Offre n°2025-08969

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

Type de contrat : Fixed-term contract

Niveau de diplôme exigé : Graduate degree or equivalent

Fonction : PhD Position

A propos du centre ou de la direction fonctionnelle

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).

Contexte et atouts du poste

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

Mission confiée

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.

Principales activités

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

Compétences

Languages : English (french not mandatory)

Programming language: Python

Deep learning and statistics background

Knowledge of logic and symbolic AI is a plus

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 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

Rémunération

2100 € (gross monthly salary)

Informations générales

- **Thème/Domaine :** Data and Knowledge Representation and Processing Statistics (Big data) (BAP E)
- **Ville :** Villeneuve d'Ascq
- **Centre Inria :** Centre Inria de l'Université de Lille
- **Date de prise de fonction souhaitée :** 2025-09-01
- **Durée de contrat :** 3 years
- **Date limite pour postuler :** 2025-07-03

Contacts

- **Équipe Inria :** MAGNET
- **Directeur de thèse :**
  Sileo Damien / damien.sileo@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'e?orce ainsi de répondre aux enjeux de la transformation numérique de la science, de la société et de l'économie.

L'essentiel pour réussir

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

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.