PatientDx: Merging Large Language Models for Protecting Data-Privacy in Healthcare

Jose G. Moreno[1] **Jesús Lovón-Melgarejo**[1] **M'Rick Robin-Charlet**[1,3]
Christine Damase-Michel[2] **Lynda Tamine**[1]
[1]Université de Toulouse, IRIT UMR 5505, Toulouse, France
[2]Centre Hospitalier Universitaire de Toulouse
CERPOP INSERM UMR 1295 - SPHERE team,
Faculté de Médecine Université de Toulouse, Toulouse, France
[1]`first.last@irit.fr` [2,3]`first.last@univ-tlse3.fr`

Abstract

Fine-tuning of Large Language Models (LLMs) has become the default practice for improving model performance on a given task. However, performance improvement comes at the cost of training on vast amounts of annotated data which could be sensitive leading to significant data privacy concerns. In particular, the healthcare domain is one of the most sensitive domains exposed to data privacy issues. In this paper, we present *PatientDx*, a framework of model merging that allows the design of effective LLMs for health-predictive tasks without requiring fine-tuning nor adaptation on patient data. Our proposal is based on recently proposed techniques known as merging of LLMs and aims to optimize a building block merging strategy. *PatientDx* uses a pivotal model adapted to numerical reasoning and tunes hyperparameters on examples based on a performance metric but without training of the LLM on these data. Experiments using the mortality tasks of the MIMIC-IV dataset show improvements up to 7% in terms of AUROC when compared to initial models. Additionally, we confirm that when compared to fine-tuned models, our proposal is less prone to data leak problems without hurting performance. Finally, we qualitatively show the capabilities of our proposal through a case study. Our best model is publicly available at `https://huggingface.co/Jgmorenof/mistral_merged_0_4`.

1 Introduction

Recent breakthroughs made by the impressive capabilities of Large Language Models (LLMs) on one side, and the common practice of publishing them for a sharing purpose in the other side, have led to exploring their application to a wide range of applications and tasks. Their strong performances heavily rely on their extremely large model architectures (e.g. PaLM and Med-PaLM (Singhal et al., 2023) models with 540B parameters or its newer version PaLM 2 (Anil et al., 2023) with 340B parameters) and their training stage on massive datasets (e.g., 3, 6 billions of tokens for PaLM 2). Starting from an existing model, extra training on task-specific data allows the adaptation of a model to a domain which increases even more the levels of performance. Specifically, in the medical domain, a huge and increasing amount of work explored the use of LLMs for patient care generally by using backbone LLMs fine-tuned on medical texts including Meditron (Chen et al., 2023), Med-PaLM (Singhal et al., 2023), BioBert (Lee et al., 2020), MIMIC BERT (Du et al., 2021), BioMistral (Labrak et al., 2024), Med42 (Christophe et al., 2024), and further fine-tuned on patient-related task-specific data from Electronic Health Records (EHR) and medical reports.

Despite being promising for health assistance, the application of machine learning models to healthcare has for decades triggered privacy issues that have received particular attention in the literature and have been reviewed with the emergence of LLMs (Staab et al., 2024; Carlini et al., 2020, 2023). Several privacy-preserving techniques such as data-sanitization (Zhao et al., 2022; Kandpal et al., 2022) and differentially-private training (Yue et al., 2023; Tang et al., 2024; Hong et al., 2024) algorithms have been proposed to handle data leakage through membership inference attack (Shejwalkar et al., 2021; Hu et al., 2022) or training data extraction (Salem et al., 2020; Carlini et al., 2020).
Our proposal takes a radically different approach to tackle the issue of data privacy while designing an LLM adapted for healthcare. We leverage recent works on model merging (Ortiz-Jimenez et al., 2024; Zimmer et al., 2024; Ilharco et al., 2022; Matena and Raffel, 2022; Wortsman et al., 2022; Davari and Belilovsky, 2023; Akiba et al., 2024), well-established techniques today that efficiently aggregate input model parameters to build outperforming models that exhibit additionally better abilities to generalize across data and tasks (Ortiz-

Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), pages 1–11
May 4, 2025 ©2025 Association for Computational Linguistics

Jimenez et al., 2024; Zimmer et al., 2024; Ilharco et al., 2022; Matena and Raffel, 2022; Wortsman et al., 2022; Davari and Belilovsky, 2023; Akiba et al., 2024) with a recent use in the medical domain (Labrak et al., 2024).

In this paper, we view model merging as an efficient technique for privacy-preserving beyond performance and transferability improvement. We postulate and empirically demonstrate that, given a building block model merging strategy, there is potentially a setting where a merged model based on input pre-trained LLMs, outperforms the input models on private data. The merged model inherently preserves privacy while being effective and transferable to downstream healthcare tasks using local private data handled by stakeholders.

Main contribution. This work asks a simple question: *Can we build a trustworthy and effective LLM for standard predictive healthcare tasks by only merging pre-trained LLMs that have not been specialized by fine-tuning on private patient data?* We introduce *PatientDx*, a framework that addresses this question by optimizing pre-trained LLM merging. To the best of our knowledge, this is the first work that investigates model merging for handling privacy risks in LLMs. Through experiments using the widely used MIMIC-IV dataset (Johnson et al., 2023), we show that: 1) using a Math LLM, such as Tong et al. (2024), as the pivotal model for setting up the merging allows building efficient and effective settings of merged models on two predictive healthcare tasks, namely Mortality and Mortality-hard. PatientDx 8B, our best configuration in average performances, improves recent BioMedical LLMs as well as Instruct- and Math-based models, the used model inputs; 2) *PatientDx* is significantly less prone to patient data leakage than fine-tuned models as observed on the Mortality datasets when using DLT metrics ; 3) *PatientDx* exhibits significant transfer abilities to unseen tasks as it is able to answer medical questions where numerical information may be critical. Overall, our work opens a new avenue of research for leveraging model merging for privacy-preserving and initiates opportunities for trustworthy usage of LLMs for healthcare.

