Sujet de stage — M2 / dernière année d'école d'ingénieur

Conception d'un pipeline agentique à base de LLM pour la détection et l'annotation de citations dans des textes historiques

Domaines : Intelligence artificielle, LLM, systèmes agentiques, NLP, recherche d'information, humanités numériques
Laboratoires : LIRIS/CIHAM
Durée : 5 à 6 mois
Niveau : Master 2 informatique ou dernière année d'école d'ingénieur

Contexte

L'identification des citations et réemplois textuels constitue une tâche importante pour l'analyse et l'édition scientifique de textes historiques. Dans les corpus médiévaux, cette tâche est particulièrement complexe : une source peut être citée littéralement, mais également abrégée, modifiée, paraphrasée ou simplement évoquée par une allusion.

Le projet **VulgatIA** financé par la FIL a pour objectif de développer un outil d'assistance capable d'identifier automatiquement les citations bibliques dans des textes latins médiévaux et d'en retrouver précisément la source. Le projet est mené conjointement par le **CIHAM**, laboratoire spécialisé notamment dans l'étude des mondes médiévaux, et le **LIRIS**, laboratoire de recherche en informatique.

Les partenaires disposent déjà de ressources importantes : deux versions de la Vulgate encodées en XML-TEI, dont une version lemmatisée, des scripts Python de détection de citations ainsi que plusieurs corpus de textes latins médiévaux permettant de développer et d'évaluer les méthodes proposées.

L'enjeu informatique du projet consiste à dépasser une approche reposant sur un modèle unique pour mettre en œuvre une **architecture agentique**, dans laquelle plusieurs agents spécialisés coopèrent et utilisent dynamiquement différents outils : recherche dans un corpus de référence, méthodes de similarité textuelle, scripts existants, modèles de langage, validation ou normalisation des résultats.

Objectif du stage

L'objectif du stage sera de concevoir, implémenter et évaluer un prototype de pipeline agentique permettant de détecter et d'identifier des citations ou allusions bibliques dans des textes latins.

Une attention particulière sera portée aux cas difficiles qui échappent aux méthodes classiques de rapprochement de chaînes de caractères : citations modifiées, citations partielles, paraphrases et allusions.

Le système devra combiner les capacités de raisonnement et d'analyse contextuelle des LLM avec des méthodes déterministes ou algorithmiques plus classiques afin de produire des résultats à la fois performants, explicables et structurés.

Le ou la stagiaire participera à la définition de l'architecture du système et travaillera notamment sur les aspects suivants :

- prise en main et préparation des corpus de textes latins et de la Vulgate encodée en XML-TEI ;
- étude et intégration des outils de détection de réemploi textuel existants dans le projet ;
- conception d'une architecture composée de plusieurs agents spécialisés : détection d'un passage candidat, recherche de sources potentielles, comparaison des passages, désambiguïsation et validation ;
- définition des outils accessibles aux agents et des mécanismes d'orchestration permettant de sélectionner dynamiquement la méthode adaptée à chaque cas ;

- expérimentation de différentes stratégies fondées sur les LLM : prompting, sorties structurées, tool calling, retrieval, agents de vérification, etc. ;
- développement d'une méthode permettant d'associer une citation détectée à une référence biblique structurée livre / chapitre / verset ;
- mise en place d'un protocole d'évaluation et comparaison avec des méthodes non-LLM ou avec les scripts existants ;
- analyse des erreurs et des limites du système.

Dans un second temps, le prototype pourra être relié à la chaîne éditoriale du projet afin de générer automatiquement des annotations exploitables dans des documents XML-TEI. Cette production d'un texte enrichi constitue l'une des spécificités de VulgatIA.

