Can Continual Pretraining Bridge the Performance Gap between General-purpose and Specialized Language Models in the Medical Domain?

Anonymous ACL submission

Abstract

This paper narrows the performance gap between small, specialized models and significantly larger general-purpose models through domain adaptation via continual pre-training and merging. We address the scarcity of specialized non-English data by constructing a high-quality German medical corpus (*FineMed-de*) from *FineWeb2*. This corpus is used to continually pre-train and merge three well-known LLMs (ranging from 7B to 24B parameters), creating the *DeFineMed* model family. A comprehensive evaluation confirms that specialization dramatically enhances 7B model performance on German medical benchmarks. Furthermore, the pairwise win-rate analysis of the *Qwen2.5*-based models demonstrates an approximately 3.5-fold increase in the win-rate against the much larger *Mistral-Small-24B-Instruct* through domain adaptation. This evidence positions specialized $7B$ models as a competitive, resource-efficient solution for complex medical instruction-following tasks. While model merging successfully restores instruction-following abilities, a subsequent failure mode analysis reveals inherent trade-offs, including the introduction of language mixing and increased verbosity, highlighting the need for more targeted fine-tuning in future work. This research provides a robust, compliant methodology for developing specialized LLMs, serving as the foundation for practical use in German-speaking healthcare contexts.

1 Introduction

Large language models (LLMs) have demonstrated transformative potential across various domains, including healthcare (Paass and Giesselbach, 2023; Zhang et al., 2024). Their ability to process and generate human-like text enables applications such as diagnostics, personalized treatment plans, and efficient information retrieval (Meyer et al., 2024). However, significant gaps remain in the integration of LLMs into clinical workflows, where general-purpose models often fail to capture domain-specific knowledge and terminology with sufficient accuracy (Klug et al., 2024; Alonso et al., 2024).

The effective application of LLMs in the medical field faces two significant, interconnected challenges. First, stringent data protection regulations necessitate on-premise solutions, making the use of large, API-based LLM services impractical and favoring smaller, computationally efficient models (Belcak et al., 2025). Second, these smaller models, while regulatory compliant, struggle to capture the complex, nuanced medical terminology due to a scarcity of high-quality, domain-specific datasets, a problem acutely felt in non-English languages (German Federal Ministry of Health, 2024). This creates a critical trade-off: regulatory constraints demand small models, underscoring the necessity of a targeted, specialized knowledge base to achieve competitive clinical performance.

This work directly addresses this challenge by investigating the core research question: *Can domain-adaption of $7B$ language models, achieved via continual pre-training and merging, close the performance gap sufficiently to compete with significantly larger, general purpose models on complex medical tasks?* Beyond technical constraints, this study aims to quantify how effectively domain knowledge, gained without massive retraining, can be internalized to transform resource-efficient models into viable clinical tools.

This paper directly addresses these challenges by developing a specialized German medical LLM through a systematic approach to dataset creation and model adaptation, as illustrated in Figure 1. Our key contributions are as follows:

- A robust and scalable filtering methodology for creating high-quality, domain-specific datasets from general-purpose corpora. This methodology combines LLM-based annota-

tion with machine learning techniques and is applicable to various domains and languages.

- A comprehensive evaluation, utilizing knowledge-intensive benchmarks (*MMLU-de* and *MedQA-de*) and a pairwise win-rate analysis, demonstrating that specialized $7B$ models drastically close the performance gap toward $24B$ models. This validates the competitive viability of smaller, resource-efficient LLMs, while a simultaneous failure mode analysis highlights the critical trade-offs inherent to the specialization methodology.

The remainder of this paper is structured as follows: Section 2 provides an overview of the state of the art. Section 3 outlines the construction of the German medical pre-training dataset. Further, Section 4 describes the model adaptation techniques employed. The evaluation methodology and results are presented in Section 5, followed by a discussion of the observed results in section 6. Finally, Section 7 concludes the paper with a summary of our findings and potential future directions.

2 Related Work

This work intersects several key areas of research in medical LLM development, including model adaptation through continual pre-training and model merging techniques, as well as dataset curation through document filtering. To provide a comprehensive context, we review relevant literature in each of these domains, highlighting advancements and methodologies that align with our approach.

2.1 Model Adaptation through Continual Pretraining

Recent studies have explored continual pretraining as a means to adapt general-purpose LLMs to specialized domains, but results remain inconsistent. Öncel et al. (2024) showed that additional pretraining can fail, or even harm performance, when domain data diverges from the model's original distribution, underscoring the importance of corpus quality and alignment. In contrast, research in the legal and financial domains (Niyogi and Bhattacharya, 2024; Siriwardhana et al., 2024) suggests that combining continual pretraining with model merging can yield competitive improvements while preserving general reasoning capabilities. Similarly, Yang et al. (2024b) showed that targeted continual pretraining with efficient adaptation strategies enhances domain fluency, suggesting that success depends less on scale and more on data selection and adaptation design.

2.2 Medical Large Language Models

Recent progress in Medical LLMs has been achieved through pretraining (Luo et al., 2022; Peng et al., 2023), fine-tuning (Singhal et al., 2025; Toma et al., 2023; Han et al., 2023; Zheng et al., 2025), and prompting (Nori et al., 2023; Liu et al., 2023), yielding models that match or surpass human experts on medical *QA* tasks (Singhal et al., 2025; Nori et al., 2023).

Labrak et al. (2024) advanced this line with *BioMistral*, a continually pretrained biomedical model based on Mistral (Jiang et al., 2023) and PubMed-Central data (National Library of Medicine, 2003), outperforming open-source baselines. Our work fundamentally differs from *BioMistral* not only in targeting the German language but in its primary research objective. Where *BioMistral* aims for performance improvements on established, knowledge-intensive benchmarks, our study investigates the competitive viability of specialized small LMs against significantly larger general-purpose models and extends the evaluation beyond knowledge-intensive benchmarks.

More recently, Zheng et al. (2025) introduced *Apollo-2*, a multilingual medical LLM family covering 50 languages, focusing on instruction-tuning and modular routing rather than continual pretraining for domain adaptation.

2.3 Merging Large Language Models

Model merging provides an efficient alternative to computationally expensive fine-tuning and ensembling by combining pretrained LLMs fine-tuned on specialized tasks. The survey by Yang et al. (2024a) presents a taxonomy of model merging approaches, their applications in various domain subfields, and their challenges. Additionally, Nobari et al. (2025) propose *Activation-Informed Merging*, leveraging activation-space information to improve the robustness and efficiency of merging, showing up to a 40% performance gain across benchmarks.

