PatientDx: 合并大型语言模型以保护医疗保健中的数据隐私

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

大型语言模型（LLM）的微调已经成为提高特定任务模型性能的默认方法。然而，性能的提升伴随着在大量标注数据上进行训练的代价，这些数据可能是敏感的，从而导致显著的数据隐私问题。尤其是，医疗保健领域是最容易受到数据隐私问题侵袭的领域之一。在本文中，我们提出了PatientDx，这是一种模型合并框架，允许设计出不需要微调或在患者数据上进行适配的健康预测任务的有效 LLM。我们的提议基于最近提出的 LLM 合并技术，并旨在优化一种构建模块合并策略。PatientDx 使用一个适应于数值推理的核心模型，并根据性能指标在实例上调整超参数，但不需要在这些数据上训练 LLM。在使用 MIMIC-IV 数据集的死亡率任务的实验中，与初始模型相比，在 AUROC 方面提高了最多 7%。此外，我们证实，与微调模型相比，我们的提议不容易出现数据泄漏问题，同时不会损害性能。最后，我们通过一个案例研究定性展示了我们提案的能力。我们最好的模型公开可在 https://huggingface.co/Jgmorenof/mistral_merged_0_4 获取。

一方面，大型语言模型（LLM）令人印象深刻的能力的最新突破，另一方面，为共享目的而普遍发布它们的做法，已导致探索其在广泛应用和任务中的应用。它们的强大性能在很大程度上依赖于其极其庞大的模型架构（例如，具有 5400 亿参数的 PaLM 和 Med-PaLM (Singhal et al., 2023) 模型或其更新版本 PaLM 2 (Anil et al., 2023) 具有 3400 亿参数）以及它们在海量数据集上的训练阶段（例如，PaLM 2 使用 3,6 亿个 token）。从现有模型开始，在特定任务的数据上进行额外训练，可以让模型适应特定领域，从而进一步提高性能水平。具体来说，在医学领域，大量且不断增加的研究探索了使用大型语言模型（LLMs）进行患者护理，通常使用在包括 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) 等在内的医学文本上微调的基础 LLMs，并进一步在来自电子健康记录（EHR）和医疗报告的与患者相关的特定任务数据上微调。尽管在健康辅助方面有很大前景，但几十年来，机器学习模型在医疗保健中的应用一直引发隐私问题，这些问题在文献中受到特别关注，并且随着大型语言模型的出现而被重新审视 (Staab et al., 2024; Carlini et al., 2020, 2023) 。为应对通过成员推断攻击 (Shejwalkar et al., 2021; Hu et al., 2022) 或训练数据提取 (Salem et al., 2020; Carlini et al., 2020) 造成的数据泄漏，提出了几种隐私保护技术，例如数据清理 (Zhao et al., 2022; Kandpal et al., 2022) 和差分隐私训练 (Yue et al., 2023; Tang et al., 2024; Hong et al., 2024) 算法。

我们的建议采用了一种根本不同的方法来解决在设计适用于医疗保健的大型语言模型时的数据隐私问题。我们利用了最近在模型合并 (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) 方面的工作，这些技术今天已被公认为能够有效地聚合输入模型参数以构建更强大的模型，这些模型在数据和任务之间显示出更好的泛化能力，并且最近在医学领域得到应用 (Labrak et al., 2024) 。

在本文中，我们将模型合并视为一种超越性能和可转移性提升的有效隐私保护技术。我们假设并通过实验证明，给定一个构建模块的模型合并策略，存在一种潜在的设置，其中基于输入预训练 LLMs 的合并模型在私有数据上优于输入模型。合并模型在固有地保护隐私的同时，能够使用由利益相关者处理的本地私有数据，效能强大且可转移到下游的医疗任务中。

主要贡献。这项工作提出了一个简单的问题：我们能否仅通过合并未通过对私人患者数据进行微调而专门化的预训练 LLM，来构建一个可信且有效的 LLM 用于标准预测性医疗任务？我们引入了 PatientDx，一个通过优化预训练 LLM 合并来解决这个问题的框架。据我

