Inria

# Offer #2024-07866

## PhD Position F/M Code and Proof Generation with Large Language Models

**Contract type :** Fixed-term contract

**Level of qualifications required :** Graduate degree or equivalent

**Fonction :** PhD Position

## Assignment

Generative AI is gaining momentum and has raised significant interest in tackling more and more problems from linguistics, maths, commonsense reasoning, biology, physics, etc.
Transformers introduced in [1] have quickly become the state-of-the-art neural network architecture for sequence processing with applications ranging from natural language processing and computer vision to code generation [2]. Transformers performances scale with the number of parameters and the number of training data [3], and with modern GPU/TPU chips it is now possible to train very large Transformers models with billions of parameters.
Large Language Models (LLMs), like GPT-4, are extremely large Transformers models trained for natural processing tasks on huge datasets containing billions of words. After the initial training, LLMs can be specialized for a specific task using various techniques:

- Fine-tuning consists in adjusting the parameters of an LLM by re-training the model, or part of the model, on a specialized dataset starting from pre-trained parameters. In addition, direct preference optimization [4] can fine-tune LMs to align with human preferences, achieving precise control of the behavior of LLMs.

- Prompt augmentation techniques leverage the capabilities of general-purpose LLMs to learn and adapt by adding context directly in the user input thanks to a prompt. Retrieval Augmented Generation (RAG) [5] is an advanced form of prompt augmentation where, given a prompt, relevant data are retrieved from an external database, and added to the original prompt.

Beyond natural language, general-purpose LLMs quickly demonstrated emergent programming abilities due to the presence of code in the training dataset. There has been an explosion of specialized LLMs either entirely trained or fine-tuned on code: AlphaCode [6], StarCoder [7], Codex [2], CodeT5 [8], Code LLaMa [9], etc. Researchers are only beginning to explore the capabilities of LLMs for software development and many challenges need to be addressed.
In this thesis, we will explore new research in neural code generation and applications to formal verification. To improve the reliability of LLM based code assistants, we will explore possible interactions between the LLM and external tools like a Python interpreter, a test framework, or a proof assistant. While LLM based interactive tools are only nascent, they have the potential to improve software development at every level.
LLMs have shown promise in proving formal theorems using interactive theorem provers (ITP) such as Isabelle, Lean or Coq. While full proof automation remains challenging, one of our goals in this thesis is to build a tool to enable the triple interaction human-ITP-LLM for Coq. We will explore various fine-tuning and prompt augmentation techniques in this context and then focus more precisely on the verification of generated code. We want to use an LLM to formalize a specification in Coq, and generate both the corresponding code, and a proof of correctness using existing formalized semantics. The proof assistant then tries the proof to accept or reject a program, and the human can validate the formal specification, or refine it if necessary.

**References:**

- [1] Attention Is All You Need, Vaswani et al., 2017 https://arxiv.org/abs/1706.03762

- [2] Evaluating Large Language Models Trained on Code, Chen et al., 2021 https://arxiv.org/abs/2107.03374

- [3] Training Compute-Optimal Large Language Models, Hoffmann et al., 2022 https://arxiv.org/abs/2203.15556

- [4] Direct Preference Optimization: Your Language Model is Secretly a Reward Model, Rafailov et al., 2023 https://arxiv.org/abs/2305.18290

- [5] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, Lewis et al., 2020 https://arxiv.org/abs/2005.11401

- [6] Competition-Level Code Generation with AlphaCode, Li et al., 2022 https://arxiv.org/abs/2203.07814

- [7] StarCoder: may the source be with you!, Li et al., 2023 https://arxiv.org/abs/2305.06161

- [8] CodeT5: Identifier-aware Unified Pre-trained Encoder-Decoder Models for Code Understanding and Generation, Wang et al., 2021, https://arxiv.org/abs/2109.00859

- [9] Code Llama: Open Foundation Models for Code, Rozière et al., 2023 https://arxiv.org/abs/2308.12950

## 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
- Flexible organization of working hours (after 12 months of employment)
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training
- Social security coverage

## General Information

- **Theme/Domain :** Optimization, machine learning and statistical methods
  Statistics (Big data) (BAP E)
- **Town/city :** Paris
- **Inria Center :** Centre Inria de Paris
- **Starting date :** 2024-09-01
- **Duration of contract :** 3 years
- **Deadline to apply :** 2024-07-21

## Contacts

- **Inria Team :** ARGO
- **PhD Supervisor :**
  Lelarge Marc / Marc.Lelarge@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.

**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

**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.