Le juste usage des LLM: Méthodes hybrides explicables *et sobres* pour le NLP en cancerologie

Laboratoire : Laboratoire de Biomécanique Appliquée (LBA)
Projet : DuraXell (Durabilité et eXplicabilité des méthodes de NLP en santé)
Tuteur : Akram Redjdal (bureau 6407)

Filières visées :

— e-santé

— Data science et intelligence artificielle

— Artificial Intelligence and Cybersecurity

1) Contexte scientifique du projet

Dans le domaine de l'oncologie, plusieurs informations comme le stade, la classification TNM, et les biomarqueurs sont cruciales à la fois pour le pronostic et pour le choix thérapeutique. La plupart de ces informations ne sont pas disponibles sous forme structurée ; elles sont disséminées dans les comptes rendus d'anatomopathologie, opératoires, comptes rendus de *réunions de concertation pluridisciplinaire* (RCP)...etc. Il est donc important de pouvoir les extraire automatiquement pour :

— des fins cliniques (préparation RCP, continuité des soins, réduction des erreurs de retranscription),

— des fins organisationnelles (indicateurs qualité, registres, pilotage),

— des fins scientifiques (recherche clinique, études "real-world", présélection de patients pour les essais cliniques).

Le développement actuel des entrepôts de données de santé (EDS) suscite un enthousiasme en tant que mine d'informations liées à l'expérience « en vie réelle » des patients, collectées massivement dans le cadre des soins **(1)**. Ces EDS cherchent à améliorer le pilotage de l'activité hospitalière et faire avancer la recherche, basés notamment sur les nouvelles technologies d'intelligence artificielle (IA). Deux obstacles s'opposent à une utilisation en routine de ces données de *big data* : leur complétude **(2)** et leur format de stockage textuel et non structuré, donc non exploitable automatiquement **(3)**.

Certaines initiatives cherchent à développer des algorithmes de traitement automatique du langage naturel (TALN ou NLP) afin de pallier ces deux problèmes **(4–5)**. Les systèmes d'extraction textuelle traditionnels basés sur des règles se sont vus, depuis quelques années, remplacés par les réseaux de neurones artificiels, en raison du bond sans précédent des puissances de calcul informatique et de l'augmentation massive des données disponibles pour leur entraînement

(6). Ce domaine a tiré pleinement parti du développement récent des modèles d'apprentissage machine (machine learning, ML) et, notamment, des *large language models* (LLM) **(7)(8)**.

Or, l'entraînement et la validation des modèles d'apprentissage machine posent deux problèmes : celui de la durabilité car leur consommation de ressources est importante (empreinte carbone, temps expert d'annotations de documents, nombre de documents annotés nécessaires au développement des modèles), et celui de l'explicabilité de l'IA car l'architecture du squelette du réseau de neurones demeure largement inaccessible aux développeurs (les tentatives d'explicabilité de type XAI comme les *SHAP diagrams* reposent, eux-mêmes, sur des modèles gourmands en ressources) **(9)**.

Hypothèse. L'hypothèse du projet « DuraXell » est qu'il serait possible de garantir des niveaux d'explicabilité et de durabilité optimaux des techniques de TALN, même quand elles impliquent un recours à des LLM. Il s'agit d'optimiser la consommation en ressources humaines et carbones pour une tâche d'extraction d'information textuelle donnée, en :

1. hybridant les différentes méthodes de TALN (règles, ML, LLM),
2. réservant l'utilisation des techniques consommatrices d'énergie et les moins explicables (ex. : LLM) aux cas complexes,
3. favorisant l'utilisation de LLM de petite taille, quand ils sont jugés nécessaires.

2) Objectifs

DuraXell vise, à terme, l'exploitation des données réelles de l'EDS de l'AP-HP ; **dans le cadre du tremplin recherche**, l'objectif est de préparer le terrain en travaillant sur des données *open source*. Concrètement, il s'agit de :

1. constituer un jeu de données centré sur le cancer du sein,
2. opérationnaliser les pipelines à base de règles existantes (traduction FR→EN et enrichissement : variantes d'écriture, négation, temporalité, normalisation clinique) **(10)**,
3. comparer de manière contrôlée trois approches d'extraction — règles, ML léger et LLM — au moyen de métriques communes de performance, d'explicabilité et de coût en ressources (temps, mémoire, empreinte carbone, effort d'annotation),
4. sur cette base, mettre en place un mécanisme de triage capable de quantifier et prédire quand recourir à un modèle de plus grande taille, afin de réserver cet usage aux cas complexes et minimiser le coût en ressources, tout en maintenant la qualité des extractions. Cette méthodologie, "proof of concept" sur données ouvertes pendant le tremplin, sera déployée et validée en stage sur l'EDS de l'AP-HP.

3) Méthodologie

Principe : pour chaque entité cible (parmi une sélection d'entités qui sera définie) dans le domaine cancer du sein (TNM, ER/PR/HER2, Ki-67), on évalue la complexité d'extraction en amont, on choisit la méthode la plus sobre et explicable (règles → ML léger → LLM), puis on escalade seulement si nécessaire.

