CareerCross by JAC Recruitment

PROGRE

No Japanese Required AI Engineer (RAG Specialist)

Engineer specializing in RAG in English

Job Information

Recruiter
PROGRE Ltd

Hiring Company
AI company leading Japan AI market

Job ID
1554604

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
7 million yen ~ 13 million yen

Refreshed
August 10th, 2025 06:18

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

Key Responsibilities

RAG System Design & Operation

- Architect and implement RAG-based systems
- Build and optimize vector databases (e.g., FAISS, Elasticsearch, Pinecone)

- Develop document preprocessing and chunking strategies

Maintenance & Monitoring

- Operate and monitor RAG systems in production
- Analyze performance, identify bottlenecks, ensure stability

Accuracy Improvement

- Evaluate and improve retrieval/response quality
- Apply prompt engineering and model fine-tuning
- Define and implement evaluation metrics

R&D and Technical Validation

- Explore and validate cutting-edge RAG methods
- Build POCs (Proof of Concept)
- Contribute to architecture and tool decisions

Development

- Build and extend RAG-related features
- Design and implement APIs
- Contribute to both frontend/backend development when needed

Team Structure

Our development team consists of approximately 65 members, structured into the following groups:

- Client-Facing Solution Development
- In-house AI SaaS Development
- Common Platform Development (Infra / Data / AI R&D)

Required Skills

Minimum Qualifications

- Bachelor’s degree or equivalent experience in CS, AI, ML, Mathematics, or related fields
- Hands-on experience designing, building, and operating RAG systems
- Experience with vector DBs (e.g., FAISS, Elasticsearch, Pinecone)
- Document preprocessing and chunking strategies
- Production monitoring and incident response
- System optimization and performance tuning
- Evaluation and improvement of search/generation quality
- Prompt engineering & fine-tuning
- Metric design and implementation
- API development (e.g., RESTful APIs)
- Python development (including ML/NLP libraries)
- Experience with cloud (AWS, GCP, etc.)
- Team collaboration and communication skills
- English: Business level or above

Preferred Qualifications

- Frontend/backend experience (React, Vue.js, Flask, FastAPI, etc.)
- LLM tuning and evaluation experience
- CI/CD experience (GitHub Actions, Jenkins)
- Containerization (Docker, Kubernetes)
- Familiarity with RAG-related academic papers and trends
- POC development experience
- Data engineering knowledge (ETL, pipelines)
- Strong reading skills of English technical documents

Tech Stack / Tools

- Languages: Python (Backend), TypeScript / React / Next.js (Frontend) / NX
- Infrastructure: GCP (Kubernetes), Docker
- Tools: Slack, Confluence, Linear, Google Workspace, GitHub, Notion
- Hardware: Mac (Apple Silicon), dual monitors

Company Description