

A Modular Approach for Clinical SLMs Driven by Synthetic Data with Pre-Instruction Tuning, Model Merging, and Clinical-Tasks Alignment

Jean-Philippe Corbeil^{1*}, Amin Dada⁴, Jean-Michel Attendu¹, Asma Ben Abacha¹
Alessandro Sordoni^{2,3}, Lucas Caccia², François Beaulieu¹, Thomas Lin¹
Jens Kleesiek^{4†}, Paul Vozila¹

¹Microsoft Healthcare & Life Sciences ²Microsoft Research Montréal, Canada
³Mila, Université de Montréal, Canada ⁴IKIM, University Hospital Essen, Germany

Abstract

High computation costs and latency of large language models such as GPT-4 have limited their deployment in clinical settings. Small language models (SLMs) offer a cost-effective alternative, but their limited capacity requires biomedical domain adaptation, which remains challenging. An additional bottleneck is the unavailability and high sensitivity of clinical data. To address these challenges, we propose a novel framework for adapting SLMs into high-performing clinical models. We introduce the **MediPhi** collection of 3.8B-parameter SLMs developed with our novel framework: pre-instruction tuning of experts on relevant medical and clinical corpora (PMC, Medical Guideline, MedWiki, etc.), model merging, and clinical-tasks alignment. To cover most clinical tasks, we extended the CLUE benchmark to CLUE+, doubling its size. Our expert models deliver relative improvements on this benchmark over the base model without any task-specific fine-tuning: 64.3% on medical entities, 49.5% on radiology reports, and 44% on ICD-10 coding (outperforming GPT-4-0125 by 14%). We unify the expert models into MediPhi via model merging, preserving gains across benchmarks. Furthermore, we built the **MediFlow** collection, a synthetic dataset of 2.5 million high-quality instructions on 14 medical NLP tasks, 98 fine-grained document types, and JSON format support. Alignment of MediPhi using supervised fine-tuning and direct preference optimization achieves further gains of 18.9% on average.

1 Introduction

Advances in natural language processing (NLP) have enabled large language models (LLMs) like

*Corresponding author: jcorbeil@microsoft.com

†Other affiliations: Cancer Research Center Cologne Essen (CCCE), German Cancer Consortium (DKTK, Partner site Essen) and Department of Physics of TU Dortmund (Dortmund, Germany).

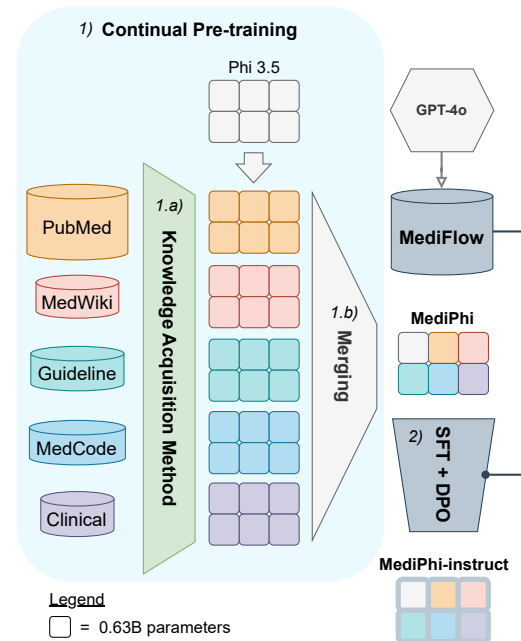


Figure 1: Our approach in two steps: 1) continual pre-training; 2) alignment. 1) Starting from *Phi3.5-mini-instruct*: 1.a) We leverage *knowledge acquisition methods* such as pre-instruction tuning on diverse medical and clinical corpora to obtain domain-specific experts. 1.b) We recur to *model merging* to merge experts first with the base model to combine new knowledge as well as recover skills degraded by the previous step, then together to form a unified model, MediPhi. 2) We generate MediFlow, a synthetic instruction dataset MediFlow for clinical tasks. We align our model on MediFlow using Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO) and obtain MediPhi-Instruct. *Segmentation of model parameters as equal block size are only for illustrative purposes.*

GPT-4 to excel in medical tasks, especially medical exams (Nori et al., 2023; Ben Abacha et al., 2024; Nori et al., 2024). However, their deployment in clinical settings¹ faces many challenges,

¹We define the medical field as encompassing medical knowledge (e.g. anatomy, genetics, biology), while the clinical

including high latency and cost (Yang et al., 2023; Dennstädt et al., 2025). As LLM progress faces diminishing returns from scaling (Udandarao et al., 2024; Longpre et al., 2024; Muennighoff et al., 2023; Villalobos et al., 2022), specialized small language models (SLMs) provide a viable alternative (Sardana et al., 2024), when optimized for domain-specific performance, lower computational requirements, and real-world clinical integration.

Developing high-quality clinical language models is hindered by the unavailability of clinical data which are sensitive and tightly licensed, e.g. protected health information under HIPAA. Current medical LLMs perform well on multiple-choice question datasets but struggle with real-world clinical complexities (Dada et al., 2024; Chen et al., 2024; Liu et al., 2024; Jeong et al., 2024a,b). Furthermore, both the inaccessibility of clinical data and the misalignment of current continual pre-training methods for clinical tasks are critical limitations in the context of SLMs, which have limited capacity. Addressing these gaps require innovative training strategies tailored to small models.

This work introduces a modular framework for building high-performance medical SLMs, leveraging pre-instruction tuning (Jiang et al., 2024), model merging, and clinical alignment. Using pre-instruction tuning, we adapt *Phi3.5 mini* with 3.8B parameters (Abdin et al., 2024a) into experts trained on diverse medical and clinical corpora. We unify these models through model merging into one SLM which preserves benchmark improvements. We complete training by aligning the model with MediFlow, a new synthetic instruction dataset on clinical tasks. A representation of this approach is given in Figure 1.

Our contributions include:

- Introducing the **MediPhi family**, the first collection of high-performance SLMs for medical and clinical applications under a commercially permissive license². The collection includes both a generalist expert model and specialized variants.
- Releasing of the **MediFlow** dataset of 2.5 million high-quality synthetic instructions generated with a GPT-4o based agentic pipeline,

field is related to direct patient care and healthcare practice (e.g. doctor-patient dialog, discharge summary, and clinical note).

²<https://huggingface.co/collections/microsoft/mediphi-6835cb830830dfc6784d0a50>

also under a commercially permissive license³. This dataset for the clinical domain contains 14 tasks categories, 98 fine-grained input documents, 6 difficulty levels, and 2 output formats (JSON or plain text), filling a gap in the clinical NLP resources. Additionally, we release synthetic validation sets designed to guide the model merging algorithm.

- An extension of CLUE to the **CLUE+ benchmark** by doubling its size from 6 to 12 datasets including complementary clinical tasks and input documents, allowing a comprehensive evaluation of medical and clinical capabilities of language models (e.g. radiology reports, medications, medical error detection, doctor-patient dialog summarization, information extraction of social determinants of health, and medical coding).
- A demonstration of the **effectiveness of pre-instruction tuning** for medical domain adaptation extending the method beyond question-answering with named entity recognition, relation extraction, and summarization.
- A **study on ICD10CM medical coding**, in terms of domain adaptation and benchmarking of medical models, with relative improvements up to 44% over the base model, surpassing GPT-4-0125 by 14%.

2 Previous Work

2.1 Medical Large Language Models

Researchers have developed various open-weight LLMs with diverse capabilities and research licenses in the medical NLP domain. Examples include ClinicalCamel (Toma et al., 2023), Med42 (Christophe et al., 2023), PMC-Llama (Wu et al., 2024), BioMedGPT (Zhang et al., 2024), Meditron (Chen et al., 2023), BioMistral (Labrak et al., 2024) and Asclepius (Kweon et al., 2024). Most recent medical LLMs (Chen et al., 2023; Christophe et al., 2024a; Gururajan et al., 2024; Ankit Pal, 2024) are based on Llama3 (Dubey et al., 2024) — 8B and 70B parameters. Google also trained their own medical LLM named Med-PaLM 2 (Singhal et al., 2023) with 340B parameters, which is not publicly available.

³<https://huggingface.co/datasets/microsoft/mediflow>

2.2 Medical Instruction Tuning

Recently, authors have released instruction-tuned models for the medical domain: Aloe (Gururajan et al., 2024), Hippocrates (Acikgoz et al., 2024), and Med42 v2 (Christophe et al., 2024a,b). The alignment phase of these three models includes similar instruction datasets such as medical question-answering databases, non-medical alignment data (e.g. UltraChat by Ding et al.), and benchmark training sets — e.g. MedQA (Jin et al., 2021), PubMedQA (Jin et al., 2019) and ACI-Bench (Yim et al., 2023). While this mix of data has contributed to improvements, it also introduces imbalances in task and document coverage, as well as in-distribution evaluations in some cases, i.e. closer to evaluation in fine-tuning setting instead of zero-shot/few-shot setting.

2.3 Synthetic Instructions

Phi 1 & 2 (Gunasekar et al., 2023; Li et al., 2023), with 1.3B and 2.7B parameters, respectively, showed strong reasoning performance using synthetic textbook-like data. Phi-3 and Phi-3.5 (Abdin et al., 2024a) scaled model size to 3.8B parameters and topic coverage. Phi4 (Abdin et al., 2024b) at 14B parameters is trained using an iterative data generation process. Similarly, Orca (Mukherjee et al., 2023; Mitra et al., 2024) focused on reasoning data, while UltraChat (Ding et al., 2023) targeted multi-turn prompts. Zhang et al. (2023b) generated 52,000 medical question-answering instructions based on forms filled by experts. In the clinical domain, Kweon et al. (2024) generated 158k short question-answering instructions for 8 tasks with synthetic clinical documents as inputs, seeded from the PMC-Patients corpus (Zhao et al., 2023). MagPie (Xu et al., 2024) proposed the generation of synthetic instructions in general domains using LLMs’ chat template.

2.4 Model Merging

Model Soup (Wortsman et al., 2022) demonstrated improvements over single checkpoint evaluation by averaging training checkpoints, leading to methods like spherical linear interpolation (SLERP), enabled by linear-mode connectivity (Frankle et al., 2020; Mirzadeh et al.). To extend to multiple models, authors have proposed *task arithmetic* (Ilharco et al.) while techniques such as TIES (Yadav et al., 2024a), DARE (Yu et al., 2024), and BreadCrumbs (Davari and Belilovsky, 2024) address the param-

eter interference issue. Hammoud et al. (2024) highlighted the importance of validation sets for an optimal merge. Large-scale experiments (Yadav et al., 2024b) showed the effectiveness of merging multiple experts, especially from large instruction models. Merging has also proven as effective as data-mix strategies during pre-training (Ahmadian et al., 2024; Na et al., 2024).

2.5 Clinical Benchmarking

Medical LLMs are commonly benchmarked on medical knowledge through multiple-choice datasets such as MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), and MMLU-medical (Hendrycks et al., 2020). However, recent studies (Liu et al., 2024; Dada et al., 2024; Chen et al., 2024; Jeong et al., 2024a,b) revealed a gap on the clinical domain from medical LLMs.

3 Methodology

Our method for clinical SLMs as illustrated in Figure 1 includes two steps: 1) continual pre-training composed of 1.a) domain knowledge acquisition methods and 1.b) model merging, and 2) post-training with supervised fine-tuning and direct preference optimization on generated synthetic data.