1. **Construire d'abord un jeu de données (JDD) "cancer du sein".** Agréger un petit corpus ouvert : 40 notes d'oncologie **CORAL** (EN) pour des exemples annotés, des cas cliniques en français du corpus **CAS** pour la variété FR, et des patients synthétiques générés avec **Synthea** pour couvrir des scénarios rares et contrôler l'échelle. Constituer ensuite un échantillon FR annoté (60–80 extraits) avec *spans* de preuve, dates et sources. (`https://physionet.org`)
2. **Poser une baseline par famille de méthodes.**
   (a) *Règles :* opérationnaliser les pipelines existantes **(10)**, les traduire FR→EN et les enrichir (variantes, négation, temporalité, normalisation clinique).
   (b) *ML léger :* entraîner un modèle compact (CRF / LogReg) pour affiner frontières et attributs sur le JDD annoté.
   (c) *LLM :* évaluer un petit LLM (4–7B, quantifié) en extraction guidée par *prompt* court (2–3 indices cités).
3. **Proposer des métriques d'évaluations et de décisions :** précision, rappel, F1 par cible (stade/TNM, ER/PR/HER2, Ki-67), taux d'erreurs cliniquement critiques, le % d'extractions justifiées, les ressources (temps/mémoire, % d'appels LLM, Wh/CO_2e estimé).
4. **Évaluer les baselines sur le même JDD** pour une comparaison.
5. **Proposer une version hybride minimale (LLM "on-demand").** Implémenter une cascade : règles → ML → LLM, avec des déclencheurs simples (faible confiance du ML, mentions en conflit, ambiguïté temporelle, ambiguïté linguistique, valeur manquante).

4) Perspectives

Le projet sera prolongé en stage sur des données réelles de l'Entrepôt de Données de Santé, avec les objectifs suivants : (i) généraliser à davantage de documents (anapath, imagerie, RCP), (ii) valider à l'échelle et en contexte hospitalier (robustesse et generalisabilité), (iii) affiner le triage (seuils *data-driven*, détection de conflits plus fine), (iv) étendre le périmètre vers plus de cancers et plus d'entités cliniques.

5) Références

1. Pastorino R, De Vito C, Migliara G, Glocker K, Binenbaum I, Ricciardi W, et al. Benefits and challenges of Big Data in healthcare : an overview of the European initiatives. 2019 ;29 :23–7.
2. Priou S, Lame G, Jankovic M, Chatellier G, Bey R, Tournigand C, et al. Why Are Data Missing in Clinical Data Warehouses ? A Simulation Study of How Data Are Processed (and Can Be Lost). *Stud Health Technol Inform.* 2023 May ;302 :202–6.
3. Foran DJ, Chen W, Chu H, Sadimin E, Loh D, Riedlinger G, et al. Roadmap to a Comprehensive Clinical Data Warehouse for Precision Medicine Applications in Oncology. *Cancer Inform.* 2017 ;16 :1176935117694349.
4. Jung HA, Jeong O, Chang DK, Park S, Sun JM, Lee SH, et al. Real-time autOmatically updated data warehOuse in healThcare (ROOT) : an innovative and automated data collection system. *Transl Lung Cancer Res.* 2021 Oct ;10(10) :3865–74.

5. Saha A, Burns L, Kulkarni AM. A scoping review of natural language processing of radiology reports in breast cancer. *Frontiers in Oncology*. 2023 ;13 :1160167.

6. Petch J, Kempainnen J, Pettengell C, Aviv S, Butler B, Pond G, et al. Developing a Data and Analytics Platform to Enable a Breast Cancer Learning Health System at a Regional Cancer Center. *JCO Clin Cancer Informatics*. 2023 Mar ;7 :e2200182.

7. Vincent M, Douillet M, Lerner I, Neuraz A, Burgun A, Garcelon N. Using Deep Learning to Improve Phenotyping from Clinical Reports. *Stud Health Technol Inform*. 2022 Jun 6 ;290 :282–6. Disponible sur : `https://pubmed.ncbi.nlm.nih.gov/35673018/`

8. Kehl KL, Xu W, Gusev A, Bakouny Z, Choueiri TK, Riaz I Bin, et al. Artificial intelligence-aided clinical annotation of a large multi-cancer genomic dataset. *Nat Commun*. 2021 Dec ;12(1) :7304.

9. Tjoa E, Guan C. A Survey on Explainable Artificial Intelligence (XAI) : Toward Medical XAI. *IEEE Trans Neural Networks Learn Syst*. 2021 Nov ;32(11) :4793–813.

10. Redjdal A, Novikava N, Kempf E, Bouaud J, Seroussi B. Leveraging Rule-Based NLP to Translate Textual Reports as Structured Inputs Automatically Processed by a Clinical Decision Support System. *Stud Health Technol Inform*. 2024 ;316 :1861–1865. doi : 10.3233/SHTI240794.