们所知，这是首个研究通过模型合并来处理 LLM 隐私风险的工作。通过使用广泛使用的 MIMIC-IV 数据集 (Johnson et al., 2023) 进行实验，我们显示：1) 使用一个数学 LLM，比如 Tong et al. (2024)，作为设置合并的关键模型，可以在两个预测性医疗任务——死亡率和死亡率困难上，构建高效且有效的合并模型设置。PatientDx 8B 是我们在平均性能上最好的配置，提高了最近的生物医学 LLM 以及指令和基于数学的模型所使用的模型输入；2) PatientDx 比微调模型在使用 DLT 指标观察死亡率数据集时显著减少了患者数据泄露的可能性；3) PatientDx 展现了显著的迁移能力，能够回答可能涉及关键数字信息的医疗问题。总的来说，我们的工作为利用模型合并进行隐私保护开辟了新的研究方向，并为可信的医疗 LLM 使用创造了机会。

1 相关工作

1.1 处理大型语言模型的隐私风险

LLM 的强大功能引发了关于隐私问题的争论和研究工作的增加 (Yan et al., 2024; Neel and Chang, 2023)。研究表明，LLM 确实能够记住其训练数据的私有部分，被称为逐字记忆，这可能导致在推理时的数据泄漏风险 (Staab et al., 2024; Carlini et al., 2020, 2023)。Carlini et al. (2020) 实证表明记忆、模型大小和训练数据重复之间存在对数线性关系。潜在威胁包括成员推断 (Shejwalkar et al., 2021; Hu et al., 2022) 和训练数据提取 (Salem et al., 2020; Carlini et al., 2020)。早期用于保护数据隐私的方法是数据清理（例如，匿名化）(Zhao et al., 2022; Kandpal et al., 2022)。然而，除了这些方法需要明确提及和保护先前的敏感数据之外，已经证明数据保护并不一定会导致自然语言的隐私保护，因为隐私是依赖于上下文的 (Brown et al., 2022)。差分隐私 (Li et al., 2021; Bu et al., 2024) 则专注于通过在微调阶段部署的几种技术，例如在训练数据中注入随机噪声 (Yue et al., 2023)，或通过私有小样本生成 (Tang et al., 2024) 或隐私保护提示 (Hong et al., 2024) 的情境学习，在推理阶段为数据添加形式噪声，以避免访问个人数据。联邦学习是处理 LLM 中数据隐私的另一种方法 (McMahan et al., 2016)，最初为那些数据分布存储在不同设备上的场所设计的模型训练。它们固有地为一种新的训练范式提供了机会，允许构建能保护用户隐私的模型。一些工作结合了差分隐私与本地联邦学习（FL）(McMahan et al., 2016; Kairouz et al., 2021) 来添加正式的保证。只有少数工作探讨了与 LLM 结合的联邦学习 (Ye et al., 2024)。通过设计 OpenFedLLM 框架，Ye et al. (2024) 表明在各种环境下，FL 算法显著优于本地 LLM 训练模型。适应 LLM 到特定任务是当前使用 LLM 的一种方式。尽管零样本能力在 LLM 上显示出强大的表现，但经过微调的小型模型也获得了类似的性能。微调模型通常比其基础版本或更大型的模型更强，因为在特定任务数据上有额外的暴露，代价是额外的计算能力。例如，训练 BLOOM 模型的计算成本估计为 1.08 百万 GPU 小时，而微调该模型则显著下降到数百小时。因此，尽管微调提升了 LLM 的性能，它仍然意味着重要的计算成本。为了解决这个问题，提出了参数高效微调（PEFT）技术。这些技术，例如低秩（LoRA）分解，允许进行微调过程，但需要更少的参数，因此也降低了训练计算成本。适配器网络是另一种在执行微调时减少参数数量的方法。类似 LoRa，适配器向网络中添加了额外的参数，但与完全微调相比所需的内存使用量明显减少。最后，基于前缀的模型向 transformers 模块的 V 和 K 矩阵添加额外的参数以执行微调。对 PEFT 模型的文献详细回顾可见于 Xu et al. (2023)。最近，越来越多的研究集中在模型合并 (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)，这主要涉及结合多种预训练或微调过的相同架构的模型，以高效构建一种比输入模型更有效的模型，同时具有高水平的数据和任务转移能力。最基本的模型合并方法是线性插值，也称为模型汤 (Wortsman et al., 2022)。这包括使用模型系数在具有相同架构的模型权重之间进行线性组合。尽管这种策略看似简单，但它在多项任务中取得了有希望的结果。其基本思路是，多种微调模型的组合比单一微调模型具有更好的性能。更复杂的合并策略是球面线性插值，简称为 SLerp (Jang et al., 2024)。这种策略基于模型的角度组合。虽然它最近在生物医学领域中应用 (Labrak et al., 2024)，但这是首次成功地将其应用于患者数据的贡献。

