≪みなとみらい駅直結≫推論システムエンジニア｜AI課題を解決するエンジニア｜クラウド・AI開発経験歓迎

Job Information

Hiring Company
Unsung Fields Corp.

Job ID
1573146

Industry
Software

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

Job Type
Permanent Full-time

Location
Kanagawa Prefecture, Yokohama-shi Nishi-ku

Salary
8 million yen ~ 14 million yen

Work Hours
09:00 ~ 18:00（休憩時間 60分）

Holidays
週休2日制（土日祝）、年末年始休暇、慶弔休暇、年間休日120日以上

Refreshed
March 5th, 2026 13:00

General Requirements

Minimum Experience Level
Over 6 years

Career Level
Mid Career

Minimum English Level
Daily Conversation

Minimum Japanese Level
Native

Minimum Education Level
Bachelor's Degree

Visa Status
No permission to work in Japan required

Job Description

推論システムエンジニア（LLM サービングランタイム + パフォーマンス）

【職務概要】

推論システムエンジニアとして、本番環境におけるLLM推論を支えるサービングランタイムを担当します。本ポジションは非常に技術的で、システムのパフォーマンスと安定性に焦点を当てています。具体的には、リクエストのライフサイクル動作の最適化、ストリーミングの正確性、バッチ処理やスケジューリング戦略、キャッシュおよびメモリ挙動、ランタイム実

行効率などを改善します。マルチテナント負荷下においても正確性と信頼性を維持しながら、TTFT（Time to First Token）、p95／p99 レイテンシ、スループット、コスト効率を改善させる変更を実装していきます。
プラットフォーム／インフラ運用、ネットワーキング、API／コントロールプレーンの各チームと密接に連携し、サービングシステムが本番環境で予測可能に振る舞い、インシデント発生時にも迅速にデバッグできる状態を確保します。本ポジションは、推論パイプライン全体を俯瞰して論理的に考え、厳密な計測に基づいて改善効果を検証し、本番環境レベルの慎重さと規律をもって運用できるエンジニアを対象としています。

【業務内容】

- サービングランタイムのエンドツーエンドな挙動を担当：リクエストのライフサイクル管理、ストリーミンセマンティクス 、キャンセル処理、リトライとの相互作用、タイムアウト設定、一貫した障害モードの設計・運用。
- バッチ処理およびスケジューリング戦略の設計・実装を担当：動的バッチ処理、アドミッションコントロール、複数テナントが混在する環境下での公平性、優先レーン、連鎖的障害を防ぐためのバックプレッシャー機構。
- システムレベルでの性能最適化：Time-to-first-tokenの短縮、テールレイテンシの安定化、tokens/secスループットの向上、現実的なワークロード下でのアクセラレータ利用効率の改善。
- メモリ挙動およびキャッシュ効率の改善：KVキャッシュのポリシー設計、フラグメンテーション制御、エビクション戦略、OOM（メモリ不足）やパフォーマンス低下を防ぐための安全策の設計。
- ランタイム実行最適化の推進：オペレータ単位での改善、量子化手法の統合、必要に応じたコンパイル／チューニング経路、環境へのデプロイにおいて安定した性能を生み出すパラメータ設計。
- 性能計測に関する規律の確立：再現可能なベンチマーク、現実的なトラフィックトレース、ワークフローのプロファイリング、回帰検知ゲート、本番での成果指標と連動したダッシュボード。
- システムへ本番運用に耐えうる仕組の組み込み：機能フラグを用いた段階的ロールアウト、カナリアリリース、安全な設定変更、MTTR（平均復旧時間）を短縮するインシデント対応手順書。
- ネットワーキングおよびインフラ運用チームと連携し、デプロイメントトポロジー、障害ドメイン、容量制約を性能および信頼性目標に適合させます。
- プロダクトおよびAPIチームと協力し、サービングレイヤーの保証事項が外部インターフェースおよび顧客の期待に正確に反映されるよう確保します。

【雇用形態】
正社員
※試用期間あり、3ヶ月

【給与】
年収：800万円～1,400万円
月収：66.6万円～116.6万円（月額基本給：66.6万円～116.6万円）
■昇給：あり

【就業時間】
09:00 ~ 18:00（休憩時間 60分）

【勤務地】
〒220-6010　神奈川県横浜市西区みなとみらい2丁目3番1号 クイーンズタワーA 10階
■アクセス：各線 桜木町 駅から徒歩7分、東横線・みなとみらい線 みなとみらい 駅から直結
■就業場所全面禁煙
■勤務地変更範囲：会社の定める事業所
■転勤・出向：無し

【休日休暇】

- 年間休日 120 日
- 完全週休二日制
- 年間有給休暇（入社7ヶ月目には最低10日以上）

【待遇・福利厚生】

- 交通費 一部支給 (上限月1万5千円)
- 社会保険（健康保険、厚生年金、雇用保険、労災保険）
- 残業手当：通常の残業代

Required Skills

【必須要件】

- 高性能システムの構築経験5年以上（モデルサービング、GPU システム、パフォーマンスエンジニアリング、低レイテンシ分散システムなど）。
- LLM 推論におけるトレードオフへの深い理解：バッチ処理とレイテンシの関係、プリフィルとデコードの動特性、キャッシュ挙動、メモリプレッシャー、テールレイテンシの要因
- Python／C++スタックを横断して使用し、本番環境でのプロファイリング・デバッグツールを活用した業務に慣れていること。
- 本番環境の変動や運用上の制約下でも持続する性能改善をリリースしてきた実績。
- 高いサイバーハイジーン意識に基づくエンジニアリング規律：テスト、計測、ドキュメンテーション、慎重なロールアウト運用の徹底。
- チーム横断で明確にコミュニケーションでき、インシデント時にも冷静に対応できる能力。

Company Description