Amy (Chujun) Tao

919-813-8187 | chujunt@andrew.cmu.edu | linkedin.com/in/chujun-tao | https://amytao.github.io/

Education

Carnegie Mellon University GPA: 4.0/4.0 — **Pittsburgh, USA**
Master of Science in Electrical and Computer Engineering — May 2026

University of California, Berkeley: 3.7/4.0 — **Berkeley, USA**
Exchange Student in the EECS Department — Aug 2023

The Chinese University of Hong Kong GPA: 3.6/4.0 — **Shenzhen, China**
Bachelor of Engineering in Computer Engineering — July 2024

Skills

Languages: C/C++, PHP, Java, Go, Python, JavaScript, TypeScript, Bash, SQL, Verilog
Frameworks & Tools: React.js, Node.js, gMock, GraphQL, Redis, Kafka, RabbitMQ, Docker, Kubernetes, AWS, Azure, Nginx, JUnit, Postman, Maven, Git, Jenkins, Prometheus, Grafana

Work Experience

Meta | Software Engineer Intern — 05/2025-08/2025

- Unified revenue-sharing workflow orchestration engines, reducing system complexity and cutting engineering onboarding time by 50%.
- Migrated and refactored 7 tasks of the digital balance spend workflow and created unit tests using gMock.
- Enhanced workflow extensibility by integrating it with the Action Config Factory, enabling reuse across five types of Facebook digital products.
- Implemented asynchronous DAG-based workflow execution with step-level tracing, providing end-to-end visibility through error logging and ensuring idempotent execution.
- Dogfooded the fan-funding subscription purchase flow using the new workflow, validating end-to-end transactions on Google Pay and Apple Pay.

Projects

CourPilot: A Multi-Source Deep Research Assistant for Course Materials (Agentic systems)

- Orchestrated a multi agent system using Claude Sonnet 4.5 for intelligent tool selection and decision-making, with semantic routing to optimize tool chains across multiple MCP servers.
- Built BERTopic-based topic modeling system integrated with Milvus vector database for semantic clustering, automatic topic reduction, and dynamic knowledge base updates with hot-reloading capabilities.
- Implemented a hybrid search architecture with fine-tuned Mistral-7B, web search (Tavily API), and academic paper retrieval, enabling hallucination detection on LLM responses via diversion decoding to measure model resistance to semantic deviation.

Sparse Autodiff Framework: Efficient Sparse Matrix Operations for Graph Neural Networks (AI Infra)

- Extended the Needle Autodiff framework with sparse matrix support, implementing COO and BSR storage formats with format conversion capabilities.
- Designed and implemented a sparse-dense matrix multiplication operator (SpMM) integrated with the autograd computation graph, enabling gradient computation through sparse operations.
- Built an end-to-end GCN recommendation model on MovieLens 100K, constructing bipartite graph adjacency matrices with 97% sparsity and achieved order-of-magnitude improvements in memory efficiency (26 to 3.9MB) and computational efficiency (0.22 to 0.006 GFLOPs) by processing only non-zero entries.

Movie Recommender for Streaming Platform (ML Infra)

- Designed and deployed a modular movie recommender system for 10K users and 27K movies, featuring real-time data ingestion with Kafka, automated model updates via Docker, cold-start handling, popularity-aware re-ranking, model versioning and lineage with MLflow, and observability using Prometheus and Grafana.
- Applied popularity-aware re-ranking to promote mid-/long-tail content for exposure diversity and conducted post-hoc fairness audits to mitigate bias.
- Enabled traceable and reproducible deployments by leveraging MLflow for model registry and pipeline versioning, with logged performance assessment and experiment tracking.
- Enhanced zero-downtime model retraining by triggering on both user behavior drift and scheduled intervals, ensuring stable performance and system reliability.