Token–tag alignment for the example sequence:** An example input sample - *Medikation bei Aufnahme: Ramipril 10mg 1-0-0, HCT 25mg 1-0-0* and the gold sequence of tags per token.

Class type	CRF			BERT		
	Pr	**Re**	F_1	**Pr**	**Re**	F_1
ActiveIng	0.84	0.83	0.83 (0.60)	0.80	0.91	**0.85** (0.86)
Drug	0.80	0.75	0.77 (0.77)	0.81	0.87	**0.84** (0.81)
Duration	0.80	0.73	0.77 (0.60)	0.78	0.89	**0.83** (0.59)
Form	0.47	0.33	0.39 (0.41)	0.57	0.71	**0.63** (0.60)
Frequency	0.97	0.97	0.96 (0.94)	0.96	0.98	**0.97** (0.94)
Strength	0.93	0.96	0.94 (0.92)	0.93	0.97	**0.95** (0.93)
Micro avg.	0.92	0.91	0.92 (0.87)	0.90	0.95	**0.93** (0.88)

Table 4.11 **Results medication information extraction:** Token-wise precision (Pr), recall (Re) and F_1-score (F_1) results for medication information extraction per class and per model, including entity-wise F_1-score in brackets and the micro-average F_1-score in the last row.

Class type	Precision	Recall	F_1-score
Drug (short)	0.13	0.67	0.21
Drug (long)	0.81	0.80	0.81

Table 4.12 **Results medication information extraction GGPONC:** Token-wise precision, recall and F_1-scores on CARDIO:DE100 for CLINICAL DRUG class of GGPONC NER model.

precision and recall. Token-wise results for precision, recall and F_1-score of the GGPONC NER classifier are listed in Table 4.12.

Considering token-wise and entity-wise micro-average F_1-score, BERT shows slightly higher results than the CRF. However, the CRF often outperforms the neural model in terms of precision. Regarding token-wise F_1-score BERT outperforms the CRF over all classes. This is also true for the low-frequency class FORM. Furthermore, this class achieved the overall lowest F_1-score for both models.

Results of GGPONC NER show the highest F_1-score for the long mapping (0.81), along with a balanced precision and recall score. The short mapping shows an overall much lower F_1-score (0.21) along with a much lower precision (0.13) than recall score (0.67). Considering different domains (oncology vs. cardiology) and document types (guidelines vs. doctor's letters), the cross-domain GGPONC NER baseline showed impressive results on the CARDIO:DE100 corpus split. The short mapping achieved a recall of 0.67, while the precision score was only 0.13. This was particularly due to issues while mapping our medication information classes based on our guidelines to the more generic GGPONC SNOMED CT class CLINICAL DRUG. We frequently observed, that GGPONC NER annotates token sequences

such as *20 mg* (STRENGTH) or *1-0-0* (FREQUENCY) as CLINICAL DRUG, resulting in a high amount of false positives. Hence, our long mapping using the more comprehensive mapping of CLINICAL DRUG to four classes DRUG/ACTIVEING/FREQUENCY/STRENGTH achieved both a high precision score (0.81) and a high recall score (0.80). These results show the potential of cross-domain models for German MIE. We therefore leave it to future work, to systematically evaluate the performance of publicly available German MIE models on already distributable German medical corpora (for further evaluation results, see Section A.4). Current SOTA results for medication information extraction on English datasets achieve up to 0.95 F_1-score (Hahn et al. 2020). Considering different data sets and hyperparameters a comparable German model for medication extraction for the classes ACTIVEING/DRUG (CLINICAL DRUG) recognition in the GGPONC 2.0 corpus achieves 0.91 F_1-score, thus, outperforms our more fine-grained baseline distinguishing between DRUG (0.81) and ACTIVEING (0.86).

4.5.2 Section classification

Equal to the medication information extraction task, we evaluated a statistical and a neural model for section classification on CARDIO:DE v. 1.0 (for a detailed class count statistics of CARDIO:DE v. 1.0, cf. Section 5.3.1 and (Richter-Pechanski et al. 2023)). For the statistical model we opted for a SVM (Cortes et al. 1995). Our neural model is based on a well-documented Hugging Face encoder-based BERT language model for sequence classification, pre-trained on different publicly available German language corpora (`deepset/gbert-base`) (Devlin et al. 2019; Chan et al. 2020). The objective of this task was to assign a set of fourteen section types (ANREDE, AKTUELLDIAGNOSEN, DIAGNOSEN, ALLERGIENUNVERTRÄGLICHKEITENRISIKEN, ANAMNESE, AUFNAHMEMEDIKATION, KUBEFUNDE, BEFUNDE, ECHOBEFUNDE, LABOR, ZUSAMMENFASSUNG, MIX, ENTLASSMEDIKATION, ABSCHLUSS) to each input sample. An input sample consists of a paragraph of text (a paragraph is defined by the MS WORD "¶" character) extracted from a doctor's letter with no further context information. Table 4.13 shows a tokenized example of an input sample, containing 48 tokens assigned to the ANREDE class. We evaluated this task using the F_1-score per class and the macro-average F_1-score per classifier (Table 4.14).

Analyzing the macro-average F_1-score the BERT model outperforms the baseline by 0.02 pp. Taking the per class F_1-score into account, BERT achieves a better score in nine section classes. In four section classes both models achieve the same score, while for the class EntlassMedikation SVM achieves a higher F_1-score.

Both models worst performing classes are related to medication sections. We observe a very low recall score for AUFNAHMEMEDIKATION for the SVM and for ENTLASSMEDIKA-

Input sample	Section
`['über', 'Ihren', 'Patienten', 'B', '-', 'SALUTE', 'B', '-', 'PER', 'I', '-', 'PER', 'geboren', 'am', '<', '[', 'Pseudo', ']', ',', '24', '/', '06', '/', '1977', '>', ',', 'wohnhaft', 'in', 'B', '-', 'PLZ', 'B', '-', 'LOC', 'I', '-', 'ADDR', 'I', '-', 'ADDR', 'der', 'sich', 'vom', 'bis', 'in', 'unserer', 'stationären', 'Behandlung', 'befand', '.']`	ANREDE

Table 4.13 **Training sample for section classification:** Tokenized input sample (left column) including its section class (right column). The machine learning model assigns a single section class to a given input sample.

Section type	SVM			BERT		
	Pr	**Re**	F_1	**Pr**	**Re**	F_1
ANREDE	0.99	1.00	0.99	0.99	1.00	0.99
AKTUELLDIAGNOSEN	0.73	0.51	0.60	0.67	0.67	**0.67**
DIAGNOSEN	0.69	0.78	0.73	0.78	0.79	**0.79**
ALLERGIENUNVERTRÄGLICHKEITENRISIKEN	0.97	0.94	0.95	0.97	0.96	**0.96**
ANAMNESE	0.90	0.81	0.85	0.81	0.93	**0.87**
AUFNAHMEMEDIKATION	0.92	0.10	0.18	0.43	0.95	**0.59**
KUBEFUNDE	0.98	0.97	0.98	0.98	0.97	0.98
BEFUNDE	0.84	0.80	0.82	0.95	0.78	**0.86**
ECHOBEFUNDE	0.89	0.89	0.89	0.85	0.96	**0.90**
LABOR	0.97	1.00	0.98	0.98	0.99	0.98
ZUSAMMENFASSUNG	0.90	0.95	0.92	0.94	0.94	**0.94**
MIX	0.87	0.64	0.74	0.65	0.88	**0.75**
ENTLASSMEDIKATION	0.64	0.91	**0.75**	0.79	0.26	0.39
ABSCHLUSS	0.99	0.98	0.98	0.97	0.98	0.98
Macro avg.	0.88	0.80	0.81	0.84	0.86	**0.83**

Table 4.14 **Results section classification:** Precision (Pr), recall (Re) and F_1-score (F_1) results per class and macro-average F_1-score per model for section classification.

TION for the BERT model. The SVM frequently classifies AUFNAHMEMEDIKATION instances as ENTLASSMEDIKATION. The BERT model, on the other hand, frequently misclassifies ENTLASSMEDIKATION instances as AUFNAHMEMEDIKATION, but to a lesser extent (confusion matrices for both models, see Figure A.2 and Figure A.3). Due to the fact that this task was performed only as a simple text classification task without further context information, these errors cannot be easily avoided, only by merging these class types or by adding context information to each input sample (for context enriched section classification, cf. Chapter 5).
At the time of the experiments, our baseline classifiers, considering different datasets, languages and annotation guidelines, our final macro-average F_1-scores for SVM and BERT were comparable to similar published section classification results (Lohr et al. 2018b).

4.6 Data accessibility

CARDIO:DE must be formally requested via the open research database HEIDATA following three steps (Richter-Pechanski et al. 2022):

1. Sending a data request mail to HEIDATA (`data@uni-heidelberg.de`) including a signed data usage agreement (DUA) form, addressing information about correct data usage and security standards. Under this licence, it is clearly prohibited, to identify individuals or try to contact or advertise them.

2. Including a group description and a project description (details, see below).

3. After a positive decision by the CARDIO:DE study director Christoph Dieterich (CARDIO:DE supervisor), the data requestor will receive detailed instructions how to download the corpus via HEIDATA.

Data approval requires at least one week. The data request needs to contain the following information:

- A signed DUA, signed by each data user individually.

- A group description including the requestor's (data user's) name, affiliation, position and email address and website of the institution.

- A project description of the research purpose (max. 150 words).

- Name, affiliation, position, email address and signature of the responsible person to administer and manage the infrastructure on which the corpus will be stored.

The data request gets approved, if it contains all requested information and if it is in line with the DUA. The data request will be rejected, if it contains incomplete or incorrect information or if it violates regulations of the DUA.

4.7 Conclusion

With CARDIO:DE we demonstrate the feasibility of distributing a clinical routine corpus under strict European and German data protection regulations. The corpus is the first and, since its release in 2023, remains the only native-language German clinical routine corpus containing 500 structurally coherent doctor's letters covering a broad range of cardiovascular clinical routine documents. The corpus was collected and prepared entirely within the hospital infrastructure. We share CARDIO:DE via a request-based DUA process to enable compliant and reproducible research. As the corpus is the main data source for our core experiments in this thesis in Chapter 5 and 6 and supports downstream evaluation on real-world data in Chapter 7 its contributions to our clinical constraints (cf. Section 2.1) and research questions are often indirect, enabling further research rather than standalone. Nevertheless, to emphasize its impact on our experiments, below we will explicitly summarize the role of CARDIO:DE along our clinical constraints (Section 2.1) and RQs (Section 1.2).

Domain expertise and staff time (addressing RQ 1, RQ 3)
We distribute the corpus with two curated annotation layers (section classes, medication information), initial benchmark results and detailed annotation guidelines. This allows the research community to conduct on-premise machine learning experiments with high-quality gold standard data rather than building datasets and pipelines from scratch, thereby reducing the demand for domain expertise and staff time.

Local compute resources (addressing RQ 1, RQ 2)
While the corpus is not directly contributing to reduce compute demands, CARDIO:DE makes German clinical routine texts and annotations available outsite our department. A DUA access allows research groups to run their own on-premise experiments on their local compute infrastructure across token- and paragraph-level tasks.

Native-language barrier (addressing RQ 1, RQ 2)
We provide the first German clinical corpus with annotations and benchmarks across task complexities (paragraph- and token-level MIE), enabling native-language training and evaluation of MIE tasks on authentic German cardiology doctor's letters.

Transparency requirements (addressing RQ 4)
CARDIO:DE contains carefully de-identified real-world doctor's letters from clinical routine. We store the data on the open research database HEIDATA and make it available via a clearly defined DUA process supporting a reproducible and transparent research. Furthermore, our prospective GDPR-compliant study design can serve as a template for other hospitals to create new clinical datasets.

Chapter 5

Clinical Section Classification using Pretrained Language Models and Prompting

5.1 Outline and contributions

This chapter is based on our published work presented in (Richter-Pechanski et al. 2024). We evaluated prompt-tuned encoders for multi-class section classification on paragraphs in German doctor's letters (CARDIO:DE). We compared general-domain German PLMs, domain- and task-adapted PLMs and medical PLMs. We evaluated PET and its prompt optimizing tool pattern-exploiting training with automatic labels (PETAL) in few-shot settings with null prompts and contextualized prompts using Shapley values for data engineering and model analysis.

Section 5.2 motivates the task and presents solutions to clinical constraints presented in Chapter 2. Section 5.3 introduces key methodologies, including PET, selected PLMs and Shapley values. Section 5.4 presents the data used for pre-training and fine-tuning experiments. Section 5.5 defines metrics, data preparation and experimental design. Section 5.6 presents all section classification results including in-depth evaluations using Shapley values. Section 5.7 finalizes with a comprehensive discussion of the results with respect to clinical constraints and research questions. The chapter concludes this study in Section 5.8.

The chapter contributes to RQ 1 by comparing various further-pretraining and fine-tuning strategies in a few-shot setup using encoder PLMs in different sizes (110 and 340M parameters) on German language clinical texts completely inside clinical compute infrastructure (cf. challenges 1.a. *Domain expertise*, 1.b. *Staff time*, 1.c. *Compute restrictions* and 1.d. *Native*

language).
We contribute to RQ 2 by showing that resource-aware pretraining and prompting strategies for small-sized encoders are sufficient for a German section classification task (cf. challenges 1.c. *Compute restrictions* and 1.d. *Native language*). We reduce manual efforts during model development by applying few-shot learning methods (PET), automatic prompting methods (PETAL and null prompts) and data sampling, hence directly contributing to RQ 3 (cf. challenges 1.a. *Domain expertise* and 1.b. *Staff time*). Finally, we contribute to RQ 4 by leveraging Shapley value attributions for an encoder-based classification task to support error analysis, data sampling and model transparency (cf. challenge 2. *Transparency*). All relevant code of this project is published on GitHub.[1]

5.2 Introduction and background

Doctor's letters are typically divided into sections, such as anamnesis (patient medical history), diagnosis or medication, containing semantically related sentences. Typically, it is not necessary to consider all sections to obtain specific medical information (Richter-Pechanski et al. 2021) or medication information (Uzuner et al. 2010a). Instead, MIE tasks, such as medication extraction or patient cohort retrieval, can be improved by contextualizing the information in a doctor's letter (Edinger et al. 2018). However, automatic section classification is non-trivial due to a high variability of the structuring of information across physicians and time periods (Lohr et al. 2018b).

We aim to evaluate an NLP model to classify paragraphs in doctor's letters to their corresponding clinical section and evaluate best-practice strategies to identify an ideal setup to address our clinical constraints discussed in Section 2.1. Specifically, for each constraint, we identify and propose the following solutions:

1. **On-premise resource constraints**

    1.a. **Domain expertise** We reduce the demand for clinical expertise in MIE by exploiting existing domain knowledge available in hospitals, such as clinical routine documents. We evaluate domain- and task-adapted (Gururangan et al. 2020) general-domain PLMs, as well as PLMs pretrained on clinical data from scratch (Bressem et al. 2024) in combination with prompt-based learning methods (Schick et al. 2021a), which require only limited training data.

[1] see, https://github.com/dieterich-lab/section_classification_pet/, accessed: 05.12.2025.

1.b. **Staff time** To reduce time investment and costs of manual data annotation through clinical staff, we apply few-shot learning (Lake et al. 2015) and context-enriched training data using prompt-based fine-tuning with PET + PETAL (Schick et al. 2021a; Schick et al. 2020) and compare the results with supervised sequence classification methods. We further evaluate the feasibility of null prompts (Logan et al. 2022), which have been shown to alleviate the search for effective prompts while achieving improved results.

1.c. **Local compute resources** While LLMs have shown impressive medical capabilities (Singhal et al. 2023), in 2023 their demands of compute power along with unsolved issues regarding automatic evaluation, faithfulness control, and trustworthiness make their use in clinical contexts often impractical (Parnami et al. 2022; Thirunavukarasu et al. 2023). We, therefore, focus on smaller PLMs (110 and 340 million learnable parameters) in a few-shot learning setting. Notably, prompt-based fine-tuning already achieved higher accuracy with smaller, encoder-based PLMs compared to PLMs fine-tuned for sequence labeling with a full-fledged training dataset in German (Schick et al. 2021a).

1.d. **Native-language barrier** We leverage internal clinical documents in German language for domain- and task-adaptation (Gururangan et al. 2020) of publicly available PLMs mostly pre-trained on English data, as well as PLMs pretrained on German clinical data from scratch (Bressem et al. 2024).

2. **Transparency requirements** To address the need for transparent and trustworthy model predictions in clinical routine, we used well-established masked-language-models. They allow application of SOTA interpretability methods out-of-the-box that rely on saliency features computed with, for example, Shapley values (Lundberg et al. 2017), to explain our model predictions.

In what follows we conducted in-depth evaluations of these proposed solutions in a real-world section classification task, applied to German doctor's letters from the cardiovascular domain. To our knowledge, this was the first in-depth evaluation of a prompt-based fine-tuning method such as PET on real-world clinical routine data in German language.

Section type	CARDIO:DE400 Training set	CARDIO:DE100 Test set
ANREDE (*Salutation*)	402	99
DIAGNOSEN (*Diagnosis*)	8,023	1,738
ALLERGIEN (*Allergies*)	1,031	236
ANAMNESE (*Patient Medical History*)	1,188	281
MEDIKATION (*Medication*)	6,148	1,627
BEFUNDE (*Findings*)	15,396	3,914
ZUSAMMENFASSUNG (*Summary*)	3,645	843
MIX (*Mix*)	945	242
ABSCHLUSS (*Closing Remarks*)	2,805	695
Total	**39,583**	**9,675**

Table 5.1 **Distribution of section classes:** Number of samples per section class per corpus split. English translation in round brackets.

5.3 Methods

5.3.1 Patter-exploiting training

In our experiments, we systematically evaluated methods for few-shot learning, that is, using minimal training data, in a lower-resource domain and language, in our case German clinical routine (Hahn et al. 2020; Jantscher et al. 2023; Idrissi-Yaghir et al. 2024). Specifically, we evaluate PET, a semi-supervised prompting method optimized for few-shot learning scenarios (Schick et al. 2021a) which is designed to recast classical text classification or information extraction tasks as a language modeling problem. In our study, we classify paragraphs of German doctor's letters (CARDIO:DE v. 1.0) into a set of nine section categories (Table 5.1). The objective is to accurately categorize, for instance, a paragraph such as *The patient reports pressure pain in the left chest* under the section class ANAMNESE.

To perform PET experiments, we need a pre-trained masked language model M with a vocabulary V, a few-shot data set with training instances $x_i \in X$ and target labels $y_i \in Y$. We further need a pattern function P that maps instances to a set of cloze sentences (templates) $P : X \mapsto V^*$, and a verbalizer function $v : Y \mapsto V$ that maps each label to a single token from the vocabulary of M.

The PET workflow contains three basic steps (cf. Figure 2.1 in Section 2.4.2): (1) applying P to each input instance x_i and fine-tune a model M for each template to obtain the most likely token for the *MASK* token $v(y)$, (2) use the ensemble of fine-tuned models M from the

previous step and annotate a large unlabeled dataset D with soft labels and (3) train a final classifier C with a traditional sequence classification head on the labeled dataset D.

In Figure 2.1 we follow a running example of a PET workflow. (Step 1) We select the input sample *ASS 100mg 1-0-0* as x_i. The goal is to classify this input sample as either MEDIKATION or ANAMNESE. Next, we select as a template `SAMPLE Sektion: [MASK]` and apply a pattern function P to the x_i and get the final input to our model M (*ASS 100mg 1-0-0 Sektion: [MASK]*). Now, the objective for the model is to predict the most likely [MASK] token given its vocabulary V, e.g. *Arznei*. As this token is not equivalent to our set of classes (MEDIKATION or ANAMNESE) we use a verbalizer function $v : Y \mapsto V$ to map our categories to the output token *Arznei*. In our example the class MEDIKATION is mapped to the token *Arznei*. The PET workflow fine-tunes a separate model M for each created template on a few-shot data set X. (Step 2) After fine-tuning, we use this ensemble of prompt-based fine-tuned models to annotate a large unseen dataset D with soft labels. (Step 3) We then train a final classifier C with a standard sequence classification head on this soft labeled dataset D.

Creating templates

Template engineering is a crucial hyperparameter in a PET experiment. For the *core experiments* we used four different template types (including examples and English translations (in brackets)):

- Null prompt: `SAMPLE [MASK]`
  Keine peripheren Ödeme `[MASK]`
  (*No peripheral edema* `[MASK]`)
- Punctuation: `SAMPLE : [MASK]` and `SAMPLE - [MASK]`
  Keine peripheren Ödeme : `[MASK]`
  (*No peripheral edema :* `[MASK]`)
- Prompt: `SAMPLE Sektion [MASK]`
  Keine peripheren Ödeme Sektion `[MASK]`
  (*No peripheral edema Section* `[MASK]`)
- Q&A: `SAMPLE Frage: Zu welcher Sektion gehört dieser Text? Antwort: [MASK]`
  Keine peripheren Ödeme Frage: Zu welcher Sektion gehört dieser Text? Antwort: `[MASK]`

(*No peripheral edema Question: To which section does this text belong? Answer:* `[MASK]`)

To minimize engineering costs we also evaluated the feasibility of using exclusively null prompts, by removing all tokens from prompt templates, as proposed by (Logan et al. 2022) (cf. challenges 1.a. *Domain expertise* and 1.b. *Staff time*). We defined three null prompt templates: (1) `SAMPLE [MASK]`; (2) `[MASK] SAMPLE` and (3) `[MASK] SAMPLE [MASK]`.

Verbalizer

Defining the verbalizer token can be tedious, because domain knowledge and technical expertise about the used PLM is required. This can be a significant issue, as such a comprehensive knowledge is uncommon in the clinical setting. Moreover, PET restricts the verbalizer token to a single token. Hence, an appropriate and intuitive token may not be applicable for a label mapping, if it is not included in the PLM's vocabulary. For instance, the word ANAMNESE is not part of the `gbert` vocabulary. This makes a verbalizer search for clinicians quite challenging. Therefore, we use PET with automatic labels (PETAL) for all our experiments, except for the zero-shot baselines (Schick et al. 2020). This can reduce engineering costs and makes our experimental setup more comparable and reproducible (cf. challenges 1.a. *Domain expertise* and 1.b. *Staff time*). As visualized in Figure B.2 PETAL calculates the most likely verbalizer token per label, given the few-shot training data for each pattern and given a PLM. We created a separate verbalizer for each few-shot size for each training set.

5.3.2 Pretrained language models

To evaluate the feasibility of exploiting existing clinical domain knowledge by further-pretraining, we used a set of three language models, all based on the BERT architecture (Devlin et al. 2019) and available at HUGGINGFACE Hub: (1) `deepset/gbert-base` (Chan et al. 2020), (2) `deepset/gbert-large`, (3) `Smanjil/German-MedBERT` (Bressem et al. 2024). The largest model `gbert-large` contains 340 million parameters. In our clinical infrastructure (year: $2022-23$), which contains a maximum of two NVIDIA RTX6000 GPUs or a single NVIDIA A40, we were able to perform all further-pretraining experiments within a reasonable timeframe (cf. Section B.3). Compared to current foundation LLMs with billions of parameters, we consider these models as lightweight (cf. challenge 1.c. *Compute restrictions*).

For both `gbert` and `medbertde` we create medical-adapted variants by further-pretraining, as proposed by (Gururangan et al. 2020) to assess the impact of different pretraining datasets

on section classification results (Figure 5.1). We defined datasets for three different pretraining approaches:

1. *task-adaptation* Using CARDIO:DE (cf. Section 4). This data set contains unlabeled data extracted from the same source as the training and test data of the section classification task. It is relatively small, only 5.8*MB* (megabytes). (PLMs appended with suffix `-task`)

2. *domain-adaptation* Using 179,000 doctor's letters from the Cardiology department at the University Hospital (cf. Section 5.4.2). This data set contains a broad range of texts from clinical routine in cardiovascular domain. With 1.3*GB* (gigabytes) it is significantly larger than the task-adaptation data set. (PLMs appended with suffix `-domain`)

3. *combination of both approaches* Further-pretrain a domain-adapted PLM on our task specific data (PLMs appended with suffix `-comb`)

We performed all pre-training experiments using a masked language modeling objective, a standard pretraining task for transformer-based architectures like BERT.[2] For hyperparameters and further training details see Section B.3.

5.3.3 Shapley values

We follow the interpretability setup introduced in Section 2.5 and use Shapley value attributions (SHAP) to interpret our section classification models (RQ 4). We estimate token-level contributions per input sample to the predicted section probability and visualize them as custom heatmaps. Token negatively contributing to a section class are shown in red shades, token positively contributing to a section class are shown in blue shades. We did not conduct any further faithfulness tests.

5.4 Data

5.4.1 Annotated corpus

For our experiments, we used CARDIO:DE v. 1.0, which includes 500 doctor's letters from the Cardiology Department at the Heidelberg University Hospital (further details, cf. Section

[2] cf. https://huggingface.co/docs/transformers/main/tasks/masked_language_modeling, accessed 11.09.2025.

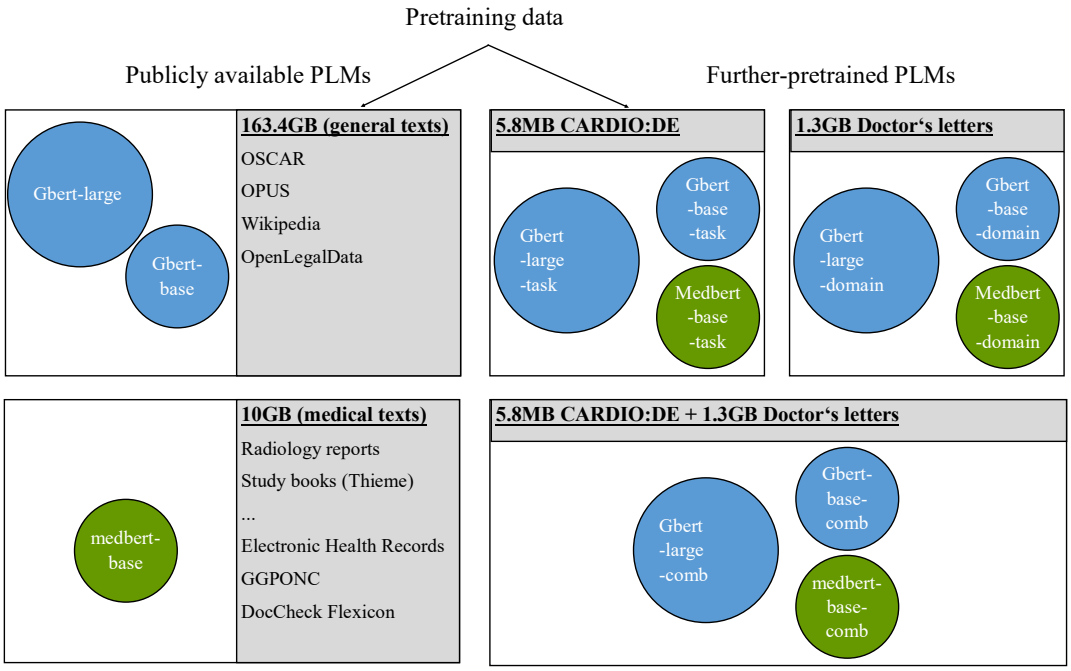


Fig. 5.1 **Pretrained language models:** We used two publicly available PLMs: `gbert` and `medbertde`. We evaluated base and large `gbert` models. Four pretraining methods were used: (1) publicly available, (2) task-adapted, (3) domain-adapted and (4) task- and domain-adapted combined.

4 and (Richter-Pechanski et al. 2023)). We used CARDIO:DE400 to sample training data and CARDIO:DE100 for evaluation. To increase readability and semantic consistency and to decrease the chance for re-identification, all PHI placeholders were replaced with semantic surrogates, as proposed in (Lohr et al. 2021).

We split the corpus by newline characters, which are part of the MS DOC source documents. Using publicly available sentence splitting methods or pattern heuristics to split the corpus produced unsatisfactory results. Furthermore, sequence length of newline split paragraphs rarely exceed 512 token (min: 3, max: 599, mean: 30.9, median: 16, 99th percentile: 205), thus, complying with most PLM sequence length restrictions. If a paragraph exceeds the maximum sequence length of the PLM we trim the sample accordingly.

We reduced the section classes in CARDIO:DE to the most significant sections. We removed the LABOR section, as it contains flattened tables that result in a large amount of relatively well-structured and short numeric samples. Internal experiments showed that they can be sufficiently identified using regular expressions and patterns. Furthermore, we merged seven semantically similar classes in CARDIO:DE annotations to three meta classes: (1) DIAGNOSEN: (*AktuellDiagnosen* + *Diagnosen*), (2) MEDIKATION: (*AufnahmeMedikation* + *EntlassMedikation*) and (3) BEFUNDE: (*KUBefunde* + *EchoBefunde* + *Befunde*). The restricted input length of our PLMs, even when combined with our context-enriched samples (cf. Table 5.2) was insufficient to reliably distinguish between *AufnahmeMedikation* and *EntlassMedikation* (for implications for downstream tasks, cf. Chapter 7.4). Our final dataset contains 49,258 paragraphs annotated with 9 section classes (Tab. 5.1).

During annotation, human annotators of CARDIO:DE were presented the whole document (for further annotation details, see Subsection 4.4.4 and (Richter-Pechanski et al. 2023)). To mimic this setup for our automatic section classifiers in this study, we introduced basic information about document structure to the model without introducing additional pre-processing steps or external knowledge. In addition to our training data containing single paragraph samples we assessed two types of context-enriched datasets for our experiments (examples, cf. Table 5.2):

- no-context (a single paragraph to be classified)
- context (previous paragraph + main paragraph + subsequent paragraph)
- prevcontext (previous paragraph + main paragraph)

The context-enriched samples still mostly comply with sequence length restrictions of PLMs (minimum 7, maximum 967, mean length 90.2, median length 63 and 99th percentile 371 sub tokens). If the sequence length of the context enriched sample is exceeded, we trim the sequence of the context to fit the maximum sequence length of the PLM.

Context type	Example
nocontext	Cvrf: Hypertonie, Nikotinkonsum, Hypercholesterinämie *Cardiovascular risk factors: high blood pressure, smoker, high cholesterol*
context	- OP am 02.01.2011 `[SEP]` Cvrf: Hypertonie, Nikotinkonsum, Hypercholesterinämie `[SEP]` Anamnese: *- Surgery on January 2, 2011* `[SEP]` *Cardiovascular risk factors: high blood pressure, smoker, high cholesterol* `[SEP]` *Patient medical history:*
prevcontext	- OP am 02.01.2011 `[SEP]` Cvrf: Hypertonie, Nikotinkonsum, Hypercholesterinämie *- Surgery on January 2, 2011* `[SEP]` *Cardiovascular risk factors: high blood pressure, smoker, high cholesterol*

Table 5.2 **Contextualized paragraphs:** A sample annotated as ALLERGIESINTOLERANCES-RISKS with three different context types, each separated by the `[SEP]` token. English translation in italics.

5.4.2 Pretraining data

For all pretraining experiments we used an internal clinical routine corpus containing approximately 179,000 German doctor's letters in a binary MS DOC format covering the time period 2004 to 2020. We collected letters from the Cardiology Department of the University Hospital Heidelberg. The pretraining corpus is disjoint from the annotated corpus. We conducted the following pre-processing steps: each letter was converted into a `UTF-8` encoded raw text file using the LIBREOFFICE command line tool `soffice` (version 6.2.8). We chose LIBREOFFICE, as it best preserved the structure of newlines and blanklines. We automatically de-identified all letters using our deep learning method presented in Section 2.6.1 (Richter-Pechanski et al. 2019). Similarly to the annotated corpus, we replaced PHI tokens with semantic surrogates (Lohr et al. 2021). All doctor's letters were concatenated into a single raw text file. We separated each new letter by the sequence `###BEGINN`. All empty lines and all tables containing laboratory values were removed. The corpus is sentence-split using NLTK's (version 3.7) `PunktSentenceTokenizer`.

The doctor's letters were further supplemented by the GGPONC corpus, which contains German oncology guidelines, with a total of 2 million token (Borchert et al. 2022). The final corpus covers 1.3 GB of raw text, approximately 218,084,190 token and 667,903 unique token.

5.5 Experimental setup

5.5.1 Metrics

We measure section classification performance with accuracy for per-model results. In a multi-class text classification task, the accuracy is defined as the ratio of text documents correctly classified to their respective classes over the total number of text documents:

$$\text{Accuracy} = \frac{\sum_{i=1}^{n} \text{TP}_i}{\text{Total Number of Texts}} \tag{5.1}$$

where TP_i represents the true predictions for each class i and n is the total number of classes.

To measure section classification performance per-section class, we use the F_1-score. It is defined as the harmonic mean of precision and recall given by

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5.2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5.3}$$

Hence, the F_1-score is defined by:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.4}$$

where TP, FP, and FN represent true positives, false positives, and false negatives, respectively.

We used approximate randomization tests (Koehn 2004; Padó 2006) to measure statistical significance for accuracy and F_1-score results. Results are considered significant if $p < 0.05$.[3]

5.5.2 Creating Few-Shot Data

To conduct PET experiments we created six few-shot datasets. Each dataset contains N paragraphs per section class with size $N = 10, 20, 50, 100, 200$ and 400 randomly selected from the CARDIO:DE400 data (random seed 42). Each paragraph included the previous and subsequent context paragraph. All other context types (NOCONTEXT, PREVCONTEXT) are derived from this dataset. Each few-shot set includes three labeled training files and three unlabeled files with the remaining samples from the CARDIO:DE400 dataset (for an example folder sturcture, cf. Figure B.3). All experiments were evaluated on the complete CARDIO:DE100 held-out dataset.

[3] cf. https://github.com/smartschat/art, accessed 19.09.2025.

5.5.3 Core Experiments

We conducted *core experiments* to assess the performance of different section classification models along three dimensions to compare: (1) sequence classifier (SC) variants fine-tuned under the pretrain-then-finetune paradigm to pretrain-then-prompt models using PET (Figure 2.1), (2) four different pre-training methods for clinical adaptation, and (3) six different few-shot sizes: $10 - 400$.

The SC models are trained using BERT-architecture with an additional output layer for a sequence-classification task as described in (Devlin et al. 2019). We use the SC implementation of the PET framework, defined by the parameter `--method sequence_classifier`.

For all *core experiments* we used base-sized BERT models (`gbert-base-*` and `medbertde-base-*`) using all five templates combined and nocontext samples (cf. Table B.2). To measure the standard deviation in all experiments we used three disjoint training sets including their unlabeled sets for each few-shot set. Furthermore, we conducted all experiments with two random initial seeds (123 and 234).

5.5.4 Additional experiments

In *additional experiments* we investigate the effectiveness of additional parameters, using the model that performed best in *core experiments*, with reduced few-shot sets: $20, 50, 100$ and 400. We investigate the impact of (1) *model size* comparing BERT-large and BERT-base models, (2) *null prompt patterns*, and (3) *contextualization*. In *core* and *additional experiments* we further perform *class-based evaluations* on two primary classes, which were selected with clinical experts: (1) ANAMNESE (mostly unstructured) and (2) MEDIKATION (semi-structured).

Model size : We evaluated the impact of adding model parameters, by comparing `gbert-base` (110 million) vs `gbert-large` (340 million) PLMs. We limited this setup to `gbert` PLMs, since a large `medbertde` was not published.

Null prompts : (Logan et al. 2022) discovered that the usage of *null prompts*, prompts without manually crafted templates achieve competitive accuracy to manually tuned prompt templates on a wide range of tasks. This is of particular interest in the clinical domain, to further reduce costly engineering efforts.

Adding context : To provide more information on the structure of the document, we added context paragraphs to each input sample in order to evaluate their effect. We evaluated three types of context (Tab. 5.2).

5.6 Results

5.6.1 Baselines

We defined two baselines to assess model performance in our *core* and *additional experiments*: as **lower bound** we use a *zero-shot prompting* approach; as **upper bound** we use a *fine-tuned sequence classifier* trained on the *full* size of the training corpus. Figure 5.2 shows the accuracy results for both baselines. The upper bound results exceed 96% accuracy for both models. The further-pretrained `gbert` models yield a minimal (statistically significant) advance of 0.4-0.6 accuracy points above the original `gbert-base`. For `medbertde` no such difference is observed.

The *zero-shot results* are all below 16% accuracy, except for the public `medbert-base` that with 28.3% achieves a great advance over `gbert-base` with 7.2% accuracy. However, the `gbert` models further-pretrained on both task- and domain-specific data more than double the performance of the original model to 15% accuracy, beyond `gbert` pre-trained on domain-specific data only (`*-domain`). All performance differences for `gbert` are statistically significant, except `gbert-base` and `gbert-domain`.

5.6.2 Core experiments

Figure 5.3 presents our *core experiment* results compared to the baselines introduced above.

Evaluation

PET vs. SC The PET model variants significantly outperform SC models at shot sizes ≤ 100 in 31 out of 32 setups when comparing the same pretraining methods. Only SC `medbertde-base-comb` outperforms all PET models with shot size 100.

Few-shot size Both PET and SC models benefit from an increase in few-shot size. We observed statistical significance at shot sizes $\leq$200. The smaller the shot size, the greater the relative performance gain of PET over SC models.

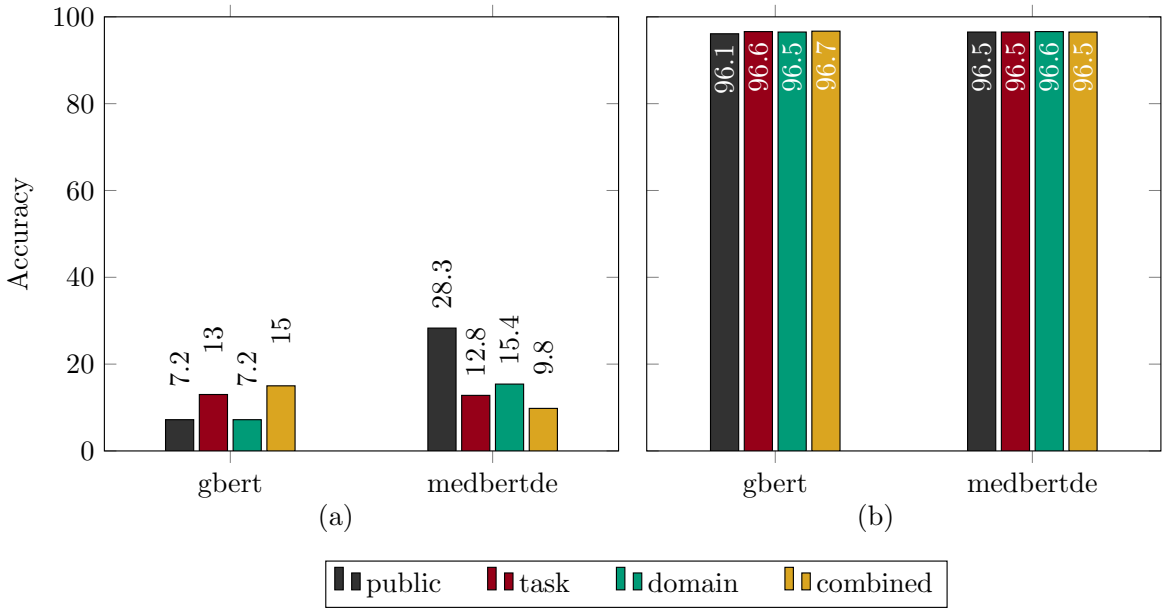


Fig. 5.2 **Section classification baseline results (lower/upper bound):** We show accuracy scores in percentage per pretraining method (public, task-adapted, domain-adapted and combination of both) per model: `gbert-base` and `medbertde-base`. (a) Lower-bound: used in zero-shot prompting (b) Upper bound: *full* training set.

Further-pretraining We observe notably different results for further-pretrained `gbert` and `medbertde` PLMs.

Gbert PET models benefit significantly from further-pretraining with ≤ 100 shots. Accuracy gradually increases with task-specific, domain-specific and combined pretraining, in that order. `Gbert` SC models also benefit significantly from domain-adapted models over all shot sizes (except 10 and 400 shots), but not from task-adaptation or their combination. Overall, we observe a more consistent effect of further-pretraining for PET models compared to SC models.

Medbertde Further-pretraining shows no consistent performance improvement for `medbertde` model variants. In particular, with 20 shots, the `medbertde-base` PET model outperforms the further-pretrained models, achieving a statistically significant 79.1% accuracy. For few-shot sizes 10 and 50-400, the best performing model alternates between the `medbertde-domain` and `medbertde-comb` PET models. Similar to `gbert` models, the relative gain of pretraining decreases with increasing shot sizes. It appears that our pretraining method using cardiovascular doctor's letters has no impact or may even impair the `medbertde` model. A possible reason could be that the public `medbertde` model was only pretrained on 10GB of clinical and medical texts, primarily from the oncology domain.

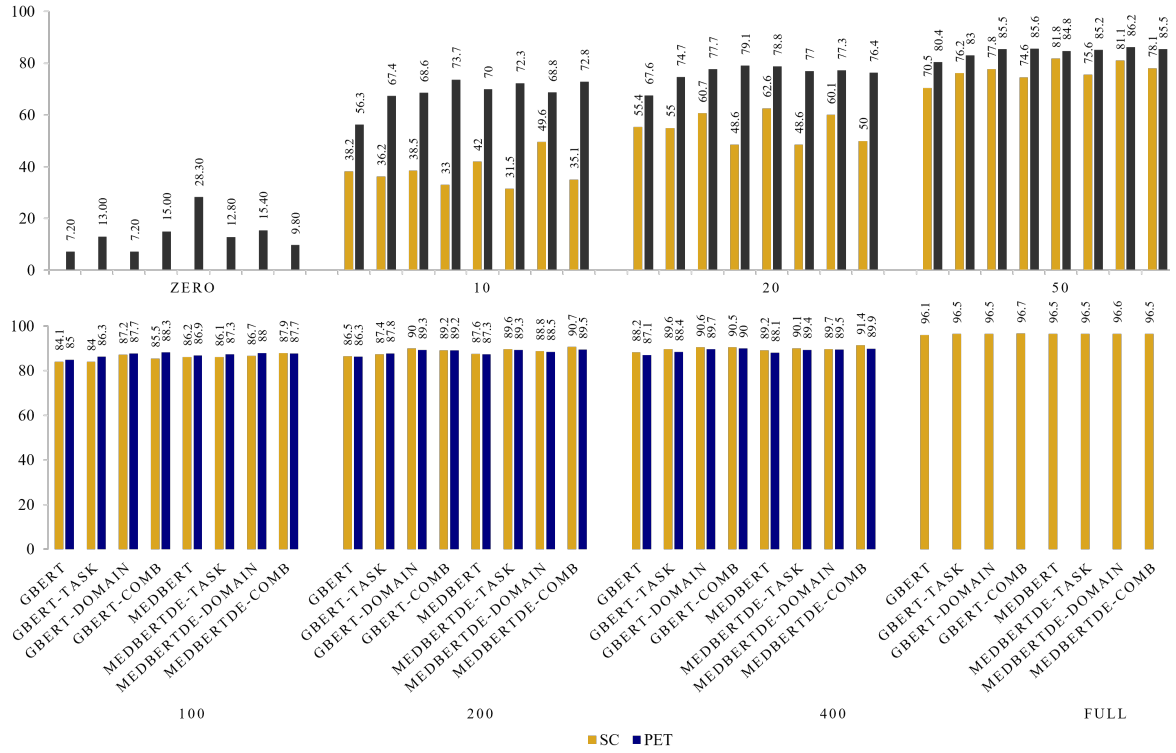


Fig. 5.3 **Accuracy scores in percentage for *core experiments* and lower/upper bound:** Comparing prompting using PET vs. SC, few-shot sizes $10-400$ and pretraining methods using base BERT models. For reference, lower-bound PET baselines trained with zero-shots (ZERO) and upper-bound SC models trained on complete training set (FULL).

However, future research is needed for further investigation (pretraining data information cf. Fig. 5.1).

Best-performing model variant According to our core experiments, the overall best-performing model is the `gbert` domain- and task-adapted model (`gbert-base-comb`). This model achieved best accuracy scores with shot sizes ≤ 100 compared to other pretraining methods and to fine-tuned SC models with shot sizes ≤ 400. When using only 20 shots, this model outperforms the SC model by 30.5 pp. and the public `gbert-base` PET model by 11.5 pp. Hence, we select this model for all *additional experiments*. If not further pretrained `medbertde-base` outperforms public `gbert-base`: this is similar to our baseline experiments. However, further-pretraining does not improve the performance of `medbertde-base`, possibly due to the relatively small pretraining data size of `medbertde-base` (10GB).

Robustness Experiments were performed using three training sets and two initial random seeds. For smaller shot sizes (≤ 50 shots) standard deviation was low ($\sim 2.5\%$) decreasing to less than 1% for larger sizes. We observed this for `gbert` and `medbertde` with no impact of different pretraining methods.

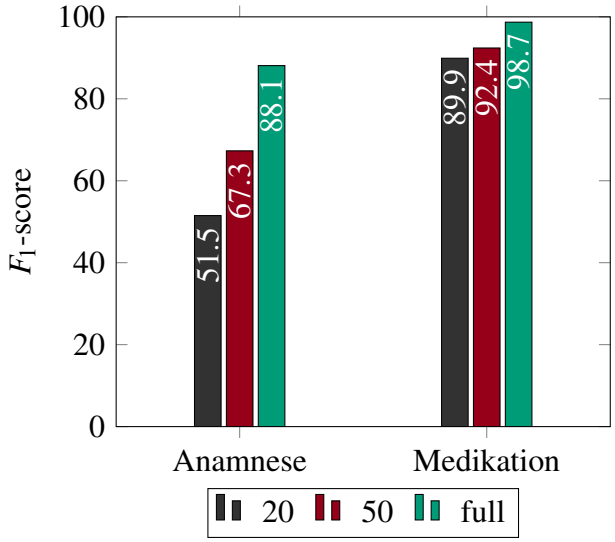
Inspecting primary classes

We investigated the impact of shot size on the accuracy of predicting the selected primary section classes (Figure 5.4a). Across shot sizes $20-50$, the F_1-scores of both classes increase in average by 9.2% pp. ANAMNESE, with a lower F_1-score, benefits more from larger few-shot sizes. However, the SC model trained on the full training set significantly outperforms the 50-shot models. This is especially true for the ANAMNESE class. Even if shot size is increased to our maximum of 400 shots, the results still differ significantly: (ANAMNESE: 82.4%, MEDIKATION: 97.5%). Results for more semi-structured classes like MEDIKATION are closest to the performance of the full model. For results of all shot sizes cf. Figure B.4.

While our primary classes benefit from further-pretraining, F_1-score of ANREDE slightly decreased. A possible explanation could be that ANREDE often contains non-clinical terminology that describes a patient's place of residence, date of birth and name (cf. Figure B.5).

Inspecting Shapley values

To better understand model predictions in a few-shot setting, we further analyzed Shalpey values of the 20-shot model for the lower-performing class ANAMNESE. We chose a false



(a)

True label	Prediction (probability)	Shapley values
Zusammenfassung	Anamnese (0.77)	Die Aufnahme der Patientin erfolgte bei akutem Myokardinfarkt -LRB- STEMI -RRB-
Zusammenfassung	Zusammenfassung (0.18)	Die Aufnahme der Patientin erfolgte bei akutem Myokardinfarkt -LRB- STEMI -RRB-

(b)

Fig. 5.4 **Core experiments:** primary class F_1-score in percentage and selected Shapley values: (a) F_1-score scores per few-shot sizes for primary classes with using `gbert-base-comb nocontext`. (b) Shapley value analysis for `gbert-base-comb nocontext` with respect to ANAMNESE and ZUSAMMENFASSUNG prediction. First column: true label of the sample, second column: predicted label including label probability, third column: selected Shapley values. We used 20 training shots. For readability reasons, we grouped some token sequences. Further details, see Figure B.7. Legend: **Blue: positive contribution**, **Red: negative contribution**.

positive sample as the running example for the remainder of this study, because ANAMNESE belongs to our primary classes and often suffers from a low precision rate (for 20-shots, `gbert-base-comb` achieves 44.6% precision and 62.2% recall (cf. confusion matrix Figure B.6). Table 5.4b illustrates selected Shapley values per token for the sample: *Die Aufnahme der Patientin erfolgte bei akutem Myokardinfarkt -LRB- STEMI -RRB- . (English: The patient was admitted due to an acute myocardial infarction -LRB- STEMI -RRB-.)* towards the classes ANAMNESE and ZUSAMMENFASSUNG, respectively.

