Offer #2024-08406

PhD Position F/M LLM4Code and SopraSteria: Software migration and modernization with LLMs

Contract type : Fixed-term contract

Level of qualifications required : Graduate degree or equivalent

Fonction : PhD Position

About the research centre or Inria department

The Inria Centre at Rennes University is one of Inria's eight centres and has more than thirty research teams. The Inria centre is a major and recognized player in the field of digital sciences. It is at the heart of a rich R&D and innovation ecosystem: highly innovative PMEs, large industrial groups competitiveness clusters, research and higher education players, laboratories of excellence, technological research institute, etc

Context

The PhD subject (thèse CIFRE) is a collaboration between SopraSteria (in Nantes) and DiverSE Inria research team (in Rennes).

The candidate will be employee of SopraSteria and spend part of the time in SopraSteria and in DiverSE.

The work is also part of a Défi Inria LLM4Code "Reliable and productive Code Assistants based on large language models" with more than 10 research teams working on several aspects of LLMs and code. Hence the candidate will have the opportunity to collaborate with numerous researchers and experts, as well as to leverage computational infrastructure and the SoftwareHeritage project.

More details here: https://project.inria.fr/llm4code/

Generative AI, in particular the recent Large Language Models (LLMs), show great promise for software developments. Specialized models are now able to perform an impressive variety of programming tasks: solving programming exercises, assisting software developers, or even generating mechanized proofs. Yet, many challenges still need to be addressed to build reliable and productive LLM-based coding assistants: improving the quality of the generated code, increasing the developers' confidence in the generated code, enabling interaction with other software development tools (verification, test), and providing new capabilities (automated migration and evolution of software).

The goal of the Défi Inria LLM4Code is to leverage LLM capabilities to build code assistants that can enhance both reliability and productivity. The défi is organized along three work packages: Self-improving code generation, Evolution of existing software (WP2), Interactive tools with AI-in-the-loop.

The specific subject lies in the WP2 "migration and modernization of existing software"

Assignment

The candidate will be employee of SopraSteria and spend part of the time in SopraSteria and in DiverSE.

The work is also part of a Défi Inria LLM4Code "Reliable and productive Code Assistants based on large language models" with more than 10 research teams working on several aspects of LLMs and code. Hence the candidate will have the opportunity to collaborate with numerous researchers and experts, as well as to leverage computational infrastructure and the SoftwareHeritage project.

Main activities

A vast portion of the software used nowadays in the critical sectors of the industry is written in legacy languages (e.g. Fortran, COBOL, Ada, etc.) that are prone to be outdated. These languages do not profit from modern software engineering tools, do not adhere to the latest standards of quality or security, and are famous for blocking developers in their everyday work. However, there is no standard solution to migrate an existing code base to newer technologies that would be stable, secure, and affordable from the time/value perspective.
We propose to leverage LLMs' capabilities for software migration. While LLMs excel at

translation tasks for natural language, programming language migration is still challeng- ing [Zhu et al., 2022, Pan et al., 2023, Yan et al., 2023]. Incorporating fine-grained examples into the training of LLMs is essential to capture the nuances of different programming paradigms and semantics. These examples provide detailed, context-rich scenarios that help LLMs understand and adapt to various programming structures and logic. Furthermore, leveraging compilers or transpilers to generate synthetic data can be effective in creating a diverse
training dataset. On the other hand, LLMs can enhance existing compilers or migration tools by broadening their scope to cover more diverse and complex corner cases. This results in tools that are not only more robust, but also capable of addressing a wider range of migration scenarios. Besides, LLMs are not solely beneficial for translating programs; they also play a crucial role in comprehending existing codebases, documenting system architectures, or synthesizing test cases to validate the migration. Both activities are essential for software migration, and our strategy includes utilizing LLMs to efficiently address these specific tasks.
We will first experiment on the open challenge of converting Fortran-77 to C. From many perspectives, the gap between Fortran-77 (as the most spread version of Fortran) and C is significant. Furthermore, the lack of a reference dataset matching Fortran-77 to C code, and the validation of the results generated by the LLM raise multiple challenges. In addition to the challenging case of converting Fortran-77 to C, we aim to explore the problem of migrating old codebases written in programming languages such as 4GL or old Java [Fleurey et al., 2007, Verhaeghe et al., 2019]. Software migration involves resolving many tasks and related problems: reverse engineering (e.g., understanding the existing codebase and functionality, documenting the system's architecture), translating code to a modern platform or programming language, testing (from unit to user acceptance) to ensure the new migrated system fits original requirements, etc. For each task, LLMs can be of interest [Xie et al., 2023, Fan et al., 2023, Hou et al., 2023]

Angela Fan, Beliz Gokkaya, Mark Harman, Mitya Lyubarskiy, Shubho Sengupta, Shin Yoo, and Jie M Zhang. Large language models for software engineering: Survey and open problems. arXiv preprint arXiv:2310.03533, 2023.

