AI QA Specialist (LLM Evaluation)

Job Information

Recruiter
PROGRE Ltd

Hiring Company
AI QA Specialist (LLM Evaluation)

Job ID
1560023

Industry
Internet, Web Services

Company Type
Small/Medium Company (300 employees or less)

Non-Japanese Ratio
About half Japanese

Job Type
Permanent Full-time

Location
Tokyo - 23 Wards, Shinjuku-ku

Salary
7 million yen ~ 14 million yen

Work Hours
10:00〜19:00

Refreshed
April 1st, 2026 06:00

General Requirements

Minimum Experience Level
Over 6 years

Career Level
Mid Career

Minimum English Level
Business Level (Amount Used: English usage about 25%)

Minimum Japanese Level
None

Minimum Education Level
Technical/Vocational College

Visa Status
Permission to work in Japan required

Job Description

As an AI QA Specialist, you will lead the design, construction, and operation of the quality evaluation infrastructure for AI agents.

- Own the entire process from evaluation metric selection and design to integrating automated evaluation pipelines into CI/CD
- Plan and execute red teaming to detect safety risks before release

- Quantitatively verify the effectiveness of quality improvements through A/B test analysis based on statistical experimental design
- Feed evaluation signals back to the research and development teams, creating a compound-interest loop for model improvement
- Ensure the quality of products used in production by ~200 companies through a "science of quality" approach

Job Description

- **Evaluation Infrastructure Design & Development**
  - Design, build, and maintain evaluation sets (synthetic data + real logs)
  - Select and design evaluation metrics (win rate, task success, factuality, harm detection)
  - Build automated evaluation pipelines and integrate them into CI/CD
  - Design agent harnesses (multi-turn, tool use, long-context support)
- **Safety & Quality Verification**
  - Plan and execute red-teaming (adversarial testing)
  - Build safety and policy compliance verification frameworks
  - Design and run prompt/tool regression tests
  - Analyze and improve issues related to hallucination, bias, and output quality
- **Statistical Analysis & Reporting**
  - Design and analyze statistical experiments (A/B tests, significance testing)
  - Create quality reports and improvement proposals
  - Visualize regression detection and quality trends
  - Feed evaluation signals back to research and development teams

Required Skills

You May Be a Good Fit If You

- Bachelor's degree or equivalent practical experience in Computer Science, Software Engineering, Artificial Intelligence, Machine Learning, Mathematics, Physics, or related fields
- 3+ years of practical experience as a software engineer or QA engineer
- Knowledge of LLM / generative AI evaluation methods (prompt evaluation, quantitative output quality measurement, hallucination detection, etc.)
- Foundational knowledge of statistics and experimental design
- Experience building evaluation pipelines in Python
- Experience integrating tests into CI/CD pipelines
- Experience designing prompt / tool regression tests

Strong Candidates May Also Have

- NLP / ML evaluation benchmark design experience
- Knowledge of AI safety / Responsible AI
- Red teaming / penetration testing experience
- Experience evaluating multi-agent workflows, tool use, and long-context scenarios
- Large-scale data processing experience (Spark / BigQuery, etc.)
- Ability to read, comprehend, and reproduce research papers
- Technical communication ability in English

Company Description