Comprehensive Study on German Language Models for Clinical and Biomedical Text Understanding

Ahmad Idrissi-Yaghir[1,2*], **Amin Dada**[3*], **Henning Schäfer**[1,4*], **Kamyar Arzideh**[3], **Giulia Baldini**[3,11], **Jan Trienes**[3], **Max Hasin**[3], **Jeanette Bewersdorff**[5], **Cynthia S. Schmidt**[3,4], **Marie Bauer**[3], **Kaleb E. Smith**[6], **Jiang Bian**[7,8], **Yonghui Wu**[7,8], **Jörg Schlötterer**[9,10], **Torsten Zesch**[5], **Peter A. Horn**[4], **Christin Seifert**[9], **Felix Nensa**[3,11], **Jens Kleesiek**[3,12,13,14], **Christoph M. Friedrich**[1,2]

[1] Department of Computer Science, University of Applied Sciences and Arts Dortmund, Dortmund, Germany
[2] Institute for Medical Informatics, Biometry and Epidemiology, University Hospital Essen, Essen, Germany
[3] Institute for AI in Medicine (IKIM), University Hospital Essen (AöR), Essen, Germany
[4] Institute for Transfusion Medicine, University Medicine Essen, Essen, Germany
[5] Computational Linguistics, CATALPA FernUniversität in Hagen, Germany
[6] NVIDIA, Santa Clara, CA, USA
[7] Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL, USA
[8] Cancer Informatics and eHealth core, University of Florida Health Cancer Center, University of Florida, Gainesville, FL, USA
[9] University of Marburg, Marburg, Germany
[10] University of Mannheim, Mannheim, Germany
[11] University Hospital Essen, Institute of Interventional and Diagnostic Radiology and Neuroradiology, Essen, Germany
[12] Cancer Research Center Cologne Essen (CCCE), West German Cancer Center Essen, University Hospital Essen (AöR), Essen, Germany
[13] German Cancer Consortium (DKTK, Partner site Essen), Heidelberg, Germany
[14] Department of Physics, TU Dortmund, Dortmund, Germany
{ahmad.idrissi-yaghir, christoph.friedrich}@fh-dortmund.de
{amin.dada, henning.schaefer, jens.kleesiek}@uk-essen.de

Abstract

Recent advances in natural language processing (NLP) can be largely attributed to the advent of pre-trained language models such as BERT and RoBERTa. While these models demonstrate remarkable performance on general datasets, they can struggle in specialized domains such as medicine, where unique domain-specific terminologies, domain-specific abbreviations, and varying document structures are common. This paper explores strategies for adapting these models to domain-specific requirements, primarily through continuous pre-training on domain-specific data. We pre-trained several German medical language models on 2.4B tokens derived from translated public English medical data and 3B tokens of German clinical data. The resulting models were evaluated on various German downstream tasks, including named entity recognition (NER), multi-label classification, and extractive question answering. Our results suggest that models augmented by clinical and translation-based pre-training typically outperform general domain models in medical contexts. We conclude that continuous pre-training has demonstrated the ability to match or even exceed the performance of clinical models trained from scratch. Furthermore, pre-training on clinical data or leveraging translated texts have proven to be reliable methods for domain adaptation in medical NLP tasks.

Keywords: German-centric NLP, Clinical Language Models, Domain Adaptation

1. Introduction

In recent years, pre-trained language models (PLMs) such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have become crucial in the field of natural language processing (NLP). These models have significantly enhanced the performance of a wide range of tasks, including document and token classification, as well as machine translation and text summarization. The success of these models is primarily attributed to their transformer-based architecture (Vaswani et al., 2017) and their training on large amounts of unlabeled data. However, since these models are often trained on general data sources such as Wikipedia, news articles, and books, their effectiveness may be limited in specific domains, such as

[*] These authors contributed equally to this work

LREC-COLING 2024, pages 3654–3665

medicine or finance, which have distinct terminologies and writing styles. To achieve better results in these specialized fields, it is essential to use language models that are tailored to these specific domains. Building on this concept, language models can be adapted to specialized domains through two methods. The first approach involves training models from scratch on unlabeled data from the desired domain. This method ensures that the model is grounded in the unique characteristics of the target domain from the beginning. The second approach relies on continuous pre-training. Instead of starting from scratch, existing general-purpose pre-trained models can be used and refined through further pre-training on the domain-specific unlabeled data (Gururangan et al., 2020). This allows a transition that shifts the focal point of the model from a broad scope to one specific to the particularities of the target domain.

Particularly in the medical domain, such specialized models can potentially improve the practice of medicine by providing accurate and relevant insights from vast amounts of textual data. These specialized models are highly valuable given the complex nature of medical terminology and the critical importance of accurate information in healthcare. For example, they can help analyze patient records, extract critical information from medical literature, and facilitate real-time clinical decision-making by understanding patient queries or medical notes.