3.1 Continual Pre-Training

3.1.1 Datasets

In Table 1, we list the medical and clinical corpora with permissive licenses, used to adapt our SLM into five experts. We separate these into five groups: *PubMed*, *Clinical*, *MedCode*, *Guidelines*, and *MedWiki*. We describe these dataset groups with their licensing status in detail in Appendix A.1.

Table 1: Medical and clinical data sources separated into five groups: *PubMed*, *Clinical*, *MedCode*, *Guidelines*, and *MedWiki*.

Groups	Source	Document Type	#Docs	#Tokens
PubMed	PMC	Scientific Articles	3.8M	42B
	PMC Abstract	Scientific Abstracts	36M	6B
Clinical	PMC-Patients	Patient summaries	156k	130M
	Asclepius	Clinical Documents	80k	44M
	NoteChat	Conversations	80k	72M
	MTSamples	Clinical Documents	5k	4M
MedCode	ICD9/10&ATC	Webpages	206k	257M
Guidelines	Guidelines	Websites	37k	90M
MedWiki	MedWiki	Encyclopedia	80k	80M

3.1.2 Domain-Specific Pre-Training

We consider three methods to enhance domain knowledge of language models: domain adaptation pre-training (DAPT), textbook-like synthetic material (Explainer), and pre-instruction tuning (PIT). The latter showed considerable improvements in our experiments. We apply these techniques on the base model to obtain five experts in the medical and clinical domain from our five dataset groups. We describe continual pre-training optimization strategies in Appendix A.2.

DAPT (Ganin and Lempitsky, 2015; Gururangan et al., 2020) Domain Adaptation Pre-Training is a technique for adapting a deep learning model to a specific domain by performing next-token prediction on domain-specific corpora. The expert that is trained on the *PubMed* group follows standard DAPT, because it is orders of magnitude larger than the others. We also train on all the data a *DataMix* baseline using DAPT, which is outperformed by our approach. For this method, we trained models for one epoch with a cosine scheduler at a maximum learning rate of $5e-5$.

Explainer (Gunasekar et al., 2023) Textbook-like material demonstrated the effectiveness of training SLMs on textbook-quality data generated by a strong LLM. Therefore, we generate a textbook-like "explainer" using GPT-4o-0806 for the *MedCode* dataset group, for which the dense format of the webpages impedes model learning. For this method, we trained models for two epochs with a cosine scheduler at a maximum learning rate of $1e-4$.

PIT (Jiang et al., 2024) Pre-Instruction Tuning⁴ (PIT) demonstrated significant improvements over the conventional training paradigm, by first fine-tuning on instruction-like data, followed by training on a concatenation of the instruction data and pre-training corpus.

PIT requires the generation of task data for each document in the corpus. While Jiang et al. (2024) originally used question-answering (QA) as the primary task, we expand the approach to include summarization, named entity recognition, and relation extraction. We use GPT-4o-0806 to generate outputs for all four tasks on the ICD10CM subset of the *MedCode* dataset as an initial case study. Based on the results, we extend this process

⁴Despite its name, PIT does not involve explicit instruction data or chat-style formatting.

to the remaining four dataset groups — Clinical, *MedCode*, Guidelines, and *MedWiki* — to train multiple expert models.

Each expert undergoes sequential training in two phases. The first phase involves fine-tuning on the generated outputs from a single task for one epoch using a cosine scheduler (peak learning rate at $1e-4$). If the task contains multiple elements, such as several question-answer pairs, they are concatenated into a single sequence, separated by end-of-sentence (EOS) tokens. In the second phase, the model is fine-tuned for two epochs with another cosine scheduler (peak learning rate at $3e-4$) on the concatenation of the task data and the original documents, with EOS tokens acting as separators.

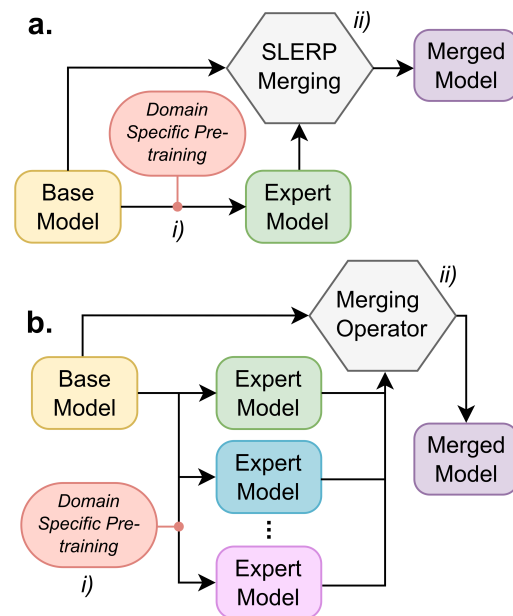


Figure 2: Overview of model merging techniques. **a.** *Domain-specific merging via SLERP*: An expert model is obtained by fine-tuning the base model on domain-specific data (step i). The expert and base models are then merged using spherical linear interpolation (SLERP) to produce the final merged model (step ii). **b.** *Multi-expert merging*: Multiple expert models are independently derived from the base model via domain-specific pretraining (step i). These are then combined using a merging operator (e.g., Task Arithmetic, Ties, or BreadCrumbs) to produce a unified merged model (step ii).

3.1.3 Domain-Specific Model Merging

We train five expert models specializing in different aspects of medical and clinical knowledge, each derived from a distinct dataset group: *PubMed*, *Clinical*, *MedCode*, *Guidelines*, and *MedWiki*.

Following the success of BioMistral (Labrak et al., 2024), the first approach involves merging each expert model individually with the base model using SLERP⁵ as in step *a* of Figure 2. We determined the merging proportions (10%, 25% or 50%) via validation sets (see below). We apply merging after PIT since these techniques demonstrate a synergistic effect. While PIT enhances domain-specific learning, it also leads to catastrophic forgetting — degrading the model’s initial abilities such as instruction following, long context handling, and multilingual support (Scialom et al., 2022). This issue is particularly evident in zero-shot and few-shot settings, where instruction-tuned models generally outperform their base counterparts (Longpre et al., 2023; Zhang et al., 2023a). Merging with the original instruction model after PIT mitigates these degradations, preserving general capabilities while maximizing domain adaptation⁶.

3.1.4 Multi-Expert Merging into MediPhi

Our second set of experiments combines all five experts into a unified SLM forming *MediPhi* as in step *b* of Figure 2. Multi-model merging involves three primary techniques: Task Arithmetic, TIES, and BreadCrumbs. Given the vast configuration space of multi-model merges, we employ an evolutionary algorithm via MergeKit (Goddard et al., 2024) to optimize the merging process.

However, optimization on our benchmark is not feasible due to a lack of validation data and framework incompatibilities. To address this, we generate synthetic validation sets aligned with our benchmark tasks. Specifically, we prompt GPT-4o-0806 to create multiple-choice question sets covering 12 medical and clinical topics relevant to our benchmark (e.g., doctor-patient interactions, medical coding, discharge summaries). These sets maintain contextual consistency with our evaluation tasks. The evolutionary algorithm is set to terminate after 500 evaluations, guided by average accuracy on these validation sets. Further details about validation set creation procedures are provided in Appendix A.3.

⁵A method optimized for merging two models, often yielding high-performing hybrids (Hammoud et al., 2024; Ahmadian et al., 2024; Labrak et al., 2024).

⁶Alternative approaches, such as KL-divergence regularization in RLHF (Ouyang et al., 2022), exist for maintaining generalization, but recent studies suggest model merging can further optimize reward alignment in RL settings (Ramé et al., 2024).

3.1.5 Evaluation Metrics

To identify the top-performing model, we primarily measure the *average accuracy* on CLUE+. However, we also consider important that the expert achieves uniform improvements, especially among experts with similar accuracies. For this purpose, we use $\#DG$, the number of datasets on which the model achieves gains, and $CV \Delta$, the *coefficient of variations of gains/losses* as defined in equation 1.

$$CV \Delta = \frac{\sqrt{\mathbb{E}_{d \sim \mathcal{D}} [(\delta_d - \mu_d)^2]}}{|\mu_d|} \quad (1)$$

where δ_d is the expert accuracy minus the baseline one for the d^{th} dataset in the benchmark \mathcal{D} and $\mu_d = \mathbb{E}_{d \sim \mathcal{D}}[\delta_d]$. A small $CV \Delta$ indicate uniform gains or losses across datasets, while a high value indicates gains or losses on a narrow subset of datasets.

3.2 Clinical Alignment

3.2.1 Data Generation Pipelines

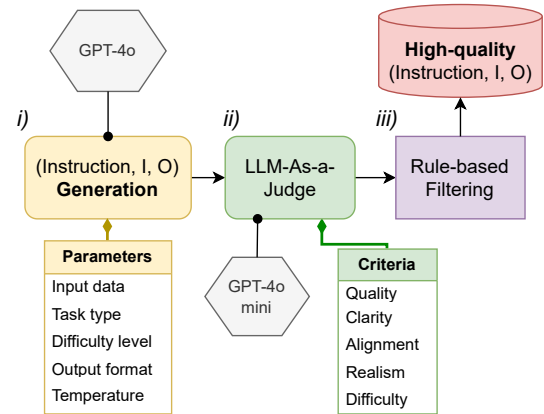


Figure 3: Schema of MediFlow generation. i) Given a randomly sampled set of parameter values, we prompt *GPT-4o-0806* for N instructions to obtain triplets (*instruction, inputs, outputs*) in which we have P pairs of inputs (I) and outputs (O) each. ii) We prompt *GPT-4o mini* (2024-07-18) with LLM-as-a-Judge and self-consistency on M samples. iii) We define a heuristic to filter a diverse, high-quality subset for alignment purposes.

Generation of MediFlow In Figure 3, we illustrate our agentic pipeline to generate the MediFlow dataset.

At step *i*), we prompt *GPT-4o-0806* with a meta-prompt to generate several triplets

(*instruction, inputs, outputs*) at once conditioned on five parameters: input-data type, task type, difficulty level, output format, and temperature. For this step, we also apply temperatures of 1.0 (70% of the dataset) or 1.25 (30% of the dataset), to strike a balance between accuracy or diversity, respectively. Specifically, we request 10 instructions at a time with four input-output pairs each.

In step *ii*), we prompt *GPT-4o mini* (2024-07-18) with a LLM-as-a-Judge approach — using self-consistency with chain-of-thought across $M = 5$ samples at temperature of 1.0 — to provide a critical assessment of the synthetic instruction. We provide five criteria in the prompt on a scale of 1 to 4: quality, clarity, alignment, realism, and difficulty. We compute the final score $S_j \in [1, 4]$ of the j^{th} criterion by summing individual scores $s_{ij} \in \{1, 2, 3, 4\}$ multiplied by their respective counts $c_i \in [0, M]$ (constrained by $\sum_{i=1}^M c_i = M$) as in $S_j = \frac{1}{M} \sum_{i=1}^M c_i \cdot s_{ij}$.

At the final step *iii*), we use a heuristic to trim the collection down to its top- K highest quality samples based on the quality criteria. We added details of the process in Appendix A.5.

Generation of MediFlow-DPO From MediFlow, we filter the 130k top-quality samples stratified by task type, input-data type and output format to further align the SLM with direct preference optimization (DPO) after SFT. We generate marginally wrong outputs (i.e. rejected outputs) prompting GPT-4o-0806 as an error inducer. We provide the detail in Appendix A.7.