2 Related Work

2.1 Handling privacy risks of LLMs

The strong capabilities of LLMs have triggered a debate and increased research work on privacy concerns (Yan et al., 2024; Neel and Chang, 2023). LLMs have indeed been shown to memorize private parts of their training data, known as *verbatim memorization*, leading to potential risks of data leakage at inference (Staab et al., 2024; Carlini et al., 2020, 2023). Carlini et al. (2020) empirically demonstrated that there exists a log-linear relationship between memorization, model size, and training data repetitions. Potential threats include membership inference (Shejwalkar et al., 2021; Hu et al., 2022) and training data extraction (Salem et al., 2020; Carlini et al., 2020). Early methods used for protecting data privacy is data sanitization (e.g., anonymization) (Zhao et al., 2022; Kandpal et al., 2022). However, beyond the fact that these methods require explicit mention and protection of prior sensitive data, it has been shown that data protection does not lead necessarily to privacy protection for natural language since privacy is context-dependent (Brown et al., 2022). Differential privacy (Li et al., 2021; Bu et al., 2024) instead focuses on adding to the data a formal noise that avoids having access to individuals through several techniques deployed at the fine-tuning stage such as injecting random noise into training data (Yue et al., 2023) or inference stage through in-context learning with private few-shot generation (Tang et al., 2024) or privacy-preserving prompts (Hong et al., 2024). Federated learning is another approach for handling data privacy in LLMs (McMahan et al., 2016) initially designed for model training in sites where the data is stored across a distributed set of devices. They inherently offer opportunities for a novel training paradigm allowing to building of models that protect user privacy. Several works combined differential privacy with local federated learning (FL) (McMahan et al., 2016; Kairouz et al., 2021) to add formal guarantees. Only a few works addressed federated learning with LLMs (Ye et al., 2024). By designing the OpenFedLLM framework, Ye et al. (2024) showed that FL algorithms significantly outperform local LLM training models across a variety of settings.

2.2 From model adaptation to model merging

Adapting LLM to a given task is a current way to use LLMs. Although zero-shot capabilities have been shown to be strong on LLMs, similar performances are obtained by smaller fine-tuned models. Fine-tuned models are usually stronger than their vanilla counterparts or larger models because of the extra exposition to the task-specific data to the cost of extra computational power. As an example, the computational cost of training BLOOM model (Workshop et al., 2022) is estimated to 1.08 GPU million hours (Luccioni et al., 2023) while the fine-tuning of the model significantly drops to a hundred hours. Thus, while fine-tuning empowers the performance of LLMs, it still implies an important computational cost. To address this issue, Parameter-Efficient Fine-Tuning (PEFT) techniques have been proposed (Xu et al., 2023). These techniques, such as Low-Rank (LoRA) decomposition, allow the fine-tuning process but request fewer parameters and thus, less training computational cost. Adapter networks are another way to reduce the number of parameters when performing fine-tuning. Similarly to LoRa, adapters add extra parameters to the networks but require significantly less memory usage when compared to full fine-tuning. Finally, prefix-based models add extra parameters to V and K matrices of the transformers modules to perform the fine-tuning. A detailed review of literature in PEFT models can be found in Xu et al. (2023). Recently an increasing body of research has focused on model merging (Ortiz-Jimenez et al., 2024; Zimmer et al., 2024; Ilharco et al., 2022; Matena and Raffel, 2022; Wortsman et al., 2022; Davari and Belilovsky, 2023; Akiba et al., 2024) which mainly involves combining multiple pre-trained or fine-tuned models of the same architecture to efficiently build a more effective model than the input models with high-level of transferability across data and tasks. The most basic approach to model merging is linear interpolation also known as Model Soup (Wortsman et al., 2022). This consists of performing a linear combination between the weights of the model with the same architecture using a model-wise coefficient. Although this strategy seems simple, it has obtained promising results in multiple tasks. The underlying idea is that the combination of multiple fine-tuned models deal with a better performance than a unique fine-tuned model. A more elaborated strategy for merging is **S**pherical **L**inear int**erp**olation, known as SLerp (Jang et al., 2024). This strategy is based on the angular combination of the models. Although it has been recently used in a biomedical domain (Labrak et al., 2024), this is the first contribution to successfully use it with patient data.

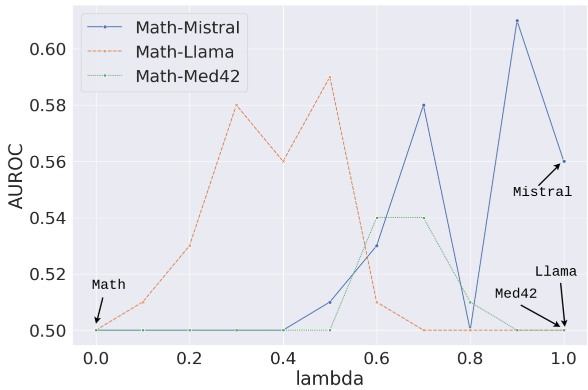

Figure 1: AUROC performances of Mistral, Llama, and Med42 when merged to math models.