However, building these specialized medical models presents unique challenges. Medical data is characterized not only by specialized terminology but also by the sensitive nature of the information. Patient confidentiality and other ethical considerations are paramount, which can complicate the acquisition of large datasets for training.

Particular focus has been set on German medical data because of its data sparsity when compared to English datasets (Névéol et al., 2018; Schneider et al., 2020). While resources for languages like French (Labrak et al., 2023) and Spanish (Carrino et al., 2022) have been increasingly made available, German medical data still remains notably underrepresented and has only recently been pushed further (Bressem et al., 2024).

In this work, several new German biomedical and clinical language models are introduced and extensively evaluated on multiple downstream tasks. All model variants are continuously pre-trained on two different data streams, resulting in public models benefiting from translations of biomedical and medical datasets into German, while private models use internal data from a large German hospital. Translation-based models will be made publicly available.[1]

[1] https://huggingface.co/ikim-uk-essen

2. Related Work

Following the success of transformer-based language models, several language models have been developed for biomedical and clinical domains, mainly for English. One of the earliest models is BioBERT (Lee et al., 2019), which was initialized from a general BERT model and further pre-trained using biomedical data such as PubMed abstracts. This approach demonstrated the effectiveness of continuous pre-training on domain-specific data, allowing the model to more accurately capture the challenging nature of the biomedical domain textual data. Several other specialized models soon followed. One such specialized model is ClinicalBERT (Alsentzer et al., 2019). While it used a similar continuous pre-training approach to BioBERT, it differed in its integration of clinical data, particularly from sources such as the Medical Information Mart for Intensive Care (MIMIC-III, Johnson et al. 2016), a public dataset of de-identified medical records for over 40,000 patients in the intensive care units of Beth Israel Deaconess Medical Center from 2001–2012. Furthermore, Gu et al. (2022) introduced PubMedBERT, which was not based on a previously pre-trained model but was trained from scratch on biomedical data. The training dataset consisted of both PubMed abstracts and the full text from PMC, resulting in models that were able to achieve improved performance on a wide range of biomedical tasks.

In languages other than English, it is more challenging to build such specialized models due to the lack of available data, as is the case for German (Zesch and Bewersdorff, 2022). However, significant advancements have occurred in this field recently, such as BioGottBERT (Lentzen et al., 2022), a model based on the GottBERT model (Scheible et al., 2020). GottBERT is a German RoBERTa base model that underwent training utilizing general domain information. BioGottBERT enhanced its medical capabilities by undergoing further pre-training on public German medical texts from Wikipedia and scientific abstracts. This resulted in a significant improvement in performance on medical tasks when compared to GottBERT. In addition to BioGottBERT, the authors trained an ELECTRA (Clark et al., 2020) small and basic models from scratch with the aim of evaluating the effectiveness of training new models using only a limited amount of biomedical data. However, the authors reported that this strategy was unsuccessful, and the resulting models were inferior to existing general models. Another German medical model is MedBERTde (Bressem et al., 2024). This BERT-based model was trained from scratch using various public German medical datasets such as

GGPONC 1.0 (Borchert et al., 2020), PubMed abstracts, and doctoral dissertations. In addition, the training also incorporated real-world clinical data, such as radiology reports from the Charité University Hospital in Berlin. By integrating such a wide range of clinical and biomedical data, medBERTde aims to provide a comprehensive understanding of the medical field tailored to the German context.

Recently, French has also seen a surge in specialized biomedical and clinical pre-trained language models. Among them is DrBERT (Labrak et al., 2023), which is based on the RoBERTa architecture and has been trained on both public web data and specialized private medical data from the University Hospital of Nantes. The public web data is a large text corpus called NACHOS (opeN crAwled frenCh Healthcare cOrpuS), crawled from several online biomedical sources. By training on the different datasets, a set of models was obtained, which were then compared by evaluating their performance on a wide range of public and private tasks. Another developed model is AliBERT (Berhe et al., 2023), a model designed specifically for the French biomedical domain. It was trained using a regularized unigram tokenizer on different sub-corpora of French biomedical textual documents, such as biomedical articles from ScienceDirect, thesis manuscripts, and articles from the Cochrane database. The model excels in F1 and accuracy scores, proving its capabilities in this domain. Interestingly, despite a smaller amount of training data and a shorter pre-training period, AliBERT manages to surpass some notable general French models, highlighting its capabilities.

3. Pre-Training Datasets

This section details the datasets utilized for pre-training, including clinical data and public medical/biomedical data. While the clinical dataset captures real-world patient insights, the medical data encompasses a wide range of scientific information. These data sources provide the foundation for our models. An in-depth description of these datasets follows.