Alternative strategies further explore computational efficiency and generalization. Gupta and Gupta (2024) introduce a task-vector-based approach using LoRA-derived representations and geometric median aggregation, enabling effective merging with reduced computational costs. Meanwhile, Yadav et al. (2024) conducted large-scale

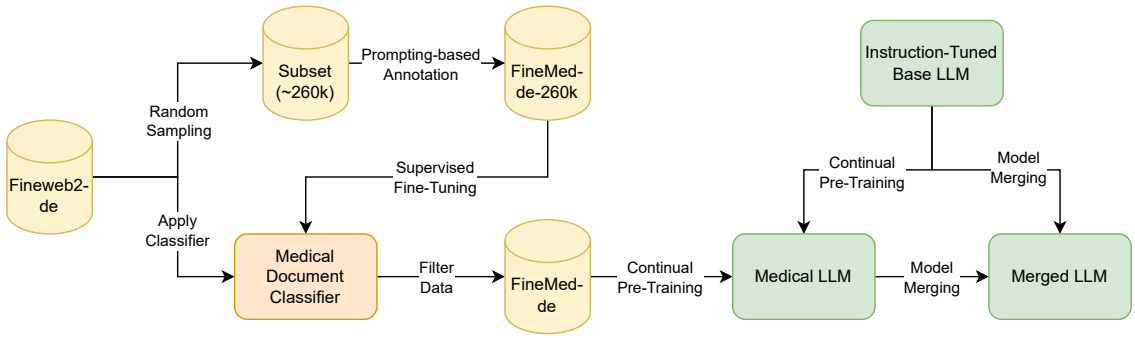

Figure 1: High-level illustration of the Data Filtering and Model Adaptation workflow: a subset of the German *FineWeb2* dataset is annotated into medical and non-medical documents. The resulting annotated dataset, *FineMed-de-260k*, is used to train a classifier, which is then applied to the full German *FineWeb2* split, producing the *FineMed-de* dataset. This dataset is subsequently used for continual pretraining - resulting in a Medical LLM - which is then merged with the initial instruction-tuned checkpoint.

experiments, merging models up to 64B parameters, showing that merging enhances zero-shot generalization and is particularly effective when starting from strong base models. Their findings suggest that larger models facilitate more successful merging, and different merging strategies perform similarly at scale. Collectively, these works highlight the potential of model merging for creating high-performing, resource-efficient LLMs.

In this work, we employ Spherical Linear Interpolation (SLERP) (Shoemake, 1985) for model merging, a technique shown by Goddard et al. (2024) to achieve the best performance in the medical domain.

3 Medical Filtering Pipeline

This section details the creation of the *FineMed-de* corpus. The high level process is illustrated in Figure 1. Unlike existing curation approaches that rely solely on either classical ML or LLMs, we apply a hybrid method. By using LLMs to generate high-quality labels and classical ML techniques to scale the filtering process to millions of documents, our approach aims to achieve both quality and efficiency in medical dataset curation.

3.1 Source Data

The source dataset utilized in this study is *FineWeb2* (Penedo et al., 2024), a web-crawled dataset designed to improve accessibility for training large language models. *FineWeb2* is built upon *CommonCrawl*[1], encompassing a wide variety of web content including forum posts, blog articles, and news articles. The original *FineWeb2* dataset is partitioned by language using the *GlotLID* language classifier (Kargaran et al., 2023), with the German subset being a substantial portion, ranking as the third largest at $640GB$. *FineWeb2* is licensed under the *Open Data Commons Attribution License (ODC-By) v1.0*, allowing for its open use and distribution. Our objective is to filter this subset to extract exclusively medical documents for pre-training specialized German medical LLMs.

3.2 Medical Document Classifier

To create a robust medical pre-training corpus, we first developed a medical document classifier. This involved constructing a supervised dataset by sampling approximately $260k$ documents from the German *FineWeb2* dataset. We then employed the *Mixtral-8x7B-Instruct-v0.1* (Jiang et al., 2024a) model to partition these documents into medical and non-medical categories using a zero-shot prompting approach. At the time of our study, this model offered a combination of strong multilingual performance and computational efficiency, making it a practical choice for our classification task. The prompt used is provided in Appendix A.1. The output of the LLM was validated through manual inspection of 100 random samples by three human annotators, yielding an F1 score of 91.1 ± 2.5. The manual inspection process and further evaluation of the LLM's performance is detailed in Appendix A.1. The resulting labeled dataset, named *FineMed-de-260k*, was then randomly split into training and testing sets, as detailed in Table 1.

To achieve a cost-effective and scalable solution for classifying the entire pre-training corpus, we train a significantly smaller and more efficient clas-

[1] https://commoncrawl.org/

	#Documents		#Words	
	Med	**Other**	**Med**	**Other**
Test	$4.9k$	$21.5k$	$8.7M$	$20.3M$
Train	$44.1k$	$193.5k$	$81.9M$	$181.8M$
Total	$49.0k$	$215.0k$	$90.6M$	$202.1M$

Table 1: Label distribution between medical (Med) and non-medical (Other) domains of the *FineMed-de-260k* dataset in terms of number of documents and words.

Dataset	#Documents	#Words
FineMed-de	$7.3M$	$5.1B$
FineWeb2-de	$427.7M$	$234.8B$

Table 2: Dataset size of *FineMed-de* compared to the original *FineWeb2-de* in number of documents and number of words.

sifier based on this annotated dataset. Specifically, we fine-tune a $279M$ parameter *xlm-roberta* model (Conneau et al., 2020) as a medical document classifier, achieving a precision of 0.95 and recall of 0.8 on the test set. Details of the training procedure can be found in Appendix A.2.

3.3 Medical Pre-Training Corpus

The application of the classifier to the entire German *FineWeb2* dataset required approximately 400 GPU hours, distributed across 8 A100 (40GB) GPUs on the Karolina cluster.[2] The resulting dataset, which we named *FineMed-de*, contains roughly 7.3 million documents. The statistics of the resulting dataset are presented in Table 2.

4 Model Adaptation

To create specialized German medical LLMs, we select three source models for adaptation via continual pre-training and model merging. We name this family of adapted models *DeFineMed*, a combination of "De" for German and *FineMed*, our dataset used for continual pre-training. The decision to start from instruction-tuned models for our adaptation process is motivated by both methodological precedent and empirical evidence. Following the setup of the BioMistral model family (Labrak et al., 2024), which also relies on instruction-tuned checkpoints as the basis for continual pre-training and model merging, we adopt a similar strategy to ensure alignment with established practices in the domain. At the same time, we acknowledge that the optimal sequencing of instruction tuning and continued pre-training remains an open research question, with recent studies reporting conflicting findings on whether instruction tuning should precede or follow domain adaptation (Jindal et al., 2024; Jiang et al., 2024b).

4.1 Source Models

The foundation of our model adaptation process is the selection of appropriate source models. We prioritize models with strong multilingual capabilities to establish a robust base for understanding and generating text in our target language. Specifically, we focus on *Qwen2.5-7B-Instruct* and *Mistral-7B-Instruct*, both of which offer a balance between performance and computational efficiency within the $7B$ parameter range. To investigate how our process scales with model size, we also included *Mistral-Small-24B-Instruct* with 23.6B parameters. The complete list of models used in this work is provided in the Table 3.

4.2 Continual Pre-training

To adapt the source models to the medical domain, we perform continual pre-training using the *FineMed-de* dataset from Section 3. We leverage the Hugging Face *Transformers* library (Wolf et al., 2020) alongside the *Accelerate* library (Gugger et al., 2022) for efficient distributed training. To optimize both memory usage and computational performance, we incorporate several advanced techniques such as *Fully Sharded Data Parallelism* (FSDP) (Zhao et al., 2023), *Flash Attention* (Dao, 2023), *Activation Checkpointing*, *Sequence Packing* (Ding et al., 2024), and mixed-precision training with *bfloat16*. We train all models for a total of two epochs using the *AdamW* optimizer (Loshchilov and Hutter, 2019) with a linear learning rate decay. In order to address training instabilities, we implement an extended warmup phase of 500 steps. A detailed description of the pre-training procedure is given in Appendix A.3.

