Offre de Stage IPSL 2024

(soutenu par le programme EUR IPSL-*Climate Graduate School)*

Titre du sujet de stage : **Leveraging Large Language Models for Climate information retrieval**

Description du sujet :

Ce stage vise à exploiter les récents progrès dans les Grands Modèles de Langage (Large Language Models ; LLM) pour concevoir un chatbot intelligent facilitant l'accès et la navigation au sein de la documentation interne du Centre de Modélisation du climat de l'IPSL. L'objectif est de créer un assistant virtuel capable de comprendre les requêtes des utilisateurs, de tirer des informations pertinentes de la documentation et de fournir des réponses précises. Un autre but est aider les chercheurs et ingénieurs de l'IPSL à trouver rapidement des informations à propos des données climatiques accessibles sur le mésocentre de l'IPSL https://mesocentre.ipsl.fr/ .

Un challenge particulier réside dans le souhait de se passer des modèles closed source comme GPT d'OpenAI, et donc travailler uniquement avec des modèles open source. Les modèles publiés par l'entreprise française Mistral peuvent être un premier point de départ. Suivant les compétences du stagiaires, nous décideront si nous partirons sur une approche type RAG (Retrieval Augmented Generation) ou bien sur du finetuning. Pour le front end on pourra utiliser Chat UI de Hugging Face.

Ce stage de Master 2 offre une opportunité unique de contribuer à aider la communauté des sciences du climat pour bénéficier au maximum des différents datasets (provenant de simulations numériques ou des observations) mais aussi les aider dans la recherche d'information pertinentes dans des publications ou de la documentation de données.

L'IPSL mettra à disposition du stagiaire ses moyens de calculs (clusters CPU et GPU). Il sera également envisagé un accès au super calculateur Jean Zay de l'IDRIS.

Comme référence deux projets intéressants sur des sujets proches:

- ChatECMWF https://github.com/ECMWFCode4Earth/ChatECMWF
- ClimateQ&A https://huggingface.co/spaces/Ekimetrics/climate-question-answering

Résumé en anglais (5 lignes) :

This internship focuses on utilizing Large Language Models (LLM) to develop an intelligent chatbot for enhanced access to the IPSL Climate Modeling Center's internal documentation. The goal is to create a virtual assistant capable of understanding user queries and providing accurate information, particularly aiding IPSL researchers in quickly accessing climate data. An interesting challenge is the preference for open-source models, with Mistral's models as potential starting points. The chosen approach, whether Retrieval Augmented Generation (RAG) or fine-tuning, will depend on the intern's skills.

Responsable du stage (Nom/prénom/statut) : **Redouane Lguensat, IRD/IPSL**

Laboratoire concerné : **IPSL**

Adresse à laquelle a lieu le stage : **4, place de Jussieu, 75005, Paris**

Equipe de recherche concernée (si pertinent) ou autre participant à l'encadrement du stage:

- **Sébastien Gardoll, IRD/IPSL**
- **Guillaume Levavasseur, SU/IPSL**

Niveau du stage (M1, M2, internship) : **M2**

Thème scientifique de l'IPSL concerné : **SAMA**

Durée du stage : **5 ou 6 mois**

Période : **01/03/2024 30/09/2024**

Rémunération de l'ordre de **580 euros par mois**

Compétences souhaitées: **Python, bibliothèques LLMs comme LangChain ou LLamaindex ou autre, familiarité avec les outils Git/Bash/JupyterNotebooks. Des connaissances en sciences de climat sont souhaitables mais pas nécessaires.**

Est-il prévu une thèse dans le prolongement du stage ? **Non, mais d'autres opportunités peuvent être considérés suivant les résultats du stage**