Université de Rennes

Direction de la Recherche et de l'Innovation

Appel à projets - Année 2025

Contrats doctoraux «Ordinaires»

ECOLE DOCTORALE : MATISSE **POLE DE RECHERCHE : Mathématique-Numérique**

Unité de recherche d'accueil

Unité de Recherche : IRISA

Equipe : LACODAM

Encadrement

Directeur de thèse : Sébastien Ferré **Unité de recherche** : IRISA

Employeur/Établissement d'origine : Université de Rennes

HDR : X oui ☐ non **Nb de thèses en cours** : 1

Mail : ferre@irisa.fr **Tél.** : 02 99 84 75 70

Co-directeur de thèse : Peggy Cellier **Unité de recherche** : IRISA

Employeur/Établissement d'origine : INSA Rennes

HDR : X oui ☐ non **Nb de thèses en cours** : 1

Mail : peggy.cellier@irisa.fr **Tél.** : 02 99 84 22 84

Sujet

Titre : LLM pour l'interrogation de données du web sémantique

Mots clés : LLM, web sémantique, graphe de connaissances, question answerin

Cotutelle internationale (oui/non) : non **Si oui, préciser l'établissement et le pays de rattachement** :

Descriptif (15-30 lignes) :
Contexte socioéconomique et scientifique ; Les hypothèses et questions posées ; Les grandes étapes de la thèse et démarche ; Approches méthodologiques et techniques envisagées

Les grands modèles de langues (LLM) tels que GPT [1], Mistral [2], ou encore Llama [3] connaissent depuis quelques années un engouement justifié. Ils ont permis de faire des avancées considérables en traitement de la langue (traduction, résumé de textes, question answering, etc.) mais ils ont aussi eu des répercussions importantes dans d'autres domaines comme le traitement d'images par exemple. Malgré des résultats spectaculaires, les LLM ont certaines limites bien connues, en particulier le phénomène d'hallucination, c'est-à-dire le fait que le LLM va inventer une réponse plausible mais fausse à une question, ou encore des difficultés à mener un raisonnement [4]. Plusieurs équipes de recherche s'intéressent à essayer de dépasser ces limites. Par exemple, pour contrer l'effet hallucinatoire plusieurs approches ont été proposées. Une première méthode est de fine-tuner, c'est à dire de spécialiser le modèle LLM pour un certain type de données/documents/domaines afin de le rendre plus performant pour ce domaine. D'autres approches nommées RAG (Retrieval-Augmented Generation) fournissent au LLM en plus de sa connaissance interne, une connaissance externe (documents textuels ou données) le rendant plus pertinent pour répondre à des questions sur ce texte. D'autres approches, comme Chameleon [5] ou ReAct [6], utilisent et combinent des outils spécialisés spécifiques (ex., bases de données externes) pour répondre à des questions.

Même si ces approches sont prometteuses, leurs taux d'échec à répondre à des questions restent élevés comme le montre cette étude [4]. Dans l'équipe Lacodam, nous développons des systèmes pour interroger des données de façon fiable [7,8]. En particulier le système Sparklis [8] permet d'interroger des données au format du web sémantique (graphes de connaissances) dans un langage semi-structuré. L'avantage de cette approche par rapport aux LLM est la fiabilité du résultat ainsi que la possibilité d'expliciter comment l'information a été obtenue et de donner la ou les sources utilisées. Ce dernier point est particulièrement crucial et fait partie des obligations du AI Act [9] voté par l'Union européenne.

Afin de tirer parti des deux approches à savoir, d'un côté la flexibilité et les connaissances générales inhérentes aux LLM et de l'autre la fiabilité et la capacité à tracer le résultat des approches d'interrogation de données, nous souhaitons développer une approche hybride. Des premiers tests d'utilisation des LLM pour générer directement des requêtes en langage formel (SPARQL [14]) afin d'interroger ensuite une base de connaissances ont montré de nombreuses limites [10,11]. Nous souhaitons travailler sur ces limites en proposant de piloter à l'aide d'un LLM un système d'interrogation de base de connaissances, tel que Sparklis.

Objectifs de la thèse

Dans un premier temps l'étudiante ou l'étudiant devra établir une bibliographie comprenant les notions de base du web sémantique [15,14], les méthodes existantes d'interrogation de ces bases de connaissances en langue naturelle [12] et les approches d'hybridation entre graphes de connaissances et LLM [16]. On étudiera aussi des approches LLM intégrant des plugins, telles que Chameleon [5] ou ReAct [6]. Ensuite, nous souhaitons mettre en place une plateforme de comparaison d'approches d'interrogation de bases de connaissances sur le modèle de ToolQA [4] qui servira à mesurer les différentes approches existantes ou développées pendant le stage. Enfin nous souhaitons étudier l'utilisation d'un LLM pour le pilotage pas à pas d'un outil d'interrogation de données en s'inspirant du principe de Chain of Thought [13]. Dans un premier temps les outils envisagés sont Sparklis pour l'interrogation de données.

Bibliographie

[1] https://openai.com/index/gpt-4/ [2] https://mistral.ai/fr/ [3] https://llama.meta.com/

[4] Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2024. ToolQA: a dataset for LLM question answering with external tools. In Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23). Curran Associates Inc., Red Hook, NY, USA, Article 2180, 50117–50143.

[5] Lu, Pan, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu and Jianfeng Gao. "Chameleon: Plug-and-Play Compositional Reasoning with Large Language Models." International Conference on Neural Information Processing Systems (NeurIPS '23).

[6] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, and Y. Cao. ReAct: Synergizing reasoning and acting in language models. In The Eleventh International Conference on Learning Representations, 2023.

[7] Ferré, Sébastien. "Camelis: a logical information system to organise and browse a collection of documents." _International Journal of General Systems_ 38 (2009): 379 - 403.

[8] Sébastien Ferré. 2017. Sparklis: An expressive query builder for SPARQL endpoints with guidance in natural language. Semant. web 8, 3 (2017), 405–418. [9] AI Act. https://artificialintelligenceact.eu/fr/

[10] Alexandra Padonou, Peggy Cellier and Sébastien Ferré, Étude de l'utilisation des modèles de langages pour l'interrogation en langue naturelle des graphes de connaissances. EGC 2024.

[11] Jens Lehmann, Sébastien Ferré, Sahar Vahdati: Language Models as Controlled Natural Language Semantic Parsers for Knowledge Graph Question Answering. ECAI 2023: 1348-1356

[12] Affolter, Katrin, Kurt Stockinger, and Abraham Bernstein. "A comparative survey of recent natural language interfaces for databases." The VLDB Journal 28.5 (2019): 793-819.

[13] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22). Curran Associates Inc., Red Hook, NY, USA, Article 1800, 24824–24837. [14] SPARQL 1.1 Query Language. W3C, 2013. https://www.w3.org/TR/sparql11-query/ [15] RDF 1.1 Primer. W3C, 2014. https://www.w3.org/TR/rdf11-primer/

[16] Pan, Shirui, et al. "Unifying large language models and knowledge graphs: A roadmap." IEEE Transactions on Knowledge and Data Engineering (2024).