

**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:**

AI Engineer (w/m/d)

**Ihr Aufgabengebiet:**

- Spannende und innovative GenAI-Projekte für unsere Kunden: von Kreativ-Workshops über Proof-of-Concepts bis zur gemeinsamen Umsetzung produktiver Lösungen
- Konzeption und Implementierung von LLM-basierten Applikationen
- Design und Umsetzung von Retrieval-Augmented-Generation-Systemen (RAG)
- Umsetzung von Structured Information Extraction aus unstrukturierten Dokumenten
- Design und Realisierung von agentischen Systemen und AI-Workflows
- Prompt- und Context-Engineering zur gezielten Steuerung von LLM-basierten Lösungen
- Sorgfältige Evaluation der Lösungen nach dem Prinzip des Evaluation Driven Development (EDD), inklusive Observability über AI-Evaluation- und Observability-Plattformen
- Einbringen der eigenen Expertise in die Angebotsphase und fachliche Mitgestaltung von Offerten
- Optionale Zusammenarbeit mit klassischem Software Engineering und MLOps, eng mit den Software- und ML-Engineers

**Was Sie ausmacht:**

- Einige Jahre Erfahrung in der Entwicklung von GenAI-Anwendungen, idealerweise im produktiven Einsatz für Kunden oder eigene Produkte
- Abgeschlossenes Masterstudium in Informatik, Data Science oder einem verwandten Fachgebiet
- Nachgewiesene Expertise bei der Implementierung und dem Betrieb von Retrieval-Augmented-Generation-Systemen (RAG)
- Erfahrung mit agentischen Systemen und AI-Workflows sowie dem Model Context Protocol (MCP)
- Sorgfältiger Umgang mit der Evaluation von LLM-Lösungen, vertraut mit Evaluation Driven Development (EDD) und AI-Evaluation- und Observability-Plattformen (AEOPs)
- Vertraut mit modernen Vektordatenbanken, z. B. Qdrant, sowie mit Embedding- und Re-Ranking-Modellen
- Verständnis von Open-Weight Large Language Models, inklusive multimodaler Modelle
- Erfahrung mit LLM-Runtimes sowie LLM-Proxies bzw. Gateways
- Sicherer Umgang mit Python und Software-Engineering-Best-Practices
- Ausgeprägte Kommunikationsstärke und Freude an der Zusammenarbeit mit Team und Kunden
- Verhandlungssicheres Deutsch und Englisch in Wort und Schrift
- Von Vorteil: Cloud-Erfahrung in Infrastructure as Code (IaC), idealerweise Azure sowie Erfahrung mit Guardrail von LLM-Anwendungen

**Was diese Stelle ausmacht:**

- Fortwährendes Lernen und stetige Weiterbildung im Bereich GenAI
- Einblicke in die verschiedenen Branchen unserer Kunden
- Abwechslung und spannende Zusammenarbeit in internen Projektteams sowie vor Ort bei Kunden