Zhenlin Wang

(412) 726-9719 • zhenlin.wang.criss@gmail.com • linkedin.com/in/zhenlin-wang • github.com/Criss-Wang

EDUCATION

Carnegie Mellon University

M.S. Machine Learning • GPA: 3.8 08/2022 - 12/2023

National University of Singapore

B.S. Applied Mathematics and Computer Science • GPA: 4.75 08/2018 - 06/2022

WORK EXPERIENCE

Aisera

Palo Alto, CA • 02/2024 - Present

Software Engineer - AI/ML

- Streamlined model migration analysis via A/B testing, reducing 40% cost and 50% latency at inference time by shifting from GPT-4 to finetuned Llama3-70B without compromising accuracy and reliability.
- Developed online + offline LLM evaluation framework using Kafka, Postgres, AWS, and Python, enabling robust feedback data collection and capturing 30% more model/prompt drift cases via evaluation-based alerts.
- Built datasets ETL tools to enable efficient large-scale I/O across S3, local and production DB for faster model prototyping.
- Facilitate AI Agent function calling via LLM gateway using a LangChain Adapter in Java, reduced 20% code redundancy.

J.P. Morgan

New York, NY • 06/2023 - 08/2023

Quantitative Research & Development Intern

- Developed an event-driven signal generation system incorporating momentum indicators with RNN models to deliver real-time signals. The strategy outperformed baseline by 12% PnL in backtesting.
- Led the development of real-time coupon/swap MA calculation and validation tool in Athena Data Analytics library to facilitate online pricing model execution.
- Designed an RL (PPO + A3C) based trading strategy via PyTorch & RLlib, ranked 3rd in the internal algo-trading hackathon.

Institute for Mathematical Science - NUS

Singapore • 05/2022 - 08/2022

Machine Learning Research Engineer

- Developed Schizophrenia relapses pattern recognition system with 1D CNN, LSTM and Temporal Transformer models, resulting in 0.23 higher R-square than statistical models, contributing to advanced healthcare analytics for MOHT.
- Enhanced training efficiency by 2.4x using distributed and mixed-precision training, leading to accelerated research timelines and increased productivity in model development.
- Built AWS data pipelines with Kinesis and Glue, reducing data processing time by 35% and enabling real-time analytics.

Emporio Analytics

Singapore • 05/2020 - 08/2020

Data Science and Engineering Intern

- Engineered a price prediction ML pipeline end-to-end using GMM and Facebook Prophet, reducing RMSLE by 15% when deployed on the analytics app used by over 1,000 merchants.
- Invented a two-stage training workflow via PySpark + Ray, resulting in 150% ETL speedup on billion-level transaction data and 30% training speedup through online/offline separation and secondary calibration.
- Improved model endpoint throughput by 3x via asynchronous calls and request-based caching using Redis and RabbitMQ.

PROJECTS

ResumeAssist

11/2023 - Present

- Developed a resume-enhancing AI using agent-based LLM architecture to expedite college students' tech job-hunting process.
- Finetuned Mistral-7B and Llama3-8B and curated graph-enabled knowledge bases using 200+ resumes collected from Discord, reducing 85% latency against GPT-4 function-calling use cases when deployed with LangChain and SageMaker.
- Applied Chain-of-Thought, n-shot prompting with continual in-context learning to boost agents' performance by 25%.

SKILLS

Language: Python, C++, Java, TypeScript, Go

AI/ML: PyTorch, Hugging Face, NumPy, Pandas, Sklearn, TensorFlow, PySpark, Ray, SageMaker, Jupyter

Software Development: FastAPI, Spring Boot, React, Kafka, MySQL, AWS, Redis, Jenkins, Prometheus, Grafana, Docker