Michael Page www.michaelpage.co.jp

Senior AI Engineer & Architect - Up to 9.5M

Senior AI Engineer & Architect - Tokyo

Job Information

Recruiter
Michael Page

Job ID
1577483

Industry
Software

Job Type
Temporary

Location
Tokyo - 23 Wards

Salary
8.5 million yen ~ 10 million yen

Refreshed
February 9th, 2026 18:13

General Requirements

Career Level
Mid Career

Minimum English Level
Fluent

Minimum Japanese Level
Basic

Minimum Education Level
Bachelor's Degree

Visa Status
Permission to work in Japan required

Job Description

This position is ideal for a senior AI engineer who wants to architect and deliver highly scalable AI systems-spanning multi-agent frameworks, RAG pipelines, NLP/NLU, and MLOps. You will be a core technical driver, guiding design, development, deployment, optimization, and long-term evolution of an enterprise AI automation platform.

Client Details

Our client is a **large, technology-driven telecommunications & digital services company** undergoing rapid expansion in AI-powered automation. Operating at national scale and known for its advanced network capabilities, the company is investing heavily in internal AI platforms to transform operational efficiency and intelligent automation.

Description

In this senior role, you will architect, build, and optimize production-grade AI systems with a focus on **multi-agent orchestration**, **LLM integration**, **RAG pipelines**, and enterprise-ready **MLOps**. Key responsibilities include:

- Design end-to-end AI architecture for multi-agent systems and LLM orchestration (e.g., LangChain, LangGraph, Agno).

- Lead development of the in-house AI automation product, ensuring scalability, reliability, and performance.
- Build and optimize RAG systems using vector databases and graph-based retrieval approaches.
- Integrate and fine-tune LLM APIs from multiple providers for diverse use cases.
- Develop NLP/NLU pipelines for semantic understanding, intent classification, and entity extraction.
- Implement AI agent frameworks for complex automated workflows.
- Design CI/CD for ML models, monitoring systems, retraining pipelines, and A/B test environments.
- Manage model versioning, experiment tracking, and deployment automation.
- Architect scalable backend systems for AI workloads and resource optimization.
- Collaborate with product, data science, and engineering teams to convert business needs into technical solutions.
- Establish best practices for AI system development, testing, and production reliability.
- Mentor team members and lead technical discussions.
- Evaluate emerging AI technologies and integrate new frameworks.
- Document architecture, design decisions, and operational processes.
- (Optional) Develop front-end components for AI applications using React.

Job Offer

- Strategic ownership in a high-investment AI transformation program
- Opportunity to build advanced, scalable AI systems used across the organization
- Work with cutting-edge LLM, RAG, and AI orchestration technologies
- Collaborative, innovation-driven engineering culture
- Career progression into architecture leadership and AI platform strategy

To apply online please click the 'Apply' button below. For a confidential discussion about this role please contact Serena Wu on +81 3 6627 6137.

Required Skills

You will thrive in this role if you bring strong AI/ML engineering experience, deep technical ownership, and a passion for large-scale AI infrastructure. Core requirements:

- Bachelor's degree in Computer Science, AI, ML, or equivalent experience.
- 5+ years AI/ML engineering experience, including 2-3+ years deploying production AI systems.
- Proven experience leading AI/ML projects or technical teams.
- Expert Python skills and strong experience with LLM orchestration frameworks (LangChain, LangGraph, etc.).
- Hands-on experience building RAG systems and working with vector databases (Pinecone, Weaviate, ChromaDB, etc.).
- Strong experience integrating LLM APIs (OpenAI, Anthropic, etc.).
- Solid understanding of NLP/NLU methods and AI agent architectures.
- MLOps experience including CI/CD, monitoring, retraining workflows, and A/B test frameworks.
- Experience with cloud platforms (AWS/Azure/GCP) and backend development (FastAPI, Flask, Django).
- Strong debugging ability and communication skills.

Preferred:

- Master's/PhD in AI/ML; experience with Agno, React front-end, transformer fine-tuning, MLOps platforms (MLflow, Kubeflow, Vertex AI), AI safety, distributed computing (Kubernetes/Docker), multimodal AI, or open-source contributions.

Languages:
English fluent (mandatory). Japanese is a plus.

Company Description

Our client is a large, technology-driven telecommunications & digital services company undergoing rapid expansion in AI-powered automation. Operating at national scale and known for its advanced network capabilities, the company is investing heavily in internal AI platforms to transform operational efficiency and intelligent automation.