4.3 Model Merging

Inspired by the observations from Labrak et al. (2024), where merging models led to improved performance in various tasks, we employ model merging after training. Following continual pre-training, we apply model merging to mitigate catastrophic forgetting and restore instruction-following abilities. By merging it with its original instruction-tuned version we expect to recover the instruction

[2] https://www.it4i.cz/en/infrastructure/karolina

Model	Parameters	Vocab Size	License	Source
Mistral-7B-Instruct-v0.3	$7.25B$	$32k$	*Apache-2.0*	(Jiang et al., 2023)
Qwen2.5-7B-Instruct	$7.62B$	$152k$	*Apache-2.0*	(Qwen et al., 2025)
Mistral-Small-24B-Instruct	$23.6B$	$131k$	*Apache-2.0*	(Mistral AI, 2025)

Table 3: Source model information, including model size, vocabulary size, license citations.

following ability (Yang et al., 2024a). Each continually pre-trained model is merged with its corresponding instruction-tuned base model using the Mergekit framework (Goddard et al., 2024). Specifically, we employ SLERP (Shoemake, 1985) with both checkpoints weighted equally ($t = 0.5$), as it has been shown to outperform alernative strategies in the medical domain Goddard et al. (2024). This approach provides an efficient means of preserving generalization capabilities without additional fine-tuning.

5 Evaluation

This chapter presents a comprehensive evaluation of the model checkpoints across three dimensions: benchmark performance, pairwise competitive analysis, and quantitative failure mode assessment. Our analysis systematically investigates the impact of continual pre-training and model merging on both domain-specific accuracy and output quality.

We use *BioMistral* as a comparative baseline throughout our evaluation due to its similar domain adaptation technique. *Apollo-2* is excluded as a direct baseline due to its focus on instruction-tuning as its primary specialization mechanism.

5.1 Knowledge-Intensive Benchmark Evaluation

We evaluate model performance across three checkpoints for each selected source model—the initial instruction-tuned model, the continually pre-trained model, and the final merged model. Performance is assessed using the LM Evaluation Harness (Gao et al., 2024; Biderman et al., 2024) on two established medical benchmarks: (1) *MMMLU*, specifically focusing on the German medical tasks (Anatomy, Clinical Knowledge, and College Medicine), and (2) *MedQA-de*, a machine-translated German version of the *MedQA* dataset (Jin et al., 2021) consisting of 500 medical exam questions. To ensure that domain adaptation does not compromise general reasoning or linguistic capabilities, we additionally evaluate all models on a set of general knowledge benchmarks, with the corresponding results reported in Appendix A.5 and in Table 10.

Base Models achieve very different performance across the benchmarks. *Qwen2.5-7B-Instruct* outperforms the *Mistral-7B-Instruct* model by 9.34% on average accross all benchmarks, likely due to more modern data pipelines and pre-training methodologies. Notably, the *Mistral-Small-24B-Instruct* model achieves the best overall results, demonstrating a significant increase in performance, particularly on *MedQA-de*, where it outperforms the second-best model by 18.60%. We observed that even our weakest base model outperforms the *BioMistral* baseline on three out of four benchmarks. This is likely due to the primarily English corpus used to adapt the model.

Continual Pre-Training on domain-specific data consistently improves performance across multilingual models. *DeFineMed-Mistral-7B* exhibits the strongest relative performance increase from continual pre-training, with an average improvement of 6.61%. *DeFineMed-Mistral-Small-24B* also benefits from continual pre-training, though to a lesser extent than smaller models, with an average gain of 1.57%. The impact of using domain-specific data is validated through an ablation study designed to isolate and quantify the benefit of our medical corpus compared to an unfiltered baseline, as detailed in Section A.4.

Model Merging shows mixed results. While it consistently improved performance on the *MedQA* benchmark and showed consistent, though in some cases marginal, gains for the *DeFineMed-Mistral-Small-24B-SLERP* model, other combinations of models and benchmarks did not reflect a clear trend. For instance, the performance of *DeFineMed-Qwen-2.5-7B* on the *Clinical Knowledge* and *College Medicine* benchmarks dropped after merging. On average, model merging slightly improved performance across the benchmarks, offering the advantage of partially restoring instruction-following abilities that were lost during continual pre-training.

Model	Anatomy	Clinical Knowledge	College Medicine	MedQA-de	Average
BioMistral-7B (baseline)	44.44 ± 4.29	52.08 ± 3.07	40.46 ± 3.74	37.20 ± 2.16	43.55 ± 1.70
BioMistral-7B-SLERP (baseline)	45.93 ± 4.30	58.49 ± 3.03	50.87 ± 3.81	37.60 ± 2.17	48.22 ± 1.71
Mistral-7B-Instruct-v0.3	42.22 ± 4.27	61.13 ± 3.00	53.18 ± 3.80	42.40 ± 2.21	49.73 ± 1.71
DeFineMed-Mistral-7B	57.78 ± 4.27	65.28 ± 2.93	55.49 ± 3.79	46.80 ± 2.23	56.34 ± 1.70
DeFineMed-Mistral-7B-SLERP	53.33 ± 4.31	65.66 ± 2.92	56.65 ± 3.78	50.20 ± 2.24	56.46 ± 1.70
Qwen2.5-7B-Instruct	54.07 ± 4.30	69.06 ± 2.85	63.01 ± 3.68	50.20 ± 2.24	59.08 ± 1.68
DeFineMed-Qwen2.5-7B	62.22 ± 4.19	76.60 ± 2.61	68.21 ± 3.55	52.60 ± 2.24	64.91 ± 1.62
DeFineMed-Qwen2.5-7B-SLERP	65.19 ± 4.12	75.85 ± 2.63	65.90 ± 3.61	54.20 ± 2.23	65.28 ± 1.62
Mistral-Small-24B-Instruct	66.67 ± 4.07	78.87 ± 2.51	73.41 ± 3.37	68.80 ± 2.07	71.94 ± 1.55
DeFineMed-Mistral-Small-24B	68.15 ± 4.02	79.62 ± 2.48	77.46 ± 3.19	68.80 ± 2.07	73.51 ± 1.52
DeFineMed-Mistral-Small-24B-SLERP	68.89 ± 4.00	82.64 ± 2.33	78.03 ± 3.16	70.20 ± 2.05	74.94 ± 1.49

Table 4: Model performance on different German benchmarks in one-shot setting. The average accuracy is the unweighted mean of accuracies, whereas the standard error of the average is the square root of the unweighted mean of variances. Best score per model family is underlined.

In terms of overall performance, *DeFineMed-Mistral-Small-24B-SLERP* emerges as the strongest model, achieving an average accuracy of 74.94% and significantly outperforming all other models. Notably, *DeFineMed-Qwen2.5-7B-SLERP* demonstrates exceptionally strong performance for its size, performing comparable to the much larger non-specialized *Mistral-Small-24B-Instruct* model on the *Anatomy* and *Clinical Knowledge* benchmarks.

5.2 Pairwise Win-Rate Evaluation

This section presents the comparative performance of model checkpoints through pairwise win-rate analysis, quantified using the LLM-as-a-Judge methodology using *GPT-4.1-mini* (OpenAI, 2025) on the a machine-translated version of the *MedAlpaca* dataset (Han et al., 2025). Given the requirement for instruction-following abilities, our analysis concentrates on the base instruction-tuned models and the final merged checkpoints.

The results are summarized in Figure 2. The analysis confirms that model scale is the most significant determinant of output quality. The two $24B$ models overwhelmingly dominate all $7B$ models, establishing a high performance ceiling for the task.

Domain adaptation is highly successful in the $7B$ range. Pairwise comparisons show that the merged checkpoints consistently and significantly outperform their instruction-tuned base counterparts. For instance, *DeFineMed-Qwen2.5-7B-SLERP* wins against *Qwen2.5-7B-Instruct* by a rate of 0.66 versus 0.33. This validates model merging as a highly effective technique for applying domain knowledge while retaining core instruction capabilities at this scale. Among the $7B$ models, the *Qwen2.5* architecture demonstrates superior base performance over the *Mistral* architecture.

