

**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:**

Schaffhausen (X Kilometer und X Minuten von Ihrer Wohnadresse entfernt)

**Jobbeschreibung:**

AI/ML Engineer (w/m/d)

**Ihr Aufgabengebiet:**

- Design, Implementierung und Deployment produktionsreifer, skalierbarer AI/ML-Systeme, inklusive multimodaler Foundation-Model-Lösungen
- Aufbau robuster APIs, Evaluationspipelines, Infrastruktur und Observability-Layer
- Rasches Prototyping und Bewertung neuer AI-Frameworks, Bibliotheken und Tools zur Verbesserung von Systemperformance, Entwicklungsgeschwindigkeit oder Produktfähigkeiten, mit anschliessender Überführung in gut architektierte, wartbare Systeme
- Zusammenarbeit mit Product-, Data- und Engineering-Teams zur Übersetzung von Business-Problemen in AI-gestützte Lösungen
- Mitwirkung am Auf- und Ausbau der internen ML-Plattform und des Toolsets, inklusive CI/CD für ML sowie Experiment-Tracking und -Evaluation
- Umsetzung von spec-driven Development, bei dem Spezifikationen und Design-Intent selbst verantwortet werden, während LLMs einen Grossteil der Implementierung übernehmen, stets mit Mensch im Kontrollprozess

**Was Sie ausmacht:**

- Leidenschaft für den Aufbau robuster Systeme, schnelles Experimentieren und das Ausliefern von Mehrwert durch AI
- Ausgeprägter AI-Optimismus: Sie greifen als Erstes zu diesen Tools, loten deren Grenzen aus und lassen sich von den Möglichkeiten begeistern
- Starke Python-Programmierenkenntnisse sowie Erfahrung mit ML/AI-Bibliotheken wie PyTorch, Hugging Face Transformers, LangChain und LangGraph
- Erfahrung im Deployment und Betrieb produktionsreifer ML-Systeme, inklusive Monitoring, Testing und Continuous Delivery
- Expertise mit LLMs, Vektordatenbanken, Retrieval-Augmented Generation (RAG) sowie im Aufbau agentischer Systeme mit Tool-Nutzung und Orchestrierung
- Strukturierter Ansatz zur Bewertung der LLM-Performance und zum Verständnis des Prompt-Verhaltens über verschiedene Modellvarianten und -versionen hinweg
- Praktische Erfahrung mit Cloud-Plattformen (GCP, AWS oder Azure); Kenntnisse in Infrastructure-as-Code oder Container-Orchestrierung von Vorteil
- Fähigkeit, rasch mit Prototypen zu iterieren und diese anschliessend in gut architektierte, wartbare Systeme zu überführen
- Kollaborative Denkweise und ausgezeichnete Kommunikationsfähigkeiten für die funktions- und disziplinübergreifende Zusammenarbeit

**Was diese Stelle ausmacht:**

- Zahlreiche individuelle Benefits