2 PatientDx：患者数据隐私保护模型合并

2.1 动机

让我们考虑一下在患者数据上的医疗预测任务的标准设置：给定患者 p 的电子健康记录，表示为 EHR 表 T，任务 τ 的目标对于 LLM $\mathcal{M}$ 是通过生成患者结果 $y \in \mathcal{Y}$ 来进行医学预测，其中 $\mathcal{Y}$ 是一组类别，例如，“预测患者 P 的死亡率”，用 $y =$ “是” 或 $y =$ “否”。通过使用生成模型，一个常见的做法是使用序列化技术 (Hegselmann et al., 2022; Lovon-Melgarejo et al.,

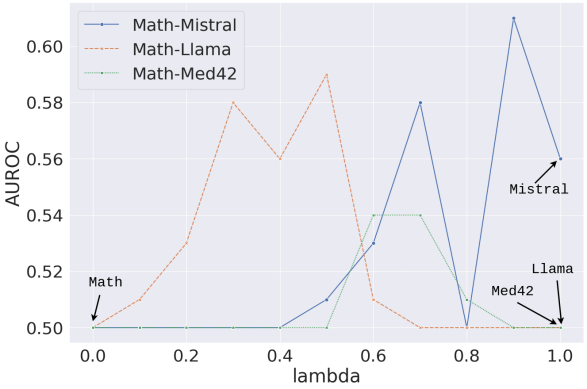

Figure 1: 将 Mistral、Llama 和 Med42 合并到数学模型时的 AUROC 表现。

2024; Lovon et al., 2025) 将表 T 转换成文本输入，然后使用提示将其输入到 LLM。

我们的提议是由两个主要观察驱动的：

- 观察 1。患者数据包括人口统计和临床特征，包括年龄、实验室测量、诊断和程序，其中时间序列临床特征（例如血压、心率）具有细粒度的值，时间戳（秒、分钟）和多种格式（范围、值、字符串）。我们认为，鉴于 LLM 需要在没有接受过此类数据训练的情况下，从特征名称和数值的角度理解患者数据结构和内容，无论是汇总形式（例如平均值）还是时间序列，一个针对数值推理调整的骨干 LLM $\mathcal{M}$ （例如 DART-math (Tong et al., 2024) ）将是使模型在与患者数据相关的数值预测任务中有效的关键。- 观察 2。图 1 展示了合并 LLMs 在死亡率任务上的 AUROC 性能变化，左侧性能对应于仅使用数学模型，如 Tong et al. (2024)，右侧性能对应于强大的 LLMs，如 Mistral、Med42 或 Llama 在 MIMIC-IV 患者数据集 (Johnson et al., 2023) 上的表现。我们可以有趣地看到，中间性能是初始模型（曲线的极端）。这表明，有一个探索空间可以在没有患者数据的情况下找到最佳模型合并设置，但可以改善输入模型的性能。

基于这些主要观察，我们假设包括一个适应于数学推理的 LLM 的模型合并提供了一个在高效和有效的同时处理隐私风险的机会。

2.2 PatientDx 框架

我们在下文中描述了驱动 PatientDx 实现两个主要目标的关键理念。

处理隐私风险：合并仅在输入预训练的 LLM 或在非隐私数据上微调的 LLM 进行设置，这些 LLM 具有相同架构和参数 $\theta_1\ \theta_2 \ldots, \theta_n$ 。本质上，输入模型 $\mathcal{M}_i$ 在训练和推理时均不处理隐私风险。

