Lead AI Engineer / Architect (m/w/d) - Agentic Systems

Mehr als Labor.
Die Limbach Gruppe besteht derzeit aus über 30 Einzellaboratorien. Die ärztlich geführten Einzellaboratorien haben sich durch kompetente medizinische Beratung, hochspezialisierte Diagnostik, eine umfassende Angebotspalette und ein breites Dienstleistungsspektrum als führende Unternehmensgruppe etabliert.

Die Laboratorien sind ein verlässlicher Partner für niedergelassene Ärzte, Krankenhäuser und andere medizinische Einrichtungen.

Aufgaben

In unserer Laborgruppe werden täglich tausende Laborparameter gemessen, Befunde freigegeben und Anforderungsscheine verarbeitet. Diese Prozesse werden von hochqualifizierten MTAs und Fachärzten manuell durchgeführt.Unsere neu gegründete, direkt an die Unternehmensleitung angebundene KI-Einheit entwickelt produktive ML-Systeme, die diese Kernprozesse digitalisieren: von der computergestützten Vorbefundung medizinischer Proben über automatisierte Auftragserfassung bis zur standortübergreifenden Skalierung dieser Lösungen auf über 30 Labore.Der Einsatz erfolgt in einer ISO-15189-zertifizierten Umgebung unter den Anforderungen der DSGVO, IVDR und des EU AI Act. Unser Anspruch ist es, Lösungen vom ersten Piloten bis zum stabilen Betrieb über die gesamte Laborgruppe zu verantworten.In deiner Rolle verantwortest du die Architektur und treibst folgende Themen gemeinsam mit unserem Team voran:

- Multi-Agenten-Systeme & Tech-Scouting: Architektonische Transformation und Skalierung unseres internen Agenten-Ökosystems. Du orchestrierst spezialisierte KI-Agenten zur maßgeblichen Beschleunigung sowie Automatisierung interner Workflows und treibst durch proaktives Tech-Scouting im Open-Source-Markt (z. B. Evaluierung von LangGraph-Updates) die kontinuierliche Verbesserung unserer Entwicklungszyklen in agilen Sprints voran
- Hybride LLM-Infrastruktur: Evaluation, Verknüpfung und produktiver Einsatz von lokalen vLLM-Inferenzclustern und kommerziellen API-Modellen. Du designst effiziente Architekturen für Review- und Planungs-Tasks, Dokumenteninterpretation und interne Wissensabfragen
- Systemintegration: Entwicklung von MCP-Servern als standardisiertes Interface für unsere Agenten-Ökosysteme. Dies umfasst die deterministische Bereitstellung von Tools, Prompts und Kontextdaten für die sichere Inferenz durch unsere LLMs
- Klinische KI & Konzernweite ML-Lösungen: Entwicklung produktiver ML-Modelle als zentraler Baustein unserer Wertschöpfung. Neben unserem aktuellen Fokus auf den medizinischen Kernbereich evaluierst und implementierst du perspektivisch auch passgenaue ML-Architekturen für operative und kaufmännische Prozesse im gesamten Unternehmensverbund
- ML-Pipelines, Monitoring & Retraining: Aufbau und Betrieb robuster Data- und ML-Pipelines. Du stellst sicher, dass unsere produktiv eingesetzten Modelle an den verschiedenen Laborstandorten performant laufen, kontinuierlich auf Data Drift gemonitort und qualitätsgesichert einem Retraining unterzogen werden. Dies schließt auch die asynchrone Evaluierung unserer Modelle mit ein
- Skalierung im Verbund: Übertragung erprobter, wertschöpfender ML-Architekturen und KI-Systeme von einzelnen Standorten in den gesamten Laborverbund, unter Berücksichtigung technischer, regulatorischer und organisatorischer Anforderungen

Profil

- Modell-Agnostik & Systemarchitektur: Du bewertest und wählst die optimalen Architekturen für unsere diversen Use-Cases aus. Auf Basis fundierter Trade-off-Analysen (Kosten, Latenz, Qualität) designst du deterministische Routing-Logiken – sei es die Orchestrierung kommerzieller API-LLMs, der Einsatz lokaler Open-Weight-Modelle oder die Nutzung klassischer ML-Methoden für strukturierte Labordaten
- Erfahrung im "Agentic Engineering": Du hast nicht nur mit LLMs gechattet, sondern komplexe Workflows, State-Machines oder Multi-Agenten-Systeme programmatisch umgesetzt (idealerweise mit LangChain/LangGraph oder in purem Python) und weißt, wie man "Agentic Loops" in der Produktion stabil und berechenbar hält
- Advanced Tech-Stack & Lokales LLM-Hosting: Hervorragende Python-Kenntnisse und sicherer Umgang mit dem ML-Ökosystem (HuggingFace, PyTorch o.ä.). Du

