Sujet de Thèse : Approches neurosymboliques pour le traitement de données hospitalières

Encadrement : Sandrine Mouysset (MCF, HDR), Laurent Molinier (Pr, CHU), Eric de la Clergerie (DR, Inria), Jean-Marc Alliot (DR, CHU)

Contexte :

L'essor des modèles de langage à grande échelle (Large Language Models, LLM) constitue un catalyseur majeur pour la transformation numérique du secteur de la santé. En automatisant la rédaction de compte-rendus, en facilitant le triage des patients et en assistant la décision clinique, ces architectures ouvrent de nouvelles perspectives opérationnelles (Rajpurkar et al., 2022). Des travaux récents démontrent leur capacité à alléger la charge administrative des praticiens et à synthétiser rapidement de grands volumes de littérature médicale (Clusmann et al., 2023). Néanmoins, le déploiement effectif de ces outils se heurte à un verrou technologique et réglementaire. En effet, les modèles de pointe de la littérature — à l'instar de la famille GPT (OpenAI), de Med-PaLM (Singhal et al., 2023) ou de MedGemma — reposent majoritairement sur des jeux de données d'entraînement anglophones, créant un biais linguistique et culturel face aux terminologies médicales francophones (comme le CIM-10 ou la CCAM) et sont hébergés dans des environnements non souverains.

Ce travail de thèse s'inscrit dans le cadre du projet FrenchGen4Health qui vise à combler ce manque en développant le premier modèle de langage génératif médical francophone, entraîné sur des données de santé françaises anonymisées et hébergé sur des serveurs HDS souverain.

Objectifs de la thèse :

L'objectif principal du projet est de constituer un commun numérique national fondé sur le développement d'un modèle de langage génératif (LLM) de très grande taille (*giga-modèle*), spécifiquement entraîné sur la terminologie biomédicale francophone et les spécificités des écosystèmes hospitaliers. En s'appuyant sur une architecture mutualisée et ouverte, cette initiative vise à démocratiser l'accès aux technologies d'IA générative avancées pour l'ensemble des acteurs de santé.

Les objectifs secondaires sont de :

- Développer un modèle de langage open-source entrainé sur des données publiques de santé,

- Entraîner un modèle « communautaire » spécialisé sur des données réelles pseudonymisées issues de 11 établissements partenaires,

- Tester un cas d'usage clinique (rédaction automatisée de comptes rendus de consultations) pour évaluer l'impact sur la qualité des soins et la charge administrative,

- Définir une gouvernance éthique et technique garantissant la conformité RGPD et la sécurité des données,

- Mettre en place une infrastructure pérenne (Data Lab LLM) favorisant la reproductibilité scientifique.

Références :

Rajpurkar, P. et al. (2022). *AI in health and medicine*. Nature Medicine, 28(1), 31-38.

Clusmann, J. et al. (2023). *The future landscape of large language models in medicine*. Communications Medicine, 3(1), 141

Singhal, K. et al. (2023). *Large language models encode clinical knowledge*. Nature, 620(7972), 172-180