优化任务性能：给定一个使用度量标准 m 可测量表现的试验任务 τ ，PatientDx 构建一个具有参数 θ^* 的单一合并模型 $\mathcal{M}_e^*$ ，达到最佳表现 $m(\tau)^*$ 。因此，为构建模型 $\mathcal{M}_e^*$ ，PatientDx 依赖于核心参数合并函数 f ，引入了如 $\mathcal{M}_e^* = f(\lambda^*, \mathcal{M}_{i=1}^n)$ 和 $\lambda^* = argmax_{\lambda_i\ i=1\ldots n} m(\tau)$ 的标量特定超参数 λ_i 。应强调 PatientDx 需要一个度量标准来优化合并超参数，如 $m(\tau^e)^* \geq m(\tau)_i$ ，而不需在私人数据上训练 $\mathcal{M}_e^*$ 或在合并后进一步微调它。

虽然学习最佳合并函数值得探索，但这留待未来的工作中进行。我们仅考虑在不损失普遍性的情况下使用的最新合并函数，并专注于从任务性能的角度识别最佳超参数。我们具体考虑 $n = 2$ 和以下两个合并函数：

- 模型汤 (Wortsman et al., 2022)：包含使用模型系数对输入模型权重进行线性组合。形式上写为 $\theta^* = \sum_{i=1}^n \lambda_i \theta_i$ ，其中 $\sum_{i=1}^n \lambda_i = 1$ 和 $\forall_i \lambda_i > 0$ 。
- SLerp (Jang et al., 2024)：与模型汤不同，SLerp 基于输入模型的角度组合，如 $\theta^* = \sum_{i=1}^n \frac{\sin(\lambda_i \Omega)}{\sin(\Omega)} \theta_i$ ，其中 $\sum_{i=1}^n \lambda_i = 1$ 和 $\forall_i \lambda_i > 0$ 。对于 $n = 2$ ，Ω 是由向量 $\overrightarrow{\theta_1}, \overrightarrow{\theta_2}$ 和 $\cos(\Omega) = \overrightarrow{\theta_1} \times \overrightarrow{\theta_2}$ 形成的弧所夹的角度。

3 实验和结果

我们进行了实验来回答以下研究问题：

- RQ1。合并模型在患者诊断（死亡率）方面是否比输入模型更有效？如果患者描述包含更多的数值数据，其表现是否相同？
- 研究问题 2：合并模型是否比微调模型更少受到数据泄露现象的影响？
- RQ3. 合并模型在下游任务中的有效性是否与输入模型相当？它们能够回答与患者相关的问题吗？它们在信息检索导向的任务中有用吗？

为了回答 RQ1 和 RQ2，我们选择了 MIMIC-IV (Johnson et al., 2023) 数据集，这是一个公开可用的医疗领域的关于患者数据的信息的数据集，而 RQ3 则通过从医疗领域的研究文章中提取的问题进行探索。

3.1 数据集和实验设置

MIMIC-IV 数据集 (Johnson et al., 2023) 用于运行我们的实验。特别是，我们选择了数据集中心

提供的死亡率配置[1]，如 Lovon-Melgarejo et al. (2024) 中所述。该死亡率数据集采用文本表示的患者信息，如在章节 2.1 中显示，由六个主要文本信息组成：人口统计、诊断、ChartEvents、药物、程序和 OutputEvents。此外，输入被修改为侧重于输入的数值，即 CharEvents 和 Medications 部分。这个更数字化的数据集在我们的实验中被重命名为 Mortality-hard。在这两种情况下，任务都是预测患者的描述是否对应一个已经去世或幸存的患者。两个数据集的统计信息显示在表格 **??** 中。需要注意的是，删除更多文本信息的效果极大地影响了输入中的数字的数量，因为比例从 9.86 % 变为 13.51 %，而字母的下降和空格保持在类似的比例（≈ 15 %）。