3.2.2 Alignment Process SFT + DPO

We train *MediPhi*, the multi-expert merged SLM, with supervised fine-tuning (SFT) on MediFlow to obtain *MediPhi-SFT*. We then align *MediPhi-SFT* with DPO which leads to *MediPhi-Instruct* (SFT + DPO). We provide the hyperparameters for both settings in Tables 7 and 8 of Appendix A.2.

4 Experiments

4.1 CLUE+ Benchmark

The CLUE Benchmark (Dada et al., 2024) covers six datasets: MedNLI (Romanov and Shvade, 2018), MeQSum (Ben Abacha and Demner-Fushman, 2019), Problem List Summarization (Gao et al., 2023), LongHealth (Adams et al., 2024),

MeDiSumQA (Dada et al., 2025), and MeDiSumCode. For these datasets, we implement the same configuration (i.e. prompts, metrics, and few shots) as Dada et al. (2024). While CLUE focused on six tasks using clinical notes and discharge summaries as input data, we extend it with additional clinical input documents (e.g., radiology reports, doctor-patient dialog) and tasks (e.g. information extraction, medical error detection) to obtain a broader assessment of clinical abilities. We introduce the CLUE+ benchmark, with six additional datasets: MedicationQA (Ben Abacha et al., 2019), MEDIQA-RRS QA (Ben Abacha et al., 2021), MEDEC (Ben Abacha et al., 2024), ACI-Bench (Yim et al., 2023), Social Determinant of Health (Lybarger et al., 2023), and MedConceptsQA ICD10CM (Shoham and Rappoport, 2024). We provide further information in the Appendix A.8 in Tables 9 and 10.

4.2 Continual Pre-Training

To assess the efficiency of different continual pre-training methods, we focus on ICD10CM medical coding webpages from the *MedCode* dataset. After continual pre-training on this subset, we test the models on ICD10CM questions from the MedConceptsQA dataset (Shoham and Rappoport, 2024). We left out the easy and medium difficulty questions to focus on the most challenging questions.

4.2.1 DAPT vs. Explainer vs. PIT

In Figure 4, we plot the accuracy of different continual pre-training methods DAPT, Explainer and PIT, as described in Sec. 3.1.2. First, we note that DAPT diminishes the performance to random (*Webpage*). We hypothesize that the webpages have peculiar implicit format. One solution is to reformulate them as textbook-like explainers. The model fine-tuned on explainers is improving upon the baseline by 6% (*Explainers*). We see that using PIT by generating summaries (*Summary*) increases performance further by 8% percent.

4.2.2 Domain-Specific Merging

In Figure 5, we experiment with merging back into the base model after fine-tuning on the explainers or training with PIT.

We observe significant improvements across tasks for SLERP-merged models, with the Summary model performing best. Although question answering (*QA*), named entity recognition (*Entities*), and relation extraction (*Relations*)

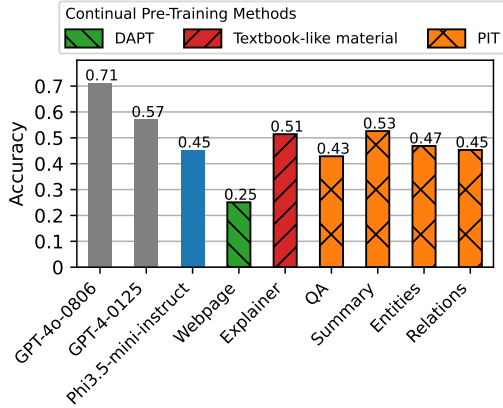


Figure 4: Performance of different continual pre-training methods: domain adaptation pre-training (DAPT) on ICD10CM webpages, fine-tuning on synthetic textbook-like material generated with GPT-4o-0806 (Explainer), and pre-instruction tuning (PIT) with different synthetically generated tasks: QA, summarization, NER, and relation extraction.

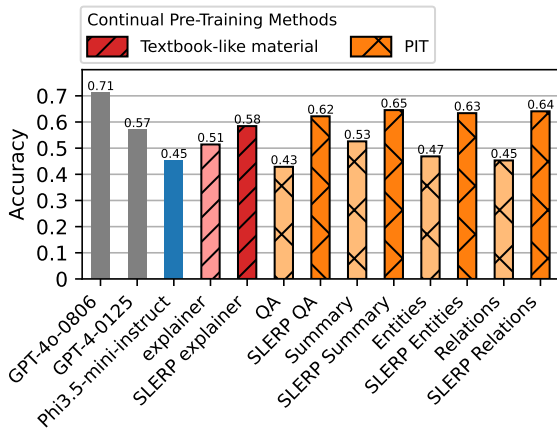


Figure 5: Impact of SLERP merging on the performance of ICD10CM. Merging back with the base model (SLERP 50%) systematically results in gains and these are more pronounced for PIT.

initially decline compared to the base model before merging, SLERP not only boosts the explainer model to 58% (a 28% relative improvement over the base model) but also enhances all PIT-trained models further—reaching 62% (38% relative) for QA and 65% (44% relative) for Summary, surpassing GPT-4-0125 by 8% (14% relative).

Based on these results, the rest of the paper will apply PIT with summaries, unless otherwise stated.

4.2.3 Multi-Domain Merging

We present the results on the CLUE+ benchmark of the five medical and clinical experts, adapted on the dataset groups and merged back with the base

model with SLERP in Table 2. While the *DataMix* merged model trained on all the corpora (i.e. similar to BioMistral) achieves gains over the base model, the individual experts realize further gains, except for *MedCode*. The latter improves over the base models but mostly only on coding datasets (i.e. MeDiSumCode and MedConceptsQA, see Tables 11 and 12 in Appendix A.8 for detailed benchmark). The highest enhancement comes from *MedWiki* on 10 out of 12 datasets, with an average improvement of 3.2% (8.8% relative). One possible explanation is that the model is learning better on educational contents — such as encyclopedic material — with vast coverage of medical concepts. We observe that removing PIT and/or merging on the *Guideline* group lead to the worst outcomes below the baselines.

Table 2: Average performances on CLUE+ of SLERP-merged experts obtained with PIT on the five dataset groups. We provide *DataMix* as a baseline SLERP expert trained on all dataset groups. We also provide an ablation study on *Guideline*. We indicate gains and losses over the base model. #DG stands for datasets with gains (out of 12 datasets). CV Δ stands for coefficient of variations of gains/losses.

		AVG \uparrow	#DG \uparrow	CV $\Delta\downarrow$
Baselines	<i>Phi-3.5 mini</i>	36.5	-	-
	<i>DataMix</i>	37.5	10	1.2
SLERP Experts	<i>PubMed</i>	37.7	9	1.8
	<i>Clinical</i>	39.6	10	2.0
	<i>MedWiki</i>	39.7	10	1.5
	<i>MedCode</i>	36.7	5	39.6
	<i>Guideline</i>	39.2	10	1.8
Guideline Ablations	w/o SLERP	27.2	4	1.0
	w/o PIT	33.0	6	0.9
	w/o SLERP&PIT	25.2	1	2.8

We present the results on CLUE+ of combining all experts using three multi-model merging techniques (Task-Arithmetic, Ties-merging, and BreadCrumbs) in Table 3. While Task-Arithmetic yields the highest score of 39.4 with gains on 9 out of 12 benchmark datasets, the models from Ties and BreadCrumbs remain competitive, demonstrating more uniform gains across the benchmark, as indicated by their lower CV Δ . Since the goal is to select the most robust model for the align-

Table 3: Average performance on CLUE+ of unifying experts into one SLM. We leverage Task-Arithmetic (TA), TIES-merging and Breadcrumbs (BC) for merging 5 experts: *PubMed*, *Clinical*, *MedWiki*, *MedCode* and *Guideline*. We provide statistics across dataset performances of the SLERP experts, translating into worst case (minimum), average, and best case (maximum). We indicate **gains** and **losses** over the baseline. #DG stands for datasets with gains (out of 12 datasets). CV Δ stands for coefficient of variations of gains/losses.

		AVG \uparrow	#DG \uparrow	CV Δ \downarrow
Baseline	<i>Phi-3.5 mini</i>	36.5	-	-
	Minimum	34.5	4	1.9
	Average	38.5	7	2.1
	Maximum	43.1	11	1.1
	<i>DataMix</i>	37.5	10	1.2
Unified SLM	Task-Arithmetic	39.4	9	1.9
	Ties	39.3	7	1.7
	BreadCrumbs	39.3	11	1.5

ment phase, we find that the *BreadCrumbs* expert offers the best trade-off between high-amplitude improvements and consistent gains on 11 datasets. We also see that all unified SLMs are above the average expectation of the SLERP performances by 0.9%, but below the maximum values and below the top-performing *MedWiki* expert by 0.3%. Yet, this expert improves only on 10 datasets, one less than the BreadCrumbs-merged expert. We also observe for *BreadCrumbs* a stronger improvement over *MedWiki* of 10.6% on ICD10CM medical coding (see Table 12 in Appendix A.8), which is still lower than the specialized *MedCode* expert with an improvement of 36.9% over the *MedWiki* expert. We select the *BreadCrumbs* expert as *MediPhi* for its strong, balanced average score on CLUE+.

4.3 Post-Training

We show the results of aligning SLMs with the MediFlow dataset, as well as MediFlow-DPO in Table 4. To begin, we note that the alignment of *MediPhi* on all instructions (2.5M) leads to an accuracy of 41.9, an improvement of 5.4% over *Phi3.5*. By filtering MediFlow for top-quality 800k, we push the gain further by 6.5% to 43.0%. Then, we apply DPO using MediFlow-DPO to realize an improvement of 6.9% (18.9% relative), of which the maximum average ends up at 43.4%. Notable

Table 4: Average performance on CLUE+ of aligning *MediPhi* with MediFlow using SFT followed by DPO. We indicate **gains** and **losses** over the baseline. #DG stands for datasets with gains (out of 12 datasets). CV Δ stands for coefficient of variations of gains/losses.

		AVG \uparrow	#DG \uparrow	CV Δ \downarrow
Alignment of SLMs	<i>Phi3.5 mini</i>	36.5	-	-
	+SFT 800K	42.2	9	1.4
	+DPO	42.2	8	1.4
	<i>MediPhi</i>	39.3	11	1.5
	+SFT 2.5M	41.9	9	1.6
	+SFT 800K	43.0	9	1.4
	+DPO	43.4	9	1.4

gains in relative are: medical entities (SDoH) with 64.3% and on radiology reports (RRS QA) by 49.5% (see Table 12). If we align *Phi3.5* on MediFlow using SFT, we reach an accuracy of 42.2%. While surpassing *MediPhi-SFT* 2.5M by 0.3%, *MediPhi-SFT* 800k surpasses it by a margin of 2% (relative). Despite achieving gains overall, we observe a diminution in #DG from 11 to 9 for all models. The lower-performing datasets for all aligned models are Problem List Summarization and MediSumCode, which we hypothesize this result from a specific bias in MediFlow towards listing tasks like extracting problems from clinical notes or medical codes from discharge summaries (see Tables 11 and 12 in Appendix A.8).

Dada et al. (2024) highlighted a performance gap regarding medical LLMs’ performances in clinical settings based on the CLUE benchmark. Out of the twelve medical LLMs in their study, only two are improving upon their base model as shown in Table 5: BioMistral with DARE model merging, and Med42 with SFT alignment. By applying PIT, merging, and clinical alignment, *MediPhi* yields the highest improvement over its base model on CLUE+, achieving a +6.9% gain, compared to +1.1% for BioMistral and +1.2% for Med42.

