No Japanese Required Software Engineer, AI Platform

English speaking role

Job Information

Recruiter
PROGRE Ltd

Hiring Company
AI company

Job ID
1561704

Industry
Internet, Web Services

Company Type
Large Company (more than 300 employees)

Non-Japanese Ratio
About half Japanese

Job Type
Permanent Full-time

Location
Tokyo - 23 Wards, Shinjuku-ku

Salary
8 million yen ~ 14 million yen

Salary Bonuses
Bonuses included in indicated salary.

Work Hours
10:00 ~ 19:00

Holidays
完全週休二日制 所定休日：土・日・祝日 休暇：年次有給休暇、夏季休暇（3日）、年末年始休暇（12月31日〜1月3日）、慶弔休暇

Refreshed
June 24th, 2026 07:00

General Requirements

Minimum Experience Level
Over 6 years

Career Level
Mid Career

Minimum English Level
Business Level (Amount Used: English Only)

Minimum Japanese Level
None

Minimum Education Level
Bachelor's Degree

Visa Status
Permission to work in Japan required

Job Description

Role & Expectations

As a Software Engineer (AI Platform), you will power the reliability, performance, and cost efficiency of the entire AI platform through backend engineering.

- Design, implement, and operate backend services while also optimizing Kubernetes clusters and cloud infrastructure
- Design and build observability infrastructure (tracing, logging, metrics) to rapidly detect and resolve failures unique to AI agents
- Deliver improvements with direct business impact through inference cost and infrastructure cost optimization
- Maintain 99.9% uptime through SLI/SLO design and operations, on-call, and incident response
- Improve developer experience for in-house engineers through CI/CD pipeline construction and development environment improvements

Why You'll Love This Role

- **At the intersection of Backend × Infrastructure** — A new domain where you support the entire platform through the power of backend engineering.
- **Platform engineering for the AI era** — Go beyond traditional infrastructure / SRE to tackle AI-specific challenges: inference cost optimization, GPU management, agent tracing, and more.
- **Large-scale cloud infrastructure design** — Gain experience designing and operating large-scale distributed systems with Kubernetes, event-driven architectures, and autoscaling.
- **Cost optimization with real impact** — Inference and infrastructure cost optimization directly translates to business impact. Improving $/request ripples across all products.
- **Powering every product** — Support 99.9% uptime for a production environment used by ~200 companies. Every AI agent runs on the infrastructure you build.
- **Rapid-growth environment** — In a startup that has grown to 200+ people and 9 products in just 3 years, you will have significant autonomy in technical decision-making.

Job Description

- **Backend Services & Platform Development**
  - Design, implement, and operate backend services for the AI platform
  - Design, build, and operate Kubernetes clusters
  - Architect and optimize cloud infrastructure (GCP)
  - Codify and automate infrastructure with IaC (Terraform)
  - Cost/performance optimization (autoscaling, caching, batch processing, GPU management)
- **Observability & Governance**
  - Design and build the observability stack (tracing, logging, metrics)
  - Implement AI agent-specific tracing (inference request tracking, tool call visualization)
  - Build data access and permission management infrastructure
  - Address security requirements
- **SRE & Reliability**
  - Maintain platform uptime of ≥99.9%
  - Design and operate SLIs / SLOs
  - On-call, incident response, and post-mortems
  - Continuously improve incident MTTR
- **Developer Experience**
  - Build and improve CI/CD pipelines
  - Maintain development and staging environments
  - Create and maintain infrastructure documentation for internal engineers

Work Style

Hybrid work model: 3 days in the office, 2 days remote.

Required Skills

Minimum Qualifications

- Bachelor's degree or equivalent practical experience in Computer Science, Software Engineering, Artificial Intelligence, Machine Learning, Mathematics, Physics, or related fields
- 3+ years of practical experience as a backend engineer
- Production product development experience in Python
- Design and operations experience on cloud platforms (AWS / GCP / Azure)
- Understanding and operational experience with Kubernetes / container orchestration
- Distributed system design and operations experience

Preferred Qualifications

- IaC practical experience (Terraform / Pulumi, etc.)
- GPU cluster operations and optimization experience
- ML infrastructure / MLOps construction experience
- AI workload operations experience (inference servers, model serving)
- Event-driven architecture experience (Kafka / RabbitMQ, etc.)

- SRE / DevOps practices (SLI / SLO design, Chaos Engineering, etc.)
- Security engineering experience
- Technical communication ability in English

Company Description