3.1. Clinical Data

The first dataset was sourced from a major German hospital, providing a comprehensive clinical dataset that spans from 2002 to 2023. It includes various clinical documents, including clinical notes, different reports, and doctor's letters. Each text was divided into paragraphs, which were then filtered. Paragraphs with a ratio of letters to other characters below 60% and paragraphs with an average number of words per line below three were filtered out. The resulting dataset consists of 3,060,845,169 tokens from 25,023,489 documents. To the best of our knowledge, this is the largest German clinical text dataset compiled for pre-training.

3.2. Public Data

The second dataset is derived from publicly available biomedical data. It was initiated with approximately 16K German abstracts from PubMed. Recognizing the limited size of this dataset, it was necessary to expand it to improve reliability and coverage. To achieve this, approximately 6 million English PubMed abstracts, along with MIMIC-III clinical notes (Johnson et al., 2016), were translated using the Fairseq WMT'19 English to German translation model[2] (Ng et al., 2019). Although the translation of medical content can be complex and potentially lead to inaccuracies due to specialized terminologies, it provides a way of augmenting the corpus. The decision to use this particular translation model was based on a physician's assessment. They were provided with different translations of 10 PubMed abstract samples and 20 MIMIC notes samples generated by various translation models, including WMT19-en-de, M2M-100 (Fan et al., 2021), NLLB (Team et al., 2022), T5 (Raffel et al., 2020), MBart-50 (Tang et al., 2021), and OPUS-MT-en-de (Tiedemann and Thottingal, 2020). Their evaluation guided the final model selection.

Prior to translation, a preprocessing step was performed. All documents were tokenized into individual sentences using the Stanza library (Qi et al., 2020). These sentences were then grouped into specific segments, each limited to a maximum of 128 tokens. Due to this segmentation, the number of documents increased substantially. For instance, the initial 6 million English PubMed abstracts were divided into approximately 21 million segments or documents for translation. The token limit was chosen based on the observation that segments with more than 128 tokens often suffered from poor translation quality. The number of tokens was determined using the tokenizer of the translation model. This process yielded approximately 45M documents, detailed in Table 3.2.

4. Base Models

In this section, the models included in our benchmark are described. As a baseline for our evaluations, we use models that have been pre-trained on extensive datasets and have demonstrated strong performance on a variety of general and medical NLP tasks. The general Ger-

[2] https://huggingface.co/facebook/wmt19-en-de, last accessed: 2023-10-13

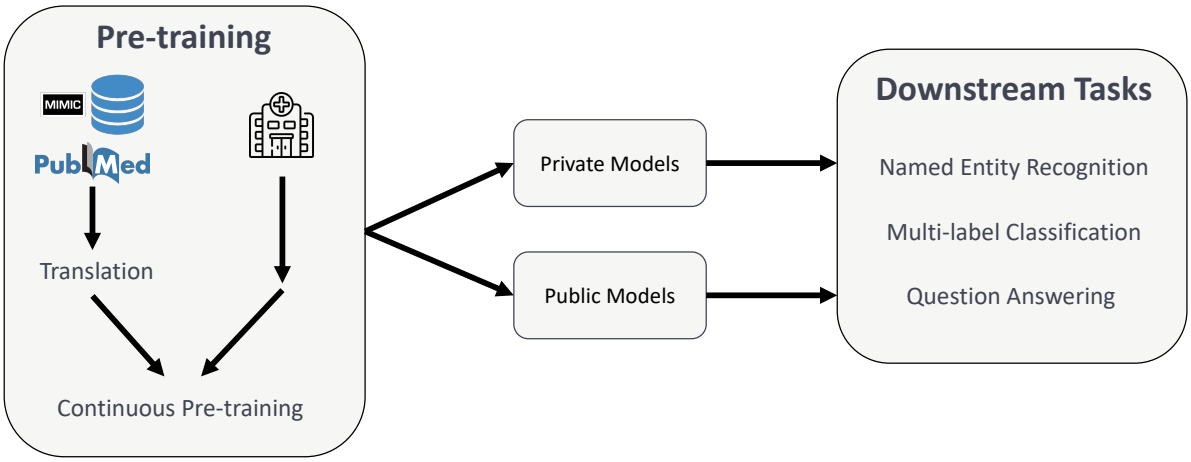

Figure 1: Workflow for Continuous Pre-training of different Publicly Available General Domain Models with Private and Public Datasets: The private dataset comes from a major German hospital and yields 25M documents. The public dataset, sourced from English PubMed abstracts and MIMIC-III clinical notes, is preprocessed, tokenized into sentence segments, and translated into German via the Fairseq WMT'19 English to German model, resulting in approximately 45M documents. Both model setups are subject to model fine-tuning and evaluation across various biomedical and clinical German language downstream tasks.