Crucially, specialization drastically closes the performance gap toward the largest model. The merged *DeFineMed-Qwen2.5-7B-SLERP* model achieves a win-rate of 0.31 against the much larger *Mistral-Small-24B-Instruct*, representing a significant, 3.5-fold increase from the base *Qwen2.5-7B-Instruct* win-rate of 0.09 against the same model. This demonstrates that the fusion of domain knowledge makes the smaller $7B$ model substantially more competitive against $24B$ models.

Conversely, the applied domain adaptation technique proved detrimental at the $24B$ scale, where the base *Mistral-Small-24B-Instruct* model outperforms its merged counterpart. This suggests that the performance drop at larger models may be partly attributable to limitations in the applied merging method, although additional factors related to large-scale adaptation cannot be excluded.

5.3 Quantitative Evaluation of Failure Modes

This section presents a quantitative analysis of model output quality by analyzing failure modes. We again make use of the machine-translated version of the *MedAlpaca* dataset (Han et al., 2025) and employ *GPT-4.1* to systematically quantify failure modes of LLM responses. Specifically, we analyze the frequency of failure modes including language mixing, typos, contradiction, hallucination and verbosity among others (see Table 11 in the appendix for detailed descriptions). Figure 3 summarizes the distribution of these failure modes across the selected model checkpoints. Overall, the results indicate that specialization via continual pre-training and subsequent model merging often mitigates a majority of common failure modes.

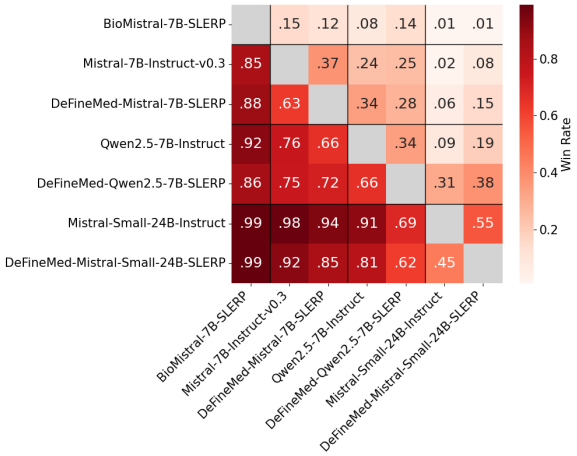

Figure 2: Matrix of model win-rates on the German *MedAlpaca* dataset, where the value represents the win-rate of the row model over the column model in a head-to-head comparison.

Across the *Mistral-7B* and *Qwen2.5-7B* families, a clear trend of mitigation is observed in failure modes such as hallucination, omission, overgeneralization, and repetition after the merging step. Furthermore, the integration of domain knowledge via merging significantly improves the models' epistemic calibration. Both overconfidence and contradiction errors are dramatically reduced. This suggests the specialized models are less likely to confidently state incorrect or conflicting information, leading to more factually grounded outputs.

Conversely, pre-training and merging can promote or introduce specific failure modes. The most extreme case is the language mixing failure mode in the *DeFineMed-Mistral-7B-SLERP* model, which exhibited 207 of the 216 instances, a drastic increase from the 12 instances in the base *Mistral-7B-Instruct-v0.3*. This phenomenon aligns with prior qualitative observations and indicates a strong, yet localized, negative impact of multilingual domain adaptation on language separation. However, this trend is not observed in the *Qwen-7B* models.

We also observe an inverse relationship between omission and verbosity among the $7B$ models. The desire to be more comprehensive (reducing omission) often results in outputs that are significantly more verbose. This increase in verbosity is expected, as the models were continually pre-trained on a rich, verbose corpus of specialized text. The merging technique is not potent enough to fully counteract this learned behavior, suggesting that more involved instruction-tuning methods may be necessary to shape the desired output.

The occurrence of typos showed a contradictory pattern: a drastic decrease from 118 to 41 for the *Qwen2.5-7B* family following specialization and merging, but a drastic increase from 85 to 125 for the *Mistral-7B* family. This suggests that the impact on low-level language fluency is highly dependent on the underlying base model.

Finally, the $24B$ model family demonstrates remarkable stability. The failure mode counts for the merged checkpoint show no significant changes from its base model counterpart, with all counts remaining low. This observation is consistent with findings that larger models are more robust to failure modes introduced during post-pre-training adaptation, including perturbations from parameter merging or fine-tuning (Wei et al., 2022; Yadav et al., 2024).

6 Discussion

A key element enabling our results is the construction of a large-scale German medical pre-training corpus from the *FineWeb2 dataset*, achieved through a combination of LLM-based filtering and traditional ML techniques. This process directly addresses the limited availability of high-quality datasets and provides the necessary knowledge foundation for small model effectiveness.

The evidence from our evaluation directly addresses the question of whether small, domain-adapted models can challenge the performance of significantly larger counterparts. While a fully decisive answer requires further investigation, our findings represent a major step forward, demonstrating that the $7B$ models, enhanced via continual pre-training and merging, significantly closed the knowledge and performance gap against the $24B$ model.

In our quantitative benchmark analysis, the *DeFineMed-Qwen2.5-7B-SLERP* model demonstrated performance comparable to the much larger *Mistral-Small-24B-Instruct* on three out of four domain-specific benchmarks. More critically, the pairwise win-rate evaluation quantified this competitive gain: the specialization and merging process enabled the *DeFineMed-Qwen2.5-7B-SLERP* to increase its win-rate against the *Mistral-Small-24B-Instruct* by approximately 3.5-fold. This marked reduction in the performance disparity supports the viability of adopting smaller, specialized models as a competitive, resource-efficient alternative to deploying large, general-purpose models for complex

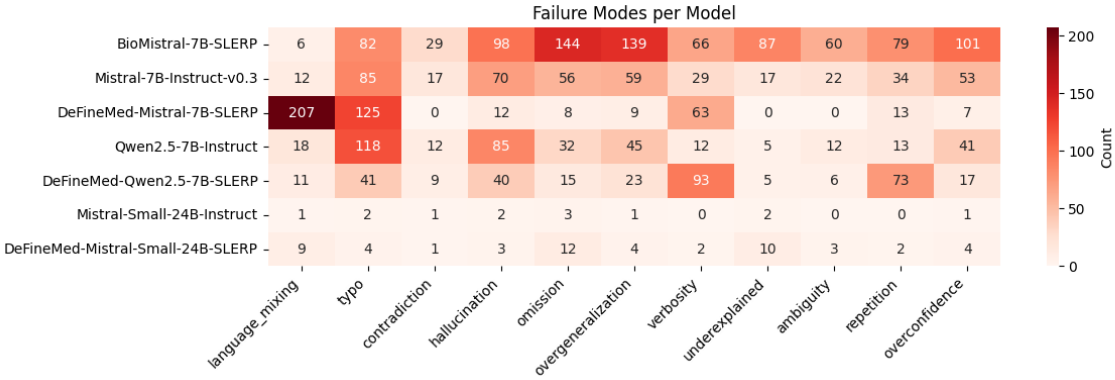

Figure 3: Frequency count of distinct failure modes for base instruction-tuned and merged models, quantified using *GPT-4.1* methodology on 216 instances from the German-translated *MedAlpaca* dataset.

medical instruction-following tasks.

Despite the successful performance gains, the specialization process highlighted critical trade-offs and limitations of the current techniques. The benchmark evaluation demonstrated that model merging, while successfully restoring instruction-following, yielded only marginal average improvements and inconsistent benefits, suggesting that merging alone is insufficient to fully harness the potential of domain-adapted checkpoints.

