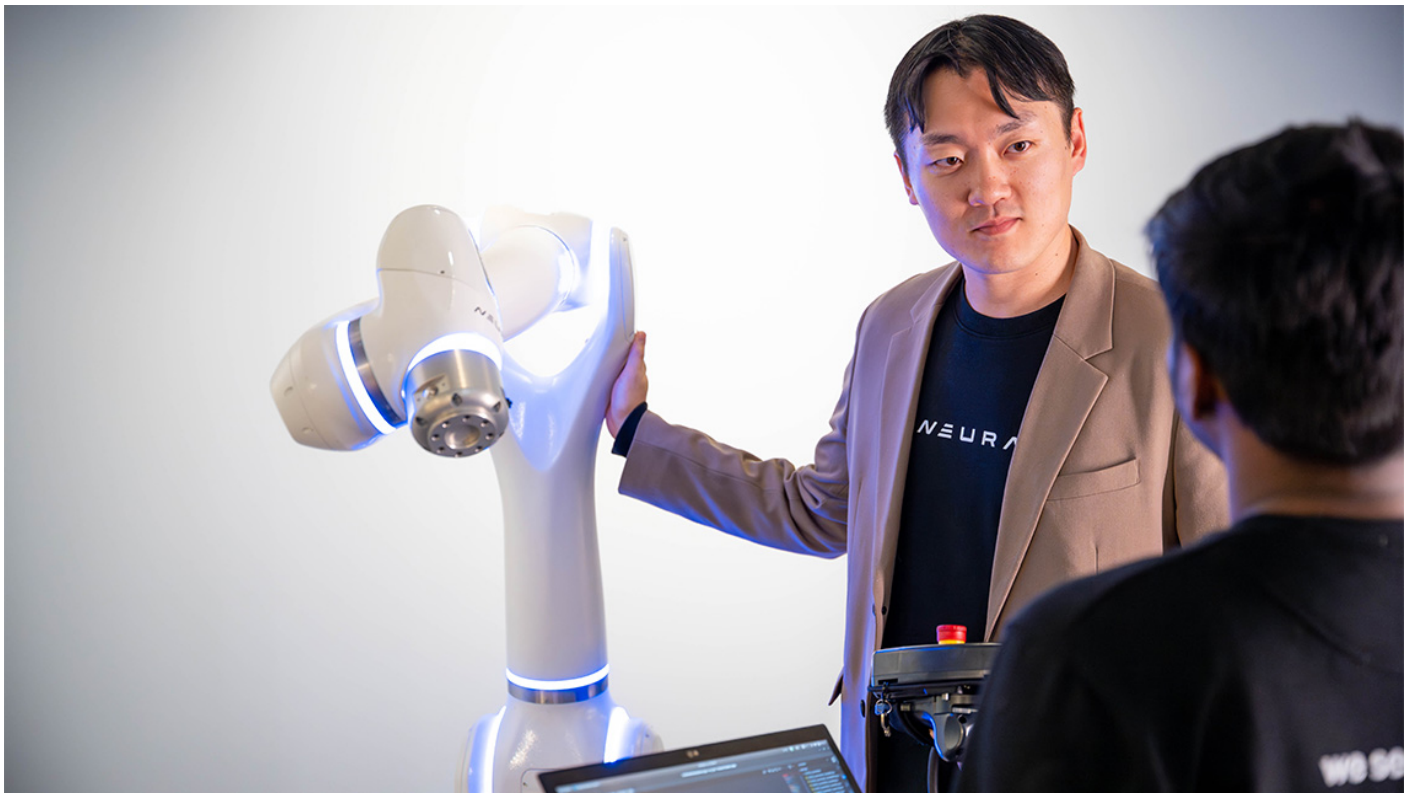

Welcome to **NEURA Robotics**, the innovator of the robotics world. Our goal is to equip collaborative robots with groundbreaking cognitive capabilities to enable safe and intuitive collaboration with humans. Under the leadership of founder David Reger, we have spent the first years of **NEURA Robotics** laying the foundations for humans and robots to work hand in hand.

"We serve humanity" is not just a motto, but our mission. Become part of our ambitious, international company and shape the future of robotics with us.

Welcome to **NEURA Robotics** - where innovation meets team spirit.

Your mission & challenges

- You are the go-to expert for NEURA's GPU cluster infrastructure - a large-scale AWS HyperPod environment running cutting-edge GPU instances for foundation model training and customer fine-tuning workloads. You design the operational framework, build self-service tooling for ML teams, and work directly with AWS to influence the platform at the hyperscaler level.
- Your focus is on cluster engineering and operations — not on ML research itself, but on making sure the people doing that research have rock-solid, efficient, and accessible infrastructure under them.
- Setting up, configuring, and continuously evolving NEURA's HyperPod clusters, including HyperPod/Slurm and HyperPod/EKS orchestration models.
- Designing and implementing strategies for cluster stability: node failure detection, automated job recovery, checkpoint coordination, and fault-tolerant multi-node training workflows.

- Providing a workload priority management framework that allows multiple teams and use cases like foundation model pretraining, fine-tuning, customer workloads, to share cluster capacity efficiently and fairly.
- Optimizing end-to-end GPU utilization: identifying and resolving bottlenecks across compute, GPU memory, EFA networking, and storage throughput.
- Working directly and closely with the AWS HyperPod product and solutions engineering teams, escalating operational issues, sharing learnings from one of the platform's largest deployments, and placing concrete requirements on the roadmap.
- Providing self-service tooling that allows ML researchers and engineers to launch, monitor, and manage training jobs independently, without requiring infrastructure intervention for routine operations.
- Developing onboarding documentation, training materials, and internal workshops that enable users to operate efficiently, follow best practices, and understand cost implications of their workloads.
- Infrastructure as Code is a given for you. Every cluster configuration, every operational change, every new environment is code first.
- Owning the cost and capacity strategy: Spot instance management, Reserved Instance planning, Savings Plans, and ongoing commitment negotiations with AWS.

What we can look forward to

- 5+ years of experience in infrastructure or systems engineering, with a strong focus on GPU cluster or HPC operations.
- Deep hands-on experience with AWS HyperPod and AWS instances; direct prior experience with HyperPod is a strong differentiator.
- Solid understanding of both Slurm and Kubernetes as cluster orchestration layers, and the ability to evaluate their trade-offs for large-scale GPU workloads.
- Practical knowledge of distributed training - you understand what affects throughput and how to debug it.
- Experience building self-service tooling and operational documentation for technical end users.
- You make complex infrastructure accessible, not just functional.
- Strong understanding of cloud cost management at scale: Spot interruption handling, capacity reservations, cost attribution across teams and workloads.
- Comfort working across organizational boundaries — your primary partners are ML researchers, but you'll also work closely with product, finance, and cloud vendor teams.
- Strong English communication skills. German is a plus.

What you can look forward to

- Become part of an agile company, actively shape topics and benefit from flat hierarchies in a highly motivated team
- Enjoy an attractive salary, flexible working hours and 30 days of vacation
- The freedom to contribute your own ideas and drive them forward
- Celebrate successes together with company events
- Take advantage of our corporate benefits program
- And even more fun with great colleagues

Apply

We are looking forward to meeting you and shaping the future of robotics together. Are you in?

Couldn’t find a suitable position? Please send us an unsolicited application.
We are always looking for passionate tech enthusiasts to help us revolutionize the world of robotics!

NEURA ROBOTICS