在超参数选择方面，对于我们的模型和微调后的模型，进行了将数据集分为 k 折的处理，其中 k 等于 2，[2]。我们将所有配置的提示词固定为在 Lovon-Melgarejo et al. (2024) 中提出的提示词，该提示词直接向大语言模型提问并建议输出格式。完整提示词是："您是一位极为乐于助人的医疗助理。您仅使用是或否来回答问题，并考虑到患者的医院档案：{ patient_data }。问题：患者是否死亡？答案（是或否）："。

对于死亡率集合，使用了标准度量指标，即接收者操作特征曲线下面积（AUROC）和精确率-召回曲线下面积（AUPRC）。在二分类任务的失衡条件下，这两个指标都很有用，因为其他指标可能会产生误导，而 AUPRC 对类别失衡更加敏感。针对表格 **??** 中的两个数据集，低于 0.5 和 0.1 的性能分别不比随机效果更好（针对 AUROC 和 AUPRC）。最后，由于 LLMs 的预测是原始文本，对于 AUROC 的计算，我们将输出限制为两个标记，并验证生成的答案中关于问题中是否有正面（"是"，"死亡"，"1"）或负面（"否"，"存活"，"活着"，"0"）的词语。对于 AUPRC 的计算，我们使用了仅 "是" 和 "否" 词语的标准化概率，如 Zhuang et al. (2024) 中所建议的。

为了合并这些模型，我们使用了一个名为 MergeKit 的公开可用工具。为了简化起见，我们选择了两个基础模型，Mistral 和 Llama，以及基于三个类别的后续模型。

请注意，这些模型的多种组合是可能的。然而，我们专注于基于数学模型的组合，因为观察 1（参见 § 2.1）。对于我们提出模型的每种组合，我们重新命名了 θ^* 如下：

- PatientDx 7B：此配置探索了 Mistral 模型（Instruct 和 Math）的组合。
- PatientDx 8B：此配置探索 Llama 模型（Instruct 和 Math）的组合。
- PatientBioDx 8B：此配置还探索组合 Llama 模型，但在医学文本（生物医学和数学）中进行预训练。

我们的主要结果呈现在表格 **??** 中。LLM 类别 BioMedical、Instruct 和 Math 代表了在训练[3] 期间按其专业化分组的强大 LLM 基准。最后一类，合并模型，代表了我们的贡献（表中给出了每个 θ^* 模型的 λ^* 值）。对于死亡率任务，重要的是要注意，大多数模型在 AUROC 指标上的表现接近 0.5，包括 BioMistral、Llama Instruct、Med42、Mathstral 和 DART math。只有 Meditron 和 Mistral Instruct 模型才能获得大于 0.55 但小于 0.6 的值。在 AUPRC 方面，Med42 是一个强大的基准（0.20），与其他基准明显不同（<0.16）。

然而，我们提出的 PatientDx 和 PatientBioDx 模型在 AUROC 方面的表现优于所有以前的基线。特别是，PatientDx 8B 配置提高了 0.07 个绝对点，是最强的基线。还要注意，与输入模型 Llama3 和 DART math 相比，PatientDx 8B 模型的提升超过 0.1（从 0.5005-0.5015 到 0.63），这表明合并模型的提议允许大幅改进。该结果使我们能够回答 RQ1 的第一部分，即 PatientDx 模型可以超越输入模型。对于 Mortality-hard，在某些差异下，观察到了与 Mortality 类似的行为。总体而言，基线和我们的贡献表现下滑，除了少数例外。对于基线，AUROC 的最大降幅出现在 Mistral 7B Instruct 模型（-0.0656），而 AUPRC 则在 Med42 8B 模型（-0.0881）中观察到。对于我们的模型，AUROC 的较大降幅出现在 PatientDx 7B 模型（-0.1057），而 AUPRC 则出现在 PatientBioDx 8B 模型（-0.0703）。这些证据表明了 Mortality-hard 数据集的难度，并且还指出在我们的模型中，PatientDx 8B 模型似乎更加鲁棒，并且在文本信息减少后影响较小。两个数据集之间的平均表现显示在 "Average" 列中。这些列证实在 AUROC 和 AUPRC 方面，我们的模型 PatientDx 8B 与最近的生物医学基线（如 Meditron 7B 和 Med42 8B）相比相当有竞争力。这个关于 Mortality-hard 的数据集的结果完成了 RQ1，因为更多的数值患者数据会对各个基线和我们的模型的表现产生负面影响，只有 PatientDx 8B 在该数据集上在 AUROC 和 AUPRC 方面保持一致的表现（Meditron 7B 和 PatientDx 7B 在一个指标上更好，要么是 AUROC，要么是 AUPRC，但在另一个指标上表现急剧下降）。