The quantitative analysis of failure modes further detailed these shortcomings. While specialization effectively mitigated factual errors like hallucination and overconfidence, it introduced mode-specific failures, most notably language mixing and increased verbosity due to the nature of pre-training. These findings underscore the inherent trade-off between domain adaptation and instruction-following fidelity. Furthermore, at the $24B$ scale, the merged model underperformed relative to the base model, indicating possible limitations of the applied merging technique in this setting. This suggests limitations in applying simple merging to complex, highly-capable models. Collectively, these findings highlight that merging must be complemented by more targeted instruction-tuning technique to fully harness the potential of specialized models.

7 Conclusion

In this paper, we successfully demonstrated a systematic approach to creating specialized German medical large language models, addressing the core question of whether small, resource-efficient models can be adapted to compete with significantly larger, general-purpose counterparts.

Our approach involved the creation of *FineMedde*, a large-scale German medical dataset, which was then used to successfully adapt three state-of-the-art LLMs ranging from $7B$ to $24B$ parameters. Our comprehensive evaluation confirmed the competitive viability of the $7B$ models. The pairwise win-rate analysis was particularly decisive, showing that domain adaptation and subsequent merging enabled the *Qwen2.5*-based $7B$ model to achieve an approximately 3.5-fold increase in its win-rate against the *Mistral-Small-24B-Instruct* model. These results confirm that well-adapted smaller models represent a competitive and resource-efficient alternative for environments constrained by computation and regulation.

However, the evaluation also revealed critical trade-offs. While model merging successfully restored instruction-following capabilities, our failure mode analysis highlighted side effects such as language mixing and increased verbosity. These findings suggest that merging alone is insufficient to fully realize the potential of domain-adapted models. Moving forward, future research should prioritize domain-specific instruction tuning to mitigate observed failure modes observed.

Limitations

Despite the advancements achieved through continual pre-training and model merging, several limitations must be considered when interpreting the results.

One key limitation is the pre-training corpus, which was derived from *FineWeb2* by filtering for German medical content. While this approach allowed us to construct a domain-specific dataset tailored to our needs, it also means that the dataset inherits the limitations of *FineWeb2*. These include

potential biases, misinformation, and the presence of harmful content. Notably, these issues persist despite the existing filtering efforts of the *FineWeb2* authors, highlighting the challenges of fully eliminating such limitations. Consequently, our work inherits these limitations.

Another challenge lies in the process of model merging. While merging helped restore the instruction-following abilities that were lost during continual pre-training, its benefits were not consistent across different tasks. Moreover, the merging process introduced new failure modes. These unintended side effects underscore the need for further refinement after model merging to ensure more reliable and predictable improvements in model performance.

In addition, our benchmark evaluation primarily relied on the German subsets of two widely used benchmarks: *MMMLU* and *MedQA*, which assess medical language understanding. However, certain tasks within these benchmarks, such as *College Medicine* in *MMMLU*, are known to contain a high proportion of incorrectly labeled samples (Gema et al., 2025).

While the pairwise win-rate and failure mode analyses provide valuable insights into relative model behavior and output quality, they also entail several limitations. First, the LLM-as-a-Judge evaluation relies on *GPT-4.1-mini* as the scoring model, which introduces potential bias and may not perfectly align with human judgment, particularly for multilingual or domain-specific content. Furthermore, pairwise comparisons capture relative preference, not absolute accuracy, and may favor stylistic similarity to the judge model over true factual correctness. Second, the failure mode quantification depends on automated classification of response types, which may not fully capture nuanced language or reasoning mistakes. The analysis is also limited by the scope of the *MedAlpaca* dataset, which may not capture the full range of medical or general instruction-following scenarios. Future work should include human expert evaluations, cross-judge consistency checks, and domain-specific annotation frameworks to validate and extend these findings.

References

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Artificial Intelligence in Medicine*, 155:102938.

Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. 2025. Small language models are the future of agentic ai. *Preprint*, arXiv:2506.02153.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, and 11 others. 2024. Lessons from the trenches on reproducible evaluation of language models.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

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. *Preprint*, arXiv:1911.02116.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *Preprint*, arXiv:2307.08691.

Hantian Ding, Zijian Wang, Giovanni Paolini, Varun Kumar, Anoop Deoras, Dan Roth, and Stefano Soatto. 2024. Fewer truncations improve language modeling. *arXiv preprint arXiv:2404.10830*.

DiscoResearch, Occiglot, DFKI, and hessian.Ai. 2024. Llama3-german-8b: A german language model based on meta's llama3-8b. `https://huggingface.co/DiscoResearch/Llama3-German-8B`. Accessed: 2025-03-11.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. A framework for few-shot language model evaluation.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani,

Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile van Krieken, and Pasquale Minervini. 2025. Are we done with mmlu? *Preprint*, arXiv:2406.04127.

German Federal Ministry of Health. 2024. Machbarkeit einer deutschen mimic. https://www.bundesgesundheitsministerium.de/service/publikationen/details/mimic.html. Accessed: 2025-03-11.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vladimir Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. Arcee's MergeKit: A toolkit for merging large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 477–485, Miami, Florida, US. Association for Computational Linguistics.

Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. 2022. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate.

Siddharth Gupta and Aakash Gupta. 2024. Model merging using geometric median of task vectors. In *LLM Merging Competition at NeurIPS 2024*.

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexei Figueroa, Alexander Löser, Daniel Truhn, and Keno K. Bressem. 2025. Medalpaca – an open-source collection of medical conversational ai models and training data. *Preprint*, arXiv:2304.08247.

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K. Bressem. 2023. Medalpaca – an open-source collection of medical conversational ai models and training data. *Preprint*, arXiv:2304.08247.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024a. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Zhengbao Jiang, Zhiqing Sun, Weijia Shi, Pedro Rodriguez, Chunting Zhou, Graham Neubig, Xi Lin, Wen-tau Yih, and Srini Iyer. 2024b. Instruction-tuned language models are better knowledge learners. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5421–5434, Bangkok, Thailand. Association for Computational Linguistics.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14).

Ishan Jindal, Chandana Badrinath, Pranjal Bharti, Lakkidi Vinay, and Sachin Dev Sharma. 2024. Balancing continuous pre-training and instruction fine-tuning: Optimizing instruction-following in llms. *Preprint*, arXiv:2410.10739.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. GlotLID: Language identification for low-resource languages. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Katrin Klug, Katharina Beckh, Dario Antweiler, Nilesh Chakraborty, Giulia Baldini, Katharina Laue, René Hosch, Felix Nensa, Martin Schuler, and Sven Giesselbach. 2024. From admission to discharge: a systematic review of clinical natural language processing along the patient journey. *BMC Medical Informatics and Decision Making*, 24(1):238.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *Preprint*, arXiv:2402.10373.

Zhengliang Liu, Yue Huang, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Yiwei Li, Peng Shu, Fang Zeng, Lichao Sun, Wei Liu, Dinggang Shen, Quanzheng Li, Tianming Liu, Dajiang Zhu, and Xiang Li. 2023. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *Preprint*, arXiv:2303.11032.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for

biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Annika Meyer, Janik Riese, and Thomas Streichert. 2024. Comparison of the performance of gpt-3.5 and gpt-4 with that of medical students on the written german medical licensing examination: Observational study. *JMIR Med Educ*, 10:e50965.

Team Mistral AI. 2025. Mistral small 3: Apache 2.0, 81% mmlu, 150 tokens/s. https://mistral.ai/news/mistral-small-3. Accessed: 2025-03-11.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. 2023. Scaling data-constrained language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 50358–50376. Curran Associates, Inc.

