

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:

(X Kilometer und X Minuten von Ihrer Wohnadresse entfernt)

Jobbeschreibung:

Staff / Principal Machine Learning Engineer, Serving - Switzerland (w/m/d)

Ihr Aufgabengebiet:

- Optimierung moderner Serving-Frameworks und Inferenz-Techniken wie vLLM oder TRT-LLM
- Modellbeschleunigung mittels Quantisierung, Distillation, Caching-Strategien, Continuous Batching, Paged Attention und Speculative Decoding
- Entwicklung und Optimierung hochperformanter Systeme, inkl. Profiling und Ausschöpfen der GPU-Performance
- Arbeiten mit verteilten Systemen und Skalierung ueber Kubernetes, Ray, Multi-GPU/Multi-Node-Inferenz sowie zuverlaessige Handhabung tausender gleichzeitiger Verbindungen
- Uebernahme von Modellen aus dem Research-Team, Containerisierung, Optimierung des Serving-Prozesses und Sicherstellung eines zuverlaessigen Produktivbetriebs
- Enge taegliche Zusammenarbeit mit dem in den USA ansaessigen Leadership- und Engineering-Team

Was Sie ausmacht:

- Tiefes Verstaendnis moderner Serving-Frameworks wie vLLM oder TRT-LLM
- Praktische Erfahrung mit Modellbeschleunigungstechniken wie Quantisierung, Distillation und Caching-Strategien
- Fundierte Kenntnisse in C++, CUDA, Rust oder hochoptimiertem Python sowie im Profiling von Code
- Erfahrung mit verteilten Systemen und Skalierung, z. B. Kubernetes, Ray und Multi-GPU/Multi-Node-Inferenz
- Nachweisbare eigene Projekte, Open-Source-Beitraege oder technische Fachartikel
- Faehigkeit, Modelle end-to-end von der Forschung bis zum stabilen Produktivbetrieb zu verantworten
- PhD in Informatik, Physik, Mathematik oder gleichwertige praktische Erfahrung im Aufbau von Backend- oder ML-Systemen
- Verhandlungssichere Englischkenntnisse in Wort und Schrift

Was diese Stelle ausmacht:

- Vollzeit-Festanstellung mit Remote-Arbeit innerhalb der Schweiz
- Flache Strukturen, schnelle Iterationen und minimaler Prozess-Aufwand
- Moeglichkeit zur Mitwirkung an Open-Source-Beitraegen und sichtbarer Forschungsarbeit
- Bei Interesse Unterstuetzung bei Umzug und Visum in die San Francisco Bay Area moeglich