Full job description

- 3+ years of non-internship professional software development experience - 2+ years of non-internship design or architecture (design patterns, reliability and scaling) of new and existing systems experience - Experience programming with at least one software programming language

The Artificial General Intelligence (AGI) team is looking for a passionate, talented, and inventive Software Development Engineer(SDE)/Machine Learning Engineer(MLE) to play pivotal role in the development of industry-leading multi-modal and multi-lingual Large Language Models (LLM). As our SDE/MLE superstar, you'll have the power to lead the charge in developing mind-blowing algorithms and modeling techniques that will push the boundaries of large model training using cutting-edge hardware like GPUs and AWS Trainium. Your groundbreaking work will directly impact our customers' lives through game-changing products and services powered by your Generative AI breakthroughs! Get ready to dive into Amazon's vast and diverse data sources and harness the immense power of our large-scale computing resources to turbocharge the development of multi-modal Large Language Models (LLMs) and other awe-inspiring Generative Artificial Intelligence (Gen AI) applications. Your expertise and insights will be invaluable in defining data strategies, enrichment processes, model optimizations, and evaluation methods that will set new standards in the industry! So, if you're passionate about pushing the limits of AI, thrive in a fast-paced and innovative environment, and are ready to make a lasting impact on the world, this is your chance! Join us on this exhilarating adventure and let's revolutionize AGI together! #MLE Key job responsibilities * Ability to quickly learn cutting-edge technologies and algorithms in the field of Generative AI to participate in our journey to build the best LLMs. * Responsible for the development and maintenance of key platforms needed for developing, evaluating and deploying LLM for real-world applications. * Work with other team members to investigate design approaches, prototype new technology and evaluate technical feasibility. * Work closely with Applied scientists to process massive data, scale machine learning models while optimizing. A day in the life As a SDE/MLE with the AGI team, you will be responsible for leading the development of novel algorithms and modeling techniques to advance the state of the art of large model training using hardware like NVDIA GPUs. Your work will directly impact our customers in the form of products and services that make use of Generative AI innovations. You will leverage Amazon's heterogeneous data sources and large-scale computing resources to accelerate development with multi-modal Large Language Models (LLMs) and other Generative Artificial Intelligence (Gen AI) applications. As a key player in our team, you'll have a significant influence on our overall strategy, shaping the future direction of AGI at Amazon. You'll be the driving force behind our system architecture and the champion of best practices that will ensure an unparalleled infrastructure of the

highest quality. Work in an Agile/Scrum environment to move fast and deliver high quality software. About the team Join our AGI team and work at the forefront of AI. Collaborate with top minds pushing boundaries in deep learning, reinforcement learning, and more. Gain valuable experience and accelerate your career growth. This is a unique opportunity to create history and shape the future of artificial intelligence. Mission of the team: We leverage our hyper-scalable, general-purpose large model training and inference systems to develop and deploy cutting-edge sensory AI foundational models that revolutionize machine perception, interpretation and interaction, with humans and with the physical world.

- 3+ years of full software development life cycle, including coding standards, code reviews, source control management, build processes, testing, and operations experience
- Master's degree in computer science or equivalent
- Experience in techniques like kernel fusion and custom kernels to improve GPU utilization, mixed precision training using lower precision and dynamic loss scaling while leveraging hardware specific mixed precision capabilities and/or demonstrated ability to implement efficient memory management like gradient (activation) checkpointing, gradient accumulation, offloading optimizer states, and smart prefetching.

Los Angeles County applicants: Job duties for this position include: work safely and cooperatively with other employees, supervisors, and staff; adhere to standards of excellence despite stressful conditions; communicate effectively and respectfully with employees, supervisors, and staff to ensure exceptional customer service; and follow all federal, state, and local laws and Company policies. Criminal history may have a direct, adverse, and negative relationship with some of the material job duties of this position. These include the duties and responsibilities listed above, as well as the abilities to adhere to company policies, exercise sound judgment, effectively manage stress and work safely and respectfully with others, exhibit trustworthiness and professionalism, and safeguard business operations and the Company's reputation. Pursuant to the Los Angeles County Fair Chance Ordinance, we will consider for employment qualified applicants with arrest and conviction records.