Although *Phi3.5* already surpasses Mistral models, the *MediPhi* SLMs further widen this gap. Despite being less than half the size of LLaMA3 (8B), *MediPhi-Instruct* (3.8B) achieves near-parity on CLUE+, with a performance difference below 1%. *MediPhi-Instruct* outperforms LLaMA3 on four key datasets⁷: ICD10CM (+29.2%), MeDiSumCode (+13.9%), RRS QA (+5.8%), and MeQ-

⁷See Tables 11 and 12 in Appendix A.8

Table 5: Performances on CLUE+ of other medical LLMs compared with MediPhi models. We indicate **gains** and **losses**. #DG stands for datasets with gains (out of 12 datasets). CV Δ stands for coefficient of variations of gains/losses.

Model	AVG \uparrow	#DG \uparrow	CV Δ \downarrow
Mistral-7B-Instruct-v0.1	33.6	-	-
BioMistral-7B-DARE	34.7 (+1.1)	8	3.4
Phi-3.5 mini (3.8B)	36.5	-	-
MediPhi (3.8B)	39.3 (+2.8)	11	1.5
MediPhi-SFT (3.8B)	43.0 (+6.5)	9	1.4
MediPhi-Instruct (3.8B)	43.4 (+6.9)	9	1.4
Meta-Llama-3-8B-Instruct	44.1	-	-
Llama3-Med42-8B*	45.3 (+1.2)	5	7.8

*fine-tuned on ACI-Bench, i.e. not a few-shot setting.

Sum (+3.3%). Med42 improves over LLaMA3 by +1.2%, mainly due to a +27.7% boost on ICD10CM as well as being fine-tuned on the trainset of ACI-Bench. However, MediPhi surpasses Med42 in relative percentages on four tasks: ICD10CM by 2.1%, on RRS QA by 13.9%, SDoH by 5.2%, and MeDiSumCode by 47.6%. Moreover, it is competitive within 1% in absolute on three other tasks: MeQSum, MeDiSumQA and MEDEC. MediPhi models (3.8B) also achieve wider improvements over its base model with a #DG between 9 and 11, compared to 5 for Med42 (8B).

We summarized the overall progression of our approach in Figure 6.

5 Conclusion

In this work, we introduced MediPhi, the first clinical-focused SLM collection, alongside MediFlow, a large-scale synthetic instruction dataset for clinical alignment. Our results show that PIT significantly enhances domain adaptation, especially when combined with model merging. Notably, our medical coding expert surpasses GPT-4-0125 on the ICD10CM benchmark, and aligning MediPhi with MediFlow improves CLUE+ performance by 18.9% on average. By releasing these resources, we aim to enhance reproducibility, drive SLM adoption in clinical settings, and foster MediPhi’s continued development through its modular design.

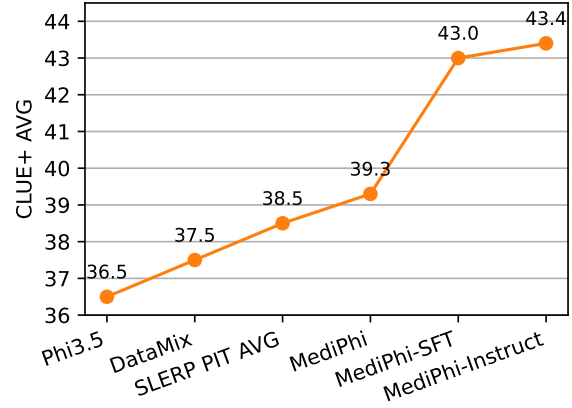


Figure 6: Summary of improvements from Phi3.5 to *MediPhi-Instruct* (SFT+DPO). *DataMix* is adapted on all corpora. *SLERP PIT AVG* is the average of five experts trained with PIT on each dataset. *MediPhi* is the unified expert from the five experts. *MediPhi-SFT* is an instruct model based on SFT only.

Limitations

This study necessitated a vast amount of resources in terms of 8x80GB A100 GPUs on Azure Machine Learning for approximately 12,000 GPU-hours, of which close to 3,600 GPU-hours were dedicated to achieve the final model with evaluations. It also required access to Azure OpenAI services for GPT-4o, GPT-4o mini and text-embedding-3-large (i.e. close to 25B input-output tokens).

The MediFlow corpus has few limits. The first limitation is that we set our scope to one-turn instructions instead of multi-turn conversation dataset like UltraChat (Ding et al., 2023). The second issue is the limited amount of very high complexity tasks (i.e. multi-step reasoning tasks) with also very long inputs and outputs. The third limit is that MediFlow is an English-only clinical corpus. To broaden the scope of MediFlow, future works might use seed data from clinical corpora as well as expanding our agentic pipeline. While MediPhi and MediPhi-instruct might conserve abilities from *Phi3.5-mini-instruct* (i.e. multilingual, conversational, safety alignment, etc.) by model merging (Hammoud et al., 2024), we hypothesize that these could be affected for which the impact was not studied outside medical and clinical abilities as well as instruction-following abilities. Thus, our recommendation is to use the MediPhi collection exclusively on clinical NLP tasks. We also strongly advice a verification of the model output by an expert in the specific medical field of the task.

Our CLUE+ benchmark expands upon the CLUE coverage in terms of tasks, input data

and datasets. While the coverage of the clinical field is large, a couple of gaps still remain. The information-extraction tasks are only represented by the SDoH dataset, and there is no input data on the nursing subfield (e.g. nurse-patient dialogs or notes).

Given the quick evolutions of OpenAI’s GPT-4o models as well as their stochastic nature, future exact reproduction of parts of this work may become impossible if the mentioned versions are no longer maintained.

Ethics Statement

We acknowledge that the substantial computational resources required for this work, including GPU hours and API calls, contribute to carbon emissions as well as limit the reproduction of this work. However, we believe that the knowledge and artifacts produced — such as the MediPhi SLM collection and the MediFlow corpus — offer significant positive impacts. These include enabling the adoption of SLMs in clinical settings, potentially reducing carbon emissions in the medium to long term, and fostering further research on continual pre-training in medical, clinical, and other domains.

Additionally, our modular approach with releases of models and datasets allows researchers and practitioners to build directly on this work, promoting accessibility and collaboration. In the future, strong language models in the clinical field can positively impact public health and support the work of healthcare providers, enhancing patient care and operational efficiency.

References

2019. Mtsamples. <https://mtsamples.com/>. Accessed: 2024-11-24.
- Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024a. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.
- Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024b. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.
- Emre Can Acikgoz, Osman Batur İnce, Rayene Bech, Arda Anıl Boz, Ilker Kesen, Aykut Erdem, and Erkut Erdem. 2024. Hippocrates: An open-source framework for advancing large language models in health-care. In *GenAI for Health: Potential, Trust and Policy Compliance*.
- Lisa Adams, Felix Busch, Tianyu Han, Jean-Baptiste Excoffier, Matthieu Ortala, Alexander Löser, Hugo JWL Aerts, Jakob Nikolas Kather, Daniel Truhn, and Keno K Bressem. 2024. Longhealth: A question answering benchmark with long clinical documents. *CoRR*.
- Arash Ahmadian, Seraphina Goldfarb-Tarrant, Beyza Ermis, Marzieh Fadaee, Sara Hooker, et al. 2024. Mix data or merge models? optimizing for diverse multi-task learning. In *Safe Generative AI Workshop*.
- Malaikannan Sankarasubbu Ankit Pal. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. <https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B>.
- Asma Ben Abacha and Dina Demner-Fushman. 2019. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28th - August 2*.
- Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis Goodwin, Sonya E. Shooshan, and Dina Demner-Fushman. 2019. Bridging the gap between consumers’ medication questions and trusted answers. In *MEDINFO 2019*.
- Asma Ben Abacha, Yassine M’rabet, Yuhao Zhang, Chaitanya Shivade, Curtis Langlotz, and Dina Demner-Fushman. 2021. Overview of the mediqua 2021 shared task on summarization in the medical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 74–85.
- Asma Ben Abacha, Wen-wai Yim, Yajuan Fu, Zhaoyi Sun, Meliha Yetisgen, Fei Xia, and Thomas Lin. 2024. Medec: A benchmark for medical error detection and correction in clinical notes. *arXiv preprint arXiv:2412.19260*.
- Yekun Chai. 2019. eval4ner: An all-round evaluation for named entity recognition. <https://github.com/cyk1337/eval4ner>.
- Canyu Chen, Jian Yu, Shan Chen, Che Liu, Zhongwei Wan, Danielle Bitterman, Fei Wang, and Kai Shu. 2024. Clinicalbench: Can llms beat traditional ml models in clinical prediction? In *GenAI for Health: Potential, Trust and Policy Compliance*.
- Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

- Clément Christophe, Avani Gupta, Nasir Hayat, Praveen Kanithi, Ahmed Al-Mahrooqi, Prateek Munjal, Marco Pimentel, Tathagata Raha, Ronnie Rajan, and Shadab Khan. 2023. Med42 - a clinical large language model.
- Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024a. Med42-v2: A suite of clinical llms. *arXiv preprint arXiv:2408.06142*.
- Clément Christophe, Tathagata Raha, Svetlana Maslenskova, Muhammad Umar Salman, Praveenkumar Kanithi, Marco Pimentel, and Shadab Khan. 2024b. Beyond fine-tuning: Unleashing the potential of continuous pretraining for clinical llms. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10549–10561.
- Jean-Philippe Corbeil. 2024. Iryonlp at mediqua-corr 2024: Tackling the medical error detection & correction task on the shoulders of medical agents. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 570–580.
- Amin Dada, Osman Alperen Koras, Marie Bauer, Amanda Butler, Kaleb E Smith, Jens Kleesiek, and Julian Friedrich. 2025. Medisumqa: Patient-oriented question-answer generation from discharge letters. *arXiv preprint arXiv:2502.03298*.
- Amin Dada, Osman Alperen Koras, Marie Bauer Contreras, Amanda Butler, Kaleb E Smith Seibold, Marc Constantin, and Jens Kleesiek. 2024. Does biomedical training lead to better medical performance? *arXiv preprint arXiv:2404.04067v4*.
- MohammadReza Davari and Eugene Belilovsky. 2024. Model breadcrumbs: Scaling multi-task model merging with sparse masks. In *European Conference on Computer Vision*, pages 270–287. Springer.
- Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDondald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.
- Fabio Dennstädt, Janna Hastings, Paul Martin Putora, Max Schmerder, and Nikola Cihoric. 2025. Implementing large language models in healthcare while balancing control, collaboration, costs and security. *npj Digital Medicine*, 8(1):143.
- Hantian Ding, Zijian Wang, Giovanni Paolini, Varun Kumar, Anoop Deoras, Dan Roth, and Stefano Soatto. 2024. Fewer truncations improve language modeling. In *Forty-first International Conference on Machine Learning*.
- Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. 2020. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, pages 3259–3269. PMLR.
- Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France. PMLR.
- Yanjun Gao, Timothy Miller, Majid Afshar, and Dmitriy Dligach. 2023. Bionlp workshop 2023 shared task 1a: Problem list summarization. In *Proceedings of the 22nd Workshop on Biomedical Language Processing*.
- 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.
- Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.
- Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Jordi Bayarri-Planas, Adrián Tormos, Daniel Hinjos, Pablo Bernabeu-Perez, Anna Arias-Duart, Pablo Agustin Martin-Torres, Lucia Urcelay-Ganzabal, Marta Gonzalez-Mallo, et al. 2024. Aloe: A family of fine-tuned open healthcare llms. *CoRR*.
- Suchin Gururangan, Ana Marasović, 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*, pages 8342–8360, Online. Association for Computational Linguistics.
- Hasan Hammoud, Umberto Michieli, Fabio Pizzati, Philip Torr, Adel Bibi, Bernard Ghanem, and Mete Ozay. 2024. Model merging and safety alignment: One bad model spoils the bunch. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13033–13046.

- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.
- Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.
- Daniel Jeong, Saurabh Garg, Zachary C Lipton, and Michael Oberst. 2024a. Medical adaptation of large language and vision-language models: Are we making progress? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12143–12170.
- Daniel P Jeong, Pranav Mani, Saurabh Garg, Zachary C Lipton, and Michael Oberst. 2024b. The limited impact of medical adaptation of large language and vision-language models. *arXiv preprint arXiv:2411.08870*.
- Zhengbao Jiang, Zhiqing Sun, Weijia Shi, Pedro Rodriguez, Chunting Zhou, Graham Neubig, Xi Lin, Wen-tau Yih, and Srini Iyer. 2024. 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.
- 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):6421.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.
- Achintya Kundu, Rhui Dih Lee, Laura Wynter, Raghu Kiran Ganti, and Mayank Mishra. 2024. Enhancing training efficiency using packing with flash attention. *arXiv preprint arXiv:2407.09105*.
- Sunjun Kweon, Junu Kim, Jiyou Kim, Sujeong Im, Eunbyeol Cho, Seongsu Bae, Jungwoo Oh, Gyubok Lee, Jong Hak Moon, Seng Chan You, et al. 2024. Publicly shareable clinical large language model built on synthetic clinical notes. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5148–5168.
- Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickaël Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5848–5864.
- Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.
- Chin-Yew Lin. 2004. *ROUGE: A package for automatic evaluation of summaries*. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Fenglin Liu, Zheng Li, Hongjian Zhou, Qingyu Yin, Jingfeng Yang, Xianfeng Tang, Chen Luo, Ming Zeng, Haoming Jiang, Yifan Gao, Priyanka Nigam, Sreyashi Nag, Bing Yin, Yining Hua, Xuan Zhou, Omid Rohanian, Anshul Thakur, Lei Clifton, and David A. Clifton. 2024. *Large language models are poor clinical decision-makers: A comprehensive benchmark*. *medRxiv*.
- Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: designing data and methods for effective instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 22631–22648.
- Shayne Longpre, Robert Mahari, Ariel Lee, Campbell Lund, Hamidah Oderinwale, William Brannon, Nayan Saxena, Naana Obeng-Marnu, Tobin South, Cole Hunter, et al. 2024. Consent in crisis: The rapid decline of the ai data commons. In *NEURIPS*.
- Kevin Lybarger, Meliha Yetisgen, and Özlem Uzuner. 2023. The 2022 n2c2/uw shared task on extracting social determinants of health. *Journal of the American Medical Informatics Association*, 30(8):1367–1378.
- Leland McInnes, John Healy, Steve Astels, et al. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205.
- Seyed Iman Mirzadeh, Mehrdad Farajtabar, Dilan Gorur, Razvan Pascanu, and Hassan Ghasemzadeh. Linear mode connectivity in multitask and continual learning. In *International Conference on Learning Representations*.
- Arindam Mitra, Luciano Del Corro, Guoqing Zheng, Shweti Mahajan, Dany Rouhana, Andres Coda, Yadong Lu, Wei-ge Chen, Olga Vrousos, Corby Rosset, et al. 2024. Agentinstruct: Toward generative teaching with agentic flows. *arXiv preprint arXiv:2407.03502*.
- 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. *Advances in Neural Information Processing Systems*, 36:50358–50376.
- Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. *Orca: Progressive learning from*

- complex explanation traces of gpt-4. *Preprint*, arXiv:2306.02707.
- Clara Na, Ian Magnusson, Ananya Harsh Jha, Tom Sherborne, Emma Strubell, Jesse Dodge, and Pradeep Dasigi. 2024. Scalable data ablation approximations for language models through modular training and merging. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21125–21141.
- National Library of Medicine. 2003. [Pmc open access subset \[internet\]](#). Bethesda (MD): National Library of Medicine.
- Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *Medicine*, 84(88.3):77–3.
- Harsha Nori, Naoto Usuyama, Nicholas King, Scott Mayer McKinney, Xavier Fernandes, Sheng Zhang, and Eric Horvitz. 2024. From medprompt to o1: Exploration of run-time strategies for medical challenge problems and beyond. *arXiv preprint arXiv:2411.03590*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikanan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.
- Alexandre Ramé, Johan Ferret, Nino Vieillard, Robert Dadashi, Léonard Hussenot, Pierre-Louis Cedo, Pier Giuseppe Sessa, Sertan Girgin, Arthur Douillard, and Olivier Bachem. 2024. Warp: On the benefits of weight averaged rewarded policies. *CoRR*.
- Alexey Romanov and Chaitanya Shivade. 2018. [Lessons from natural language inference in the clinical domain](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.
- Nikhil Sardana, Jacob Portes, Sasha Doubov, and Jonathan Frankle. 2024. Beyond chinchilla-optimal: accounting for inference in language model scaling laws. In *Proceedings of the 41st International Conference on Machine Learning*, pages 43445–43460.
- Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. 2022. Fine-tuned language models are continual learners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6107–6122.
- Ofir Ben Shoham and Nadav Rappoport. 2024. Med-conceptsq: Open source medical concepts qa benchmark. *Computers in Biology and Medicine*, 182:109089.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.
- Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. 2023. Clinical camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. *CoRR*.
- Vishaal Udandara, Ameya Prabhu, Adhiraj Ghosh, Yash Sharma, Philip Torr, Adel Bibi, Samuel Albanie, and Matthias Bethge. 2024. No "zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and An Chang Ho. 2022. [Will we run out of data? limits of llm scaling based on human-generated data](#).
- Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2023. Notechat: a dataset of synthetic doctor-patient conversations conditioned on clinical notes. *arXiv preprint arXiv:2310.15959*.
- Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR.
- C Wu, W Lin, X Zhang, Y Zhang, W Xie, and Y Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association: JAMIA*, pages ocae045–ocae045.
- Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *CoRR*.
- Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. 2024a. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36.
- Prateek Yadav, Tu Vu, Jonathan Lai, Alexandra Chronopoulou, Manaal Faruqi, Mohit Bansal, and Tsendsuren Munkhdalai. 2024b. What matters

for model merging at scale? *arXiv preprint arXiv:2410.03617*.

Rui Yang, Ting Fang Tan, Wei Lu, Arun James Thirunavukarasu, Daniel Shu Wei Ting, and Nan Liu. 2023. Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4):255–263.

Wen-wai Yim, Yajuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Acibench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*, 10(1):586.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*.

Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, et al. 2024. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, pages 1–13.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023a. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. 2023b. Alpacre: Instruction-tuned large language models for medical application. *CoRR*.

Zhengyun Zhao, Qiao Jin, Fangyuan Chen, Tuorui Peng, and Sheng Yu. 2023. A large-scale dataset of patient summaries for retrieval-based clinical decision support systems. *Scientific data*, 10 1:909.

A Appendix

A.1 Information on Dataset Groups

The *PubMed* group is the largest by 2 orders of magnitude with about 48B tokens. PubMed Central ([National Library of Medicine, 2003](#)) have a segment for commercial use, while abstracts are public.

On the general medical side, we have medical Wikipedia known as *MedWiki* ([Corbeil, 2024](#)). We also fetch the open medical *Guidelines* ([Chen et al., 2023](#)) which comes from multiple recognized health organizations (e.g. WHO). Then, we gather medical coding as *MedCode* corpora from public websites: ICD9CM, ICD9PROC, ICD10CM, ICD10PROC and ATC. The ICD coding is a wide taxonomy of diseases and medical conditions as

well as procedure codes, while *Anatomical Therapeutic Chemical* (ATC) codes are a classification of medical drugs.

In the *Clinical* group, we leverage PMC-Patients v2 ([Zhao et al., 2023](#)) — subset with distribution- and commercial-friendly CC licenses — which contains clinical cases. We also include synthetic derivative datasets under same licensing conditions: *Asclepius* ([Kweon et al., 2024](#)) containing clinical documents and *NoteChat* ([Wang et al., 2023](#)) containing doctor-patient dialogues. Furthermore, we include *MTSamples* ([mts, 2019](#)), a public database filled with various de-identified clinical documents.

A.2 Hyperparameters & Pre-Training Optimization Strategies

We use the HuggingFace ecosystem (*transformers*, *trl*, *accelerate*, and *datasets*) to train models leveraging a multi-node configuration. The hyperparameters of the continual pre-training are listed in Table 6. The hyperparameters of the alignment process are provided in Tables 7 and 8 for SFT and DPO, respectively.

[Ding et al. \(2024\)](#) highlighted the importance of longer context windows, minimal truncations, and restricted cross-document attention during pre-training. Many prior works in medical NLP rely on concatenate-and-truncate strategies, despite introducing significant truncation ([Toma et al., 2023](#); [Christophe et al., 2023, 2024a](#); [Labrak et al., 2024](#)). Instead, we implement Best-Fit Packing ([Ding et al., 2024](#)) for the *PubMed* expert and *DataMix* baseline, ensuring efficient token utilization. The other experts are trained with PIT one document at a time. For all expert models, we avoid padding and cross-document attention ([Kundu et al., 2024](#))

Table 6: Hyperparameters for CPT excluding PIT.

Hyperparameter	
Maximum Tokens	4,096
Optimizer	AdamW
LR Scheduler	Linear Warmup - Cosine
# Warmup Steps	35
Maximum LR	1×10^{-4}
Epochs	1
NEFTune α	5
Batch Size per GPU	16
# GPUs	16
Gradient Accumulation	2
Effective Batch Size	512

in trainings.

A.3 Validation Sets

The twelve validation sets cover subjects such as clinical case, clinical knowledge, medication, ICD10CM code definitions, radiology report, clinical NLI, QA on discharge letter, medical codes from discharge letter, problem list from clinical notes, summarization of patient inquiry, QA on medical consultation and QA on multiple EHR documents. We control its diversity by generating the embeddings of the question with its context, and applying the density-based clustering method HDBSCAN (McInnes et al., 2017). The final valid sets are the combination of the outliers (i.e. not assigned to a cluster) and one sample per cluster for a total up to 1,200 samples each.

A.4 Generation of MediFlow triplets

We parametrized the prompt to generate MediFlow based on the task type, input data, output format (plain text or JSON), difficulty level (moderate, moderate-hard, hard, very hard, or extreme), and number of input-output example pairs (3 or 4 per instruction). For the difficulty level, we favored the sampling of hard, very hard and extreme levels by a ratio of 3:1.

The **14 task types** are: summarization, question-answering, multiple-choice question-answering, named entity recognition, relation extraction, classification, reasoning & diagnosis, textual entailment, text simplification, text expansion, abbreviation expansion, aspect-oriented keyword extraction, error detection & correction, and note scoring.

The 36 input-data types with various levels of granularity (1 (only complete document) up to 7 (complete document with 6 individual sections))

Table 7: Hyperparameters for SFT.

Hyperparameter	
Optimizer	AdamW
LR Scheduler	Linear Warmup - Cosine
# Warmup Steps	40
Maximum LR	2×10^{-5}
Epochs	2
NEFTune α	5
Batch Size per GPU	16
# GPUs	8
Gradient Accumulation	2
Effective Batch Size	256

Table 8: Hyperparameters for DPO.