National Library of Medicine. 2003. Pmc open access subset. https://pmc.ncbi.nlm.nih.gov/tools/openftlist/. [cited 2025 03 14].

Mitodru Niyogi and Arnab Bhattacharya. 2024. Paramanu-ayn: Pretrain from scratch or continual pretraining of llms for legal domain adaptation? *Preprint*, arXiv:2403.13681.

Amin Heyrani Nobari, Kaveh Alimohammadi, Ali ArjomandBigdeli, Akash Srivastava, Faez Ahmed, and Navid Azizan. 2025. Activation-informed merging of large language models. *Preprint*, arXiv:2502.02421.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *Preprint*, arXiv:2311.16452.

OpenAI. 2025. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/. Accessed: 2025-10-07.

Gerhard Paass and Sven Giesselbach. 2023. *Foundation Models for Natural Language Processing: Pretrained Language Models Integrating Media*.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Martin Jaggi, Leandro von Werra, and Thomas Wolf. 2024. FineWeb2: A sparkling update with 1000s of languages.

Cheng Peng, Xi Yang, Aokun Chen, Kaleb E. Smith, Nima PourNejatian, Anthony B. Costa, Cheryl Martin, Mona G. Flores, Ying Zhang, Tanja Magoc, Gloria Lipori, Duane A. Mitchell, Naykky S. Ospina, Mustafa M. Ahmed, William R. Hogan, Elizabeth A. Shenkman, Yi Guo, Jiang Bian, and Yonghui Wu. 2023. A study of generative large language model for medical research and healthcare. *npj Digital Medicine*, 6(1).

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Justus J. Randolph. 2005. Free-Marginal Multirater Kappa (multirater K [free]): An Alternative to Fleiss' Fixed-Marginal Multirater Kappa. In *Joensuu Learning and Instruction Symposium*.

Ken Shoemake. 1985. Animating rotation with quaternion curves. In *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '85, page 245–254, New York, NY, USA. Association for Computing Machinery.

Karan Singhal, Tien Tu, Johannes Gottweis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*.

Shamane Siriwardhana, Mark McQuade, Thomas Gauthier, Lucas Atkins, Fernando Fernandes Neto, Luke Meyers, Anneketh Vij, Tyler Odenthal, Charles Goddard, Mary MacCarthy, and Jacob Solawetz. 2024. Domain adaptation of llama3-70b-instruct through continual pre-training and model merging: A comprehensive evaluation. *Preprint*, arXiv:2406.14971.

Augustin Toma, Patrick R. Lawler, Jimmy Ba, Rahul G. Krishnan, Barry B. Rubin, and Bo Wang. 2023. Clinical camel: An open expert-level medical language model with dialogue-based knowledge encoding. *Preprint*, arXiv:2305.12031.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct Distillation of LM Alignment. *Preprint*, arXiv:2310.16944.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Preprint*, arXiv:2206.07682.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Prateek Yadav, Tu Vu, Jonathan Lai, Alexandra Chronopoulou, Manaal Faruqui, Mohit Bansal, and Tsendsuren Munkhdalai. 2024. What matters for model merging at scale? *Preprint*, arXiv:2410.03617.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2024a. Model merging in LLMs, MLLMs, and Beyond: Methods, Theories, Applications and Opportunities. *Preprint*, arXiv:2408.07666.

Guoxing Yang, Xiaohong Liu, Jianyu Shi, Zan Wang, and Guangyu Wang. 2024b. Tcm-gpt: Efficient pretraining of large language models for domain adaptation in traditional chinese medicine. *Computer Methods and Programs in Biomedicine Update*, 6:100158.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *Preprint*, arXiv:1905.07830.

Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. 2024. A comprehensive survey of scientific large language models and their applications in scientific discovery. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8783–8817, Miami, Florida, USA. Association for Computational Linguistics.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. 2023. Pytorch fsdp: Experiences on scaling fully sharded data parallel. *Preprint*, arXiv:2304.11277.

Guorui Zheng, Xidong Wang, Juhao Liang, Nuo Chen, Yuping Zheng, and Benyou Wang. 2025. Efficiently democratizing medical llms for 50 languages via a mixture of language family experts. In *Proceedings of the International Conference on Learning Representations (ICLR)*. Poster.

Furkan Öncel, Matthias Bethge, Betul Ermis, Mirco Ravanelli, Cem Subakan, and Çağlar Yıldız. 2024. Adaptation odyssey in llms: Why does additional pretraining sometimes fail to improve? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 19834–19843.

A Appendix

A.1 Medical Document Pre-Classifier

To classify documents into medical and non-medical categories, we apply the *Mixtral-8x7B-Instruct-v0.1* model with a zero-shot prompting approach. The exact prompt used for classification is as follows:

> You are a medical expert tasked with identifying whether the provided content is both medical and high quality. Analyze the content carefully based on the following criteria:
> - The content is related to health, medicine, or healthcare.
> - The information needs to be scientific, high-quality, accurate, and well-structured.
> Only allow the highest quality data. If it doesn't seem scientific, then leave it out. Also, leave out news stories or similar styles of text that are not written by medical experts. Restrict your response to 'yes' or 'no'. Do not include any other explanation.
> —
> Content: <DOCUMENT>

This prompt is used to filter a subset of approximately $260k$ documents from the German *FineWeb2* dataset. The model is constrained to generate only a single token, a document is classified as medical if the output token is "yes"; otherwise, it is categorized as non-medical.

A.1.1 Performance Evaluation of Medical Document Pre-Classifier

To evaluate the performance of the LLM-based medical document classifier, we randomly sampled 100 documents from the dataset, consisting of 50 documents classified as medical and 50 classified as non-medical by the LLM. These were annotated by three human annotators, who independently assessed the class of each document. We achieved an inter-annotator agreement of 84.7%, measured in *Fleiss' Kappa* (Randolph, 2005).

The evaluation results, presented in Table 5, demonstrate a strong overall performance of the LLM-based classifier, with an F1-score of 91.1 ± 2.5. The overall performance suggests that the LLM effectively distinguishes between medical

Metric	Score
Precision	92.0 ± 2.9
Recall	90.3 ± 2.6
F1-score	91.1 ± 2.5
Accuracy	91.0 ± 2.4

Table 5: Evaluation scores of the LLM's performance against three human annotators on a random subset of 100 samples.

and non-medical documents, although minor discrepancies with human annotators were observed. To closer evaluate the disagreement between the LLM and human annotators, we inspect the false positives in table 6. We find that, while some false positives are clearly non-medical, others suggest a degree of relevance to the medical domain.

Given these findings, we consider the LLM's performance sufficiently robust to proceed with training a more specialized classifier using the resulting labeled *FineMed-de-260k* dataset.

A.2 Training of Medical Classifier

To efficiently classify documents as medical or non-medical, we fine-tune a *XLM-RoBERTa-base* model using the labeled *FineMed-de-260k* dataset. This fine-tuning process aims to optimize the model for high precision in distinguishing medical content. The model was trained for $5k$ steps, approximately two epochs, with a batch size of 96, utilizing the *AdamW* (Loshchilov and Hutter, 2019) optimizer and a learning rate of 2×10^{-5}, with a weight decay of 0.01. Given the emphasis on precision over recall, we used the F_β score with $\beta = 0.7$ as the early stopping criterion.

