Offer #2024-08529

6-month research internship: SemWebRAG (Semantic Web Retrieval Augmented Generation)

Contract type : Internship

Level of qualifications required : Graduate degree or equivalent

Fonction : Internship Research

About the research centre or Inria department

The Inria center at Université Côte d'Azur includes 42 research teams and 9 support services. The center's staff (about 500 people) is made up of scientists of different nationalities, engineers, technicians and administrative staff. The teams are mainly located on the university campuses of Sophia Antipolis and Nice as well as Montpellier, in close collaboration with research and higher education laboratories and establishments (Université Côte d'Azur, CNRS, INRAE, INSERM ...), but also with the regional economic players.

With a presence in the fields of computational neuroscience and biology, data science and modeling, software engineering and certification, as well as collaborative robotics, the Inria Centre at Université Côte d'Azur is a major player in terms of scientific excellence through its results and collaborations at both European and international levels.

Context

This internship is a collaboration between the WIMMICS team (Université Côte d'Azur, Inria, CNRS, I3S) and the Forgeron3 company. It will take place on the premises of the WIMMICS team in Sophia Antipolis, in collaboration with Forgeron3 and under the supervision of:

- 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) is a joint research team at Université Côte d'Azur, Inria, CNRS, I3S, whose research lies at the intersection of artificial intelligence and the Web. Wimmics members work on methods to extract, control, query, validate, infer, explain and interact with knowledge.

Forgeron3 develops Marcus, a platform of collaborative intelligent assistants, based on open source LLMs such as those of Meta and Mistral. Forgeron3's goal is to democratize AI for European SMEs, allowing employees to focus on what matters while repetitive tasks are handled by intelligent assistants, improving every human interaction.

Assignment

Context

The emergence of Large Language Models (LLMs) has recently accelerated the use and advanced integration of Artificial Intelligence in business. However, a major issue lies in the possibility of hallucinations, i.e. unfounded responses from LLMs. These hallucinations represent a significant risk limiting the use of LLMs for business tasks, and their mitigation is therefore a particularly active research direction. Recently, the concepts of Retrieval Augmented Generation (RAG) [1] and GraphRAG [2] have been proposed and aim to enrich LLM prompts with adequate contextual elements extracted from documents available in a document database. These techniques have been able to mitigate hallucinations, but they highlight two new challenges:

1. The need to correctly index and retrieve adequate contextual elements
2. And therefore, the need for LLMs to have access to business vocabulary / expressions / definitions not necessarily seen in training.

Knowledge graphs and ontologies of the Semantic Web have been mentioned as a source of knowledge to complement LLMs and mitigate their hallucinations [3,4]. In particular, publicly available graphs and ontologies such as Wikidata4 or LOV5 constitute extensive and therefore particularly interesting directories to provide the vocabulary, definitions and business context necessary to index and retrieve documents, as well as directly enrich prompts.

Bibliography

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

Main activities

In this internship, to address the two identified issues, we propose to study the creation of a Semantic Web Retrieval Augmented Generation pipeline. In particular, the internship will include the following tasks:

1. State of the art and skills development on LLMs, RAG, GraphRAG, Semantic Web
2. Design of an extensible SemWebRAG pipeline extending the concepts of RAG and GraphRAG requiring
   flexible and adequate selection of Semantic Web resources to
   1. Improve document indexing by annotating them semantically.
   2. Improve the retrieval of adequate documents by better interpretation of prompts and / or an improved retrieval process.
   3. Enrich prompts with more contextual elements retrieved from the Semantic Web, in addition to
      elements from documents.
3. Experiment and evaluation of results.

Skills

You are proficient in:

- Python programming
- Machine Learning / Deep Learning, especially with frameworks like PyTorch or Tensorflow
- Knowledge of LLMs, frameworks like LangChain, and (Graph)RAG would be appreciated.
- Knowledge of the Semantic Web (RDF, RDFS, OWL, SPARQL, knowledge graphs and ontologies) would be appreciated.
- Ability to read and write in English

Benefits package

- Subsidized meals
- Partial reimbursement of public transport costs
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.)
- Possibility of teleworking and flexible organization of working hours
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training
- Contribution to mutual insurance (subject to conditions)

Remuneration

Traineeship grant depending on attendance hours.

General Information

- **Theme/Domain :** Data and Knowledge Representation and Processing
- **Town/city :** Sophia Antipolis
- **Inria Center :** Centre Inria d'Université Côte d'Azur
- **Starting date :** 2025-03-01
- **Duration of contract :** 6 months
- **Deadline to apply :** 2025-02-28

Contacts

- **Inria Team :** WIMMICS
- **Recruiter :**
  Monnin Pierre / pierre.monnin@inria.fr

About Inria

Inria is the French national research institute dedicated to digital science and technology. It employs 2,600 people. Its 200 agile project teams, generally run jointly with academic partners, include more than 3,500 scientists and engineers working to meet the challenges of digital technology, often at the interface with other disciplines. The Institute also employs numerous talents in over forty different professions. 900 research support staff contribute to the preparation and development of scientific and entrepreneurial projects that have a worldwide impact.

The keys to success

You are studying in Master Year 2 / final year of engineering school, with a specialty in computer science or applied mathematics. You are curious, like to learn, face challenges, experiment and discover by yourself.

Warning : you must enter your e-mail address in order to save your application to Inria. Applications must be submitted online on the Inria website. Processing of applications sent from other channels is not guaranteed.

Instruction to apply

Applications must be submitted online on the Inria website. Collecting applications by other channels is not guaranteed.

Defence Security :
This position is likely to be situated in a restricted area (ZRR), as defined in Decree No. 2011-1425 relating to the protection of national scientific and technical potential (PPST).Authorisation to enter an area is granted by the director of the unit, following a favourable Ministerial decision, as defined in the decree of 3 July 2012 relating to the PPST. An unfavourable Ministerial decision in respect of a position situated in a ZRR would result in the cancellation of the appointment.

Recruitment Policy :
As part of its diversity policy, all Inria positions are accessible to people with disabilities.