Offre de Stage : Automatisation de Méta-Analyse bibliographique par l'Intelligence Artificielle (IA) et les Large Language Models (LLMs)

Durée : 6 mois

Lieu : Tours, Faculté des Sciences et Techniques de Grandmont

Niveau requis : Master 2 ou équivalent en informatique, bio-informatique, data science ou domaine connexe.

Encadrement : Alexandre Chanson (LIFAT), Stéphane Boyer (IRBI), Nicolas Labroche (LIFAT) (prenom.nom@univ-tours.fr)

Contexte et Objectifs du Stage

La méta-analyse est une tâche d'analyse de la littérature scientifique visant à collecter l'ensemble des études portant sur un même phénomène (p. ex. effet d'un herbicide sur le système nerveux d'un insecte), puis d'en extraire les éléments qualitatif et quantitatifs permettant la réalisation d'une étude statistique s'appuyant sur l'ensemble des résultats collectés.
Cette tâche tout comme l'analyse systématique de la littérature repose sur la lecture et l'extraction d'information d'un grand nombre de textes scientifiques. Rendant ces tâches longues et complexes.
L'émergence des modèles de langage massif (LLM) a participé à démocratiser l'usage de l'intelligence artificielle. Elle a permis à tout un chacun d'interagir et d'exploiter l'information textuelle via une interface en langue naturelle ne nécessitant aucune connaissance préalable. Néanmoins ces outils comportent des risques : quand ils sont confrontés à une question portant sur une connaissance précise, les LLMs tendent à 'halluciner' présentant comme réponse des informations complètement fausses [4]. Cette phénomène tend à disparaître avec les modèles désormais entraînés à répondre qu'ils ne disposent simplement pas d'une information plutôt que de l'inventer [5]. Une des techniques visant à pallier ce manque de 'connaissance' est d'extraire l'information d'un document source et de la fournir au LLM en plus de la requête originelle. Dans sa version la plus simple l'utilisateur lui-même peut identifier un texte source et le fournir au modèle de langue (e.g. [2]). Un process plus formel et complexe vise à construire un pipeline ou l'information pertinente de réponse à une question est automatiquement localisée et fournie au LLM. Ces méthodes dites de RAG (Retrival Augmented Generation) permettent une plus grande flexibilité et puisque le système détermine de façon autonome les parties de documents nécessaire pour compléter la requête de l'utilisateur il permet de puiser dans des milliers de documents sans intervention préalable de l'utilisateur.

Ce stage de recherche se propose d'explorer le potentiel des LLMs, notamment en combinaison avec des techniques de RAG, pour automatiser et améliorer certaines tâches liées à la méta-analyse. Nous nous baserons sur une méta-analyse coordonnée par Stéphane Boyer et

portant sur le thème des échantillonnages ADN dits ‘non-invasifs’ pour l’étude des animaux [6]. Plus précisément, nous allons nous concentrer sur :

- L'extraction des données : une fois les études pertinentes identifiées, les LLMs peuvent être utilisés pour extraire les données nécessaires à la méta-analyse, en particulier 1) la méthodologie employée et la nature des échantillons ADN collectés, 2) le caractère invasif ou non des prélèvements réalisés, et 3) le cas échéant le type ‘d’erreur’ réalisé par les auteurs dans leur utilisation du terme ‘non-invasive DNA sampling’ [6].
- La synthèse des résultats : les LLMs peuvent être utilisés pour générer des résumés synthétiques des résultats de la méta-analyse, en langage naturel, et pour identifier les tendances et les conclusions principales d’une étude. L’analyse des 380 articles scientifiques étudiés en 2022 (articles publiés entre 2013 et 2018) permettra de comparer les résultats obtenus par l’approche manuelle à ceux produits par les LLMs, et d’affiner le protocole afin d’obtenir les résultats les plus précis possible.
- La mise à jour de la méta-analyse sera ensuite réalisée en appliquant notre meilleur protocole LLM sur un nouveau lot d’articles, publiés entre 2019 et 2024. Cette mise à jour pourra être soumise pour publication dans un journal à comité de lecture.

[1] Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., Dou, Z., & Wen, J. (2023). Large Language Models for Information Retrieval: A Survey. ArXiv, abs/2308.07107.
[2] https://chatgpt.com/share/671fb24d-dec8-8012-9857-760539b1390f
[3] Yun, H., Pogrebitskiy, D., Marshall, I.J., & Wallace, B.C. (2024). Automatically Extracting Numerical Results from Randomized Controlled Trials with Large Language Models. ArXiv, abs/2405.01686. https://arxiv.org/pdf/2405.01686
[4] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ArXiv, abs/2311.05232.
[5] Tonmoy, S.M., Zaman, S.M., Jain, V., Rani, A., Rawte, V., Chadha, A., & Das, A. (2024). A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models. ArXiv, abs/2401.01313.
[6] Lefort, M. C., Cruickshank, R. H., Descovich, K., Adams, N. J., Barun, A., Emami-Khoyi, A., ... & Boyer, S. (2022). Blood, sweat and tears: a review of non-invasive DNA sampling. Peer Community Journal, 2, e16. https://peercommunityjournal.org/articles/10.24072/pcjournal.98/