Job Title: Research Intern (LLM foundations) – Falcon LLM Team

Location: Technology Innovation Institute, Abu Dhabi, UAE

Technology Innovation Institute (TII) is a publicly funded research institute, based in Abu Dhabi, United Arab Emirates. It is home to a diverse community of leading scientists, engineers, mathematicians, and researchers from across the globe, transforming problems and roadblocks into pioneering research and technology prototypes that help move society ahead.

Artificial Intelligence Cross-Center Unit

The Artificial Intelligence Cross-Center Unit is the machine learning powerhouse of TII, working in close collaboration with our other research centers to harness the full benefits of AI across our projects – and drive innovation from new computing paradigms, designing and delivering new AI methodologies, technologies, solutions, and systems that address challenging issues across multiple sectors of the economy – from technology to healthcare, cybersecurity, and government, among others.

We incorporate core elements of intelligence (perception, sensing, planning, and language) in the ideation, design, and prototyping of next-generation systems with human-like intelligence. We build advanced AI computing and scalable AI-based software stacks and hardware systems to deliver significant enhancements in systems infrastructure. Our AI researchers, scientists, and engineers collaborate to ensure innovative outcomes, from AI theory to AI technologies towards better intelligence.

Falcon LLM Team

The Falcon LLM team at the Technology Innovation Institute (TII) is at the forefront of developing cutting-edge generative AI and language models. Our Falcon models have garnered significant open-source adoption, and we are committed to pushing the boundaries of AI performance, alignment, and safety. Join our dynamic team to advance the capabilities of our foundational models and make impactful contributions to the AI community.

Role Overview:

As a Research Intern with the Falcon team, you will work on advancing the state-of-the-art in generative AI, with a focus on novel architecture designs for next-generation large language models (LLMs). Your efforts will aim to improve the efficiency and scalability of our Falcon models by exploring and optimizing emerging LLM architectures. You will play a pivotal role in pioneering advanced architectures, filling gaps in current approaches, and pushing the boundaries of what's possible in the domain.

You will conduct research on model scaling, benchmarking, and architecture-specific optimization strategies, while developing tailored training and data methodologies to maximize performance across various designs. This role provides the opportunity to collaborate with researchers and cross-functional teams, contributing to high-quality research outcomes that impact both academic knowledge and real-world applications.

Key Responsibilities:

- Explore and develop novel architecture designs for next-generation large language models (LLMs), aiming to improve upon existing Transformer-based approaches.
- Conduct research on model scaling laws and benchmarking to evaluate the performance of different LLM architectures, identifying areas for optimization.
- Develop and apply tailored training strategies and data preparation techniques to optimize the performance of diverse LLM architectures.
- Contribute to research publications or technical reports, showcasing advancements in LLM architectures and their practical implications.
- Stay up-to-date with the latest developments in generative AI and language model research to incorporate emerging techniques and approaches into the work.

Minimum Qualifications:

- Ph.D. student or last-year Master student in Computer Science, Artificial Intelligence, Machine Learning, Generative AI, or a related technical field.
- Proficiency in programming languages such as Python, C++, Java, or similar, with a solid understanding of software development.
- Experience with machine learning frameworks such as PyTorch, TensorFlow, or JAX.
- Deep understanding of generative models, strong engineering skills to implement and validate research ideas quickly.
- Eligible for internship of 6 months or longer, agreed by supervisors.

Preferred Qualifications:

- Proven track record of significant contributions to AI research, demonstrated by first-authored publications in top conferences (e.g., NeurIPS, ICML, ICLR, ACL, EMNLP, CVPR, ECCV, KDD, VLDB, etc.) or patents.
- Experience with distributed training frameworks, large-scale data processing, high-dimensional data analysis from different sources or formats.

- Experience with CUDA programming, training or inference accelerations, system-level optimizations, etc.
- Contributions to open-source AI projects, libraries, or frameworks.
- Ability to work effectively in a collaborative, cross-functional environment, with a demonstrated ability to communicate complex ideas clearly.

What we offer:

- Competitive Benefits: Enjoy competitive compensation, access to state-of-the-art computational resources, and the chance to work with some of the brightest minds in the AI field. Our collaborative and inclusive work culture is centered on innovation and personal growth.
- Internship Perks: We provide round-trip flights, accommodations, health insurance, and access to advanced equipment and a cutting-edge working environment.
- Mentorship and Project Involvement: Benefit from close mentoring and active participation in exciting AI projects that will help you grow your skills.
- Equal Opportunity: We are committed to creating a diverse and inclusive workplace. TII values diversity and does not discriminate based on race, religion, gender, age, national origin, sexual orientation, marital status, veteran status, or disability.