Hyperparameter	
Optimizer	AdamW
LR Scheduler	Linear Warmup - Cosine
# Warmup Steps	50
Maximum LR	1×10^{-6}
Epoch	1
Batch Size per GPU	8
# GPUs	16
Gradient Accumulation	1
Effective Batch Size	128
β	0.1

resulting in a total of 98 fine-grained input-data types. Decisions for the granularity levels were made based on the document length — e.g. taking the complete short documents compared to segmenting long documents. The data types are:

1. Discharge Summary (6)
2. SOAP Clinical Note (5)
3. Clinical Note (6)
4. Progress Note (4)
5. Admission Note (6)
6. Scientific Article (8)
7. Clinical Case (3)
8. Nursing Note (1)
9. Monitoring Data of Vital Signs (1)
10. Referral Letter (1)
11. Emergency Department Note (7)
12. Laboratory Report (1)
13. Radiology Report (3)
14. Doctor-Patient Conversation (5)
15. Nurse-Patient Dialog (5)
16. Operative Note (8)
17. Consultation Note (1)
18. Pathology Report (1)
19. Prescription Note (1)
20. Preoperative Assessment (1)
21. Postoperative Note (1)
22. Therapy Notes (5)
23. Immunization Record (1)
24. Screening Report (1)
25. Consent Form (1)
26. Care Plan (1)
27. Dietary Notes (1)
28. Psychiatric Evaluation (5)
29. Social Work Note (1)
30. End-of-Life Care Documentation (1)
31. Triage Note (1)
32. Dental Record (1)

33. Home Health Care Report (1)
34. Genetic Testing Report (1)
35. Incident Report (1)
36. Patient Education Material (1)

In Figures 7, 8 and 9, we present the histograms of tokens for generated instructions, inputs and outputs, respectively.

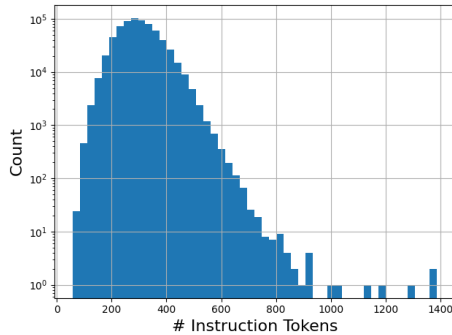


Figure 7: Distribution of instruction tokens in MediFlow with y-axis in log scale. The average is 301 ± 295 tokens.

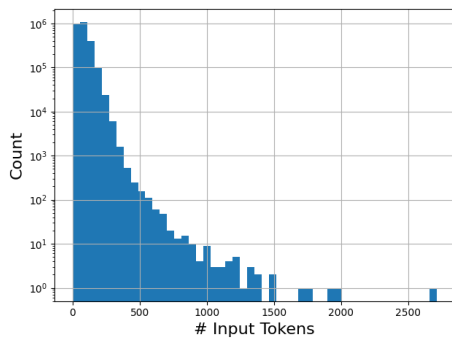


Figure 8: Distribution of input tokens in MediFlow with y-axis in log scale. The average is 76 ± 67 tokens.

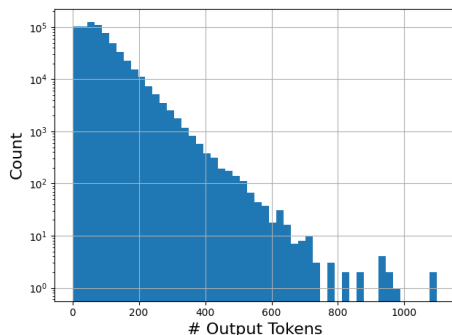


Figure 9: Distribution of output tokens in MediFlow with y-axis in log scale. The average is 79 ± 66 tokens.

MediFlow Prompt

You are an expert user querying about the medical and clinical domain in natural language processing. You will define instructions for a precise task with clear constraints in the medical/clinical domain. You must be very detailed in the instructions regarding input data, the `{{task_type}}` task and the desired output format which is `{{output_format}}`. You must use `{{input_data}}` as the task's input in some ways, especially `{{input_data_granular}}`. You must put these between the tags `<instruction>...</instruction>`. You must define clearly based on these parameters a specific task from the given type, specific expected input data and output format. You must make a task with a `{{difficulty}}` difficulty on a scale of 6 levels (low, moderate, moderate-hard, hard, very hard and extreme). You must not mention the task level in your own instructions. You must only write the instructions, i.e. do not use markdown, no extra comment, etc.

Then, you must give `{{number_examples}}` examples between tags `<examples>` `<example>` `<input>...</input>` `<output>...</output>` `</example>` `</examples>` containing input and output. You must give a complete example with input and output at the end.

You must use interesting and complex examples requiring abstractive medical capabilities to infer the output from the input. So, you must avoid any obvious input and output, and you must favor very difficult pairs. You must strictly use the format `<instruction>...</instruction>` followed by `<examples>` `<example>` `<input>...</input>` `<output>...</output>` `</example>` `</examples>` with exactly those tag names. You must use synonyms for all headers (or no header at all) to avoid leaking current vocabulary into the instructions, also use different ways to structure (or not) and detail the instructions (e.g. bullet points, sections, narrative form, or else).

A.5 Generation of MediFlow Judge Scores

We used GPT-4o-mini to score all MediFlow on 5 criteria: quality, alignment with instruction requirements, coherence, realism and difficulty. We applied a temperature of 1.0 with chain-of-thought for each criterion and average scores across 5 generated samples for self-consistency. All scores are an integer on a scale from 1 to 4 as defined below.

LLM-as-a-Judge MediFlow Prompt

You are the best instruction designer for language models in the medical/clinical field. You will be given instructions with constraints for a task to perform on clinical documents. Your task is to give a critical assessment of the instructions as a nested JSON object. For each criteria as a key, the value is a JSON object containing a "rationale" along a "score" on a scale of 1 to 4. The criteria are: quality, alignment with instruction requirements, coherence, realism and difficulty.

INSTRUCTION REQUIREMENTS::
Here are the instruction requirements:

- Defined very detailed instructions for a precise task with clear constraints in the medical/clinical domain.
- Clearly defined task type, specific input data and output format. It can also contain examples.
- Only write the instructions, i.e. do not use markdown, no extra comment, etc.

SCALES:::
 {{criteria_definitions}}
 INSTRUCTIONS:::
 {{instruction}}
 You must only output the JSON object for the critical assessment. Remember valid scores are: 1, 2, 3 and 4.
 OUTPUT:::

A.5.1 Judge Score Distributions

In Figure 10, we displayed the histograms (with a logarithm y axis) of the judge scores predicted on nearly 700k unique instructions. For all criteria, we note that a large amount of generated instructions have a peak at 3 with large distribution between 3 and 4 (on a scale from 1 to 4), which is considered on the high end. We observe a different trend with respect to the difficulty criterion, where a larger peak is still at 3 (i.e. difficulty level of hard) but some portion of the dataset (i.e. less than a 100k instruction) have scores between 2 and 3.

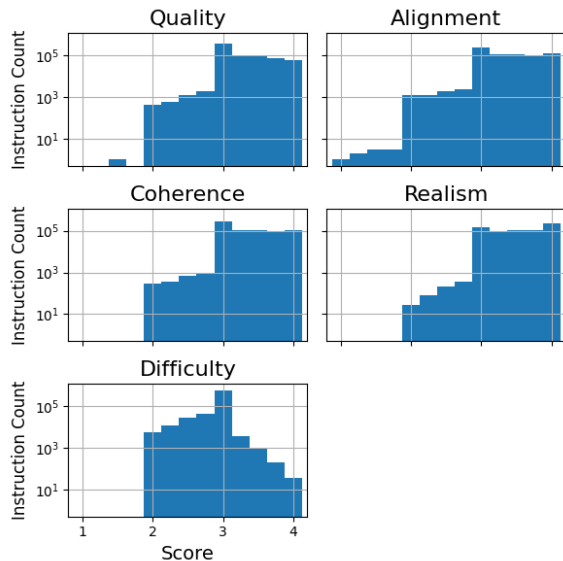


Figure 10: Judge Score Distributions across MediFlow. Y-axis is set to log scale.

A.6 MediFlow Scatterplots

We generated embedding with OpenAI *text-embedding-3-large* (truncated at 256 dimensions) for each instruction in MediFlow. Then, we applied PCA at 50 dimensions and t-SNE to get 2 dimensions which we displayed on scatterplots with

other labels (output formats, difficulty levels, task type and input-data type) in Figures 13, 11 and 12. While the scatter plot on the difficulty levels do not exhibit clear clustering patterns, we see that the others have distinctive patterns at different scales. The output format is affecting local clustering patterns often appearing at small scale as two closed blobs. The task and input-data types affect the macrostructure at similar large scale.

A.7 Generation of Marginally Wrong Outputs for DPO

DPO requires to have a prompt (corresponding in our framework to *instruction* with *input*) along a chosen response (i.e. *output*) and a rejected response. The rejected response must be less preferable compared to the chosen one. First, we filter MediFlow to around 85k instructions keeping only the top triplets on the *quality* metric along with a stratification by task type, input-data type and output format. To have the best trade-off between diversity and high quality, we sampled 3 input-output pairs for the first top high-quality 20k instructions, followed by 2 pairs for the next 25k and one pair for the last 40k. After filtering, this resulted the MediPhi-DPO dataset of 130,852 triplets. Then, we prompted GPT-4o-0806 with a triplet at a temperature of 1.0 with a randomly sampled error type on the following prompt to generate a marginally wrong output as rejected output. The error types are : ambiguity, partial correctness, over-verbosity, brevity, unbalanced detail, stylistic issues, factual inaccuracy, logical flaws, misinterpretation, simplistic reasoning, grammatical errors, and spelling errors.

Marginally Worst Output Prompt

You are a subtle flaw introducer, trained to degrade a high-quality response by introducing a specific type of error without making it obviously incorrect. You will receive the triplet (instructions, input, output). The instructions are what to do to the input to get the output, which is the good response. Your task is to provide a wrong output but in a subtle way. Here's the definitions of error types:
 {{error_type_definitions}}
 INSTRUCTIONS {{instruction}}
 INPUT {{input}}
 OUTPUT {{output}}
 Generate a wrong response that is marginally worse than the good output. Introduce an error of type {{error_type}}, ensuring the response still seems reasonable at first glance.