The model incorrectly classified this sample as ANAMNESE, with 76.8% probability, while the correct class is predicted with 18.2% probability score. Tokens such as *Die (the), Aufnahme (admission), Patient (patient), erfolgte (took place)* positively contributed to the ANAMNESE class, while the tokens *Aufnahme* and *Patient* negatively contributed to the correct ZUSAMMENFASSUNG class. Analyzing the 20-shot training dataset, we observe that these keywords occur more frequently in samples for ANAMNESE (*Die (13x), Aufnahme (6x), Patient (7x), erfolgte (8x)*) than in samples from ZUSAMMENFASSUNG (*Die (5x), Aufnahme (2x), Patient (5x), erfolgte (6x)*). The token *Myokardinfarkt (acute myocardia)* positively contributes to both section classes, and to a higher extent to ANAMNESE, even though we only observe this token in instances from ZUSAMMENFASSUNG. The token sequences representing brackets *-LRB-* and *-RRB-* contribute strongly positively to ANAMNESE. Analyzing the training data showed a higher frequency of these tokens in ANAMNESE samples (11x) compared to ZUSAMMENFASSUNG (5x).

Note on interpreting Shapley values Shapley values are additive: they sum up all token contributions along with the base value to yield the prediction probability. Shapley values towards different classes and of different models cannot be compared by absolute value, but only relative to other tokens for the same prediction and the same model.

5.6.3 Additional experiments

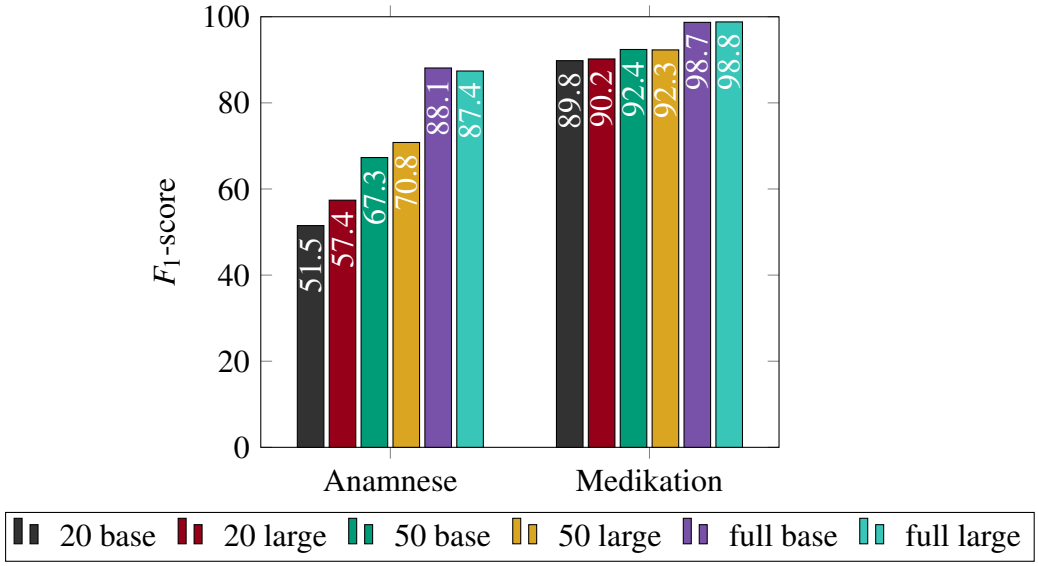
Model size

Given the limited computational resources in clinical infrastructures, we investigated how model size affects performance and investigate its impact with finer-grained analyzes. Since there is no `medbertde`-large model available, we compared `gbert-large` and `gbert-base` models.

Larger model size increases accuracy significantly, by an average of 7.2 pp. for SC models ≤ 100. PET models, by contrast, benefit less from larger model size than SC models. We

Shot size	SC (base)	SC (large)	PET (base)	PET (large)
20	48.6	61.7	**79.1**	78.2
50	74.6	81.2	85.6	**86.7**
100	85.5	87.4	88.3	**88.6**
400	90.5	90.7	89.7	**90.4**
full	96.7	96.6	-	-

(a)



(b)

Fig. 5.5 **Model size:** (a) Accuracy scores in percentage for `gbert-comb nocontext` PLMs using all templates on four few-shot sizes. (b) F_1-scores in percentage for primary classes for `gbert-comb no context` PLMs using all templates on various few-shot sizes.

even observe a slight performance decrease for shot size 20 (Table 5.5a). The only significant increase, of 1.1 points accuracy, we observed for shot size 50.

Primary classes `Gbert-large` yields an increased F_1-score for ANAMENSE with both shot sizes $(20, 50)$, by an average of $+4.7$ pp., but this is only significant for shot size 50. By contrast, the difference in F_1-score $(0.1\% - 0.4\%)$ for MEDIKATION is not statistically significant (Figure 5.5b).

Shapley values Both models, `gbert-base-comb` and `gbert-large-comb` incorrectly classify our running example belonging to ZUSAMMENFASSUNG as ANAMNESE. We do not observe significant differences in the respective token contributions (cf. Fig. B.8).

Null prompts

Inspired by insights of (Logan et al. 2022), who removed all tokens from prompt templates, using null prompts instead, with comparable classification results, we evaluated the `gbert-base-comb` model using only three null prompt templates (cf. Section 5.3.1).

Null prompts slightly decrease accuracy scores for shot sizes ≤ 50 by approximately one percentage point. For shot sizes 100 and 400 we note a slight accuracy increase. We only observed statistically significant differences in accuracy for shot-size 50 (template-based model: 85.6%, null-prompt model: 84.6%) (cf. Tab. B.3).

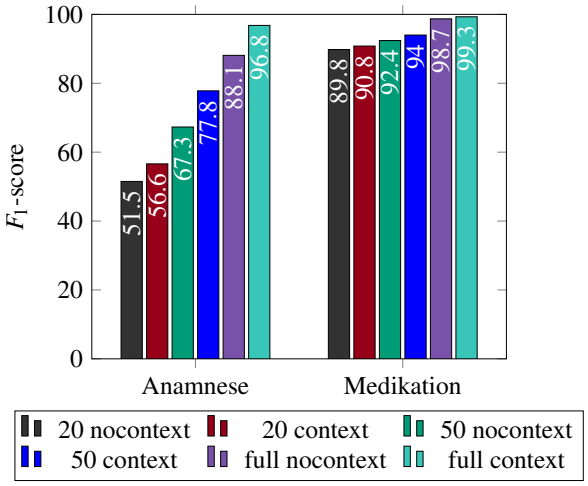
Primary classes For our primary classes we did not observe a consistent pattern. Null prompts have a slightly negative impact on F_1-scores for ANAMNESE and MEDIKATION with 20 shots. By contrast, with 50 shots, accuracy significantly decreased for ANAMNESE, but slightly increases for MEDIKATION (92.4% vs. 95.9%).

Adding context

Predicting section classes is difficult for tokens that frequently occur in different classes, as discussed for the example in Figure 5.4. To reduce the degree of ambiguity of individual tokens, we experimented with two types of *contextualization* of classification instances: Adding (1) the previous and subsequent paragraph (*context*) and (2) only the previous paragraph (*prevcontext*). Figure B.9 shows that across all few-shot sizes, (1) *context* (with mean $+2.4$ accuracy pp.) and (2) *prevcontext* (with mean $+1.6$ accuracy pp.) both achieved significantly higher accuracy than *nocontext* models (cf. Section 5.5.4):

Primary classes *Context* models improve the F_1-scores for both primary classes (by mean $+7.8$ points for ANAMNESE and $+1.3$ for MEDIKATION) (Figure 5.6a). For ANAMNESE, statistically significant improvement is only reached using 50 shots.

Shapley values `gbert-base-comb context` correctly classified our running example with 86.6% probability (Table 5.6b). Most highly contributing tokens belong to the context (previous or following, with Shapley values: $0.057 + 0.596$), while the main paragraph has an accumulated Shapley value of 0.106. The previous context contains the sequence: *Zusammenfassende Beurteilung*, a frequent section-specific title. The subsequent paragraph is the longest paragraph (37 tokens). Previously negatively contributing tokens (*Aufnahme* and *Patient*) are now positively contributing to the correct class: *Zusammenfassung*.



(a)

True label	Prediction (probability)	Shapley values
Zusammenfassung	Zusammenfassung (0.18)	Die Aufnahme der Patientin erfolgte bei akutem Myokardinfarkt -LRB- STEMI -RRB-
Zusammenfassung	Zusammenfassung (0.87)	PREVIOUS CONTEXT TOKENS [SEP] Die Aufnahme der Patientin erfolgte bei akutem Myokardinfarkt -LRB- STEMI -RRB- [SEP] SUBSEQUENT CONTEXT TOKENS

(b)

Fig. 5.6 **Additional experiments:** (context) - primary classes F_1-scores and selected Shapley values: (a) F_1-scores in percentage per few-shot sizes for primary classes with *nocontext* and *context* using *gbert-base-comb*. Comparing to `gbert-base-comb` trained on full training data with *nocontext* and *context*. (b) Shapley value analysis for `gbert-base-comb nocontext` and `gbert-base-comb context`. First column: true label of the sample, second column: predicted label including label probability, third column: selected Shapley values. We used 20 training shots. For readability reasons, we grouped some token sequences. Further details, see Figure B.10.
Legend: **Blue: positive contribution**, **Red: negative contribution**.

Shot size	Base *nocontext*	Large *nocontext*	Base *context*	**Large *context***
20	79.1	78.2	80.5	**84.3**
50	85.6	86.7	89.2	**89.4**
100	88.3	88.6	90.9	**91.3**
400	90	90.4	92.8	**93.4**
full (SC)	96.7	96.6	98.6	98.6

Table 5.3 **Combining and evaluating best performing methods:** Accuracy scores in percentage for `gbert-large-comb context` evaluated on few-shot sizes [20, 50, 100, 400] with `base vs. large` model sizes in *context vs. nocontext* settings using PET. Comparison to corresponding SC model fine-tuned on full training set.

Combining best-performing methods

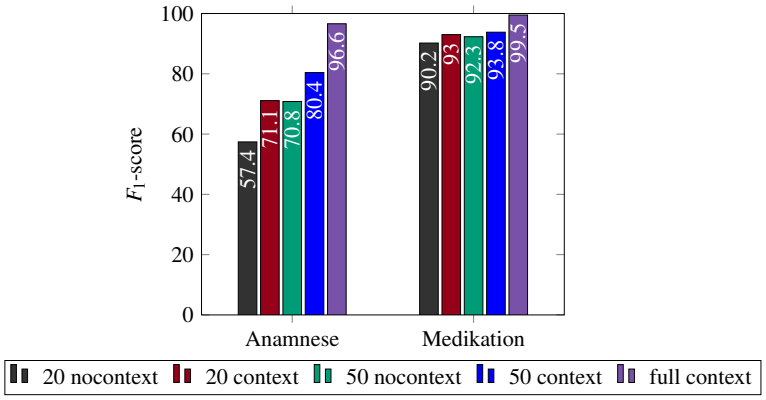
Our *core experiments* indicated that the `gbert-base-comb` model performed best of all tested models. The *additional experiments* showed that models using all five templates (cf. Section 5.3.1), a BERT-large architecture and contextualization often achieved the best performance. Hence, we investigated whether this combination (`gbert-large-comb context` trained with all templates) could further close the performance gap to a model trained on full training set.

Table 5.3 shows that `gbert-large-comb context` significantly outperforms both `gbert-base-comb` and `gbert-large-comb nocontext`. Moreover, `gbert-large-comb context` statistically significantly outperforms `gbert-base-comb context` for 20, 100 and 400 shots. Overall, the `gbert-large-comb context` outperforms *nocontext* and *base* models over all shot-sizes, yielding best results with 400 shots. Yet, PET still lags behind the *full* SC setting, with a minimal gap of -5.2 points accuracy.

Primary classes For our primary classes, `gbert-large-comb context` now outperforms `gbert-large-comb nocontext` by large margin (Figure 5.7a). Only the 50-shot results for ANAMNESE are not statistically significant (F_1-score of all shot-sizes cf. Figure B.11).

We also compared the `large` and `base` versions of `gbert-*-comb context`. The F_1-score for ANAMNESE is significantly increased by $+14.6$ points with 20 shots, and by $+2.6$ points with 50 shots. Performance for MEDIKATION is significantly increased by $+2.2$ points with 20 shots, but insignificantly decreased with 50 shots. (cf. Figure B.12)

Shapley values : We tested whether the token contributions differ between the `large` and `base` `gbert-*-comb context` models (Table 5.7b). The large model predicts the true class ZUSAMMENFASSUNG with 99.2% probability, $+12.7$ points above the `base` model. The



(a)

True label	Prediction (probability)	Shapley values
Zusammenfassung	Zusammenfassung (0.87)	PREVIOUS CONTEXT TOKENS [SEP] Die Aufnahme der Patientin erfolgte bei akutem Myokardinfarkt -LRB- STEMI -RRB- [SEP] SUBSEQUENT CONTEXT TOKENS
Zusammenfassung	Zusammenfassung (0.99)	PREVIOUS CONTEXT TOKENS [SEP] Die Aufnahme der Patientin erfolgte bei akutem Myokardinfarkt -LRB- STEMI -RRB- [SEP] SUBSEQUENT CONTEXT TOKENS

(b)

Fig. 5.7 **Additional experiments** (combined methods) - primary classes F_1-scores and selected Shapley values: (a) F_1-scores in percentage per few-shot sizes for primary classes with *nocontext* and *context* using `gbert-large-comb`. Comparing to `gbert-large-comb` trained on full training data with *context*. (b) Shapley value analysis for `gbert-base-comb context` and `gbert-large-comb context`. First column: true label of the sample, second column: predicted label including label probability, third column: selected Shapley values. We used 20 training shots. For readability reasons, we grouped some token sequences. More detailed results, see Figure B.13.
Legend: **Blue: positive contribution**, **Red: negative contribution**.

`large context` model now also places greater emphasis on the main paragraph, as opposed to the context. The ratio of the accumulated Shapley values ($\frac{classified\ instance}{context\ paragraphs}$, higher is better) is 0.36 for `gbert-large-comb context` and 0.16 for `gbert-base-comb context`.

5.7 Discussion

In this section, we discuss our empirical findings in light of the clinical constraints and proposed solutions outlined in Section 5.2 and our research questions presented in Section 1.2.

1. **On-premise resource constraints**

    1.a. **Domain expertise** (relevant RQs: RQ 1 and RQ 3)
    In in-depth evaluations we compared four pretraining approaches using PET and SC for two public German-language models (Gururangan et al. 2020): (1) *initial pre-training* using general German texts with *gbert* vs. exclusively medical and clinical data with *medbertde* (Fig. 5.1); and *further-pretraining* of these PLMs for (2) *task-adaption*, (3) *domain-adaptation* and (4) combined *task and domain-adaptation*.

    - **Finding** `Gbert` overall accuracy gradually improved with further-pretraining. The task- and domain-adapted `gbert-base-comb` performs best compared to all models, and with only *20 shots* outperforms `gbert-base` by $+11.5$ accuracy points. Also, the positive effect of further-pretraining was more consistent for PET compared to SC models. By contrast, further-pretrained `medbertde`-based SC and PET models did not achieve consistent performance improvements.
    - **Finding** Pre-training from scratch with sufficient clinical and medical data can benefit various MIE tasks. However, when pretraining data is limited and/or concentrated on a narrow domain, e.g., oncology, as in the case of `medbertde`, further-pretraining was found not to enhance performance.
    - **Finding** While `medbertde-base` without further-pretraining outperformed `gbert-base` in all shot sizes, and similarly when trained on the *full* dataset (Figure 5.3), it did not improve performance if further pre-trained and was outperformed by further-pretrained `gbert-base`.

    We could further-pretrain and finetune all models fully on-premise. By using in-house native-language, unlabeled clinical texts for further-pretraining in

combination with PET we could reduce need for clinical expertise for feature engineering, extensive manual annotation, while achieving robust performance in few-show scenarios on German clinical data.

1.b. **Staff time** (relevant RQs: RQ 1, RQ 3)
We evaluated prompt-based fine-tuning with PET (including PETAL, null prompts and optional contextualization) versus SC in few-shot learning scenarios.

- **Finding** We observed a steady increase in the performance of PET compared to SC models with decreasing few-shot training set sizes ($400 - 10$ shots). Using 20 shots, the PET `gbert-base-comb nocontext` model outperforms the corresponding SC model by $+30.5$ pp. The same `gbert-base-comb nocontext` PET model with 50 shots even rivals the SC model trained on *full* data, leaving a gap of -11.1 pp. Especially semi-structured section classes, such as MEDIKATION, perform close to the full model by -6.3 pp. (Figure 5.4a). Our few-shot models are also *robust* as measured by standard deviation.
- **Finding** *Null prompts* exhibit comparable results with no significant difference in performance, especially with few-shot sizes exceeding 100 reducing manual prompt-engineering.
- **Finding** Contextualized data with surrounding *context* paragraphs improved classification results for most section classes, especially primary classes. It allowed our base models to correctly predict our running false-positive sample as ZUSAMMENFASSUNG. However, compared to the base models interpretability analysis using SHAP revealed that the large model places greater emphasis on main paragraph tokens rather than on context paragraphs. Contextualization further reduced the accuracy gap between `gbert-*-comb context`-based PET models trained on 50 shots to the *full* SC model to -9 to -9.5pp; for classes such as MEDIKATION even to -5 to -6 pp. Contextualization does not require complex pre-processing or manual annotation.

- **Finding** While our study focused on few-shot sizes up to 400 instances, further increasing shot sizes to the full training set revealed important insights and highlighted the need for future research. For our primary classes the `gbert-large-comb context` model demonstrated little variance in performance between 50 and 400 shots. However, a substantial performance gain was observed when training on the full training set, most notably for the

ANAMNESE class (cf. Fig. B.11). This suggests that more complex classes with higher semantic complexity benefit from larger few-shot training sets to achieve accuracy comparable to the full training set.
Class-wise analysis for selected setups (cf. B.6 and B.16) indicated that more regular and semi-structured classes appear to approach stable performance with comparatively few training samples, whereas semantically more complex free-text classes continue to benefit from additional supervision. From a practical perspective, these findings suggest that annotation effort under clinical constraints may be concentrated more efficiently by prioritizing more complex section classes, while more regular classes may require less intensive annotation to achieve robust performance.

Prompt-based fine-tuning of compact encoders with PET (combined with PETAL and null prompts) and lightweight contextualization supports reducing manual annotation and engineering efforts, while achieving strong performance in a paragraph-level section classification task.

1.c. **Local compute resources** (relevant RQs: RQ 1, RQ 2)
Using smaller models saves computational resources. We therefore compared classification performances of `base` and `large` BERT PLMs for PET and SC.

- **Finding** Large PLMs achieve better classification results. However, model size has a lower impact on the performance of PET compared to SC models (Fig. 5.5a). For classes such as MEDIKATION the further-pretrained `gbert-base-comb` PLM performs almost on par with `gbert-large-comb` (Figure 5.5b).
- **Finding** For complex sections with free text such as ANAMNESE, `gbert-large` PLMs achieved better performance. They also better recognize contextualized instances (Table 5.7b and Section B.1.2).

In compute-restricted clinical environments `base` PLMs offer an optimal balance between accuracy and compute-resource demands. However, `large` PLMs are beneficial for complex section classes.

1.d. **Native-language barrier** (relevant RQs: RQ 1 and RQ 2)
We evaluated further-pretraining and fine-tuning of general-domain German PLMs and medical PLMs on German clinical texts.

- **Finding** Combining domain- and task-adaptation of general-domain German PLMs on German clinical corpora achieved strongest results using PET in a few-short learning scenario.

- **Finding** A German clinical PLM pretrained from scratch (`medbert.de`) outperformed a general-domain `gbert-base` PLM using similar few-shot training samples emphasizing the relevance of native-language clinical pretraining. However, further-pretraining did not improve performance of `medbertde`, probably due to the relatively small pretraining data.
- **Finding** PET combined with null prompts and contextualization achieves robust performance in German, which supports deployment of NLP pipelines under clinical resource constraints.

To tackle the native-language constraint we achieved the best performance by further-pretraining and fine-tuning German PLMs on German clinical texts on-premise.

2. **Transparency requirements** (relevant RQs: RQ 4)
   Shapley values (Lundberg et al. 2017), an interpretability method based on saliency features, helped in error analysis by identifying problems in *training data quality* and *model decisions*. We identified tokens that frequently occur in false-positive classes by analyzing model predictions (Figure 5.6).

   - **Finding** The use of Shapley features is especially beneficial in few-shot scenarios, as it enables data engineers to select few-shot samples with high precision.
   - **Finding** Shapley values highlighted that with very small shot sizes, and for section classes with short spans, the model prioritized the context over the instance to be classified.
   - **Finding** Our analysis of Shapley values showed that `gbert-large-comb` makes more reliable predictions than `gbert-base-comb`, by prioritizing features of instances to be classified over context (Table 5.7b).

   Shapley values are beneficial in increasing trust by supporting in-depth error analysis and data selection for encoder-based models.

5.8 Conclusion

In this study we have presented best-practice strategies to identify an ideal setup to address the multi-faceted challenges of conducting a MIE task, such as paragraph-level clinical section classification, under on-premise and transparency constraints in a lower-resource domain and language. In summary, under clinical constraints, our best performing setup used a task-

and domain-adapted BERT-large architecture trained with PET on contextualized samples using all five template types. However, while BERT-large achieved best performance for complex free-text sections, BERT-base encoders offered the overall best accuracy-efficiency trade-off.

Domain expertise To reduce the demand for clinical knowledge in MIE we showed that few-shot prompting performed particularly well with further-pretrained general-domain PLMs, and helped to reduce the demand of clinical expert knowledge for manual data annotation. `Gbert-base-comb` showed the highest performance gains, with only 20 shots, outperforming the non-adapted `gbert-base` by $+11.5$ pp. (cf. 5.1). This reduced the need for clinical expertise and large annotated datasets. In contrast, PLMs pre-trained on domain-specific data from scratch, such as *medbertde* may outperform *gbert* if not further pre-trained, but may not benefit from further pre-training.
To reduce the need for manual annotation efforts, we therefore recommend, to domain- and task adapt small-sized general-domain encoders on in-house native language clinical texts and finetune them using PET. Furthermore, if further-pretraining is not feasible due to IT or data limitations, we recommend choosing clinical PLMs like *medbertde* over non-adapted general PLMs.

Staff time Our study indicated that prompt-based learning methods improve classification results if annotated data are rare and effectively reduce time investment and costs of manual data annotation. With 20 shots, PET using `gbert-base-comb nocontext` outperformed the SC model by 30.5 pp. and with 50 shots almost competing the *full* SC baseline (-11.5 pp.). The larger the amount of annotated data, the higher the efficiency of *null prompts*, which further saves engineering time. Moreover, contextualizing classification instances improves performance, especially for the primary classes, and further closes the gap to *full* models. We therefore recommend to use prompt-based learning and apply lightweight contextualization to further save staff time and, if few-shot size≥ 100 are available, use null prompts to minimize prompt-engineering efforts.

Local compute resources We found that in the case of limited computing resources, prompting methods allow to employ smaller PLMs in a few-shot scenario, while achieving classification results comparable to larger models. Domain- and task-adapted `gbert-base-comb` performed close to `gbert-large-comb` for semi-structured section classes (e.g. MEDIKATION) offering the best accuracy-efficiency compromise in on-premise clinical environments. However, free-text sections, such as ANAMNESE may still benefit from

larger model architectures (Fig. 5.5b). We recommend to start classification experiments with further-pretrained base-sized encoder PLMs using PET to lower demand for compute demands and only apply `large` encoders to classify more complex free-text samples.

Native-language barrier Adapting PLMs to German clinical jargon texts was crucial. Domain- and task-adaptation of `gbert` models achieved the strongest PET results across shot-sizes. However, if data are scarce, `medbertde-base` outperformed `gbert-base` using similar few-shot training samples. Applying *null prompts* and *contextualization* further improved performance in a German clinical setting. If further-pretraining is not feasible, we therefore recommend to prefer native-language clinical PLMs, such as `medbertde` over general-domain PLMs.

Transparency requirements Finally, we addressed the need for transparent and trustworthy model predictions in lower-resource German clinical NLP, and possible use cases for *interpretability* methods. Our study demonstrates that the analysis of Shapley values can help improve training data quality, which is especially important with small shot sizes. Examining Shapley values, or similar interpretability methods, can also inform model selection, by revealing tokens that contribute to classification errors in specific model types. Finally, model interpretability is crucial in safety-critical domains such as clinical routine, to enhance the trustworthiness of model predictions. We recommend to leverage attribution methods such as Shapley values during model development, to support efficient error analysis, data sampling and improve model prediction transparency.

This study presents strategies and best-practice approaches for optimizing a paragraph-level section classification task in lower-resource clinical language settings. It highlights the benefits of few-shot prompting with further-pretrained PLMs as a measure to reduce the demand for manual annotation by clinicians. We further demonstrate that prompt-based learning and contextualization significantly enhance classification accuracy, especially in low-resource scenarios, while keeping demands on computing resources low.

Chapter 6

Medication Information Extraction using Local Large Language Models

6.1 Outline and contributions

This chapter is based on our published work presented in (Richter-Pechanski et al. 2025) and builds on our experiments using prompt-based learning for section classification in German doctor's letters (Richter-Pechanski et al. 2024) presented in Chapter 5 by extending it from a text classification task to a more complex token-level medication information extraction task combining NER and RE using generative LLMs.

Section 6.2 motivates this study by addressing the potential of SOTA generative LLMs under clinical constraints. Section 6.3 describes the datasets and preprocessing methods, along with fine-tuning approaches of local LLMs, evaluation and interpretability strategies. Section 5.6 presents all results by comparing the zero-shot and SOTA results with the performance of the fine-tuned LLMs. We also present novel evaluation strategies and use cases of Shapley values for generative LLMs. Section 6.5 discusses these results and findings in the context of research questions defined in Chapter 1 and clinical constraints defined in Chapter 2. Section 6.6 summarizes all findings and gives guidelines for MIE projects in clinical routine.

While prompt-based learning with encoders performs strongly on text classification tasks, performance on more complex tasks remain limited (Ma et al. 2022; Shen et al. 2023). Thus, after the *ChatGPT moment* in 2022, several studies showed promising results on token-level NER and RE tasks using generative LLMs (cf. Section 2.4.3). The release of various publicly available mid-sized LLMs (e.g. Llama and Mistral) led to updates of GPU resources across several clinical infrastructures in Germany. After our department acquired additional 4×

NVIDIA P40 GPUs, we were able to build on-premise MIE pipelines leveraging LLMs with billions of parameters.

This chapter contributes to RQ 1 by demonstrating that generative LLMs can be fine-tuned and deployed on-premise with average GPU resources for a native-language medication information extraction task (cf. challenges 1.a. *Domain expertise*, 1.b. *Staff time*, 1.c. *Compute restrictions* and 1.d. *Native language*). We address RQ 2 by demonstrating that compared to traditional optimization strategies, generative LLMs fine-tuned with PEFT methods show optimal balance between performance and resources for a complex MIE task in a clinical setup (cf. challenges 1.c. *Compute restrictions*, 1.d. *Native language*). We reduce manual efforts during model development by (i) leveraging clinical knowledge of LLMs and PEFT fine-tuning methods, (ii) applying format-constraint prompting strategies to support reliable structured outputs and (iii) integrating feedback LLMs to reduce evaluation efforts of MIE models, directly contributing to RQ 3 (cf. challenges 1.a. *Domain expertise* and 1.b. *Staff time*). Furthermore, to support transparency and clinical trust, we address RQ 4 by integrating format-restricting prompts and by leveraging well-established attribution methods based on Shapley values for generative LLMs (cf. challenge 2. *Transparency*). The models from these experiments are essential to address RQ 5, as they will be utilized to demonstrate their clinical applicability in Chapter 7. All relevant code of this project is published on GitHub.[1]

6.2 Introduction and background

Medication information is crucial in clinical routine, e.g., to ensure safe medication reconciliation during patient admission; to maintain medication continuity for care transitions or to guide life-saving treatment decisions in clinical emergencies. Importantly, medication information is a valuable data source for clinical research.

However, much of medication information is stored in unstructured text, such as doctor's letters. Figure 6.1 shows a snippet of a diagnosis section in a doctor's letter containing medication names and other medication-related information, e.g., duration of a medication and reasons for prescription. Typically, this information must be manually extracted by clinical domain experts, a process which is not scalable, time-consuming and error-prone. Automatic medication information extraction methods could facilitate the extraction from large volumes of data, improve data quality, and save valuable time of clinical staff. Recent

[1] https://github.com/dieterich-lab/medication_information_extraction_using_llms, last accessed 24.11.2025.

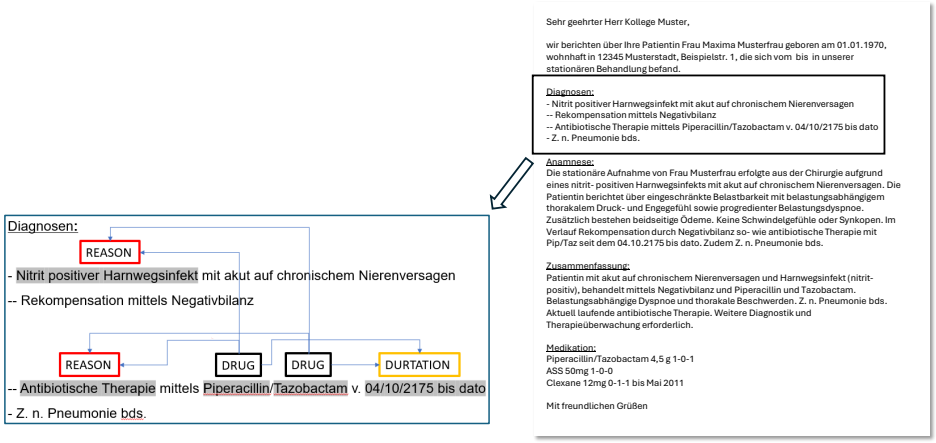


Fig. 6.1 **Doctor's letter:** Snippet of a doctor's letter containing medication information in the diagnosis section.

advances in NLP and machine learning have demonstrated the potential of generative LLMs and PEFT methods to automate various information extraction tasks.

We aim to evaluate generative LLMs to extract medication information from doctor's letters and evaluate best-practice strategies to identify an ideal setup to address our clinical constraints discussed in Section 2.1. Specifically, for each constraint, we propose the following solutions:

1. **On-premise resource constraints**

   1.a. **Domain expertise** Our first contribution focuses on incorporating clinical domain expertise. We utilize the findings from (Singhal et al. 2023) claiming that foundation LLMs can encode clinical knowledge. Furthermore, we efficiently evaluate domain-adapted medical LLMs and use clinical knowledge stored in small-sized high-quality annotated gold standard data sets in English and German to fine-tune and evaluate our models yielding SOTA on difficult evaluation categories (Hu et al. 2021).

   1.b. **Staff time** To optimize time resources for pipeline development and maintenance we define our MIE task as a one-step joint NER and RE task. To support automatic evaluation we use well-defined structured output formats with established evaluation metrics (Singhal et al. 2023; Dagdelen et al. 2024). To address the

often-complex process of evaluation of LLMs we utilize feedback LLMs leveraging findings of (Chiang et al. 2023; Sharif et al. 2025) (cf. Section 3.4). We establish best-practice approaches for structured output generation in a clinical domain for two languages using a feedback LLM for automatic evaluation.

1.c. **Local compute resources** To address the computational costs of fine-tuning LLMs, we use SOTA PEFT methods based on low-rank adaptations (LoRA), parameter quantization and smaller expert MIE LLMs in a clinical environment (Hu et al. 2021; Dettmers et al. 2023; Lehman 2024). Thus, we showcase that smaller fine-tuned expert LLMs reliably generate structured output and yield new SOTA results in a MIE task.

1.d. **Native-language barrier** We fine-tune LLMs on limited amounts of German language clinical documents (CARDIO:DE) (Chen et al. 2024b). Thereby demonstrating the applicability of publicly available language-adapted LLMs for a MIE task.

2. **Transparency requirements** As previous studies showed, saliency maps can support users in the clinical domain to explain model predictions (Kayser et al. 2024). To increase transparency and to support our evaluation results we use Shapley values that are optimized for generative LLMs using the model interpretability library Captum (Lundberg et al. 2017; Miglani et al. 2023) (cf. Section 3.5). We present two use cases to support model evaluation by investigating relation extraction and clinical reasoning capabilities using attributions of LLMs. Furthermore, our MIE approach extracts structured information from unstructured texts, thereby inherently enhancing data interpretability and enabling more effective information retrieval.

We present a novel, light-weight, fine-tuned e2e LLM pipeline for medication information extraction, achieving new SOTA on English data and setting a new performance benchmark on German clinical data. We leverage feedback-LLMs to facilitate automated evaluation and apply Shapley values to improve transparency, offering a practical application under real-world clinical constraints.

6.3 Methods

We define our MIE experiments as a one-step e2e joint NER and RE task. We created an evaluation framework that includes four main steps: (1) data preprocessing (cf. Section 6.3.2), (2) fine-tuning an LLM on training data/applying it zero-shot (cf. Section 6.3.4), (3)

generating the JSON output (cf. Section 6.3.4), and (4) evaluating the overall performance of the model (using micro average F_1-score (cf. Section 6.3.3) and per relation class (using F_1-score) supported by feedback LLMs (cf. Section 6.3.5). For inference, we apply (a) data preprocessing, (b) generating the JSON output (cf. Section 6.3.4), and (c) to support interpretability we use saliency maps to visualize input token contributions to the JSON output (cf. Section 6.3.6). To increase reproducibility and transparency, all experiments where conducted using deterministic decoding (cf. Section C.6).

Figure 6.2 illustrates the e2e MIE pipeline for (top) evaluation and (bottom) inference. We illustrate the evaluation pipeline (top) with the example sentence: *Metoprolol 50mg IV once, then transitioned to PO twice daily for hypertension control.* (i) From the gold standard annotations, we derive gold JSON strings that encode all medication entities and their relations (joint NER+RE). In our final dataset every paragraph is now paired with such a gold JSON. (ii) Using format-restricting prompts, we either train a LoRA model on this dataset or conduct a zero-shot baseline with the instruct model. (iii) Given the input paragraph, the LLM outputs a predicted JSON using the same schema as the gold JSON. (iv) To evaluate the final model, we convert both gold and predicted JSONs into sets of triplets and compute exact and lenient F_1-score. In our example the model predicted a STRENGTH of 50mg for the MEDICATION *Metoprolol*. Hence, this counts as a true positive. For pattern mismatches (e.g. *IV and PO* vs. *[IV, PO]*), a feedback LLM validates clinical correctness and we update metrics accordingly. At inference time (bottom), we apply the same preprocessing and generate structured JSON output. Additionally, we compute token-level Shapley value attributions to support model interpretability.

6.3.1 Data

All experiments were conducted on publicly available, manually annotated clinical routine corpora in English (N2C2 2018)(Henry et al. 2020) and German (CARDIO:DE)(Richter-Pechanski et al. 2023) language.

n2c2 2018 (track 2)

The corpus was distributed for the 2018 N2C2 shared task on adverse drug events and medication extraction in electronic health records and contains 505 English doctor's letters from the MIMIC-III clinical care database (Johnson et al. 2016). It was manually annotated with nine medication entity classes (DRUG, STRENGTH, FORM, DOSAGE, FREQUENCY, ROUTE, DURATION, REASON, ADVERSE DRUG EVENT) and eight medication relation classes (DRUG-STRENGTH, DRUG-FORM, DRUG-DOSAGE, DRUG-FREQUENCY, DRUG-

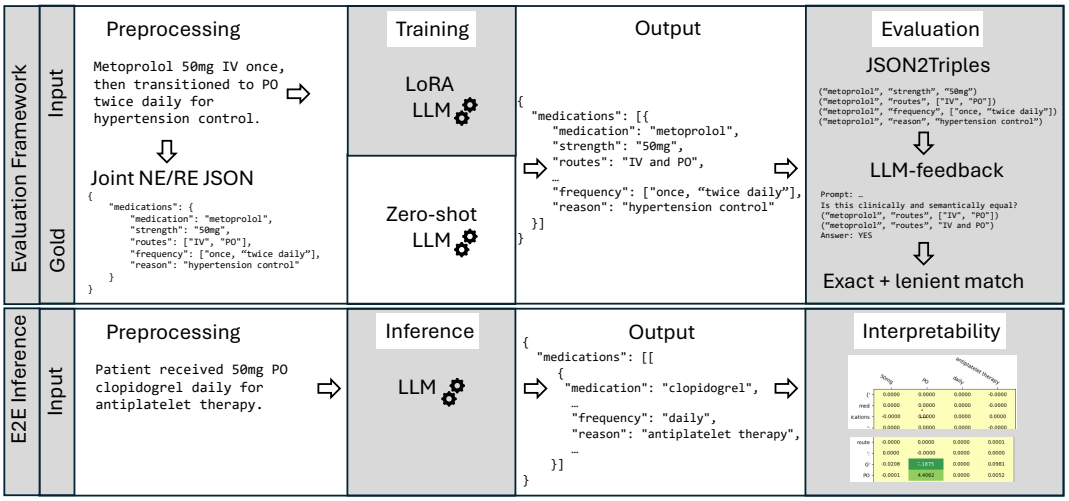


Fig. 6.2 **Medication information extraction:** (top) Model training/evaluation (preprocessing gold labels, model output generation, JSON2triplets, feedback LLM). (bottom) Inference (JSON output, interpretability). For step-by-step description, see Section 6.3.

ROUTE, DRUG-DURATION, DRUG-REASON, DRUG-ADVERSE DRUG EVENT) (Henry et al. 2020). The corpus is split into 303 letters in the training set and 202 letters in the test set. Overall, the corpus contains 83,869 entities and 59,810 relation annotations (cf. Table C.1, Table C.2).

CARDIO:DE v1.1

CARDIO:DE v. 1.1 contains 500 German doctor's letters from the cardiology department at the Heidelberg University Hospital (cf. Section 4 and (Richter-Pechanski et al. 2023)). In summary, the corpus was manually annotated with nine medication entity classes (ACTIVEINGREDIENT, DRUG, STRENGTH, FORM, DOSAGE, FREQUENCY, ROUTE, DURATION, REASON) and seven medication relation classes (MEDICATION–STRENGTH, MEDICATION–FORM, MEDICATION–DOSAGE, MEDICATION–FREQUENCY, MEDICATION–ROUTE, MEDICATION–DURATION, MEDICATION–REASON). For our experiments we merged the DRUG (brand name) and ACTIVEINGREDIENT (active pharmaceutical ingredient) class. The corpus is split into 400 letters in the training set and 100 letters in the test set. Overall CARDIO:DE contains 27,155 entities and 19,336 relation annotations (further statistics cf. Table 4.6).

6.3.2 Data preprocessing

Following best-practice in (Dagdelen et al. 2024), we converted all annotations of the English and German datasets into uniform JSON strings. The datasets contain a single doctor's letter per file. We split each letter by newline (`\n`). We avoided sentence splitting, as manual evaluations showed frequent splitting errors for clinical texts. A JSON string with a single key `medications` and an empty list as a value is created. For each medication in the input text a JSON object with the medication name (entity) and all related information classes and their assigned entities is created and appended to the `medications` list (a comprehensive example, cf. Figure C.1). However, related information of a medication name can appear across newline borders. Thus, we merged newline samples, if related medication information is contained in a neighboring sample (cf. Table C.3).

This results in a data set which contains duplicate medications and medication relations. For the English n2c2 data set we count approximately 2.6% more relation information instances.

We also experimented with character offset information in the generated JSON object to define word boundaries and annotation boundaries, similar to traditional annotation schemes (eXtensible Markup Language (XML), Brat Rapid Annotation Tool (BRAT), etc.). However, manual evaluations showed that our LLMs did not reliably count correct character offsets for each medication information. To overcome this problem, for each medication in the input text we appended a medication object in the same textual order as in the text to the JSON string. However, due to the lack of character offset information we could not use the official evaluation scripts of the N2C2 shared task, as they require this information.

Therefore, due to duplicates in the data set and missing character offset information, our results are not exactly comparable to the shared task baselines. We added a comprehensive data analysis comparing the official N2C2 format vs. our JSON format in Appendix C.1.

Despite these drawbacks, following similar recent publications, the JSON representation allowed us to represent the complex information stored in e2e (concepts + relations) annotations in a relatively compact format. The JSON strings can store multiple medications, with multiple related medication information. Furthermore, each relation information can be a list of strings containing multiple information values (e.g., `'strength': ['5mg', '10mg']`).

6.3.3 Metrics

Generative LLMs return free text, not token offset information. Recent IE work therefore adopted soft/overlap metrics instead of exact metrics (Jiang et al. 2024; Hu et al. 2024a;

Sharif et al. 2025; Moral-González et al. 2025). Our custom evaluation script (published on GitHub) follows these best-practices by first normalizing the generated output and credit token overlap.

For all experiments we adopted the primary evaluation metric of the N2C2 shared task, lenient micro average F_1-score per model and lenient F_1-score per relation class. Lenient F_1-score is 100% for relation class X if the drug name and the relation value is exactly matching or if there is an overlap between the gold relation value and the predicted relation value (lenient match examples, cf. Table C.8). Furthermore, we present exact F_1-score results in Table C.9 and Table C.10 and a comprehensive analysis between exact and lenient matching in Appendix C.3.1.

6.3.4 Local large language models

We used the official Llama 3.1 base model with 8 and 70 billion parameters (`meta-llama/Meta-Llama-3.1-[8b|70b]`) published by META PLATFORMS for fine-tuning experiments, the respective instruct models (`meta-llama/Llama-3.1-[8b|70b]-Instruct`) for zero-shot experiments and OPENBIOLLM with 8 billion parameters (`aaditya/Llama3-OpenBioLLM-8b`), a `meta-llama/Llama-3.1-fine-tuned (FT)-Instruct` model further-pretrained on biomedical texts. At the time of experiments, OPENBIOLLM was the SOTA of medical domain-adapted LLMs (all used models, cf. Table 6.1) (Zhou et al. 2023).

	Zero-shot	**Fine-tuned (FT)**
Model	Llama 3.1 Instruct 8b/70b (4-bit) OPENBIOLLM 8b	Llama 3.1 Base 8b/70b (4-bit) OPENBIOLLM 8b
Training set	No training	Full training set
Training method	–	LoRA (r=16)

Table 6.1 **Local large language models:** LLMs used for zero-shot and fine-tuning MIE experiments including training set and training method.

For LoRA and QLoRA fine-tuning, we used the `FastLanguageModel.get_peft_model` implementation of the UNSLOTH library (cf. Section 6.3.4) (Daniel Han et al. 2023). In this work we utilize (i) QLoRA for training Llama 70b and (ii) LoRA to train all 8b LLMs. For fine-tuning we used the complete training data set of N2C2 for the English experiments and CARDIO:DE for the German experiments (cf. 6.3.1).

Low-rank adaptation and quantized low-rank adaptation

Due to the restricted computational resources available in a clinical environment and strict data protection regulations, it is necessary to employ efficient methods for training and inference of LLMs which possess billions of parameters.

LoRA is a PEFT method which freezes the LLM's weights and injects trainable rank decomposition matrices into each transformer layer. This reduces the number of trainable parameters for fine-tuning and enables efficient task-switching by requiring only small, task-specific parameter updates. This makes LoRA a scalable solution for deploying large models in resource-constrained environments, such as the clinical domain (further details, cf. Section 2.4.3 and (Hu et al. 2021)).

QLoRA further reduces memory usage during fine-tuning and extends LoRA by quantizing precision of the weight parameters of a LLM to 4-bit precision (for further information, cf. Section 2.4.3 and (Dettmers et al. 2023)).

System prompt & structured output

We used a system prompt with instructions for the one-step e2e medication information extraction task. The model is prompted to extract all drug names including all related medication information in the order they appear in the paragraph. If > 1 medications with the same name appear in the paragraph, we added a counter to the medication name in the order they appear in the input (system prompt cf. Figure 6.3).

```
You are a physician.  Your task is to extract ALL drug names (active
ingredients or drug names) and their related information, such as ADE,
strength, frequency, duration, route, form, dosage, and reason from a given
text snippet of a doctoral letter.
Please make sure to extract the medications in the order they appear in the
text.  Maintain this order in the JSON response.
If a medication occurs more than once in the text, append a unique count in
parentheses to its name, starting from (1).
```

Fig. 6.3 **System prompt for MIE:** system prompt used for all MIE experiments (all LLMs, all datasets) except the ADE information, which is not available in the DE dataset.

Following current SOTA approaches, to steer LLMs to structured output, we added a pattern definition for the JSON output (Dagdelen et al. 2024). We used well-defined PYDANTIC object definitions as a format-restricting instruction in the systems prompt (cf. Section 2.4.3 and https://pydantic.dev/articles/llm-intro). Using PYDANTIC classes simplifies

structure definition and maintenance. Furthermore, we were able to add demonstrations in natural language for each medication information class (cf. Figure C.5).

6.3.5 Evaluation and feedback LLMs

For a thorough evaluation we compared the prediction of each medication information class assigned to a medication in the JSON output with the gold standard annotation. For easier processing, we consistently converted each JSON of the gold standard and the prediction into a set of triplets, containing the medication class value (triplet head), the relation class (triplet predicate/class) and the relation value (triplet tail). Based on this structure we calculated precision, recall and F_1-score per model (micro average) and per relation information class (cf. Table 6.2).

Predicted JSON	**Convert into triplets**

`{`
`  "medications": [`
`  {`
`  "medication": "metoprolol",`
`  "strength": "50mg",`
`  "routes": "IV and PO",`
`  "frequency": ["once", "twice daily"],`
`  "reason": "hypertension control"`
`  }`
`  ]`
`}`
	`("metoprolol", "strength", "50mg")`
	`("metoprolol", "routes", "IV and PO")`
	`("metoprolol", "frequency", ["once", "twice daily"])`
	`("metoprolol", "reason", "hypertension control")`

Table 6.2 **JSON to triplets:** Converting predicted JSON strings into triplets for evaluation.

In contrast to traditional classification models, LLMs generate a sequence of tokens (in our setup: JSON strings). During manual evaluation using exact and lenient F_1-scores, we discovered several patterns of false negative or false positive instances during manual analysis. For example, the model predicts

```
("metoprolol", "routes", "IV and PO")
```

and the gold standard is

```
("metoprolol", "routes", ["IV", "PO"])
```

The gold standard defines the medication route as a list of two route strings, while the LLM predicted a simple string containing both ROUTE values connected by a coordinating conjunction *and*. However, the predicted value and the gold standard are clinically similar.

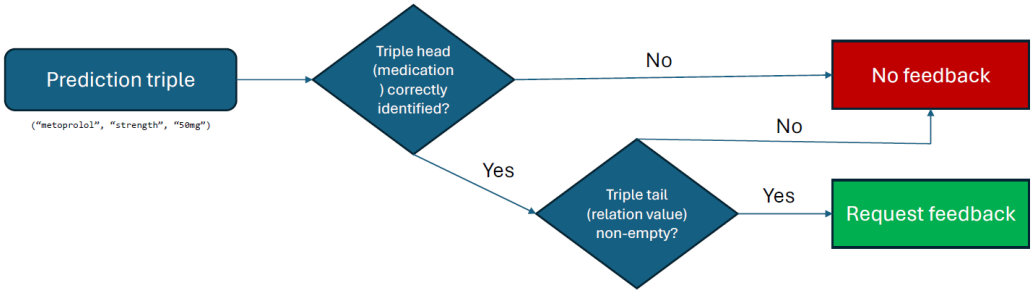


Fig. 6.4 **Feedback LLM:** Pipeline to re-evaluate lenient false positive or false negative predictions of the best performing model (fine-tuned Llama 70b) for English and German data.

