Zum Klienten:
Weitere Informationen zum Klienten und laufend neue Jobchancen erhalten Sie nach Ihrer Anmeldung und dem Onboarding durch Ihren Fachberater per Mail oder online. Wir freuen uns, Sie persönlich kennenzulernen!

Arbeitsort:
Zürich (X Kilometer und X Minuten von Ihrer Wohnadresse entfernt)

Jobbeschreibung:
Machine Learning Platform Engineer (w/m/d)

Ihr Aufgabengebiet:
- Aufbau und Betrieb der ML-Infrastruktur und Plattformen für die KI-Produkte von A1
- Design von Systemen für Modelltraining, Evaluation, Deployment, Inferenz und Experimentierung
- Aufbau und Optimierung von Model-Serving- und Inferenz-Infrastruktur für Workloads mit hohem Durchsatz und niedriger Latenz
- Verbesserung von Zuverlässigkeit, Skalierbarkeit, Latenz und Kosteneffizienz der KI-Systeme
- Entwicklung robuster Pipelines für Datenaufbereitung, Training, Evaluation, Modell-Release und kontinuierliche Verbesserung
- Aufbau von Plattformen und Tooling, damit AI Engineers und Researcher schneller experimentieren, evaluieren und Modelle ausliefern können
- Entwicklung von Evaluations- und Benchmarking-Infrastruktur zur Messung von Modellqualität, Performance und Regressionen
- Aufbau von Produktions-Observability, Monitoring, Tracing und Alerting für KI/ML-Workloads
- Verbesserung der KI-Systeme hinsichtlich Zuverlässigkeit, Skalierbarkeit, Latenz, Durchsatz und Kosten
- Identifikation von Engpässen im ML-Stack und kontinuierliche Verbesserung der Systemperformance
- Enge Zusammenarbeit mit AI Engineers, Researchern und Produktteams, um sich wandelnde Modellanforderungen in produktionsreife Infrastruktur zu überführen

Was Sie ausmacht:
- Fundierte Software-Engineering-Grundlagen und Erfahrung im Aufbau produktiver Systeme
- Erfahrung im Aufbau von ML-Infrastruktur, Plattformen oder produktiven Machine-Learning-Systemen
- Erfahrung mit Modell-Deployment, Inferenz, Evaluation oder Datenpipelines
- Fundiertes Verständnis verteilter Systeme und Systemzuverlässigkeit
- Fähigkeit, sauberen, wartbaren Code in Produktionsqualität zu schreiben
- Sicherer Umgang mit mehrdeutigen, schnelllebigen Arbeitsumgebungen
- Ausgeprägtes Verantwortungsbewusstsein sowie Freude an Experimentierfreude und kontinuierlicher Verbesserung
- Kenntnisse in Python sowie PyTorch/JAX
- Erfahrung mit LLM- und ML-Serving-Infrastruktur wie vLLM, SGLang oder TensorRT-LLM
- Vertrautheit mit Cloud-Infrastruktur, verteilten Systemen, ML/Daten-Pipelines, Workflow-Orchestrierung sowie GPU-Infrastruktur und Vektordatenbanken

Was diese Stelle ausmacht:

- Zahlreiche individuelle Benefits