Dataset	Tokens	Documents
German PubMed	5M	16K
PubMed	1,700M	21M
MIMIC-III	695M	24M
Total	2,400M	45M

Table 1: Public dataset composition. The increase in the number of documents for PubMed and MIMIC-III compared to the original source is due to the segmentation of the content into chunks of 128 tokens or less for the translation process.

man language models GBERT$_{base}$, GBERT$_{large}$ and GELECTRA$_{large}$ (Chan et al., 2020) were primarily considered. They were trained on four different datasets: the German portion of the *Open Super-large Crawled ALMAnaCH coRpus* (OSCAR) (Ortiz Suárez et al., 2020), German Wikipedia dump, *The Open Parallel Corpus* (OPUS) (Tiedemann, 2012), and Open Legal Data (Ostendorff et al., 2020). Another model is GottBERT (Scheible et al., 2020), a German RoBERTa$_{base}$ model trained on the German part of OSCAR data. Furthermore, two multilingual models XLM-R (Conneau et al., 2020) and mDeBERTa V3 (He et al., 2023) were also considered. Additionally, GeBERTa$_{base}$ and GeBERTa$_{large}$ (Dada et al., 2023) were explored, with these DeBERTa v2 (He et al., 2021) based models pre-trained from scratch on a combination of data sources ranging from Wikipedia to medical datasets. Finally, the German biomedical and clinical models BioGottBERT (Lentzen et al., 2022) and medBERTde (Bressem et al., 2024) were also examined to provide additional comparative insights.

5. Pre-training

Continuous pre-training of several publicly available general domain models was performed with the aforementioned datasets. First, we continued the pre-training of the GeBERTa$_{base}$ and GeBERTa$_{large}$ models with the clinical dataset. The objective was to quantify the contribution of clinical data to the performance of the models. While the original model has already been trained on medical texts obtained mainly from MIMIC-III, translated PubMed abstracts, and filtered CC100 (Wenzek et al., 2020), it has not seen any real-world German clinical data. Both models were trained for 200k AdamW (Loshchilov and Hutter, 2019) optimization steps with a batch size of 512. The learning rate was set to 3e^{-5} for the base model and 2e^{-5} for the large model. Additionally, the GBERT$_{base}$ and GBERT$_{large}$ models were further pre-trained on the clinical data with the same parameters. These models were not explicitly pre-trained on medical texts before.

In order to separately quantify the influence of translation-based medical texts and clinical documents, additional pre-training experi-

ments were conducted with the GBERT$_{base}$ and GBERT$_{large}$ models. Both models underwent further pre-training on the translated dataset. The GBERT$_{large}$ was trained for 73k steps with a learning rate of 5e^{-5} and a batch size of 144. On the other hand, the GBERT$_{base}$ was trained for 75k steps using a batch size of 336. Due to hardware limitations, the initial experiments were conducted separately, reflecting the differences in batch sizes between the experiments. The AdamW optimizer was also employed during this pre-training for the optimization process.

6. Downstream Datasets & Tasks

To evaluate the pre-trained models, the models were fine-tuned and subsequently evaluated across a range of downstream tasks, aiming to determine their efficacy and adaptability in specialized biomedical and clinical domains.

BRONCO

The Berlin-Tübingen-Oncology Corpus (BRONCO) (Kittner et al., 2021) is a comprehensive, freely accessible German corpus derived from 200 oncology discharge summaries of cancer patients. These summaries have been manually de-identified and annotated to highlight key entities such as diagnoses, treatments, and medications. This annotated corpus contains 11,434 sentences and 89,942 tokens, with 11,124 annotations identifying medical entities suitable for named entity recognition (NER). While the authors have released 75% (or 150 summaries) of the dataset to the public, they have kept 25% (or 50 summaries) as a held-out set to ensure unbiased evaluation or data contamination. Given the unavailability of the BRONCO50 dataset, a 5-fold cross-validation was performed to train and evaluate models on the BRONCO150 dataset.

GGPONC 2.0

The German Guideline Program in Oncology NLP Corpus (GGPONC) 2.0 (Borchert et al., 2022) represents a significant advance in German medical language resources and offers a large corpus for NER applications. Based on the top-level hierarchies of the SNOMED CT concept model, its annotation scheme distinguishes between several subclasses of entities. The “Finding” category includes entities such as diagnosis, pathology, and other relevant findings. The “Substance” category delves into clinical drugs, nutrients or body substances, and external substances. In addition, the “Procedure” category houses entities associated with therapeutic and diagnostic procedures. Recognizing the complex nature of clinical texts, where entity boundaries are often ambiguous, GGPONC 2.0 is complemented by a comprehensive annotation guide that clarifies the definition of each entity class.

