PatientDx : Fusion des grands modèles de langue pour la protection de la confidentialité des données dans le domaine de la santé

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 - équipe SPHERE,
Faculté de Médecine Université de Toulouse, Toulouse, France
[1] first.last@irit.fr, [2,3] first.last@univ-tlse3.fr

RÉSUMÉ

L'affinage des grands modèles de langue (abrégé LLM de l'anglais *large language model*) est devenu la pratique courante pour améliorer la performance des modèles sur une tâche donnée. Cependant, cette amélioration de performance s'accompagne d'un coût : l'entraînement sur de vastes quantités de données annotées potentiellement sensibles, ce qui soulève d'importantes préoccupations en matière de confidentialité des données. Le domaine de la santé constitue l'un des domaines les plus sensibles exposés aux problèmes de confidentialité des données. Dans cet article, nous présentons "PatientDx", une architecture de fusion de modèles permettant de concevoir des LLM efficaces pour les tâches prédictives en santé sans nécessiter d'affinage ni d'adaptation sur les données des patients. Notre proposition repose sur des techniques récemment proposées connues sous le nom de fusion de LLM et vise à optimiser une stratégie de fusion modulaire. "PatientDx" utilise un modèle pivot adapté au raisonnement numérique et ajuste les hyperparamètres sur des exemples en fonction d'une métrique de performance, mais sans entraîner le LLM sur ces données. Les expériences utilisant les tâches de prédiction de mortalité de l'ensemble de données MIMIC-IV montrent des améliorations jusqu'à 7% en termes d'AUROC par rapport aux modèles initiaux. De plus, nous confirmons que, comparée aux modèles affinés, notre proposition est moins sujette aux problèmes de fuite de données sans nuire à la performance. Enfin, nous démontrons qualitativement les capacités de notre proposition à travers une étude de cas. Notre meilleur modèle est publiquement disponible : https://huggingface.co/Jgmorenof/mistral_merged_0_4. Ceci est le résumé de l'article publié "PatientDx : Merging Large Language Models for Protecting Data-Privacy in Healthcare" dans l'atelier CL4Health, NAACL 2025 (Moreno *et al.*, 2025).

ABSTRACT

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

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. This is the summary of the published paper "PatientDx : Merging Large Language Models for Protecting Data-Privacy in Healthcare" in the CL4Health 2025 Workshop proceedings, NAACL 2025 (Moreno *et al.*, 2025).

MOTS-CLÉS : grands modèles de langue, fusion des modèles, confidentialité des données.

KEYWORDS: large language models, model merging, data privacy.

ARTICLE : **Accepté à** l'atelier CL4Health (NAACL) 2025 (https://bionlp.nlm.nih.gov/cl4health2025/) .

Références

MORENO J. G., LOVON-MELGAREJO J., ROBIN-CHARLET M., DAMASE-MICHEL C. & TAMINE L. (2025). PatientDx : Merging large language models for protecting data-privacy in healthcare. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health) @ NAACL*, p. 1–11, Albuquerque, New Mexico : Association for Computational Linguistics.