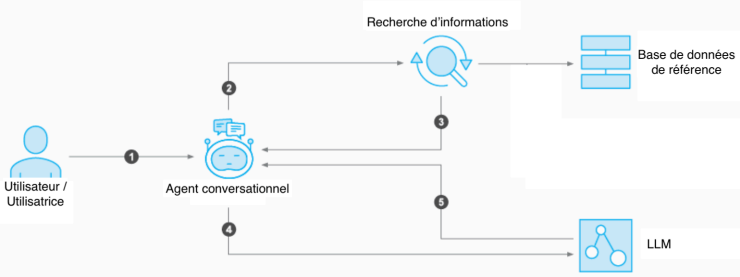
Agents conversationnels RAG pour l'évaluation des risques professionnels dans les unités du CNRS

Contexte du stage :

L'utilisation d'agents conversationnels a littéralement explosé ces dernières années, suite à la mise à disposition massive de grand modèles de langage (LLMs) tels que chatGPT. Bien que très performants, ces outils sont généralement mis à disposition sous forme d'applications web ou bien de fonctionnalités intégrées à des logiciels. Dans ce second cas, les logiciels communiquent quasi-systématiquement avec des serveurs distants pour les tâches d'inférence ou d'amélioration de l'expérience utilisateur. La plupart des LLMs librement disponibles ne permettent alors pas de garantir la confidentialité et souveraineté des données qui leurs sont accessibles. Ceci est un réel frein à leur adoption dans de nombreuses entreprises ou structures publiques. Dans ce contexte, une solution largement émergeante est l'utilisation de réseaux de neurones de type LLM embarqués sur des serveurs locaux qui questionnent des bases de données internes aux structures (voir schéma ci-dessous). On parle alors de stratégie RAG pour *Retreieval Augmented Generation,* ce qui sera au coeur de ce stage. Le stage consistera en effet à mettre en place un agent conversationnel de type RAG pour l'évaluation des risques professionnels dans les unités CNRS. Un ensemble de documents de référence internes au CNRS seront utilisés pour alimenter le LLM. Le livrable sera une IA conversationnelle intégrée à une interface web, capable d'assister les unités dans la rédaction et la mise à jour du DUERP (Document Unique d'Évaluation des Risques Professionnels). Cette IA s'appuiera sur :

- La réglementation en matière de santé, sécurité et prévention des risques professionnels.
- Des modèles d'évaluation issus des unités du CNRS.

Le tout devra être simple d'utilisation et répondre aux exigences de sécurité informatique de l'établissement.



Description technique du projet :

Le projet inclura l'étude et la mise en œuvre de différentes stratégies de "chunking" (division de texte en segments pertinents) afin d'optimiser la recherche dans la base de données.

Plusieurs modèles de langage (LLM) seront évalués, aussi bien pour la partie embedding (indexation vectorielle) que pour la génération de réponses (inférence).

Des méthodes automatiques d'évaluation seront utilises pour mesurer la performance du système sur les deux étapes clés : la recherche (retrieval) et la réponse finale.

Enfin, le *fine-tuning* des modèles d'embedding et de LLM avec Pytorch pourra être envisagé afin d'adapter le système à des jeux de données spécifiques du CNRS. La solution sera ensuite intégrée et

conteneurisée afin de pouvoir être déployée et utilisée sur un serveur distant. La solution sera codée en Python avec l'utilisation d'Ollama.

Compétences et apprentissages du stage :

- Utilisation d'un cluster HPC (soumission et suivi de jobs, optimisation des performances, exécution sur GPU, accès distant via SSH, etc)
- RAGs (Retrieval-Augmented Generation)
    - Stratégies de *chunking*
    - Gestion des *embeddings*
    - Évaluation des performances de différents LLMs
    - Fine-tuning des modèles
    - Chiffrement et déchiffrement de bases de données
- Développement web
    - Intégration du RAG dans un serveur web
    - Mise à disposition via une interface en ligne sécurisée

Pour la montée en compétence sur l'ensemble de ces domaines et l'implémentation des solutions, le ou la stagiaire sera accompagné(e) par deux ingénieurs calcul ayant déjà travaillé sur des problématiques similaires.

Mots-clés: RAG (Retrieval-Augmented Generation); LLM (Large Language Models); HPC (Calcul Haute Performance); Python / PyTorch

Enjeux du stage au CNRS :

Les unités de recherche ont l'obligation d'élaborer et de mettre à jour au moins une fois par an leur Document Unique d'Évaluation des Risques Professionnels (DUERP).

Cette obligation s'inscrit dans un cadre règlementaire dans le domaine de la prévention des risques professionnels, et repose sur les articles L 4121-1 à L 4121-5 du Code du travail. Ces articles imposent à l'employeur de prendre toute disposition pour préserver la santé mentale et physique de chacun de ses salariés et définissent les principes généraux de prévention.
L'élément central de cette démarche est le Document Unique d'Évaluation des Risques Professionnels (DUERP), instauré par le décret n°2001-1016 du 5 novembre 2001, en application de la directive européenne 89/391/CEE du 12 juin 1989 et transposé dans le code du travail aux articles R 4124-1 à R 4121-4.
Le DUERP est obligatoire dans toute structure employant au moins un agent ou salarié, y compris les unités de recherche ou d'appui qui compose le CNRS. Or dans les unités, cette démarche est particulièrement complexe, car les risques sont multiples et souvent évolutifs. L'accès aux DUERP requiert cependant un certain niveau de confidentialité dû aux informations qu'ils contiennent.

Pour aider les unités à faire face à cette obligation, le CNRS a conçu l'application EvRP il y a une dizaine d'année, afin d'aider les Assistants de Prévention des unités dans la démarche d'évaluation des risques.

Avec l'essor de l'intelligence artificielle générative, il est aujourd'hui envisageable de concevoir un outil spécifiquement dédié à l'évaluation des risques professionnels, adapté aux besoins des unités de recherche. L'objectif de cet outil serait de simplifier la démarche tout en guidant les utilisateurs tout au long du processus d'évaluation à l'aide d'un LLM. L'enjeu principal du stage consiste alors à amener une preuve de concept que la spécialisation d'un tel LLM sur l'évaluation des risques pourrait se faire à l'aide d'une stratégie RAG qui interroge le cadre réglementaire ainsi que les nombreux DUERP déjà établis avec l'outil EvRP aux seins des unités du CNRS.

Aspects pratiques :

Le stage sera hébergé à l'Institut de Mathématiques de Toulouse (IMT : Laurent Risser) en collaboration forte avec le mésocentre de calcul CALMIP (Alejandro Estana), la délégation Occitanie-Ouest du CNRS (DR14 : Stéphane Leblanc) et ANITI (Bhupen Dabholkar). L'IMT se situe à Toulouse, à proximité du métro B. La personne recrutée sera issue d'une formation ingénieur ou universitaire en sciences des données (informatique, mathématiques appliquées) avec le souhait fort de s'orienter vers un métier de type MLOps (Machine Learning Operations). Un profil de niveau M2 est idéalement recherché, mais une personne de niveau M1 particulièrement motivée pourrait être aussi recrutée. Le stage se déroulera enfin en présentiel sur une durée de 4 à 6 mois.

Vous pouvez envoyer votre CV, notes récentes et une lettre de motivation courte à Laurent Risser (lrisser@math.univ-toulouse.fr) pour candidater.