
DOMAINE DU STAGE : DATA SCIENCE (BUSINESS UNIT BIOMETRIE & DATA SCIENCE)

Réf.: ST025_DS

LIEU : Saint Herblain (44)

STAGIAIRE 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) stagiaire en Data Science au sein de la Business Unit Biométrie et Data Science pour une durée de 6 mois.

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 évaluer et comparer les performances de différentes solutions LLM (par exemple des modèles de type GPT), dans la génération de protocoles et rapports (et résumés) d'études cliniques sur des architectures privées et sécurisées.

Plus spécifiquement, les objectifs sont les suivants :

- Identification des solutions LLM les plus pertinentes.
- Installation et configuration de solutions LLM sur des architectures privées, sécurisées et proposant des ressources de calculs (type serveurs de calculs internes, cloud AWS, ...).
- Évaluer la qualité des outputs générés par chaque solution LLM.
- Analyser la capacité de chaque solution LLM à traiter des données cliniques complexes et variées, y compris des données structurées et non structurées.
- Proposer des recommandations pour l'amélioration des performances des solutions LLM (fine-tune) 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.

Profil recherché

- **Formation**
 - Master 2 ou dernière année d'école d'ingénieur en intelligence artificielle, data science, bioinformatique ou dans un domaine connexe.
- **Compétences techniques**
 - Compétence en programmation (Python, API d'IA, NLP).
 - Bonne maîtrise des modèles LLM et des concepts associés (fine-tuning, évaluation des performances, prompt engineering).
 - Expérience avec au moins un framework IA/ML comme TensorFlow, PyTorch, Hugging Face Transformers.
 - Compréhension des bases des architectures RAG et leur application.
 - Intérêt pour la sécurité et confidentialité des données dans un environnement réglementé (RGPD).
- **Souhaitable :**
 - Connaissance des architectures cloud / serveurs de calculs.
 - Expérience avec LangChain pour le développement d'applications LLM.
 - Connaissance des flux de travail de la documentation de la recherche clinique.
 - Compréhension de la science de la nutrition et de la recherche liée au microbiome.
- **Compétences générales :**
 - Anglais scientifique lu et écrit.
 - Communication, adaptabilité, rigueur, dynamisme, bon relationnel, esprit d'équipe, de synthèse et d'initiative.

Date de début : dès que possible

Durée : 6 mois

Horaire hebdomadaire de travail : 35h

Indemnité de stage : Rémunération convention de stage + 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