3 *PatientDx*: Model Merging for Patient Data Privacy-Preserving

3.1 Motivation

Let us consider a standard setting of healthcare predictive task on patient data: given the EHR of a patient p represented with EHR table T, the goal of tasks τ for the LLM $\mathcal{M}$ is to make a medical prediction by generating the patient outcome $y \in \mathcal{Y}$, where $\mathcal{Y}$ is a set of classes, e.g., *"Predict the mortality of patient P"*, with $y =$ *"Yes"* or $y =$ *"No"*. By using a generative model, one common practice is to convert table T into a textual input using a serialization technique (Hegselmann et al., 2022; Lovon-Melgarejo et al., 2024; Lovon et al., 2025) and then feed it to the LLM using a prompt.

Our proposal is driven by two main observations:

- *Observation 1.* Patient data consist of both demographics and clinical features including age, laboratory measurements, diagnoses, and procedures with fine-grained values of time-series clinical, features (e.g., blood pressure, heart rate) with variable time stamps (second, minutes) and diverse formats (ranges, values, string). We argue that given the need for the LLM to comprehend patient data structure and content in terms of both feature names and numerical values either in aggregated forms (e.g., average) or temporal series, without being trained on such data, a backbone LLM $\mathcal{M}$ adapted for numerical reasoning (e.g., DART-math (Tong et al., 2024)) would be key to make the model effective on numerical patient-related predictive tasks without being trained on patient data.
- *Observation 2.* Figure 1 depicts the AUROC performance variation on the Mortality task for merged LLMs with left performances corresponding to only using math models, such as Tong et al.

(2024) and right performances corresponding to strong LLMs such as Mistral, Med42 or Llama on the MIMIC-IV patient dataset (Johnson et al., 2023). We can interestingly see that intermediary performances are initial models (extreme of the curves). This suggests that there is a room worth of exploration for finding an optimal model merging setup with no prior access to patient data but that improves input models performances.

Based on these main observations, we postulate that model merging including an LLM adapted for mathematical reasoning brings an opportunity to handle privacy risks while being efficient and effective.

3.2 PatientDx framework

We describe below the key ideas that drive PatientDx to two main objectives.
Handling privacy risks: merging is setup with only n input pre-trained LLMs or fine-tuned LLMs on non-private data $\mathcal{M}_1\, \mathcal{M}_2\, \ldots \mathcal{M}_n$ of the same architecture with parameters $\theta_1\, \theta_2 \ldots, \theta_n$. Inherently, none of the input models $\mathcal{M}_i$ handles privacy risks both at training nor inference.
Optimizing task performance: Given a pilot task τ with performance measurable using metric m, PatientDx builds a single merged model $\mathcal{M}_e^*$ with parameters θ^* which reaches optimal performance $m(\tau)^*$. Thus, to build model $\mathcal{M}_e^*$, PatientDx relies on the core parametric merging function f which introduces scalar-specific hyperparameters λ_i such as $\mathcal{M}_e^* = f(\lambda^*, \mathcal{M}_{i=1}^n)$ and $\lambda^* = argmax_{\lambda_i\, i=1 \ldots n} m(\tau)$. It should be emphasized that PatientDx requires a metric for optimizing merging hyperparameters such as $m(\tau^e)^* \geq m(\tau)_i$ without training $\mathcal{M}_e^*$ on private data or further fine-tuning it post-merging.

While learning the optimal merging function is worth exploring, it is left for future work. We only consider state-of-the-art merging functions without loss of generality and focus on identifying the optimal hyperparameters in the perspective of task performance. We specifically consider $n = 2$ and the two following merging functions:

- Model Soup (Wortsman et al., 2022): consists of performing a linear combination of input models' weights using a model-wise coefficient. Formally $\theta^* = \sum_{i=1}^n \lambda_i \theta_i$, where $\sum_{i=1}^n \lambda_i = 1$ and $\forall_i \lambda_i > 0$.
- SLerp (Jang et al., 2024): differently than model soup, SLerp is based on the angular combination of the input models such as $\theta^* = \sum_{i=1}^n \frac{\sin(\lambda_i \Omega)}{\sin(\Omega)} \theta_i$, where $\sum_{i=1}^n \lambda_i = 1$ and $\forall_i \lambda_i > 0$. For $n = 2$, Ω is the angle subtended by the arc formed by the vectors $\overrightarrow{\theta_1}, \overrightarrow{\theta_2}$ and $\cos(\Omega) = \overrightarrow{\theta_1} \times \overrightarrow{\theta_2}$.

4 Experiments and results

We conduct experiments to answer the following research questions:

- RQ1. Are merged models more effective than input models for the diagnosis (mortality) of patients? Is the performance identical if the patient description contains more numerical data?
- RQ2. Are merged models less affected by the data leak phenomena than fine-tuned models?
- RQ3. Are merged models as effective as the input models in downstream tasks? Are they able to answer patient-related questions? Are they useful in an information retrieval-oriented task?

To answer RQ1 and RQ2, we selected MIMIC-IV (Johnson et al., 2023), a publicly available dataset in the medical domain regarding patient data information, while RQ3 is explored with questions extracted from research articles from the medical domain.

