No Japanese Required Agent Harness Engineer

English speaking role

Job Information

Recruiter
PROGRE Ltd

Hiring Company
AI company

Job ID
1561703

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
9 million yen ~ 16 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
October 28th, 2026 03:00

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 an Agent Harness Engineer, you will design and implement the agent control and execution infrastructure, leveraging your AI/ML knowledge.

- Design and implement the execution engine (Graph Runtime / State Machine) with deep understanding of LLM / AI agent operating principles
- Own AI-specific system design including model routing, context management, and memory infrastructure (long-term memory, working memory)
- Design and develop the Agent SDK used by 120 in-house engineers
- Build the guardrail / policy execution engine to safely control agent behavior
- Collaborate with Research Engineers to integrate the latest research outcomes into the production infrastructure

Job Description

- **Agent Harness design & implementation**
  - Design and implement the agent execution engine (Graph Runtime / State Machine)
  - Design and develop the Agent SDK — the interface for in-house engineers to build agents
  - Implement session management, checkpoint, and recovery mechanisms
  - Build the guardrail / policy execution engine — a rule execution infrastructure that controls agent behavior
- **AI/ML System Integration**
  - Model routing — optimal routing of inference requests across multiple LLM providers and model types
  - Design context management and memory infrastructure (long-term memory, working memory, RAG integration)
  - Optimize inference pipelines (latency reduction, cost efficiency, caching strategies)
  - Integrate latest research findings into the production infrastructure in collaboration with Research Engineers
- **Orchestration & performance**
  - Develop workflow orchestration and queuing systems
  - Cost/performance optimization (autoscaling, caching, batch processing)
  - Inference request routing and load balancing
- **Reliability & Operations**
  - Maintain platform uptime of ≥99.9%
  - Incident response and post-mortems
  - Design data access and permission management infrastructure

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
- 5+ years of practical experience as a backend engineer
- Production product development experience in Python
- Experience designing and implementing production systems that leverage LLM / AI agents
- Experience designing and implementing distributed systems (including design and coding, not just operations)
- Experience designing and implementing RESTful APIs / gRPC

Preferred Qualifications

- Agent Framework / Agent Harness design and implementation experience (LangChain / LangGraph / AutoGen, etc.)
- Production operations experience on cloud platforms (AWS / GCP / Azure)
- Understanding of RAG systems, vector databases, and memory architectures
- Model routing and inference optimization experience
- Foundation software development experience in Go (SDKs, runtimes, frameworks, etc.)
- Deep understanding of Kubernetes / container orchestration
- Event-driven architecture experience (Kafka / RabbitMQ, etc.)
- Experience implementing safety guardrails, policy execution, and AI observability
- ML infrastructure / MLOps construction experience
- Technical communication ability in English

Company Description