Job Title: Senior Research Engineer (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 Senior Research Engineer on the Falcon team, you will help shape the future of generative AI by transforming cutting-edge research into scalable, usable solutions. Your role will focus on bridging the gap between academic innovation and real-world deployment by optimizing the architecture, training, and inference performance of Falcon models. You'll work closely with researchers and engineers to ensure our models are efficient, deployable, and impactful across platforms and use cases.

In this role, you will also engage with the open-source community and end-users to promote the usability, adoption, and continuous improvement of Falcon models. This position offers a unique opportunity to contribute to both technical excellence and community-driven impact.

Key Responsibilities:

- Drive optimizations in large-scale LLM training pipelines, from infrastructure efficiency to model architecture refinements.
- Design and develop custom Triton/CUDA kernels and advanced training strategies to improve performance and scalability.
- Collaborate with cross-functional teams to ensure Falcon models are production-ready, memory-efficient, and performant across diverse hardware environments.
- Contribute to open-source integration efforts and continuously improve Falcon's accessibility and adoption based on community feedback.
- Explore and apply advanced data processing and training strategies to unlock model performance across multiple domains.
- Stay current with the latest trends in LLMs, generative AI, and system-level ML optimizations, and translate them into practical enhancements.

Minimum Qualifications:

- Master's or Ph.D. in Computer Science, AI, ML, or a related technical field.
- Strong proficiency in Python and C++, with proven experience in scalable software development.
- Hands-on experience with distributed training frameworks such as Megatron-LM, DeepSpeed, FSDP, etc.
- Experience with large-scale data processing and optimizing model throughput and memory efficiency in deployment.
- Strong analytical and problem-solving skills with a drive to understand and improve the underlying mechanics of model performance.

Preferred Qualifications:

- Experience with low-level optimization: CUDA, Triton kernels, low-precision training and inference acceleration.
- Demonstrated contributions to open-source projects related to ML, deep learning, or systems optimization.

- Strong communication and collaboration skills, with a willingness to lead engineering efforts and work across diverse teams.
- Deep curiosity and ownership mindset—willing to dig into metrics, logs, and performance reports to drive improvements.

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.
- 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.