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

[2] 仅在测试部分考虑计算成本。

[3] 一般的训练，即使有些是完整训练而其他是持续的预训练。

我们对三种 PatientDx 配置进行了消融研究。在这种情况下，我们分析了与数学模型合并的影响，以及 SLerp 合并策略的影响（当 $\lim_{\Omega \to 0}$ 时，使用线性合并作为 SLerp 的等效替代）。这一探索的结果在表 1 中给出。正如我们的结果所示，与数学模型合并的实用性是一个关键特征，而平均下降 13.7 % 和其他策略（除 SLerp 外）则平均负面影响为 14.4 %。对于我们表现更好的模型 PatientDx 8B，与数学模型结合似乎比使用 SLerp 作为组合策略更关键。排除这两个特征对模型的负面影响平均下降 17.3% 。

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

Table 1: PatientDx 配置的死亡率任务消融研究的 AUROC 结果。w/o SLerp 对应于输入模型的线性组合（模型混合），w/o Math 对应于不使用数学 LLM。

为了评估我们方案在调优期间保护患者数据的能力，我们使用了新的指标，Δ_1 和 Δ_2 ，称为数据泄露测试（DLT）(Wei et al., 2023)，它可以测量训练数据上的预期数据泄露。Δ_1 通过计算用于训练的文本（$\mathcal{P}_{train}$ ）与作为参考的文本（$\mathcal{P}_{ref}$ ）之间的困惑度差异来评估数据泄露的风险。注意，较大的值表明模型泄露数据的风险较低。类似地，Δ_2 计算训练（$\mathcal{P}_{train}$ ）和测试数据集（$\mathcal{P}_{test}$ ）之间的困惑度差异，较小的值表明训练数据没有被调优（无论是训练还是测试），而较大的值则表示任何分区都有某种过拟合。请注意，直观上 Δ 指标的行为不依赖于最终任务，而是依赖于全文的困惑度。对于参考文本的生成，我们使用了 Mistral 和 Llama 来自动生成文本。微调采用了 LoRa 优化策略，通过最佳超参数对相应集合进行处理。

数据泄露评估结果如表 **??** 所示。在此次评估中，我们包含了 PatientDx 8B 和在 Zero-shot 及微调配置下评估的强基线。注意，Δ_1 表明在所有未微调模型（NoFT）的死亡率和死亡率困难任务中，相似的数值（介于 2.20 到 4.30 之间）出现在这两个集合中。较大的数值则出现在 Med42 8B 和 PatientDx 8B，表明在 Zero-shot 条件下，这些模型较不易泄露病人信息。所有未微调模型的低 Δ_2 值也证实了这一点。另一方面，所有微调模型在死亡率数据集中显示的泄露风险都比未微调模型大。对于死亡率困难任务，只有 Mathstral 7B 在未微调模型的数值范围内。然而，Δ_2 指标表明该模型存在某种过拟合，这可能是由于数据集中较多的数字以及模型的数学专长所致。关于研究问题 2（RQ2），我们明显观察到微调模型相比未微调模型，包括 PatientDx，具有更高的泄露风险。

该问题被选中以在输入（患者年龄）和输出（剂量信息）中包含数字数据。我们更稳定的模型 PatientDx 8B 以及表现最好的基线模型 Meditron 7B 和 Med42 8B 的输出如在表 **??** 中所示。每个输出被限制为 200 个符号，提示类似于在章节 **??** 中使用的，并在表 **??** 中完整显示。Meditron 的预测是一个与任务无关的问题回答问题的完成，然后它偏离到一个不同的患者描述（44 岁的女性）。另一方面，Med42 在回答中更加连贯，包含警告和关于答案的常规信息。两个数学模型都提供了较短的答案并包含了更多相关的数字信息。我们有趣地看到 PatientDx 8B 比 DART math 提供了一个更具上下文的答案，并且保持一致，包括了数字数据。经过仔细检查，结论是 Med42 8B 是最完整的[4] 答案，因为它在推理中包含了患者的病情。PatientDx 8B 包括了有用的计算，但未能包括患者的病情。然而，这个结果显然显示了将模型与数字数据结合用于数值相关问题的潜力。