Classifying such a case as false positive or false negatives would reflect mainly a technical issue, with little relevance for clinical routine. While handling such a simple case with heuristics and pattern matching seams feasible, we identified numerous patterns of false positives and negatives showcasing instances where a simple pattern matching would fail (cf. Table C.15).

Because manual re-evaluation to identify these patterns is tedious and time-consuming, we followed recent research approaches using feedback LLMs for evaluation (Chiang et al. 2023; Sharif et al. 2025). We employed the 4bit quantized Mistral-Large 123b (`unsloth/Mistral-Large-Instruct-2407-bnb-4bit`) as an expert LLM, to ensure diversity from our main MIE model. We performed a binary classification task: Given the input paragraph, the gold and the predicted value of false positives and false negatives, we prompted the model to classify the sample as clinically SIMILAR/NOT SIMILAR (system prompt cf. Table C.16 and C.17). Cases of unrecognized or hallucinated medications or relation values are treated as valid false negatives or positives (Figure 6.4). We applied this approach for our best performing model (Llama 70b FT) for the English and German data set. A clinical expert manually reviewed a random sample of the classification from a medical perspective (further details, cf. Section C.4.1).

6.3.6 Shapley values

We follow the interpretability setup introduced in Section 2.5 and Section 3.5 and use SHAP to interpret our medication information extraction models.

We use interpretability functionalities that are specifically designed to analyze the behavior of generative LLMs introduced in CAPTUM v 0.7. (cf. Section 2.5). This implementation offers various perturbation-based attribution methods and saliency map visualizations, including Shapley values (Miglani et al. 2023). The method generates Shapley values of each input token to each generated output token. Finally, we provide further insights in model performance by presenting two use-cases using Shapley values for MIE model interpretability.

6.4 Results

In this section we present the one-step e2e joint NER and RE task results. We first present the LLM performance of zero-shot and SOTA experiments. Next, we present performance of fine-tuning experiments and further refine our results using feedback LLMs. All hyperparameters for fine-tuning, inference and interpretability are presented in Section C.6.

6.4.1 Baselines

For the English and German data we conducted zero-shot experiments as a lower-bound baseline. Furthermore, for the English data we compared to two best-performing SOTA results for the MIE task, as reported by (Modi et al. 2024), on the N2C2 data set, as our baseline. For the German CARDIO:DE data set, no MIE SOTA results had been reported so far. Thus, we conducted experiments using the encoder-based MEDBERT.DE model, which, at the time of our experiments, was the only German BERT model pre-trained on 4.7 million German medical documents and achieved SOTA results on various medical information extraction tasks (Bressem et al. 2024).

Zero-shot

N2C2 On N2C2, Llama 8b achieved the worst overall micro average F_1-score with 61.6% (Table 6.3). Llama 70b increased performance by $\approx$ 15 pp. Both models showed worst performance for the complex classes adverse drug event (ADE) (27.2%;37.4%) and REASON (28.5%;53.1%). However, for classes DOSAGE, FREQUENCY, ROUTE and STRENGTH, F_1-score was above 82%. Best-performing classes for the Llama 8b model were FREQUENCY (82.8%) and ROUTE (80.0%).

CARDIO:DE The micro average F_1-score of Llama 8b on CARDIO:DE was only 5.4%, achieving a maximum of 14.3% for FREQUENCY and STRENGTH (13.9%) (Table 6.4). While micro average recall was 68.0%, precision only achieved 3.0%. A closer analysis confirmed this observation, demonstrating that Llama 8b exhibits substantial hallucinations, attributing medication information to nearly all input samples. In contrast, Llama 70b achieved a micro average F_1-score of 71.0%. The mean performance underperformed compared to the English data set. However, FREQUENCY (86.7%) and STRENGTH (80.6%) achieved best F_1-scores.

For Llama 8b and 70b the amount of malformed JSON outputs for either data set was marginal ($0.002 - 0.2\%$). We were unable to reliably evaluate OPENBIOLLM results, as the model frequently generated malformed JSON strings on the English and German dataset making automatic evaluation within our framework infeasible due to its dependence on well-structured JSON output.

SOTA

N2C2 (Wei et al. 2019) submitted the best performing system combining an RNN and CNN architecture in the N2C2 2018 shared task, with a micro average F_1-score of 89.1%. Except ADE, DURATION and REASON all class-wise F_1-scores are $> 93\%$. For ADE the system achieved 47.6%, for REASON 57.9%. This is only five percentage points above our Llama 70b in a zero-shot setting. (El-allaly et al. 2021) used a SCIBERT model and validated their results with the official N2C2 ADE-RE evaluation script (!not e2e), outperformed the system with 91.3% micro average F_1-score. They especially improved results of ADE (64.6%), DURATION (82.2%) and REASON (78.0%).

CARDIO:DE We developed a pipeline including NER and RE using MEDBERT.DE to (1) classify all medication entities and (2) all relations between the extracted entities. We achieved a micro average F_1-score of 74.1%. This is only three percentage points above the Llama 70b zero-shot model. Best performing classes are frequency (86.4%), ROUTE (78.4%) and STRENGTH (83.4%). All other classes remained below 55%. The DOSAGE class performed by far the weakest with 19.3% F_1-score. Further analysis revealed that, unlike the English dataset, the German dataset typically includes dosage information within FREQUENCY information (e.g., *1 tablet in the morning* as *1-0-0*). Furthermore, with only 143 instances and a relatively low IAA of 73%, this class remains difficult to extract across all methods.

6.4.2 Fine-tuned

N2C2 Fine-tuning Llama 8b on the N2C2 training set, the model achieved a micro average F_1-score of 91.2%, outperforming the zero-shot baseline by ≈ 29 pp. and on-par result with the baseline of (El-allaly et al. 2021) (Table 6.3). However, Llama 8b outperforms this baseline for ADE (+5.5 pp.) and REASON (+1.5 pp.). Llama 70b outperforms the SOTA with 91.7% micro-average F_1-score. Particularly, for the complex classes ADE (+8.2 pp.) and REASON (+4.1 pp.) and the DURATION class (+3.8 pp.). For the classes DOSAGE, FORM and ROUTE Llama 70b performs $\approx 1-2$ pp. worse than SOTA.

OPENBIOLLM 8b did not improve results of Llama 8b (Table C.11). In contrast to the zero-shot experiments, all models, including OPENBIOLLM, observed to the JSON format.

CARDIO:DE Fine-tuning Llama 8b on the CARDIO:DE training set, the model achieved a micro average F_1-score of 88.2%, substantially improving the results of the zero-shot model (+83 pp.) and outperforming the BERT baseline by four percentage points (Table 6.4). Particularly for the complex REASON class (+25 pp.). Llama 70b further increased micro average F_1-score to 88.9% and REASON to 66.2%.

OPENBIOLLM 8b achieved a micro average F_1-score of just 7%, showing severe hallucinations, hence not benefiting from fine-tuning (Table C.12). However, similar to the English data, all models observed the JSON format.

We calculated confidence intervals for the English and German data for Llama 70b FT in Appendix C.3.3.

6.4.3 Using feedback pipeline

According to manual analysis of domain experts of false predictions, supposedly false positives and false negatives, in fact, frequently contained arguably clinically correct results. To automate manual evaluation of such instances, we used an external expert model (`Mistral-Large` 123b) to perform a binary classification for these samples compared to the gold standard (SIMILAR/NOT SIMILAR). We applied this approach for our best performing model (Llama 70b FT) for the English and German data set. Further details, see Section C.4.1.

N2C2 By using a feedback LLM to support automatic evaluation, micro-average F_1-score could be increased to 92.8% (Table 6.3). Filtering false positives and false negatives particularly increased results for complex classes ADE (+2.1 pp.) and REASON (+2 pp.), but as

Medication information	Baselines		Llama 8b		Llama 70b		
	Wei et al.	El-Allaly et al.	zero	FT	zero	FT	FT (feedback)
ADE	47.6	64.6	27.2	70.1	37.4	72.8	**74.9**
DOSAGE	93.6	**94.3**	68.4	92.7	81.9	92.5	93.1
DURATION	78.5	82.2	51.1	82.2	54.7	86.0	**86.4**
FORM	95.1	95.3	56.6	93.8	64.9	94.1	**95.8**
FREQUENCY	95.8	94.9	82.8	94.8	87.4	95.3	**96.1**
REASON	57.9	78.0	28.5	79.5	53.1	82.1	**84.1**
ROUTE	94.2	93.9	80.0	93.6	87.3	93.3	**94.3**
STRENGTH	**97.2**	95.8	78.2	96.0	87.9	96.2	96.6
Micro avg.	89.1	91.3	61.6	91.2	76.8	91.7	**92.8**

Table 6.3 **Lenient F_1-scores for N2C2 corpus for e2e MIE task:** Comparing two SOTA baselines with zero-shot (zero) and FT (fine-tuned) Llama 8b and 70b and optimized evaluation using feedback LLM (FT feedback) for Llama 70b.

well for FORM (+1.7 pp.) and FREQUENCY (+0.8 pp.). Hence, Llama 70b established a new SOTA for complex classes ADE (74.9%) and REASON (84.1%), and further for DURATION (86.4%) and FORM (95.8%).

CARDIO:DE The overall result was slightly increased to micro average F_1-score 89.7%. Similar to N2C2, the biggest impact we observed for the complex class REASON (+3.5 pp.) and FORM (+1.5 pp.) (Table 6.4). Hence, re-evaluation established a new benchmark for the German data set.

6.4.4 Interpretability

Our results show that LLMs achieved new SOTA results for MIE. However, due to their black-box nature, the interpretation of results remains challenging, which limits their adoption in the clinical domain (Kayser et al. 2024). While interpretability methods such as Shapley values are well-studied to explain ML models for structured data sources, they are increasingly used to interpret deep learning models for NLP tasks (further details, cf. Section 6.3.6) (Richter-Pechanski et al. 2024; Richter-Pechanski et al. 2025). However, their usage for generative LLMs remains widely understudied (Miglani et al. 2023). We therefore identified two use cases to apply Shapley values to further analyze LLMs in the context of a generative MIE task: (1) assessing the contributions of input tokens to relation information output tokens, (2) uncovering implicit knowledge on relation information. Below, we present two representative

Medication information	Baseline (medBERT.de)	Llama 8b		Llama 70b		
		zero	FT	zero	FT	FT (feedback)
DOSAGE	19.3	0.0	27.8	6.1	50.0	**50.0**
DURATION	48.9	1.9	76.3	48.0	76.7	**77.8**
FORM	54.5	0.0	73.1	17.3	84.2	**85.7**
FREQUENCY	86.4	14.3	95.0	86.7	**96.4**	**96.4**
REASON	40.8	1.5	64.3	26.9	66.2	**69.7**
ROUTE	78.4	3.2	**91.6**	76.0	89.2	89.3
STRENGTH	83.4	13.9	93.1	80.6	93.6	**94.1**
Micro avg.	74.1	5.4	88.2	71.0	88.9	**89.7**

Table 6.4 **Lenient F1-scores on the CARDIO:DE corpus for the e2e MIE task:** We compare a SOTA baseline with zero-shot (zero) and fine-tuned (FT) Llama 3.1 8b and 70b, and an optimized evaluation using a feedback LLM (FT (feedback)) for Llama 70b.

examples for each use case (further examples, cf. Appendix C.7 and (Richter-Pechanski et al. 2025)).

Use case 1 To evaluate whether input token contributions align with the relation information output tokens, we investigated a key edge case in MIE. Since our MIE task, aside from JSON syntax, reproduces input tokens in the output, their contribution to each output token may appear trivial. For instance, if the input contains the relation information for strength *5 mg* and the model correctly predicts `'strength': '5 mg'` in the JSON string, the contribution alignment seems straightforward. However, when the same strength relation value (cf. Figure 6.5 top) appears twice in the input, each linked to a different medication name, we can analyze whether the first occurrence correctly maps to the first medication, and the second to the corresponding second medication.

Investigating the contributions of input tokens, as indicated via Shapley values (cf. Figure 6.5 bottom) for the Llama 70b FT model, we observed that the first occurrence of the strength token correctly contributes to the strength value of the first medication (*Amlodipine*) in the JSON output. Specifically, the first *5 mg* token is linked to *Amlodipine*, showing the highest contribution to its JSON entry, while the second *5 mg* corresponds to *Lisinopril*, with the highest contribution to its respective JSON entry. In both cases we observed that the strength input token has the strongest positive or negative contribution to the initializing quote of the respective strength value. The contribution of the respective digit is significantly lower.

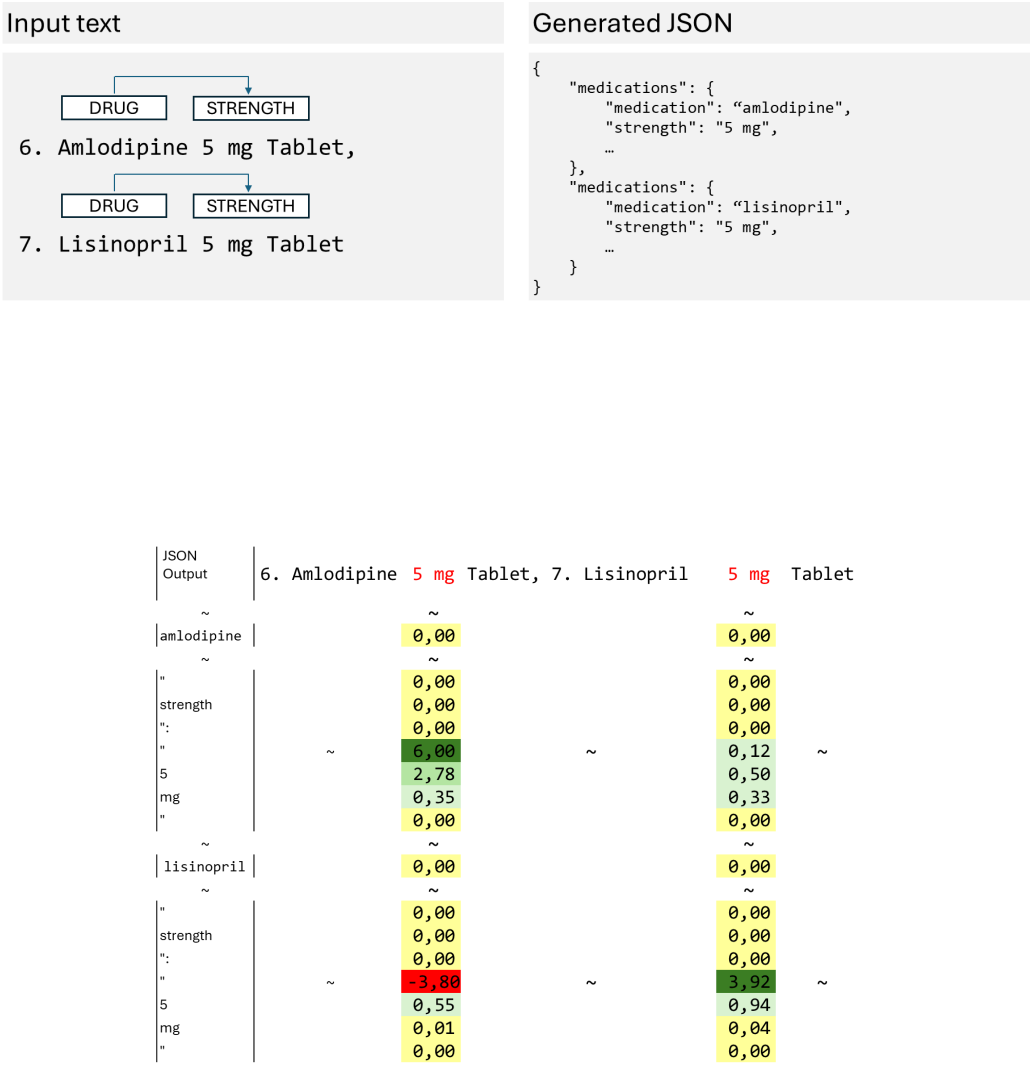


Fig. 6.5 **Shapley value use case 1:** (top) input data—example text containing two medications (*Amlodipine, Lisinopril*) each related to a similar strength value (*5 mg*), and the corresponding generated JSON snippet with medication names and strength values. (bottom) Shapley values—approximate attributions for the strength tokens in the input text *6. Amlodipine 5 mg Tablet, 7. Lisinopril 5 mg Tablet* and the generated JSON token for the strength relation class and value. Complete output cf. (Richter-Pechanski et al. 2025) Appendix C.

Use case 2 The second use case uncovers that LLMs possess implicit knowledge regarding relation information, even when they do not explicitly generate this information in the JSON output. We selected a representative instance of the N2C2 dataset. This instance contains an ADE (*acute renal insufficiency*) which was not generated in the JSON output by our Llama 8b FT model (Figure 6.6 top).

We therefore calculated Shapley values only of the ADE tokens for this instance (Figure 6.6 bottom). These tokens contribute negatively to the empty ADE output of the model. Upon further analysis, we found that the output probability of the empty string decreased to 61%. This suggests that while the model correctly recognizes the input as an ADE, the probability assigned to the corresponding ADE relation in the output remains insufficient for its explicit generation. We conducted additional manual evaluations, which revealed a similar behavior across various relation classes, as shown in Section C.7.2 and conducted a quantitative evaluation across all false negative predicted ADE and REASON instances in the N2C2 data (cf. Section C.7.3).

Based on our findings, in a future study, we plan to use Shapley values to uncover implicit relation information in LLMs in a clinical scenario. Through a graphical user interface, physicians will be presented with instances where input tokens have strong negative contributions to empty relation information outputs. These instances will be provided for further analysis, helping to identify and investigate potential false negative predictions.

6.5 Discussion

In this section, we discuss our empirical findings in light of the clinical constraints and proposed solutions outlined in Section 6.1 and Section 6.2 and our research questions presented in Section 1.2.

1. **On-premise resource constraints**

    1.a. **Domain expertise** (relevant RQs: RQ 1 and RQ 3)
    Based on the observation that LLMs contain clinical knowledge, we evaluated both open-source foundation LLMs based on Llama and the domain-adapted biomedical OPENBIOLLM, comparing their performance in zero-shot and fine-tuned setups, evaluated on well-curated English and German datasets.

    - **Finding** In a zero-shot setup on English data, Llama 8b and 70b showed strong performance but clearly lagged behind SOTA. On German data, Llama 8b showed strong hallucinations, while Llama 70b performs more stably but is still inadequate for a clinical setup (Section. 6.4.1).

Fig. 6.6 **Shapley value use case 2:** (top) input data: (left) Example input text containing ADE of the medication *Tetracycline*. (right) Corresponding generated JSON output snippet containing the medication name and the empty ADE value. (bottom) Shapley values: Visualizing approximate Shapley values for the ADE token of the input text *Also in the setting of his acute renal insufficiency, his tetracycline was held as it was possible that this could be a ...* and the generated JSON token for the ADE relation class and value. Complete output cf. (Richter-Pechanski et al. 2025) Appendix C.

- **Finding** In a fine-tuning setup on English data, Llama models define a new SOTA for complex classes such as ADE, REASON, and DURATION. On German data, both models show an overall lower performance in comparison to English data, however, they set a new benchmark for most classes on German data (Section. 6.4.2).
- **Finding** In a zero-shot setup OPENBIOLLM frequently generated malformed structured format predictions in English and German. In a fine-tuning setup OPENBIOLLM showed reliable structured output format and comparable performance to Llama on English data. However, while structured output was generated reliably on German data, the model consistently produced hallucinations (Table C.11 and C.12).

Leveraging clinical knowledge of generative LLMs combined with fine-tuning on well-curated gold standard data in a complex token-level e2e task, we could reduce the demand for clinical expertise while achieving superior or competitive performance to MIE baselines.

1.b. **Staff time** (relevant RQs: RQ 1, RQ 3)
We defined our MIE task as a one-step e2e joint NER and RE task on two gold standard data sets with well-defined structured output formats and used a feedback LLM to support automatic evaluation.

- **Finding** In a zero-shot setup, both Llama models reliably produced structured output. In a fine-tuning setup, all models consistently followed the defined output structure reducing post-processing efforts.
- **Finding** The generative capabilities of LLMs simplified the e2e pipeline, significantly reducing both development and maintenance efforts. However, all LLMs occasionally deviate from the gold standard, while frequently remaining clinically correct. This motivated us to leverage feedback LLMs to assist in assessing critical predictions.
- **Finding** Feedback LLMs helped to speed-up the oftentimes complex evaluation of LLM outputs and further helped identify frequent false positives and negatives that were actually correct. Additionally, they supported our findings that LLMs surpass the current SOTA (Section 6.4.3).

By combining format-restricting prompts and generative feedback LLMs in a lightweight e2e pipeline achieve robust SOTA results on a token-level MIE task. Furthermore, manual annotation and evaluation efforts can be minimized.

1.c. **Local compute resources** (relevant RQs: RQ 1, RQ 2)
PEFT fine-tuning using LoRA and quantization on limited training data, enabled us to deploy moderate-sized open-source LLMs for MIE tasks inside the clinical infrastructure.

- **Finding** We could fine-tune a Llama 70b model using 4bit QLoRA on our clinical infrastructure using a single NVIDIA H100 (using 48*GB* of 96*GB* VRAM). 8b models could be fine-tuned using a single NVIDIA RTX6000 (using 14*GB* of 24*GB* VRAM).
- **Finding** Fine-tuning LLMs as clinical expert LLMs achieved new SOTA results for the 70b model, while the smallest 8b model performs only slightly worse offering the best accuracy-efficiency trade-off for most relation classes. This further reduces IT infrastructure requirements, while generating reliable structured output.

Overall, PEFT with quantization achieved new SOTA with limited training data on our e2e MIE task under clinical compute restrictions. However, 70b Llama is beneficial for more complex classes.

1.d. **Native-language barrier** (relevant RQs: RQ 1, RQ 2)
We fine-tuned generative LLMs on English (N2C2) and German (CARDIO:DE) clinical texts.

- **Finding** In a zero-shot setup, Llama 8b produced a substantial amount of hallucinations on German data. Fine-tuning significantly reduced hallucinations and improved LLM performance in both languages particularly for complex classes such as REASON and DURATION. However, in German language the clinical OPENBIOLLM continued to produce hallucinations.
- **Finding** Fine-tuning improved performance in both languages. However, performance on German data remains lower.
- **Finding** In zero-shot setup, using format-restricting prompts, Llama models stick to the JSON format, while OPENBIOLLM frequently produced malformed output. One possible explanation is that further-pretraining of OPENBIOLLM affected Llamas stronger structured generation capabilities. However, since this effect was not isolated in our experiments, this interpretation remains hypothetical.

In practice, prefer Llama models with PEFT fine-tuning on small, high-quality German gold standards: fine-tuning minimizes hallucinations and preserves the JSON schema (already respected by Llama in zero-shot), whereas OPEN-

BIOLLM remains unreliable on German (hallucinations, malformed output); despite these performance gains, German performance still is below English.

2. **Transparency requirements** (relevant RQs: RQ 4)
   We used Shapley values optimized for generative LLMs to support model evaluation and to support transparency of model predictions. Furthermore we combined deterministic outputs and format-restricting prompts to support reproducibility and transparency of model outputs.

   - **Finding** Similar to traditional applications, Shapley values remain useful for measuring token-level input contributions to generative LLM predictions, providing faithfulness at the token level.
   - **Finding** In our scenario Shapley values highlight input token contributions to extracted structured output predictions. For our two use cases this supports understanding the relation extraction performance of LLMs and allowed us to identify implicit clinical knowledge, in cases where the model makes false predictions.
   - Fine-tuning general-domain LLMs using deterministic outputs and format-restricting prompts on gold standard data resulted in reliable structured output in both languages.

   Combined with format-restricting prompts and deterministic outputs, Shapley values and saliency maps improve the prediction transparency of LLMs. Importantly, saliency maps can aid physicians in evaluating predictions, supporting informed clinical decision-making (Section 6.4.4).

6.6 Conclusion

In this study, we presented best-practice strategies for conducting an e2e MIE task with generative LLMs under on-premise and transparency constraints in a lower-resource domain and language. In summary, fine-tuning moderate-sized Llama models with PEFT methods on well-curated gold standard data produced reliable structured JSON outputs and established new SOTA results on English and German data. Using feedback LLMs further supported our SOTA results and decreased manual evaluation efforts, while Shapley values supported transparency of model capabilities.

Domain expertise We could show that clinical knowledge in LLMs combined with PEFT fine-tuning on small but high-quality gold standards can reduce the demand for clinical expertise during pipeline development and evaluation. Llama 70b achieved top performance, optimizing the return of invest by reducing the demand for clinical expertise for creating training data. In contrast, OPENBIOLLM lacks robustness in structured output generation and frequently produces hallucinations. We therefore recommend the use of strong foundational Llama models fine-tuned with PEFT on compact training data sets.

Staff time We showed that all Llama models adhered to the output structure using format-restricting instructions. OPENBIOLLM needed fine-tuning for reliable structured output. While the generative nature of LLMs allowed us to build e2e MIE solutions requiring less development and maintenance effort, it also made evaluation more laborious. Thus, to support automatic evaluation, we presented feedback LLMs to assist in in fine-grained output evaluation. We therefore recommend combining format-restricting prompts with feedback LLMs to reduce manual post-processing efforts.

Local compute restrictions PEFT methods enable fine-tuning and deployment of moderate sized LLMs (8-70b parameters) on realistic local clinical hardware, while achieving SOTA performance with modest VRAM usage. We recommend to use PEFT fine-tuned 8b models for token-level MIE tasks, and only consider 70b models for more complex classes.

Native language Fine-tuning Llama models minimized hallucinations while keeping structured output format and substantially improved performance in English and German. Zero-shot experiments showed critically low performance in German which could be substantially increased after fine-tuning. However, German performance is still overall lower than English. Clinical LLMs such as OPENBIOLLM showed unstable performance on German despite fine-tuning. We recommend to use strong foundation models for MIE tasks and fine-tune them on local high-quality native-language gold standard data.

Transparency requirements To guide generative LLMs structured output we apply deterministic generation and format-restricting prompts. Furthermore, we apply Shapley values adapted to generative LLMs to highlight token-level contributions of input token to structured output. This supports error analysis and supports model interpretability, which is presented in this study in two use cases. To increase model trust and debugging, we recommend to integrate attribution methods and output probabilities during model development and

evaluation.

In summary, our thorough evaluation of local LLMs for a complex token-level NER+RE task using well-established evaluation metrics and large-scale high-quality gold standard datasets and interpretability methods highlight the strong potential of LLMs to support clinical research and improve decision-making in practical clinical settings. Our approach demonstrates a broad applicability across languages, information extraction tasks and clinical use cases, highlighting its value as a foundation for real-world clinical settings (cf. Chapter 7).

Chapter 7

Clinical Application: Medication Trends and Polypharmacy

7.1 Outline and contributions

In this chapter we are using the e2e MIE pipeline presented in Chapter 6, to extract structured medication information from unseen doctor's letters. We then apply these outputs in two real-world clinical use cases: (1) identifying guideline-driven shifts in oral anticoagulation (OAC) treatment and (2) quantifying trends in patient-specific polypharmacy over time.

In Section 7.2 we motivate both clinical applications based on clinical literature. Section 7.3 details the datasets used for our experiments (OAC: CARDIO:DE + doctor's letters from 2012; Polypharmacy: longitudinal letters from a 20-patient cohort (2008 – 2016)). Section 7.4 describes the Llama-3.1-70b FT pipeline including pre-processing steps and statistical analysis. Section 7.5 presents results and analysis for both applications. Section 7.6 discusses our findings and key limitations. Finally, Section 7.7 concludes these findings and presents next steps.

Our experiments directly contribute to RQ 5 under clinical constraints (cf. challenges 1.a. *Domain expertise*, 1.b. *Staff time*, 1.c. *Compute restrictions* and 1.d. *Native language* and 2. *Transparency*) by demonstrating that our MIE pipeline can be applied on unseen German doctor's letters in a clinical routine setup to support downstream tasks. We show that on-premise Llama models fine-tuned on CARDIO:DE generalize to German routine letters of different time periods and identify the OAC shift between 2012 and 2020 – 2021. Furthermore, we demonstrate that these models can identify patients with polypharmacy risk over time. All analyses are conducted within the hospital IT infrastructure, without additional manual annotation work beyond CARDIO:DE.

This chapter does not aim to provide an exhaustive clinical evaluation, but instead to showcase the feasibility and robustness of our e2e MIE pipeline when applied to unseen data. We performed representative preliminary experiments that highlight both the strengths and the constraints of the pipeline for downstream applications. Further in-depth clinical information can be found in relevant guidelines and research studies (Hindricks et al. 2021; Steffel et al. 2021; Delara et al. 2022; Unlu et al. 2020). Analysis presented in this chapter should be considered as a proof of concept.

7.2 Introduction and background

OAC shift In cardiology, OAC such as vitamin K antagonist (VKA) and direct oral anticoagulants (DOAC) are prescribed to patients who require long-term anticoagulation to prevent stroke and systemic embolism in atrial fibrillation and to treat or prevent venous thromboembolism. Evidence in the literature (Aebersold et al. 2024) and cardiology guidelines indicate a shift from VKAs (e.g. *Marcumar/Phenprocoumon*) towards DOACs, (e.g. *Apixaban, Rivaroxaban, Dabigatran, Edoxaban*) between 2012 and 2020-2021 (Van der Hulle et al. 2014; John Camm et al. 2012; Hindricks et al. 2021; Steffel et al. 2021). Our objective was to detect this shift directly from unstructured doctor's letters by extracting medication related to anticoagulation using our e2e pipeline.

Prior work has shown that guideline-driven treatment shifts can be detected automatically from English discharge summaries using NLP (Bean et al. 2019). However, to our knowledge, we are the first to demonstrate this for German doctor's letters.

Polypharmacy Although there is no standardized definition of polypharmacy, it typically describes the concurrent use of five or more medications by a patient. Recent studies have shown that polypharmacy is associated with various adverse health outcomes and re-hospitalization risks (Delara et al. 2022; Unlu et al. 2020). However, manually identifying polypharmacy in doctor's letters is time-consuming and tedious (König et al. 2019). We therefore quantify polypharmacy longitudinally by counting unique medication mentions per patient's doctor's letter over time using our e2e MIE pipeline.

Automatic detection of polypharmacy from unstructured clinical texts has been shown on English data using NLP approaches (Kadra et al. 2015; Socrates et al. 2025), but we found no evidence of comparable automatic approaches for German doctor's letters.

We included clinical documents without gold-standard annotations. In addition to manual evaluation on stratified samples, we adopted two evaluation strategies: for the OAC

experiments, we assessed plausibility against recommendations in OAC treatment in clinical guidelines across time; for the polypharmacy experiments we used descriptive analysis and random sampling to investigate patient medication in doctor's letters.

7.3 Data

OAC shift We use two datasets to detect OAC treatment shift: (1) all 500 doctor's letters from CARDIO:DE covering a time period of $2020-2021$. (2) a random sample of 538 letters from our unannotated local 2012 corpus ($n > 20,000$). The 2012 data contains letters of higher heterogeneity; many brief outpatient letters and short notes, which resulted in fewer anticoagulation mentions overall.

Polypharmacy We selected longitudinal doctor's letters from 20 randomly selected patients with ≥ 18 letters over a time range of $2008-2016$ of our complete local corpus ($n > 200,000$). In addition to plain-text documents, we parsed the original MS DOCX files to obtain admission date information per letter for temporal alignment. We focused on medications mentioned in semi-structured medication sections (cf. Chapter 4.3). The final corpus contains 462 doctor's letters (mean: ≈ 25, min: 18, max: 68 per patient).

7.4 Methods

We applied the one-step e2e joint MIE setup from Chapter 6 including format-restricting prompts and well-defined PYDANTIC object definitions, matching on-premise constraints (challenges 1. *On premise*, cf. Section 2.1). For all experiments, we selected the best-performing Llama 3.1 70b FT model fine-tuned on the medication layer in CARDIO:DE v. 1.1 (cf. Chapter 4).

OAC shift We retrieved all instances of ACTIVEING/DRUG whose REASON relation contained the span *Antikoagulation* (German variants matched case-insensitive). For CARDIO:DE we used the gold standard medication annotation layer, for 2012 data we applied the e2e pipeline to automatically annotate medications (Richter-Pechanski et al. 2025). Medications were mapped to three OAC classes using curated lists extracted from clinical guidelines following (Calkins et al. 2019; Van Gelder et al. 2024): VKA, DOAC and OTHER (*heparins, non-OACs*, etc.) (cf. Table 7.1). We focused on DOAC and VKA, and report OTHER descriptively.

Treatment	Text instances
VKA	*marcumar, macumar, marcumer, phenprocoumon, cumarin, marcumar-therapie, marcumartherapie*
DOAC	*apixaban, eliquis, rivaroxaban, xarelto, edoxaban, endoxaban, lixiana, lixana, dabigatran, pradaxa, noac, noak, doac*

Table 7.1 **OAC class normalization:** Normalized strings used to map extracted strings into (1) VKA and (2) DOAC, following cardiology guidelines (Calkins et al. 2019; Van Gelder et al. 2024) and frequent spelling variants in CARDIO:DE. Non-oral or non-OAC mentions are classified as OTHER.

For each year we calculated the class shares of all oral anticoagulants as $p_{DOAC} = x_{DOAC}/(x_{DOAC}+x_{VKA})$ and $p_{VKA} = x_{VKA}/(x_{DOAC}+x_{VKA})$. Medications classified as OTHER were excluded. We report Wilson 95% confidence intervals for each p. To measure statistical significance we used two-proportion z-tests comparing 2012 vs. $2020-2021$.

We also analyzed the active ingredient composition of the DOAC class (e.g. *Apixaban, Rivoroxaban, Edoxaban, Dabigatran*). We used a lexicon to map brand names to their active ingredients (e.g. Eliquis $\rightarrow$ Apixaban). Figure 7.1 visualizes the complete workflow.

Polypharmacy First, we annotated the selected corpus with medication information using our e2e MIE pipeline. Subsequently, we followed a rule-based approach to identify relevant medication mentions per letter. We did not include our section classification model presented in Chapter 5, as currently it cannot distinguish admission and discharge medication sections. We (i) retrieve the admission date from the MS DOCX metadata, (ii) split paragraphs by whitespace, (iii) focus on regular semi-structured medication sections by only counting a medication mention if the line starts with the medication name predicted in the JSON output, (iv) keep the last mention of a unique medication per letter (to broadly distinguish admission and discharge medication mentions) and (v) count unique medication mentions per letter. Our pipeline explicitly extracts medication information from the letter without any brand/active-ingredient normalization step.

We report polypharmacy per doctor's letter (threshold > 5 and > 10 medication mentions) and at patient level (fraction of patients who ever passed the threshold). We provide a descriptive summary of polypharmacy using medication mention counts and proportions. Additionally, for each proportion (letter- and patient-level) we report Wilson 95% confidence intervals. Hypothesis testing was not applied in this feasibility study.

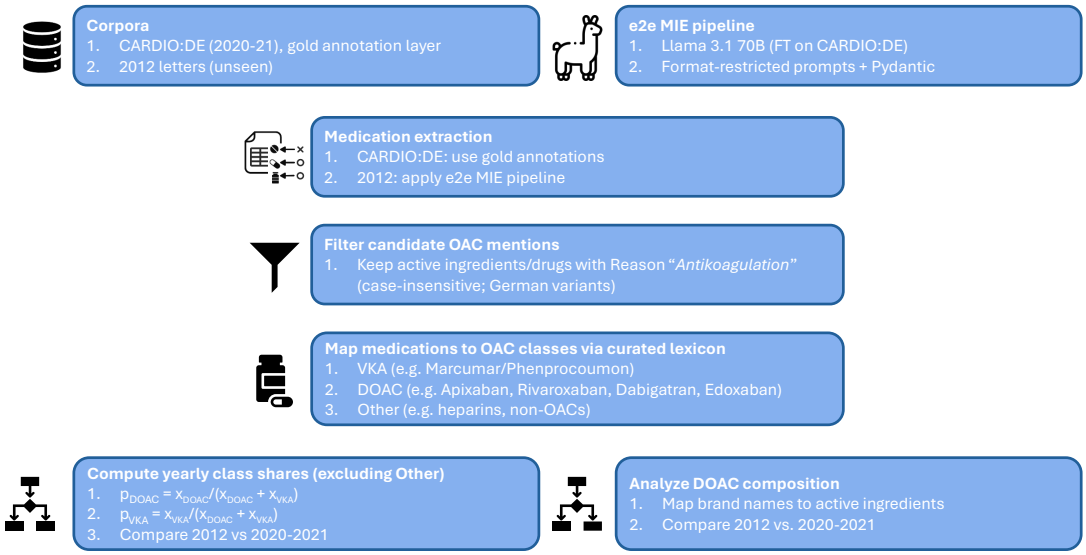


Fig. 7.1 **OAC shift workflow:** Detecting the OAC treatment shift from VKA to DOAC in doctor's letters from 2012 and 2020-21 using the e2e MIE pipeline.

7.5 Results

OAC shift The model overall extracted $x_{OAC,2012} = 77$ oral anticoagulant mentions in the 2012 dataset. The CARDIO:DE gold standard annotations contained $x_{OAC,2021} = 262$ relevant mentions. In 2012 the model extracted $x_{VKA,2012} = 29$ VKAs, $x_{DOAC,2012} = 13$ DOACs and $x_{Other,2012} = 35$ medications. The gold standard in 2021 contains $x_{VKA,2021} = 26$ VKAs, $x_{DOAC,2021} = 157$ DOACs and $x_{Other,2021} = 79$. The relative distribution is shown in Figure 7.2.

The VKA proportion among oral anticoagulants decreased by 27.8% pp. from $p_{2012} = 37.7\%$ (95% CI: $27.7\% - 48.8\%$) to $p_{2021} = 9.9\%$ (95% CI: $6.9\% - 14.1\%$). In contrast, the DOAC proportion increased by 43 pp. from $p_{2012} = 16.9\%$ (95% CI: $10.1\% - 26.8\%$) to $p_{2021} = 59.9\%$ (95% CI: $53.9\% - 65.7\%$).[1] This observation strongly supports the guideline-driven shifts from VKA to DOAC documented between 2012 and 2021 (John Camm et al. 2012; Hindricks et al. 2021; Steffel et al. 2021).

In addition to class-level changes, the composition of DOACs shifted over time (cf. Figure 7.3). In 2012, with 85% the primary DOAC active ingredient was *Rivaroxaban*, while only 15% were linked to *Dabigatran*. In 2021 *Apixaban* dominated (49%), *Rivaroxaban* decreased

[1] Two-proportion z-tests: DOAC $z = -6.64, p = 3.14e-11$ ($k_{2012} = 13, n_{2012} = 77; k_{2021} = 157, n_{2021} = 262$), VKA $z = 5.80, p = 6.49e-09$ ($k_{2012} = 29, n_{2012} = 77; k_{2021} = 26, n_{2021} = 262$).

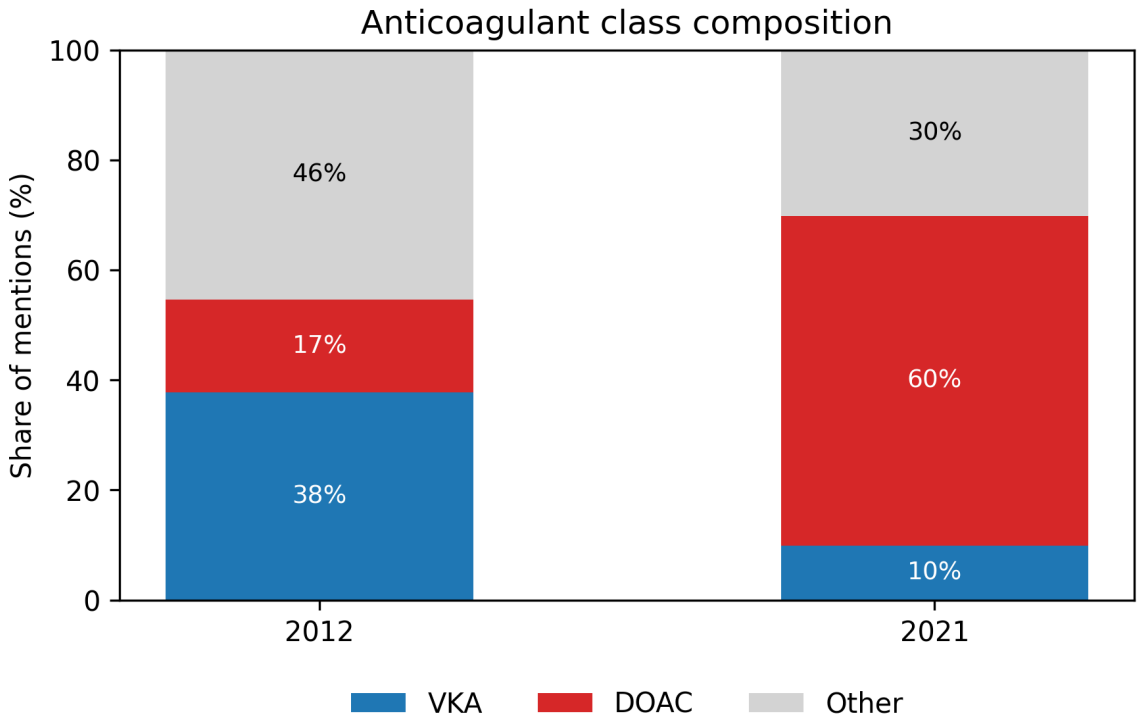


Fig. 7.2 **Anticoagulant class composition:** Composition of OAC treatment classes comparing 2012 vs. 2021 (VKA, DOAC, OTHER).

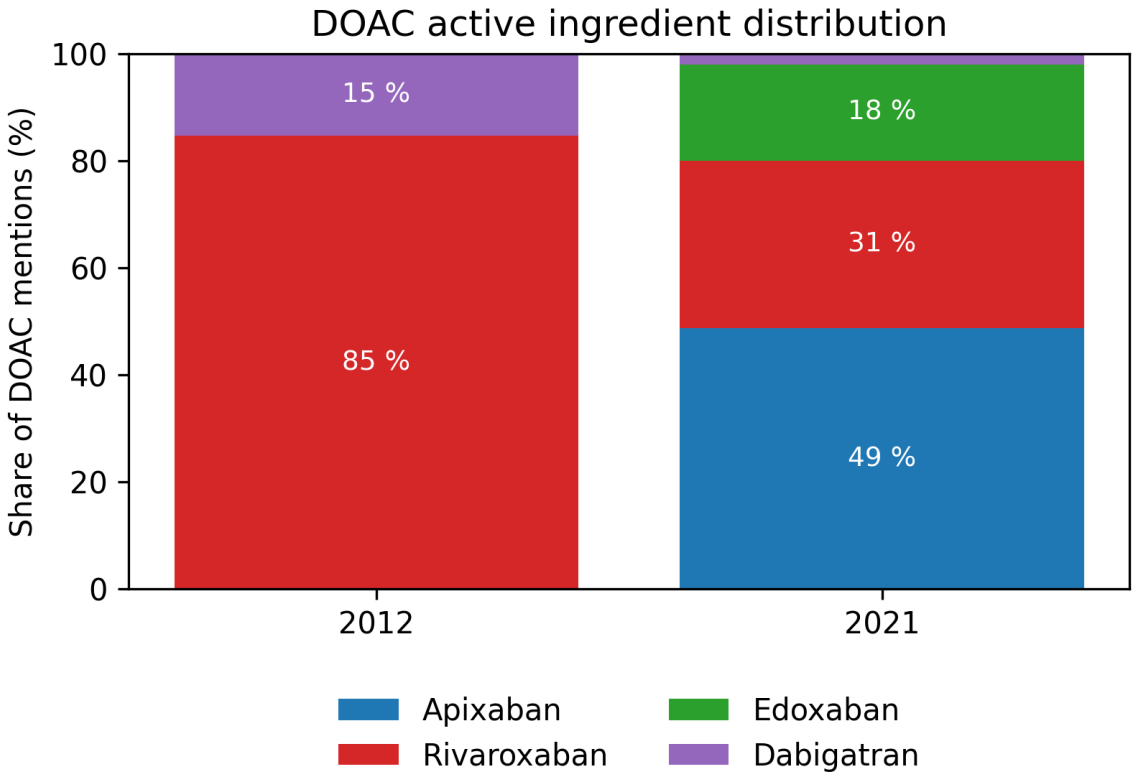


Fig. 7.3 **DOAC active ingredient distribution:** Composition of DOAC active ingredients comparing 2012 vs. 2020 – 21 (CARDIO:DE).

to 31%, while only 18% was linked to *Edoxaban* and 5% *Dabigatran*. This correspondents to large-scale pharmacological surveys, where *Apixaban* surpassed *Rivaroxaban* by 2018 (Wheelock et al. 2021).

We manually reviewed a random set of 417 letters. Out of a total of 55,664 paragraphs, in 86 the model extracted any medication (`medication` value != "") and the predicted JSON contained the string *antikoagulation*. On this predicted set the paragraph-level positive predictive value (PPV) regarding classification into DOAC/VKA was $53/86 = 61.6\%$, where $46/86 = 53.5\%$ were medications currently prescribed to the patient. We then excluded unclassifiable samples with generic medication values (e.g. `medication` = *Antikoagulation*) $28/86$, thus the final classifiable data set contained 58 paragraphs. Hence, the performance of the OAC classification achieved a PPV of $53/58 = 91.4\%$ where $46/58 = 79.3\%$ were currently prescribed or resumed. Most issues among the 86 paragraphs were related to non-classifiable generic medication names (32.6%), and missing `reason` = *Antikoagulation* relations while the correct medication was extracted (11.6%). We identified 9 paragraphs with false negative predictions. In summary, when a medication is present, medication

extraction and DOAC/VKA assignment is highly reliable, while most issues stem from generic medication mentions or missing relation links (further details and examples cf. Appendix D.1).

Polypharmacy Three in four letters exceed the threshold 5 and almost more than 40% exceed medication threshold 10 (cf. Table 7.2). At patient level, all patient had a least one doctor's letter with > 5 and 80% surpassed > 10 medications.[2]

Threshold	per Letter	per Patient
> 5 drugs	348 (75.3%)	20 (100.0%)
> 10 drugs	202 (43.7%)	16 (80.0%)

Table 7.2 **Results polypharmacy:** (first column) Count of doctor's letters containing polypharmacy and (second column) polypharmacy on patient level for two thresholds. Wilson 95% CIs reported in footnote 2.

Figure 7.4 visualizes the development of polypharmacy per patient over time. Once patients crossed the > 5 threshold, medication counts rarely decreased afterwards. Several patients remained ≥ 10 for several years.

We manually reviewed selected edge-cases (cf. Figure 7.4) to investigate performance of our setup.

- For patient 0001664709 we observed a sudden drop in medication from 16 to 2 medications in a letter end of 2013. This occurred as the medication section contained only a reference to an external document (e.g. *siehe I-Kurve*).
- For patient 0000979592 we detected an outlier right for the first letter in 2010. We found, that this letter contained a flattened table with two initial columns (*Wirkstoff, Fertigarzneimittel*), hence each medication was counted twice. This could be handled by adding a normalization step to map brand names to active ingredients. The following letter in end of 2011 did not contain a medication section at all. Hence there was a severe drop in medication count. This occurred in beginning of 2012 again.
- For patient 0000739028 we again identified a flattened table issue, which doubled the actual medication count in the beginning of 2009. The high amount of letters over a

[2]Letter level: $348/462$ exceeded > 5 medication mentions (75.3%, 95% CI: $71.2 - 79.0\%$) and $202/462$ exceed > 10 medication mentions (43.7%, 95% CI: $39.3 - 48.3\%$),
patient-level: $20/20$ exceeded > 5 medication mentions (100.0%, 95% CI: $83.9 - 100.0\%$) and $16/20$ exceed > 10 medication mentions (80.0%, 95% CI: $58.4 - 91.9\%$).

