



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:

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

Jobbeschreibung:

Voice AI Engineer (w/m/d)

Ihr Aufgabengebiet:

- Verbesserung der Gesprächsdynamik: Sprechen, Warten, Übergeben oder Unterbrechen des Agenten, inklusive Spracherkennung und Erkennung des Gesprächsendes
- Reduktion der End-to-End-Latenz über Spracherkennung, Sprachmodelle und Sprachsynthese sowie nahtloses Failover zwischen Anbietern
- Optimierung des LLM-Einsatzes im Gespräch: Prompting, Tool-Aufrufe, Streaming sowie spekulative oder parallele Anfragen
- Betrieb eigener Sprachmodelle (STT und TTS) auf GPUs, Messung von Durchsatz und Zeit bis zum ersten Audio-Output
- Fine-Tuning von Sprach- und Sprachmodellen für die Sprachen und Stimmen der Kunden
- Aufbau von Evaluationsprozessen: Testanrufe, Benchmarks und Qualitätsmetriken, um den Nutzen von Änderungen zu beurteilen

Was Sie ausmacht:

- Mehrjährige oder fortgeschrittene Erfahrung in Produktion mit Echtzeit-Audio, Conversational AI oder ML-Modellbetrieb
- Fundierte Python-Kenntnisse, inklusive asyncio und Streaming-Pipelines
- Praktische Erfahrung mit Sprachtechnologien (STT, TTS, VAD) und LLM-APIs in latenzsensitiven Systemen
- Erfahrung im GPU-basierten Modellbetrieb, z. B. mit vLLM oder vergleichbaren Lösungen, sowie Grundkenntnisse in Batching und Speicherlimits
- Von Vorteil: Erfahrung mit LiveKit Agents, pipecat oder WebRTC
- Von Vorteil: Fine-Tuning (LoRA oder vollständig) von Sprach- oder Sprachmodellen
- Von Vorteil: Kenntnisse in Audiosignalverarbeitung wie Echo-Unterdrückung oder Rauschunterdrückung
- Von Vorteil: Deutschkenntnisse oder Erfahrung mit deutschsprachigen Sprachdaten
- Sehr gute Englischkenntnisse in Wort und Schrift, da Englisch die Arbeitssprache ist
- Analytisches Vorgehen mit Vergleich von Vorher-Nachher-Werten, datenbasierte Entscheidungsfindung
- Hohes Verantwortungsbewusstsein von der Problemerkennung bis zur produktiven Lösung

Was diese Stelle ausmacht:

- Flexible Arbeitsgestaltung: vollständig remote oder im Co-Working-Space
- Firmenlaptop und -telefon werden gestellt

- Kleines, agiles Team mit direktem Einfluss der eigenen Arbeit
- Entwicklungsmöglichkeiten über den gesamten Voice-Stack hinweg, vom Modellbetrieb bis zum Nutzererlebnis am Telefon