CareerCross by JAC Recruitment

PROGRE

No Japanese Required AI Engineer (RAG Specialist)

Engineer specializing in RAG in English

募集職種

人材紹介会社
株式会社PROGRE

採用企業名
AI company leading Japan AI market

求人ID
1554604

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
700万円 ~ 1300万円

更新日
2025年08月10日 06:18

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

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)

スキル・資格

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

会社説明