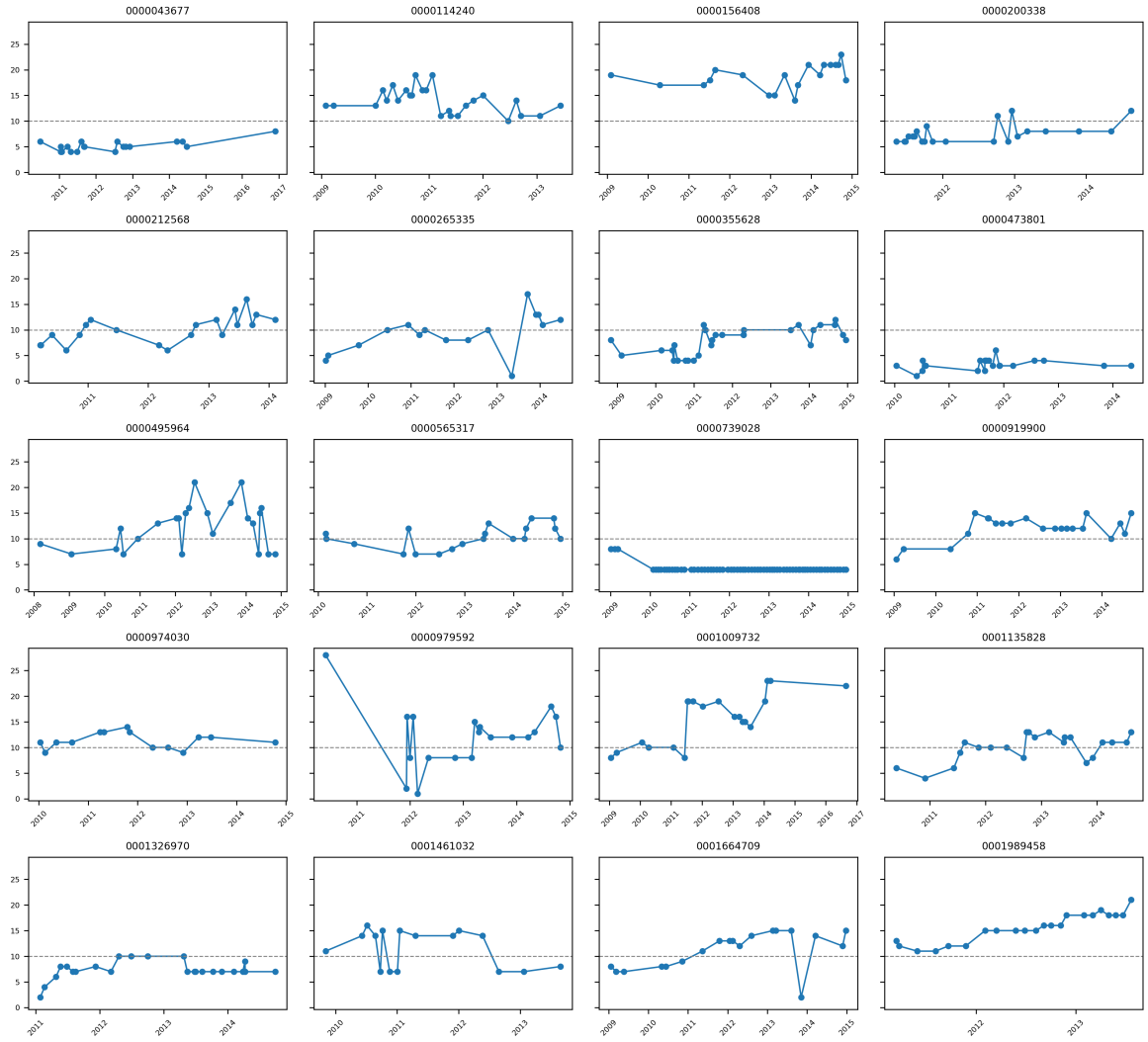


Fig. 7.4 **Polypharmacy per patient:** Number of unique medications per letter and per patient. The numbers above each plot represent a dummy patient ID. Y-axis shows the number of unique medication mentions. The x-axis represents the observation period by year. Blue dots represent a single letter and the amount of medications. The dashed line marks the 10 medications threshold.

time period of six years is recognized correctly and rooted in a regular *immunoadsorption therapy*.

- For patient 0000265335 we observed a drop in medication in 2013, due to a short report about a cardiac device. However, it follows a steep increase to almost 18 medications, due to a normalization issue: in the admission medication list, all medications are documented as active ingredients, while in the discharge section all medications are brand names. Another issue that a normalization step could resolve.

Most artifacts reflect document formatting rather than model performance issues that motivate further normalization steps. Overall, the data indicate a persistent polypharmacy issue within our cardiology cohort. Additionally, the findings suggest that polypharmacy is often a long-term occurrence. This emphasizes the importance of future clinical evaluations and studies.

7.6 Discussion

OAC shift Our e2e MIE pipeline identified a transition from VKA to DOAC between 2012 and 2021 (Figure 7.2). VKA mentions fell by 27.8 pp. while DOAC mentions rose by 43 pp. (both p-value $< 10^{-8}$). Despite corpora of different time spans our e2e extraction robustly reflected a treatment shift in European clinical guidelines.

Within DOAC medications *Rivaroxaban* was predominant in 2012, while in 2021 *Apixaban* represented approximately half of the DOAC mentions (Figure 7.3). This is supported by large-scale pharmacological studies showing the dominance of *Apixaban* among DOACs over time.

A manual review on a stratified 2012 set of classifiable paragraphs ($n = 58$) supported extraction performance (PPV 91.4% on classifiable paragraphs). Annotation issues were mostly related to generic mentions of anticoagulation (33%) and missing relation information (12%) (cf. Section D.1).

Our design choice to use the e2e pipeline with format-restricted prompts enabled a lightweight setup under clinical constraints, hence, yielded well-formatted output allowing to detect treatment shifts consistent with clinical guidelines (addressing RQ 5).

Limitations The cohort selection for 2012 is a stratified sample of a larger local corpus containing brief outpatient notes with limited medication details. Furthermore, we counted unique medication mentions in the notes, not unique prescriptions. Furthermore, the data set included generic medication mentions, e.g. *Antikoagulations-Therapie*, making a

VKA/DOAC class distinction impossible. However, the magnitude of change is robust and supported by European clinical guidelines and independent research studies.

Polypharmacy The e2e pipeline identified a high and persistent polypharmacy issue in the selected cardiology cohort. 75% of all doctor's letters contain > 5 unique medications and 44% exceed the > 10 threshold. 80% of patients reach the 10 medication threshold at least once. These results reflect previous studies that frequently identified polypharmacy risks in cardiovascular populations (Delara et al. 2022; Unlu et al. 2020). Hence, our on-premise automatic detection gives a first indication of polypharmacy in clinical routine without additional manual annotations.

Limitations Our descriptive analysis reflects unique medication mentions, not prescriptions, in a small cardiology-skewed cohort. Moreover, our cohort can be biased toward higher polypharmacy, as patients with many successive doctor's letters typically appear more frequently in the hospital, due to complex cardiovascular diseases along with more complex medication plans.

Letter date extraction relied on heterogeneous DOCX metadata (`USERAUF`/`USERENT`), necessitating rare manual supervision. In addition, as the context-enriched PLMs for section classification proved insufficient to distinguish admission and discharge paragraphs (cf. Chapter 5.4.1) we applied heuristics as described in Section 7.4. However, manual evaluation still indicated formatting issues (flattened tables, external document references) leading to moderate under- or over-estimation of unique medication counts per letter. Furthermore, since our pipeline does not include a normalization step, when medications are documented with both brand name and active ingredient in the same letter, our setup overestimates medication counts.

However, combined with additional normalization steps, our pipeline showed robust performance over time, requiring only minimal manual review. This makes the pipeline a suitable on-premise tool for clinical decision support systems by identifying polypharmacy issues in its early stages.

7.7 Summary

Our on-premise, format-restricted e2e pipeline for medication information extraction enabled identification of guideline-driven shifts from VKAs to DOACs over time directly from unstructured doctor's letters. The proportion of DOACs increased significantly from 16.9% to 59.9%, while VKAs decreased from 37.7% to 9.9%. Furthermore, within DOACs active

ingredients also shifted towards *Apixaban* dominating in 2021. Manual evaluations further supported these findings. Our method supplements hypothesis driven medical research with robust data-driven methods. Hence, these findings demonstrate the ability of our MIE pipeline to generalize on unseen clinical routine doctor's letters and to identify guideline-driven medication treatment shifts directly contributing to RQ 5.

In the longitudinal patient cohort, the same pipeline quantified polypharmacy directly from doctor's letters. 75% of doctor's letters contained more than 5 medications and 80% ever exceed 10 unique medication mentions. This further demonstrates generalization even on older doctor's letters, directly addressing RQ 5. The manual review identified outliers in medication counting due to document-formatting issues over time rather than performance issues with the e2e pipeline. Adding a normalization step would significantly prevent most of these issues.

To our knowledge, this is the first on-premise NLP application that automatically parses German doctor's letters to detect guideline-driven OAC treatment shifts and longitudinal polypharmacy under real-world clinical constraints.

As future work, we plan to extend patient cohorts and integrate interpretability methods to further investigate model behavior (cf. Chapter 6). Moreover, we plan to extend the e2e pipeline by a normalization step and the PET-based section classification model (cf. Chapter 5) to automatically distinguish admission and discharge paragraphs, by leveraging longer context capabilities of generative LLMs (cf. 5.4.1) (Beltagy et al. 2020; Li et al. 2022b; Li et al. 2023).

Chapter 8

Discussion and Conclusion

8.1 Discussion

In this thesis, we investigated the state-of-the-art NLP methods for representative clinical tasks under strict on-premise (cf. challenges: 1.a. *Domain expertise*, 1.b. *Staff time*, 1.c. *Compute restrictions*, 1.d. *Native languages*) and transparency constraints (cf. 2. *Transparency*). We followed the methodological evolution from supervised encoders through prompt-based encoders (PET) to fine-tuned generative LLMs and investigated their implications for each of these constraints. In particular, we leveraged (i) few-shot capabilities of compact encoders via PET for paragraph-level section classification and (ii) LLMs with format-restricting prompts, parameter-efficient fine-tuning (LoRA/QLoRA), deterministic decoding and feedback LLMs for a complex NER+RE e2e medication information extraction task. We conducted both tasks completely on-premise within the clinical infrastructure primarily on the German clinical routine corpus CARDIO:DE, using two well-curated gold-standard annotation layers (CDA-compliant section types, medication information), which enabled native-language evaluation and reproducible research. Additionally, we evaluated our best-performing medication extraction model using unseen German data for two real-world clinical tasks.

My time of working in the field of clinical information extraction coincided with the massive shift in computational linguistics methodology (Hahn et al. 2020; Liu et al. 2023; Singhal et al. 2023). With the rise and spread of deep learning, the complexity and performance of the models increased heavily, at the cost of transparency and increasing compute requirements (Hahn et al. 2020). In the beginning of my work in 2017, I lacked sufficient computing resources, NLP tools, and annotated clinical data. Meanwhile, the computing conditions improved and I could conduct initial MIE annotation and ML experiments on internal data. In 2023 we published the first freely available coherent German clinical corpus,

the foundation for the experiments and research presented in this thesis. This work, along with the work of the German NLP research community, has established foundations and novel benchmarks in German medical information extraction research. However, due to the scarcity of previously published results and data in German clinical NLP, direct comparability with other studies is limited.

All of our results were discussed separately in the corresponding chapters. Below, we discuss the results in context of our research questions, focusing on why certain design choices worked and identifying their limitations, and what this implies under our clinical constraints.

RQ1: On-premise model adaptation under clinical constraints Our general goal was to test the feasibility of the use of evolving model architectures within the clinical firewall.
All our experiments were conducted entirely on-premise leveraging two annotation layers of CARDIO:DE. Zero-shot capabilities of PLMs and LLMs on German clinical texts were still unreliable. Hence, we further-pretrained and fine-tuned compact German encoders up to 340 million parameters on middle-class GPU resources (max. 2x NVIDIA RTX6000). To benefit from prompting strategies for generative LLMs, we exploited recently developed PEFT methods including 4-bit quantization in combination with robust format-restricting prompts (8b: 1x NVIDIA RTX6000; 70b: 1x NVIDIA H100).
This addresses the on-premise constraints by (i) leveraging PLMs/LLMs for domain knowledge (cf. challenge 1.a. *Domain expertise*), (ii) improving performance by fine-tuning on existing gold-standard annotations (cf. challenge 1.b. *Staff time*), (iii) inside a limited compute environment (cf. challenge 1.c. *Compute restrictions*) and (iv) optimizing these models for native languages (German) (cf. challenge 1.d. *Native language*).

Our results show that model adaptation and deployment are feasible when (i) compact German encoders are further-pretrained on in-house text and fine-tuned with PET in a few-shot setup, and (ii) generative LLMs are optimized with PEFT methods in combination with format-restricted prompting. However, feasibility remains dependent on access to unlabeled clinical texts and/or at least a small set of high-quality gold-standard annotations. Zero-shot performance, particularly in German, remains unreliable. Large generative LLMs can require advanced GPU resources at training time.

RQ2: Resource-aware model choice across MIE task complexity The choice of the model architecture is crucial, particularly within the limited hospital IT infrastructure (Richter-Pechanski et al. 2024). For certain tasks, the use of extremely complex LLMs is an overkill,

highlighting the need to balance performance with resource efficiency (Schick et al. 2021b). Our findings identified a clear compute-performance trade-off influenced by the complexity of a clinical task. For section classification, task- and domain-adapted compact prompt-based German PLMs fine-tuned with few-shot PET reduced the gap to fully supervised fine-tuning. Scaling to larger encoders and adding context further improved performance especially for more complex section classes such as ANAMNESE.
However, prompt-based encoders showed limited performance on more complex token-level NLP tasks (Reynolds et al. 2021). PEFT-optimized Llama 70b achieved new SOTA results on English and surpassed German baselines. Llama 70b+feedback achieved the best overall results, especially on complex relation classes (e.g. ADE, REASON) but comes with higher computational demands.

Overall, the results indicate that native language supervised compact encoders are preferable for paragraph-level and simple token-level classification tasks if larger amounts of annotations are available. PET encoders are suitable for few-shot text classification tasks, while PEFT-optimized and language-adapted LLMs are efficient for joint token-level NER+RE tasks. Llama 8b performs strongly under on-premise constraints and Llama 70b+feedback improves performance for more complex relation classes (cf. challenge 1.c. *Compute restrictions*, 1.d. *Native language*).

RQ3: Reducing manual effort during development and evaluation Clinical NLP projects require involvement of local research staff for development and deployment and physicians for data annotation and model validation. Particularly, organizing sufficient physician resources is often a bottleneck in project and resource planning, due to the scarcity of their specialized expertise and high costs (Tamang et al. 2023).
PET uses prompt-based methods and showed SOTA results for paragraph-level section classification with further-pretrained compact encoder-based PLMs while substantially reducing manual annotation and engineering efforts via few-shot prompting and PETAL's automatic verbalizer search.
For token-level NER+RE, fine-tuned generative LLMs simplified development by reliably producing schema-constrained outputs with limited gold data, reducing the need for task-specific NLP pipelines. We used format-restricting prompts to yield schema-constrained JSON, minimizing manual post-processing. Feedback LLMs accelerated the often complex evaluation of LLM outputs by detecting semantically similar mismatches (e.g., *IV and PO* vs. [*IV*,*PO*]).

While PLMs/LLMs encode clinical knowledge, zero-shot performance remained unreliable for our tasks. However, we could minimize annotation and evaluation efforts while reducing the gap to fully supervised SOTA methods using further-pretraining, few-shot learning, optimized prompting strategies, and feedback LLMs (cf. challenge 1.a. *Domain expertise* and 1.b. *Staff time*).

RQ4: Transparency, interpretability and trust The accelerated increase in model performance came at the cost of transparency, resulting in black-box systems that are especially unacceptable in the clinical domain (Hanif et al. 2021).
To enable transparent and reproducible research, allowing independent verification of results, we released a prospective clinical corpus containing 500 doctor's letters from the cardiology domain compliant with data protection regulations.
We compared different attribution methods (FERRET) and, consistent with faithfulness criteria, adopted Shapley-based token attributions as most reliable for our setup (Jacovi et al. 2020; Attanasio et al. 2023). We leveraged computationally optimized implementations (SHAP) and used CAPTUM to calculate Shapley values for generative LLMs (Miglani et al. 2023).
For PET encoders, Shapley value analysis guided training data sampling in few-shot setups and revealed tokens that contribute to misclassification. In our e2e MIE experiments, CAPTUM-based Shapley values applied to structured JSON outputs enabled further analysis of relation-extraction capabilities and implicit clinical knowledge of LLMs. Deterministic decoding and format-restricting prompts further stabilized outputs, supporting trust in generative LLMs (cf. 2. *Transparency*).

Overall, our interpretability methods uncovered model reasoning processes, supported error analysis, and guided data sampling in few-shot setups. However, measuring the quality of faithfulness remains difficult and attribution calculations produce significant computational costs, highlighting that these methods support human judgment rather than replacing it.

RQ5: Clinical applicability on German data The final validation of any clinical NLP system lies not in automatic metrics, but in demonstrating robust and generalizable performance when applied to unseen, real-world clinical data (Elangovan et al. 2024). Hence, we evaluated generalization and applicability of our models in real-world downstream tasks on unseen German doctor's letters ($2008 - 2021$) under strict on-premise constraints.
Our best e2e MIE model recovered the expected OAC treatment shift (VKA→DOAC): VKA medications fell from 37.7% to 9.9%, while DOACs rose from 16.9% to 59.9% (PPV for

VKA/DOAC assignment in classifiable instances 91.4%).
For polypharmacy, we identified that 75.3% of the letters exceeded > 5 and 43.7% > 10 unique medication mentions. Once a patient exceeded these thresholds, they rarely fell below that level again. Most model errors were caused by generic medication mentions, missing relation links, inconsistent letter structures (flattened tables, admission/discharge ambiguities) or non-standardized naming.

These results, obtained fully on-premise on unseen native-language letters, show robust performance of our MIE system on retrospective analysis of treatment trends and polypharmacy. However, a normalization step for medication names and improved section classification by adding a medication-section disambiguation step could minimize most errors and increase reliability.

8.2 Conclusion

In this thesis, we demonstrate that clinical NLP applications with state-of-the-art performance including interpretability features can be built and deployed entirely on-premise. We assessed these applications in the context of the evolution of NLP methods, from supervised encoders, through prompt-based encoders supported by PET and finally to PEFT-optimized generative LLMs. All experiments were evaluated on local clinical corpora, particularly German CARDIO:DE. We studied two distinct task complexities, paragraph-level section classification and token-level medication information (NER+RE), to show where compact encoders are sufficient and where larger generative LLMs are necessary. We consistently followed the defined clinical constraints, on-premise and transparency, throughout all experiments. This subchapter is structured in two parts: (i) conclusions aligned with our research questions (cf. Section 1.2), and (ii) practical guidelines for local clinical NLP projects aligned with our clinical constraints defined in Section 1.3.

8.2.1 Research questions

We address the need of on-premise model adaptation (**RQ 1**), by demonstrating the feasibility of optimizing (i) compact encoders using further-pretraining and prompt-based PET and (ii) generative LLMs using PEFT (LoRA/QLoRA) optimization and inference fully on-premise on local hospital infrastructure (middle-class GPUs) and German clinical corpora, both local and distributable (CARDIO:DE). To balance performance and resource efficiency

(**RQ 2**), we present a resource-aware guideline across MIE task complexities demonstrating that domain- and task-adapted German prompt-based encoders delivered the best compute-performance trade-off for paragraph-level section classification. For token-level NER+RE tasks, PEFT-optimized generative LLMs were necessary to reach SOTA. Addressing **RQ 3**, we reduced the demand for domain expertise and staff time, sharing well-curated annotation guidelines together with the first distributable annotated German clinical corpus. Moreover, we minimized manual efforts by leveraging (i) PET in few-shot learning settings supported by automatic prompt optimization (PETAL) and null prompts, achieving a new benchmark for a German section classification task and (ii) by exploiting intrinsic clinical knowledge of LLMs combined with PEFT-optimization, format-restricting prompts and feedback LLMs for a complex token-level task, achieving a new SOTA on English data and a new benchmark on German data. We address **RQ 4** by distributing a German clinical corpus with two well-curated annotation layers to support reproducible and transparent research. We leveraged Shapley values to explain model reasoning, support training data sampling and in-depth error analysis (identifying input token contributions, relation extraction capabilities, intrinsic clinical knowledge). Additionally, we successfully applied deterministic decoding and format-restricting prompts to reliably guide LLMs to generate structured output. We leverage our e2e MIE pipeline to validate it on unseen German clinical data (**RQ 5**). To our knowledge, we are the first to demonstrate recovery of guideline-driven changes in oral anti-coagulation treatment and the indication of polypharmacy in retrospective data ($2008-2021$) on German clinical data. While both experiments serve as a proof-of-concept, they offer initial evidence supporting clinical and research applicability of our proposed methods.

8.2.2 Practical guidelines under clinical constraints

Subsequently, we translate our results into practical guidelines. For each clinical constraint, we propose best practices and highlight remaining limitations for on-premise clinical NLP projects.

Domain expertise To reduce the demand for manual annotation and feature engineering in paragraph-level (e.g. section classification) tasks, leverage in-house, unlabeled native-language clinical texts for further-pretraining and fine-tune with PET on well-curated gold-standard labels (e.g. CARDIO:DE). If further-pretraining is not feasible, we recommend choosing clinical PLMs like `medbertde` over general PLMs. For complex NER+RE tasks (e.g. medication information extraction) use strong foundational generative LLMs PEFT fine-tuned with format-restricting prompts on compact well-curated gold-standard annotations.

Limitations: Zero-shot capabilities of both PLMs and LLMs remain unreliable, especially in German. Furthermore, real-world clinical experiments (cf. Chapter 7) showed that additional clinical supervision is still required, e.g. to address medication name normalization or data heterogeneity issues.

Staff time To enable clinical experts to train and evaluate local models quickly, we distributed two well-curated annotation layers with CARDIO:DE. Use PET and PETAL in combination with contextualization to reduce manual annotation and engineering efforts for section classification. If larger few-shot sets are available, apply null prompts to minimize prompt engineering efforts. Fine-tune LLMs with format-restricting prompts in a lightweight e2e NER+RE workflow in combination with feedback LLMs to reduce manual development and evaluation effort.
Limitations: For more complex section classes, such as ANAMNESE, few-shot results still lag behind those from training on the full dataset, indicating the need for additional data or manual supervision. Furthermore, our feedback LLM is exclusively integrated as a post-hoc validator, therefore, manual evaluation remains necessary.

Compute restrictions For section classification and medication information extraction use smaller models (`bert-base`, Llama 8b) to achieve SOTA results for most classes. For more complex classes prefer larger models (`bert-large`, Llama FT 70b).
Limitations: In particular, Llama FT 70b required a NVIDIA H100 for PEFT optimization, which may exceed most hospital compute capabilities. However, this applies only during training; model inference demands significantly fewer compute resources.

Native language We support German clinical NLP research by distributing the first German clinical routine corpus. For section classification further-pretrain and fine-tune German PLMs on German clinical texts. If further-pretraining is not feasible, prefer German clinical PLMs. For NER+RE tasks apply strong general-domain LLMs, PEFT-optimized with format-restricting prompts on high-quality German gold data.
Limitations: However, we found that German medication extraction quality is still behind that of English, likely reflecting strong English pretraining bias of our selected LLMs.

Transparency We distributed CARDIO:DE, which contains coherent and realistic doctor's letters from clinical routine, supporting reproducible and transparent research. Use Shapley-based attributions to support model transparency, error analysis, and training data sampling in few-shot scenarios with encoders. Apply Shapley values with CAPTUM to gener-

ative models to investigate relation extraction capabilities and intrinsic clinical knowledge of LLMs. In combination with format-restricting prompts, they improve the transparency of model predictions in a clinical MIE setup.
Limitations: However, Shapley-based attributions, even with SHAP optimization, are computationally expensive, especially for generative tasks. Furthermore, it remains challenging to evaluate how accurately Shapley values represent the internal workings of a model (Jacovi et al. 2020).

Overall, under the clinical constraints of on-premise and transparency (cf. 1. *On-premise* and 2. *Transparency*) defined in Section 2.1 this thesis addresses our research questions and provides a process-oriented guideline for lower-resource languages and domains for MIE projects following the evolution of NLP methods over time.

8.3 Future work

Although we achieved substantial progress towards robust, transparent, on-premise MIE, we are not ready yet for a seamless deployment in clinical routine. Future work could target current key limitations identified in this thesis: (i) unreliable zero-shot performance, (ii) medication name normalization, (iii) data heterogeneity, (iv) few-shot performance still behind full-fledged training, (v) feedback LLM used only as post-hoc module, (vi) Llama 70b models difficult to train on hospital compute resources, (vii) German MIE performance still behind English, (viii) computationally expensive SHAP explanations and (ix) difficulty to measure quality/faithfulness of interpretability methods. We propose selected future work aligned with state-of-the-art research to extend our on-premise, transparent setup, while staying applicable under clinical constraints.

We plan to scale CARDIO:DE into CARDIO:DE++ by (i) incorporating data from a second hospital site, to address site-specific biases and support the development of more generalizable models across diverse clinical environments, (ii) adding temporal information by collecting at least two distinct letters per patient, and (iii) expanding the corpus into a multimodal corpus, integrating paired data from the same patients covering three clinical routine modalities: doctor's letters, structured tabular patient data, and biosignal data. To ensure an efficient project start, we aim to adapt and harmonize the existing de-identification, annotation, and access protocols established for CARDIO:DE across our partner sites. (addresses **(iii) and (vii) above**)

To further optimize native-language capabilities of PLMs/LLMs, we recommend investigating continued further-pretraining and fine-tuning on German medical texts (e.g. carefully

curated translated English clinical data)(Gururangan et al. 2020; Idrissi-Yaghir et al. 2024; Nag et al. 2025) and consider tokenizer and vocabulary adaptations (Pfister et al. 2025; Ahia et al. 2023). (**(i) and (vii)**)

Future research is needed to further reduce the need for manual domain expertise and staff time and increase the amount of annotated data, by using fine-tuned MIE models in an active learning setup to focus domain experts on the most relevant samples during manual annotation (Şapcı et al. 2024). To support model evaluation, we recommend integrating a feedback LLM into the training process (not just post-hoc) to enable e2e optimization during fine-tuning and reduce manual evaluation effort (Gu et al. 2024). (**(i), (iv) and (v)**)

To overcome computational restrictions (Llama 70b deployment, Shapley value computation), further research is needed on distillation of PLMs/LLMs into smaller expert models (Rohanian et al. 2024; Obamuyide et al. 2022) and on quantization and memory optimization approaches for model training and inference (Jin et al. 2024). To further reduce compute demands of SHAP, we recommend assessing hierarchical attribution approaches (Paes et al. 2025) and evaluating runtime quantization (e.g., 4-bit) and memory optimizations (e.g., *PagedAttention, FlashAttention-2*) on middle-class GPUs. (**(vi) and (viii)**)

To support interpretability and error analysis we recommend integrating token-level Shapley-based attributions into a domain-expert evaluation interface (Kayser et al. 2024) and comparing Shapley values with alternative interpretability methods (Zhao et al. 2024). In addition, investigations are needed on the combination of interpretability with uncertainty estimates and consistency checks (Chen et al. 2024a; Harsha Tanneru Chirag Agarwal Himabindu Lakkaraju 2024; Liu et al. 2024b). (**(viii)**)

Based on our clinical application experiments (cf. Chapter 7), we recommend addressing medication name variability and data heterogeneity by integrating medical ontologies (e.g. SNOMED CT, ICD-10) via retrieval-augmented generation (RAG) techniques to support data normalization (Berkowitz et al. 2025). Along with data normalization, we recommend to integrate a section name disambiguation step (e.g. to distinguish admission and discharge sections) using long-context PLMs (Beltagy et al. 2020; Li et al. 2022b; Li et al. 2023). To optimize development and evaluation, we suggest involving physicians for systematic validation of treatment shift and polypharmacy results, including metrics such as factuality, comprehension, reasoning, potential harm and bias (Singhal et al. 2023). Ultimately, we recommend to evaluate a combined pipeline of section classification and medication information extraction and to integrate it into a visual advisory dashboard for advanced clinical decision support and research under clinical constraints. (**(ii) and (iii)**)

The generative abilities of LLMs have considerably broadened the spectrum of applications for NLP, particularly their conversational capabilities allow for more intuitive question

answering interactions between physicians and LLMs in the form of a dialogue. We propose assessing the applicability and quality of a combination of MIE methods and QA-based LLMs to support daily clinical routine and research. Specifically, two representative clinical tasks are regularly recurring in cardiology and require significant manual effort: (i) calculating cardiovascular disease risk scores from unstructured clinical patient variables stored in doctor's letters, and (ii) deriving therapy recommendations based on these scores. This scenario could be solved by combining MIE and question-answering LLMs in addition with RAG methods, to integrate current clinical guidelines and drug-related risks databases.

Despite remaining challenges and the need for further research, this thesis highlights the robust empirical performance of on-premise information extraction models. Along increasing evidence that NLP models can substantially improve clinical routine workflows and medical research, it emphasizes the realistic potential for the application of these methods into robust decision support systems in clinical routine (Artsi et al. 2025).

List of figures

List of tables

List of abbreviations

3S three step approach 29, 30, 159

BERT bidirectional encoder representations from transformers 2, 5, 15–17, 23, 27, 31–34, 36, 37, 45, 65, 66, 68–71, 80, 81, 86, 89, 96, 100, 102, 116, 118, 152, 159

BRAT Brat Rapid Annotation Tool 111

BRONCO Biomedical entity Relation ONcology COrpus 40

CC cardiovascular concept 31–34, 36, 37, 151

CCE cardiovascular concept extraction 31–34, 66, 151

CDA clinical document architecture 5, 41, 62, 63, 141, 160

CLEF Clinical E-Science Framework 40

CNN convolutional networks 15, 16, 45, 117

CPU central processing unit 27, 30

CRF conditional random field 1, 15, 16, 24, 26, 27, 29–31, 33, 34, 36, 66, 68, 159

CVD cardiovascular desease 50

DOAC direct oral anticoagulants 130–136, 138, 139, 144, 145, 154, 161

DUA data usage agreement 71–73

e2e end-to-end 7, 108, 109, 111, 113, 116, 117, 119, 120, 124–127, 129–133, 138–141, 144, 146, 147, 149, 154, 161

ELMo embeddings from language model 27, 28, 151

ESC European Society of Cardiology 50

FFNN feed forward neural network 35, 151

FT fine-tuned 112, 115, 118–120, 122, 129, 131, 147, 161

GDPR general data protection regulation 26, 40, 49, 73

GGPONC German Guideline Program in Oncology NLP Corpus . 40, 66, 68, 69, 84, 160

GPT generative pre-trained transformer 15, 20, 43, 44

GPU graphics processing unit 1, 12, 21, 31, 34, 80, 105, 106, 142, 145, 149

GraSCCo Graz Synthetic Clinical Corpus 40

HIPAA health insurance portability and accountability act 26, 40

i2b2 Informatics for Integrating Biology and the Bedside 40, 45, 65

IAA inter-annotator agreement score 56–63, 65, 117, 152, 159, 160

IE information extraction 21, 33, 45, 46, 111

IOB inside–outside–beginning 15, 66

IT information technology 1, 6, 44, 49, 102, 125, 129, 142

IULIA Institut Universitari de Lingüística Aplicada 40

JSON javascript object notation 6, 17, 21, 23, 45–47, 109–111, 113, 114, 117, 118, 120–123, 125, 126, 132, 135, 143, 144, 154, 161

JSYNCC Jena SYnthetic CliNical Corpus 40

LLM large language model 3–7, 11, 15–17, 20–23, 37, 39, 41, 44–48, 50, 77, 80, 105–116, 118–120, 122, 124–128, 140–150, 154, 161

LoRA low-rank adaptation 6, 7, 21, 44, 108, 109, 112, 113, 125, 141, 145

LSTM long short-term memory network 1, 16, 27, 28, 34, 36, 151, 159

MADE Medication and Adverse Drug Events from Electronic Health Records 45

MERLOT Medical Entity and Relation LIMSI annOtated Text 40

MIE medical information extraction 1–4, 6, 7, 20, 39, 41, 43, 45, 46, 51, 66, 69, 72, 76, 98, 101, 102, 105–109, 112, 113, 115, 116, 119, 120, 124–127, 129–133, 138, 140–142, 144–146, 148–151, 154, 161

MIMIC Medical Information Mart for Intensive Care 40, 45

ML machine learning 49, 119, 141

MMLU Massive Multitask Language Understanding 44

n2c2 National NLP Clinical Challenges . . . 40, 45, 109, 111, 112, 116–119, 122, 125, 161

NE named entity 15, 16

NER named entity recognition . . 3, 5, 15–17, 20, 21, 26, 31, 33, 34, 37, 39, 45, 46, 66, 68, 105, 107–109, 116, 117, 124, 128, 141, 143, 145–147, 160

NLP natural language processing 1–6, 8, 11–17, 20, 22, 23, 25–27, 31, 39–44, 48–51, 76, 101, 103, 107, 119, 130, 140–150

NN neural network 29–31, 159

OAC oral anticoagulation 129–135, 138, 140, 144, 154, 161

PaLM Pathways Language Model 44

PEFT parameter-efficient fine-tuning 21, 44, 46, 106–108, 113, 125–127, 142, 143, 145–147

PET pattern-exploiting training . . 5, 15, 18–20, 42, 43, 75–80, 85–90, 92, 93, 96, 98–103, 140–147, 151, 152, 160

PETAL pattern-exploiting training with automatic labels 75–77, 80, 99, 100, 143, 146, 147

PHI protected health information 24–27, 29, 30, 40, 55, 83, 84

PLM pre-trained language model 11, 17–19, 31, 33, 37, 41–44, 75–77, 80–83, 86, 88, 93, 98, 100–103, 139, 142–144, 146–149, 151–153

pp. percentage points 30, 36, 42, 44, 69, 90, 92–94, 99, 102, 116, 118, 119, 133, 138

QA question answering 20, 21, 44, 45, 150

QLoRA quantized low-rank adapters 21, 44, 112, 113, 125, 141, 145

RAFT Real-world Annotated Few-shot Tasks 42

RAG retrieval-augmented generation 149, 150

RE relation extraction 3, 5, 16, 17, 20, 21, 37, 39, 45, 46, 105, 107–109, 116, 117, 124, 128, 141, 143, 145–147

RNN recurrent neural networks 27, 32, 45, 117

RQ research question . . 3, 7, 20, 22, 31, 33, 37, 41, 44, 46–50, 72, 73, 75, 76, 81, 98–101, 106, 122, 124–126, 129, 138, 140, 142–146

SC sequence classifier 86–90, 92, 93, 96, 98–100, 102, 152, 160

SHAP SHapley Additive exPlanations 23, 48, 81, 99, 116, 144, 148, 149

SNOMED CT Systematized Nomenclature of Medicine 66, 68, 149

SOTA state of the art 16, 20, 22, 24, 26, 27, 31, 39, 41, 43–46, 49, 50, 69, 77, 105, 107, 108, 112, 113, 116, 118–120, 122, 124–127, 143, 144, 146, 147, 161

SVM support vector machine 16, 69–71

THYME Temporal Histories of Your Medical Events 40

UIE unified information extraction 21

UMLS Unified Medical Language System 16

VKA vitamin K antagonist 130–136, 138, 139, 144, 145, 154, 161

VRAM video random-access memory 34, 125, 127

XML eXtensible Markup Language 111

References

Aebersold, H., F. Foster-Witassek, S. Aeschbacher, J. H. Beer, E. Blozik, M. Blum, L. Bonati, G. Conte, M. Coslovsky, M. L. De Perna, M. D. Valentino, S. Felder, C. A. Huber, G. Moschovitis, A. Mueller, R. E. Paladini, T. Reichlin, N. Rodondi, A. Stauber, C. Sticherling, T. D. Szucs, D. Conen, M. Kuhne, S. Osswald, M. Schwenkglenks, and M. Serra-Burriel (Jan. 2024). "Patients on vitamin K treatment: is switching to direct-acting oral anticoagulation cost-effective? A target trial on a prospective cohort". In: *Open heart* 11.1. ISSN: 2053-3624. DOI: 10.1136/OPENHRT-2023-002567. URL: https://pubmed.ncbi.nlm.nih.gov/38302139/.

Afshar, M., S. Adelaine, F. Resnik, M. P. Mundt, J. Long, M. Leaf, T. Ampian, G. J. Wills, B. Schnapp, M. Chao, R. Brown, C. Joyce, B. Sharma, D. Dligach, E. S. Burnside, J. Mahoney, M. M. Churpek, B. W. Patterson, and F. Liao (Apr. 2023). "Deployment of Real-time Natural Language Processing and Deep Learning Clinical Decision Support in the Electronic Health Record: Pipeline Implementation for an Opioid Misuse Screener in Hospitalized Adults". In: *JMIR Med Inform* 11, e44977. ISSN: 2291-9694. DOI: 10.2196/44977. URL: http://www.ncbi.nlm.nih.gov/pubmed/37079367.

Ahia, O., S. Kumar, H. Gonen, J. Kasai, D. R. Mortensen, N. A. Smith, and Y. Tsvetkov (2023). "Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models". In: *EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings*. Association for Computational Linguistics (ACL), pp. 9904–9923. ISBN: 9798891760608. DOI: 10.18653/V1/2023.EMNLP-MAIN.614. URL: https://aclanthology.org/2023.emnlp-main.614/.

Akbik, A., T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf (2019). "FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP". In: *Proceedings of the 2019 Conference of the North*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 54–59. DOI: 10.18653/V1/N19-4010. URL: https://aclanthology.org/N19-4010/.

El-allaly, E. d., M. Sarrouti, N. En-Nahnahi, and S. Ouatik El Alaoui (May 2021). "MTTLADE: A multi-task transfer learning-based method for adverse drug events extraction". In: *Information Processing & Management* 58.3, p. 102473. ISSN: 0306-4573. DOI: 10.1016/J.IPM.2020.102473.

Alsentzer, E., J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. B. A. Mcdermott (July 2019). "Publicly Available Clinical BERT Embeddings". In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics (ACL), pp. 72–78. DOI: 10.18653/V1/W19-1909. URL: https://aclanthology.org/W19-1909/.

Ankit Pal and M. Sankarasubbu (2024). *aaditya/Llama3-OpenBioLLM-70B · Hugging Face*. URL: https://huggingface.co/aaditya/Llama3-OpenBioLLM-70B#.

Artsi, Y., V. Sorin, B. S. Glicksberg, P. Korfiatis, G. N. Nadkarni, and E. Klang (Sept. 2025). "Large language models in real-world clinical workflows: a systematic review of

applications and implementation". In: *Frontiers in Digital Health* 7, p. 1659134. ISSN: 2673253X. DOI: 10.3389/FDGTH.2025.1659134/BIBTEX. URL: https://www.crd.york.ac.uk/PROSPERO/.

Attanasio, G., E. Pastor, C. Di Bonaventura, and D. Nozza (2023). "ferret: a Framework for Benchmarking Explainers on Transformers". In: *EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of System Demonstrations*. Association for Computational Linguistics (ACL), pp. 256–266. ISBN: 9781959429456. DOI: 10.18653/V1/2023.EACL-DEMO.29. URL: https://aclanthology.org/2023.eacl-demo.29/.

Bashyam, V. and R. K. Taira (2005). "Indexing Anatomical Phrases in Neuro-Radiology Reports to the UMLS 2005AA". In: *AMIA Annual Symposium Proceedings*. Vol. 2005, p. 26. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC1560562/.

Bean, D. M., J. Teo, H. Wu, R. Oliveira, R. Patel, R. Bendayan, A. M. Shah, R. J. Dobson, and P. A. Scott (Nov. 2019). "Semantic computational analysis of anticoagulation use in atrial fibrillation from real world data". In: *PloS one* 14.11. ISSN: 1932-6203. DOI: 10.1371/JOURNAL.PONE.0225625. URL: https://pubmed.ncbi.nlm.nih.gov/31765395/.

Becker, M., M. Krumscheid, A. Knobelspies, M. Seydel, P. Richter-Pechanski, and A. Karl (Nov. 2025). "Extending CARDIO:DE: Additional annotation guidelines and evaluation of NLP approaches for clinical applications". In: *International Journal of Medical Informatics* 203, p. 106009. ISSN: 1386-5056. DOI: 10.1016/J.IJMEDINF.2025.106009. URL: https://www.sciencedirect.com/science/article/pii/S1386505625002266#ab005.

Beltagy, I., K. Lo, and A. Cohan (2019). "SciBERT: A Pretrained Language Model for Scientific Text". In: *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. Association for Computational Linguistics, pp. 3615–3620. ISBN: 9781950737901. DOI: 10.18653/V1/D19-1371. URL: https://aclanthology.org/D19-1371/.

Beltagy, I., M. E. Peters, and A. Cohan (Apr. 2020). "Longformer: The Long-Document Transformer". In: *arxiv preprint*. URL: https://arxiv.org/pdf/2004.05150.

Berkowitz, J. S., A. Srinivasan, J. M. Acitores Cortina, Y. Fatapour, and N. P. Tatonetti (July 2025). "Biomedical text normalization through generative modeling". In: *Journal of Biomedical Informatics* 167, p. 104850. ISSN: 1532-0464. DOI: 10.1016/J.JBI.2025.104850. URL: https://www.sciencedirect.com/science/article/pii/S1532046425000796.

Bodenreider, O. (Jan. 2004). "The Unified Medical Language System (UMLS): Integrating biomedical terminology". In: *Nucleic Acids Research* 32.DATABASE ISS. ISSN: 03051048. DOI: 10.1093/nar/gkh061.

Borchert, F., C. Lohr, L. Modersohn, J. Witt, T. Langer, M. Follmann, M. Gietzelt, B. Arnrich, U. Hahn, and M.-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*, pp. 3650–3660. URL: https://aclanthology.org/2022.lrec-1.389/.

Bose, P., S. Srinivasan, W. C. Sleeman, J. Palta, R. Kapoor, and P. Ghosh (Sept. 2021). "A Survey on Recent Named Entity Recognition and Relationship Extraction Techniques on Clinical Texts". In: *Applied Sciences 2021, Vol. 11, Page 8319* 11.18, p. 8319. ISSN: 2076-3417. DOI: 10.3390/APP11188319. URL: https://www.mdpi.com/2076-3417/11/18/8319/htm%20https://www.mdpi.com/2076-3417/11/18/8319.

Bressem, K. K., J. M. Papaioannou, P. Grundmann, F. Borchert, L. C. Adams, L. Liu, F. Busch, L. Xu, J. P. Loyen, S. M. Niehues, M. Augustin, L. Grosser, M. R. Makowski,

H. J. Aerts, and A. Löser (Mar. 2024). "medBERT.de: A comprehensive German BERT model for the medical domain". In: *Expert Systems with Applications* 237, p. 121598. ISSN: 0957-4174. DOI: 10.1016/J.ESWA.2023.121598. URL: https://www.sciencedirect.com/science/article/abs/pii/S0957417423021000.

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. Mccandlish, A. Radford, I. Sutskever, and D. Amodei (2020). "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Vol. 33, pp. 1877–1901. URL: https://commoncrawl.org/the-data/.

Buitinck, L., G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux (2013). "{API} design for machine learning software: experiences from the scikit-learn project". In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122.

Calkins, H., A. Lin, Y. Chen, J. E. Cigarroa, J. C. Cleveland, P. T. Ellinor, M. D. Ezekowitz, M. E. Field, K. L. Furie, P. A. Heidenreich, K. T. Murray, J. B. Shea, C. M. Tracy, and C. W. Yancy (2019). "ACC/HRS Guideline for the Management of Patients With Atrial Fibrillation Circulation". In: *Circulation* 140, pp. 125–151. DOI: 10.1161/CIR.0000000000000665. URL: http://ahajournals.org.

Campillos, L., L. Deléger, C. Grouin, T. Hamon, A. L. Ligozat, and A. Névéol (June 2018). "A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT)". In: *Language Resources and Evaluation* 52.2, pp. 571–601. ISSN: 15728412. DOI: 10.1007/S10579-017-9382-Y/FIGURES/11. URL: https://link.springer.com/article/10.1007/s10579-017-9382-y.

Chan, B., S. Schweter, and T. Möller (2020). "German's Next Language Model". In: *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*. Association for Computational Linguistics (ACL), pp. 6788–6796. ISBN: 9781952148279. DOI: 10.18653/V1/2020.COLING-MAIN.598. URL: https://aclanthology.org/2020.coling-main.598/.

Chapman, W. W., P. M. Nadkarni, L. Hirschman, L. W. D'Avolio, G. K. Savova, and O. Uzuner (Sept. 2011). "Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions". In: *Journal of the American Medical Informatics Association* 18.5, pp. 540–543. ISSN: 1067-5027. DOI: 10.1136/AMIAJNL-2011-000465. URL: https://dx.doi.org/10.1136/amiajnl-2011-000465.

Che, W., Y. Liu, Y. Wang, B. Zheng, and T. Liu (2018). "Towards Better UD Parsing: Deep Contextualized Word Embeddings, Ensemble, and Treebank Concatenation". In: *CoNLL 2018 - SIGNLL Conference on Computational Natural Language Learning, Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics (ACL), pp. 55–64. ISBN: 9781948087827. DOI: 10.18653/V1/K18-2005. URL: https://aclanthology.org/K18-2005/.

Chen, J. and J. Mueller (2024a). "Quantifying Uncertainty in Answers from any Language Model and Enhancing their Trustworthiness". In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Vol. 1. Association for Computational

Linguistics (ACL), pp. 5186–5200. ISBN: 9798891760943. DOI: 10.18653/V1/2024.ACL-LONG.283. URL: https://aclanthology.org/2024.acl-long.283/.

Chen, N., Z. Zheng, N. Wu, M. Gong, D. Zhang, and J. Li (2024b). "Breaking Language Barriers in Multilingual Mathematical Reasoning: Insights and Observations". In: *EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Findings of EMNLP 2024*, pp. 7001–7016. DOI: 10.18653/V1/2024.FINDINGS-EMNLP.411. URL: https://aclanthology.org/2024.findings-emnlp.411/.

Chen, Z., A. H. Cano, A. Romanou, A. Bonnet, K. Matoba, F. Salvi, M. Pagliardini, S. Fan, A. Köpf, A. Mohtashami, A. Sallinen, A. Sakhaeirad, V. Swamy, I. Krawczuk, D. Bayazit, A. Marmet, S. Montariol, M.-A. Hartley, M. Jaggi, and A. Bosselut (Nov. 2023). "MEDITRON-70B: Scaling Medical Pretraining for Large Language Models". In: *arxiv preprint*. URL: https://arxiv.org/pdf/2311.16079.

Chiang, C. H. and H. Y. Lee (2023). "A Closer Look into Using Large Language Models for Automatic Evaluation". In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics (ACL), pp. 8928–8942. ISBN: 9798891760615. DOI: 10.18653/V1/2023.FINDINGS-EMNLP.599. URL: https://aclanthology.org/2023.findings-emnlp.599/.