由于我们的方案暗示了模型参数平均化，一个直接的直觉是最终模型在未见任务上可能会有意外表现。因此，我们对我们模型提议的答案进行了定性评估，并将其与强基线进行比较。在医学领域，可以基于患者问题进行定性或定量（基于专家）的评估。为了定性评估这一影响，我们使用文献中可用的一个生物医学相关问题。

最后，信息检索性能使用从 Zhao et al. (2023) 中提出的医学文章中提取的患者数据集进行评估。我们特别关注 ReCDS-PPR 任务，该任务在由 155.2k 候选患者和 2.9k 患者描述作为查询组成的语料库中寻找相似患者。在查询扩展设置中，使用大语言模型作为关键词生成器。为了获得关键词，我们使用了以下提示语：“你是一位高效的信息检索助手。有哪些最相关但缺失的关键词（通过同义词或逻辑推理）应当添加到以下患者档案中以帮助识别相似患者？患者：{ patient_data }。关键词：”。使用 BM25 检索与原始和扩展查询相似的患者，因为此词汇排序器在此任务中表现强劲 (Zhao et al., 2023)。使用标准信息检索指标的评估结果如表 **??** 所示。出于计算原因，我们使用 4 位量化版本的 PatientDx 8B 进行扩展评估，并将生成的令牌大小限制为 200。通过 RRF 进行与 BM25 的排名融合也使用了 Bassani (2022)。结果显示，只有 RRF 组合略微改进了 BM25 基线，但统计测试显示两者之间无显著性。在 RQ3 的结论中，虽然 PatientDx 8B 作为医学计算的数学

[4]这基于 法国医学法规（访问于 2024 年 10 月 15 日）。

工具显得有用，但其在使用查询扩展框架的信息检索中的表现仍需研究。

在这篇论文中，我们研究了将大语言模型（LLM）合并作为一种竞争策略，以获得具有竞争预测能力的新共享模型，同时没有数据隐私泄露的风险。我们在患者数据上的结果显示，将一个数学模型与指令或生物医学模型合并，可以在死亡率任务上获得改进。作为一个主要观察，我们可以突显出一个显著的 7% 提升，当将 PatientDx 8B 与输入 LLM 进行比较时。此外，同样的模型编码的训练信息比微调的替代方案更少，显示出我们提出的合并是一种可靠的策略，可以以最低泄漏风险将“微调”权重共享给数据集。最后，我们展示了 PatientDx 8B 可能的用途，比如回答医学问题和检索相似患者。尽管这篇论文有进展，但仍然存在一些局限性。主要的局限性是我们的框架需要离散和详尽的评估来生成新模型，但也存在其他局限性，如与替代方案相比性能较低，以及在其他以患者为导向的任务中的更广泛评估。然而，我们的提议可以快速受益于可以直接作为输入的新 LLM。与微调不同，我们的方案在计算能力方面相对较轻。未来的工作可能会集中在更优化的方式来结合权重，提高性能而不增加计算成本。诸如 Akiba et al. (2024) 之类的工作可能是探索更复杂合并策略的有趣方法。

主要的伦理考量是医疗大型语言模型被误用的后果。注意，此项工作旨在在学术环境中使用，并支持医疗工作队伍和研究 [5] 。为了评估我们模型的泛化能力，可以在整个训练集上进行超参数选择（没有 3.1 节中描述的 k 次折叠测试），但这将导致显著更高的计算成本。

这项工作得到了由 HDH（法国）和 FRQS（加拿大）资助的 In-Utero 项目的支持。

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.

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

[5] 对于任何医疗问题，请咨询专家。

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 classication 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.