# Offre n°2024-07866

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

*Le descriptif de l'offre ci-dessous est en Anglais*

**Type de contrat :** CDD

**Niveau de diplôme exigé :** Bac + 5 ou équivalent

**Fonction :** Doctorant

## Mission confiée

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

## Avantages

- 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

## Informations générales

- **Thème/Domaine :** Optimisation, apprentissage et méthodes statistiques
  Statistiques (Big data) (BAP E)
- **Ville :** Paris
- **Centre Inria :** Centre Inria de Paris
- **Date de prise de fonction souhaitée :** 2024-09-01
- **Durée de contrat :** 3 ans
- **Date limite pour postuler :** 2024-07-21

## Contacts

- **Équipe Inria :** ARGO
- **Directeur de thèse :**
  Lelarge Marc / Marc.Lelarge@inria.fr

## A propos d'Inria

Inria est l'institut national de recherche dédié aux sciences et technologies du numérique. Il emploie 2600 personnes. Ses 215 équipes-projets agiles, en général communes avec des partenaires académiques, impliquent plus de 3900 scientifiques pour relever les défis du numérique, souvent à l'interface d'autres disciplines. L'institut fait appel à de nombreux talents dans plus d'une quarantaine de métiers différents. 900 personnels d'appui à la recherche et à l'innovation contribuent à faire émerger et grandir des projets scientifiques ou entrepreneuriaux qui impactent le monde. Inria travaille avec de nombreuses entreprises et a accompagné la création de plus de 200 start-up. L'institut s'efforce ainsi de répondre aux enjeux de la transformation numérique de la science, de la société et de l'économie.

**Attention**: Les candidatures doivent être déposées en ligne sur le site Inria. Le traitement des candidatures adressées par d'autres canaux n'est pas garanti.

## Consignes pour postuler

**Sécurité défense :**
Ce poste est susceptible d'être affecté dans une zone à régime restrictif (ZRR), telle que définie dans le décret n°2011-1425 relatif à la protection du potentiel scientifique et technique de la nation (PPST). L'autorisation d'accès à une zone est délivrée par le chef d'établissement, après avis ministériel favorable, tel que défini dans l'arrêté du 03 juillet 2012, relatif à la PPST. Un avis ministériel défavorable pour un poste affecté dans une ZRR aurait pour conséquence l'annulation du recrutement.

**Politique de recrutement :**
Dans le cadre de sa politique diversité, tous les postes Inria sont accessibles aux personnes en situation de handicap.