Chowdhery, A., S. Narang, J. Devlin, M. Bosma, G. Mishra Adam Roberts Paul Barham Hyung Won Chung, C. Sutton Sebastian Gehrmann, P. Schuh Sasha Tsvyashchenko Joshua Maynez Abhishek Rao Parker Barnes Yi Tay, N. Shazeer, V. Prabhakaran Emily Reif Nan Du Ben Hutchinson Reiner Pope, J. Bradbury Jacob Austin Michael Isard Guy Gur-Ari, P. Yin Toju Duke Anselm Levskaya Sanjay Ghemawat Sunipa Dev Henryk Michalewski Xavier Garcia Vedant Misra Kevin Robinson Liam Fedus, D. Zhou Daphne Ippolito David Luan, H. Lim Barret Zoph, A. Spiridonov Ryan Sepassi, D. Dohan, S. M. Agrawal Mark Omernick Andrew Dai Thanumalayan Sankaranarayana Pillai Marie Pellat Aitor Lewkowycz, E. Moreira Rewon Child, O. Polozov Katherine Lee Zongwei Zhou Xuezhi Wang Brennan Saeta Mark Diaz Orhan Firat Michele Catasta, J. Wei, K. Meier-Hellstern Douglas Eck Jeff Dean Slav Petrov Noah Fiedel Google Editor, and R. Salakhutdinov (2023). "PaLM: Scaling Language Modeling with Pathways". In: *Journal of Machine Learning Research* 24, pp. 1–113. URL: http://jmlr.org/papers/v24/22-1144.html.

Chung, J. and S. Murphy (2005). "Concept-Value Pair Extraction from Semi-Structured Clinical Narrative: A Case Study Using Echocardiogram Reports". In: *AMIA Annual Symposium Proceedings* 2005, p. 131. ISSN: 15594076. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC1560613/.

Collobert, R., J. Weston, J. Com, M. Karlen, K. Kavukcuoglu, and P. Kuksa (2011). "Natural Language Processing (Almost) from Scratch". In: *Journal of Machine Learning Research* 12, pp. 2493–2537.

Cortes, C., V. Vapnik, and L. Saitta (Sept. 1995). "Support-vector networks". In: *Machine Learning 1995 20:3* 20.3, pp. 273–297. ISSN: 1573-0565. DOI: 10.1007/BF00994018. URL: https://link.springer.com/article/10.1007/BF00994018.

Dagdelen, J., A. Dunn, S. Lee, N. Walker, A. S. Rosen, G. Ceder, K. A. Persson, and A. Jain (Feb. 2024). "Structured information extraction from scientific text with large language models". In: *Nature Communications 2024 15:1* 15.1, pp. 1–14. ISSN: 2041-1723. DOI: 10.1038/s41467-024-45563-x. URL: https://www.nature.com/articles/s41467-024-45563-x.

Daniel Han, M. H. and U. Team (2023). *Unsloth*. URL: http://github.com/unslothai/unsloth.

DeepSeek-AI et al. (Dec. 2024). "DeepSeek-V3 Technical Report". In: *arXiv preprint*. URL: https://arxiv.org/pdf/2412.19437.

Delara, M., L. Murray, B. Jafari, A. Bahji, Z. Goodarzi, J. Kirkham, Z. Chowdhury, and D. P. Seitz (July 2022). "Prevalence and factors associated with polypharmacy: a systematic review and meta-analysis". In: *BMC Geriatrics 2022 22:1* 22.1, pp. 1–12. ISSN: 1471-2318. DOI: 10.1186/S12877-022-03279-X. URL: https://bmcgeriatr.biomedcentral.com/articles/10.1186/s12877-022-03279-x.

Denny, J. C., R. A. Miller, K. B. Johnson, and A. Spickard (2008). "Development and Evaluation of a Clinical Note Section Header Terminology". In: *AMIA Annual Symposium Proceedings* 2008, p. 156. ISSN: 1942597X. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC2656032/.

Dettmers, T., A. Pagnoni, A. Holtzman, and L. Zettlemoyer (2023). "QLORA: efficient finetuning of quantized LLMs". In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. NIPS '23. Red Hook, NY, USA: Curran Associates Inc.

Devlin, J., M.-W. Chang, K. Lee, K. T. Google, and A. I. Language (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North*, pp. 4171–4186. DOI: 10.18653/V1/N19-1423. URL: https://aclanthology.org/N19-1423/.

Edinger, T., D. Demner-Fushman, A. M. Cohen, S. Bedrick, and W. Hersh (2018). "Evaluation of Clinical Text Segmentation to Facilitate Cohort Retrieval". In: *AMIA Annual Symposium Proceedings*. Vol. 2017. NLM (Medline), p. 660. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC5977655/.

Elangovan, A., J. He, Y. Li, and K. Verspoor (2024). "Principles from Clinical Research for NLP Model Generalization". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024* 1, pp. 2293–2309. DOI: 10.18653/V1/2024.NAACL-LONG.127. URL: https://aclanthology.org/2024.naacl-long.127/.

Fan, F. L., J. Xiong, M. Li, and G. Wang (Nov. 2021). "On Interpretability of Artificial Neural Networks: A Survey". In: *IEEE Transactions on Radiation and Plasma Medical Sciences* 5.6, pp. 741–760. ISSN: 24697311. DOI: 10.1109/TRPMS.2021.3066428.

Faruqui, M., S. Padó, and M. Sprachverarbeitung (2010). "Training and Evaluating a German Named Entity Recognizer with Semantic Generalization". In: *Proceedings of KONVENS 2010*. Saarbrücken. URL: http://nlp.stanford.edu/software/.

Finkel, J. R., T. Grenager, and C. Manning (2005). "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling". In: *ACL-05 - 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*. Association for Computational Linguistics (ACL), pp. 363–370. ISBN: 1932432515. DOI: 10.3115/1219840.1219885. URL: https://aclanthology.org/P05-1045/.

Fornasiere, R., N. Brunello, V. Scotti, and M. J. Carman (2024). "Medical Information Extraction with Large Language Models". In: *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pp. 456–466. URL: https://aclanthology.org/2024.icnlsp-1.47/.

Gao, T., A. Fisch, and D. Chen (2021). "Making Pre-trained Language Models Better Few-shot Learners". In: *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. Vol. 1. Association for Computational

Linguistics (ACL), pp. 3816–3830. ISBN: 9781954085527. DOI: 10.18653/V1/2021.ACL-LONG.295. URL: https://aclanthology.org/2021.acl-long.295/.

Garvin, J. H., S. L. DuVall, B. R. South, B. E. Bray, D. Bolton, J. Heavirland, S. Pickard, P. Heidenreich, S. Shen, C. Weir, M. Samore, and M. K. Goldstein (Sept. 2012). "Automated extraction of ejection fraction for quality measurement using regular expressions in unstructured information management architecture(uima) for heart failure". In: *Journal of the American Medical Informatics Association* 19.5, pp. 859–866. ISSN: 10675027. DOI: 10.1136/AMIAJNL-2011-000535,. URL: https://pubmed.ncbi.nlm.nih.gov/22437073/.

Ge, Y., Y. Guo, S. Das, M. A. Al-Garadi, and A. Sarker (Aug. 2023). "Few-shot learning for medical text: A review of advances, trends, and opportunities". In: *Journal of Biomedical Informatics* 144, p. 104458. ISSN: 1532-0464. DOI: 10.1016/J.JBI.2023.104458. URL: https://www.sciencedirect.com/science/article/pii/S153204642300179X.

Geng, S., M. Josifoski, M. Peyrard, and R. West (2023). "Grammar-Constrained Decoding for Structured NLP Tasks without Finetuning". In: *EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings*. Association for Computational Linguistics (ACL), pp. 10932–10952. ISBN: 9798891760608. DOI: 10.18653/V1/2023.EMNLP-MAIN.674. URL: https://aclanthology.org/2023.emnlp-main.674/.

Ghosh, S., M. Schneider, C. Reinicke, and C. Eickhoff (June 2025). "Cohort Discovery: A Survey on LLM-Assisted Clinical Trial Recruitment". In: *arXiv preprint*. URL: https://arxiv.org/pdf/2506.15301.

Grattafiori, A. et al. (July 2024). "The Llama 3 Herd of Models". In: *arxiv preprint*. URL: https://arxiv.org/pdf/2407.21783.

Gu, J., X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, S. Wang, K. Zhang, Z. Lin, Y. Wang, L. Ni, W. Gao, and J. Guo (Nov. 2024). "A Survey on LLM-as-a-Judge". In: *arxiv preprint* 1. URL: https://arxiv.org/pdf/2411.15594.

Gurulingappa, H., A. M. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius, and L. Toldo (Oct. 2012). "Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports". In: *Journal of Biomedical Informatics* 45.5, pp. 885–892. ISSN: 1532-0464. DOI: 10.1016/J.JBI.2012.04.008. URL: https://www.sciencedirect.com/science/article/pii/S1532046412000615.

Gururangan, S., A. Marasovic, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith (2020). "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks". In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics (ACL), pp. 8342–8360. ISBN: 9781952148255. DOI: 10.18653/V1/2020.ACL-MAIN.740. URL: https://aclanthology.org/2020.acl-main.740/.

Hahn, U. and M. Oleynik (Aug. 2020). "Medical Information Extraction in the Age of Deep Learning". In: *Yearbook of Medical Informatics* 29.1, pp. 208–220. ISSN: 23640502. DOI: 10.1055/S-0040-1702001/ID/JRHAHN-34/BIB. URL: http://www.thieme-connect.com/products/ejournals/html/10.1055/s-0040-1702001%20http://www.thieme-connect.de/DOI/DOI?10.1055/s-0040-1702001.

Han, R., C. Yang, T. Peng, P. Tiwari, X. Wan, L. Liu, and B. Wang (Aug. 2024). "An Empirical Study on Information Extraction using Large Language Models". In: *arxiv preprint*. URL: https://arxiv.org/abs/2409.00369v3.

Hanif, A., X. Zhang, and S. Wood (2021). "A Survey on Explainable Artificial Intelligence Techniques and Challenges". In: *Proceedings - IEEE International Enterprise Distributed*

Object Computing Workshop, EDOCW, pp. 81–89. ISSN: 15417719. DOI: 10.1109/EDOCW52865.2021.00036.

Harsha Tanneru Chirag Agarwal Himabindu Lakkaraju, S. (2024). "Quantifying Uncertainty in Natural Language Explanations of Large Language Models". In: *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*. Vol. 238.

Hendrycks, D., C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt (Sept. 2020). "Measuring Massive Multitask Language Understanding". In: *ICLR 2021 - 9th International Conference on Learning Representations*. URL: https://arxiv.org/pdf/2009.03300.

Henry, S., K. Buchan, M. Filannino, A. Stubbs, and O. Uzuner (Jan. 2020). "2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records". In: *Journal of the American Medical Informatics Association : JAMIA* 27.1, pp. 3–12. ISSN: 1527-974X. DOI: 10.1093/JAMIA/OCZ166. URL: https://pubmed.ncbi.nlm.nih.gov/31584655/.

Hindricks, G. et al. (Feb. 2021). "2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS): The Task Force for the diagnosis and management of atrial fibrillation of the European Society of Cardiology (ESC) Developed with the special contribution of the European Heart Rhythm Association (EHRA) of the ESC". In: *European Heart Journal* 42.5, pp. 373–498. ISSN: 0195-668X. DOI: 10.1093/EURHEARTJ/EHAA612. URL: https://dx.doi.org/10.1093/eurheartj/ehaa612.

Hochreiter, S. and J. Schmidhuber (Nov. 1997). "Long Short-Term Memory". In: *Neural Computation*. Vol. 9. MIT PressPUB1010Cambridge, MA, USA, pp. 1735–1780. DOI: 10.1162/NECO.1997.9.8.1735. URL: https://dl.acm.org/doi/10.1162/neco.1997.9.8.1735.

Honnibal, M., I. Montani, M. Honnibal, H. Peters, S. V. Landeghem, M. Samsonov, J. Geovedi, J. Regan, G. Orosz, S. L. Kristiansen, P. O. McCann, D. Altinok, Roman, G. Howard, S. Bozek, E. Bot, M. Amery, W. Phatthiyaphaibun, L. U. Vogelsang, B. Böing, P. K. Tippa, jeannefukumaru, GregDubbin, V. Mazaev, R. Balakrishnan, J. D. Møllerhøj, wbwseeker, M. Burton, thomasO, and A. Patel (2019). "explosion/spaCy: v2.1.7: Improved evaluation, better language factories and bug fixes". In: *Zenodo*. DOI: 10.5281/ZENODO.3358113. URL: https://zenodo.org/records/3358113.

Hsu, E. and K. Roberts (Dec. 2025). "Leveraging large language models for knowledge-free weak supervision in clinical natural language processing". In: *Scientific Reports* 15.1, pp. 1–10. ISSN: 20452322. DOI: 10.1038/S41598-024-68168-2;SUBJMETA. URL: https://www.nature.com/articles/s41598-024-68168-2.

Hu, E., Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen (June 2021). "LoRA: Low-Rank Adaptation of Large Language Models". In: *ICLR 2022 - 10th International Conference on Learning Representations*. URL: https://arxiv.org/pdf/2106.09685.

Hu, Y., Q. Chen, J. Du, X. Peng, V. K. Keloth, X. Zuo, Y. Zhou, Z. Li, X. Jiang, Z. Lu, K. Roberts, and H. Xu (Sept. 2024a). "Improving large language models for clinical named entity recognition via prompt engineering". In: *Journal of the American Medical Informatics Association : JAMIA* 31.9, pp. 1812–1820. ISSN: 1527-974X. DOI: 10.1093/JAMIA/OCAD259. URL: https://pubmed.ncbi.nlm.nih.gov/38281112/.

Hu, Y., X. Zuo, Y. Zhou, X. Peng, J. Huang, V. K. Keloth, V. J. Zhang, R.-L. Weng, Q. Chen, X. Jiang, K. E. Roberts, and H. Xu (Nov. 2024b). "Information Extraction from Clinical

Notes: Are We Ready to Switch to Large Language Models?" In: *arxiv preprint*. URL: https://arxiv.org/pdf/2411.10020.

Huang, Q., M. Tao, C. Zhang, Z. An, C. Jiang, Z. Chen, Z. Wu, and Y. Feng (May 2023). "Lawyer LLaMA Technical Report". In: *arxiv preprint*. URL: https://arxiv.org/pdf/2305.15062.

Idrissi-Yaghir, A., A. Dada, H. Schäfer, K. Arzideh, G. Baldini, J. Trienes, M. Hasin, J. Bewersdorff, C. S. Schmidt, M. Bauer, K. E. Smith, J. Bian, Y. Wu, J. Schlötterer, T. Zesch, P. A. Horn, C. Seifert, F. Nensa, J. Kleesiek, and C. M. Friedrich (2024). "Comprehensive Study on German Language Models for Clinical and Biomedical Text Understanding". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 3654–3665. URL: https://aclanthology.org/2024.lrec-main.324/.

IV, W. F. S., S. Bethard, S. Finan, M. Palmer, S. Pradhan, P. C. d. Groen, B. Erickson, T. Miller, L. Chen, G. Savova, and J. Pustejovsky (Dec. 2014). "Temporal Annotation in the Clinical Domain". In: *Transactions of the Association for Computational Linguistics*. Vol. 2. MIT Press - Journals, pp. 143–154. DOI: 10.1162/TACL{_}A{_}00172. URL: https://aclanthology.org/Q14-1012/.

Jacovi, A. and Y. Goldberg (2020). "Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?" In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205. ISSN: 0736587X. DOI: 10.18653/V1/2020.ACL-MAIN.386. URL: https://aclanthology.org/2020.acl-main.386/.

Jagannatha, A. N., F. Liu, W. Liu, and H. Yu (Jan. 2019). "Overview of the First Natural Language Processing Challenge for Extracting Medication, Indication, and Adverse Drug Events from Electronic Health Record Notes (MADE 1.0)". In: *Drug safety* 42.1, pp. 99–111. ISSN: 1179-1942. DOI: 10.1007/S40264-018-0762-Z. URL: https://pubmed.ncbi.nlm.nih.gov/30649735/.

Jagannatha, A. N. and H. Yu (2016a). "Bidirectional RNN for Medical Event Detection in Electronic Health Records". In: *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*. Association for Computational Linguistics (ACL), pp. 473–482. ISBN: 9781941643914. DOI: 10.18653/V1/N16-1056. URL: https://aclanthology.org/N16-1056/.

— (2016b). "Structured prediction models for RNN based sequence labeling in clinical text". In: *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*. Association for Computational Linguistics (ACL), pp. 856–865. ISBN: 9781945626258. DOI: 10.18653/V1/D16-1082. URL: https://aclanthology.org/D16-1082/.

Jantscher, M., F. Gunzer, R. Kern, E. Hassler, S. Tschauner, and G. Reishofer (Dec. 2023). "Information extraction from German radiological reports for general clinical text and language understanding". In: *Scientific Reports* 13.1, pp. 1–12. ISSN: 20452322. DOI: 10.1038/S41598-023-29323-3;SUBJMETA. URL: https://www.nature.com/articles/s41598-023-29323-3.

Jiang, A. Q., A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed (Oct. 2023). "Mistral 7B". In: *arXiv preprint*. URL: https://arxiv.org/pdf/2310.06825.

Jiang, P., J. Lin, Z. Wang, J. Sun, and J. Han (2024). "GenRES: Rethinking Evaluation for Generative Relation Extraction in the Era of Large Language Models". In: *Pro-*

ceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024. Vol. 1. Association for Computational Linguistics (ACL), pp. 2820–2837. ISBN: 9798891761148. DOI: 10.18653/V1/2024.NAACL-LONG.155. URL: https://aclanthology.org/2024.naacl-long.155/.

Jin, D., E. Pan, N. Oufattole, W. H. Weng, H. Fang, and P. Szolovits (July 2021). "What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams". In: *Applied Sciences 2021, Vol. 11, Page 6421* 11.14, p. 6421. ISSN: 2076-3417. DOI: 10.3390/APP11146421. URL: https://www.mdpi.com/2076-3417/11/14/6421/htm%20https://www.mdpi.com/2076-3417/11/14/6421.

Jin, Q., B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu (2019). "PubMedQA: A Dataset for Biomedical Research Question Answering". In: *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. Association for Computational Linguistics, pp. 2567–2577. ISBN: 9781950737901. DOI: 10.18653/V1/D19-1259. URL: https://aclanthology.org/D19-1259/.

Jin, R., J. Du, W. Huang, W. Liu, J. Luan, B. Wang, and D. Xiong (2024). "A Comprehensive Evaluation of Quantization Strategies for Large Language Models". In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics (ACL), pp. 12186–12215. ISBN: 9798891760998. DOI: 10.18653/V1/2024.FINDINGS-ACL.726. URL: https://aclanthology.org/2024.findings-acl.726/.

Joachims, T. (1998). "Text categorization with Support Vector Machines: Learning with many relevant features". In: *Machine Learning: ECML-98*. Ed. by C. N{\'e}dellec. Springer, Berlin, Heidelberg, pp. 137–142. ISBN: 978-3-540-69781-7. DOI: 10.1007/BFB0026683. URL: https://link.springer.com/chapter/10.1007/BFb0026683.

John Camm, A., G. Y. Lip, R. De Caterina, I. Savelieva, D. Atar, S. H. Hohnloser, G. Hindricks, P. Kirchhof, J. J. Bax, H. Baumgartner, C. Ceconi, V. Dean, C. Deaton, R. Fagard, C. Funck-Brentano, D. Hasdai, A. Hoes, J. Knuuti, T. McDonagh, C. Moulin, B. A. Popescu, Ž. Reiner, U. Sechtem, P. A. Sirnes, M. Tendera, A. Torbicki, A. Vahanian, S. Windecker, P. Vardas, N. Al-Attar, O. Alfieri, A. Angelini, C. Blömstrom-Lundqvist, P. Colonna, J. De Sutter, S. Ernst, A. Goette, B. Gorenek, R. Hatala, H. Heidbüchel, M. Heldal, S. D. Kristensen, P. Kolh, J. Y. Le Heuzey, H. Mavrakis, L. Mont, P. P. Filardi, P. Ponikowski, B. Prendergast, F. H. Rutten, U. Schotten, I. C. Van Gelder, and F. W. Verheugt (Nov. 2012). "2012 focused update of the ESC Guidelines for the management of atrial fibrillation: An update of the 2010 ESC Guidelines for the management of atrial fibrillationDeveloped with the special contribution of the European Heart Rhythm Association". In: *European Heart Journal* 33.21, pp. 2719–2747. ISSN: 0195-668X. DOI: 10.1093/EURHEARTJ/EHS253. URL: https://dx.doi.org/10.1093/eurheartj/ehs253.

Johnson, A. E. W., T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark (May 2016). "MIMIC-III, a freely accessible critical care database". en. In: *Scientific Data* 3.1, p. 160035. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.35. URL: https://www.nature.com/articles/sdata201635.

Johnson, A. E., L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, B. Moody, B. Gow, L. w. H. Lehman, L. A. Celi, and R. G. Mark (Dec. 2023). "MIMIC-IV, a freely accessible electronic health record dataset". In: *Scientific Data* 10.1, pp. 1–9. ISSN: 20524463. DOI: 10.1038/S41597-022-01899-X;SUBJMETA=174,228,308,478,

692,700;KWRD=EPIDEMIOLOGY,HEALTH+SERVICES,PUBLIC+HEALTH. URL: https://www.nature.com/articles/s41597-022-01899-x.

Jurafsky, D. and J. H. Martin (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd. Stanford. URL: https://web.stanford.edu/~jurafsky/slp3/.

Kadra, G., R. Stewart, H. Shetty, R. G. Jackson, M. A. Greenwood, A. Roberts, C. K. Chang, J. H. MacCabe, and R. D. Hayes (July 2015). "Extracting antipsychotic polypharmacy data from electronic health records: developing and evaluating a novel process". In: *BMC Psychiatry 2015 15:1* 15.1, pp. 1–7. ISSN: 1471-244X. DOI: 10.1186/S12888-015-0557-Z. URL: https://bmcpsychiatry.biomedcentral.com/articles/10.1186/s12888-015-0557-z.

Kaspar, M., C. Morbach, G. Fette, M. Ertl, L. K. Seidlmayer, J. Krebs, G. Dietrich, L. Liman, F. Puppe, and S. Störk (Nov. 2019). "Information Extraction from Echocardiography Reports for a Clinical Follow-up Study-Comparison of Extracted Variables Intended for General Use in a Data Warehouse with Those Intended Specifically for the Study". In: *Methods of Information in Medicine* 58.4-5, pp. 140–150. ISSN: 2511705X. DOI: 10.1055/S-0039-3402069/ID/JR190033-8/BIB. URL: http://www.thieme-connect.de/products/ejournals/html/10.1055/s-0039-3402069%20http://www.thieme-connect.de/DOI/DOI?10.1055/s-0039-3402069.

Kayser, M., B. Menzat, C. Emde, B. Bercean, A. Novak, A. Espinosa, B. W. Papiez, S. Gaube, T. Lukasiewicz, and O. M. Camburu (2024). "Fool Me Once? Contrasting Textual and Visual Explanations in a Clinical Decision-Support Setting". In: *EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 18891–18919. DOI: 10.18653/V1/2024.EMNLP-MAIN.1051. URL: https://aclanthology.org/2024.emnlp-main.1051/.

Kester, S., Y. Hwee, T. Ng, and K. Y. Ngiam (2016). "Automated Anonymization as Spelling Variant Detection". In: *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, pp. 99–103. URL: https://aclanthology.org/W16-4214/.

Khalifa, A. and S. Meystre (Dec. 2015). "Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes". In: *Journal of Biomedical Informatics* 58, S128–S132. ISSN: 1532-0464. DOI: 10.1016/J.JBI.2015.08.002. URL: https://www.sciencedirect.com/science/article/pii/S1532046415001690.

Kim, Y. (2014). "Convolutional Neural Networks for Sentence Classification". In: *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 1746–1751. DOI: 10.3115/V1/D14-1181. URL: https://aclanthology.org/D14-1181/.

Kittner, M., M. Lamping, D. T. Rieke, J. Götze, B. Bajwa, I. Jelas, G. Rüter, H. Hautow, M. Sänger, M. Habibi, M. Zettwitz, T. D. Bortoli, L. Ostermann, J. Ševa, J. Starlinger, O. Kohlbacher, N. P. Malek, U. Keilholz, and U. Leser (Apr. 2021). "Annotation and initial evaluation of a large annotated German oncological corpus". In: *JAMIA Open* 4.2, pp. 1–9. ISSN: 25742531. DOI: 10.1093/JAMIAOPEN/OOAB025. URL: https://dx.doi.org/10.1093/jamiaopen/ooab025.

Klie, J.-C., M. Bugert, B. Boullosa, R. E. d. Castilho, and I. Gurevych (2018). *The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation*. URL: https://aclanthology.org/C18-2002.

Koehn, P. (2004). *Statistical Significance Tests for Machine Translation Evaluation*. URL: https://aclanthology.org/W04-3250/.

Kojima, T., S. Shane Gu, M. Reid Google Research, Y. Matsuo, and Y. Iwasawa (2022). "Large Language Models are Zero-Shot Reasoners". In: *36th Conference on Neural Information Processing Systems (NeurIPS 2022).*

König, M., A. Sander, I. Demuth, D. Diekmann, and E. Steinhagen-Thiessen (Nov. 2019). "Knowledge-based best of breed approach for automated detection of clinical events based on German free text digital hospital discharge letters". In: *PLOS ONE* 14.11, e0224916. ISSN: 1932-6203. DOI: 10.1371/JOURNAL.PONE.0224916. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0224916.

Krippendorff, K. (Jan. 2004). *Content Analysis : An Introduction to its Methodology*. 2. ed. SAGE, pp. 97–149. ISBN: 0761915443. URL: http://www.ncbi.nlm.nih.gov/pubmed/24980578.

Lafferty, J. D., A. McCallum, and F. C. N. Pereira (2001). "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". In: *Proceedings of the Eighteenth International Conference on Machine Learning*. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 282–289. ISBN: 1558607781.

Lake, B. M., R. Salakhutdinov, and J. B. Tenenbaum (Dec. 2015). "Human-level concept learning through probabilistic program induction". In: *Science* 350.6266, pp. 1332–1338. ISSN: 10959203. DOI: 10.1126/SCIENCE.AAB3050. URL: /doi/pdf/10.1126/science.aab3050.

Landolsi, M. Y., L. Hlaoua, and L. Ben Romdhane (Feb. 2023). "Information extraction from electronic medical documents: state of the art and future research directions". In: *Knowledge and Information Systems* 65.2, pp. 463–516. ISSN: 02193116. DOI: 10.1007/S10115-022-01779-1/METRICS. URL: https://link.springer.com/article/10.1007/s10115-022-01779-1.

Leaman, R., R. Khare, and Z. Lu (Oct. 2015). "Challenges in Clinical Natural Language Processing for Automated Disorder Normalization". In: *Journal of biomedical informatics* 57, p. 28. ISSN: 15320464. DOI: 10.1016/J.JBI.2015.07.010. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC4713367/.

Lee, J., W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang (Feb. 2020). "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* 36.4, pp. 1234–1240. ISSN: 1367-4803. DOI: 10.1093/BIOINFORMATICS/BTZ682. URL: https://dx.doi.org/10.1093/bioinformatics/btz682.

Lehman, E. (May 2024). "Practical Considerations For the Deployment of Clinical NLP Systems". PhD thesis. MASSACHUSETTS INSTITUTE OF TECHNOLOGY.

Lehman, E., E. Hernandez, D. Mahajan, J. Wulff, M. J. Smith, Z. Ziegler, D. Nadler, P. Szolovits, A. Johnson, A. J. Ca, E. Alsentzer, and E. H. Edu (2023). "Do We Still Need Clinical Language Models?" In: *Proceedings of Machine Learning Research* 209, p. 2023. URL: https://github.com/elehman16/clinical.

Lentzen, M., S. Madan, V. Lage-Rupprecht, L. Kühnel, J. Fluck, M. Jacobs, M. Mittermaier, M. Witzenrath, P. Brunecker, M. Hofmann-Apitius, J. Weber, and H. Fröhlich (Oct. 2022). "Critical assessment of transformer-based AI models for German clinical notes". In: *JAMIA Open* 5.4, pp. 1–10. ISSN: 25742531. DOI: 10.1093/JAMIAOPEN/OOAC087. URL: https://dx.doi.org/10.1093/jamiaopen/ooac087.

Li, F., Y. Jin, W. Liu, B. P. S. Rawat, P. Cai, and H. Yu (Sept. 2019). "Fine-Tuning Bidirectional Encoder Representations From Transformers (BERT)-Based Models on Large-Scale Electronic Health Record Notes: An Empirical Study." In: *JMIR medical informatics* 7.3, e14830. ISSN: 2291-9694. DOI: 10.2196/14830. URL: http://www.ncbi.nlm.nih.

gov/pubmed/31516126%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6746103.

Li, Q., H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, and L. He (Apr. 2022a). "A Survey on Text Classification: From Traditional to Deep Learning". In: *ACM Transactions on Intelligent Systems and Technology* 13.2, p. 31. ISSN: 21576912. DOI: 10.1145/3495162/ASSET/DB0353BB-EC92-40C0-8261-FE76FCBE4FAF/ASSETS/IMAGES/LARGE/TIST1302-31-T04.JPG. URL: https://dl.acm.org/doi/pdf/10.1145/3495162.

Li, Y., R. M. Wehbe, F. S. Ahmad, H. Wang, and Y. Luo (Jan. 2022b). "Clinical-Longformer and Clinical-BigBird: Transformers for long clinical sequences". In: *arxiv preprint*. URL: https://arxiv.org/pdf/2201.11838.

— (Jan. 2023). "A comparative study of pretrained language models for long clinical text". In: *Journal of the American Medical Informatics Association* 30.2, pp. 340–347. ISSN: 1527974X. DOI: 10.1093/JAMIA/OCAC225. URL: https://dx.doi.org/10.1093/jamia/ocac225.

Ling, W., T. Luís, L. Marujo, R. F. Astudillo, S. Amir, C. Dyer, A. W. Black, and I. Trancoso (2015). "Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation". In: *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics (ACL), pp. 1520–1530. ISBN: 9781941643327. DOI: 10.18653/V1/D15-1176. URL: https://aclanthology.org/D15-1176/.

Liu, M. X., F. Liu, A. J. Fiannaca, T. Koo, L. Dixon, M. Terry, and C. J. Cai (May 2024a). ""We Need Structured Output": Towards User-centered Constraints on Large Language Model Output". In: *Conference on Human Factors in Computing Systems - Proceedings*. DOI: 10.1145/3613905.3650756/SUPPL{_}FILE/3613905.3650756-TALK-VIDEO.VTT. URL: /doi/pdf/10.1145/3613905.3650756?download=true.

Liu, P., W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig (Sept. 2023). "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing". In: *ACM Computing Surveys* 55.9. ISSN: 15577341. DOI: 10.1145/3560815/SUPPL{_}FILE/3560815-APP.PDF. URL: https://dl.acm.org/doi/pdf/10.1145/3560815.

Liu, S., Z. Li, X. Liu, R. Zhan, D. F. Wong, L. S. Chao, and M. Zhang (2024b). "Can LLMs Learn Uncertainty on Their Own? Expressing Uncertainty Effectively in A Self-Training Manner". In: *EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 21635–21645. DOI: 10.18653/V1/2024.EMNLP-MAIN.1205. URL: https://aclanthology.org/2024.emnlp-main.1205/.

Liu, Y., D. Li, K. Wang, Z. Xiong, F. Shi, J. Wang, B. Li, and B. Hang (Sept. 2024c). "Are LLMs good at structured outputs? A benchmark for evaluating structured output capabilities in LLMs". In: *Information Processing & Management* 61.5, p. 103809. ISSN: 0306-4573. DOI: 10.1016/J.IPM.2024.103809. URL: https://www.sciencedirect.com/science/article/abs/pii/S0306457324001687.

Liu, Z., B. Tang, X. Wang, and Q. Chen (Nov. 2017). "De-identification of clinical notes via recurrent neural network and conditional random field". In: *Journal of Biomedical Informatics* 75, S34–S42. ISSN: 1532-0464. DOI: 10.1016/J.JBI.2017.05.023. URL: https://www.sciencedirect.com/science/article/pii/S1532046417301223.

Logan, R. L., I. Balažević, E. Wallace, F. Petroni, S. Singh, and S. Riedel (2022). "Cutting Down on Prompts and Parameters: Simple Few-Shot Learning with Language Models". In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics (ACL), pp. 2824–2835. ISBN: 9781955917254.

DOI: 10.18653/V1/2022.FINDINGS-ACL.222. URL: https://aclanthology.org/2022.findings-acl.222/.

Lohr, C., S. Buechel, and U. Hahn (2018a). "Sharing Copies of Synthetic Clinical Corpora without Physical Distribution — A Case Study to Get Around IPRs and Privacy Constraints Featuring the German JSYNCC Corpus". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. URL: https://aclanthology.org/L18-1201/.

Lohr, C., E. Eder, and U. Hahn (July 2021). "Pseudonymization of PHI Items in German Clinical Reports". In: *Public Health and Informatics: Proceedings of MIE 2021*, pp. 273–277. DOI: 10.3233/SHTI210163. URL: https://ebooks.iospress.nl/doi/10.3233/SHTI210163.

Lohr, C., S. Luther, F. Matthies, L. Modersohn, D. Ammon, K. Saleh, A. G. Henkel, M. Kiehntopf, and U. Hahn (2018b). "CDA-Compliant Section Annotation of German-Language Discharge Summaries: Guideline Development, Annotation Campaign, Section Classification". In: *AMIA Annual Symposium Proceedings* 2018, p. 770. ISSN: 1942597X. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC6371337/.

Lohr, C., F. Matthies, J. Faller, L. Modersohn, A. Riedel, U. Hahn, R. Kiser, M. Boeker, and F. Meineke (Aug. 2024). "De-Identifying GRASCCO – A Pilot Study for the De-Identification of the German Medical Text Project (GeMTeX) Corpus". In: *Studies in Health Technology and Informatics* 317, pp. 171–179. ISSN: 18798365. DOI: 10.3233/SHTI240853. URL: https://ebooks.iospress.nl/doi/10.3233/SHTI240853.

Lohr, C., L. Modersohn, J. Hellrich, T. Kolditz, and U. Hahn (June 2020). "An Evolutionary Approach to the Annotation of Discharge Summaries". In: *Studies in Health Technology and Informatics* 270, pp. 28–32. ISSN: 18798365. DOI: 10.3233/SHTI200116. URL: https://ebooks.iospress.nl/doi/10.3233/SHTI200116.

Long, W. (2005). "Extracting Diagnoses from Discharge Summaries". In: *AMIA Annual Symposium Proceedings*. Vol. 2005, p. 470. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC1560678/.

Lu, A., H. Zhang, Y. Zhang, X. Wang, and D. Yang (2023). "Bounding the Capabilities of Large Language Models in Open Text Generation with Prompt Constraints". In: *EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Findings of EACL 2023*. Association for Computational Linguistics (ACL), pp. 1982–2008. ISBN: 9781959429470. DOI: 10.18653/V1/2023.FINDINGS-EACL.148. URL: https://aclanthology.org/2023.findings-eacl.148/.

Lu, Y., Q. Liu, D. Dai, X. Xiao, H. Lin, X. Han, L. Sun, and H. Wu (2022). "Unified Structure Generation for Universal Information Extraction". In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Vol. 1. Association for Computational Linguistics (ACL), pp. 5755–5772. ISBN: 9781955917216. DOI: 10.18653/V1/2022.ACL-LONG.395. URL: https://aclanthology.org/2022.acl-long.395/.

Lundberg, S. M., P. G. Allen, and S.-I. Lee (2017). "A Unified Approach to Interpreting Model Predictions". In: *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*. DOI: 10.5555/3295222.3295230. URL: https://dl.acm.org/doi/pdf/10.5555/3295222.3295230.

Ma, R., X. Zhou, T. Gui, Y. Tan, L. Li, Q. Zhang, and X. Huang (2022). "Template-free Prompt Tuning for Few-shot NER". In: *NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*. Association for Computational Linguistics

(ACL), pp. 5721–5732. ISBN: 9781955917711. DOI: 10.18653/V1/2022.NAACL-MAIN.420. URL: https://aclanthology.org/2022.naacl-main.420/.

Manning, C. D., M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky (2014). "The Stanford CoreNLP Natural Language Processing Toolkit". In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Vol. 2014-June. Association for Computational Linguistics (ACL), pp. 55–60. ISBN: 9781937284794. DOI: 10.3115/V1/P14-5010. URL: https://aclanthology.org/P14-5010/.

Marimon, M., J. Vivaldi, and N. ´. Uria Bel (July 2017). "Annotation of negation in the IULA Spanish Clinical Record Corpus". In: *Proceedings of the Workshop Computational Semantics Beyond Events and Roles*. Association for Computational Linguistics (ACL), pp. 43–52. DOI: 10.18653/V1/W17-1807. URL: https://aclanthology.org/W17-1807/.

Martin, P., W. E. Haefeli, and M. Martin-Facklam (Sept. 2004). "A drug database model as a central element for computer-supported dose adjustment within a CPOE system". In: *Journal of the American Medical Informatics Association* 11.5, pp. 427–432. ISSN: 10675027. DOI: 10.1197/JAMIA.M1296/2/M{_}JAMIAM1296F01.JPEG. URL: https://dx.doi.org/10.1197/jamia.M1296.

Meineke, F., L. Modersohn, M. Loeffler, and M. Boeker (May 2023). "Announcement of the German Medical Text Corpus Project (GeMTeX)". In: *Studies in health technology and informatics* 302, pp. 835–836. ISSN: 1879-8365. DOI: 10.3233/SHTI230283. URL: https://pubmed.ncbi.nlm.nih.gov/37203512/.

Mekaj, Y. H., A. Y. Mekaj, S. B. Duci, and E. I. Miftari (June 2015). "New oral anticoagulants: Their advantages and disadvantages compared with vitamin K antagonists in the prevention and treatment of patients with thromboembolic events". In: *Therapeutics and Clinical Risk Management* 11, pp. 967–977. ISSN: 1178203X. DOI: 10.2147/TCRM.S84210;WGROUP:STRING:PUBLICATION. URL: https://www.tandfonline.com/doi/pdf/10.2147/TCRM.S84210.

Miglani, V., A. Yang, A. H. Markosyan, D. Garcia-Olano, and N. Kokhlikyan (2023). "Using Captum to Explain Generative Language Models". In: *3rd Workshop for Natural Language Processing Open Source Software, NLP-OSS 2023, Proceedings of the Workshop*. Association for Computational Linguistics (ACL), pp. 165–173. ISBN: 9798891760455. DOI: 10.18653/V1/2023.NLPOSS-1.19. URL: https://aclanthology.org/2023.nlposs-1.19/.

Modersohn, L., S. Schulz, C. Lohr, and U. Hahn (Aug. 2022). "GRASCCO — The First Publicly Shareable, Multiply-Alienated German Clinical Text Corpus". In: *Studies in Health Technology and Informatics* 296, pp. 66–72. ISSN: 18798365. DOI: 10.3233/SHTI220805. URL: https://ebooks.iospress.nl/doi/10.3233/SHTI220805.

Modi, S., K. A. Kasmiran, N. Mohd Sharef, and M. Y. Sharum (Mar. 2024). "Extracting adverse drug events from clinical Notes: A systematic review of approaches used". In: *Journal of Biomedical Informatics* 151. ISSN: 15320464. DOI: 10.1016/j.jbi.2024.104603. URL: https://pubmed.ncbi.nlm.nih.gov/38331081/.

Moral-González, R. del, H. Gómez-Adorno, and O. Ramos-Flores (Dec. 2025). "Comparative analysis of generative LLMs for labeling entities in clinical notes". In: *Genomics & informatics* 23.1. ISSN: 1598-866X. DOI: 10.1186/S44342-024-00036-X. URL: https://pubmed.ncbi.nlm.nih.gov/39915888/.

Mosca, E., F. Szigeti, S. Tragianni, D. Gallagher, and G. Groh (Oct. 2022). "SHAP-Based Explanation Methods: A Review for NLP Interpretability". In: *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 4593–4603. URL: https://aclanthology.org/2022.coling-1.406/.

Mujtaba, G., L. Shuib, N. Idris, W. L. Hoo, R. G. Raj, K. Khowaja, K. Shaikh, and H. F. Nweke (Feb. 2019). "Clinical text classification research trends: Systematic literature review and open issues". In: *Expert Systems with Applications* 116, pp. 494–520. ISSN: 0957-4174. DOI: 10.1016/J.ESWA.2018.09.034.

Mykowiecka, A., M. Marciniak, and A. Kupść (Oct. 2009). "Rule-based information extraction from patients' clinical data". In: *Journal of Biomedical Informatics* 42.5, pp. 923–936. ISSN: 1532-0464. DOI: 10.1016/J.JBI.2009.07.007. URL: https://www.sciencedirect.com/science/article/pii/S1532046409001002.

Nag, A., S. Chakrabarti, A. Mukherjee, and N. Ganguly (June 2025). "Efficient Continual Pretraining of LLMs for Low-resource Languages". In: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*. Association for Computational Linguistics (ACL), pp. 304–317. DOI: 10.18653/V1/2025.NAACL-INDUSTRY.25. URL: https://aclanthology.org/2025.naacl-industry.25/.

Nath, C., M. S. Albaghdadi, and S. R. Jonnalagadda (Apr. 2016). "A Natural Language Processing Tool for Large-Scale Data Extraction from Echocardiography Reports". In: *PLOS ONE* 11.4, e0153749. ISSN: 1932-6203. DOI: 10.1371/JOURNAL.PONE.0153749. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0153749.

Névéol, A., H. Dalianis, S. Velupillai, G. Savova, and P. Zweigenbaum (Mar. 2018). "Clinical Natural Language Processing in languages other than English: Opportunities and challenges". In: *Journal of Biomedical Semantics* 9.1, pp. 1–13. ISSN: 20411480. DOI: 10.1186/S13326-018-0179-8/TABLES/2. URL: https://link.springer.com/articles/10.1186/s13326-018-0179-8%20https://link.springer.com/article/10.1186/s13326-018-0179-8.

Obamuyide, A. and B. Johnston (2022). "Meta-Learning Adaptive Knowledge Distillation for Efficient Biomedical Natural Language Processing". In: *2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing - Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*. Association for Computational Linguistics (ACL), pp. 131–137. ISBN: 9781959429043. DOI: 10.18653/V1/2022.FINDINGS-AACL.12. URL: https://aclanthology.org/2022.findings-aacl.12/.

OpenAI (2023). "GPT-4 Technical Report". URL: https://cdn.openai.com/papers/gpt-4.pdf.

Otter, D. W., J. R. Medina, and J. K. Kalita (Feb. 2021). "A Survey of the Usages of Deep Learning for Natural Language Processing". In: *IEEE Transactions on Neural Networks and Learning Systems* 32.2, pp. 604–624. ISSN: 21622388. DOI: 10.1109/TNNLS.2020.2979670.

Padó, S. (2006). *User's guide to \texttt{sigf}: Significance testing by approximate randomisation*.

Paes, L. M., D. Wei, H. J. Do, H. Strobelt, R. Luss, A. Dhurandhar, M. Nagireddy, K. N. Ramamurthy, P. Sattigeri, W. Geyer, and S. S. Ghosh (Aug. 2025). "Multi-Level Explanations for Generative Language Models". In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 1, pp. 32291–32317. DOI: 10.18653/V1/2025.ACL-LONG.1553. URL: https://aclanthology.org/2025.acl-long.1553/.

Pal, A., L. K. Umapathi, and M. Sankarasubbu (Apr. 2022). *MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering*. URL: https://proceedings.mlr.press/v174/pal22a.html.

Parcalabescu, L. and A. Frank (2024). "On Measuring Faithfulness or Self-consistency of Natural Language Explanations". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 1, pp. 6048–6089. DOI: 10.18653/V1/2024.ACL-LONG.329. URL: https://aclanthology.org/2024.acl-long.329/.

Parnami, A. and M. Lee (Mar. 2022). "Learning from Few Examples: A Summary of Approaches to Few-Shot Learning". In: *arxiv preprint*. URL: https://arxiv.org/pdf/2203.04291.

Patrick, J. and M. Li (Sept. 2010). "High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge". In: *Journal of the American Medical Informatics Association : JAMIA* 17.5, p. 524. ISSN: 10675027. DOI: 10.1136/JAMIA.2010.003939. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC2995676/.

Patterson, O. V., M. S. Freiberg, M. Skanderson, J. S. Fodeh, C. A. Brandt, and S. L. DuVall (June 2017). "Unlocking echocardiogram measurements for heart disease research through natural language processing". In: *BMC Cardiovascular Disorders* 17.1, pp. 1–11. ISSN: 14712261. DOI: 10.1186/S12872-017-0580-8;TYPE=ARTICLE;KWRD=NATURAL. URL: https://bmccardiovascdisord.biomedcentral.com/articles/10.1186/s12872-017-0580-8.

Perkins, S. W., J. C. Muste, T. Alam, and R. P. Singh (June 2024). "Improving Clinical Documentation with Artificial Intelligence: A Systematic Review". In: *Perspectives in Health Information Management* 21.2, p. 1d. ISSN: 15594122. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC11605373/.

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer (2018). "Deep Contextualized Word Representations". In: *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* 1, pp. 2227–2237. DOI: 10.18653/V1/N18-1202. URL: https://aclanthology.org/N18-1202/.

Pfister, J., J. Wunderle, and A. Hotho (Aug. 2025). "LLäMmlein: Transparent, Compact and Competitive German-Only Language Models from Scratch". In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1. Association for Computational Linguistics (ACL), pp. 2227–2246. DOI: 10.18653/V1/2025.ACL-LONG.111. URL: https://aclanthology.org/2025.acl-long.111/.

Pomares-Quimbaya, A., M. Kreuzthaler, and S. Schulz (July 2019). "Current approaches to identify sections within clinical narratives from electronic health records: a systematic review". In: *BMC medical research methodology* 19.1, p. 155. ISSN: 14712288. DOI: 10.1186/S12874-019-0792-Y/FIGURES/3. URL: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-019-0792-y.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever (2019). "Language Models are Unsupervised Multitask Learners". In: *OpenAI blog*. URL: https://github.com/codelucas/newspaper.

Reynolds, L. and K. McDonell (May 2021). "Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm". In: *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery. ISBN: 9781450380959. DOI: 10.1145/3411763.3451760. URL: /doi/pdf/10.1145/3411763.3451760?download=true.

Ribeiro, M. T., S. Singh, and C. Guestrin (2016). ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *NAACL-HLT 2016 - 2016 Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session. Association for Computational Linguistics (ACL), pp. 97–101. DOI: 10.18653/V1/N16-3020. URL: https://aclanthology.org/N16-3020/.

Richter-Pechanski, P. et al. (2022). *CARDIO:DE*. DOI: 10.11588/data/AFYQDY. URL: https://doi.org/10.11588/data/AFYQDY.

Richter-Pechanski, P., A. Amr, H. A. Katus, and C. Dieterich (Sept. 2019). "Deep Learning Approaches Outperform Conventional Strategies in De-Identification of German Medical Reports". In: *Studies in Health Technology and Informatics* 267, pp. 101–109. ISSN: 18798365. DOI: 10.3233/SHTI190813. URL: https://ebooks.iospress.nl/doi/10.3233/SHTI190813.

Richter-Pechanski, P., N. A. Geis, C. Kiriakou, D. M. Schwab, and C. Dieterich (Nov. 2021). "Automatic extraction of 12 cardiovascular concepts from German discharge letters using pre-trained language models". In: *Digital Health* 7. ISSN: 20552076. DOI: 10.1177/20552076211057662/SUPPL{_}FILE/SJ-XLSX-9-DHJ-10.1177{_}20552076211057662.XLSX. URL: https://scholar.google.com/scholar_url?url=https://journals.sagepub.com/doi/pdf/10.1177/20552076211057662&hl=de&sa=T&oi=ucasa&ct=ufr&ei=yvh4aMf-OoOuieoPqNm12A4&scisig=AAZF9b_4BP3DIULX59jx6-rvqNla.

Richter-Pechanski, P., S. Riezler, and C. Dieterich (2018). "De-Identification of German Medical Admission Notes". In: *Studies in Health Technology and Informatics* 253, pp. 165–169. ISSN: 18798365. DOI: 10.3233/978-1-61499-896-9-165. URL: https://ebooks.iospress.nl/doi/10.3233/978-1-61499-896-9-165.

Richter-Pechanski, P., M. Seiferling, C. Kiriakou, D. M. Schwab, N. A. Geis, C. Dieterich, and A. Frank (2025). "Medication information extraction using local large language models". In: *Journal of Biomedical Informatics* 169, p. 104898. DOI: https://doi.org/10.1016/j.jbi.2025.104898. URL: https://www.sciencedirect.com/science/article/pii/S1532046425001273.