LIMBACH GRUPPE

Einstiegslevel:
Führungskraft

Standort:
Heidelberg

Tätigkeit:
nicht-medizinisch

Art:
Vollzeit

Unternehmensbereich:
KI

Kontakt:
CV genügt. Hilfreich, aber keine Pflicht, ist ein kurzes Statement zu einem System, das du tatsächlich deployt und betrieben hast: Was war das Problem, was hast du gebaut, was hat nach dem Go-Live nicht funktioniert wie geplant.

Ansprechpartner für erste Fragen ist Herr Alexander Lenard: Alexander.Lenard@limbachgruppe.com

Zum Stellenmarkt:

beherrschst zudem strukturiertes Prompt-Engineering sowie die Bereitstellung von Modellen außerhalb von Cloud-APIs (vLLM, TGI) und wendest Konzepte wie KV-Cache, Quantisierung und Tensor Parallelism auf Multi-GPU-Setups sicher an
- End-to-End MLOps & Automatisierte Evaluierung: Du hast ML-Modelle ausgerollt und anschließend betrieben, inklusive Fehleranalyse, Retraining und Weiterentwicklung nach dem Go-Live. Du nutzt Tracking- und Registrierungs-Frameworks (z. B. MLflow, Weights & Biases) sowie Model-Compiler (z. B. NVIDIA TensorRT) für den Betrieb und implementierst asynchrone Test-Pipelines sowie automatisiertes Benchmarking (z. B. LLM-as-a-Judge) zur Fehleranalyse und Qualitätssicherung
- Praxis-Fokus: Nachweisbare Produktionserfahrung in Computer Vision, NLP/Dokumentenverarbeitung oder LLM-Integration
- Kommunikationsstärke: Du besitzt die Fähigkeit, komplexe technische Systemdesign-Entscheidungen gegenüber Fachärztinnen und Fachärzten und dem Management verständlich, nutzenorientiert und auf Augenhöhe zu vertreten
- Sprachkenntnisse: Fließendes Deutsch (mind. C1) für die interne Kommunikation und sicheres technisches Englisch

Chance

- Du findest bei uns einen attraktiven und modern ausgestatteten Arbeitsplatz vor - inklusive direktem Zugriff auf ein dediziertes Bare-Metal GPU-Cluster im lokalen Rechenzentrum für ungehindertes, IVDR-konformes Prototyping
- Direkte, interdisziplinäre Zusammenarbeit auf Augenhöhe mit der medizinischen Leitung (Product Owner) beim Ausbau und der Skalierung unserer ML-Architektur
- Regelmäßige fachliche und organisatorische Entwicklungsmöglichkeiten stehen dir durch unsere Limbach Akademie zur Verfügung
- Wir bieten dir flexible Arbeitszeiten mit der Möglichkeit zum mobilen Arbeiten
- Ein unbefristetes Arbeitsverhältnis ist für uns selbstverständlich
- Durch unsere gute Verkehrsanbindung kannst du uns auch mit unserem günstigen Deutschland-Ticket oder Dienstrad-Leasing gut erreichen
- Wir bieten dir eine betriebliche Altersvorsorge inklusive Arbeitgeberzuschuss
- Dir stehen in unserem Haus eine eigene Kantine und freie Getränke zur Verfügung
- Regelmäßige Firmenveranstaltungen und Teamevents schaffen eine angenehme Teamatmosphäre

Vorteile

- Job-Ticket
- Sonder-/Zusatzurlaub
- Jubiläumszahlung/ -urlaub
- Behindertengerechter Arbeitsplatz
- Zuschuss zur Betrieblichen Altersvorsorge
- Mitarbeiterevents
- Obstkorb
- Kostenlose Getränke
- Job-Rad
- Gesundheitsförderung
- Mitarbeiterparkplätze
- Betriebsrestaurant / Verpflegungsangebot
- Moderne Ausstattung am Arbeitsplatz
- Weiterbildungsangebote

Sportangebot

Mobiles Arbeiten

Wenn wir Ihr Interesse geweckt haben, freuen wir uns über Ihre vollständige Bewerbung unter Angabe Ihrer Gehaltsvorstellung und des frühestmöglichen Eintrittstermins.