GraSCCo

Graz Synthetic Clinical text Corpus (GraSCCo) (Modersohn et al., 2022) is a synthetic German corpus consisting of about 60 clinical documents with more than 43,000 tokens. It includes a series of alienation steps to hide privacy-sensitive information in real clinical documents, the true origin of all GraSCCo texts. As a result, the data is publicly available without any legal restrictions. Within the medbert.de paper, an additional annotation of the GraSCCo data for an NER task was performed by the authors of the GGPONC 2.0. We use these annotations for our benchmark as well.

CLEF eHealth 2019

The CLEF (Conference and Labs of the Evaluation Forum) (Kelly et al., 2019) eHealth dataset is a curated collection of non-technical summaries (NTS) of animal experiments from Germany, which was used to organize the Multilingual Information Extraction Task (Task 1) in the CLEF eHealth Challenge 2019 (Dörendahl et al., 2019). These NTS have been made publicly available to increase transparency in animal research. For the identification of the primary diseases that are the focus of the experiments, each NTS in the dataset is manually annotated with the corresponding ICD-10 codes. Reports are predominantly scientific and biomedical, while some clinical jargon could also be observed. In total, the dataset contains over 8,000 NTSs for training and an additional 407 NTSs for testing. The primary objective associated with this dataset is a multi-label classification, where systems are challenged to assign the relevant ICD-10 codes to each summary automatically.

RadQA

The RadQA dataset is an extractive question-answering dataset created from 1,223 anonymized radiology reports of brain CT scans from a large hospital in Germany. For its development, three medical student assistants in their sixth and eighth semesters were assigned to annotate 29,273 question-answer pairs. The annotators were provided with a list of questions designed with the input of a radiologist. This decision was influenced by the unique nature of radiology queries and the challenges of annotating sensitive clinical data. The annotators

Task	Learning Rate	Batch Size	Epochs
	base, large	base, large	
BRONCO	$3e-5, 1e-5$	$16, 16$	20
GGPONC 2.0	$3e-5, 1e-5$	$16, 16$	5
GraSCCo	$3e-5, 1e-5$	$16, 16$	20
CLEF eHealth	$4e-5, 1e-5$	$16, 32$	20
RadQA	$3e-5, 1e-5$	$16, 16$	10

Table 2: Hyperparameters of the different downstream tasks.

generated one custom question for every third report to ensure variety.

7. Fine-tuning

In a comprehensive evaluation, the performance of the models was assessed on a variety of downstream tasks. The problem classes included NER (BRONCO, GGPONC 2.0, GraSCCO), multi-label classification (CLEF eHealth 2019), and extractive question answering (RadQA).

Hyperparameters

Choosing the right hyperparameters is crucial for optimizing model performance. However, in this study, an extensive hyperparameter search was intentionally avoided to reflect a clinical environment with limited computational resources. Instead, a mixture of standard and task-specific settings was opted for, starting from the default parameters in HuggingFace. Only the learning rate and batch size were adjusted to address training instabilities, and the number of epochs was set to ensure convergence of all models, resulting in variation across different dataset sizes in the benchmark.

While this approach may not yield the optimal results that can be achieved by an extensive grid search with hundreds of configurations, it provides valuable insights into the performance of models under standard parameter conditions, which is particularly relevant for clinical applications where computational resources are often limited. By using uniform configurations for basic and large models and an appropriate small number of epochs for fine-tuning, comparability across the models and tasks in the benchmark is aimed to be increased.

Although a more in-depth analysis of the hyperparameter optimization process could further enhance the rigor of this work, the methodology used reflects a practical scenario in clinical settings and offers a realistic assessment of model performance under resource constraints. The specific hyperparameters used for each downstream task are listed in Table 2.

Task Specifics

In the context of the CLEF eHealth 2019 fine-tuning, class imbalances among the labels are addressed with logarithmic weighting. In multi-label scenarios, there are often many classes, among which some appear rarely and others very frequently. This can lead to training not converging at all or being sensitive to hyperparameters. Some model configurations were unstable for this task, especially for large models. Using the adjusted weights led to stable results. The positive class weights, w_j, are calculated for label j and use the following logarithmic scheme:

$$w_j = \log\left(\frac{N}{1+c_j}\right)$$

Where N denotes the total number of training samples and c_j is the count of each label. The logarithmic weighting scheme adjusts weights inversely to the label frequency in the training data, ensuring balanced attention between frequent and rare labels during training.

8. Results

The fine-tuning results for all downstream tasks are split between Table 3 and Table 4.

Across all tasks, our clinically pre-trained models achieved the highest F1-Scores, with the exception of GGPONC 2.0, where one of the translation models achieved the highest score. This is especially evident in the results of the GraSCCo dataset, where clinical pre-training improved the performance of GBERT$_{base}$ by 5.4 percentage points in F1-Score and the performance of GBERT$_{large}$ by 6.9 percentage points. In addition, the GBERT-BioM-translation models were able to outperform the general models. For instance, pre-training on translated texts resulted in an improvement of 2.1 percentage points over GBERT$_{base}$.