Richter-Pechanski, P., P. Wiesenbach, D. M. Schwab, C. Kiriakou, M. He, M. M. Allers, A. S. Tiefenbacher, N. Kunz, A. Martynova, N. Spiller, J. Mierisch, F. Borchert, C. Schwind, N. Frey, C. Dieterich, and N. A. Geis (Dec. 2023). "A distributable German clinical corpus containing cardiovascular clinical routine doctor's letters". In: *Scientific Data* 10.1. ISSN: 20524463. DOI: 10.1038/S41597-023-02128-9;SUBJMETA=308,692,700;KWRD=HEALTH+CARE,MEDICAL+RESEARCH.

Richter-Pechanski, P., P. Wiesenbach, D. M. Schwab, C. Kiriakou, N. Geis, C. Dieterich, and A. Frank (Oct. 2024). "Clinical information extraction for lower-resource languages and domains with few-shot learning using pretrained language models and prompting". In: *Natural Language Processing*, pp. 1–24. ISSN: 2977-0424. DOI: 10.1017/NLP.2024.52. URL: https://www.cambridge.org/core/journals/natural-language-processing/article/clinical-information-extraction-for-lowerresource-languages-and-domains-with-fewshot-learning-using-pretrained-language-models-and-prompting/4596EA36DE0034F9A25D7576C4116BC9.

Roberts, A., R. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, I. Roberts, and A. Setzer (Oct. 2009). "Building a semantically annotated corpus of clinical texts". In: *Journal of Biomedical Informatics* 42.5, pp. 950–966. ISSN: 1532-0464. DOI: 10.1016/J.JBI.2008.12.013. URL: https://www.sciencedirect.com/science/article/pii/S1532046409000069.

Rohanian, O., M. Nouriborji, H. Jauncey, S. Kouchaki, F. Nooralahzadeh, L. Clifton, L. Merson, and D. A. Clifton (2024). "Lightweight transformers for clinical natural language

processing". In: *Natural Language Engineering* 30.5, pp. 887–914. ISSN: 1351-3249. DOI: 10.1017/S1351324923000542. URL: https://www.cambridge.org/core/journals/natural-language-engineering/article/lightweight-transformers-for-clinical-natural-language-processing/BEF81FDE6E12B9DC5AD4906AE67CDDEB.

Roller, R., N. Rethmeier, P. Thomas, M. Hübner, H. Uszkoreit, O. Staeck, K. Budde, F. Halleck, and D. Schmidt (2018). "Detecting named entities and relations in German clinical reports". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 10713 LNAI, pp. 146–154. ISSN: 16113349. DOI: https://doi.org/10.1007/978-3-319-73706-5{_}12. URL: https://link.springer.com/chapter/10.1007/978-3-319-73706-5_12.

Roller, R., L. Seiffe, A. Ayach, S. Möller, O. Marten, M. Mikhailov, C. Alt, D. Schmidt, F. Halleck, M. Naik, W. Duettmann, and K. Budde (July 2022). "A Medical Information Extraction Workbench to Process German Clinical Text". In: *arxiv preprint*. URL: https://arxiv.org/pdf/2207.03885.

Rousseeuw, P. J. and M. Hubert (Jan. 2011). "Robust statistics for outlier detection". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1.1, pp. 73–79. ISSN: 19424795. DOI: https://doi.org/10.1002/widm.2. URL: /doi/pdf/10.1002/widm.2%20https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.2%20https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.2.

Sänger, M., L. Weber, M. Kittner, and U. Leser (2019). "Classifying German Animal Experiment Summaries with Multi-lingual BERT". In: *CLEF eHealth 2019 Task 1*. URL: https://www.animaltestinfo.de/.

Şapcı, A. O. B., H. Kemik, R. Yeniterzi, and O. Tastan (May 2024). "Focusing on potential named entities during active label acquisition". In: *Natural Language Engineering* 30.3, pp. 602–624. ISSN: 1351-3249. DOI: 10.1017/S1351324923000165. URL: https://www.cambridge.org/core/journals/natural-language-engineering/article/focusing-on-potential-named-entities-during-active-label-acquisition/087DF41DC645A49AA712D55D486405DD.

Scheible, R., J. Frei, F. Thomczyk, H. He, P. Tippmann, J. Knaus, V. Jaravine, F. Kramer, and M. Boeker (2024). "GottBERT: a pure German Language Model". In: *EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 21237–21250. DOI: 10.18653/V1/2024.EMNLP-MAIN.1183. URL: https://aclanthology.org/2024.emnlp-main.1183/.

Schick, T., H. Schmid, and H. Schütze (2020). "Automatically Identifying Words That Can Serve as Labels for Few-Shot Text Classification". In: *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*. Association for Computational Linguistics (ACL), pp. 5569–5578. ISBN: 9781952148279. DOI: 10.18653/V1/2020.COLING-MAIN.488. URL: https://aclanthology.org/2020.coling-main.488/.

Schick, T. and H. Schütze (2021a). "Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference". In: *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*. Association for Computational Linguistics (ACL), pp. 255–269. ISBN: 9781954085022. DOI: 10.18653/V1/2021.EACL-MAIN.20. URL: https://aclanthology.org/2021.eacl-main.20/.

— (2021b). "It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners". In: *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings*

of the Conference. Association for Computational Linguistics (ACL), pp. 2339–2352. ISBN: 9781954085466. DOI: 10.18653/V1/2021.NAACL-MAIN.185. URL: https://aclanthology.org/2021.naacl-main.185/.

— (June 2022). "True Few-Shot Learning with Prompts—A Real-World Perspective". In: *Transactions of the Association for Computational Linguistics* 10, pp. 716–731. ISSN: 2307387X. DOI: 10.1162/TACL{_}A{_}00485. URL: https://aclanthology.org/2022.tacl-1.41/.

Schlünder, I. (2015). "Datenschutzkonforme Lösungen für die Versorgungsforschung. 14". In: *Deutscher Kongress für Versorgungsforschung*.

Shapley, L. S. (May 2016). "17. A Value for n-Person Games". In: *Contributions to the Theory of Games (AM-28), Volume II*, pp. 307–318. DOI: 10.1515/9781400881970-018/HTML.

Sharif, O., J. Gatto, M. Basak, and S. M. Preum (Feb. 2025). "REGen: A Reliable Evaluation Framework for Generative Event Argument Extraction". In: *arxiv preprint*. URL: https://arxiv.org/pdf/2502.16838.

Sharma, V., A. Thalhammer, A. Kugic, S. Schulz, and M. Kreuzthaler (2024). "Sequence-Model-Based Medication Extraction from Clinical Narratives in German". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 14844 LNAI, pp. 334–344. ISSN: 1611-3349. DOI: 10.1007/978-3-031-66538-7{_}33. URL: https://link.springer.com/chapter/10.1007/978-3-031-66538-7_33.

Sheikhalishahi, S., R. Miotto, J. T. Dudley, A. Lavelli, F. Rinaldi, and V. Osmani (Apr. 2019). "Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review". In: *JMIR Medical Informatics* 7.2, e12239. ISSN: 22919694. DOI: 10.2196/12239. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC6528438/.

Shen, Y., Z. Tan, S. Wu, W. Zhang, R. Zhang, Y. Xi, W. Lu, and Y. Zhuang (2023). "PromptNER: Prompt Locating and Typing for Named Entity Recognition". In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Vol. 1. Association for Computational Linguistics (ACL), pp. 12492–12507. ISBN: 9781959429722. DOI: 10.18653/V1/2023.ACL-LONG.698. URL: https://aclanthology.org/2023.acl-long.698/.

Shin, T., Y. Razeghi, R. L. Logan, E. Wallace, and S. Singh (2020). "AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts". In: *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. Association for Computational Linguistics (ACL), pp. 4222–4235. ISBN: 9781952148606. DOI: 10.18653/V1/2020.EMNLP-MAIN.346. URL: https://aclanthology.org/2020.emnlp-main.346/.

Si, Y., J. Wang, H. Xu, and K. Roberts (Nov. 2019). "Enhancing clinical concept extraction with contextual embeddings". In: *Journal of the American Medical Informatics Association* 26.11, pp. 1297–1304. ISSN: 1527974X. DOI: 10.1093/JAMIA/OCZ096. URL: https://dx.doi.org/10.1093/jamia/ocz096.

Singhal, K., S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, A. Babiker, N. Schärli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, B. Agüera y Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam, and V. Natarajan (Aug. 2023). "Large language models encode clinical knowledge". In: *Nature* 620.7972, pp. 172–180. ISSN: 14764687. DOI: https://doi.org/10.1038/s41586-023-06291-2. URL: https://www.nature.com/articles/s41586-023-06291-2.

Sivarajkumar, S. and Y. Wang (2023). "HealthPrompt: A Zero-shot Learning Paradigm for Clinical Natural Language Processing". In: *AMIA Annual Symposium Proceedings*. Vol. 2022, p. 972. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC10148337/.

Small, A. M., D. H. Kiss, Y. Zlatsin, D. L. Birtwell, H. Williams, M. A. Guerraty, Y. Han, S. Anwaruddin, J. H. Holmes, J. A. Chirinos, R. L. Wilensky, J. Giri, and D. J. Rader (Aug. 2017). "Text mining applied to electronic cardiovascular procedure reports to identify patients with trileaflet aortic stenosis and coronary artery disease". In: *Journal of Biomedical Informatics* 72, pp. 77–84. ISSN: 1532-0464. DOI: 10.1016/J.JBI.2017.06.016. URL: https://www.sciencedirect.com/science/article/pii/S1532046417301387.

Socrates, V., D. S. Wright, T. Huang, S. Fereydooni, C. Dien, L. Chi, J. Albano, B. Patterson, N. S. Kanaparthy, C. X. Wright, A. Loza, D. Chartash, M. Iscoe, and R. A. Taylor (Apr. 2025). "Identifying Deprescribing Opportunities With Large Language Models in Older Adults: Retrospective Cohort Study". In: *JMIR Aging* 8.1, e69504. ISSN: 25617605. DOI: 10.2196/69504. URL: https://aging.jmir.org/2025/1/e69504.

Starlinger, J., M. Kittner, O. Blankenstein, and U. Leser (Aug. 2017). "How to improve information extraction from German medical records". In: *IT - Information Technology* 59.4, pp. 171–179. DOI: https://doi.org/10.1515/itit-2016-0027.

Steffel, J., R. Collins, M. Antz, P. Cornu, L. Desteghe, K. G. Haeusler, J. Oldgren, H. Reinecke, V. Roldan-Schilling, N. Rowell, P. Sinnaeve, T. Vanassche, T. Potpara, A. J. Camm, H. Heidbüchel, G. Y. Lip, T. Deneke, N. Dagres, G. Boriani, T. F. Chao, E. K. Choi, M. T. Hills, I. D. S. Santos, D. A. Lane, D. Atar, B. Joung, O. M. Cole, and M. Field (Oct. 2021). "2021 European Heart Rhythm Association Practical Guide on the Use of Non-Vitamin K Antagonist Oral Anticoagulants in Patients with Atrial Fibrillation". In: *EP Europace* 23.10, pp. 1612–1676. ISSN: 1099-5129. DOI: 10.1093/EUROPACE/EUAB065. URL: https://dx.doi.org/10.1093/europace/euab065.

Sun, C., X. Qiu, Y. Xu, and X. Huang (2019). "How to Fine-Tune BERT for Text Classification?" In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 11856 LNAI. Springer, Cham, pp. 194–206. ISBN: 978-3-030-32381-3. DOI: 10.1007/978-3-030-32381-3{_}16. URL: https://link.springer.com/chapter/10.1007/978-3-030-32381-3_16.

Sundararajan, M., A. Taly, and Q. Yan (2017). "Axiomatic Attribution for Deep Networks". In: *ICML'17: Proceedings of the 34th International Conference on Machine Learning*. DOI: 10.5555/3305890.3306024. URL: https://dl.acm.org/doi/pdf/10.5555/3305890.3306024.

Sweeney, L. (1996). "Replacing personally-identifying information in medical records, the Scrub system". In: *Proceedings of the AMIA Annual Fall Symposium*, p. 333. ISSN: 10918280. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC2233179/.

Tamang, S., M. Humbert-Droz, M. Gianfrancesco, Z. Izadi, G. Schmajuk, and J. Yazdany (Jan. 2023). "Practical Considerations for Developing Clinical Natural Language Processing Systems for Population Health Management and Measurement." In: *JMIR medical informatics* 11.1, e37805. ISSN: 2291-9694. DOI: 10.2196/37805. URL: http://www.ncbi.nlm.nih.gov/pubmed/36595345%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC9846439.

Taylor, N., Y. Zhang, D. W. Joyce, Z. Gao, A. Kormilitzin, and A. Nevado-Holgado (2024). "Clinical Prompt Learning With Frozen Language Models". In: *IEEE Transactions on Neural Networks and Learning Systems* 35.11, pp. 16453–16463. ISSN: 21622388. DOI: 10.1109/TNNLS.2023.3294633.

Temsah, M.-H., A. Jamal, K. Alhasan, A. A. Temsah, and K. H. Malki (Oct. 2024). "OpenAI o1-Preview vs. ChatGPT in Healthcare: A New Frontier in Medical AI Reasoning". In: *Cureus* 16.10. ISSN: 2168-8184. DOI: 10.7759/CUREUS.70640. URL: https://pubmed.ncbi.nlm.nih.gov/39359332/.

Thirunavukarasu, A. J., D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting (Aug. 2023). "Large language models in medicine". In: *Nature Medicine* 29.8, pp. 1930–1940. ISSN: 1546170X. DOI: 10.1038/S41591-023-02448-8;SUBJMETA=1719,308,575,692,700;KWRD=PATIENT+EDUCATION,TRANSLATIONAL+RESEARCH. URL: https://www.nature.com/articles/s41591-023-02448-8.

Timmis, A. et al. (Feb. 2022). "European Society of Cardiology: cardiovascular disease statistics 2021". In: *European Heart Journal* 43.8, pp. 716–799. ISSN: 0195-668X. DOI: 10.1093/EURHEARTJ/EHAB892. URL: https://academic.oup.com/eurheartj/article/43/8/716/6472699.

Tjong, E. F., K. Sang, and F. De Meulder (2003). "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition". In: *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 142–147. URL: https://aclanthology.org/W03-0419/.

Toepfer, M., H. Corovic, G. Fette, P. Klügl, S. Störk, and F. Puppe (Nov. 2015). "Fine-grained information extraction from German transthoracic echocardiography reports". In: *BMC Medical Informatics and Decision Making* 15.1, pp. 1–16. ISSN: 14726947. DOI: 10.1186/S12911-015-0215-X/TABLES/6. URL: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-015-0215-x.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample (Feb. 2023). "LLaMA: Open and Efficient Foundation Language Models". In: *arXiv preprint*. URL: https://arxiv.org/pdf/2302.13971.

Turchin, A., N. S. Kolatkar, R. W. Grant, E. C. Makhni, M. L. Pendergrass, and J. S. Einbinder (Nov. 2006). "Using Regular Expressions to Abstract Blood Pressure and Treatment Intensification Information from the Text of Physician Notes". In: *Journal of the American Medical Informatics Association* 13.6, pp. 691–695. ISSN: 10675027. DOI: 10.1197/JAMIA.M2078,. URL: https://pubmed.ncbi.nlm.nih.gov/16929043/.

Unlu, O., E. B. Levitan, E. Reshetnyak, J. Kneifati-Hayek, I. Diaz, A. Archambault, L. Chen, J. T. Hanlon, M. S. Maurer, M. M. Safford, M. S. Lachs, and P. Goyal (Nov. 2020). "Polypharmacy in Older Adults Hospitalized for Heart Failure". In: *Circulation: Heart Failure* 13.11, E006977. ISSN: 19413297. DOI: 10.1161/CIRCHEARTFAILURE.120.006977/SUPPL{_}FILE/CIRCHF{_}CIRCHF-2020-006977{_}SUPP1.PDF. URL: /doi/pdf/10.1161/CIRCHEARTFAILURE.120.006977?download=true.

Uzuner, Ö., T. C. Sibanda, Y. Luo, and P. Szolovits (Jan. 2008). "A de-identifier for medical discharge summaries". In: *Artificial Intelligence in Medicine*. Vol. 42. Artif Intell Med, pp. 13–35. DOI: 10.1016/j.artmed.2007.10.001. URL: https://pubmed.ncbi.nlm.nih.gov/18053696/.

Uzuner, Ö., I. Solti, and E. Cadag (Sept. 2010a). "Extracting medication information from clinical text". In: *Journal of the American Medical Informatics Association : JAMIA* 17.5, p. 514. ISSN: 10675027. DOI: 10.1136/JAMIA.2010.003947. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC2995677/.

Uzuner, Ö., I. Solti, F. Xia, and E. Cadag (Sept. 2010b). "Community annotation experiment for ground truth generation for the i2b2 medication challenge". In: *Journal of the*

American Medical Informatics Association 17.5, pp. 519–523. ISSN: 10675027. DOI: 10.1136/JAMIA.2010.004200,. URL: https://pubmed.ncbi.nlm.nih.gov/20819855/.

Uzuner, Ö., B. R. South, S. Shen, and S. L. DuVall (Sept. 2011). "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text". In: *Journal of the American Medical Informatics Association : JAMIA* 18.5, p. 552. ISSN: 10675027. DOI: 10.1136/AMIAJNL-2011-000203. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC3168320/.

Van der Hulle, T., J. Kooiman, P. L. den Exter, O. M. Dekkers, F. A. Klok, and M. V. Huisman (Mar. 2014). "Effectiveness and safety of novel oral anticoagulants as compared with vitamin K antagonists in the treatment of acute symptomatic venous thromboembolism: a systematic review and meta-analysis". In: *Journal of Thrombosis and Haemostasis* 12.3, pp. 320–328. ISSN: 1538-7836. DOI: 10.1111/JTH.12485. URL: https://www.sciencedirect.com/science/article/pii/S1538783622038727#s0050.

Van Gelder, I. C. et al. (Sept. 2024). "2024 ESC Guidelines for the management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS): Developed by the task force for the management of atrial fibrillation of the European Society of Cardiology (ESC), with the special contribution of the European Heart Rhythm Association (EHRA) of the ESC. Endorsed by the European Stroke Organisation (ESO)". In: *European Heart Journal* 45.36, pp. 3314–3414. ISSN: 0195-668X. DOI: 10.1093/EURHEARTJ/EHAE176. URL: https://dx.doi.org/10.1093/eurheartj/ehae176.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017). "Attention is all you need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., pp. 6000–6010. ISBN: 9781510860964.

Wang, J., C. Wang, F. Luo, C. Tan, M. Qiu, F. Yang, Q. Shi, S. Huang, and M. Gao (2022). "Towards Unified Prompt Tuning for Few-shot Text Classification". In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics (ACL), pp. 524–536. ISBN: 9781959429432. DOI: 10.18653/V1/2022.FINDINGS-EMNLP.37. URL: https://aclanthology.org/2022.findings-emnlp.37/.

Wang, Y., L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, and H. Liu (Jan. 2018). "Clinical information extraction applications: A literature review". In: *Journal of Biomedical Informatics* 77, pp. 34–49. ISSN: 1532-0464. DOI: 10.1016/J.JBI.2017.11.011. URL: https://www.sciencedirect.com/science/article/pii/S1532046417302563.

Wei, Q., Z. Ji, Z. Li, J. Du, J. Wang, J. Xu, Y. Xiang, F. Tiryaki, S. Wu, Y. Zhang, C. Tao, and H. Xu (Jan. 2019). "A study of deep learning approaches for medication and adverse drug event extraction from clinical text". In: *Journal of the American Medical Informatics Association : JAMIA* 27.1, p. 13. ISSN: 1527974X. DOI: 10.1093/JAMIA/OCZ063. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC6913210/.

Wellner, B., M. Huyck, S. Mardis, J. Aberdeen, A. Morgan, L. Peshkin, A. Yeh, J. Hitzeman, and L. Hirschman (Sept. 2007). "Rapidly Retargetable Approaches to De-identification in Medical Records". In: *Journal of the American Medical Informatics Association* 14.5, pp. 564–573. ISSN: 10675027. DOI: 10.1197/JAMIA.M2435,. URL: https://pubmed.ncbi.nlm.nih.gov/17600096/.

Wheelock, K. M., J. S. Ross, K. Murugiah, Z. Lin, H. M. Krumholz, and R. Khera (Dec. 2021). "Clinician Trends in Prescribing Direct Oral Anticoagulants for US Medicare Beneficiaries". In: *JAMA Network Open* 4.12, e2137288. ISSN: 25743805. DOI: 10.1001/

JAMANETWORKOPEN.2021.37288. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC8649845/.

Wilbur, W. J., A. Rzhetsky, and H. Shatkay (July 2006). "New directions in biomedical text annotation: Definitions, guidelines and corpus construction". In: *BMC Bioinformatics* 7.1, pp. 1–10. ISSN: 14712105. DOI: 10.1186/1471-2105-7-356/TABLES/5. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-356.

Wu, C., W. Lin, X. Zhang, Y. Zhang, W. Xie, and Y. Wang (Sept. 2024). "PMC-LLaMA: toward building open-source language models for medicine". In: *Journal of the American Medical Informatics Association* 31.9, pp. 1833–1843. ISSN: 1527974X. DOI: 10.1093/JAMIA/OCAE045. URL: https://dx.doi.org/10.1093/jamia/ocae045.

Wu, S. and Y. He (Nov. 2019). "Enriching pre-trained language model with entity information for relation classification". In: *International Conference on Information and Knowledge Management, Proceedings*, pp. 2361–2364. DOI: 10.1145/3357384.3358119;WGROUP:STRING:ACM. URL: /doi/pdf/10.1145/3357384.3358119?download=true.

Wu, S., K. Roberts, S. Datta, J. Du, Z. Ji, Y. Si, S. Soni, Q. Wang, Q. Wei, Y. Xiang, B. Zhao, and H. Xu (Mar. 2020). "Deep learning in clinical natural language processing: a methodical review". In: *Journal of the American Medical Informatics Association* 27.3, pp. 457–470. ISSN: 1527974X. DOI: 10.1093/JAMIA/OCZ200. URL: https://dx.doi.org/10.1093/jamia/ocz200.

Wu, Y., M. Jiang, J. Lei, and H. Xu (2015). "Named Entity Recognition in Chinese Clinical Text Using Deep Neural Network". In: *Studies in Health Technology and Informatics* 216, pp. 624–628. ISSN: 18798365. DOI: 10.3233/978-1-61499-564-7-624. URL: https://ebooks.iospress.nl/doi/10.3233/978-1-61499-564-7-624.

Xu, D., W. Chen, W. Peng, C. Zhang, T. Xu, X. Zhao, X. Wu, Y. Zheng, Y. Wang, and E. Chen (Dec. 2024). "Large language models for generative information extraction: a survey". In: *Frontiers of Computer Science* 18.6, pp. 1–24. ISSN: 20952236. DOI: 10.1007/S11704-024-40555-Y/METRICS. URL: https://link.springer.com/article/10.1007/s11704-024-40555-y.

Yang, H., X.-Y. Liu, and C. D. Wang (June 2023). "FinGPT: Open-Source Financial Large Language Models". In: *International Joint Conference on Artificial Intelligence*. Elsevier BV. DOI: 10.2139/SSRN.4489826. URL: https://papers.ssrn.com/abstract=4489826.

Yang, Z., D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy (2016). "Hierarchical Attention Networks for Document Classification". In: *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, pp. 1480–1489. DOI: 10.18653/V1/N16-1174. URL: https://aclanthology.org/N16-1174/.

Yogarajan, V., M. Mayo, and B. Pfahringer (2018). "A survey of automatic de-identification of longitudinal clinical narratives". In: *arxiv preprint*. URL: https://arxiv.org/abs/1810.06765.

Zahl-Holmstad, B., B. H. Garcia, K. Svendsen, T. Johnsgård, R. V. Holis, E. H. Ofstad, T. Risør, E. C. Lehnbom, T. Wisløff, M. Chan, and R. Elenjord (Dec. 2023). "Completeness of medication information in admission notes from emergency departments". In: *BMC Health Services Research* 23.1, pp. 1–12. ISSN: 14726963. DOI: 10.1186/S12913-023-10371-4/FIGURES/4. URL: https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-023-10371-4%20http://creativecom-mons.org/publicdomain/zero/1.0/.

Zelenko, D., C. Aone, and A. Richardella (2002). "Kernel Methods for Relation Extraction". In: *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, EMNLP 2002*. Association for Computational Linguistics (ACL), pp. 71–78. DOI: 10.3115/1118693.1118703. URL: https://aclanthology.org/W02-1010/.

Zeng, D., K. Liu, S. Lai, G. Zhou, and J. Zhao (2014). "Relation Classification via Convolutional Deep Neural Network". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 2335–2344. URL: https://aclanthology.org/C14-1220/.

Zhai, Z., D. Q. Nguyen, and K. Verspoor (2018). "Comparing CNN and LSTM character-level embeddings in BiLSTM-CRF models for chemical and disease named entity recognition". In: *EMNLP 2018 - 9th International Workshop on Health Text Mining and Information Analysis, LOUHI 2018 - Proceedings of the Workshop*, pp. 38–43. DOI: 10.18653/V1/W18-5605. URL: https://aclanthology.org/W18-5605/.

Zhang, M., Y. Zhang, and G. Fu (2017). "End-to-End Neural Relation Extraction with Global Optimization". In: *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*. Association for Computational Linguistics (ACL), pp. 1730–1740. ISBN: 9781945626838. DOI: 10.18653/V1/D17-1182. URL: https://aclanthology.org/D17-1182/.

Zhao, H., H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, and M. Du (Feb. 2024). "Explainability for Large Language Models: A Survey". In: *ACM Transactions on Intelligent Systems and Technology* 15.2, p. 38. ISSN: 21576912. DOI: 10.1145/3639372. URL: https://dl.acm.org/doi/pdf/10.1145/3639372.

Zheng, S., J. J. Lu, N. Ghasemzadeh, S. S. Hayek, A. A. Quyyumi, and F. Wang (Apr. 2017). "Effective Information Extraction Framework for Heterogeneous Clinical Reports Using Online Machine Learning and Controlled Vocabularies". In: *JMIR Medical Informatics* 5.2, e12. ISSN: 22919694. DOI: 10.2196/MEDINFORM.7235. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC5442348/.

Zhou, G., J. Su, J. Zhang, and M. Zhang (2005). "Exploring Various Knowledge in Relation Extraction". In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pp. 427–434. URL: http://www.ldc.upenn.edu/Projects/ACE.

Zhou, H., F. Liu, B. Gu, X. Zou, J. Huang, J. Wu, Y. Li, S. S. Chen, P. Zhou, J. Liu, Y. Hua, C. Mao, C. You, X. Wu, Y. Zheng, L. Clifton, Z. Li, J. Luo, and D. A. Clifton (Nov. 2023). "A Survey of Large Language Models in Medicine: Progress, Application, and Challenge". In: *arxiv preprint*. URL: https://arxiv.org/pdf/2311.05112.

Zhou, H., M. Li, Y. Xiao, H. Yang, and R. Zhang (Sept. 2024). "LEAP: LLM instruction-example adaptive prompting framework for biomedical relation extraction". In: *Journal of the American Medical Informatics Association* 31.9, pp. 2010–2018. ISSN: 1527974X. DOI: 10.1093/JAMIA/OCAE147. URL: https://dx.doi.org/10.1093/jamia/ocae147.

Zhu, Q., Y. Gu, L. Luo, B. Li, C. Li, W. Peng, M. Huang, and X. Zhu (Dec. 2021). "When does Further Pre-training MLM Help? An Empirical Study on Task-Oriented Dialog Pre-training". In: *Proceedings of the Second Workshop on Insights from Negative Results in NLP*. Association for Computational Linguistics (ACL), pp. 54–61. ISBN: 978-1-954085-93-0. DOI: 10.18653/V1/2021.INSIGHTS-1.9. URL: https://aclanthology.org/2021.insights-1.9/.

Appendix A

CARDIO:DE - Distributing a Clinical Corpus

A.1 Annotation guidelines

This appendix contains selected annotation rules that are relevant for interpreting the CARDIO:DE annotation layers. The complete guideline documents are available via heiDATA (cf. https://doi.org/10.11588/DATA/USQLMB).

A.1.1 Section type annotation

The ANAMNESE class also includes headings such as *Spezielle Krankheitsanamnese*. The class ECHOBEFUNDE covers echocardiographic findings introduced by headings such as *Echokardiographie*, *Transthorakale Echokardiographie*, *Dynamische Stress-Echokardiographie*, or *TEE*. The class LABOR may also begin with irregular introductory phrases such as *Werte wurden ermittelt ...*, not only with explicit *Labor* headings. In addition, the class RISIKOFAKTORENALLERGIEN summarizes information on allergies, intolerances, and cardiovascular risk factors, although these may occur as separate subsections in the original letters.

The guidelines further clarify that MIX is used for content that cannot be assigned to one of the predefined section classes, for example *Procedere*, *Nächster Termin/Kontrolle*, *Reiseanamnese*, *Soziales*, or *MRE*. Moreover, laboratory values were excluded from regular section annoation when they occured in improper converted tables.

A.1.2 Medication information

The initial annotation guidelines make an explicit distinction between DRUG and ACTIVEING. DRUG is used for manufacturer or brand names, whereas ACTIVEING covers generic substance names, groups of active ingredients, and general medication classes. For example, generic names, grouped active ingredients, or expressions such as *Antibiotikum*, *OAK*, or *Statin* are annotated as ACTIVEING. The guidelines specify that multi-word drug names remain a single DRUG span even if they already contain strength or a specific active ingredient.

The guidelines define several rules for relation annotation and span boundaries. For REASON, the allowed context window is restricted to the same sentence or the immediately preceding or following sentence. In addition, therapy expressions are annotated as REASON only when they are further specified by an active ingredient. Otherwise, they are treated as ACTIVEING. Articles and prepositions are excluded from spans unless they contribute relevant meaning. Brackets are included only when they belong to a single annotation span; unpaired brackets are excluded. Compound expressions containing medication mention are

annotated as single spans when appropriate. Finally, if frequency and strength information are combined in a single token, the token is annotated as STRENGTH.

A.2 Hyperparameters

Parameter	Value
algorithm	`lbfgs`
c1	0.1
c2	0.1
max_iterations	100
all_possible_transitions	`False`

Table A.1 **Hyperparameters medication information CRF:** Selected hyperparameters for the CRF (medication information).

```
def word2features(sent, i):
    word = sent[i][0]
    postag = sent[i][1]
    features = {
        'bias': 1.0,
        'word.lower()': word.lower(),
        'word[-3:]': word[-3:],
        'word[-2:]': word[-2:],
        'word.isupper()': word.isupper(),
        'word.istitle()': word.istitle(),
        'word.isdigit()': word.isdigit(),
        'postag': postag,
        'postag[:2]': postag[:2]}
    if i > 0:
        word1 = sent[i-1][0]
        postag1 = sent[i-1][1]
        features.update({
            '-1:word.lower()': word1.lower(),
            '-1:word.istitle()': word1.istitle(),
            '-1:word.isupper()': word1.isupper(),
            '-1:postag': postag1,
            '-1:postag[:2]': postag1[:2]})
    else:
        features['BOS'] = True
    if i < len(sent)-1:
        word1 = sent[i+1][0]
        postag1 = sent[i+1][1]
        features.update({
            '+1:word.lower()': word1.lower(),
            '+1:word.istitle()': word1.istitle(),
            '+1:word.isupper()': word1.isupper(),
            '+1:postag': postag1,
            '+1:postag[:2]': postag1[:2],
        })
    else:
        features['EOS'] = True
```

Fig. A.1 **Features medication information CRF:** Selected linguistic features of the CRF (medication information).

Parameter	Value
attention_probs_dropout_prob	0.10
classifier_dropout	null
hidden_act	gelu
hidden_dropout_prob	0.10
hidden_size	768.00
initializer_range	0.02
intermediate_size	3072.00
layer_norm_eps	1.00E-12
max_position_embeddings	512.00
model_type	bert
num_attention_heads	12.00
num_hidden_layers	12.00
pad_token_id	0
position_embedding_type	absolute
transformers_version	4.21.0
type_vocab_size	1
use_cache	true
vocab_size	30000
epochs	6
batch_size	64
learning_rate	5.00E-05
train_val_test_split	360/40 (CARDIO:DE400), 100 (CARDIO:DE100)
optimizer	AdamW

Table A.2 **Hyperparameters medication information BERT:** Selected hyperparameters for BERT (medication information).

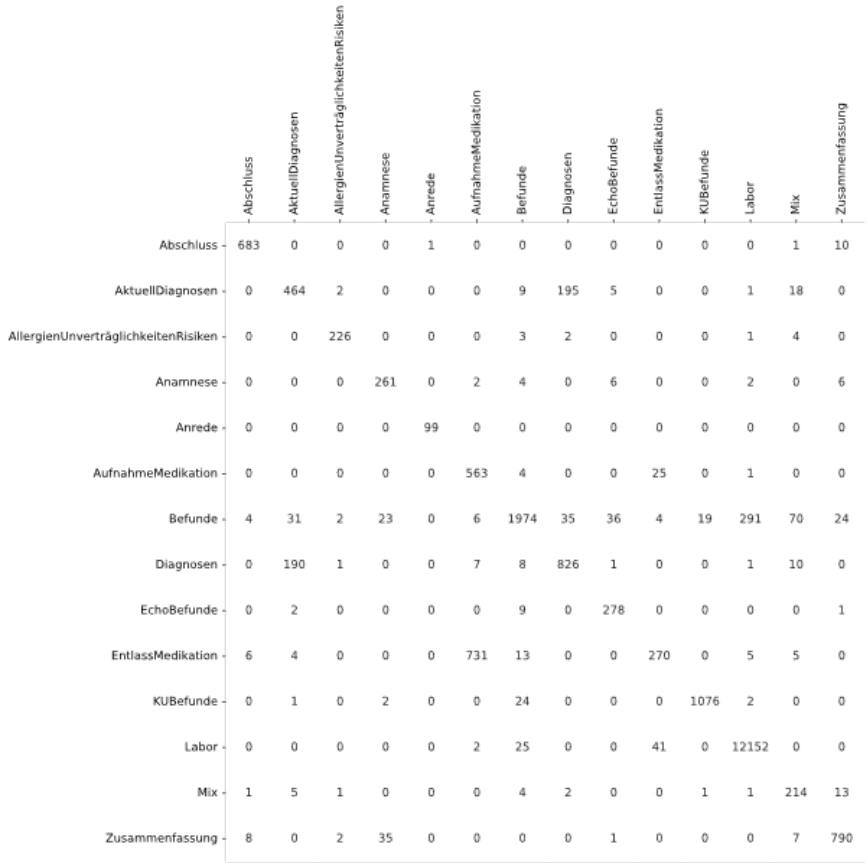
Parameter	Value
C	1
break_ties	False
cache_size	200
class_weight	None
coef0	0
decision_function_shape	ovr
degree	3
gamma	scale
kernel	rbf
max_iter	−1
probability	False
random_state	None
shrinking	True
tol	0.001
verbose	False
feature_vector_generation	TF–IDF vectorizer (scikit-learn v. 1.0.2)
filter	German stopwords using nltk.corpus (v. 3.7)

Table A.3 **Hyperparameters section classification SVM:** Selected hyperparameters for the SVM (section classification).

Parameter	**Value**
attention_probs_dropout_prob	0.1
classifier_dropout	null
hidden_act	gelu
hidden_dropout_prob	0.1
hidden_size	768
initializer_range	0.02
intermediate_size	3072
layer_norm_eps	1.00E-12
max_position_embeddings	512
model_type	bert
num_attention_heads	12
num_hidden_layers	12
pad_token_id	0
position_embedding_type	absolute
transformers_version	4.21.0
type_vocab_size	1
use_cache	true
vocab_size	30000
epochs	10
patience	1
batch_size	32
learning_rate	5.00E-05
train_val_test_split	360/40 (CARDIO:DE400), 100 (CARDIO:DE100)
optimizer	AdamW

Table A.4 **Hyperparameters section classification BERT:** Selected hyperparameters for BERT (section classification).

A.3 Additional results



	Abschluss	AktuellDiagnosen	AllergienUnverträglichkeitenRisiken	Anamnese	Anrede	AufnahmeMedikation	Befunde	Diagnosen	EchoBefunde	EntlassMedikation	KUBefunde	Labor	Mix	Zusammenfassung
Abschluss	683	0	0	0	1	0	0	0	0	0	0	0	1	10
AktuellDiagnosen	0	464	2	0	0	0	9	195	5	0	0	1	18	0
AllergienUnverträglichkeitenRisiken	0	0	226	0	0	0	3	2	0	0	0	1	4	0
Anamnese	0	0	0	261	0	2	4	0	6	0	0	2	0	6
Anrede	0	0	0	0	99	0	0	0	0	0	0	0	0	0
AufnahmeMedikation	0	0	0	0	0	563	4	0	0	25	0	1	0	0
Befunde	4	31	2	23	0	6	1974	35	36	4	19	291	70	24
Diagnosen	0	190	1	0	0	7	8	826	1	0	0	1	10	0
EchoBefunde	0	2	0	0	0	0	9	0	278	0	0	0	0	1
EntlassMedikation	6	4	0	0	0	731	13	0	0	270	0	5	5	0
KUBefunde	0	1	0	2	0	0	24	0	0	0	1076	2	0	0
Labor	0	0	0	0	0	2	25	0	0	41	0	12152	0	0
Mix	1	5	1	0	0	0	4	2	0	0	1	1	214	13
Zusammenfassung	8	0	2	35	0	0	0	0	1	0	0	0	7	790

Fig. A.2 **Confusion matrix section classification BERT:** Confusion matrix of the BERT model (section classification).

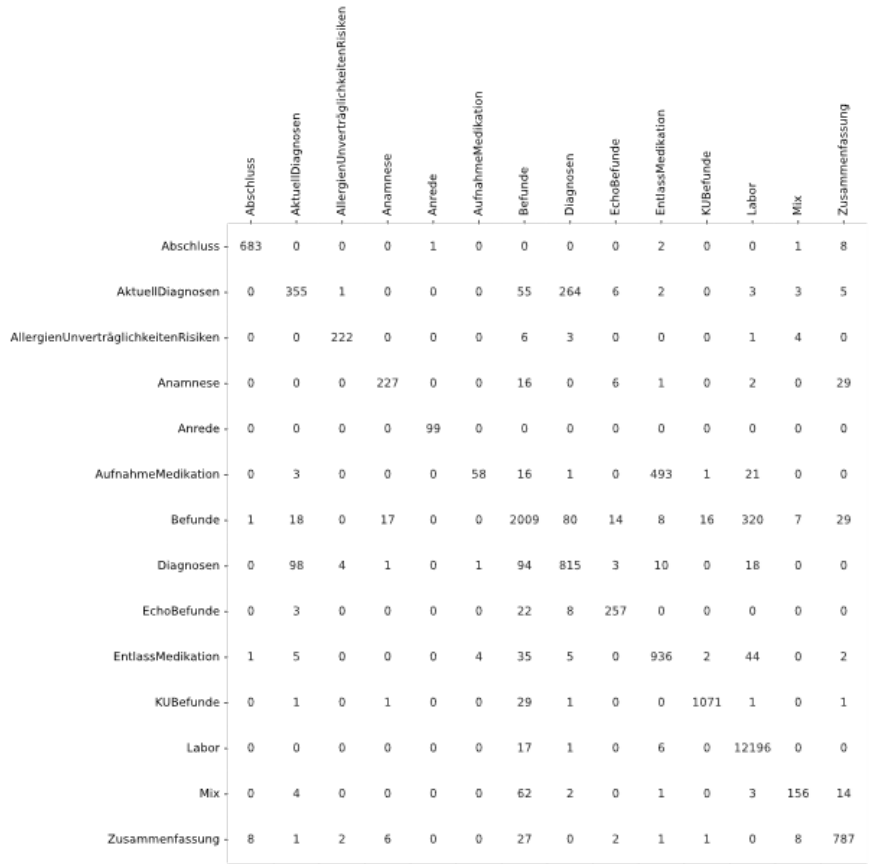


Fig. A.3 **Confusion matrix section classification SVM:** Confusion matrix of the SVM model (section classification).

A.4 GGPONC NER

A.4.1 Introduction

GGPONC NER was trained on high-quality SNOMED CT annotations of the GGPONC 2.0 corpus (Borchert et al. 2022). The class CLINICAL DRUG is a sub class of the SNOMED CT class SUBSTANCE and semantically overlaps with our DRUG/ACTIVEING class. However, CLINICAL DRUG is more general and covers a broader range of medication information, including FREQUENCY and STRENGTH.

Token	CARDIO:DE Gold Annotations	GGPONC NER Predictions
Prävastatin	ACTIVEING	CLINICAL DRUG
20	STRENGTH	CLINICAL DRUG
0	FREQUENCY	CLINICAL DRUG
-	FREQUENCY	CLINICAL DRUG
0	FREQUENCY	CLINICAL DRUG
-	FREQUENCY	CLINICAL DRUG
1	FREQUENCY	CLINICAL DRUG

Table A.5 **CARDIO:DE vs. GGPONC annotation schema:** Annotated snippet of a CARDIO:DE doctor's letter containing medication information, showing CARDIO:DE gold-standard annotations and GGPONC NER predictions.

As illustrated in Table A.5 GGPONC NER annotates the whole text snippet with CLINICAL DRUG. This indicates, that information like FREQUENCY and STRENGTH are part of CLINICAL DRUG. Therefore, we evaluated two types of mappings from our CARDIO:DE labels to CLINICAL DRUG (Table A.6).

Mapping type	GGPONC	CARDIO:DE
Short	CLINICAL DRUG	DRUG/ACTIVEING
Long	CLINICAL DRUG	DRUG/ACTIVEING/FREQUENCY/STRENGTH

Table A.6 **Mapping types from CARDIO:DE vs. SNOMED CT:** SNOMED CT *Clinical Drug* to CARDIO:DE medication information classes.

During class mapping, we removed the IOB format of all class labels. The beginning and end of an entity are based on different assumptions due to different definitions of CLINICAL DRUG and our labels, thus this schema produced various annotation errors.

A.4.2 Evaluation

In our evaluations we renamed all mapped classes including CLINICAL DRUG consistently to DRUG. GGPONC NER was released in four versions. We show the results of the best performing model on our data: `04_ggponc_fine_long` (Table 3-4).

Class	Precision	Recall	F_1-score	Support
DRUG	0.13	0.67	0.21	2,128

Table A.7 **GGPONC Results short mapping:** Precision, recall, and F_1-score for the DRUG class (short mapping, `04_ggponc_fine_long`).

Class	Precision	Recall	F_1-score	Support
DRUG	0.81	0.80	0.80	11,291

Table A.8 **GGPONC Results long mapping:** Precision, recall, and F_1-score for the DRUG class (long mapping, `04_ggponc_fine_long`).

Results of the short mapping show a low precision, while recall achieved 67%. This indicates a large amount of false positive predictions. The long mapping could clearly improve both precision and recall scores. In-depth analysis confirmed our assumptions, that the CLINICAL DRUG class of SNOMED CT covers our FREQUENCY and STRENGTH classes, too.

We further investigated frequently appearing false positive predictions of both models (Table A.9, A.10 and A.11).

Token	CARDIO:DE Annotation	GGPONC NER Prediction
Albumin	O	CLINICAL DRUG

Table A.9 **GGPONC analysis laboratory values:** Laboratory values are annotated as CLINICAL DRUG by GGPONC NER.

Token	CARDIO:DE Annotation	GGPONC NER Prediction
Septal	O	CLINICAL DRUG
Occluder	O	CLINICAL DRUG

Table A.10 **GGPONC analysis technical values:** Technical devices are annotated as CLINICAL DRUG by GGPONC NER. In GGPONC, cardiovascular devices are rarely mentioned. The model recognizes this term as medical, but classifies it to the wrong class.

Token	CARDIO:DE Annotation	GGPONC NER Prediction
Preiswertere	O	CLINICAL DRUG
,	O	CLINICAL DRUG
wirkstoffgleiche	O	CLINICAL DRUG
Präparate	O	CLINICAL DRUG

Table A.11 **GGPONC analysis generic medication mentions:** More generic text sequences about medication are annotated as CLINICAL DRUG by GGPONC NER. The whole context in the doctor's letter is: *Selbstverständlich können auch preiswertere wirkstoffgleiche Präparate anderer Hersteller verwendet werden*. In CARDIO:DE guidelines we only annotate medication information, where the patient is the experiencer. This might be due to the text type of GGPONC, guidelines. They often speak in very abstract terms about medication information.

Frequent false negatives of both models were ACTIVEING like: *ASS, Clopidogrel* or *Vitamin D*. In addition medication mentions like *Panzytrat* or *Tromcardin* were frequently not recognized as DRUG. In Table A.12 and A.13 we show results of the `02_ggponc_fine_short` model. Details, see (Borchert et al. 2022). Further investigations and experiments to compare both model types we leave for future work.

	precision	recall	f1-score	support
DRUG	0.61	0.77	0.68	2128

Table A.12 **Results ggponc short model and short mapping:** Precision, recall and F_1-score for the DRUG class (short mapping, `02_ggponc_fine_short`).

	precision	recall	f1-score	support
DRUG	0.61	0.15	0.23	11291

Table A.13 **Results ggponc short model and long mapping:** Precision, recall and F_1-score for the DRUG class (long mapping, `02_ggponc_fine_short`).

Appendix B

Clinical Section Classification using Pretrained Language Models and Prompting

B.1 Ablation tests

B.1.1 Comparing to `medbertde`

While `gbert-base-comb nocontext` was the overall best-performing model in our core experiments, the publicly available `medbertde-base nocontext` pretrained on medical data from scratch achieved superior results in frequent scenarios. Hence, we assessed, how `medbertde-base context` performs in comparison to our final model `gbert-large-comb context`. We trained a `medbertde-base context` model without further pretraining, as further pretraining did not show a consistent performance improvement in the core experiments. For 20 shots `medbertde-base context` achieved statistically significant better accuracy results than `gbert-large-comb context` (86.2% vs. 84.3%). Performance differences for 50 and 100 shots are not significant, while using 400 shots, `gbert-large-comb context` achieves better results (93.4% vs. 92.4%) (cf. Table B.4).

Primary classes With regard to the primary classes, the F_1-score of `gbert-large-comb context` is significantly better than `medbertde-base context` for the ANAMNESE class with mean +4.2 percentage points. This supports our hypothesis, that larger PLMs are superior on complex free text section classes (cf. Section 5.7). To a lesser extent, but significantly, `medbertde-base context` achieves better F_1-scores for the MEDIKATION class. (Figure B.14)

B.1.2 Inspecting `[SEP]` recognition

We observed significant performance drops for classes such as: ALLERGIENUNVERTRÄGLICHKEITENRISIKEN, ANREDE and MIX. To gain better understanding of this decline, (1) we performed *fine-grained class analysis* for samples from ANREDE and (2) *analyzed Shapley values*. (1) We found that precision dropped from 98.2% to 48.1%. 99 out of 131 instances were misclassified as DIAGNOSEN. Even if we use 400 training shots, the `gbert-base-comb context` model still achieves a low precision rate (56.8%). This can only be improved to 68.3% using a `gbert-large-comb context` model. Both precision scores are significantly below *nocontext* models with a precision of 98.2%.(2) Shapley values shed further light on typical patterns of this section samples. The three classes ALLERGIENUNVERTRÄGLICHKEITENRISIKEN, ANREDE, MIX typically contain only a single paragraph or sentence. ANREDE paragraph are typically followed by a DIAGNOSEN paragraph, containing section headers such as *Aktuelle Diagnosen:*.

