ML Inference / Serving Infrastructure Engineer

Location: San Francisco (US) or Berlin (GER)
Type: Permanent (Full-time)
Start Date: As soon as possible / by arrangement

ABOUT DAINA

DAiNA is a precision-oncology company focused on enabling personalized cancer treatment for individual patients. We combine comprehensive molecular tumor data including genomics, transcriptomics (bulk, single cell and spatial), proteomics and epigenetics, with AI-driven analysis. Our platform connects multi-omic profiling with functional ex-vivo tumor models and personalized liquid-biopsy monitoring, creating a continuous workflow from biopsy and treatment selection through to therapy monitoring and adaptation. We also operate GMP manufacturing to produce individualized N=1 therapeutics. In short, we help physicians make more informed, personalized treatment decisions based on high-dimensional molecular tumor data.

For more information, visit: www.daina.com

THE ROLE

As ML Inference / Serving Infrastructure Engineer, you will build and optimize the serving infrastructure that enables DAiNA to run AI and LLM workloads reliably, securely and efficiently.

You will focus on high-throughput inference, GPU orchestration, autoscaling, observability and deployment patterns for open-weight models. Your work will support DAiNA's AI-driven reporting, retrieval, decision-support and computational oncology workflows.

A key part of the role is to ensure that model serving is performant, reproducible and suitable for secure biomedical environments with strong requirements around privacy, reliability and operational control.

WHAT YOU'LL DO

- Build and optimize inference infrastructure for open-weight models using tools such as vLLM, TGI, TensorRT-LLM or comparable frameworks.
- Design deployment patterns for secure cloud, isolated or sovereign environments.
- Implement GPU orchestration, autoscaling, monitoring and observability for AI workloads.
- Optimize model serving for latency, throughput, cost and reliability.
- Define performance baselines, load-testing approaches and operational metrics.
- Instrument systems for latency, throughput, error rates, utilization and drift.
- Work with AI/ML, RAG, fine-tuning and platform teams to integrate serving infrastructure into DAiNA workflows.
- Document deployment patterns and operational guidance clearly for production use.

WHAT YOU BRING

- Strong infrastructure, MLOps or platform engineering background with experience serving ML or LLM systems in production.
- Strong Python skills and solid systems engineering knowledge.
- Hands-on experience with vLLM, TGI, TensorRT-LLM, Kubernetes-style orchestration or comparable tools.
- Practical understanding of GPU workloads, autoscaling, model deployment and inference optimization.
- Strong focus on observability, reliability, cost control and operational robustness.
- Experience building clean, reproducible infrastructure components and documentation.
- Ability to work closely with engineering, AI/ML and platform stakeholders.

NICE TO HAVE

- Experience with secure cloud, VPC-isolated, on-premise or sovereign deployment environments.
- Experience tuning performance across different GPU hardware.
- Familiarity with open-weight model families such as Llama, Mistral, Qwen or comparable models.
- Experience with monitoring, logging, tracing, model versioning or deployment automation.
- Background in healthcare, biotech, diagnostics, pharma or another sensitive-data environment.

WHY DAINA

- Employer contributions toward private health insurance or supplementary health coverage, as well as pension or retirement savings.
- Hybrid model with ~60% of working time **expected** on-site and the rest **remotely**, depending on team and business needs.
- Performance-based bonus opportunity, depending on company and individual performance.
- A personal development budget for conferences, courses, certifications, and training.
- The opportunity to work directly on real patient cases and help shape a first-in-class precision-oncology platform
- A proactive, collaborative team with fast decision-making and strong ownership of your domain.

HOW TO APPLY

Send your CV and a short note on why this role fits you to **recruiting@daina.com**, quoting “ML Inference / Serving Infrastructure Engineer” in the subject line. We review applications on a rolling basis and aim to reply quickly.

DAiNA is an equal-opportunity employer. We welcome applicants of every background and assess every candidate on merit, regardless of age, gender, ethnicity, religion, disability, sexual orientation or origin. If you need any adjustment to the process, let us know in your application.