Although BRONCO150 is the only NER task with real-world clinical documents, there is no observable performance difference between MedBERTde, GeBERTa$_{base}$, and our clinical base models. However, the GeBERTa-Clinical$_{large}$ model outperformed the second-best model by 0.5 percentage points. The minor differences between the various models and the lack of clear advantages of clinical models can be attributed to the small size of the dataset. Interestingly, on GGPONC 2.0, various general domain models are on par with clinical models or even better. In the case of GeBERTa$_{base}$, the additional clinical pre-training decreased its performance. It is worth noting that the dataset consists of guidelines for oncologists written in a different writing style than clinical documents. Overall, the translation-based models achieved the highest results.

In the multi-label classification CLEF eHealth 2019 task, large models generally performed better. For example, GBERT$_{large}$ has an F1-Score that is around 1.6 percentage points better than its base variant. Domain-specific continuous pre-training on the translations reached strong performance across all metrics. Prioritizing a balance between precision and recall favors the GELECTRA$_{large}$ model, while GeBERTa-Clinical$_{large}$ surpasses in precision. The general domain GBERT$_{large}$ model provides the highest recall, although its precision is lower compared to others.

Looking at the difference between GBERT-Clinical$_{base}$ and GBERT-BioM-Translation$_{base}$ in Table 4, the latter achieves comparable results to its clinical counterpart on BRONCO, GGPONC 2.0, and RadQA. Only on GraSCCo the results of the GBERT models trained on the translations are worse. Compared to the baseline GBERT general models, all further pre-trained models have an advantage. In this case, we see no clear advantage of training on German clinical data over translated texts.

In summary, across almost all tasks, clinical pre-training and translation-based pre-training led to better performance than general domain models that were not explicitly trained on medical data. While the results indicate a slight advantage for models trained on our clinical data, the performance difference tends to be small, and in some cases, the translation-based models even outperform the clinical ones.

9. Discussion and Limitations

Overall, the results show that especially the addition of medical pre-training data gives a performance advantage, but not necessarily the quality of the data. For example, a 6.9 percentage points improvement in F1-Score performance was measured on the GraSCCo dataset between GBERT$_{large}$ and GBERT-Clinical$_{large}$. However, the difference between GBERT-Clinical$_{large}$ and GBERT-BioM-Translation$_{large}$ is only 2.4 percentage points. Similarly, in RadQA the difference between GBERT$_{base}$ and GBERT-BioM-Translation$_{base}$ is 1.4 percentage points, but the difference between the two further pre-trained models is only 0.1 percentage points. We conclude that the presence of medical data is crucial for pre-training, but not necessarily its quality or proximity for the downstream task. The small differences between our GeBERTa-Clinical and standard GeBERTa models further support this hypothesis, as the standard version of GeBERTA already contained translations. In this context, it is also worth noting that the difference between GBERT-Clinical and GBERT-BioM-Translation models is particularly evident on the GraSCCo task, even though it consists of synthetically generated documents and not real clinical texts. In contrast, smaller differences are evident in clinical tasks such as RadQA.

These findings open up possibilities to train clinically applicable models in different scenarios where only limited clinical data is available or cannot be accessed for privacy reasons. This is the case, for example, with low-resource languages, where far fewer clinical documents are written. Additionally, synthetic or translated public data can also help protect patient privacy by avoiding pre-training on patient data.

Despite being less resource-intensive than training from scratch, this work was able to achieve good results on various downstream tasks. This highlights the effectiveness of transfer learning and the value of pre-trained models.

As we discuss the findings and methodologies, it is important to address some of the challenges and limitations. Determining the optimal hyperparameters can be intricate. Subtle changes, such as adjusting the batch size or seed, can affect the results. As a result, direct comparisons with previous work should be made with caution, especially when score variances are minimal. Additionally, although models grounded on translations can be made public, sharing models trained on proprietary clinical data remains prohibited. This restriction is grounded in data protection measures and concerns about potential retrieval attacks that might expose the training data.

10. Conclusion

In this study, several new German biomedical and clinical language models were introduced coming from two data streams: public translation data and private clinical data from a large German hospital, with the translation-based models made publicly available to the research community. These models were assessed on five downstream tasks and compared to a variety of German and multilingual models from the general, biomedical, and clinical domains. It was demonstrated that the translation of PubMed is a promising approach that can avoid data protection concerns. In most downstream tasks, translation-based models achieved comparable results to large-scale clinical models, although only 6 million translated abstracts were utilized. In particular, clinical downstream tasks show that domain-specific pre-training can still be worthwhile, but the general domain German models, such as GBERT, can be sufficient in many domains. Despite the performance and ease of distribution for translation-based models, it is impor-