Hence, to test the ability of PET models to recognize, that the sample to classify is between the two `[SEP]` token, we created nine artificial test samples by combining context paragraphs that are atypical for our dataset, as presented in Figure B.15.
If we use a `gbert-base-comb context` model trained on 20 shots, the first sample in Figure B.15 is still incorrectly classified with 97% as ANREDE. In contrast, the second sample is correctly classified with 99% accuracy as MEDIKATION.
Overall, 5/9 samples where still incorrectly classified as ANREDE class.
We investigated another section class such as ALLERGIENUNVERTRÄGLICHKEITENRISIKEN, which typically only contains a single paragraph, too, we observed a similar behavior. Often the context models incorrectly classify samples from the previous section DIAGNOSEN as ALLERGIENUNVERTRÄGLICHKEITENRISIKEN. E.g. *Z.n. Bandscheibenvorfall 11.09.1941 [SEP] - Z.n. Hodentorsion 11.09.1941 [SEP] Kardiovaskuläre Risikofaktoren: Arterielle Hypertonie, Hypercholesterinämie, positive Familienanamnese, Nikotinanamnese: nie (English: History of disc prolapse on September 11, 1941 [SEP] - History of testicular torsion on September 11, 1941 [SEP] Cardiovascular risk factors: arterial hypertension, hypercholesterolemia, positive family history, smoking history: never.)*. These results raised the question: Are there often misclassifications at the first or final paragraph of a section class? The confusion matrix (Figure B.16) shows that typical false positives involve such patterns. For example:

- DIAGNOSEN often misclassified as ANREDE
- DIAGNOSEN and BEFUNDE often misclassifed as ALLERGIENUNVERTRÄGLICHKEITENRISIKEN
- MEDIKATION and ZUSAMMENFASSUNG often misclassifed as ABSCHLUSS

This reveals, that contextualizing paragraphs can harm classification results for certain section classes. This is especially relevant for section classes, which usually contain single-paragraph samples. This suggests that in a few-shot learning scenario, smaller PLMs can have difficulty distinguishing testing instances from contexts, and hence do not sufficiently focus on the instances themselves.

B.1.3 Removing section titles from data

We identified that our best performing model from the core experiments `gbert-base-comb nocontext` using 20 shots frequently misclassified samples containing section titles of our primary classes. 32% of the false negative samples of the `Medikation` class contained either the text sequence *Medikation bei Aufnahme: (English: Medication on admission)* or

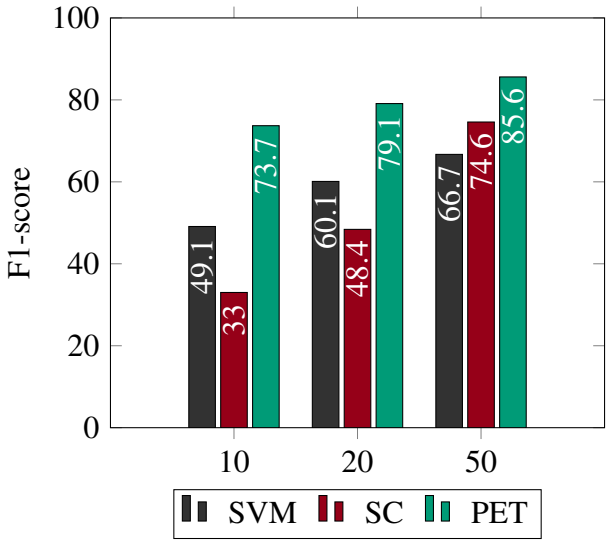
Medikation bei Entlassung: (English: Medication at discharge). A similar classification error we observed for the ANAMNESE class: 81% of false negatives contain the text sequence *Anamnese*. While in the training samples for MEDIKATION we did not identify any section titles, there was a single title *Anamnese:* in the training set of *Anamnese*.
Adding context and increasing model size could significantly avoid these kind of errors, still 5% of the false negatives of the MEDIKATION class of our final model `gbert-large-comb context` contained these kind of text sequences. Hence, we trained our final model `gbert-large-comb context` on a modified training and test set, filtered by a list of the most common section titles (Figure B.17). In Table B.5 shows, that accuracy could be increased over all few-shot sizes by approximately 2%. Figure B.18 shows, that both primary classes can improve F_1-scores for 20 and 50 shots. In contrast, the models trained on the full training set, slightly decrease in performance. This is not surprising, as in contrast to the few-shot sets, the full training set frequently contains section titles.
However, it is important to note that these results can not be compared directly to the experimental results with included section titles, since we modified the training and test data set. Considering experimental limitations in clinical routine, it may be beneficial to avoid the use of section titles as they can be often well identified through manual patterns and heuristics. This approach is especially relevant if only smaller PLMs are employed with strong sequence length restrictions due to limited resources.

B.1.4 Classifying *nocontext* samples using a *context* model

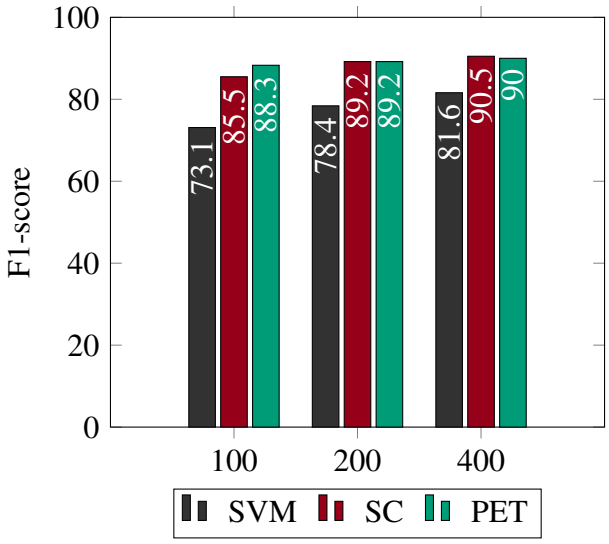
In Figure B.19 we show Shapley values of a sample without further context with the gold label ZUSAMMENFASSUNG classified by (a) `gbert-base-comb context` and (b) the `gbert-large-comb context` model. `Gbert-base-comb context` shows very similar token contributions with respect to ZUSAMMENFASSUNG as `gbert-base-comb nocontext` in Figure B.8a. But both base models incorrectly classify the sample.
In contrast, `gbert-large-comb context`, correctly assigns the ZUSAMMENFASSUNG class with a probability of 79%. Adding context paragraphs increases this to 99% (see Figure B.13b). Interestingly, most of the input token positively contribute to the correct class, with the exception of AUFNAHME. This is expected, as this token frequently negatively contributed to ZUSAMMENFASSUNG in various experimental setups.

B.2 Baseline - support vector machine

While deep learning methods became the SOTA for text classification tasks, statistical machine learning approaches such as SVM remain highly prevalent (Pomares-Quimbaya et al. 2019). Therefore, we trained a SVM for our core experimental setup to compare its performance to our neural SC baseline and our best performing PET approach in the core experiments `gbert-base-comb nocontext` (Figure B.1).
PET is always outperforming both SVM and SC for all shot sizes. Only if shot size is ≤ 20 the SVM outperforms our neural SC baseline model. If shot sizes are ≥ 50 SC and PET consistently outperform the SVM. We used the `LinearSVC` implementation and `TfidfVectorizer` for text encoding, both with default hyperparameters, as implemented in `scikit-learn` version 1.0.2. (Buitinck et al. 2013).



(a) Comparing SVM, SC and PET using few-shot sets: 10, 20, 50



(b) Comparing SVM, SC and PET using few-shot sets: 10, 20, 50

Fig. B.1 **Baseline comparison SVM, SC and PET:** Comparing model performance using core experimental setup. Comparing Support Vector Machine (SVM), BERT with a sequence classification head (SC) and PET.

B.3 Hyperparameters

Further pretraining We applied the following hyperparameters for pretraining experiments described in Section 5.3.2: vocabulary size: 30,000; maximum sequence length: 512;

1. task-adaptation:

- data: CARDIO:DE corpus
- epochs: 100, batch size: 24, fp16: True, gradient accumulation steps: 4
- 1×RTX6000 graphics processing unit (GPU) with 24GB video random access memory (VRAM)
- Training time: $\sim$ 2h

2. domain-adaptation:
   - data: 179,000 German doctor's letters + GGPONC
   - epochs: 3, batch size: 16, fp16: True, gradient accumulation steps: 1
   - 2×RTX6000 GPUs with each 24GB VRAM
   - Training time: $\sim$ 17h
3. combined:
   - data: 179,000 German doctor's letters + GGPONC + CARDIO:DE corpus
   - epochs: epochs: 100, batch size: 24, fp16: True, gradient accumulation steps: 4
   - 1×RTX6000 GPU with 24 GB VRAM
   - Training time: additional $\sim$ 2h to domain-adaptation

PET and SC experiments :

- All PET experiments were conducted on a single NVIDIA A40 GPU with 40GB VRAM. However, we also conducted PET experiments on NVIDIA P4 with 8GB VRAM using BERT-base models by only reducing evaluation batch size at inference time.
- Hyperparameters PET and SC: BERT-base models: training batch size 4, evaluation batch size: 64; BERT-large models: training batch size 4, evaluation batch size 16.
- Each experiment conducted with three different training sets and two random seeds (in total six setups). To increase comparability, we always selected models trained on training set 3 and with random seed 123 to investigate Shapley values.

B.4 Additional figures and tables

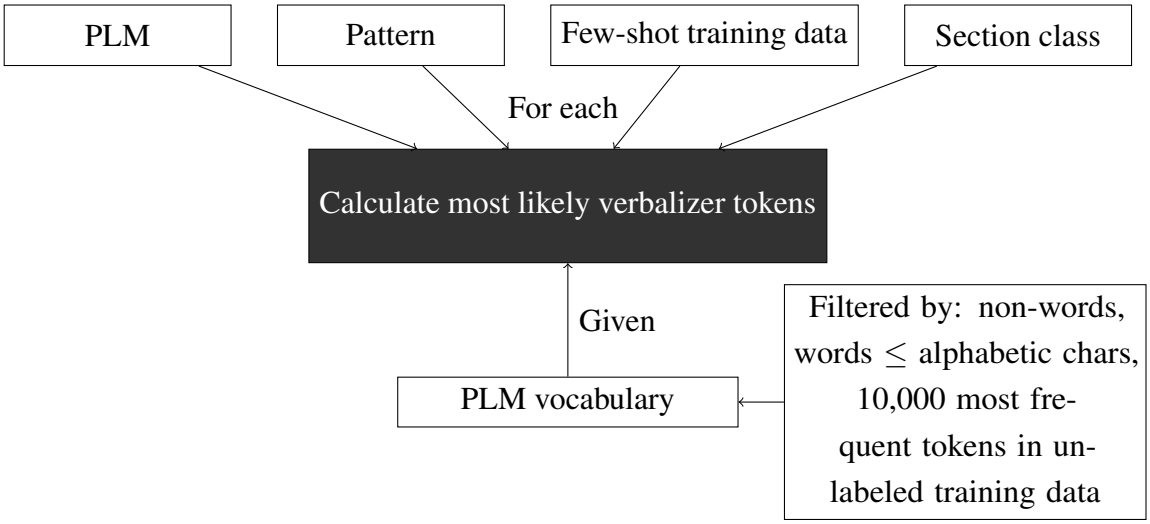


Fig. B.2 **The PETAL workflow:** PETAL calculates the most likely verbalizer token per label for each (1) PLM, (2) prompt pattern, (3) few-shot training set. The verbalizer token must be part of the PLM's vocabulary.

```
| - 10shots/
|   | - set_1.csv
|   | - set_2.csv
|   | - set_3.csv
|   | - unlabeled_1.csv
|   | - unlabeled_2.csv
|   | - unlabeled_3.csv
| - holdout/
|   | - full_holdout.csv
```

Fig. B.3 **PET few-shot data:** Example folder structure for the 10-shot data set including the heldout data set.

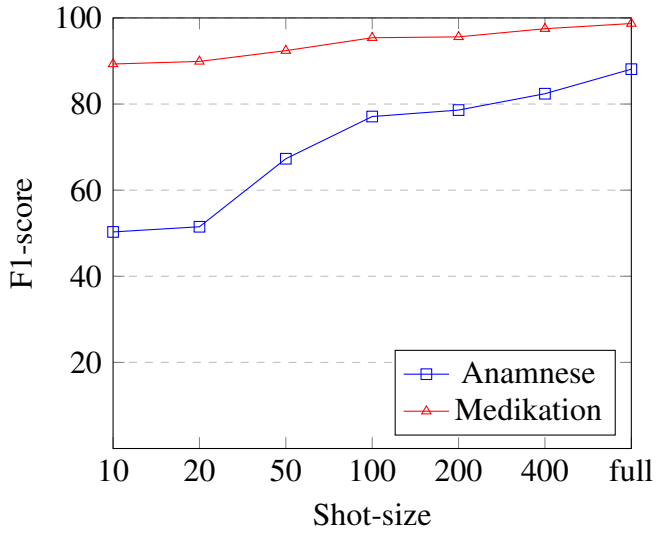


Fig. B.4 **PET core experiments:** Primary class F_1-score for all shot sizes. F_1-score per few-shot sizes for primary classes with no context using `gbert-base-comb nocontext`.

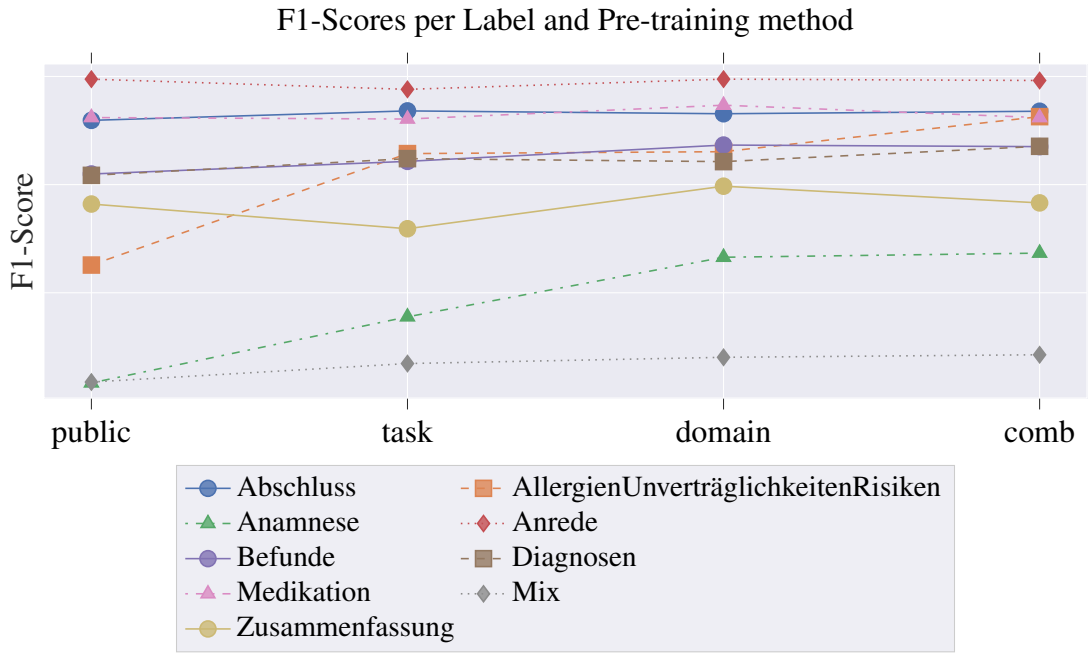


Fig. B.5 **Analyzing pre-training impact per label:** F_1-scores per label per pre-training method using *gbert-base nocontext*.

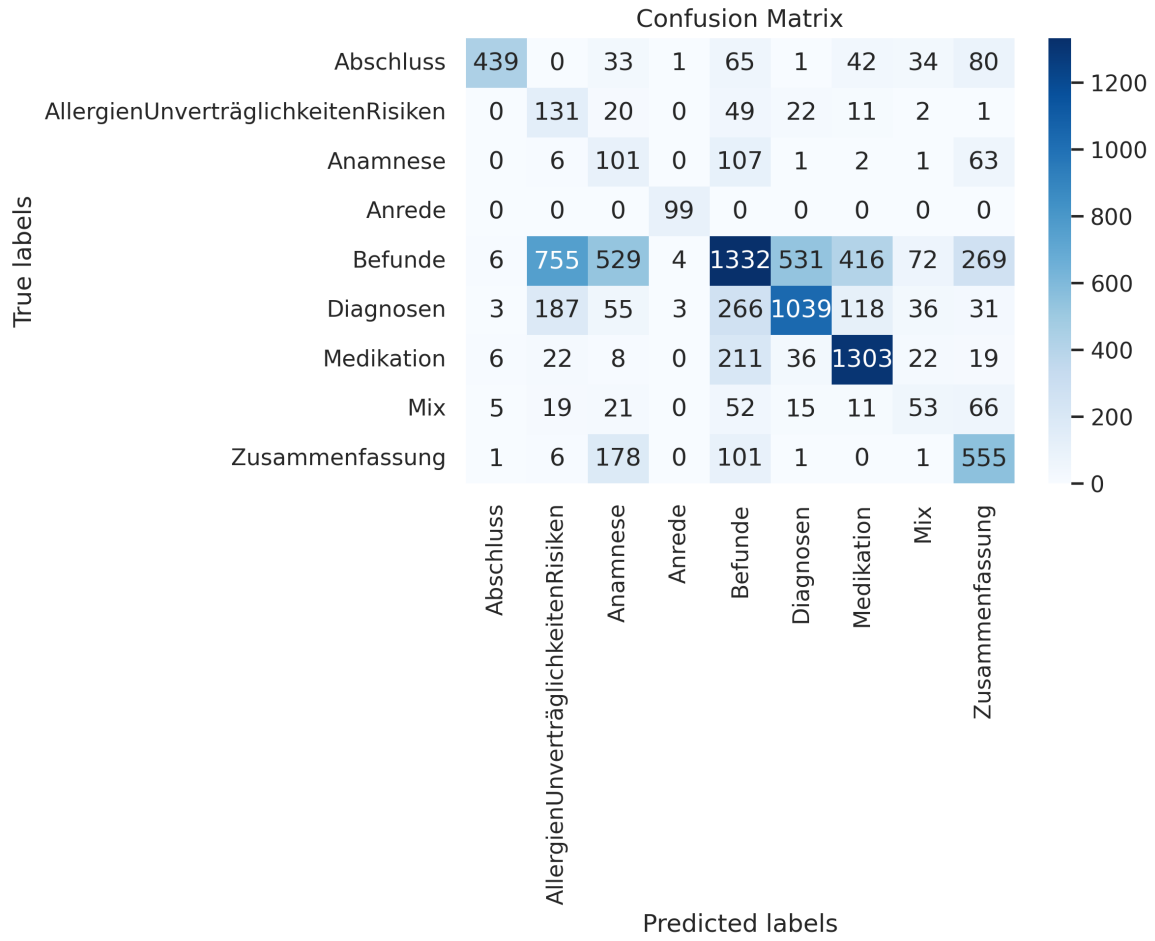
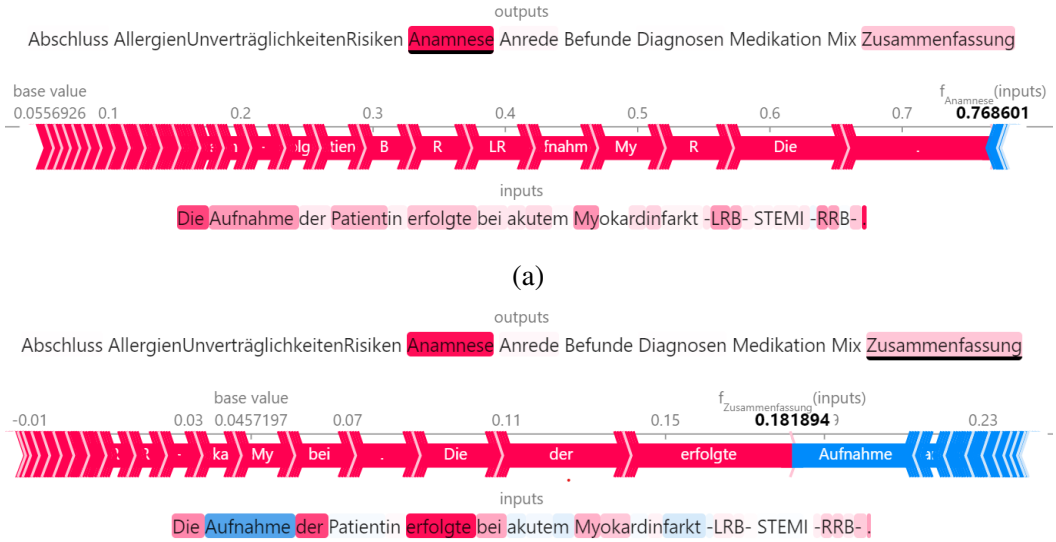


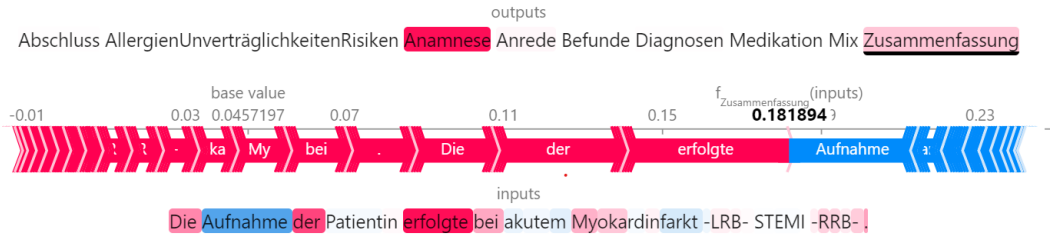
Fig. B.6 **Confusion matrix for** `gbert-base`**:** Model trained on 20 shots on training set 3 with initial seed 123.



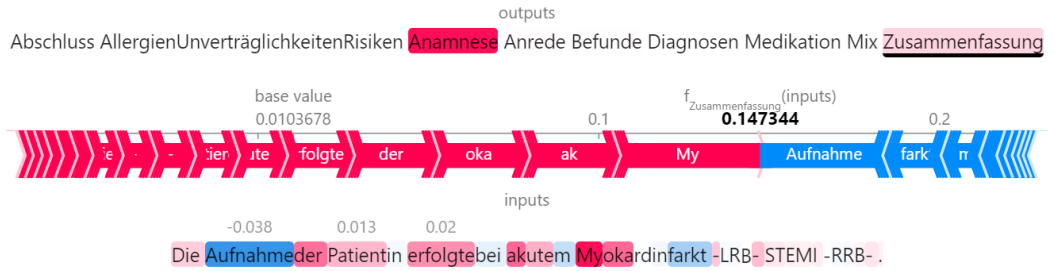
(a)

(b)

Fig. B.7 **Shapley values base nocontext:** Shapley values for `gbert-base-comb nocontext` for predicted class comparing (a) ANAMNESE and (b) ZUSAMMENFASSUNG using 20 training shots. Shapley values with respect to predicted label (underlined). Shapley values per sub tokens. Legend: Red: positive contribution, Blue: negative contribution.



(a)



(b)

Fig. B.8 **Shapley values large nocontext:** Shapley values for predicted class ZUSAMMENFASSUNG comparing (a) `gbert-base-comb nocontext` and (b) `gbert-large-comb nocontext` with 20 training shots. Shapley values with respect to predicted label (underlined). Shapley values per sub tokens.
Legend: Red: positive contribution, Blue: negative contribution.

	Training set	Test set
ANREDE *(Salutation/Greeting)*	402	99
AKTUELLDIAGNOSEN *(Current Diagnosis)*	3,298	694
DIAGNOSEN *(Diagnosis)*	4,725	1,044
ALLERGIEN *(Allergies)*	1,031	236
ANAMNESE *(Patient Medical History)*	1,188	281
AUFNAHMEMEDIKATION *(Admission Medication)*	2,058	593
KUBEFUNDE *(Body Findings)*	4,194	1,105
BEFUNDE *(Findings)*	9,636	2,519
ECHOBEFUNDE *(Echocardiogram Findings)*	1,566	290
LABOR *(Laboratory)*	55,420	12,220
ZUSAMMENFASSUNG *(Summary)*	3,645	843
MIX *(Mix)*	945	242
ENTLASSMEDIKATION *(Discharge Medication)*	4,090	1,034
ABSCHLUSS *(Closing Remarks)*	2,805	695
Total	95,003	21,895

Table B.1 **Statistics section classes CARDIO:DE:** Number of samples per section class per CARDIO:DE corpus split. English translations in round brackets.

PLM	Pretrained	Method	Few-shots
gbert-base	public task domain comb	PET&SC	10, 20, 50, 100, 200, 400
medbertde-base	public task domain comb		

Table B.2 **Setup for core experiments:** Experimental overview for our core experiments including PLMs, pretraining method, learning method and few-shot sizes.

Shot size	All templates	Null prompt templates
20	79.1	78.2
50	85.6	84.6
100	88.3	88.5
400	89.7	90
full (SC)	96.7	

Table B.3 **Results null prompts:** Accuracy scores for `gbert-base-comb nocontext` PLMs using all templates or null prompts on four few-shot sizes.

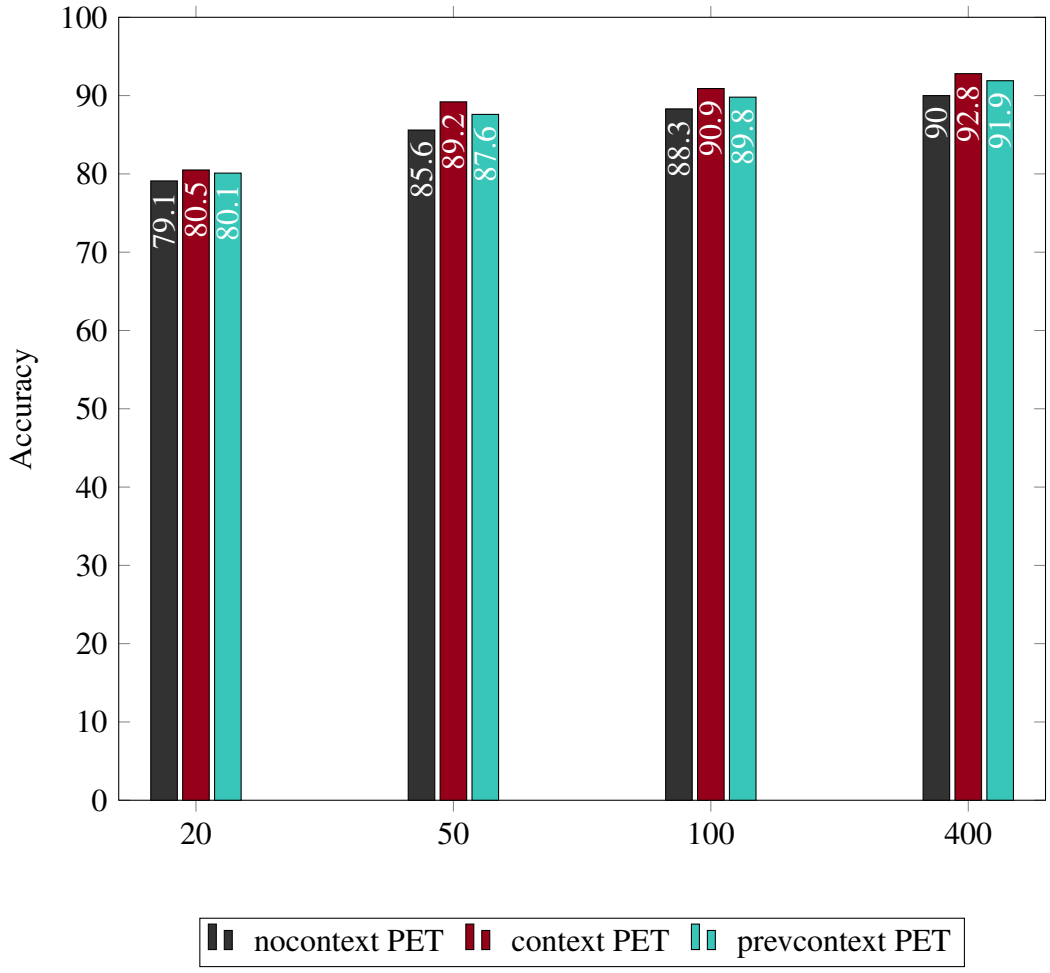


Fig. B.9 **Results context types:** Accuracy scores for different context types: (1) no context, (2) context, (3) prevcontext and few-shot sizes 20, 50, 100 and 400 using PET.

Shot size	gbert-large-comb	medbertde-base
20	84.3	86.2
50	89.4	90.3
100	91.3	91.3
400	93.4	92.4

Table B.4 **Results gbert-large-comb context vs. medbertde-base context:** Comparing accuracy of both models trained with all templates.

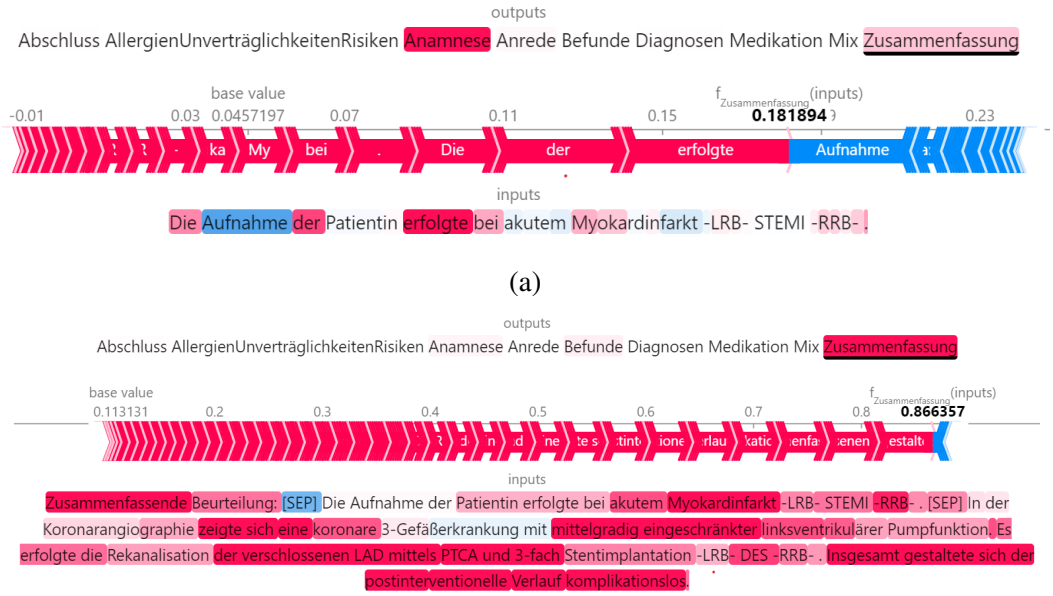


Fig. B.10 **Shapley values adding context:** Shapley values for `gbert-base-comb context` for predicted class ZUSAMMENFASSUNG comparing (a) `gbert-base nocontext` and (b) `gbert-base context` with 20 training shots. Shapley values with respect to predicted label (underlined). Shapley values per sub tokens.
Legend: Red: positive contribution, Blue: negative contribution.

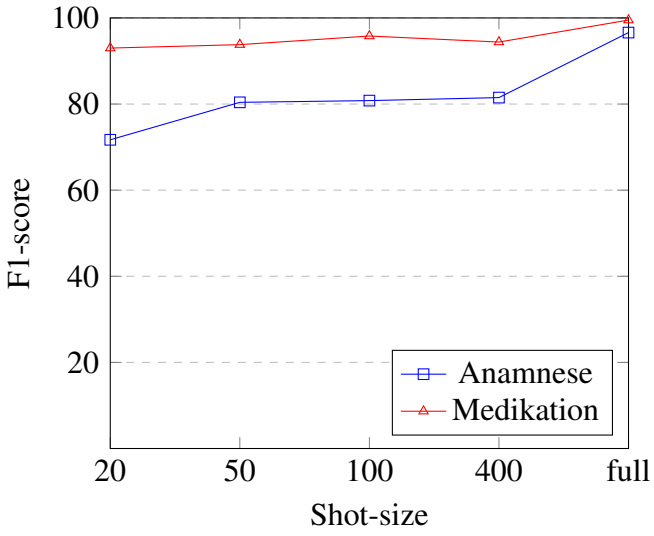


Fig. B.11 **Results additional experiments:** Primary class F_1-score for all shot sizes: Accuracy scores per few-shot sizes for primary classes using `gbert-large-comb context`.

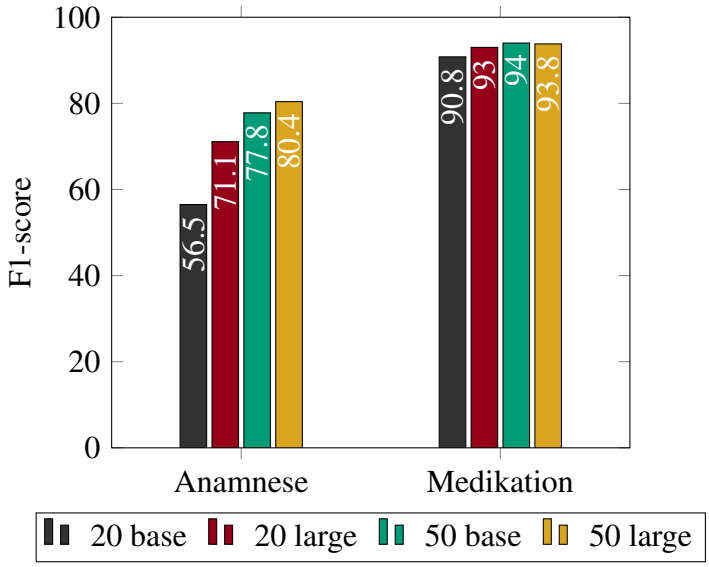


Fig. B.12 **Combining best performing methods:** comparing accuracy scores for `gbert-large-comb context` vs. `gbert-base-comb context` with all templates on two few-shot sizes for primary classes.

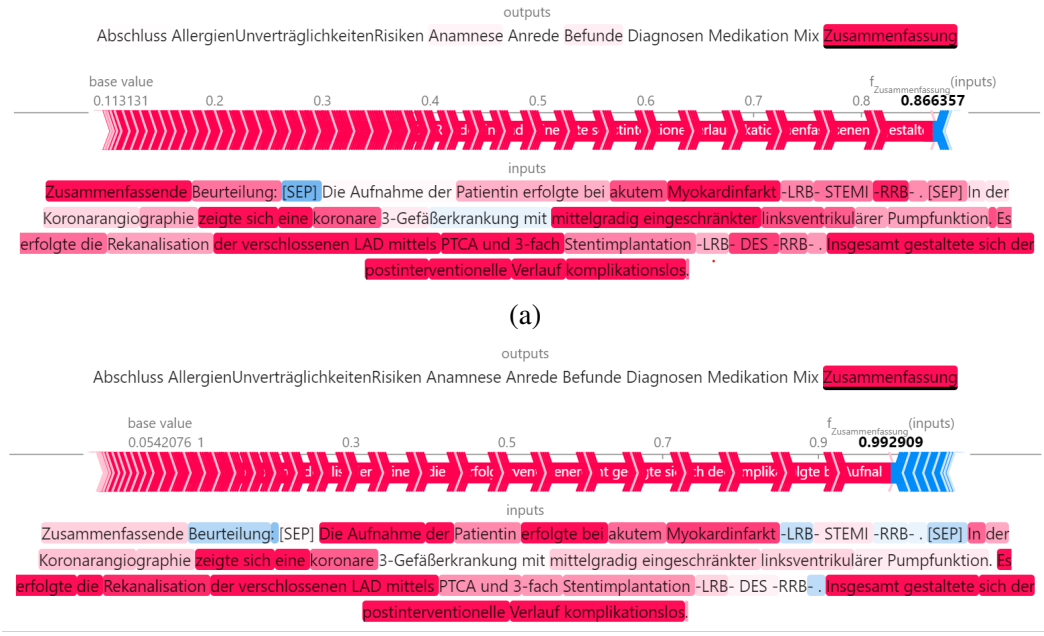


Fig. B.13 **Shapley values final model:** Shapley values for `gbert-large-comb`, context for predicted class ZUSAMMENFASSUNG comparing (a) `gbert-base context` and (b) `gbert-large context` with 20 training shots. Shapley values with respect to predicted label (underlined). Shapley values per sub tokens.
Legend: Red: positive contribution, Blue: negative contribution.

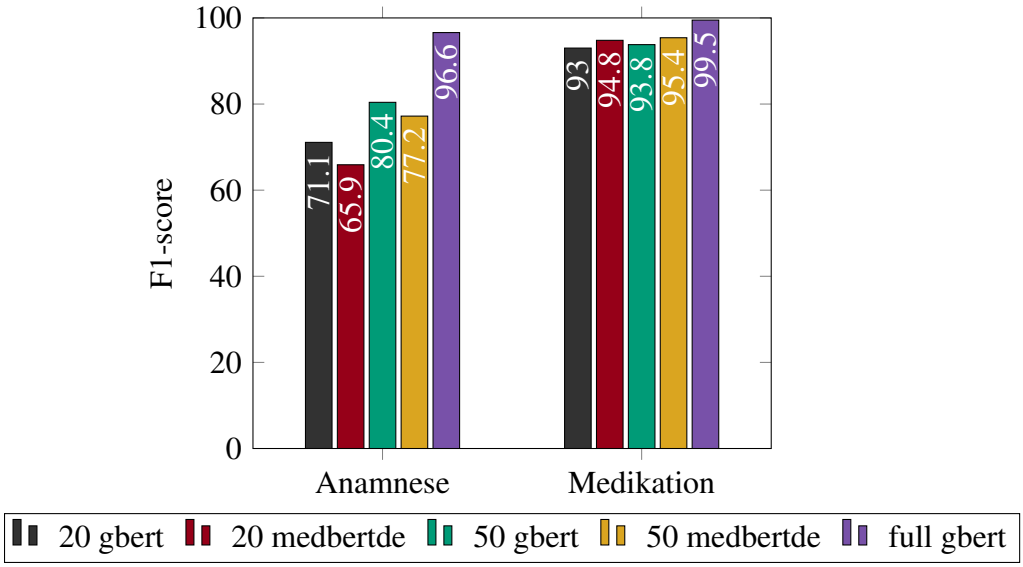


Fig. B.14 **Results gbert-large-comb context vs. medbertde-base context primary classes:** Comparing `gbert-large-comb context` and `medbertde-base context` trained with all templates. F1-score per primary label.

[SEP] über Ihre Patientin Frau Martina Mustermann geboren am 12.12.1999 wh. 2000 Musterstadt Musterstr. 1 die sich am in unserer Ambulanz vorstellte. [SEP] **Anamnese:**

English: [SEP] regarding your patient Mrs. Martina Mustermann, born on December 12, 1999, residing at Musterstr. 1, 2000 Musterstadt, who presented herself at our outpatient clinic. [SEP] Patient Medical History:

[SEP] über Ihre Patientin Herr Max Mustermann geboren am 12.12.1999 wh. 2000 Musterstadt Musterstr. 1 die sich am in unserer Ambulanz vorstellte. [SEP] **Medikation:**

English: [SEP] regarding your patient Mrs. Martina Mustermann, born on December 12, 1999, residing at Musterstr. 1, 2000 Musterstadt, who presented herself at our outpatient clinic. [SEP] Medication:

Fig. B.15 **Ablation tests context:** Two artificial training samples including English translation with atypical co-occuring context paragraphs. In the first sample, the section title ANAMNESE follows immediately after a ANREDE sample. In the second example, MEDIKATION follows after a ANREDE sample. Usually ANREDE is followed by DIAGNOSE, rarely by ANAMNESE and never in our data set by MEDIKATION.

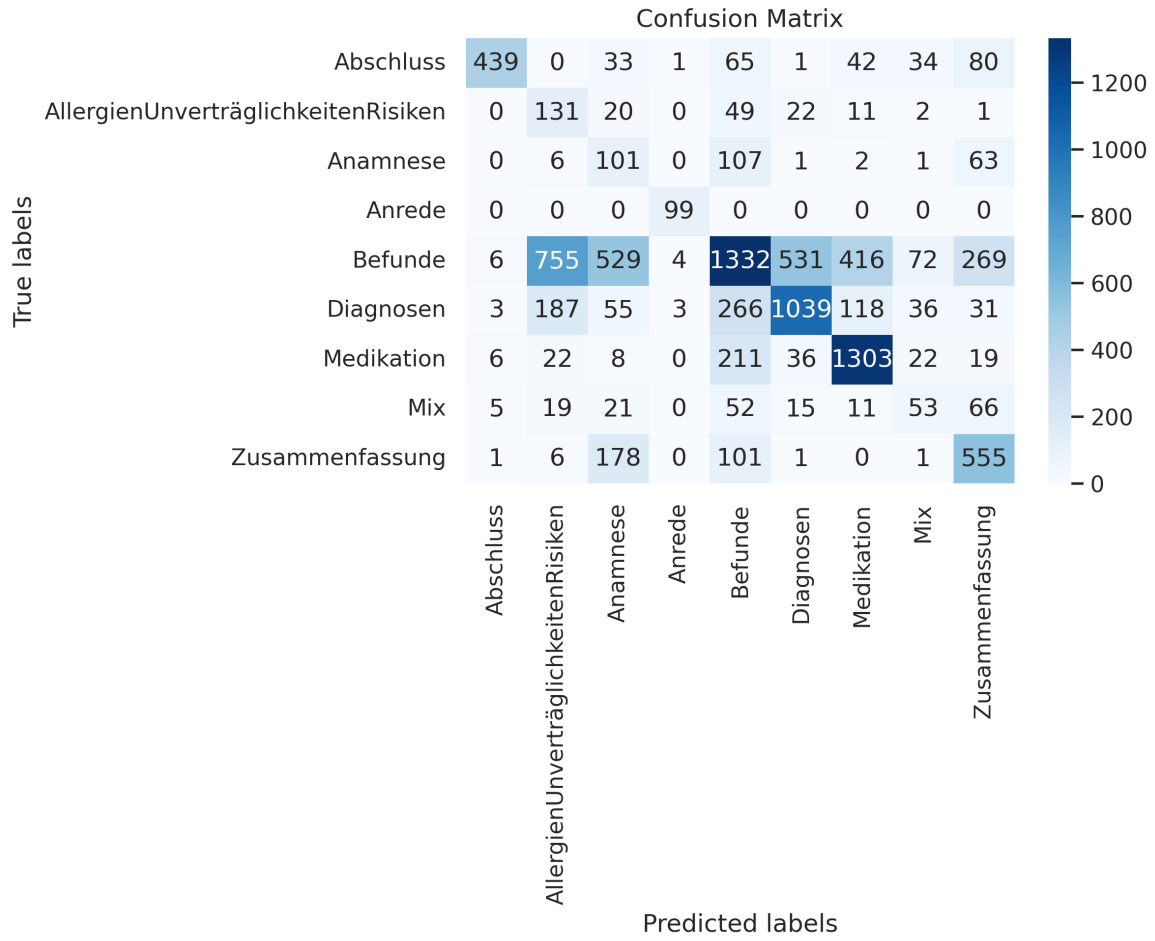


Fig. B.16 **Confusion matrix for `gbert-base-comb context`:** Confusion matrix for `gbert-base-comb context` trained on 20 shots on training set 3 with initial seed 123.

Anamnese:, Diagnosen:, Zusammenfassung:, Körperlicher Untersuchungsbefund:, Aktuell:, Labor:, Ruhe-EKG:, Procedere:, Medikation bei Aufnahme:, Therapieempfehlung:, Beurteilung:, Therapieempfehlung -LRB- von kardiologischer Seite -RRB- :, Befund und Beurteilung:, Transthorakale Echokardiographie:, Maßnahmen:, Befund:, Aktuelle Medikation:, Beurteilender Abschnitt:, Beschreibender Abschnitt:, Zusammenfassende Beurteilung:, Belastungs-EKG:, Kultureller Befund:, Medikation bei Entlassung:, Lokalbefund:, Körperliche Untersuchung:, Allergien:, Ruhe-EKG bei Aufnahme:, Oral:, Kardiovaskuläre Risikofaktoren:, Nächster Termin/Kontrolle:, Procedere/Termine:, Aktuelle Therapie:, Diagnose:, Echokardiographie:, Bisherige Medikation:, Farbduplexsonographie der Gefäße der rechten Leiste:, Kapilläre Blutgasanalyse:, Sonstige Diagnosen:, Therapieempfehlung von kardiologischer Seite:, Nächster Termin/Prozedere:, Befund/Zusammenfassung:, Kommentar:, Indikation für stationären Herzkatheter:, EKG:, Spirometrie:, Wichtig:, Medikation:, Langzeit-EKG vom B-DATE:, Aktuelle/Bisherige Medikation:, Befund / Zusammenfassung:

Fig. B.17 **List of most common section titles:** We generated this list by filtering the data set by short sequences including a single ":" at the end.

Shot size	PET including section titles	PET excl. section titles
20	84.3	86.7
50	89.4	91.1
100	91.3	93.4
400	93.4	95.8

Table B.5 **Ablation test results section titles:** F_1-score results using `gbert-large-comb context` trained with and without section titles in training and test data.

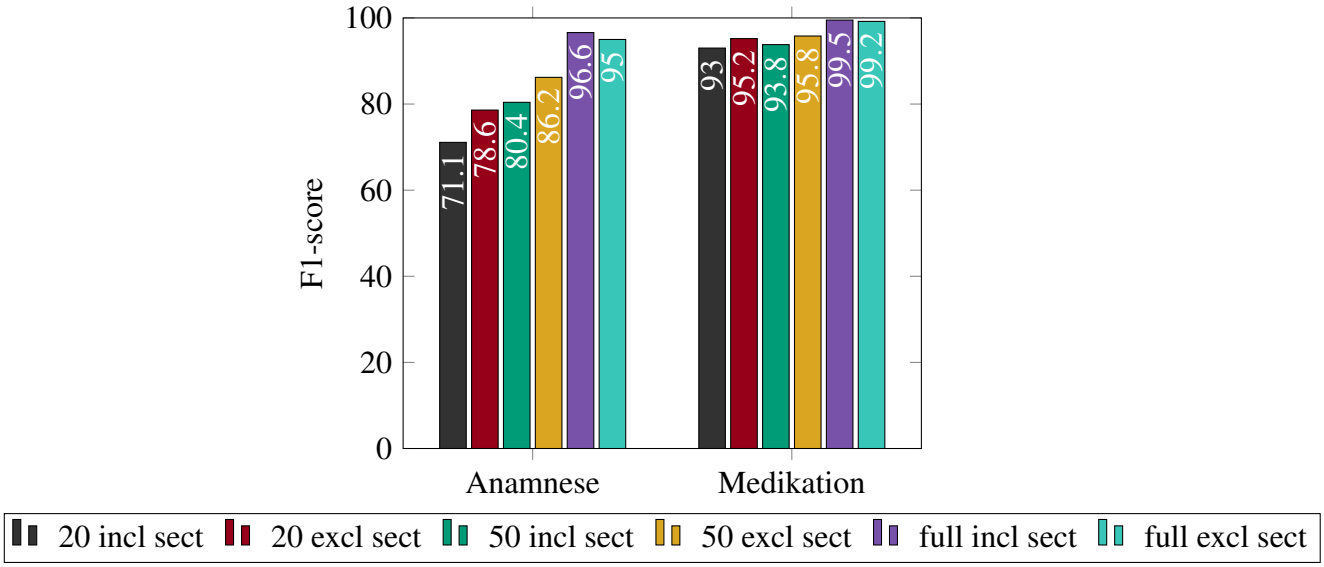
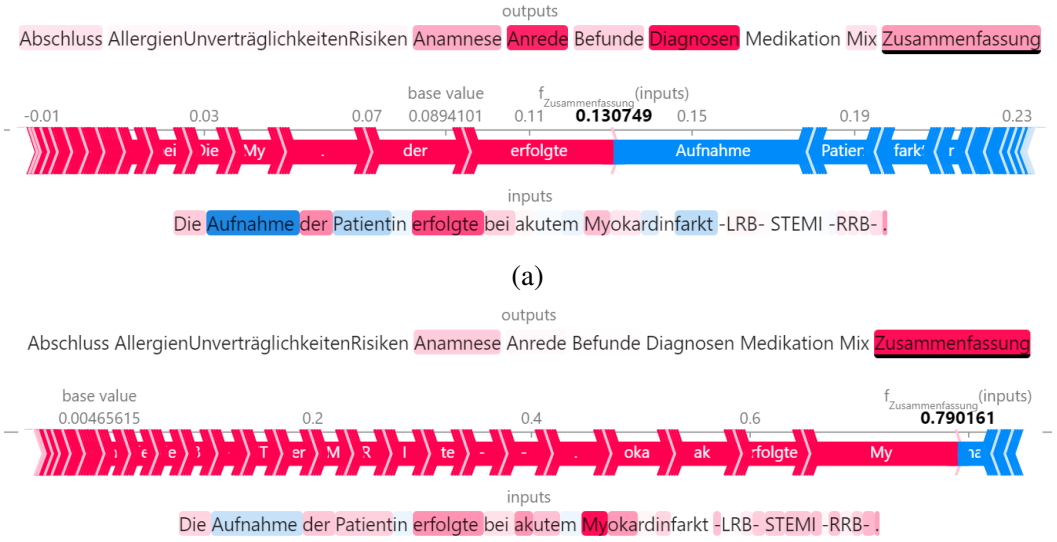


Fig. B.18 **Ablation test results section titles for primary classes:** Comparing accuracy scores for `gbert-large-comb context` including and excluding section titles. For reference we show results for SC model trained on full training sets for both scenarios.



