No Japanese Required Lead Backend Engineer AI Speech

English speaking role

Job Information

Recruiter
PROGRE Ltd

Hiring Company
AI company

Job ID
1561705

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
10 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
November 26th, 2025 00:00

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 Summary

As a **Backend Engineer (Tech Lead)** , you will serve as a technical leader within the Speech division of our Solution Architect team.
You will be responsible for defining the technical strategy, designing the architecture, leading multiple projects, mentoring engineers, and engaging in advanced technical discussions with clients.
You will play a central role in driving the success of our AI Speech solutions both technically and organizationally.

Why You'll Love This Role

Lead the Technical Strategy:
Play a key role in defining the technical direction and architecture for cutting-edge speech AI solutions, influencing the future of our organization.

Cutting-edge AI Development:
Lead the design and development of next-generation speech dialogue systems integrating ASR, TTS, and LLM technologies.

Solve Complex, High-Impact Problems:
Work directly on large-scale enterprise projects to solve real-world challenges through technology.

Social Impact:
Contribute to solutions that enhance enterprise productivity—such as call center automation and sales enablement—creating tangible value for businesses.

Leadership and Mentorship:
Guide and develop engineers, support hiring activities, and foster technical excellence across the team.

Wide Technical Exposure:
Gain deep experience across domains such as real-time speech processing, CTI integration, cloud infrastructure, and machine learning operations.

High Autonomy & Agility:
Experience the speed and decision-making flexibility unique to a startup environment.

Responsibilities

Technical Strategy & Architecture Design

- Define and execute the technical roadmap for the Speech domain
- Design architectures for large-scale, high-availability, and low-latency systems
- Make key decisions on technologies (languages, frameworks, infrastructure, tools)
- Identify and address technical debt and drive continuous improvement
- Develop strategies for security, performance, and cost optimization

AI Speech Solution Development

- Lead design and development of custom client solutions using "JAPAN AI Speech"
- Architect real-time speech processing systems (ASR, TTS, dialogue)
- Build low-latency streaming platforms using WebRTC, Twilio, or SIP
- Design and implement LLM-powered conversational systems

System Integration & API Design

- Design large-scale API integrations with existing CRM/CTI/business systems
- Architect and implement microservices-based distributed systems
- Standardize REST/gRPC/GraphQL API design best practices
- Build event-driven architectures for scalability and responsiveness

Infrastructure & Operations

- Design and optimize containerized Kubernetes environments on GCP
- Manage database architecture and query optimization (Postgres, Redis, Elasticsearch)
- Operate streaming and batch data pipelines (Kafka, Pub/Sub, etc.)
- Build and enhance CI/CD pipelines (GitHub Actions, ArgoCD)
- Design observability systems (Prometheus, Grafana, ELK, Jaeger)
- Drive SRE best practices (SLA/SLI design and reliability improvement)

AI & ML Integration

- Plan and execute production integration of ASR/TTS models
- Optimize inference performance (Triton, TorchServe, ONNX)
- Design and maintain MLOps pipelines
- Collaborate with ML and Data Science teams on applied AI initiatives

Team Leadership

- Mentor and guide engineers technically and professionally
- Conduct code and design reviews to maintain high-quality standards
- Participate in hiring and technical interviews
- Improve development processes, standards, and documentation
- Share technical knowledge through internal sessions and written materials

Client Engagement & Project Leadership

- Lead technical discussions, requirements definition, and solution proposals
- Design and execute PoC (Proof of Concept) for large-scale projects
- Provide technical direction and progress management across projects
- Ensure post-deployment maintenance, scalability, and compliance with security standards

Work Style

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

Required Skills

Minimum Qualifications

- Bachelor’s degree (or equivalent experience) in Computer Science, Software Engineering, AI, Machine Learning, Mathematics, or a related field

  5+ years of experience developing applications in **Python**

  3+ years as a **System Architect or Technical Lead**

  Strong hands-on experience with real-time communication (WebRTC, WebSocket, gRPC) or streaming systems (Kafka, Pub/Sub)

  Proven track record designing architectures for large-scale, highly available, and cost-efficient systems

  5+ years of experience with cloud services (GCP preferred; AWS/Azure acceptable) and container technologies (Docker, Kubernetes)

  Experience designing and operating RDB/NoSQL systems (Postgres, Redis) and object storage (GCS/S3)

  Experience mentoring engineers and providing technical leadership

  Experience in technical client communication, requirement definition, and custom solution delivery

  Experience working in a startup or fast-growing environment

  Strong knowledge and passion for AI-related technologies

Preferred Qualifications

- Deep experience or research background in ASR, TTS, or speech signal processing

  Production experience with model inference platforms (Triton, TorchServe, ONNX Runtime, TensorRT)

  Experience with CTI integration (Twilio, SIP, Asterisk)

  Experience operating large-scale data warehouses (BigQuery, Redshift, Snowflake) or data lakes

  Experience designing or implementing LLM- or RAG-based dialogue systems

  MLOps architecture and implementation experience

  Experience with NLP or ML-based systems development

  Hands-on experience with machine learning using Python, TensorFlow, or PyTorch

  Experience developing chatbots, speech recognition, or AI agent systems

  Full-stack experience (TypeScript/React)

  Experience launching new products or services from 0→1 phase

  Bilingual proficiency (Japanese/English)