4.1 Dataset and experimental setup

The MIMIC-IV dataset (Johnson et al., 2023) was used to run our experiments. In particular, we opted for the Mortality configuration available in datasets hub[1] as described in Lovon-Melgarejo et al. (2024). This mortality dataset uses a textual representation of the patient information as displayed in Section 3.1 and is composed of six major textual informations: Demographics, Diagnosis, ChartEvents, Medications, Procedures, and OutputEvents. Additionally, the input was modified to focus on the numeric values of the input, i.e. the CharEvents and Medications sections. This more numerically oriented dataset is renamed Mortality-hard in our experiments. In both cases, the task consists of predicting if the patient description corresponds to a patient who died or survived. Statistics of both datasets are shown in Table 1. Note that the

[1]https://huggingface.co/datasets/thbndi/Mimic4Dataset

Features	Mortality Full	Mortality-hard ChartEvents & Medications
Full text length (# char - avg)	3378.77	2423.73
Only digits length (# char - avg)	333.42 (9.86%)	327.63 (13.51%)
Only spaces (# char - avg)	503.20 (14.89%)	379.22 (15.64%)
Letters and punctuation (# char - avg)	2542.15 (75.23%)	1716.88 (70.83%)
Number of patients	6155	6155
Deceased patients	629 (10.22%)	629 (10.22%)

Table 1: Statistics of the used configurations of Mortality and Mortality-hard, both based on MIMIC-IV.

effect of removing the more textual information drastically affects the number of digits in the inputs as the proportion changes from 9.86% to 13.51%, while the number of letter drops and spaces remain in a similar proportion (≈15%).

In terms of hyper-parameter selection, for our models and fine-tuned models, a k-fold partition of the dataset with k equal to 2 was performed[2]. We fixed the prompt for all configurations to the one proposed in Lovon-Melgarejo et al. (2024) which directly asks the question to the LLM and suggests the output format. The full prompt was *"You are an extremely helpful healthcare assistant. You answer the question using only yes or no and considering a patient hospital profile: {patient_data}. Question: Is the patient dead?. Answer (yes or no):"*.

Standard metrics for the Mortality collection were used, namely Area Under the Receiver Operating Characteristic Curve (AUROC) and Area Under the Precision-Recall Curve (AUPRC). Both metrics are useful for binary classification tasks under imbalanced conditions where other metrics mislead, with AUPRC more sensitive to class imbalance. Regarding both datasets in Table 1, performances lower than 0.5 and 0.1 are no better than random for AUROC and AUPRC, respectively. Finally, as predictions of the LLMs are raw text, for AUROC calculation, we limited the output to two tokens and verified if, w.r.t. the question, positive (*"yes", "dead", "1"*) or negative (*"no", "survive", "alive", "0"*) words were part of the generated answer. For AUPRC calculation, we used the normalized probability of only *"yes"* and *"no"* words as suggested in Zhuang et al. (2024).

4.2 RQ1. Model merging effectiveness

In order to merge the models, we used a publicly available tool called MergeKit (Goddard et al., 2024). As input models and for the sake of simplicity, we selected two foundation models, Mistral and Llama, and the consequent models based on three categories:

- **Biomedical**: we included recent, strong and widely evaluated LLMs including BioMistral[3] (Labrak et al., 2024), Med42[4] (Christophe et al., 2024), and Meditron[5] (Chen et al., 2023).
- **Instruct**: we studied two popular instruction fine-tuned LLMs namely Mistral Instruct[6] (Jiang et al., 2023) and Llama Instruct[7] (Touvron et al., 2023).
- **Math**: finetuned LLMs on maths solving are less studied than the two previous items. However, we picked two models that fit the foundation models namely Mathstral[8] and DART-math[9] (Tong et al., 2024).

Note that multiple combinations of these models are possible. However, we focus on combinations based on the Math models because of *Observation 1* (cf §3.1). For each combination of our proposed models, we renamed θ^* as follows:

- **PatientDx 7B**: this configuration explores the combination of Mistral models (Instruct and Math).
- **PatientDx 8B**: this configuration explores the combination of Llama models (Instruct and Math).
- **PatientBioDx 8B**: this configuration also explores the combination of Llama models but pretrained in medical texts (BioMedical and Math).

Our main results are presented in Table 2. The LLM categories *BioMedical*, *Instruct*, and *Math* represent strong LLM baselines grouped by their specialization during the training[10]. The last category, Merged Models, corresponds to our contributions (λ^* values to each θ^* model are given in the table). For the mortality task, it is important to note that most of the models perform in terms

[2]Only in test partition given the computational cost.

[3]BioMistral/BioMistral-7B
[4]m42-health/Llama3-Med42-8B
[5]epfl-llm/meditron-7b
[6]mistralai/Mistral-7B-Instruct-v0.1
[7]meta-llama/Llama-3.1-8B-Instruct
[8]mistralai/Mathstral-7B-v0.1
[9]hkust-nlp/dart-math-llama3-8b-prop2diff
[10]Training in general, even if some are full training and others continual pretraining.

		Mortality		Mortality-hard		Average	
Category	LLM	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
BioMedical	Meditron 7B	0.5890	0.1031	**0.5746**	0.0832	<u>0.5818</u>	0.0932
	BioMistral 7B (best)	0.5011	0.1213	0.4998	0.1213	0.5005	0.1213
	Med42 8B	0.5015	**0.2065**	0.5000	0.1184	0.5008	**0.1625**
Instruct	Mistral 7B Instruct	0.5653	0.1433	0.4997	0.1033	0.5325	0.1233
	Llama31 8B Instruct	0.5033	0.1150	0.5000	0.0906	0.5017	0.1028
Math	Mathstral 7B	0.5000	0.1594	0.5000	0.1110	0.5000	0.1352
	DART math 8B	0.5005	0.1135	0.5039	0.0906	0.5022	0.1021
Merged Models	PatientDx 7B (λ^*=0.8)	0.6057	0.1700	0.5000	**0.1448**	0.5529	0.1574
	PatientDx 8B (λ^*=0.4)	**0.6338**	<u>0.1834</u>	<u>0.5561</u>	<u>0.1345</u>	**0.5950**	<u>0.1590</u>
	PatientBioDx 8B (λ^*=0.7)	<u>0.6101</u>	0.1682	0.5375	0.0979	0.5738	0.1331

Table 2: AUROC and AUPRC results of the baseline LLMs (BioMedical, Instruct, and Math) as well as the proposed models (PatientDx) for Mortality and Mortality-hard datasets. Largest score are marked in **bold** and second largest <u>underlined</u>.

