EXPLORATION DE L'UTILISATION DES GRANDS MODÈLES DE LANGAGE (LLMS) AU CENTRE DE DONNÉES ASTRONOMIQUES DE STRASBOURG

DIRECTEUR DE THÈSE : PIERRE OCVIRK (HDR), CO-DIR PAR SÉBASTIEN DERRIÈRE.
OBSERVATOIRE ASTRONOMIQUE DE STRASBOURG, 11 RUE DE L'UNIVERSITÉ, 67000 STRASBOURG
TEL : 03 68 85 24 40 ; E-MAIL : pierre.ocvirk@astro.unistra.fr

RÉSUMÉ: Ce projet de thèse vise à explorer et développer des applications de Grands Modèles de Langage (LLMs) pour les workflow du Centre de Données astronomiques de Strasbourg (CDS). Face à l'évolution rapide des technologies d'Intelligence Artificielle (IA), nous proposons d'identifier et d'implémenter des cas d'usage s'intégrant aux processus existants tout en améliorant l'efficacité du traitement des données astronomiques. L'accent sera mis sur l'utilisation de modèles et de framework open source, pour automatiser une sélection de tâches liées à l'indexation de la littérature astrophysique et la curation des données au CDS.

CONTEXTE: Le CDS gère un volume considérable de données astronomiques nécessitant un traitement manuel important pour les services VizieR et SIMBAD. Les avancées dans le domaine des LLMs sont potentiellement porteuses d'opportunités pour automatiser certaines tâches, avec un écart de performance entre modèles commerciaux et open source de moins en moins important (ex: DeepSeek). On peut aujourd'hui envisager de développer et d'exploiter des agents IA 'maison' (à l'aide de OpenManus, LMStudio-python ou ollama) qui satisfont aux besoins de confidentialité propre à l'activité du CDS. Ce projet s'inscrit dans une démarche d'innovation prudente, visant à enrichir les workflow CDS sans les perturber. L'idée est de tenter d'automatiser certaines tâches simples et répétitives, afin de libérer du temps 'humain' à investir sur les tâches plus complexes. L'expertise reconnue du CDS et les données disponibles en interne en font le lieu idéal pour mener à bien ce projet.

OBJECTIFS: Évaluer une collection de LLMs open source (ex: LLama, Gemma3, DeepSeek) pour le traitement des références astronomiques, notamment sur les tâches suivantes: (i) identification de noms d'objets et extraction de quantités physiques (ex: magnitudes, coordonnées et systèmes, tailles) à partir d'articles ou de Readme, (ii) génération de Readme et documentation (ex: explication des colonnes et détermination des unités), (iii) extraction d'informations de sources données diverses, header fits, noms de fichiers, notamment pour identifier les objets/instruments/télescopes, (iv) aide à la cross-identification par des LLMs multi-modaux image et texte, (v) détection/résolution d'erreurs dans les catalogues existants, (vi) requête/découverte de données en langage naturel.

MÉTHODOLOGIE: Pour chaque tâche identifiée, la stratégie sera: (i) définir précisément la fonctionnalité désirée, (ii) construire un jeu de données de référence pour l'évaluation, (iii) effectuer des benchmark avec différents LLMs en itérant sur les prompts, (iv) analyser les résultats et conclure sur l'applicabilité. L'objectif idéal de cette exploration sera d'identifier, parmi les diverses tâches évaluées, deux cas d'usage particulièrement prometteurs qui fonctionnent efficacement. La dernière année du doctorat se concentrera sur ces 2 cas prioritaires pour assurer la pérennisation des outils développés, leur documentation complète et leur déploiement en production. Les documentalistes et astronomes seront impliqués en continu pour valider la pertinence des résultats et des outils en développement, garantissant ainsi que les solutions répondent aux besoins réels du CDS.