Franck Fleurey, Erwan Breton, Benoit Baudry, Alain Nicolas, and Jean-Marc Jézéquel. Model-driven engineering for software migration in a large industrial context. In Model Driven Engineering Languages and Systems: 10th International Conference, MoDELS 2007, Nashville, USA, September 30-October 5, 2007. Proceedings 10, pages 482–497. Springer, 2007.

Benoıt Verhaeghe, Anne Etien, Nicolas Anquetil, Abderrahmane Seriai, Laurent Deruelle, Stéphane Ducasse, and Mustapha Derras. Gui migration using mde from gwt to angular 6: An industrial case. In 2019 IEEE 26th International Conference on Software Analysis, Evolution and Reengineering (SANER), pages 579–583, 2019. doi: 10.1109/SANER.2019.8667989.

Ming Zhu, Karthik Suresh, and Chandan K Reddy. Multilingual code snippets training for program translation. Proceedings of the AAAI Conference on Artificial Intelligence, 36 (10):11783–11790, Jun. 2022. doi: 10.1609/aaai.v36i10.21434. URL https://ojs.aaai.org/index.php/AAAI/article/view/21434.

Rui Xie, Tianxiang Hu, Wei Ye, and Shikun Zhang. Low-resources project-specific code summarization. In Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering, ASE '22, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394758. doi: 10.1145/3551349.3556909. URL https://doi.org/10.1145/3551349.3556909.

Skills

You need to:

- have (or soon receive) a Masters degree in computer science/engineering, informatics, or related fields
- be ok investing 3+ years as a "research apprentice" (aka PhD student)

The subject requires strong expertise in software engineering, including automated software engineering, program transformations (compilers, interpreters, etc.) and analysis, the mastering of numerous languages (ie being polyglot), the development of languages. The candidate should also be highly knowledgeable in LLMs, from foundations to cutting-edge tools recently developed, and excited by the use of LLMs with software (it does not exclude, of course, to be critics about LLMs and their current limits)

Benefits package

- Subsidized meals
- Partial reimbursement of public transport costs

- Possibility of teleworking (after 6 months of employment) and flexible organization of working hours
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities

Remuneration

Monthly gross salary: 2100€

General Information

- **Theme/Domain :** Distributed programming and Software engineering
  Software engineering (BAP E)
- **Town/city :** Rennes
- **Inria Center :** Centre Inria de l'Université de Rennes
- **Starting date :** 2024-02-01
- **Duration of contract :** 3 years
- **Deadline to apply :** 2025-01-15

Contacts

- **Inria Team :** DIVERSE
- **PhD Supervisor :**
  Acher Mathieu / Mathieu.Acher@irisa.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 need to:

- be really excited about our project
- be persistent (get back up and continue when things don't work out as planned -- true research rarely works out as planned)
- be fearless (e.g., be ok hacking a virtual machine, a compiler, a kernel, or implementing a complex algorithm)
- have a small child's attitude (to want to understand and learn about everything they encounter)
- have an engineer's attitude (not to take the first solution that comes to mind, but to look at the key alternatives)
- have a researcher's attitude (to want to truly understand something, and to not be satisfied with the first best explanation)
- want to look at the simple and obvious before exploring the complicated
- be able to focus (to ignore the many other cool things one could also do)
- derive pleasure from coming up with a logical and clear argument or explanation
- like to read (books, papers, papers, papers)
- like to write (prospectus, proposal, dissertation, and papers)
- like to present (at conferences, or in class)
- like to convince others using sound arguments
- be ok working hard
- under-promise and over-deliver
- be happy staying in Brittany for quite some time
- be ok traveling long distance from time to time (e.g., for conferences)

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

Please submit online : your resume, cover letter and letters of recommendation eventually

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.