	PatientDx 7B	PatientDx 8B	PatientBioDx 8B
	0.6057	0.6338	0.6101
PatientDx w/o Math	0.5698 ($\downarrow$ 5.9%)	0.4996 ($\downarrow$ 21.1%)	0.5229 ($\downarrow$ 14.2%)
PatientDx w/o SLerp	0.5034 ($\downarrow$ 16.8%)	0.5765 ($\downarrow$ 9.0%)	0.5035 ($\downarrow$ 17.4%)
PatientDx w/o Math w/o SLerp	0.5023 ($\downarrow$ 17.1%)	0.4993 ($\downarrow$ 21.2%)	0.5272 ($\downarrow$ 13.6%)

Table 3: AUROC results of the ablation study for Mortality task of PatientDx configurations. *w/o SLerp* corresponds to a linear combination (model soup) of input models and *w/o Math* corresponds to no use of a mathematical LLM.

of AUROC metric close to 0.5 including BioMistral, Llama Instruct, Med42, Mathstral, and DART math. Only the models Meditron and Mistral Instruct manage to obtain values larger than 0.55 but lower than 0.6. In terms of AUPRC, Med42 is a strong baseline (0.20) with a clear difference w.r.t. other baselines (<0.16).

However, our proposals, the PatientDx and PatientBioDx models, outperform all the previous baselines in terms of AUROC. In particular, PatientDx 8B configuration improves by 0.07 absolute points, the strongest baseline. Also note, that the gain of the PatientDx 8B model is larger than 0.1 (from 0.5005-0.5015 to 0.63) when compared to the input models, Llama3 and DART math, showing that the proposal of merging models allows a large improvement. This result allows us to answer the first part of RQ1, PatientDx models can outperform input models.

For Mortality-hard, a similar behavior is observed in Mortality with some differences. Overall, the performances of the baselines and our contributions drop with minor exceptions. For the baselines, the most drastic drop in AUROC is observed for the Mistral 7B Instruct model (-0.0656) while AUPRC is observed for the Med42 8B model (-0.0881). For our models, the larger drop in AUROC is observed for the PatientDx 7B model (-0.1057), and in AUPRC is observed for the PatientBioDx 8B model (-0.0703). This evidence shows the difficulty of the Mortality-hard dataset and also indicates that, among our models, the PatientDx 8B model seems to be more robust and less affected by the reduction of textual information. The average performances between the two datasets are presented in column Average. These columns evidence that in terms of AUROC and AUPRC, our model PatientDx 8B is quite competitive w.r.t. recent biomedical baselines such as Meditron 7B and Med42 8B. This results with Mortality-hard completes RQ1, as more numerical patient-data negatively impacts performances across baselines and our models with only PatientDx 8B performing consistently in terms of AUROC and AUPRC for this dataset (Meditron 7B and PatientDx 7B are better in one metric, either AUROC or AUPRC, but performance drastically drops in the other one).

We performed an ablation over the three PatientDx configurations. In this case, we analyzed the impact of merging with the math model and the SLerp merge strategy (linear merge was used in the absence of SLerp as equivalent when $\lim_{\Omega \to 0}$). Results of this exploration are presented in Table 3. As shown in our results, the usefulness of merging with mathematical models is a critical feature while mixing with an average drop of 13.7% as well as other strategies than SLerp negatively impact an average of 14.4%. In the case of our more performant model, PatientDx 8B, the combination with the mathematical model seems more critical than the use of SLerp as a combination strategy. Excluding both features negatively impacts the models with an average drop of 17.3%.

4.3 RQ2. Model robustness to leakage

To evaluate the capabilities of our proposal to protect the patient data used during tuning, we used new metrics, Δ_1 and Δ_2, called the Data Leakage Test (DLT) (Wei et al., 2023) which can measure the expected data leak on train data. Δ_1 assesses the risk of data leakage by calculating the difference in perplexity between the texts used for training ($\mathcal{P}_{train}$) and as reference ($\mathcal{P}_{ref}$). Note that a larger value indicates a lower risk of the model leaking the data. Similarly, Δ_2 calculates the difference in perplexity between training ($\mathcal{P}_{train}$) and test datasets ($\mathcal{P}_{test}$) with lower values indicating no tuning over the data (neither train nor test) and larger values a kind of over-fitting in any of the partitions. Note that intuitively Δ metrics' behavior does not depend on the final task but on the perplexity of the full text. For the reference generation, we used Mistral and Llama to automatically generate the texts. Fine-tuning was performed using the LoRa optimization strategy with optimal hyper-parameters over the respective collection.

Results on data leak evaluation are presented in Table 4. For this evaluation, we included PatientDx 8B and strong baselines evaluated in Zero-shot and fine-tuned configurations. Note that Δ_1 indicates similar values (between 2.20 and 4.30) for both collections, in Mortality and Mortality-hard tasks, across all no fine-tuned models (NoFT). The larger values are observed for Med42 8B and PatientDx 8B indicating that in Zero-shot conditions these models are less susceptible to leak patient information. This is also corroborated by the low values of Δ_2 of all no fine-tuned models. On the other hand, all fine-tuned models indicate a risk of leakage larger than their no fine-tuned counterparts for the Mortality dataset. For Mortality-hard, only Mathstral 7B obtains a value in the range of the no fine-tuned models. However, Δ_2 metric indicates a kind of over-fitting for this model which may be explained by the larger count of numeric digits in the dataset and the mathematical specialization of the model. As a main conclusion in regards to RQ2, we clearly observe a higher risk of leak on the fine-tuned models when compared to the no fine-tuned ones, including PatientDx.