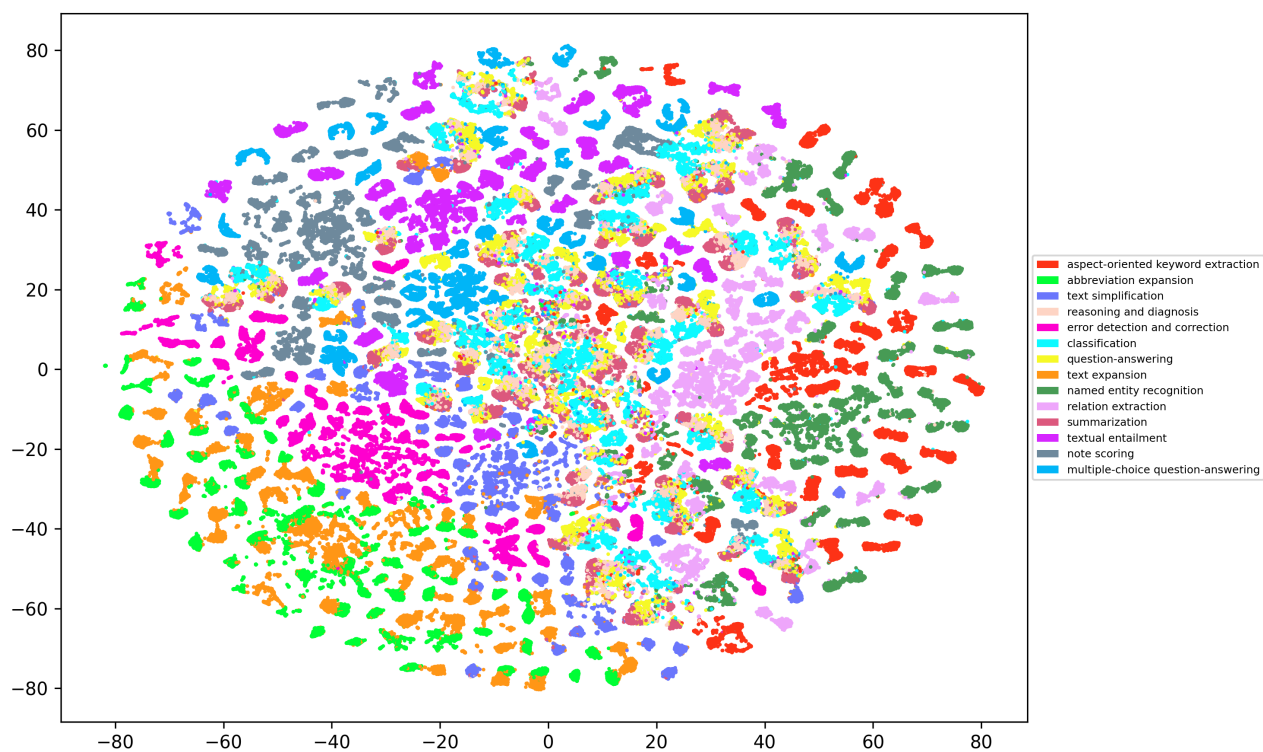


Figure 11: t-SNE 2D scatterplot of MediFlow (2.5M) using OpenAI *text-embedding-3 large* API with tasks as colors.

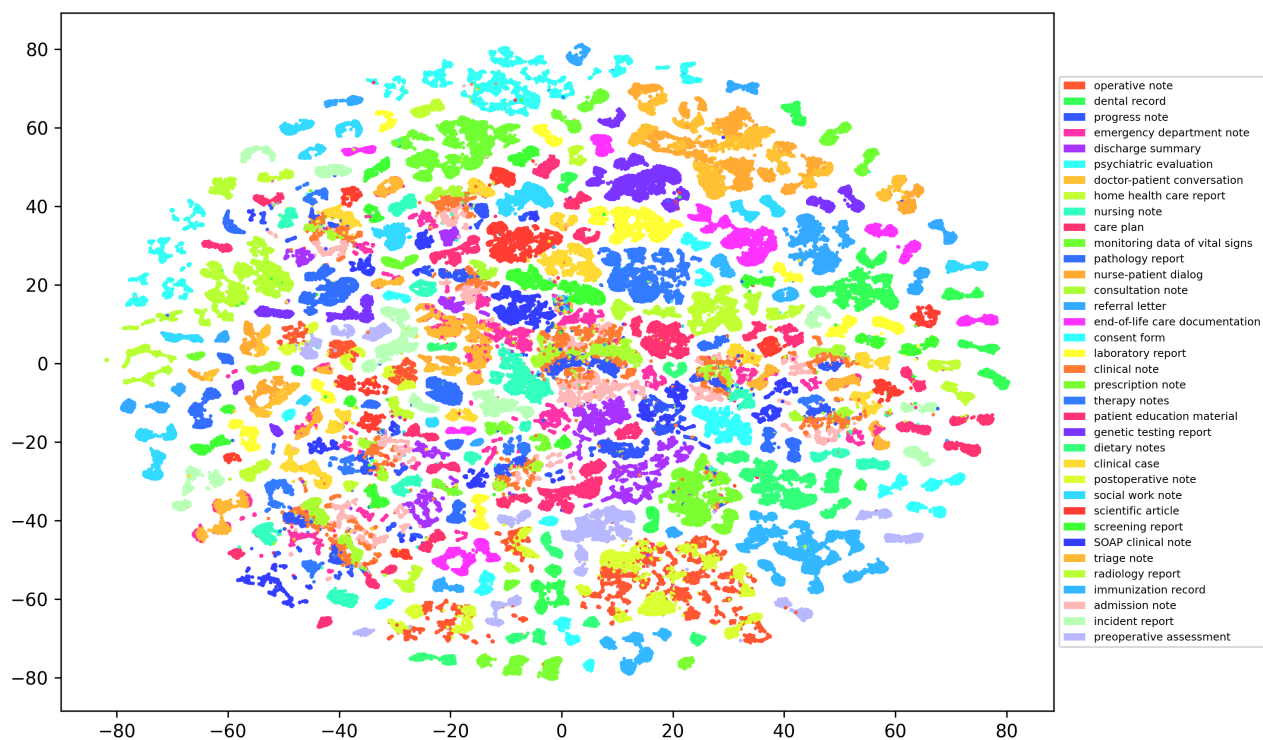


Figure 12: t-SNE 2D scatterplot of MediFlow (2.5M) using OpenAI *text-embedding-3 large* API with input-data types as colors.

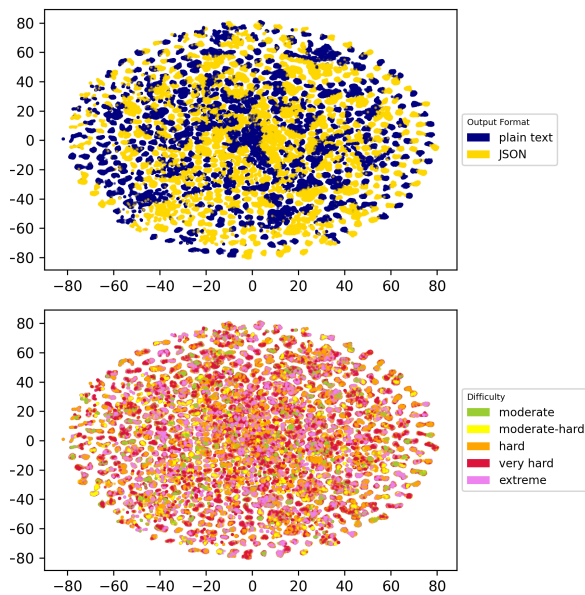


Figure 13: t-SNE 2D scatterplots of MediFlow (2.5M) with output format (top) and difficulty level (bottom) as colors.

A.8 CLUE+ Benchmark

A.8.1 MedicationQA

MedicationQA (Ben Abacha et al., 2019) consists of 674 consumer health questions collected from MedlinePlus⁸. These questions were linked to a matching excerpt from a trusted source that contains the answer.

Through manual inspection we identified two obstacles for effective few-shot evaluation of LLMs on MedicationQA:

- Some questions were poorly formulated or a search query instead of a question
- Some of the answers did not give a specific answer to the question
- Answers were often unconcise since they were not formulated as a direct answer to the question, but a retrieved excerpt from a text

Based on these observations we prompted an LLM⁹ to remove non-matching question answer pairs and formulate a direct answer based on the given excerpt. Figure 14 shows an example for a reformulated answer. This process resulted in 485 question-answer pairs.

For benchmarking we use the following system prompt:

MedicationQA System Prompt

You are a highly skilled assistant, specifically trained to assist patients with medical questions. Give a concise answer. Do not mention anything that was not explicitly asked for. Do not generate anything else.

A.8.2 MEDIQA-RRS QA

MeDIQA-RRS (Ben Abacha et al., 2021) is a summarization dataset based on findings and impressions sections of Radiology reports from the Indiana University chest X-ray dataset (Demner-Fushman et al., 2016) and reports from the Stanford Health Care system. The findings serve as model inputs while the impressions are treated as the summarization ground truth. We evaluate on the test split, which contains 600 finding-impression pairs. We observed that the impressions were often not a complete summary, but rather an answer to a question posed before the exam (e.g., "No acute cardiopulmonary findings."). Additionally, the comprehensiveness of impressions varied substantially between reports. To address this we reformulated the impressions to a series of question-answer pairs using an LLM⁹. We specified that the answers should use the same wording as in the original impression section to ensure factuality. We confirmed that answers were not changed, filtering answers with exact string matching. Figure 15 shows an example of this reformulation step.

The following system prompt was used in the evaluation:

MEDIQA-RRS QA System Prompt

You are a highly skilled assistant, specifically trained to interpret radiology reports. You will receive the findings section of a report along with specific questions. Provide concise, focused answers based solely on the information provided. Avoid adding any details not explicitly requested.

A.8.3 ACI-Bench

ACI-Bench (Yim et al., 2023) is a dataset that takes as input a doctor-patient dialog and the task is to generate a clinical note of five sections.

ACI-Bench System Prompt

Task: Generate an Extremely Detailed Clinical Note from a Doctor-Patient Conversation
 Role: You are an expert medical professional responsible for creating a highly detailed, comprehensive, and fully elaborated clinical note from a doctor-patient conversation.
 Instructions: - You will receive a doctor-patient conversation as input. - Your task is to produce a long, exhaustive clinical note covering all clinically relevant details. - The note must be extremely detailed, ensuring no important information is omitted. - Expand on every symptom, examination finding, and treatment plan to provide a complete, structured summary.

⁸<https://medlineplus.gov>

⁹meta-llama/Llama-3.3-70B-Instruct

Question
 is it nessasary to ween off of cymbalta before starting effexor?

Orginal Answer
 Switching from one antidepressant to another is frequently indicated due to an inadequate treatment response or unacceptable adverse effects. All antidepressant switches must be carried out cautiously and under close observation. Conservative switching strategies involve gradually tapering the first antidepressant followed by an adequate washout period before the new antidepressant is started. This can take a long time and include periods of no treatment with the risk of potentially life-threatening exacerbations of illness. Clinical expertise is needed for more rapid or cross-taper switching as drug toxicity, including serotonin syndrome, may result from inappropriate co-administration of antidepressants. Some antidepressants must not be combined. Antidepressants can cause withdrawal syndromes if discontinued abruptly after prolonged use. Relapse and exacerbation of depression can also occur. Gradual dose reduction over days to weeks reduces the risk and severity of complications.

Reformulated Answer
 Yes, it is necessary to taper off Cymbalta before starting Effexor to avoid potential complications, including withdrawal syndromes and serotonin syndrome. A gradual dose reduction over days to weeks is recommended to reduce the risk and severity of complications.

Figure 14: An example of a verbose answer in MedicationQA and the reformulated answer we replaced it with.

Findings
 Fracture deformity proximal right humerus. Hyperinflation lungs. No pulmonary consolidation. ____ opacity left base compatible ____ atelectasis or ____ scarring. The cardiomedialastinal silhouette appears unremarkable. Mild atherosclerotic calcification aorta. Prior chest surgery. Costophrenic ____ clear. Visualized spine vertebrae appear normal in ____ and alignment.

Impression
 Fracture deformity proximal right humerus. No pulmonary consolidation

Original Sample

Reformulated QA

Question
 What is noted in the proximal right humerus?

Answer
 Fracture deformity.

Question
 Is there any pulmonary consolidation?

Answer
 No.

Figure 15: An example of a formulation of questions based on the impression section.

