Minatomirai Station | Inference Systems Engineer @AI startup

Job Information

Hiring Company
Unsung Fields Corp.

Job ID
1573179

Industry
Software

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

Job Type
Permanent Full-time

Location
Kanagawa Prefecture, Yokohama-shi Nishi-ku

Train Description
Minatomirai Line, Minatomirai Station

Salary
8 million yen ~ 14 million yen

Work Hours
09:00 - 18:00（60-minute break）

Holidays
Two-day weekends,holidays,special leaves,120+ days off annually

Refreshed
February 5th, 2026 13:00

General Requirements

Minimum Experience Level
Over 6 years

Career Level
Mid Career

Minimum English Level
Business Level

Minimum Japanese Level
None

Minimum Education Level
Bachelor's Degree

Visa Status
No permission to work in Japan required

Job Description

Inference Systems Engineer (LLM Serving Runtime + Performance)

Role overview

As an Inference Systems Engineer, you will own the serving runtime that powers production LLM inference. This is a deeply technical role focused on system performance and stability: optimizing request lifecycle behavior, streaming correctness, batching/scheduling strategy, cache and memory behavior, and runtime execution efficiency. You will ship changes that improve TTFT, p95/p99 latency, throughput, and cost efficiency—while preserving correctness and reliability under multi-tenant load.

You will collaborate closely with platform/infrastructure operations, networking, and API/control-plane teams to ensure the serving system behaves predictably in production and can be debugged quickly when incidents occur. This role is for engineers who can reason about the entire inference pipeline, validate improvements with rigorous measurement, and operate with production-grade discipline.

Responsibilities

- Own the end-to-end serving runtime behavior: request lifecycle, streaming semantics, cancellation, retries interaction, timeouts, and consistent failure modes.
- Design and implement batching and scheduling strategy: dynamic batching, admission control, fairness under mixed tenants, priority lanes, and backpressure mechanisms to prevent cascading failures.
- Optimize performance at the systems level: reduce time-to-first-token, improve tail latency stability, increase tokens/sec throughput, and improve accelerator utilization under realistic workloads.
- Improve memory behavior and cache efficiency: KV-cache policies, fragmentation control, eviction strategies, and safeguards against OOM cliffs and performance thrash.
- Drive runtime execution optimizations: operator-level improvements, quantization integration, compilation/tuning paths where appropriate, and parameterization that produces stable performance across deployments.
- Establish a performance measurement discipline: reproducible benchmarks, realistic traffic traces, profiling workflows, regression detection gates, and dashboards tied to production outcomes.
- Build production readiness into the system: feature-flagged rollouts, canarying, safe configuration changes, and incident playbooks that reduce MTTR.
- Partner with networking and infrastructure operations to align deployment topology, failure domains, and capacity constraints to performance and reliability goals.
- Collaborate with product and API teams to ensure the serving layer’s guarantees are reflected accurately in external interfaces and customer expectations.

[Employment Type]
Full-time employee
*Probationary period: 3 months

[Salary]
Annual Salary: ¥8,000,000 - ¥14,000,000
Monthly Salary: ¥666,667 - ¥1,166,667 (Monthly Base Salary: ¥666,667 - ¥1,166,667)
■Salary Increases: Available

[Working Hours]
9:00 AM - 6:00 PM (60-minute break)

[Work Location]
Queen's Tower A, 10th Floor, 2-3-1 Minatomirai, Nishi-ku, Yokohama, Kanagawa Prefecture, 220-6010
■Access: 7-minute walk from Sakuragicho Station (all lines), direct access from Minatomirai Station (Toyoko Line, Minatomirai Line)
■Non-smoking workplace
■Changes to work location: Company-designated offices
■Transfers/Secondments: None

[Holidays and Leave]
120 days off per year Days
Full two-day weekend
Annual paid vacation (minimum 10 days after the seventh month of employment)

[Benefits]
Partial transportation allowance (up to ¥15,000 per month)
Social insurance (health insurance, employee pension insurance, employment insurance, workers' compensation insurance)
Overtime pay: Standard overtime pay

Required Skills

Requirements

- 5+ years building high-performance systems (model serving, GPU systems, performance engineering, or low-latency distributed systems).
- Strong understanding of LLM inference tradeoffs: batching vs latency, prefill vs decode dynamics, cache behavior, memory pressure, and tail latency causes.
- Comfort working across Python/C++ stacks with production profiling and debugging tools.
- Track record of shipping performance improvements that hold up under production variance and operational constraints.
- Strong engineering hygiene: tests, instrumentation, documentation, and careful rollout discipline.
- Ability to communicate clearly across teams and operate calmly during incidents.