Offre d'alternance

DOMAINE DE L'ALTERNANCE : DATA SCIENCE (BUSINESS UNIT BIOMETRIE & DATA SCIENCE)
Réf.: ALT025_DS
LIEU : Saint Herblain (44)

ALTERNANCE EN DATA SCIENCE

EXPLOITER LES LARGE LANGUAGE MODELS (LLM) POUR LA REDACTION SCIENTIFIQUE ET MEDICALE

L'entreprise

Biofortis SAS est une société de services en recherche clinique ou CRO (Contract Research Organization) dédiée à l'innovation dans les secteurs agroalimentaire, nutrition, santé, cosmétique et pharmaceutique. Biofortis accompagne le développement des produits de ses clients en offrant des services originaux et innovants allant du développement scientifique de nouveaux produits, en passant par l'apport de preuves précliniques et cliniques, jusqu'aux analyses du microbiote et leur interprétation scientifique.

Forte de ses 80 collaborateurs, notre entreprise présente à son actif plus de 500 projets, 250 essais cliniques gérés full-service en France, en Europe et à l'international.

Dans le cadre de ses activités de R&D, Biofortis recherche un(e) alternant(e) en Data Science pour une durée d'un an au sein de la Business Unit Biométrie et Data Science.

Contexte

L'essor de l'intelligence artificielle (IA) a considérablement transformé le paysage de la recherche médicale, offrant des possibilités sans précédent pour optimiser les essais cliniques (gain de temps, maximisation des chances de succès, diminution des coûts, …) et notamment dans le contexte de la rédaction scientifique et médicale.
Dans ce contexte, les grands progrès réalisés dans les modèles de langue de grande taille (LLM) ont ouvert la voie à la rationalisation et à la facilitation de tâches complexes telles que la recherche bibliographique et l'aide à la construction de design d'études, ou bien encore la rédaction de rapports et de résumés d'études cliniques.

Objectifs

Ce projet vise à développer, évaluer et comparer des solutions basées sur différents modèles de langue de grande taille (LLM, par exemple des modèles de type GPT) pour la génération de protocoles, rapports et résumés d'études cliniques sur des architectures privées et sécurisées.

Plus spécifiquement, les objectifs sont les suivants :

- Revue des outils existants (commerciaux ou libres)
- Identification des LLM les plus pertinents.

- Développement, installation et configuration de solutions LLM sur des architectures privées, sécurisées et proposant des ressources de calculs (serveurs internes, cloud AWS, …).
- Évaluation de la qualité des outputs générés par chaque solution développée.
- Analyse des performances des différentes solutions LLM dans le traitement de données cliniques hétérogènes, incluant à la fois des formats structurés et non structurés.Proposition de recommandations pour l'amélioration des performances des solutions LLM (fine-tuning) dans ce contexte spécifique.

Dans l'ensemble de ce projet, il sera nécessaire de prendre en compte les recommandations réglementaires et éthiques pour le traitement des données (RGPD) et le développement de solutions à base d'IA (notamment le European AI Act).

Ce travail sera réalisé en collaboration avec les équipes de data science et de rédaction scientifique et médicale.

Le(la) stagiaire pourra également être amené à participer aux activités de data science de l'équipe.

Activités principales

- Réaliser une veille scientifique sur les avancées en LLM et leur application en recherche clinique.
- Préparer et nettoyer les données disponibles (formater, anonymiser, structurer) pour les rendre exploitables par les modèles.
- Implémenter et déployer des pipelines de génération de protocoles et rapports en utilisant les LLM identifiés.
- Mettre en place des environnements de test sur architectures privées ou cloud (type AWS) pour évaluer les performances (temps de génération, pertinence, cohérence).
- Concevoir et exécuter des jeux de tests pour comparer différents modèles et configurations (paramètres, fine-tuning).
- Analyser les résultats générés par chaque solution (qualité linguistique, respect des contraintes cliniques, faisabilité pratique) et rédiger des rapports d'évaluation.
- Collaborer avec l'équipe de rédaction médicale et qualité pour valider la pertinence scientifique et réglementaire des contenus générés.
- Proposer des axes d'optimisation (fine-tuning, ajustements des prompts, architecture) et participer à leur implémentation.
- Documenter le code, les protocoles d'évaluation et les recommandations pour assurer la reproductibilité.
- Présenter régulièrement l'avancement du projet aux équipes de data science et de rédaction scientifique.

Profil recherché

- Formation requise :
  - Master 2 ou dernière année d'école d'ingénieur en intelligence artificielle, data science, bioinformatique ou dans un domaine connexe, avec idéalement un projet académique ou stage en IA appliquée à la santé.

- Compétences techniques requises :
  - Programmation en Python, API d'IA et traitement du langage naturel (NLP).
  - Bonne connaissance des modèles LLM et des concepts associés (fine-tuning, évaluation des performances, prompt engineering).

    - Expérience concrète avec au moins un framework IA/ML comme TensorFlow, PyTorch, Hugging Face Transformers.
    - Compréhension des architectures RAG et capacité à intégrer ces concepts dans des applications LLM.
    - Intérêt pour la sécurité et confidentialité des données dans un environnement réglementé (RGPD).
- Expérience souhaitable :
    - Expérience avec les architectures cloud (AWS, GCP) ou serveurs de calculs.
    - Connaissance de LangChain pour prototyper des cas d'usage LLM (résumé automatique, génération de protocoles, etc).
- Connaissance des normes ICH et des étapes du déroulement d'une étude clinique
- Compétences générales :
    - Anglais scientifique lu et écrit pour consulter la littérature internationale et collaborer avec des équipes non-francophones.
    - Excellentes capacités de communication et esprit d'équipe pour présenter régulièrement les résultats aux équipes pluridisciplinaires.
    - Rigueur, sens de l'initiative et autonomie pour gérer les projets techniques et proposer des améliorations concrètes.

Date de début : Septembre 2025

Durée : 1 an

Horaire hebdomadaire de travail : 35h

Gratification :
. Rémunération : calculée sur le SMC (Salaire Minimum Conventionnel) en fonction de votre âge et de votre niveau d'étude
. Participation aux titres restaurant et frais de transport en commun

Contact :
Merci d'adresser CV + lettre de motivation à :
Diego Tomassi, Senior Data Scientist : diego.tomassi@biofortis.fr