Output Format: Your clinical note must be structured into the following five required sections:

1. **CHIEF COMPLAINT** A precise and natural-language statement describing the primary reason for the visit. Usually written in the patient's own words (e.g., "Chest pain for two days.>").
 2. **HISTORY OF PRESENT ILLNESS** Provide a detailed, full narrative including: - Onset: Exact time course (sudden/gradual, exact duration). - Duration: Progression over time. - Severity: Patient's description or numeric scale (1-10). - Location: Anatomical specificity. - Modifying Factors: What worsens or improves symptoms (activities, medications). - Associated Symptoms: Describe all related symptoms. - Prior Treatments: List exact medications, dosages, patient responses.
 3. **PHYSICAL EXAM** Expand on every finding instead of using short labels. Describe specific observations for each system: - Vital Signs: BP, HR, Temp, O2 Sat, RR. - General Appearance: Patient's demeanor, level of distress. - Neurological: Reflexes, motor strength, sensory findings. - Cardiovascular: Detailed heart sounds, pulses, peripheral findings. - Pulmonary: Breath sounds, presence of wheezes/crackles. - Abdominal: Bowel sounds, tenderness, distension.
 4. **RESULTS** Include all relevant diagnostic data, explaining why each result matters. List both abnormal and pertinent normal findings.
 5. **ASSESSMENT AND PLAN (A/P)** Diagnosis & Differential Diagnosis: Explain why the most likely condition was chosen. Plan: - Medications: Name, dose, frequency, and rationale. - Additional Tests: Imaging, lab work, specialist referrals. - Follow-up Plan: Next steps, expected outcomes. - Patient Education: Instructions, lifestyle modifications. - Ensure that every plan component is justified.
- Reminder:
- You must include every detail exhaustively from the dialog.
 - You must justify every diagnosis and treatment.
 - You must provide a thorough explanation of the assessment and plan.
- Thus, you must expand on every detail from the dialog in the note.

A.8.4 Social Determinants of Health

SDoH System Prompt

Task: Entity Extraction for Social Determinants of Health (SDOH)
Your job is to extract key socio-demographic and behavioral factors from the provided text. Your output must be a list of JSON objects, where each object contains:

- "entity": The specific category of the extracted information, chosen from the predefined taxonomy below.
- "label": The corresponding label from the taxonomy.
- "value": The exact phrase from the document that represents the entity.

Entities & Taxonomy:

1. Employment

- "entity": "Employment" (general employment-related mention)
- "label": "StatusEmploy" → "employed", "unemployed", "retired", "on disability", "student", "homemaker"
- "label": "Duration" → "for the last five years", "since 2010"
- "label": "History" → "15 years ago", "in 2005"
- "label": "Type" → Specific occupations (e.g., "geologist", "registered nurse", "office work")

2. Living Status

- "entity": "LivingStatus" (mentions of where and how someone lives)
- "label": "StatusTime" → "current", "past", "future"
- "label": "TypeLiving" → "alone", "with family", "with others", "homeless"
- "label": "Duration" → "for the past ten years", "since 2015"
- "label": "History" → "moved out five years ago", "in 2010"

3. Substance Use

- "entity": "Alcohol", "Drug", "Tobacco" (mentions of substance use)
- "label": "StatusTime" → "none", "current", "past"
- "label": "Duration" → "for the past eight years"
- "label": "History" → "seven years ago", "in 2005"
- "label": "Method" → "smoke", "snort", "inhale", "inject" (for drugs), "chew", "vape" (for tobacco)
- "label": "Type" → "beer", "wine", "heroin", "marijuana", "cigarettes"

- "label": "Amount" → "# of drinks", "# of cigarettes", "# of times"
 - "label": "Frequency" → "daily", "monthly", "yearly"
- Example Output (Using Entities from the Document): ["entity": "Employment", "label": "StatusEmploy", "value": "full-time student", "entity": "LivingStatus", "label": "TypeLiving", "value": "currently lives alone", "entity": "Alcohol", "label": "StatusTime", "value": "drinks occasionally", "entity": "Drug", "label": "History", "value": "used marijuana seven years ago"]

Guidelines for Extraction:

1. Extract only explicitly mentioned entities—do not infer information.
 2. Use exact text from the document—the "value" must match the original wording.
 3. Categorize precisely—select the most appropriate "entity" and "label".
 4. Ensure valid JSON format—return structured, machine-readable output.
- Your final output should be a list of JSON objects containing only the entities present in the document.

A.8.5 MEDEC

MEDEC System Prompt

Task: Medical Error Detection in Clinical Text Role: You are an expert medical reviewer analyzing clinical text for accuracy. Your task is to determine whether there is one medical error in the provided text.

Input Format: The input consists of multiple sentences. Each sentence starts with a Sentence ID, followed by the sentence itself. Sentences are formatted one per line with a space separating the ID and the sentence text.

Types of Errors to Detect: Diagnosis errors (incorrect or conflicting diagnoses) Treatment errors (inappropriate or missing treatments) Management errors (incorrect clinical decision-making) Causation errors (incorrect understanding of disease causes or progression)

Output Format: If one sentence contains a medical error, return only the Sentence ID of that sentence. If no errors are found, return "-1" (without quotes). You must not provide any explanation, only the Sentence ID or -1.

Example with an error: input :

0 The patient was diagnosed with bacterial pneumonia and prescribed amoxicillin.

1 The recommended treatment for viral pneumonia is antibiotics.

2 The patient showed signs of improvement after three days.

output:1

Example without error:

input:0 The patient was diagnosed with bacterial pneumonia and prescribed amoxicillin.

1 The recommended treatment for bacterial pneumonia is antibiotics.

2 The patient showed signs of improvement after three days.

output:-1

Table 9: The configuration used for CLUE+ subset datasets of the benchmark regarding few shots and metric. The decoding strategy is a greedy decoding for all.

Dataset Name	Few-Shots	Metric
ICD10CM	3	AVG Accuracy
MedicationQA	3	Rouge-1 F1 (Lin, 2004)
RRS QA	3	Rouge-1 F1 (Lin, 2004)
SDoH	4	Type-Match F1 w/ boundary overlap (Chai, 2019)
ACI-Bench	1	Rouge-1 F1 (Lin, 2004)
MEDEC	2	Sentence ID Accuracy

Table 10: CLUE+ benchmark datasets with task types, input, and output specifications. Task types are: question-answering (QA), summarization, reasoning and information extraction (IE).

	Task	Dataset Name	Input	Output
CLUE	NLI	MedNLI (Romanov and Shivade, 2018)	Premise + Hypothesis	Label
	Summary	MeQSum (Ben Abacha and Demner-Fushman, 2019)	Consumer Health Question	Summary question
	Summary	Problem list summarization (Gao et al., 2023)	Progress notes	Problem list
	QA	LongHealth (Adams et al., 2024)	Clinical records + Question	Answer
	QA	MeDiSumQA (Dada et al., 2025)	Discharge letter + Questions	Answers
	Reasoning	MeDiSumCode (Dada et al., 2024)	Discharge letter	ICD10CM codes
CLUE+	QA	MedConceptsQA ICD10CM (Shoham and Rappoport, 2024)	Code + definition options	Answer
	QA	MedicationQA (Ben Abacha et al., 2019)	Question on medication	Answer
	QA	MEDIQA-RRS QA (Ben Abacha et al., 2021)	Findings + Questions	Answers
	IE	SDoH (Lybarger et al., 2023)	Clinical note	Entities
	Summary	ACI-Bench (Yim et al., 2023)	Doctor-patient dialog	Clinical note
	Reasoning	MEDEC (Ben Abacha et al., 2024)	Clinical note	Error detection

Table 11: Performances on the CLUE subset of datasets for the merged and aligned versions of MediPhi as well as other medical LLMs. PLS stands for Problem List Summary while LH refers to LongHealth.

		MedNLI	PLS	MeQSum	LH	MeDiSumQA	MeDiSumCode
Baseline	<i>Phi-3.5-mini-instruct</i>	66.6	28.4	36.7	45.9	25.9	41.1
SLERP	DataMix	68.5	29.0	37.7	45.7	26.6	41.4
	PubMed	68.3	29.2	37.6	45.7	26.3	41.0
	Clinical	69.2	29.4	38.1	43.5	26.7	40.5
	MedWiki	72.8	29.2	37.6	43.6	25.1	41.7
	MedCode	68.5	22.3	33.5	45.7	23.6	39.0
	Guideline	70.3	29.8	37.6	41.1	25.1	41.9
BreadCrumbs	MediPhi	66.9	28.8	37.9	45.7	26.1	41.7
MediFlow	MediPhi-SFT	70.6	26.9	42.8	44.2	28.8	35.0
	MediPhi-Instruct	71.0	26.0	42.8	45.0	29.1	37.2
Other Medical LLMs	Mistral-7B-Instruct-v0.1	64.8	25.0	31.1	30.0	25.5	13.9
	BioMistral-7B-DARE	66.8	28.4	34.5	30.5	25.7	21.3
	Meta-Llama-3-8B-Instruct	74.1	31.6	39.5	58.8	30.3	27.8
	Llama3-Med42-8B	77.5	32.4	42.8	57.9	29.7	25.2

Table 12: Performances on the new CLUE+ subset of datasets for the merged and aligned versions of MediPhi as well as other medical LLMs. ACI refers to ACI-Bench.

		RRS QA	MedicationQA	MEDEC	ACI	SDoH	ICD10CM
Baseline	<i>Phi-3.5-mini-instruct</i>	41.2	11.2	14.8	42.3	35.1	49.3
SLERP	DataMix	43.3	10.8	18.8	42.7	36.2	49.5
	PubMed	44.1	10.3	22.2	42.7	35.8	49.5
	Clinical	52.1	12.0	34.5	43.9	35.8	49.6
	MedWiki	46.7	12.2	28.8	44.7	43.6	50.2
	MedCode	45.6	12.0	18.1	39.0	24.8	68.7
	Guideline	48.9	11.9	28.3	44.7	41.0	49.8
BreadCrumbs	MediPhi	44.5	11.3	29.1	44.3	39.7	55.5
MediFlow	MediPhi-SFT	60.8	18.8	35.0	43.4	54.5	54.9
	MediPhi-Instruct	61.6	19.3	34.4	43.5	56.7	54.9
Other Medical LLMs	Mistral-7B-Instruct-v0.1	50.4	22.7	21.5	50.4	40.2	27.6
	BioMistral-7B-DARE	49.6	22.3	23.1	43.3	45.9	25.1
	Meta-Llama-3-8B-Instruct	55.8	26.1	46.5	50.2	63.1	25.7
	Llama3-Med42-8B	54.1	25.7	35.4	56.5	53.9	53.4

Table 13: Performances of MediPhi models on multiple-choice question-answering medical benchmarks.

	MedQA	MedMCQA	PubMedQA	MMLU-med	AVG
<i>Phi-3.5-mini-instruct</i>	0.486	0.554	0.768	0.715	0.631
PubMed	0.467	0.549	0.774	0.724	0.629
Clinical	0.500	0.551	0.772	0.727	0.638
Medical	0.519	0.535	0.740	0.701	0.624
MedWiki	0.478	0.548	0.768	0.719	0.628
MedCode	0.531	0.532	0.758	0.700	0.630
Guideline	0.459	0.556	0.770	0.724	0.627
MediPhi	0.491	0.559	0.766	0.720	0.634
MediPhi-SFT	0.536	0.552	0.766	0.716	0.642
MediPhi-Instruct	0.548	0.555	0.764	0.714	0.645