OFFRE DE STAGE EN NLP

Mission : Production de petits modèles de langue (SLM) pour des domaines de spécialité

Durée : 6 mois - Début du stage souhaité : mars 2025 - Lieu : EDF R&D Lab Saclay (91120) et LISN (Orsay)

Contexte et objectifs

La R&D d'EDF (2000 chercheurs) a pour missions principales de contribuer à l'amélioration de la performance des unités opérationnelles du groupe EDF, d'identifier et de préparer les relais de croissance à moyen et long terme. Dans ce cadre, le département Services, Economie, Outils Innovants et IA (SEQUOIA) est un département pluridisciplinaire (sciences de l'ingénieur, sciences humaines et sociales) qui fournit un appui à l'élaboration et au portage des offres, des services et des outils de relation client aux directions opérationnelles du groupe EDF.
Au sein de ce département, ce stage sera rattaché au groupe « Statistiques et Outils d'Aide à la Décision » (SOAD) : cette équipe compte une vingtaine d'ingénieurs chercheurs spécialisés en IA et data science avec des compétences fortes autour du machine learning et du deep learning, du web sémantique, de l'IA symbolique et de l'IA générative (texte, voix, image, multimodalité…), en particulier du NLP (LLM, RAG, data mining,).

Ce stage concerne le domaine du TAL (Traitement Automatique des Langues) et plus particulièrement les LLM (« Large Language Model ») popularisés par ChatGPT, Mistral, Llama, etc. Si l'utilisation des LLM est bien adaptée pour des usages généralistes, leur utilisation dans des tâches très spécifiques/spécialisées et répétitives pose des questions. Le ratio coût/performance des approches LLM généralistes n'est pas optimal. De plus, pour des tâches de classification, des articles [1] ont montré qu'une approche LLM pouvait être moins performante qu'une approche de type CamemBERT hyper optimisée (fine tunée sur du vocabulaire métier). Comme le montre [2], il y a un équilibre à trouver entre la taille du modèle utilisé (et indirectement son coût de fonctionnement) et ses performances [3].

Objectifs

Ce stage a donc pour objectifs de se focaliser sur les **petits modèles de langue (SLM) adaptés à des tâches spécifiques** [4] et [5]. Une poursuite en thèse CIFRE est envisagée, c'est pourquoi ce stage est proposé sous la forme d'une bilocalisation entre EDF R&D et le LISN[1] (laboratoire de recherche encadrant de cette thèse).

Etapes du stage :

- Partie bibliographique sur les approches existantes pour produire de petits modèles et notamment pour les domaines de spécialité.
- Partie expérimentale sur l'utilisation/implémentation d'une ou plusieurs de ces méthodes (par exemple, reproduire la méthode décrite dans un article avec ou sans code disponible).
- Partie test sur des données EDF, évaluer les performances d'un petit modèle sur une tâche particulière.
- Poser les jalons pour une thèse sur le sujet.

[1] - Le LISN est situé sur le plateau de Saclay, sur la commune d'Orsay (https://www.lisn.upsaclay.fr/)

Profil recherché :

- Formation : master machine-learning avec TAL + projet TAL.
- Niveau de langue : anglais (niveau B2 minimum), français oral et écrit.
- Compétences solides en programmation, en particulier en Python.
- Connaissance des techniques de traitement du langage naturel (NLP) et des modèles d'IA générative.
- Un goût pour la recherche est indispensable (autonomie, curiosité scientifique).
- Une procédure d'examen de candidature sera réalisée sur les deux sites (EDF et LISN).

Références

[1] Vautier, N., Héry, M., Miled, M., Truche, I., Bullier, F., Guénet, A. L., Dubuisson Duplessis, G.D., Campano, S et Suignard, P. (2024, July). Utilisation de LLMs pour la classification d'avis client et comparaison avec une approche classique basée sur CamemBERT. In APIA 2024.

[2] Grangier, D., Katharopoulos, A., Ablin, P., & Hannun, A. (2024). Specialized Language Models with Cheap Inference from Limited Domain Data. arXiv preprint arXiv:2402.01093. https://arxiv.org/pdf/2402.01093

[3] Pillet, X., Volkova, A., Greffard, N., & Dufour, R. (2024). Entre performance et frugalité en TAL: Approches pour la réduction de la taille des (L) LMs. https://hal.science/hal-04576377/document

[4] Lu, Z., Li, X., Cai, D., Yi, R., Liu, F., Zhang, X., ... & Xu, M. (2024). Small Language Models: Survey, Measurements, and Insights. arXiv preprint arXiv:2409.15790. https://arxiv.org/pdf/2409.15790

[5] Wang, F., Zhang, Z., Zhang, X., Wu, Z., Mo, T., Lu, Q., ... & Wang, S. (2024). A Comprehensive Survey of Small Language Models in the Era of Large Language Models: Techniques, Enhancements, Applications, Collaboration with LLMs, and Trustworthiness. arXiv e-prints, arXiv-2411. https://arxiv.org/pdf/2411.03350

Informations pratiques

Unité d'accueil : Groupe SOAD (Statistique et Outils d'Aide à la Décision), département SEQUOIA d'EDF Lab Paris-Saclay, 7 boulevard Gaspard Monge, 91120 Palaiseau.
Le stage sera encadré par des ingénieurs-chercheurs Data Scientist du département SEQUOIA et des chercheurs du LISN (C. Grouin et Th. Lavergne).

Transmettre par mail un CV et une lettre de motivation à:
philippe.suignard@edf.fr (Département SEQUOIA).