	Model	CLEF eHealth 2019			RadQA		GraSCCo		
		F1	P	R	F1	EM	F1	P	R
general	GBERT$_{base}$ (Chan et al., 2020)	.816	.818	.815	.794	.707	.642	.617	.676
	GBERT$_{large}$ (Chan et al., 2020)	.832	.802	**.865**	.809	.718	.647	.617	.680
	GottBERT (Scheible et al., 2020)	.791	.818	.765	.796	.712	.652	.681	.624
	XLM-R$_{large}$ (Conneau et al., 2020)	.804	.781	.829	.813	.731	.674	.655	.694
	GELECTRA$_{large}$ (Chan et al., 2020)	.827	.826	.828	.812	.725	.681	**.702**	.661
	mDeBERTa V3$_{base}$ (He et al., 2023)	.793	.786	.801	.810	.741	.675	.646	.706
medical	GeBERTa$_{base}$ (Dada et al., 2023)	.823	.817	.829	.839	.769	.684	**.702**	.667
	GeBERTa$_{large}$ (Dada et al., 2023)	.837	.848	.826	.834	.757	.669	.700	.640
	BioGottBERT (Lentzen et al., 2022)	.791	.779	.803	.797	.706	.637	.673	.605
	GBERT-BioM-Translation$_{base}$†	.825	.851	.801	.808	.716	.661	.642	.681
	GBERT-BioM-Translation$_{large}$†	.833	.860	.807	.811	.714	.692	.677	.707
clinical	MedBERTde (Bressem et al., 2024)	.836	.839	.833	.833	.761	.660	.626	.697
	GBERT-Clinical$_{base}$†	.833	.853	.815	.807	.726	.696	.670	.725
	GBERT-Clinical$_{large}$†	**.843**	**.876**	.812	.806	.710	**.716**	.692	**.742**
	GeBERTa-Clinical$_{base}$†	.804	.816	.792	.845	.762	.680	.699	.662
	GeBERTa-Clinical$_{large}$†	.833	.874	.796	**.846**	**.768**	.703	.677	.730

Table 3: Performance of Different Models on various Downstream Tasks: CLEF eHealth 2019 (Multi-label classification), RadQA (Extractive Questions Answering), and GraSCCo (NER). F1-Scores reported are micro-averaged. P denotes Precision, R denotes Recall, EM denotes Exact Match, † marks models that we pre-trained

	Model	BRONCO150			GGPONC 2.0		
		F1	P	R	F1	P	R
general	GBERT$_{base}$	.833 ± .004	.818 ± .002	.849 ± .011	.770	.761	.780
	GBERT$_{large}$	.835 ± .006	.820 ± .004	.852 ± .011	.772	.758	.786
	GottBERT	.840 ± .008	.827 ± .010	.854 ± .012	.756	.744	.768
	XLM-R$_{large}$	.841 ± .003	.823 ± .007	.860 ± .007	.775	.764	.786
	GELECTRA$_{large}$	.850 ± .006	**.835 ± .005**	.865 ± .007	.780	.769	.792
	mDeBERTa V3$_{base}$	.843 ± .005	.824 ± .007	.862 ± .007	.768	.753	.783
medical	GeBERTa$_{base}$	.848 ± .007	.830 ± .010	.866 ± .007	.772	.761	.783
	GeBERTa$_{large}$	.847 ± .008	.825 ± .003	.872 ± .001	.772	.758	.786
	BioGottBERT	.844 ± .011	.826 ± .012	.863 ± .013	.770	.756	.785
	GBERT-BioM-Translation$_{base}$†	.842 ± .006	.824 ± .007	.861 ± .007	.780	.766	**.794**
	GBERT-BioM-Translation$_{large}$†	.844 ± .007	.825 ± .007	.864 ± .009	**.786**	**.779**	.793
clinical	MedBERTde	.848 ± .005	.833 ± .008	.864 ± .006	.755	.744	.744
	GBERT-Clinical$_{base}$†	.847 ± .009	.828 ± .009	.867 ± .011	.772	.763	.781
	GBERT-Clinical$_{large}$†	.844 ± .007	.825 ± .007	.864 ± .009	.785	.778	.792
	GeBERTa-Clinical$_{base}$†	.846 ± .006	.823 ± .007	.870 ± .009	.768	.757	.780
	GeBERTa-Clinical$_{large}$†	**.855 ± .007**	.832 ± .006	**.878 ± .011**	.773	.760	.787

Table 4: Performance of Different Models on NER Downstream Tasks: BRONCO150 (5-fold cross validation), GGPONC 2.0. F1-Scores reported are micro-averaged. P denotes Precision, R denotes Recall, † marks models that we pre-trained

tant to recognize that in half of the tasks tested, models derived from private clinical data still performed better, highlighting the importance and effectiveness of large specialized data sources.

