# No Japanese Required Agent Harness Engineer

**English speaking role**

## 募集職種

**人材紹介会社**
株式会社PROGRE

**採用企業名**
AI company

**求人ID**
1561703

**業種**
インターネット・Webサービス

**会社の種類**
大手企業 (300名を超える従業員数)

**外国人の割合**
外国人 半数

**雇用形態**
正社員

**勤務地**
東京都 23区, 新宿区

**給与**
900万円 ~ 1600万円

**ボーナス**
給与： ボーナス込み

**勤務時間**
10:00 ~ 19:00

**休日・休暇**
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

**更新日**
2026年06月24日 07:00

## 応募必要条件

**職務経験**
6年以上

**キャリアレベル**
中途経験者レベル

**英語レベル**
ビジネス会話レベル (英語使用比率: 常時英語)

**日本語レベル**
無し

**最終学歴**
大学卒： 学士号

**現在のビザ**
日本での就労許可が必要です

## 募集要項

**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.

## スキル・資格

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

## 会社説明