WILLIAM TSAO

Arcadia, CA | (626) 219-2222 | williamtsao@gmail.com | http://www.linkedin.com/in/william-tsao-usc

EDUCATION

University of Southern California August 2021-Present
Computer Science

- GPA: 3.41 · Upper-Division CS GPA: 3.56

SUMMARY

Full-stack & data-platform engineer with production LLM UX, scalable search/data systems, and quantitative research. Built a real-time React (TypeScript), FastAPI, and WebSockets chat that streams tokens from multiple LLM providers. Engineered equities pipelines with Python and scikit-learn using regime clustering that reduced backtest max drawdown by 77.7%. Operated multi-region Kubernetes services to 99.9% uptime and accelerated Elasticsearch queries by 50% after migrating from MongoDB.

EXPERIENCE

Software Developer June 2024-May 2025; September 2025–Present
Kolo AI, Pasadena, CA

- Built a persona-driven chat UI in React and TypeScript with streamed Markdown and inline citations; added icons/tooltips and model defaults so users can reliably control tone/behavior.
- Shipped Google Calendar integration end-to-end (OAuth 2.0 PKCE, create/read, webhook-driven sync) with concurrency-safe handlers and idempotency to prevent duplicate events.
- Implemented a usage & cost attribution pipeline across Sendbird/Telnyx with Decimal-safe accounting, timezone-correct aggregation, webhook ingestion, and SQL/Grafana dashboards.
- Prototyped a menu-intelligence data pipeline: multithreaded/recursive crawling, image/PDF extraction, LLM-assisted cleaning, and APIs powering a restaurant-recommendation chat flow.
- Delivered core frontend features for mobile/desktop (new-conversation flow, auth, settings/sidebar, layout polish) and added moderation profiles (SHAFT) to switch safety modes per session.

Quantitative Analyst Intern May 2025-August 2025
Draco Evolution, Taipei, Taiwan

- Built a real-time intraday research stack that streamed tick/minute data from Polygon.io and executed paper trades on Alpaca; computed features and ran in-process inference for 5-second signals.
- Engineered a Python ML pipeline with scikit-learn (StandardScaler, SelectKBest[f_regression], HistGradientBoostingRegressor), plus PCA with K-Medoids regime clustering.
- Designed leak-free, walk-forward backtests tracking P&L, win rate, cumulative P&L, and max drawdown; delivered 77.7% lower max drawdown than buy-and-hold and identified regimes with 54% up-move and 54% down-move clusters.
- Ensured market-aware data handling by aligning to exchange sessions and maintaining incrementally updated historical datasets from Polygon REST, enabling reproducible forward tests.

Software Engineer Intern June 2024-August 2024
TSMC, Taichung, Taiwan

- Operated fab-support services on Kubernetes across Taiwan, Germany, and Japan, instrumented with Prometheus and Grafana, delivering 99.9% uptime for internal users.
- Migrated fab-floor ticketing data from MongoDB to Elasticsearch, tuning index/shard layout and mappings to enable semantic-style queries and 50% faster search latency.
- Extended the backend in Java to support role-based submission, cross-department routing, and export features; deployed updated services via kubectl and Kubernetes manifests.

Research Analyst February 2022-April 2023
University of Cambridge, Centre for Alternative Finance, Cambridge, UK

- Evaluated digital-asset policy, environmental footprint (PoW vs PoS), and social-safeguard considerations for aid programs and translated findings into actionable recommendations for NGOs and donors.
- Co-authored CDAP at CCAF (2023) “Cryptoasset Ecosystem in Latin America and the Caribbean”.
- Co-authored CDAP at CCAF (2023) “Considering Digital Assets for Humanitarian Cash-Based Transfers”.

LANGUAGES

English (Fluent), Chinese (Fluent)