Offre n°2024-08529

Stage de recherche de 6 mois : SemWebRAG (Semantic Web Retrieval Augmented Generation)

Type de contrat : Stage

Niveau de diplôme exigé : Bac + 5 ou équivalent

Fonction : Stagiaire de la recherche

A propos du centre ou de la direction fonctionnelle

Le centre Inria d'Université Côte d'Azur regroupe 42 équipes de recherche et 9 services d'appui. Le personnel du centre (500 personnes environ) est composé de scientifiques de différentes nationalités, d'ingénieurs, de techniciens et d'administratifs. Les équipes sont principalement implantées sur les campus universitaires de Sophia Antipolis et Nice ainsi que Montpellier, en lien étroit avec les laboratoires et les établissements de recherche et d'enseignement supérieur (Université Côte d'Azur, CNRS, INRAE, INSERM ...), mais aussi avec les acteurs économiques du territoire.

Présent dans les domaines des neurosciences et biologie computationnelles, la science des données et la modélisation, le génie logiciel et la certification, ainsi que la robotique collaborative, le Centre Inria d'Université Côte d'Azur est un acteur majeur en termes d'excellence scientifique par les résultats obtenus et les collaborations tant au niveau européen qu'international.

Contexte et atouts du poste

Ce stage est une collaboration entre l'équipe WIMMICS (Université Côte d'Azur, Inria, CNRS, I3S) et l'entreprise Forgeron3. Il aura lieu dans les locaux de l'équipe WIMMICS3 à Sophia Antipolis, en collaboration avec Forgeron3 et sous la supervision de :

- Pierre Monnin (pierre.monnin@inria.fr –https://pmonnin.github.io)
- Fabien Gandon (fabien.gandon@inria.fr –http://fabien.info)

Wimmics (Web-Instrumented Man-Machine Interactions, Communities and Semantics) est une équipe de recherche commune à Université Côte d'Azur, Inria, CNRS, I3S, qui se situe à l'intersection de l'intelligence artificielle et du Web. Les membres de Wimmics travaillent sur des méthodes pour extraire, contrôler, interroger, valider, déduire, expliquer et interagir avec les connaissances.

Forgeron3 développe Marcus, une plateforme d'assistants intelligents collaboratifs, fondée sur des LLM open source tels que ceux de Meta et Mistral. L'objectif de Forgeron3 est de démocratiser l'IA pour les PME européennes, permettant aux collaborateurs de se concentrer sur l'essentiel pendant que les tâches répétitives sont gérées par des assistants intelligents, améliorant ainsi chaque interaction humaine.

Mission confiée

Contexte
L'apparition des grands modèles de langue (Large Language Models – LLM) a récemment accéléré l'usage et l'intégration avancée de l'Intelligence Artificielle en entreprise. Néanmoins, un élément bloquant reste la possibilité d'hallucinations, c'est-à-dire des réponses non fondées des LLMs. Ces hallucinations représentent un risque significatif limitant l'utilisation des LLMs pour des tâches en entreprise, et leur atténuation constitue donc une direction de recherche particulièrement active. Récemment, les concepts de Retrieval Augmented Generation (RAG) [1] et GraphRAG [2] ont été proposés et ont pour objectif d'enrichir le prompt du LLM par des éléments contextuels adéquats extraits de documents disponibles dans une base de documents. Ces techniques ont pu atténuer les hallucinations, mais elles mettent en avant deux nouveaux défis :

1. Le besoin d'indexer et de récupérer correctement des éléments contextuels adéquats
2. Et donc, le besoin pour les LLMs d'avoir accès à du vocabulaire / expressions / définitions métier n'étant pas nécessairement vu à l'entraînement.

Les graphes de connaissances et les ontologies du Web Sémantique ont été mentionnés comme source de connaissances permettant de compléter les LLMs et d'atténuer leurs hallucinations [3,4]. Notamment, les graphes et ontologies publiquement disponibles comme Wikidata1 ou LOV2 constituent des répertoires étendus et donc particulièrement intéressants pour fournir le vocabulaire, les définitions et le contexte métier nécessaires pour indexer et retrouver les documents, ainsi qu'enrichir directement les prompts.