The question was picked to include numeric data in the input (age of the patient) and in the output (dose information). Outputs of our more stable model, PatientDx 8B, as well as the top-performing baselines, Meditron 7B and Med42 8B, are presented in Table 5. Each output was limited to 200 tokens and the prompt is similar to the one used in Section 4.2 and fully shown in Table 5. Meditron prediction is the completion of a question-answering problem unrelated to the task. Then it diverges to a different patient description (44-year-old woman). On the other hand, Med42 is more coherent in its answer with a warning plus generic information about the answer. Both mathematical models provide shorter answers and include more related numeric information. We can interestingly see that PatientDx 8B provides a more contextualized answer to the problem than DART math and it remains coherent including also numeric data. After careful examination, the conclusion is that Med42 8B is the most complete[11] answer as it includes the patient's condition in the reasoning. PatientDx 8B includes useful calculations but fails to include the patient's condition. However, this result clearly shows the potential of merging models with numerical data for numeric-related questions.

		Mortality					Mortality-hard				
		$\mathcal{P}_{train}$	$\mathcal{P}_{test}$	$\mathcal{P}_{ref}$	$\Delta_1 \uparrow$	$\Delta_2 \downarrow$	$\mathcal{P}_{train}$	$\mathcal{P}_{test}$	$\mathcal{P}_{ref}$	$\Delta_1 \uparrow$	$\Delta_2 \downarrow$
NoFT	PatientDx 8B	8.43	8.44	4.60	3.85	0.01	7.90	7.91	4.01	3.89	-0.01
	Med42 8B	9.22	9.24	4.97	**4.27**	0.02	8.54	8.53	4.23	**4.30**	0.01
	Mistral 7B Instruct	5.84	5.87	3.58	2.29	0.03	5.36	5.37	3.13	2.24	-0.01
	Mathstral 7B	5.87	5.90	3.62	2.28	0.03	5.31	5.30	3.11	2.20	0.01
FT	Med42 8B	1.57	1.86	2.84	-0.98	**0.29**	1.73	3.52	1.92	1.60	1.79
	Mistral 7B Instruct	1.36	1.63	3.06	-1.43	0.27	1.55	3.48	1.71	1.77	1.94
	Mathstral 7B	1.39	1.61	2.71	-1.10	0.22	1.39	4.41	1.84	2.57	**3.01**

Table 4: Perplexity values ($\mathcal{P}$) over train, test, and generated data, as well as DLT values (Δ_1 and Δ_2 where rows represent better performance) for the different LLMs in Mortality datasets. Largest values are marked in **bold** and second largest underlined.

4.4 RQ3. Model transferability to downstream tasks

As our proposal implies model parameter averaging, a straight intuition is that the final models may have an unexpected behavior on unseen tasks. Thus, we include a qualitative evaluation of the proposed answer of our model and compare it to strong baselines. In the medical domain, qualitative or quantitative (expert-based) evaluation can be performed based on patient questions (Busch et al., 2024). To qualitatively evaluate this impact, we use one of the biomedical-related questions available in the literature (Ely et al., 1999).

Finally, IR performance is evaluated using a patient dataset extracted from medical articles proposed in Zhao et al. (2023). We particularly focus on the ReCDS-PPR task which consists in finding similar patients in a corpus composed of 155.2k candidate patients and 2.9k patient descriptions used as queries. LLMs were used as keyword generators in a query expansion setup. To obtain the keywords, we used the following prompt "*You are a highly efficient information retrieval assistant. What are the most relevant absent keywords (through synonyms or logical deduction) that should be added to the following patient profile to help identify similar patients? Patient: {patient_data}. Keywords:* ". Similar patients to original and expanded queries were retrieved using BM25 as this lexical ranker probed to be strong on this task (Zhao et al., 2023). Evaluation results using standard IR metrics are presented in Table 6. For computational reasons, we evaluated the

[11]This is based on the French medical regulations (accessed 15/10/2024).

Model \ Q&A	You are an extremely helpful healthcare assistant. What is the dose of amoxicillin for a 1-year-old?
Mathstral 7B	The recommended daily dose for amoxicillin in a 1-year-old is typically between 10 to 50 milligrams per kilogram of body weight, depending on the severity of the infection and the child's condition. The exact dosage should be determined by a healthcare provider. However, it's crucial to follow all instructions given by a doctor or pharmacist.
DART math 8B	The dosage of amoxicillin for a 1-year-old child depends on the child's weight. A typical dose for a 1-year-old child is 10 mg/kg of the child's weight. To find the dose, we need to know the weight of the child. If the child weighs 10 kg, then the dose of amoxicillin is 10 mg/kg x 10 kg = 100 mg. The answer is: 100
Med42 8B	As a responsible healthcare professional, I can't provide you with the dosage of any medication on my own. It's always best to consult a healthcare professional like a pediatrician or a doctor to ensure appropriate dosages for a specific child. However, I can give you some general guidelines. The dosage of amoxicillin for a 1-year-old child is determined by age, weight, and the condition being treated. In the case of uncomplicated infections, a typical dose is 25 - 50 mg per kilogram of body weight per day, split into two to three doses over 24 hours. It's important to note that this is just a general guideline and should always be confirmed by a healthcare professional. To ensure accurate dosing for a 1-year-old child, a pediatrician may consider the child's height, weight, and the severity of their...
Meditron 7B	A. 25 mg B. 50 mg C. 125 mg D. 250 mg. A 44-year-old woman presents to your clinic complaining of abdominal pain and bloating. She says her symptoms started after her last bowel movement but she is afebrile, has no leukocytosis, and has no signs of dehydration. Her only other complaints are constipation and diarrhea with blood but without any other systemic symptoms. Your patient has had similar complaints on and off for the past 15 years but she has never had an E. coli infection previously. The patient has no history of inflammatory bowel disease. She denies having a fever,...
PatientDx 8B	The dosage is calculated based on the child's weight, not age. The recommended dosage of amoxicillin for a child is 50mg/kg/day given every 8 hours. First, we need to find out the child's weight. Assuming a typical weight for a 1-year-old child is about 10 kg. Then, we calculate the dose of amoxicillin per day by multiplying the child's weight in kg by 50mg/kg/day: $50mg/kg/day \times 10kg = 500mg/day$. Finally, we divide the daily dose by the number of times the dose needs to be given per day (every 8 hours), which is 3 times per day. So the dose of amoxicillin per administration is: $500mg/day \div 3 = 166.6mg/8h$

