Offer #2025-08969

PhD Position F/M Data selection techniques for LLMs reasoning improvement

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 University of Lille centre, created in 2008, employs 360 people including 305 scientists in 15 research teams. Recognised for its strong involvement in the socio-economic development of the Hauts-De-France region, the Inria University of Lille centre pursues a close relationship with large companies and SMEs. By promoting synergies between researchers and industrialists, Inria participates in the transfer of skills and expertise in digital technologies and provides access to the best European and international research for the benefit of innovation and companies, particularly in the region.

For more than 10 years, the Inria University of Lille centre has been located at the heart of Lille's university and scientific ecosystem, as well as at the heart of Frenchtech, with a technology showroom based on Avenue de Bretagne in Lille, on the EuraTechnologies site of economic excellence dedicated to information and communication technologies (ICT).

Context

Large Language Models (LLMs) have demonstrated remarkable capabilities, with reasoning models highlighting the critical role of high-quality training data. While procedural generation offers infinite training datasets in domains like logical reasoning, games, and retrieval, not all synthetic data contributes equally. Generated examples often suffer from redundancy, inappropriate difficulty, or lack meaningful signal—for instance, large number arithmetic may appear challenging but provides minimal educational value.

This PhD research addresses **optimal data selection from infinite procedural sources**, moving beyond ad-hoc metrics like diversity and difficulty. The work will develop principled methodologies for assessing training data **impact profiles** using influence techniques (influence functions, Shapley values) to quantify how individual examples contribute to model capabilities, with connections to curriculum learning principles.

The candidate will create frameworks encompassing multiple quality aspects to identify high-impact training examples, validated through **downstream performance on real-world tasks** and **computational efficiency metrics**. This research aims to establish new standards for data-efficient training and synthetic data curation.

Keywords: Large Language Models, Data Selection, Procedural Generation, Influence Functions, Training Efficiency

Assignment

This PhD student will collaborate with Damien Sileo and the Adada consortium (engineers and interns) to develop **intelligent data selection methods** for procedurally generated datasets. The research focuses on extracting high-value training examples from massive synthetic data pools, moving beyond simple similarity metrics to downstream tasks toward principled selection criteria that optimize model performance and learning efficiency.

Main activities

Data Generation & Filtering:

- Contribute marginally to synthetic problem generators to understand generation mechanisms
- Develop large-scale data filtering pipelines for procedurally generated datasets
- Explore data representation techniques for effective sample characterization

Core Research Focus:

- Extract optimal coresets from massive synthetic datasets tailored to specific downstream tasks
- Design adaptive curriculum strategies accounting for model scale (larger models requiring more challenging examples)
- Develop hyperparameter modulation techniques for controlled generation diversity and difficulty calibration
- Move beyond similarity-based metrics to develop principled selection criteria optimizing learning outcomes

Validation & Dissemination:

- Evaluate coreset extraction and curriculum strategies across diverse reasoning tasks
- Assess scalability and computational efficiency of proposed filtering methods
- Conduct controlled experiments measuring downstream performance improvements
- Write and disseminate research findings through publications and presentations

Skills

Languages : English (french not mandatory)

Programming language: Python

Deep learning and statistics background

Knowledge of logic and symbolic AI is a plus

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
- Social security coverage

Remuneration

2100 € (gross monthly salary)

General Information

- **Theme/Domain :** Data and Knowledge Representation and Processing
  Statistics (Big data) (BAP E)
- **Town/city :** Villeneuve d'Ascq
- **Inria Center :** Centre Inria de l'Université de Lille
- **Starting date :** 2025-09-01
- **Duration of contract :** 3 years
- **Deadline to apply :** 2025-07-03

Contacts

- **Inria Team :** MAGNET
- **PhD Supervisor :**
  Sileo Damien / damien.sileo@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

Strong knowledge of deep learning and ideally reinforcement learning

Autonomy, critical thinking, willingness to tackle hard problems

Interest in formal algorithms

Strong scientific background

Knowledge of NLP

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.