Future work could look at training models that use all available PubMed abstracts. In addition, we aim to explore the training of German medical large language models and leverage their capabilities.

11. Ethical Statement

This study was conducted in alignment with the principles of the Helsinki Declaration and received approval from the Institutional Review Board (IRB). Throughout the research, the team maintained transparency, integrity, and respect for the unique requirements of medical data, emphasizing patient welfare and data protection standards.

The use of language models in the medical field raises several ethical concerns. Biases in the training data can lead to poor outcomes for underrepresented groups, posing a significant issue in healthcare delivery. This highlights the need for strategies to identify and address potential biases, ensuring equitable representation and outcomes for all patient groups.

Moreover, the "black box" nature of these models complicates their use in medical decision-making, where transparency and trust are crucial. This emphasizes the importance of prioritizing interpretability and explainability in the development and application of language models, fostering trust and enabling informed decision-making.

Furthermore, the use of patient data presents challenges in informed consent and privacy, especially given the difficulties in securing individual consent for large datasets and the inability for patients to opt-out post-training. This underscores the necessity for robust data protection measures, adherence to relevant privacy regulations, and ensuring that patient data is handled with the utmost care and confidentiality.

These issues emphasize the need for careful ethical considerations in the deployment of language models in healthcare, and this study aimed to address these concerns through a comprehensive and ethically grounded approach.

12. Acknowledgement

The work of Ahmad Idrissi-Yaghir and Henning Schäfer was funded by a PhD grant from the DFG Research Training Group 2535 Knowledge- and data-based personalisation of medicine at the point of care (WisPerMed).

13. Bibliographical References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Aman Berhe, Guillaume Draznieks, Vincent Martenot, Valentin Masdeu, Lucas Davy, and Jean-Daniel Zucker. 2023. AliBERT: A pre-trained language model for French biomedical text. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 223–236, Toronto, Canada. Association for Computational Linguistics.

Florian Borchert, Christina Lohr, Luise Modersohn, Thomas Langer, Markus Follmann, Jan Philipp Sachs, Udo Hahn, and Matthieu-P. Schapranow. 2020. GGPONC: A corpus of German medical text with rich metadata based on clinical practice guidelines. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 38–48, Online. Association for Computational Linguistics.

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

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

Casimiro Pio Carrino, Joan Llop, Marc Pàmies, Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Joaquín Silveira-Ocampo, Alfonso Valencia, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. Pretrained biomedical language models for clinical NLP in Spanish. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 193–199, Dublin, Ireland. Association for Computational Linguistics.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Amin Dada, Aokun Chen, Cheng Peng, Kaleb Smith, Ahmad Idrissi-Yaghir, Constantin Seibold, Jianning Li, Lars Heiliger, Christoph Friedrich, Daniel Truhn, Jan Egger, Jiang Bian, Jens Kleesiek, and Yonghui Wu. 2023. On the impact of cross-domain data on German language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13801–13813, Singapore. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Antje Dörendahl, Nora Leich, Benedikt Hummel, Gilbert Schönfelder, and Barbara Grune. 2019. Overview of the clef ehealth 2019 multilingual information extraction. In *Working Notes of CLEF 2019-Conference and Labs of the Evaluation Forum*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Heal.*, 3(1):2:1–2:23.

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8342–8360. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1).

Liadh Kelly, Hanna Suominen, Lorraine Goeuriot, Mariana Neves, Evangelos Kanoulas, Dan Li, Leif Azzopardi, Rene Spijker, Guido Zuccon, Harrisen Scells, and João Palotti. 2019. Overview of the CLEF eHealth evaluation lab 2019. In *Lecture Notes in Computer Science*, pages 322–339. Springer International Publishing.

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

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. Drbert: A robust pre-trained model in french for biomedical and clinical domains. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, pages 16207–16221. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Luise Modersohn, Stefan Schulz, Christina Lohr, and Udo Hahn. 2022. GRASCCO — the first publicly shareable, multiply-alienated german clinical text corpus. In *Studies in Health Technology and Informatics*. IOS Press.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of biomedical semantics*, 9(1):12.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Malte Ostendorff, Till Blume, and Saskia Ostendorff. 2020. Towards an Open Platform for Legal Information. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, page 385–388, New York, NY, USA. Association for Computing Machinery.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Raphael Scheible, Fabian Thomczyk, Patric Tippmann, Victor Jaravine, and Martin Boeker. 2020. Gottbert: a pure german language model. *CoRR*, abs/2012.02110.

Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra. 2020. BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72, Online. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp

Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Torsten Zesch and Jeanette Bewersdorff. 2022. German medical natural language processing–a data-centric survey. In *The Upper-Rhine Artificial Intelligence Symposium UR-AI 2022 : AI Applications in Medicine and Manufacturing, 19 October 2022, Villingen-Schwenningen, Germany*, pages 137–145. Furtwangen University.