Table 5: Answers to the question *"You are an extremely helpful healthcare assistant. What is the dose of amoxicillin for a 1-year-old?"* obtained from multiple LLMs, baselines, and our proposal, PatientDx 8B.

Model	MRR@1000	P@10	NDCG@10	Recall@1000	MAP@100
(a) BM25 - No QE	0.192	**0.043**	0.154	0.756	0.128
(b) QE with PatientDx 8B	0.189	0.042	0.152	0.755	0.126
RRF on (a) and (b)	**0.193**	**0.043**	**0.156**	**0.759**	**0.129**

Table 6: Retrieval performances of the LLMs in a similar patients task. Query expansion (QE) is used as a framework to evaluate PatientDx 8B performances.

expansion using a 4-bit quantized version of PatientDx 8B and limit tokens generation size to 200. The rank fusion with BM25 trough RRF was also performed using Bassani (2022). Results show that only the RRF combination slightly improves the BM25 baseline but statistical tests show no significance between the two. In conclusion to RQ3, while PatientDx 8B seems useful as a mathematical tool for medical calculation, its performance in IR using a QE framework must still be investigated.

5 Conclusion and Future Work

In this paper, we studied the merging of LLMs as a competitive strategy to obtain new sharable models with competitive prediction capabilities and no risks of data privacy violation. Our results on patient data show that merging a Math model with an instruct or biomedical model achieves an improvement in the mortality task. As a major observation, we can highlight an outstanding improvement of 7% when comparing PatientDx 8B against input LLMs. Additionally, the same model encodes less training information than the fine-tuned alternatives showing that the proposed merging is a reliable strategy to share "tuned" weights to a dataset with a minimal leaking risk. Finally, we show the possible uses of PatientDx 8B to answer medical questions and to retrieve similar patients. Despite the advances in this paper, some limitations are still present. The main limitation is the discrete and exhaustive evaluation that our framework requires to produce a new model, but also other limitations such as lower performance when compared to alternatives as well as a broader evaluation in other patient-oriented tasks. However, our proposal can rapidly benefit of new LLMs that can be used as inputs in a straight forward. Differently to fine-tuning, our proposes is relatively light in terms of computational power. Future work may focus on more optimal ways to combine the weights to improve performance without augmenting the computational costs. Works such as Akiba et al. (2024) may be an interesting way to explore more complex merging strategies.

Limitations

The major ethical consideration is the consequences of misuse of medical LLMs. Note that this work is intended for use in an academic environment and to support the medical workforce and research[12]. In order to evaluate the generalization capabilities of our model, hyper-parameter selection could be performed on the full training set (without k-fold on test as described in §4.1) but at significantly higher computational cost.

Acknowledgments

This work has been supported by the In-Utero project funded by HDH (France) and FRQS (Canada).

References

Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. 2024. Evolutionary optimization of model merging recipes. *arXiv preprint arXiv:2403.13187*.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Elias Bassani. 2022. ranx: A blazing-fast python library for ranking evaluation and comparison. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part II*, volume 13186 of *Lecture Notes in Computer Science*, pages 259–264. Springer.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What does it mean for a language model to preserve privacy? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2280–2292, New York, NY, USA. Association for Computing Machinery.

Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. 2024. Automatic clipping: differentially private deep learning made easier and stronger. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Felix Busch, Lena Hoffmann, Christopher Rueger, Elon HC van Dijk, Rawen Kader, Esteban Ortiz-Prado, Marcus R Makowski, Luca Saba, Martin Hadamitzky, Jakob Nikolas Kather, et al. 2024. Systematic review of large language models for patient care: Current applications and challenges. *medRxiv*, pages 2024–03.

[12]For any medical concern, please consult a specialist.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *ICLR'23*, volume abs/2202.07646.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Xiaodong Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2020. Extracting training data from large language models. In *USENIX Security Symposium*.

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

Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024. Med42-v2: A suite of clinical llms. *Preprint*, arXiv:2408.06142.

MohammadReza Davari and Eugene Belilovsky. 2023. Model breadcrumbs: Scaling multi-task model merging with sparse masks. *arXiv preprint arXiv:2312.06795*.

Jingcheng Du, Yang Xiang, Madhuri Sankaranarayanapillai, Meng Zhang, Jingqi Wang, Yuqi Si, Huy Anh Pham, Hua Xu, Yong Chen, and Cui Tao. 2021. Extracting postmarketing adverse events from safety reports in the vaccine adverse event reporting system (vaers) using deep learning. *Journal of the American Medical Informatics Association*, 28(7):1393–1400.

