Minatomirai Station | Inference Systems Engineer @AI startup

募集職種

採用企業名
株式会社Unsung Fields

求人ID
1573179

業種
ソフトウエア

会社の種類
中小企業 (従業員300名以下)

雇用形態
正社員

勤務地
神奈川県, 横浜市西区

最寄駅
みなとみらい線、　みなとみらい駅

給与
800万円 ~ 1400万円

勤務時間
09:00 - 18:00（60-minute break）

休日・休暇
Two-day weekends,holidays,special leaves,120+ days off annually

更新日
2026年03月19日 13:00

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

日本語レベル
無し

最終学歴
大学卒： 学士号

現在のビザ
日本での就労許可は必要ありません

募集要項

Inference Systems Engineer (LLM Serving Runtime + Performance)

Role overview

As an inference & serving engineer, your objective is to build a high-performance, multi-tenant serving stack that squeezes maximum utilization out of heterogeneous hardware. This involves navigating the trade-offs between various state-of-the-art

inference frameworks and engines, selecting and optimizing the right runtime for the right workload. The scope of work is not limited to Large Language Models; it extends to the frontier of Generative AI, including high-throughput Video generation and complex Multimodal systems where memory pressure and compute requirements are significantly more demanding.

Beyond just deploying models at scale, this role is responsible for building a robust system that bridges the gap between boutique, high-performance clusters and massive, multi-node deployments as the company grows. This requires a deep understanding of the "Inference Triangle"—constantly tuning the stack to find the optimal equilibrium between low-latency (TTFT/ITL), high-throughput, and inference quality (Precision/Quantization). The ideal candidate is a hands-on engineer who views the entire GPU fleet as a single, programmable compute fabric and is eager to get their hands dirty at every level of the stack.

Responsibilities

- **Runtime Selection & Deep Optimization:** Lead the evaluation, integration, and continuous tuning of diverse inference frameworks to ensure best-in-class performance across LLM, Video, and Multimodal workloads.
- **Latency & Throughput Engineering:** Own the end-to-end performance profile of the model lifecycle, implementing advanced strategies such as disaggregated prefill/decode, speculative decoding, and continuous batching to minimize TTFT and maximize tokens-per-second.
- **Scalable Systems Evolution:** Design and implement serving architectures that function seamlessly on small experimental clusters while providing a clear, robust path to massive-scale, multi-node deployments.
- **Advanced Memory & Cache Orchestration:** Implement and optimize memory management techniques to maximize KV-cache reuse and minimize redundant computations in multi-turn or high-concurrency scenarios.
- **Day 0 Model Support:** Working with the ecosystem, craft a Day 0 model support strategy ensuring our stack provides stable, high-performance support for new models when they are released.]
- **Cross-Stack Integration:** Collaborate with the Backend/Gateway and Compute Orchestration teams to ensure the inference engine’s telemetry, failure domains, and lifecycle management are perfectly aligned with the global load balancer and API layers.
- **Hands-on Technical Leadership:** Maintain a high level of personal agency by writing production code, debugging complex distributed system "hangs," and contributing to architectural decisions in a flat, fast-moving team environment.
- **Collaborative Communication:** Function as a primary technical peer to engineering leads, translating complex hardware and model constraints into clear product and infrastructure strategies.
- **Inference Strategy & Trade-offs:** Define path forward when balancing model precision and quantization against the physical limits of HBM bandwidth and compute throughput.
- **(Optional) Kernel-Level Development:** Dive into the lowest levels of the execution stack to develop and refine custom CUDA or Triton kernels, eliminating overhead in the execution loop and optimizing for specific hardware primitives.

[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

スキル・資格

You may be a fit if you have the following skills:

- **Inference Engine:** Deep experience with the internals of modern runtimes. You are a prominent contributor to inference engine ecosystems, including but not limited to OSS projects or proprietary engines at top-tier AI labs.
- **Multimodal Domain Knowledge:** Understanding of the specific challenges involved in serving Large Language Models alongside Video and Vision-based generative models.

- **Scale-First Engineering:** A track record of building and managing distributed systems that have evolved from small-scale proofs-of-concept to large-scale production deployments.
- **Great Team Spirit:** A mission-driven approach to engineering, valuing clear communication, hands-on execution, and collective success over individual silos.

会社説明