Bibliographie

1. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. https://arxiv.org/pdf/2312.10997
2. Larson, J. & Truitt, S. GraphRAG: Unlocking LLM discovery on narrative private data. https://www.microsoft.com/en-us/research/blog/graphrag-unlocking-llm-discovery-on-narrative-private-data/
3. Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). Unifying large language models and knowledge graphs: A roadmap. IEEE Transactions on Knowledge and Data Engineering. https://ieeexplore.ieee.org/document/10387715
4. Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, Damien Graux. Large Language Models and Knowledge Graphs: Opportunities and Challenges. TGDK 1(1): 2:1-2:38 (2023) https://doi.org/10.4230/TGDK.1.1.2

Principales activités

Dans le cadre de ce stage, pour répondre aux deux problématiques identifiées, nous proposons d'étudier la création d'un pipeline de Semantic Web Retrieval Augmented Generation. En particulier, le stage comportera les tâches suivantes :

1. État de l'art et montée en compétences sur les LLMs, RAG, GraphRAG, Web Sémantique
2. Design d'un pipeline extensible de SemWebRAG étendant les concepts de RAG et GraphRAG nécessitant de sélectionner de manière flexible et adéquate les ressources du Web Sémantique pour
    1. Améliorer l'indexation des documents en les annotant sémantiquement.
    2. Améliorer la récupération de documents adéquats par une meilleure interprétation des prompts et / ou un processus de récupération amélioré.
    3. Enrichir les prompts avec davantage d'éléments contextuels récupérés du Web Sémantique, en plus des éléments issus des documents.
3. Expérimentation et évaluation des résultats.

Compétences

Vous possédez les connaissances suivantes :

- Programmation en Python
- Machine Learning / Deep Learning, notamment avec des frameworks comme PyTorch ou Tensorflow
- Une connaissance des LLMs, des frameworks comme LangChain, et du (Graph)RAG serait un avantage.
- Une connaissance du Web Sémantique (RDF, RDFS, OWL, SPARQL, graphes de connaissances et ontologies) serait un avantage.
- Capacité de lecture et de rédaction en anglais

Avantages

- Restauration subventionnée
- Transports publics remboursés partiellement
- Congés: 7 semaines de congés annuels + 10 jours de RTT (base temps plein) + possibilité d'autorisations d'absence exceptionnelle (ex : enfants malades, déménagement)
- Possibilité de télétravail (après 6 mois d'ancienneté) et aménagement du temps de travail
- Équipements professionnels à disposition (visioconférence, prêts de matériels informatiques, etc.)
- Prestations sociales, culturelles et sportives (Association de gestion des œuvres sociales d'Inria)
- Accès à la formation professionnelle
- Participation mutuelle (sous conditions)

Rémunération

Gratification selon temps de présence.

Informations générales

- **Thème/Domaine :** Représentation et traitement des données et des connaissances
- **Ville :** Sophia Antipolis
- **Centre Inria :** Centre Inria d'Université Côte d'Azur
- **Date de prise de fonction souhaitée :** 2025-03-01
- **Durée de contrat :** 6 mois
- **Date limite pour postuler :** 2025-02-28

Contacts

- **Equipe Inria :** WIMMICS
- **Recruteur :**
  Monnin Pierre / pierre.monnin@inria.fr

A propos d'Inria

Inria est l'institut national de recherche dédié aux sciences et technologies du numérique. Il emploie 2600 personnes. Ses 215 équipes-projets agiles, en général communes avec des partenaires académiques, impliquent plus de 3900 scientifiques pour relever les défis du numérique, souvent à l'interface d'autres disciplines. L'institut fait appel à de nombreux talents dans plus d'une quarantaine de métiers différents. 900 personnels d'appui à la recherche et à l'innovation contribuent à faire émerger et grandir des projets scientifiques ou entrepreneuriaux qui impactent le monde. Inria travaille avec de nombreuses entreprises et a accompagné la création de plus de 200 start-up. L'institut s'efforce ainsi de répondre aux enjeux de la transformation numérique de la science, de la société et de l'économie.

L'essentiel pour réussir

Vous étudiez en M2 / dernière année d'école d'ingénieur, avec une spécialité en informatique ou en mathématiques appliquées. Vous êtes curieux, aimez apprendre, être confronté à des défis, expérimenter et découvrir par vous-même.

Attention: Les candidatures doivent être déposées en ligne sur le site Inria. Le traitement des candidatures adressées par d'autres canaux n'est pas garanti.

Consignes pour postuler

Les candidatures doivent être déposées en ligne sur le site Inria. Le traitement des candidatures adressées par d'autres canaux n'est pas garanti.

Sécurité défense :
Ce poste est susceptible d'être affecté dans une zone à régime restrictif (ZRR), telle que définie dans le décret n°2011-1425 relatif à la protection du potentiel scientifique et technique de la nation (PPST). L'autorisation d'accès à une zone est délivrée par le chef d'établissement, après avis ministériel favorable, tel que défini dans l'arrêté du 03 juillet 2012, relatif à la PPST. Un avis ministériel défavorable pour un poste affecté dans une ZRR aurait pour conséquence l'annulation du recrutement.

Politique de recrutement :
Dans le cadre de sa politique diversité, tous les postes Inria sont accessibles aux personnes en situation de handicap.