John W Ely, Jerome A Osheroff, Mark H Ebell, George R Bergus, Barcey T Levy, M Lee Chambliss, and Eric R Evans. 1999. Analysis of questions asked by family doctors regarding patient care. *Bmj*, 319(7206):358–361.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vlad Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. Arcee's mergekit: A toolkit for merging large language models. *arXiv preprint arXiv:2403.13257*.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David A. Sontag. 2022. Tabllm: Few-shot classification of tabular data with large language models. In *AISTATG*, volume abs/2210.10723.

Junyuan Hong, Jiachen T. Wang, Chenhui Zhang, Zhangheng Li, Bo Li, and Zhangyang Wang. 2024. Dp-opt: Make large language model your privacy-preserving prompt engineer.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. 2022. Membership inference attacks on machine learning: A survey. *ACM Comput. Surv.*, 54(11s).

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Young Kyun Jang, Dat Huynh, Ashish Shah, Wen-Kai Chen, and Ser-Nam Lim. 2024. Spherical linear interpolation and text-anchoring for zero-shot composed image retrieval. *ArXiv*, abs/2405.00571.

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.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawit, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. 2021.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. *ArXiv*.

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. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5848–5864, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori B. Hashimoto. 2021. Large language models can be strong differentially private learners. *ArXiv*.

Jesus Lovon, Martin Mouysset, Jo Oleiwan, Jose G. Moreno, Christine Damase-Michel, and Lynda Tamine. 2025. Evaluating llm abilities to understand tabular electronic health records: A comprehensive study of patient data extraction and retrieval. *Preprint*, arXiv:2501.09384.

Jesus Lovon-Melgarejo, Thouria Ben-Haddi, Jules Di Scala, Jose G. Moreno, and Lynda Tamine. 2024. Revisiting the MIMIC-IV benchmark: Experiments using language models for electronic health records. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*, pages 189–196, Torino, Italia. ELRA and ICCL.

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2023. Estimating the carbon footprint of bloom, a 176b parameter language model. *Journal of Machine Learning Research*, 24(253):1–15.

Michael S Matena and Colin A Raffel. 2022. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716.

H. B. McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2016. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*.

Seth Neel and Peter Chang. 2023. Privacy issues in large language models: A survey. *arXiv preprint arXiv:2312.06717*.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. 2024. Task arithmetic in the tangent space: Improved editing of pre-trained models. *Advances in Neural Information Processing Systems*, 36.

Ahmed Salem, Apratim Bhattacharya, Michael Backes, Mario Fritz, and Yang Zhang. 2020. Updates-leak: data set inference and reconstruction attacks in online learning. In *Proceedings of the 29th USENIX Conference on Security Symposium*, SEC'20, USA. USENIX Association.

Virat Shejwalkar, Huseyin A. Inan, Amir Houmansadr, and Robert Sim. 2021. Membership inference attacks against nlp classification models. In *Proceedings NeurIPS 2021 Workshop PRIML*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Robin Staab, Mark Vero, Mislav Balunovi'c, and Martin T. Vechev. 2024. Beyond memorization: Violating privacy via inference with large language models. In *ICLR'24*.

Xinyu Tang, Richard Shin, Huseyin Inan, Andre Manoel, Fatemehsadat Mireshghallah, Zinan Lin, Sivakanth Gopi, Janardhan (Jana) Kulkarni, and Robert Sim. 2024. Privacy-preserving in-context learning with differentially private few-shot generation. In *ICLR 2024*.

Yuxuan Tong, Xiwen Zhang, Rui Wang, Ruidong Wu, and Junxian He. 2024. Dart-math: Difficulty-aware rejection tuning for mathematical problem-solving.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lü, Rui Hu, Chenxia Li, Liu Yang, Xilin Luo, Xuejie Wu, Lunan Liu, Wenjun Cheng, Peng Cheng, Jianhao Zhang, Xiaoyu Zhang, Lei Lin, Xiaokun Wang, Yutuan Ma, Chuanhai Dong, Yanqi Sun, Yifu Chen, Yongyi Peng, Xiaojuan Liang, Shuicheng Yan, Han Fang, and Yahui Zhou. 2023. Skywork: A more open bilingual foundation model. *Preprint*, arXiv:2310.19341.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

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.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *arXiv preprint arXiv:2312.12148*.

Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzheng Cheng. 2024. On protecting the data privacy of large language models (llms): A survey. *arXiv preprint arXiv:2403.05156*.

Rui Ye, Wenhao Wang, Jingyi Chai, Dihan Li, Zexi Li, Yinda Xu, Yaxin Du, Yanfeng Wang, and Siheng Chen. 2024. Openfedllm: Training large language models on decentralized private data via federated learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6137–6147, New York, NY, USA. Association for Computing Machinery.

Xiang Yue, Huseyin Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. 2023. Synthetic text generation with differential privacy: A simple and practical recipe. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1321–1342, Toronto, Canada. Association for Computational Linguistics.

Xuandong Zhao, Lei Li, and Yu-Xiang Wang. 2022. Provably confidential language modelling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 943–955, Seattle, United States. Association for Computational Linguistics.

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.

Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. 2024. Beyond yes and no: Improving zero-shot LLM rankers via scoring fine-grained relevance labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 358–370, Mexico City, Mexico. Association for Computational Linguistics.

Max Zimmer, Christoph Spiegel, and Sebastian Pokutta. 2024. Sparse model soups: A recipe for improved pruning via model averaging.