Questions de recherche

Le stage permettra notamment d'étudier plusieurs questions actuelles en intelligence artificielle :

1. **Comment combiner efficacement LLM et algorithmes spécialisés ?** Un LLM est-il pertinent pour toutes les étapes de la chaîne ou certaines opérations doivent-elles rester déterministes ?
2. **Comment organiser la coopération entre plusieurs agents ?** On pourra notamment comparer une chaîne fixe d'agents à des approches dans lesquelles le système sélectionne dynamiquement les outils et agents nécessaires.
3. **Comment détecter une paraphrase ou une allusion ?** Il s'agira d'étudier les limites des méthodes fondées uniquement sur la similarité lexicale et l'apport de représentations sémantiques ou des LLM.
4. **Comment vérifier les réponses d'un LLM ?** Le corpus biblique constituant ici une base de référence fermée, le stage offre un cadre particulièrement intéressant pour développer des mécanismes de vérification et limiter les hallucinations.

Technologies envisagées

Le développement sera principalement réalisé en **Python**. Selon les orientations retenues pendant le stage, pourront notamment être utilisés des modèles de langage open source ou accessibles par API, des frameworks d'orchestration d'agents ou des implémentations spécifiques, des techniques de recherche sémantique et d'embeddings, des outils NLP ainsi que des bibliothèques de manipulation **XML/TEI**.
Le choix précis des technologies fera partie du travail : l'objectif n'est pas de dépendre d'un framework particulier, mais d'identifier l'architecture la mieux adaptée au problème scientifique.

Profil recherché

Le stage s'adresse à un étudiant ou une étudiante de **Master 2 en informatique, intelligence artificielle ou traitement automatique des langues**, ou de **dernière année d'école d'ingénieur**. Une bonne maîtrise de **Python** est attendue, ainsi qu'un intérêt marqué pour l'intelligence artificielle et les modèles de langage.
Des connaissances dans un ou plusieurs des domaines suivants seront appréciées : NLP, LLM et modèles Transformers, systèmes agentiques, recherche d'information, embeddings et recherche vectorielle, apprentissage automatique, traitement de données structurées ou XML.
Aucune connaissance préalable du latin ou de l'histoire médiévale n'est nécessaire. Le stage sera mené en collaboration avec les chercheurs spécialistes des corpus et du domaine.

Résultats attendus

À l'issue du stage, les principaux livrables seront un prototype fonctionnel et documenté du pipeline, une évaluation expérimentale sur les corpus du projet et une analyse comparative des différentes stratégies explorées.
Le code développé dans le cadre de VulgatIA a vocation à être diffusé en **open source** et les résultats pourront donner lieu à une présentation dans un séminaire ou une conférence scientifique. La diffusion du code, l'évaluation sur corpus et la communication scientifique font explicitement partie des objectifs du projet financé.

Références

David A. Smith, et al.. *Infectious texts: Modeling text reuse in nineteenth-century newspapers*. In 2013 IEEE International Conference on Big Data, pages 86–94, 2013.

Marco Büchler, et al. *Towards a Historical Text Re-use Detection*. In Text Mining: From Ontology Learning to Automated Text Processing Applications, pages 221–238, 2014.

Chao Zhou, et al. *Paraphrase Identification With Deep Learning: A Review of Datasets and Methods*. IEEE Access, 13:65797– 65822, 2025.

Tom B. Brown, et al. 2020. *Language models are few-shot learners*. In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20). Curran Associates Inc., Red Hook, NY, USA, Article 159, 1877–1901.

Yao, Shunyu et al. *ReAct: Synergizing Reasoning and Acting in Language Models.* ICLR, 2023.

Timo Schick, et al. 2023. Toolformer: language models can teach themselves to use tools. In Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23).

Encadrants

Marjorie Burghart, CR CNRS, CIHAM – marjorie.burghart@cnrs.fr
Ludovic Moncla, maître de conférences, INSA Lyon, LIRIS UMR CNRS – ludovic.moncla@insa-lyon.fr

Modalités

Rémunération : environ 650€ par mois
Lieu : Laboratoire LIRIS INSA Lyon et laboratoire CIHAM CNRS
Date de début : février/mars 2027
Durée : 5 à 6 mois

Candidature : Envoyer un mail présentant votre parcours, vos motivations ainsi que votre CV et vos derniers relevés de notes à : marjorie.burghart@cnrs.fr et ludovic.moncla@insa-lyon.fr
Date limite de candidature : 10 décembre 2026 (traitement des candidatures au fil de l'eau)