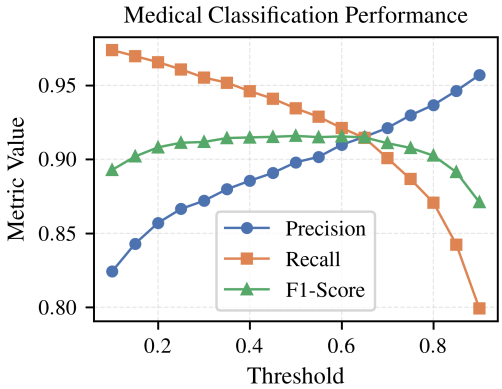

Figure 4: Performance metrics of the medical document classifier for various decision thresholds.

To ensure high precision during filtering, we employ a decision threshold of 0.9. This threshold is strategically chosen to minimize the inclusion of non-medical texts, prioritizing the precision of the resulting medical document selection. As depicted in Figure 4, this approach yields a precision of 0.95 and a recall of 0.8 at the chosen threshold of 0.9, demonstrating an effective balance between these metrics. This application process allows for the reliable extraction of medical documents, forming the foundation of our specialized dataset.

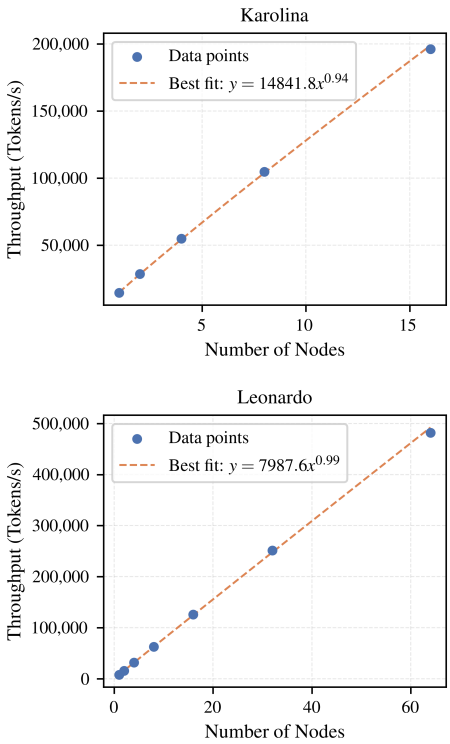

Figure 5: Weak scaling behavior on Karolina and Leonardo. The actual computation per accelerator is kept constant throughout, with a micro batch size of 1.

A.3 Continual Pre-Training Details

The following outlines specific technical details involved in the continual pre-training process for our models. We describe the data handling strategies employed for efficient batching, the optimization setup and training schedule used to ensure stable and effective learning, and the hardware resources and distributed training techniques leveraged to scale our experiments.

For batching efficiency, we adopt a bin-packing strategy, grouping sequences into constant-length batches to reduce padding and truncation. This approach not only avoids the computational overhead of padding tokens but also results in better model performance and higher throughput (Ding et al.,

Count	False Positive Example
3	Siehe auch: Mango Chicken Curry – Eine köstliche und gesunde Mahlzeit. Inhalt: 1. Erdmandeln sind reich an Ballaststoffen, 2. Erdmandeln sind reich an Vitaminen und Mineralstoffen, 3. Erdmandeln sind glutenfrei, 4. Erdmandeln sind ...
2	Es gibt sie wie Sand am Meer – die Persönlichkeitstest. Wir fragen uns, inwiefern ein solcher Test wirklich zu einem persönlichen Turnaround führen kann. Marlies Zindel führt eine eigene Beratungspraxis und nutzt den sogenannten Birkman-Test in ihren Gespr...
2	Dr. Preis ist unter den Top5:· Orthopäden in Köln (Stand 06/2017) Dr. med. Stefan Preis Arzt, Orthopäde Weiterbildungen: Chirotherapie (Manuelle Medizin), Sportmedizin Orthopädie/Sporttraumatologie Klinik am Ring Hohenstaufenring 28 50674 Köln ...
1	Video: Was versteht man unter dem Begriff Osmolarität? Osmotische Konzentration, früher bekannt als Osmolarität, ist das Maß für die Konzentration des gelösten Stoffes, definiert als die Anzahl der Osmole (Osm) des gelösten Stoffes pro Liter (L) der Lösung...
1	Viele Menschen haben den Verdacht, dass Tilapia "schlecht" oder "schmutzig" sei, was viele zu der Frage führt, ob der Fisch schlecht für Sie ist. Das liegt daran, dass Tilapia für seine Kontamination bekannt ist. In der Vergangenheit ernährten sich einige...
1	Forschungsförderung für UMG-Ernährungsprojekt Risiken früher erkennen, Komplikationen verhindern Forschende der Unimedizin Greifswald wollen die Überlebenschancen von Patienten mit entzündeter Bauchspeicheldrüse steigern. Dazu möchten sie den Einfluss des...
1	Anders als beim Sehen, können wir nur schwer störende Geräusche ausblenden. Deswegen wirkt sich unerwünschter Lärm auch auf unseren Schlafqualität aus, erklärt Prof. Dr. med. Dipl.-Psych. Manfred Beutel. Nicht nur physiologische, sondern auch seelische...

Table 6: Excerpt of the false-positives from the manually annotated 100 samples, showing the text excerpts along with the number of annotators who deemed each sample non-medical.

Parameter	PT
Learning Rate	2×10^{-5}
Weight Decay	0.01
Warmup Steps	500
Effective Batch Size	1024
Sequence Length	8192

Table 7: Hyperparameters used for continual pre-training (PT).

2024). We separate different sequences within a sample using the tokenizer's <EOS> token, allowing them to attend to each other within the same batch. This method contributed significantly to the overall efficiency of the pre-training process.

We train the models using the *AdamW* optimizer (Loshchilov and Hutter, 2019) with a linear learning rate decay. In order to address training instabilities, we implement an extended warmup phase of 500 steps, representing about half of the medical training corpus. To compensate for the longer warmup phase, we train all models for two epochs. Research by Tunstall et al. (2023) suggests that training beyond 1.5 epochs yields minimal additional benefits, which supports our training regime of 2 epochs with 0.5 epochs of warmup. Additionally, Muennighoff et al. (2023) indicate that four epochs are typically necessary before overfitting becomes a concern, further justifying our decision.

The experiments were conducted across multiple compute clusters. For most models, we utilized 8 nodes on the Karolina cluster, each equipped with 8 NVIDIA A100 (40GB) GPUs. However, the Mistral-Small-24B model was trained on the Leonardo cluster [3], leveraging 128 nodes. Table 8 provides a summary of the hardware and compute resources used during the continual pre-training phase for each model. Although all models were trained for exactly two epochs, the number of steps per model varied. This discrepancy is due to differences in the model-specific tokenizers and their fertility on the training corpus. Detailed hyperparameters are listed in Table 7.

For multi-node settings, we employed the Hybrid Sharding Data Parallel (HSDP) variant of FSDP. This approach involved distributing the data across multiple compute nodes and sharding the

[3] https://leonardo-supercomputer.cineca.eu/

Model	Cluster	Accelerator	Num GPUs	Time	Steps	Tokens
Mistral-7B-Instruct-v0.3	Karolina	A100 (40GB)	64	72h	2970	25.1B
Qwen2.5-7B-Instruct	Karolina	A100 (40GB)	64	64h	2634	22.1B
Mistral-Small-24B-Instruct	Leonardo	A100 (64GB)	128	48h	2294	19.2B

Table 8: Computational resources invested in continual pre-training of different models.

model parameters, gradients, and optimizer states within each node. By avoiding node-level sharding, we achieved near-linear scaling, as illustrated in Figure 5. In both cases, we reached a throughput of approximately 1850 tokens per second per GPU for $7B$ parameter sized models. This is consistent with findings of prior works, which report throughputs of 2350 (DiscoResearch et al., 2024) and 1950 (Labrak et al., 2024) tokens per second per GPU, respectively. Note that, due to its larger size, the *Mistral-Small-24B* model required full sharding across nodes, leading to a corresponding increase in communication overhead.

