Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

La spécialisation de domaine est-elle toujours pertinente ? Une étude de l'adaptation de modèles de langue génératifs sur un nouveau corpus biomédical français

Aidan Mannion[1] Cécile Macaire[1] Armand Violle[2] Stéphane Ohayon[2]
Xavier Tannier[2] Didier Schwab[1] Lorraine Goeuriot[1] François Portet[1]
(1) Université Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
(2) Sorbonne Université, LIMICS, 15 rue de l'École de Médecine, 75006 Paris, France
[prenom.nom]@univ-grenoble-alpes.fr, [prenom.nom]@sorbonne-universite.fr

RÉSUMÉ

Les grands modèles de langue ont démontré des capacités remarquables dans divers domaines, mais leur adaptation à des domaines spécialisés reste difficile. Cette étude examine le pré-apprentissage comme stratégie visant à spécialiser les modèles de langue de taille moyenne dans le domaine biomédical français grâce à un pré-apprentissage continu. Nous abordons des questions de recherche autour du pré-apprentissage continu spécialisé pour l'adaptation au domaine et la relation entre les gains de performance spécifiques au domaine et la dégradation des capacités générales. Nos contributions comprennent la publication d'un corpus biomédical français sous licence entièrement libre et de modèles de langue biomédicaux français spécialisés, ainsi que de nouvelles perspectives pour la mise en œuvre du pré-apprentissage spécialisé. Nos résultats suggèrent que la fusion des modèles (merging) est essentielle pour atténuer les compromis liés à la généralisation et peut même améliorer les performances sur certaines tâches spécialisées. Les données et les modèles sont accessibles à partir de la page suivante : https://huggingface.co/spaces/HealthDataHub/PARTAGES.

ABSTRACT

Is domain specialisation still worth it ? Insights from the adaptation of generative language models to a new French biomedical corpus

Large language models (LLMs) have demonstrated remarkable capabilities across diverse domains, yet their adaptation to specialized fields remains challenging. This study investigates domain-adaptive pre-training (DAPT) as a strategy for specializing small to mid-sized LLMs in the French biomedical domain through continued pre-training. Our contributions include the release of a fully open-licensed French biomedical corpus suitable for commercial and open-source applications, the training and release of specialized French biomedical LLMs, and novel insights for DAPT implementation. Our findings suggest that model merging post-DAPT is essential to mitigate generalization trade-offs, and in some cases even improves performance on specialized tasks at which the DAPT was directed.

MOTS-CLÉS : Adaptation aux domaines spécialisés, TALN biomédical.

KEYWORDS: Domain-adaptive pre-training, Biomedical NLP.

ARTICLE ACCEPTÉ À : LREC 2026 : The Fifteenth biennial Language Resources and Evaluation Conference, Palma, Mallorca, Spain, May 11-16, 2026..
URL : https://lrec.elra.info/lrec2026-main-833

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 2 : articles déjà publiés, pages 24–25.

Remerciements

Ces travaux ont été menés dans le cadre du projet PARTAGES, lauréat de l'appel à projets « Digital Commons for Generative Artificial Intelligence » de Bpifrance France 2030. Il a également été partiellement soutenu par l'agence nationale de la recherche (ANR) via la chaire MIAI « IA & Langage » (ANR-23-IACL-0006). Ces travaux ont été réalisés à l'aide des ressources HPC de GENCI à l'IDRIS, dans le cadre de l'allocation 2025-A0181016171 sur le supercalculateur Jean Zay.