(a)

(b)

Fig. B.19 **Ablation test Shapley values using context model on nocontext sample:** Comparing Shapley values for (a) `gbert-base-comb` context model vs. (b) `gbert-large-comb` context model using a sample without context. Shapley values with respect to predicted label (underlined). Shapley values per sub tokens. Legend: Red: positive contribution, Blue: negative contribution.

Appendix C

Medication Information Extraction using Local Large Language Models

Entity class	Full	Training	Test
DRUG	26,800	16,225	10,575
STRENGTH	10,921	6,691	4,230
FORM	11,010	6,651	4,359
DOSAGE	6,902	4,221	2,681
FREQUENCY	10,293	6,281	4,012
ROUTE	8,989	5,476	3,513
DURATION	970	592	378
REASON	6,400	3,855	2,545
ADE	1,584	959	625
Total	83,869	50,951	32,918

Table C.1 **N2C2 2018 (track 2) corpus:** Statistics about annotated entity classes.

Relation class	Full	Training	Test
DRUG–STRENGTH	10,946	6,702	4,244
DRUG–FORM	11,028	6,654	4,374
DRUG–DOSAGE	6,920	4,225	2,695
DRUG–FREQUENCY	10,344	6,310	4,034
DRUG–ROUTE	9,084	5,538	3,546
DRUG–DURATION	1,069	643	426
DRUG–REASON	8,579	5,169	3,410
DRUG–ADE	1,840	1,107	733
Total	59,810	36,384	23,462

Table C.2 **N2C2 2018 (track 2) corpus:** Statistics about annotated relations.

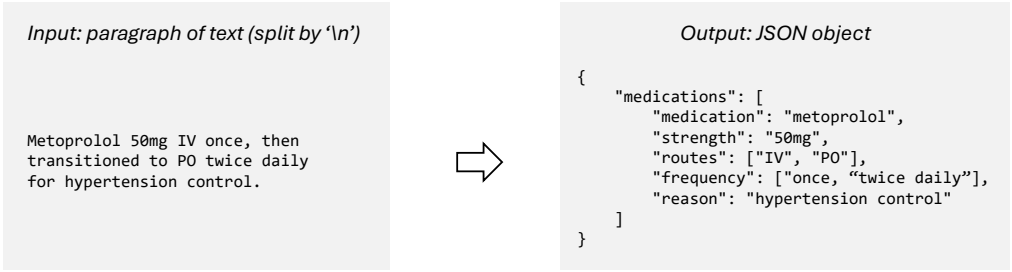


Fig. C.1 **Data pre-processing:** Converting a text snippet of a doctor's letter into a JSON string. A JSON object with a medication (entity) Metoprolol and the related information classes (e.g. strength) and their assigned entities (e.g. 50mg).

Newline split samples	**Merged sample**	**JSON**
Metoprolol to control hypertension\n 50 mg twice daily.	Metoprolol to control hypertension 50 mg twice daily.	{ "medications": [{ "medication": "metoprolol", "strength": "50mg", "routes": "", "frequency": "twice daily", "reason": "hypertension" }] }

Table C.3 **Example of a merged sample:** Strength and frequency information of Metoprolol is contained in a neighboring sample.

C.1 Data analysis

C.1.1 BRAT vs. JSON

The original N2C2 dataset comes in a standardized BRAT format (cf. Figure C.2):

Entities (spans)				
ID	Type	Start	End	Text
T1	Drug	1094	1101	Lipitor
T2	Frequency	1158	1163	q.i.d

Relations			
ID	Type	Arg1	Arg2
R1	Frequency-Drug	T1	T2

Fig. C.2 **Example dummy snippet of a n2c2 BRAT-formatted gold annotation:** Each annotated entity has a unique ID, an entity category, the exact character offsets, and the text span. Furthermore, it contains relation information between entities.

This format is designed for token-classification systems, such as encoder-based BERT models, which emit label information per input token. In contrast, recent LLMs generate a sequence of text conditioned on an input text. Manual evaluations revealed that LLMs consistently failed to generate accurate token-level character offset information, despite explicit instructions. Therefore, as discussed in Section 6.3.2, we utilize a flexible JSON format to represent both the medical entity and entity relation LLM predictions. However, due to (1) merging strategies and (2) missing character offset information, we could not use the official N2C2 evaluation script.

Character offsets Many instances of medication information appear multiple times in one sample. E.g. *The patient is discharged on ASS 100 mg daily for cardiovascular prophylaxis, with an additional prescription for ASS 100 mg as needed for acute pain management.* or equal relation information *6. Amlodipine 5 mg Tablet, Lisinopril 5 mg Tablet* . In both scenarios, implementing a rule-based mapping approach is often impossible without context or human judgment. For instance, when examining the corresponding JSON prediction: `{'medications',{'medication':'Amlodipine','ade':'','dosage':'','duration':'','form':'Tablet','frequency':'','reason':'','route':'','strength':'5mg'}}` it is unclear to which `strength` mention in the input text the `5 mg` value for `Amlodipine` corresponds. Motivated by our experiments using Shapley values we attempted to leverage interpretability to map output token to input token contributions as described

in Section 6.4.4. However, on a large data set this method was computationally intensive and yielded inaccurate results, particularly with longer input texts, a common pattern in this disambiguation task.

Due to the lack of character offset information, we investigated how many medication names and medication relation values in both datasets appear more than once per sample. These instances make it challenging to create a mapping script from BRAT to JSON (cf. Table C.4).

Entity class	**Repetitions (%)**
DRUG	1.70
ADE	0.73
DOSAGE	0.23
DURATION	0.41
FORM	0.10
FREQUENCY	0.35
REASON	5.07
ROUTE	0.70
STRENGTH	0.18

Table C.4 **Distribution of repeated medication information instances:** Repetitions within non-empty annotated gold samples (8,238 in total) of the N2C2 corpus.

Duplicates Our merging strategy, designed to capture all relevant related information for each medication per input sample, resulted in duplicate entries within out dataset (cf. Section 6.3.2). These must be removed before a BRAT conversion can take place. For a transparent and comparable setup, we would thus need to re-run all experiments using this data set.

Furthermore, there are no trivial copy and paste duplicates in our dataset: they arise because relation information of a medication is distributed over neighboring lines. Therefore, while there are duplicate annotations in the data, they always appear in different contexts. To quantify these, we counted all duplicate occurrences per medication relation class across the dataset (cf. Table C.5).

Entity class	BRAT	JSON	Percentage difference (%)
ADE	778	733	5.78
DOSAGE	2,722	2,695	0.99
DURATION	436	426	2.29
FORM	4,380	4,374	0.14
FREQUENCY	4,038	4,034	0.10
REASON	3,538	3,410	3.62
ROUTE	3,568	3,546	0.62
STRENGTH	4,258	4,244	0.33

Table C.5 **Instance count per relation information:** Comparing instance count per relation information between the BRAT-formatted and the JSON-formatted N2C2 dataset.

Due to these challenges implementing a perfect mapping script is time-consuming and complex. Recent work on information extraction therefore uses soft-matching or relaxed scoring strategies (Han et al. 2024; Hu et al. 2024a; Moral-González et al. 2025; Jiang et al. 2024), or combines these with LLMs-as-a-judge, comparable to our feedback LLM approach (Sharif et al. 2025). To provide at least a baseline comparison, we release an initial JSON to BRAT converter, based on heuristics. The shortcoming of this converter are reported in the Table C.6. Across the corpus the script fails to map 16.7% of all gold entities and 13.3% of all gold relations and introduces 7.4% additional entity artifacts.

	JSON	BRAT	missing/additional
Entities	34,507	31,040	5,760/2,293
Relations	23,718	20,561	3,157/0

Table C.6 **Statistics of the JSON to BRAT conversion:** The script produces 5,343 missing and 803 additional entities and 325 missing and 592 additional relations.

We could only find the official N2C2 2018 track 2 evaluation script for task 1 (NER) and task 2 (RE). We used this script to evaluate our best-performing Llama 70b FT model and report a lenient F_1-score for NER of 87% and 68% for RE (cf. Table C.7). We could not find any official or community script for the end-to-end task to produce comparable results to our task. We therefore can not supply these results using our converted data.

Entity class	NER	RE
DRUG	0.95	-
ADE	0.61	0.45
DOSAGE	0.79	0.65
DURATION	0.85	0.73
FORM	0.90	9.78
FREQUENCY	0.85	0.74
REASON	0.62	0.43
ROUTE	0.78	0.44
STRENGTH	0.95	0.92
Overall (micro)	0.87	0.68

Table C.7 **Lenient F_1-scores n2c2 BRAT:** Using the official N2C2 evaluation script on the BRAT-converted output of our best-performing Llama-70B FT model.

C.2 Metrics

Gold value	Predicted value
qhs	qhs (once a day at bed time)
25 mg / 3 ml	25 mg/3 ml
intraperitoneal bleeding	intraperitoneal bleeding from gist tumors
left eye	to left eye

Table C.8 **Lenient matches:** Example gold standard samples and predictions of relation values. All predicted values are considered lenient matches but not exact matches.

C.3 Additional results

Medication information	Llama 8b		Llama 70b	
	Zero	FT	Zero	FT
ADE	22.3	63.1	29.3	67.3
DOSAGE	23.9	90.4	45.0	90.3
DURATION	20.0	71.6	36.8	70.8
FORM	51.4	91.2	61.5	91.3
FREQUENCY	58.4	88.2	69.2	88.9
REASON	20.6	69.5	41.6	72.9
ROUTE	78.6	92.6	86.2	92.5
STRENGTH	73.7	91.8	82.5	91.5
MICRO AVG.	49.4	86.5	65.2	87.0

Table C.9 **Exact F_1-scores for the N2C2 corpus for the e2e MIE task:** Using Llama 3.1 (8b and 70b) in zero-shot (*Zero*) and fine-tuned (*FT*) settings.

Medication information	Llama 8b		Llama 70b	
	Zero	FT	Zero	FT
DOSAGE	0.0	27.8	3.1	50.0
DURATION	1.0	58.9	33.2	60.5
FORM	0.2	73.1	17.3	84.2
FREQUENCY	13.9	93.3	83.9	94.4
REASON	1.0	48.6	18.2	51.3
ROUTE	3.1	91.3	74.5	88.6
STRENGTH	13.7	91.8	79.9	92.7
MICRO AVG.	5.1	83.7	67.4	84.4

Table C.10 **Exact F_1-scores for the CARDIO:DE corpus for the e2e MIE task:** using Llama 3.1 (8b and 70b) in zero-shot (**Zero**) and fine-tuned (**FT**) settings.

C.3.1 Lenient vs. exact results

In this section we analyze, how much measured performance of our LLM improves when switching from exact to lenient matching, separating near-miss errors from truly wrong

predictions. In combination with the feedback-LLM these insights can help to improve data post-processing steps to support automatic evaluation.

In the N2C2 dataset (Figure C.3) the largest gain are observed for duration (+0.15) and reason (+0.10). Typical errors cover punctuation and completeness errors: *for 1 week,* vs *for one week*, *24hrs* vs. *for 24hrs* or *itchy* vs. *subjectively itchy* or *PNA* vs. *during PNA*. This shows that the model usually retrieves the correct medication-relation class name but struggles with relation value formatting or completeness. Moderate gaps for FREQUENCY and ADE point to a lesser extent for similar issues. ROUTE, DOSAGE and FORM in contrast are either fully correct or clearly incorrect.

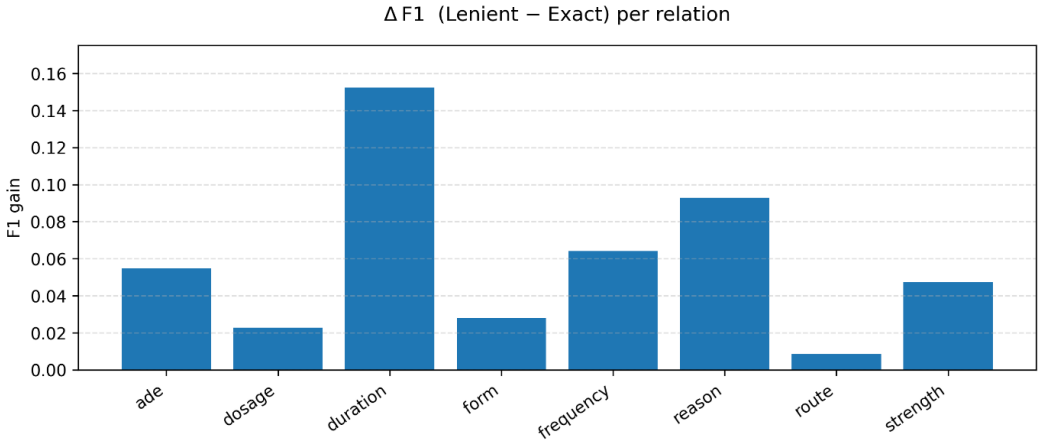


Fig. C.3 **Lenient vs. exact F_1-score:** N2C2 2018 track 2 test data using Llama 70b FT.

These patterns are similar for CARDIO:DE (Figure C.4): DURATION (+0.16) and REASON (+0.15) show the highest gap between lenient and exact matches. Whereas ROUTE, FREQUENCY and STRENGTH change only to a marginal extend.

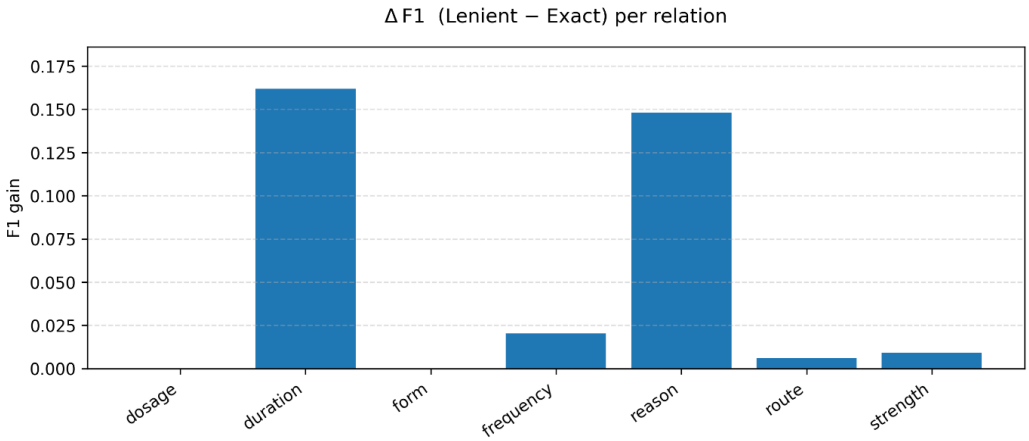


Fig. C.4 **Lenient vs. exact F_1-score:** CARDIODE test data using Llama 70b FT.

Overall, the plots suggest that for classes such as REASON and DURATION the lenient metric captures semantically similar predictions that fail under exact matching regarding normalization and completeness. However, while lenient matching, as the primary metric for the N2C2 task, uncovers these issues, semantic variations in format, abbreviations, comprehensiveness and granularity (cf. Table C.15) can be addressed more effectively through our feedback LLM.

C.3.2 OpenBioLLM

Medication information	zero	FT
ADE	–	69
DOSAGE	–	92
DURATION	–	83
FORM	–	94
FREQUENCY	–	95
REASON	–	78
ROUTE	–	93
STRENGTH	–	96
Micro avg.	–	91

Table C.11 **Lenient F1-scores for the N2C2 corpus for the e2e MIE task:** Using OPENBIOLLM 8b in zero-shot (zero) and fine-tuned (FT) settings.

Medication information	zero	FT
DOSAGE	–	0
DURATION	–	1
FORM	–	0
FREQUENCY	–	7
REASON	–	0
ROUTE	–	4
STRENGTH	–	10
Micro avg.	–	7

Table C.12 **Lenient F1-scores for the CARDIO:DE corpus for the e2e MIE task:** Using OPENBIOLLM 8b in zero-shot (zero) and fine-tuned (FT) settings.

C.3.3 Confidence intervalls

95% bootstrap confidence intervals (CI) of the lenient F_1-score of Llama-70b FT model for the N2C2 data (cf. Table C.13) and CARDIO:DE (cf. Table C.14). We used 1000 bootstrap resamples based on the complete test set (random seed: 42).

For the N2C2 dataset most classes show a tight CI (≤ 0.01). ADE and DURATION show the widest CI, because they have the fewest instances in the data (cf. Table C.2).

Entity class	95% CI	Llama 70b FT
ADE	0.70–0.76	73
DOSAGE	0.92–0.93	93
DURATION	0.83–0.89	86
FORM	0.94–0.95	94
FREQUENCY	0.95–0.96	95
REASON	0.81–0.83	82
ROUTE	0.93–0.94	93
STRENGTH	0.96–0.97	96

Table C.13 **Confidence interval N2C2:** Column 1: 95% confidence interval (CI). Column 2: Lenient F1-scores of fine-tuned Llama 70b on the N2C2 corpus for the end-to-end MIE task.

Given that all classes in the CARDIO:DE dataset have lower instance counts compared to N2C2 data, the overall higher confidence interval is expected. This is particularly noticeable for rare classes like DOSAGE and FORM. In contrast, more frequent classes such as FREQUENCY, STRENGTH, REASON and DURATION remain more stable with CI <=0.08.

Entity class	95% CI	Llama 70b FT
DOSAGE	0.27–0.69	50
DURATION	0.73–0.80	77
FORM	0.72–0.93	84
FREQUENCY	0.96–0.97	96
REASON	0.62–0.70	66
ROUTE	0.86–0.92	89
STRENGTH	0.93–0.95	94

Table C.14 **Cofidence interval CARDIO:DE:** Column 1: 95% confidence interval (CI). Column 2: Lenient F1-scores of fine-tuned Llama 70b on the CARDIO:DE corpus for the end-to-end MIE task.

C.4 Feedback LLM

Pattern type	Example
Format variance	*Input:* Her hypertension responded to 5mg/10mg doses of IV labetalol. **strength**: **gold** = "5mg/10mg" **pred** = ["5mg", "10mg"]
Abbreviations	*Text:* Antiphospholipid syndrome: on longterm anticoagulation with warfarin for history of APLS. **reason**: **gold** = "Antiphospholipid syndrome" **pred** = "APLS"
Comprehensiveness	*Text:* controlling tachypnea and dyspnea with morphine. **reason**: **gold** = "controlling tachypnea and dyspnea" **pred** = ["tachypnea", "dyspnea"]
Granularity	*Text:* Because of his severe 3VD he was started on heparin and nitroglycerine drips for optimal control of his CAD. **reason**: **gold** = "CAD" **pred** = "severe 3VD"

Table C.15 **Error patterns of false positives and false negatives:** First column: pattern type; second column: example, including the input paragraph and in bold the relation class with gold vs. predicted values.

C.4.1 Manual evaluation feedback LLM

To evaluate the performance of the feedback LLM we conducted manual evaluation of a subset of all predicted instances on the English dataset. For each relation class, and separately for false positive and false negative samples, if available, we selected five instances classified as SIMILAR and five instances classified as NOT SIMILAR by the feedback LLM (overall 149 instances). A clinical expert, presented with the gold standard, the LLM prediction and the feedback LLM classification, evaluated each instance to determine if it was classified correctly from a clinical perspective.

The clinical expert agreed with the feedback LLM's SIMILAR classification in 94% of cases. In contrast, he only agreed with the feedback LLM's NOT SIMILAR classification in 44% of cases. However, the expert often considered instances labeled as NOT SIMILAR to be SIMILAR as well. This suggests that the feedback model acts as a more cautious classifier, which can be advantageous in a clinical context, as the expert frequently reclassified NOT SIMILAR instances as SIMILAR. It also highlights the strong performance of our medication information classifier.

C.5 Prompts

```
class MedicationInfo(BaseModel):
    """A list of medication information extracted from the text."""
    medications: List[Medication] = Field(
        default_factory=list,
        description="A list of medications and their related information."
    )

class Medication(BaseModel):
    """Medication information extracted from the text."""
    medication: str = Field(description="A drug name or an active ingredient.")
    strength: Optional[Union[str, List[str]]] = Field(
        default="",
        description=("Extract the strength of the medication from the text. "
                     "Examples: 100 mg, 5 mg, 10 mg, 40 mg, 20 mg, 10mg, 5mg, "
                     "25 mg, 2.5 mg, 100mg, etc.")
    )
    ...
```

Fig. C.5 **Pydantic schema:** Schema used in the system prompts of MIE experiments.

System prompt (n2c2, EN)

You are a doctor specializing in pharmacology. You receive a text along with two triplets containing medication information: the gold standard (Gold) and the model prediction (Pred), which may include false positives and false negatives. Each triplet includes the medication name, the category of medication information, and a value.
The context of the text should be taken into account to ensure the clinical meaning of the values is fully understood and no subtleties are overlooked.
Your task is to evaluate carefully and conservatively whether the values in the two triplets are clinically comparable. A classification of *SIMILAR* should only be made if the similarity is obvious and clearly clinically meaningful!
Please proceed with the evaluation as follows:
1) Carefully read the text to understand the context of the medication information.
2) Compare the values in the Gold and Pred triplets.
3) Determine if the values are clinically equivalent or comparable in the given context.
4) If the values are clearly similar within the context of the text, provide the result as *Result: SIMILAR*.
5) If the values are not clearly comparable or have a divergent clinical meaning, provide the result as *Result: NOT SIMILAR*.
Example of structure:
Text: The patient administers insulin a day to manage blood sugar levels.
Triplets: ((‘insulin’, ‘frequency’, ”), (‘insulin’, ‘frequency’, ‘day’)): The model predicts a false positive. Day is not in the gold standard. Result: NOT SIMILAR.
Text: The patient’s methotrexate regimen includes doses on qFri and qSat, administered weekly.
Triplets: ((‘methotrexate’, ‘frequency’, ‘qFri, qSat’), (‘methotrexate’, ‘frequency’, [‘qFri’, ‘qSat’])): Both values are similar. Result: SIMILAR.

Table C.16 **English system prompt feedback LLM:** System prompt for the feedback LLM task (English, N2C2).

Systemprompt (CARDIO:DE, DE)

Sie sind ein Arzt mit Spezialisierung auf Pharmakologie. Sie erhalten einen Text und zwei Triplets, die Medikationsinformationen enthalten: der Goldstandard (Gold) und die Modellvorhersage (Pred), mit möglichen falsch-positiven und falsch-negativen Ergebnissen. Jedes Tripel enthält den Medikamentennamen, die Kategorie der Medikationsinformation und einen Wert.
Der Kontext des Textes sollte berücksichtigt werden, damit die klinische Bedeutung der Werte vollständig verstanden wird und keine Feinheiten übersehen werden.
Ihre Aufgabe ist es, vorsichtig und eher konservativ zu bewerten, ob die Werte in den beiden Tripeln klinisch eindeutig vergleichbar sind. Eine Einstufung als *ÄHNLICH* sollte nur dann erfolgen, wenn die Ähnlichkeit offensichtlich und eindeutig klinisch sinnvoll ist!
Führen Sie die Bewertung folgendermaßen durch:
1) Lesen Sie den Text sorgfältig, um den Kontext der Medikationsinformation zu erfassen.
2) Vergleichen Sie die Werte im Gold- und Pred-Tripel.
3) Bestimmen Sie, ob die Werte im gegebenen Kontext klinisch gleichwertig oder vergleichbar sind.
4) Wenn die Werte eindeutig und im Kontext des Textes klinisch ähnlich sind, geben Sie das Ergebnis als *Ergebnis: ÄHNLICH* an.
5) Wenn die Werte nicht eindeutig vergleichbar sind oder eine abweichende klinische Bedeutung haben, geben Sie das Ergebnis als *Ergebnis: NICHT ÄHNLICH* an.
Beispiel zur Struktur:
Text: Der Patient wird mit Entresto 24/26 mg behandelt.
Triplets: ((‘Entresto’, ‘strength’, [‘24’, ‘26mg’]), (‘Entresto’, ‘strength’, ‘24/26mg’)): Werte sind identisch. Ergebnis: ÄHNLICH.
Text:
Triplets: ((‘Lisinopril’, ‘reason’, ‘Augentropfen’), (‘Lisinopril’, ‘reason’, ‘BP control’)): Beide beziehen sich auf Blutdruckkontrolle. Ergebnis: ÄHNLICH.

Table C.17 **German system prompt feedback LLM:** System prompt for the feedback LLM task (German, CARDIO:DE).

C.6 Hyperparameters

Training On a single NVIDIA RTX6000 we trained Llama 8b models $\approx 20h$ on the full N2C2 training set. Maximum VRAM: $14GB$.

On a single NVIDIA H100 we trained Llama 70b models $\approx 10h$ on the full n2c2 training set. Maximum VRAM: $48GB$.

Hyperparameter	`Llama 8B`	`Llama 70B`
base model	`meta-llama/Meta-Llama-3.1-8B`	`meta-llama/Meta-Llama-3.1-70B`
max_seq_length	`2048`	`2048`
random_seed	`123`	`42`
4bit	`False`	`True`
per_device_batch_size	`4`	`4`
gradient_accumulation	`2`	`2`
num_epochs	`1`	`1`
r (LoRA)	`16`	`16`
Alpha (LoRA)	`16`	`16`

Table C.18 **Hyperparameters fine-tuning:** Fine-tuning Llama 3.1 on N2C2 and CARDIO:DE.

Inference On a single NVIDIA RTX6000 we conducted inference with Llama 8b models $\approx 14h$ on the full N2C2 test set.

On a single NVIDIA H100 conducted inference with Llama 70b models $\approx 34h$ on the full N2C2 training set.

Hyperparameter	**Llama 8b**	**Llama 70b**
max_seq_length	2048	2048
max_new_tokens	384	384
random_seed	42	42
4bit	False	False
batch_size	16	16
top_p	1.0	1.0
temperature	1.0	1.0
do_sample	False	False

Table C.19 **Hyperparameters inference:** Inference Llama 3.1 on N2C2 and CARDIO:DE.

Shapley values We used the N2C2 Llama 8b FT for all Shapley experiments.

Hyperparameter	Value
max_seq_length	2048
max_new_tokens	384
random_seed	42
4bit	False
top_p	1.0
temperature	1.0
do_sample	False

Table C.20 **Hyperparameters Shapley values:** Calculating Shapley value attributions using Llama 8b FT.

C.7 Interpretability

In this section we present further examples for the application of Shapley values for use case: (1) assessing input token contribution to relation information output token (cf. Section C.7.1), (2) uncovering implicit knowledge on relation information (cf. Section C.7.2). Lastly, we conducted a quantitative analysis to further investigate implicit knowledge capabilities of LLMs by applying advanced interpretability methods.

C.7.1 Use case 1

We aim to investigate how Shapley values can contribute to explicating implicit information within an LLM, thereby increasing the model's interpretability in the MIE task. To show the ability of LLMs if they correctly assign relation information to medications we investigate a specific edge case: the same relation value appears twice in the input, each linked to a different medication name. Thus, we can analyze whether the first occurrence correctly maps to the first medication, and the second to the corresponding second medication.

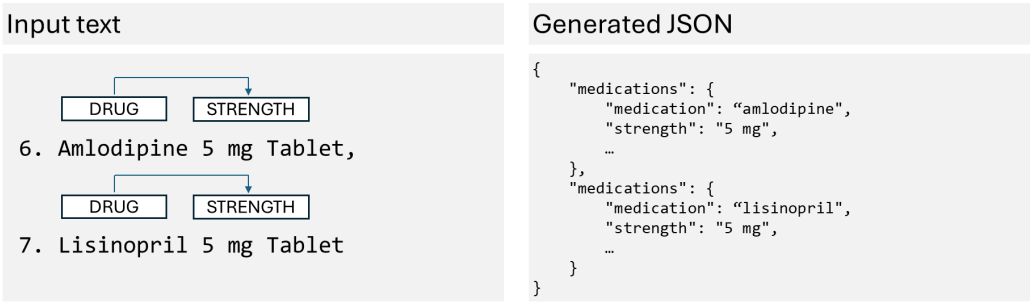


Fig. C.6 **Use case 1 example 1 input:** (left) Example input text containing two medications (Amlodipine, Lisinopril) each related to a similar strength value (5 mg). (right) Corresponding generated JSON output snippet containing the medication names and the strength values.

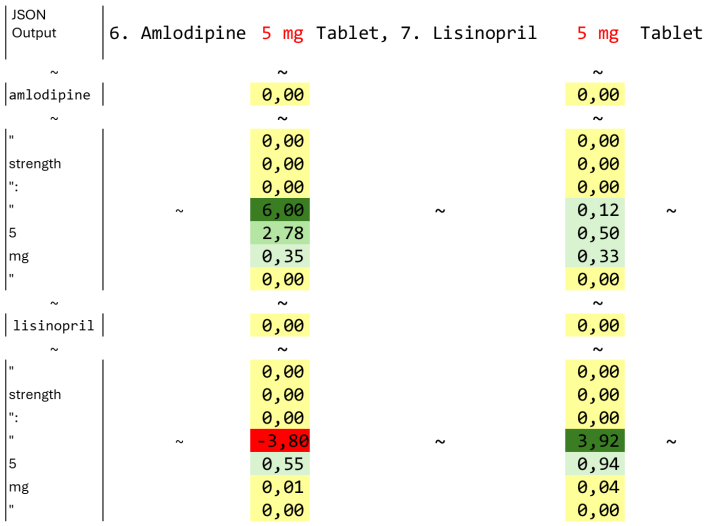


Fig. C.7 **Use case 1 example 1 attributions:** Visualizing approximate Shapley values for the strength token of the input text “6. Amlodipine 5 mg Tablet, 7. Lisinopril 5 mg Tablet” and the generated JSON token for the strength relation class and value.

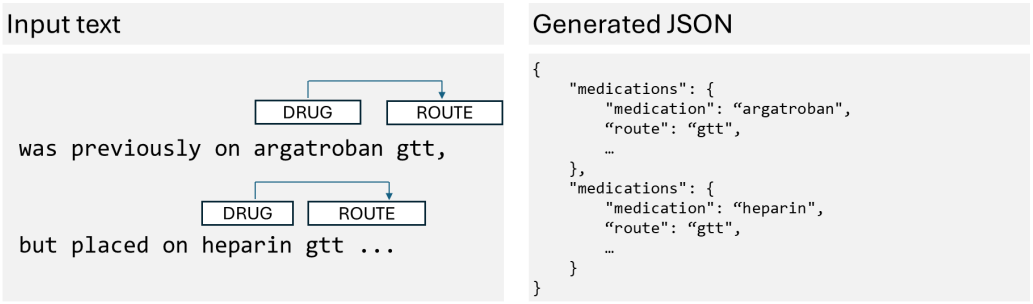


Fig. C.8 **Use case 1 example 2 input:** (left) Example input text containing two medications (Argatroban, Heparin) each related to a similar route value (gtt). (right) Corresponding generated JSON output snippet containing the medication names and the route values.

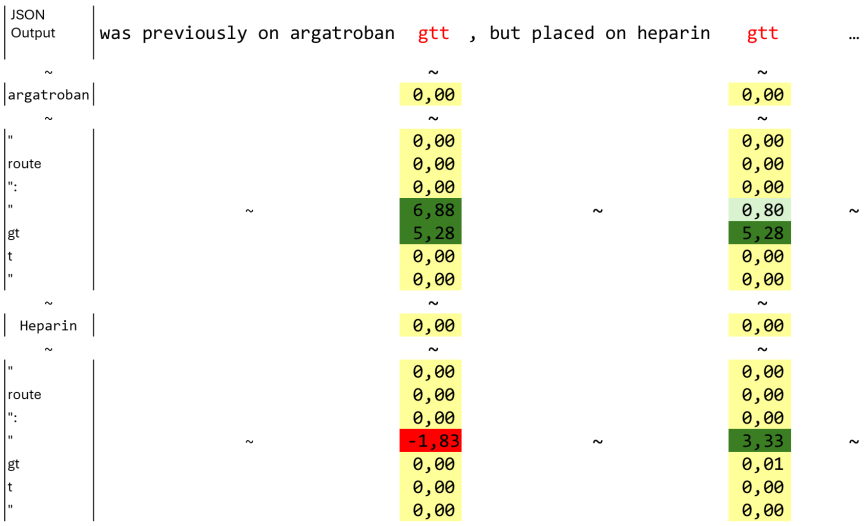


Fig. C.9 **Use case 1 example 2 attributions:** Visualizing approximate Shapley values for the route token of the input text “was previously on argatroban gtt, but placed on heparin gtt ...” and the generated JSON token for the route relation class and value.

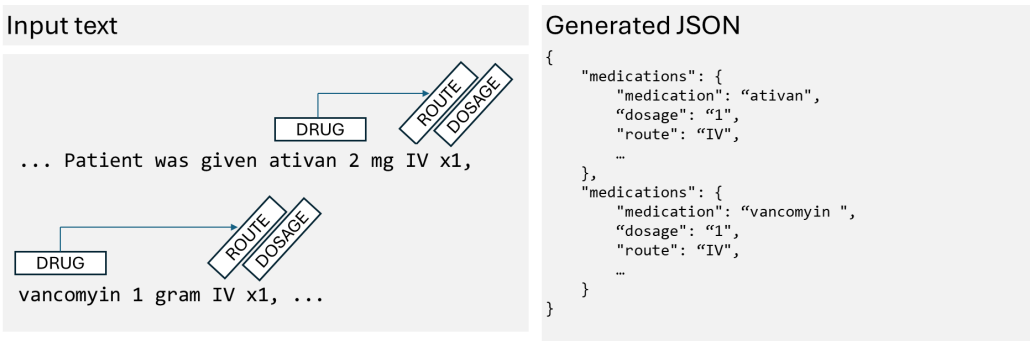


Fig. C.10 **Use case 1 example 3 input:** (left) Example input text containing two medications (Ativan, Vancomyin) each related to a similar route value (IV) and dosage value (1). (right) Corresponding generated JSON output snippet containing the medication names and the route and dosage values.

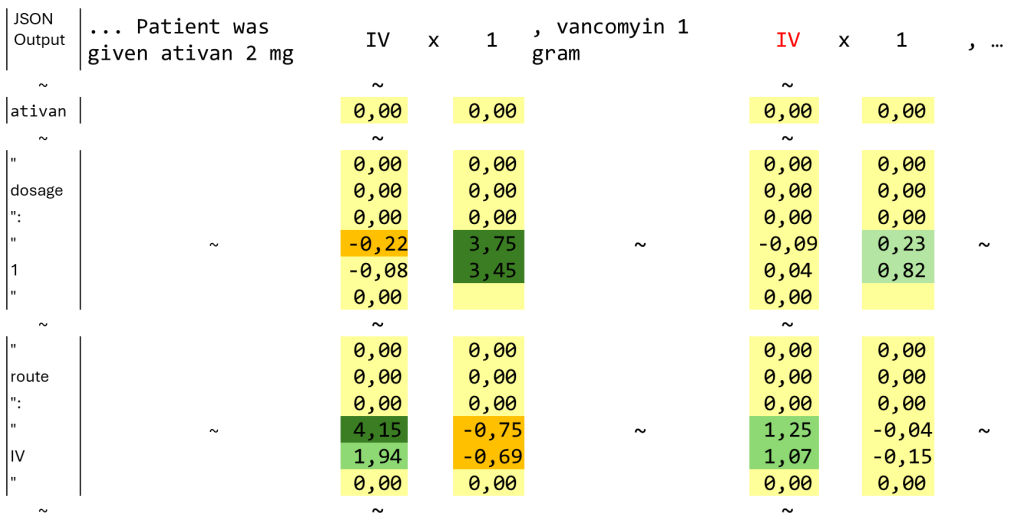


Fig. C.11 **Use case 1 example 3 attributions:** Visualizing approximate Shapley values for the dosage and route token of the input text "... Patient was given ativan 2 mg IV x1, vancomyin 1 gram IV x1, ... " and the generated JSON token for the dosage and route relation class and value.

C.7.2 Use case 2

The second use case uncovers that LLMs possess implicit knowledge regarding relation information, even when they do not explicitly generate this information in the JSON output. The following examples emphasize that tokens containing information on ADE or medication REASON negatively contribute to an empty string instead of a relation string value to the JSON output.

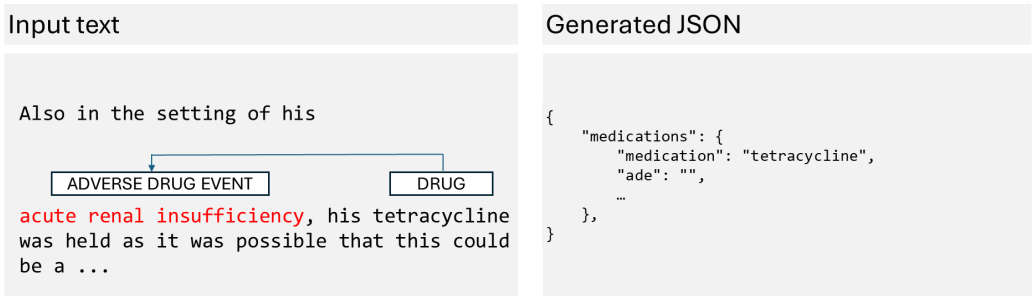


Fig. C.12 **Use case 2 example 1 input:** (left) Example input text containing ADE of the medication Tetracycline. (right) Corresponding generated JSON output snippet containing the medication name and the empty ADE value.

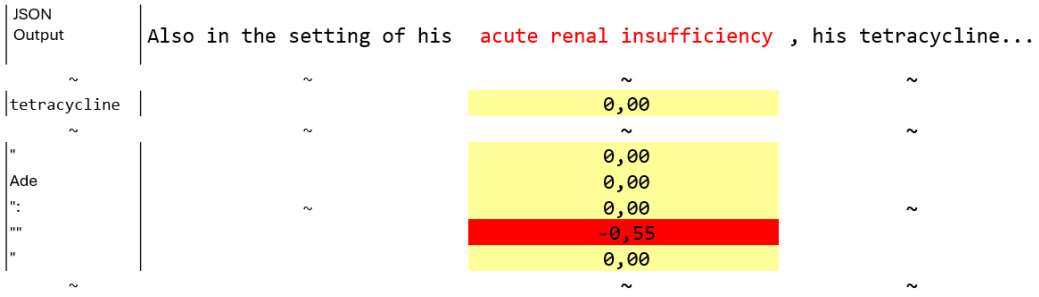


Fig. C.13 **Use case 2 example 1 attributions:** Visualizing approximate Shapley values for the ADE token of the input text “Also in the setting of his acute renal insufficiency, his tetracycline was held as it was possible that this could be a ...” and the generated JSON token for the ADE relation class and value.

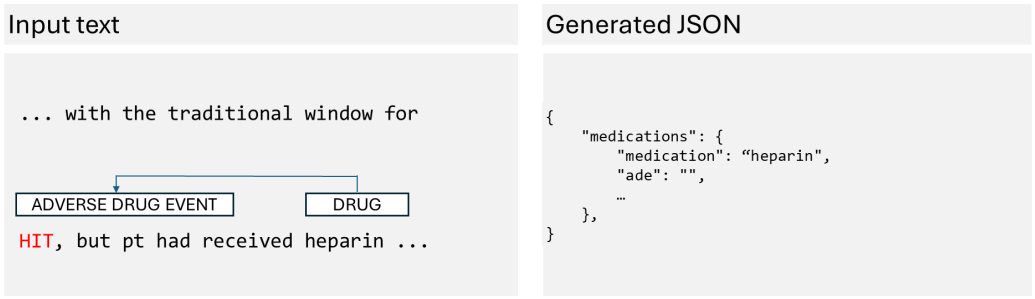


Fig. C.14 **Use case 2 example 2 input:** (left) Example input text containing ADE of the medication Heparin. (right) Corresponding generated JSON output snippet containing the medication name and the empty ADE value.

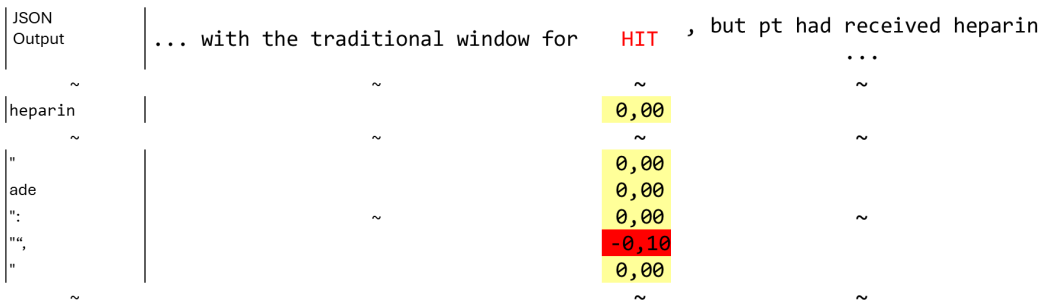


Fig. C.15 **Use case 2 example 2 attributions:** Visualizing approximate Shapley values for the ADE token of the input text “... with the traditional window for HIT, but pt had received heparin ...” and the generated JSON token for the ADE relation class and value.

Fig. C.16 **Use case 2 example 3 input:** (left) Example input text containing reason of the medication heparin. (right) Corresponding generated JSON output snippet containing the medication name and the empty reason value.

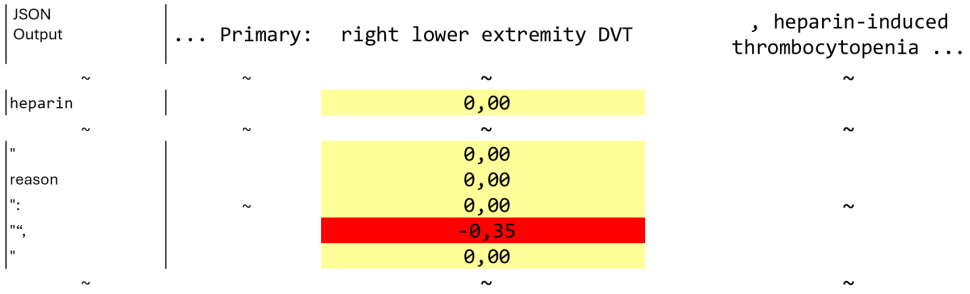


Fig. C.17 **Use case 2 example 3 attributions:** Visualizing approximate Shapley values for the reason token of the input text “... Primary: right lower extremity DVT, heparin-induced thrombocytopenia ...” and the generated JSON token for the reason relation class and value.

C.7.3 Quantitative analysis

To support our interpretability findings, a quantitative analysis was conducted regarding Use Case 2. This involved first extracting all false negative ADE and REASON instances (defined by the model's empty string prediction) from the Llama 8b FT test set. Subsequently, Shapley value contributions of input tokens representing the ADE and REASON concepts were collected for these predicted empty strings. Concurrently, contributions from these same input tokens to all other empty string predictions were also collected.

We hypothesized that a more negative contribution value of empty string ADE and REASON predictions indicates an implicit concept understanding within the model, while the model confidence is insufficient to generate the correct concept in the output.

Descriptive statistics for both ADE and REASON contribution sets are summarized in Table C.21. As shown, concept contributions for both classes exhibit notably more negative means or medians compared to their respective other empty string contributions. To statistically evaluate these observations, independent samples Welch's t-tests were performed for both classes separately. For ADE and REASON, the tests resulted in a p-value $<$ 0.0001. These findings support our hypothesis, demonstrating that Shapley values can help to effectively uncover implicit information, even if the model's prediction probability is too low to generate the correct information correctly. It confirms that Shapley values reveal the presence of a false negative prediction instance.

Shapley value type	**N (sample size)**	**Mean (std. dev)**	**Median**
ADE contributions	42	-0.17 (0.19)	-0.09
Other (ADE) contributions	270	-0.0004 (0.11)	0.00002
REASON contributions	87	-0.16 (0.18)	-0.11
Other (REASON) contributions	544	-0.0007 (0.03)	0.00003

Table C.21 **Quantitative analysis use case 2:** Descriptive statistics for ADE and REASON.

Appendix D

Clinical Application: Medication Trends and Polypharmacy

D.1 Manual analysis

We extracted 417 letters containing 55,664 paragraphs. From all predicted model outputs (cf. Chapter 6) we filtered predictions with any medication extracted (`has_med=1`) and any *antikoagulation* substring in the JSON output (`has_antikoag=1`), resulting in a manual sample of $n = 86$ paragraphs. We manually annotated each paragraph with the following metrics: (i) `correct_class`: VKA/DOAC classification, (ii) `current_oac`: OAC currently prescribed?, (iii) `only_oac_med`: generic *Antikoagulation* medication without active ingredient/brand name, (iv) `missed_oac`: model missed OAC medication and (v) `missing_relation`: OAC medication extracted, but no *Antikoagulation* relation identified.

Table D.1 summarizes all statistics of the manual review. We present PPVs per dataset.

Metric	Count
Count paragraphs	55,664
`has_med=1`	5,059
`has_antikoag=1`	86
`correct_class=1`	53
`current_oac=1`	46
`only_oac_med=1`	28
`missed_oac=1`	9
`missing_relation=1`	10

Table D.1 **Manual review OAC:** Manual review metrics and statistics of the 2012 OAC dataset.

$$PPV_{class} = \frac{53}{86} = 61.6\% \tag{D.1}$$

$$PPV_{current} = \frac{46}{86} = 53.5\% \tag{D.2}$$

Many paragraphs contained `only_oac_med`, hence no VKA/DOAC classification was possible. Hence, we as well calculated PPVs only if active ingredients/brand names are present ($n_{classiviable} = 86 - 28 = 58$).

$$PPV_{class} = \frac{53}{58} = 91.4\% \tag{D.3}$$

$$PPV_{current} = \frac{46}{58} = 79.3\% \tag{D.4}$$

We include de-identified examples for all metrics (Table D.2).

Metric	**Example**	**JSON output**
`only_oac_med=1`	*Orale Vollantikoagulation bei paroxysmalem Vorhofflimmern*	{'medications', [{'medication': 'Vollantikoagulation', 'strength': '', 'frequency': '', 'reason': 'paroxysmalem Vorhofflimmern', 'duration': '', 'route': 'Orale', 'form': '', 'dosage': ''}}
`missed_oac=1`	Antikoagulation mit Marcumar auch bei stabilem Sinusrhythmus (Ziel INR 2,0-3,0) für 3 Monate	[{'medication': 'Antikoagulation', 'strength': '', 'frequency': '', 'reason': '', 'duration': 'für 3 Monate', 'route': '', 'form': '', 'dosage': ''}}
`missing_relation=1`	Clexane 0,8 2x/d, erhöht, zur Vollantikoagulation"	*{'medications', [{'medication': 'Clexane', 'strength': '0,8', 'frequency': '2x/d', 'reason': '', 'duration': '', 'route': '', 'form': '', 'dosage': ''}}*

Table D.2 **Example paragraphs of manual review:** Representative de-identified paragraphs including JSON output per selected metrics.