The NVIDIA A100 accelerators available to us were equipped with only 40GB (Karolina) and 64GB (Leonardo) of VRAM. As a result, we had to reduce our micro batch size to 1, which lowered the arithmetic intensity compared to what would be possible on the more typical 80GB A100 GPUs.

A.4 Impact of Domain-Specific Filtering

To rigorously assess the impact of utilizing a highly domain-specific dataset, we conduct an ablation study designed to isolate and quantify the benefit of our medical corpus compared to a baseline of unfiltered data. To this end, we create a subset of the *FineWeb2* dataset, matching the size of our filtered medical corpus at $10.0B$ tokens. Applying the medical document classifier from section 3.2, reveals that only 0.89% of the sampled documents are related to the medical domain. We continually pre-trained the *Qwen2.5-7B-Instruct* model, using the exact same hyperparameters as discussed in section 4.2, for two epochs, resulting in a total of $22.1B$ training tokens and evaluate under the same conditions as the other models.

The performance results, presented in Table 9, demonstrate that the model trained on the random sample shows improvements over its base model. Interestingly, we find that the model trained on the random subset outperforms its domain-specific counterparts in three out of eight settings. This especially includes the *MedQA* benchmark. These exceptions need further investigation. On average however, it is clearly outperformed domain-specific model variants. The improvements observed in *FineWeb2-Qwen2.5-7B-Instruct* over the base model can be attributed to the inherent high quality of the *FineWeb2* dataset and the general benefit of continual pre-training on German data. However, the substantially larger gains achieved with the model trained with the filtered medical corpus demonstrate the significant impact of a domain-specific corpus.

A.5 German Evaluation in Non-medical tasks

To investigate potential trade-offs in instruction-following capabilities introduced by our domain-specific adaptation, we systematically evaluated the *DeFineMed* models on non-medical benchmarks. In particular, we compared the performance of the original pre-trained baseline models with the versions obtained after continuous pre-training on medical data and subsequent model merging. For this evaluation, we selected several widely used benchmarks outside the medical domain: MMMLU (Hendrycks et al., 2021b,a), which measures broad multi-task generalization across knowledge areas; *ARC-Easy* and *ARC-Challenge* (Clark et al., 2018) for scientific and commonsense question answering; *GSM8k* (Cobbe et al., 2021) to assess mathematical reasoning; and *HellaSwag* (Zellers et al., 2019) for evaluating contextual understanding and narrative completion. This setup allows us to quantify how well the adapted models retain general-purpose reasoning and instruction-following skills beyond their specialized medical context. As shown in Table 10, the medical models consistently outperform their base versions, suggesting that continual pretraining on our dataset not only enhances domain alignment on medical benchmarks but also improves general linguistic competence.

Model	Anatomy	Clinical Knowledge	College Medicine	MedQA	Average
Qwen2.5-7B-Instruct	54.07 ± 4.30	69.06 ± 2.85	63.01 ± 3.68	50.20 ± 2.24	59.08 ± 1.68
DeFineMed-Qwen2.5-7B	62.22 ± 4.19	76.60 ± 2.61	68.21 ± 3.55	52.60 ± 2.24	64.91 ± 1.62
FineWeb2-Qwen2.5-7B	56.30 ± 4.28	75.47 ± 2.65	61.85 ± 3.70	53.20 ± 2.23	61.70 ± 1.66
DeFineMed-Qwen2.5-7B-SLERP	65.19 ± 4.12	75.85 ± 2.63	65.90 ± 3.61	54.20 ± 2.23	65.28 ± 1.62
FineWeb2-Qwen2.5-7B-SLERP	55.56 ± 4.29	73.58 ± 2.71	68.21 ± 3.55	54.40 ± 2.23	62.94 ± 1.65

Table 9: Performance comparison of Qwen2.5-7B-Instruct models trained on domain-specific *FineMed-de* and randomly sampled *FineWeb2-de* data. Best score is underlined.

Model	MMLU	ARC-C	ARC-E	GSM8k	HellaSwag	Average
BioMistral-7B-SLERP (baseline)	46.87 ± 0.41	40.96 ± 1.44	63.43 ± 0.99	2.35 ± 0.42	54.00 ± 0.50	41.52 ± 0.38
Mistral-7B-Instruct	50.33 ± 0.40	48.29 ± 1.46	68.81 ± 0.95	11.22 ± 0.87	59.82 ± 0.49	47.70 ± 0.41
Mistral-7B-Instruct-Medical	51.95 ± 0.40	49.23 ± 1.46	71.59 ± 0.93	11.60 ± 0.88	66.59 ± 0.47	50.19 ± 0.41
Mistral-7B-Instruct-Medical-SLERP	53.39 ± 0.40	49.83 ± 1.46	71.59 ± 0.93	13.27 ± 0.93	66.08 ± 0.47	50.83 ± 0.41
Qwen2.5-7B-Instruct	63.35 ± 0.39	51.28 ± 1.46	68.81 ± 0.95	58.23 ± 1.36	61.05 ± 0.49	60.54 ± 0.46
Qwen2.5-7B-Instruct-Medical	64.44 ± 0.38	51.28 ± 1.46	70.71 ± 0.93	63.38 ± 1.33	63.22 ± 0.48	62.61 ± 0.45
Qwen2.5-7B-Instruct-Medical-SLERP	65.57 ± 0.38	52.22 ± 1.46	70.12 ± 0.94	66.94 ± 1.30	62.22 ± 0.49	63.42 ± 0.45
Mistral-Small-24B-Instruct	71.95 ± 0.35	64.08 ± 1.40	80.98 ± 0.81	64.97 ± 1.31	71.96 ± 0.45	70.79 ± 0.43
Mistral-Small-24B-Instruct-Medical	71.26 ± 0.36	62.97 ± 1.41	80.89 ± 0.81	53.75 ± 1.37	73.15 ± 0.44	68.41 ± 0.44
Mistral-Small-24B-Instruct-Medical-SLERP	73.31 ± 0.35	65.27 ± 1.39	81.40 ± 0.80	56.86 ± 1.36	73.69 ± 0.44	70.11 ± 0.44

Table 10: Model performance on different German non-medical benchmarks in one-shot setting. The average accuracy is the unweighted mean of accuracies, whereas the average standard error is the square root of the unweighted mean of variances. Best score is underlined.

Failure Mode	Description
Language Mixing	Whether the response unexpectedly switches between languages or dialects without instruction or user prompting.
Typo	Whether the response contains spelling mistakes, character errors, or malformed words that reduce readability.
Contradiction	Whether the response contains internal inconsistencies or conflicts with its own earlier statements.
Hallucination	Whether the response fabricates facts, citations, or details that are presented as truth but are not grounded in reliable sources.
Omission	Whether key details, steps, or context are missing, leading to an incomplete or misleading answer.
Overgeneralization	Whether the response makes broad, unsupported claims or ignores edge cases, resulting in oversimplification.
Verbosity	Whether the response is unnecessarily long-winded, repeating points without adding value.
Underexplained	Whether the response is too brief or lacks the necessary depth to be useful or correct.
Ambiguity	Whether the response is vague or unclear in wording, leading to multiple possible interpretations.
Repetition	Whether the response repeats phrases or sentences excessively, often due to generation loops.
Overconfidence	Whether the response states information in an authoritative tone despite being incorrect or uncertain.

Table 11: Descriptions of Failure Modes