Hani Mir

hmir95@gmail.com
linkedin.com/in/hanimir

Summary

- 8+ years of experience building distributed systems and LLM training/serving infrastructure
- Scaled 70B chat models to 20k+ QPS across 1k+ GPUs, cut inference costs by 75%, safety alignment resulted in 12% unsafe response reduction

Experience

Character.ai

San Francisco, CA
January 2025 - Present

Staff Software Engineer - ML Infrastructure

- Designed H100-based post-training pipeline to fine-tune 70B chat models on ~20B tokens; lifted RPG-character engagement ~10%
- Trained and served real-time classification and safety models (F1=0.75); scaled throughput from ~5 QPS to ~500 QPS (p95 latency <150ms)
- Profiled and optimized classifier inference engine; raised GPU utilization from 30% to 95%
- Zero to one development of parameter-efficient fine-tuning (LoRA/QLoRA) pipeline; enabled nightly retraining with 75% less GPU-hours
- Added user/character-aware model routing; increased user engagement 8% (p95 TTFT <2.5s at 20k+ QPS)
- Designed and implemented two-tower character recommendation model, feature extraction, and real-time serving pipeline; improved For You CTR by 50%

PyTorch, Kubernetes, Slurm, vLLM, WandB, DDP, Megatron-LM, Accelerate

ELI5

Remote
March 2023 - September 2024

Co-founder/CTO

- Built fine-tuning pipeline with O(10k) annotated interviews; enabled human-quality interview grading (K=0.72)
- Created automated recruiting platform, interviewed 2k candidates per month at 10x cost savings vs human-led interviews; contributed to 20 hires per month

Python, MySQL, React, Javascript, vLLM, RLHF

Jane Street Capital

New York, NY
June 2022 - March 2023

Staff Software Engineer - Positions

- Owned design + breaking ground on revamp of distributed Position Management System (PMS); improved SLA from 3 (trending downwards) -> 4 nines, saved on $1M+ of misinformed/missed trades
- Maintenance + feature dev on existing PMS, maintained positions on O(100k) securities, fed into systems firmwide
- Interviewed ~25 software engineer candidates, member of hiring committee, member of tech recruiting steering committee

OCaml, PostgreSQL

Hani Mir

hmir95@gmail.com
linkedin.com/in/hanimir

Jane Street Capital

New York, NY
June 2020 - June 2022

Senior Software Engineer - Reconciliations

- Zero to one development (DB, backend, frontend) of reconciliations platform; scaled from O(1k) to O(10k) automated reconciliations daily firmwide totaling $10M+ annually
- Owned technical development of position data ingestion pipeline; scaled from O(10) -> O(100) data sources
- Zero to one development (DB, backend, frontend) of invoicing platform; automated unscalable manual process to scale from O(1) -> O(100) clients invoiced monthly for $1M+ annually
- Interviewed ~100 software engineer candidates, member of hiring committee

OCaml, PostgreSQL

Jane Street Capital

New York, NY
June 2019 - June 2020

Senior Software Engineer - External

- Zero to one development (design, PM, DB, backend, frontend) of internal ATS to manage 100% of candidates and interviews at the firm
- Scaled recruiting processes to grow company from 700 -> 2000+ and O(10k) -> O(100k) annual applicants
- Unified data from 6+ sources into one centralized data store; enabled groups across the firm to share hiring processes and tooling
- Created (backend, frontend) internal Human Capital Management (HCM) system in response to COVID-related WFH requirements; used daily by all employees
- Interviewed ~50 software engineer candidates

OCaml, Javascript, Python, PostgreSQL

Shopify

Waterloo, ON
December 2017 - December 2018

Software Engineer - Shopify Plus

- Maintenance + feature dev on ETL pipelines for 30k+ Shopify Plus clients
- Led migration of all ops processes to Salesforce; enhanced 300+ ops employees to scale beyond 5M shops

Spark, Python, Ruby

Education

University of Waterloo

Waterloo, ON
2013

Bachelor of Computer Science (BCS)

Artificial Intelligence, Machine Learning, Data Structures and Algorithms, Functional Programming, User Interfaces