attentions.ai

Job Description for AI Engineer- LLM Architectures

Job Title: AI Engineer - LLM Architectures

Location: Pune, INDIA

Job Type: Full-time

Company Overview:

Attentions.ai is pioneering the future of artificial intelligence through advanced research and practical applications in deep learning, with a focus on LLMs (Large Language Models). We are committed to pushing the boundaries of AI to enhance natural language understanding and generative capabilities.

Job Description:

We are seeking a AI Engineer with a specialized focus on LLM architectures, experienced in CUDA programming, PyTorch, and the use of the Triton toolkit for optimizing deep learning models. The ideal candidate will have a deep understanding of GPU programming and neural network optimization techniques, contributing directly to our core products and R&D projects.

Responsibilities:

- LLM Development and Optimization: Design and implement large language models using PyTorch, focusing on scalability and efficiency. Optimize existing models using CUDA and Triton to enhance performance.

- Research and Innovation: Conduct research that contributes to the state-of-the-art in LLM architectures, including publishing results in top-tier AI conferences and journals.
- GPU Programming: Develop custom CUDA kernels and utilize Triton for creating high-performance operations tailored to the needs of neural network training and inference.
- Collaboration: Work closely with other engineers, researchers, and product teams to integrate advanced AI models into commercial applications and services.
- Technical Mentoring: Lead by example in technology and methodology, mentoring junior engineers and helping to guide the strategic direction of our AI engineering efforts.

Qualifications:

- Knowledge in AI and machine learning, with specific expertise in CUDA programming, PyTorch, and LLM architectures.
- Skills: Strong expertise in GPU programming and deep learning frameworks, particularly PyTorch.
- Proficiency in developing and optimizing CUDA kernels.
- Experience with Triton or similar tools for optimizing deep learning workloads on GPUs.
- Demonstrated ability to design and implement complex LLMs.
- Published work in AI/ML fields is highly desirable.
- Soft Skills: Excellent problem-solving skills, innovative thinking, effective communication abilities, and a strong collaborative spirit.