SIQI XU

Tel: 646-301-3247 www.linkedin.com/in/siqi-xu-8a6254255 dmt1909206@gmail.com
Web : siqixuchristina.com

EDUCATION

New York University, New York, NY| Master of Science in Computer Engineering **Sep. 2024- Present**
Awards: Scholarship for Academic Excellence (Academic Year of 2024-2026)
Xiamen University, Xiamen, China / Selangor Malaysia | Bachelor of Engineering, Digital Media Technology Sep. 2019-May 2023
Awards: Scholarship for Academic Excellence (Academic Year of 2021-2023) Ranking: top 5%; Major GPA: 3.79/4.00

PROFESSIONAL SKILLS

- Programming: C/C++, C#, Python, SQL, Java
- Tools: AWS (S3, EMR), Spark/PySpark, Docker, Kubernetes, FastAPI, Redis, MongoDB, DynamoDB, MLflow, ONNX, Triton, Prometheus/Grafana, GitHub Actions, Terraform, ArgoCD, Databricks, Tableau, PyTorch, TensorFlow

EXPERIENCE

Paradigm Study **Jun. 2025 – Sep.2025**
Intern, Software Developer *San Francisco, CA*

- Built and shipped end-to-end features in a React (Next.js) and TypeScript codebase, partnered with PM and designers, delivered 4 to 6 modules across admin dashboards, report triage, and community workflows from design to production release.
- Improved data access performance by optimizing query paths with Drizzle ORM and adding Redis caching, reducing average response latency by 25% under load and increasing throughput.
- Delivered analytics experiences for account management, exam data visualization, and progress tracking, improving usability and operational efficiency for internal stakeholders.
- Automated incident and handoff workflows via Slack and Linear integrations across 3 to 5 scenarios, implemented safe retries and idempotency safeguards, instrumented key failure metrics in Prometheus and routed alerts to Slack to speed up detection and triage.
- Added CI quality checks in GitHub Actions including lint, typecheck, and build verification to improve release reliability and support weekly release routines. Implemented basic unit and integration tests for critical flows and APIs to reduce regressions and improve release confidence, and containerized the web service with Docker with documented local run and deployment steps.

Simple Creator Network Co., Ltd **Jun. 2024 –Nov.2024**
Intern, Software Developer

- Built and optimized a C# Unity puzzle game (100 levels), applying object pooling and asset loading optimizations to cut load time by 20% and improve runtime stability (e.g., fewer frame drops/crashes during long sessions).
- Profiled and optimized A* pathfinding hotspots by tuning heuristic and tie-breaking rules, balancing path quality and compute cost for large maps and frequent replans. Reduced pathfinding CPU overhead by 30% by adding path cache reuse and invalidation logic, and validated gains via repeatable benchmarks and regression checks to prevent performance drift.

Xiamen HDW Network Co., Ltd **Jun. 2023 –Aug.2023**
Intern, Software Developer

- Developed and executed test plans for internal Python developer tools (build/packaging utilities, asset management scripts), partnering with engineers to triage failures and improve toolchain reliability.
- Built automated smoke/regression checks with pytest + requests to validate build artifacts and critical API endpoints, added timeouts/retries and structured failure logs to speed up debugging and reduce manual verification.
- Integrated CI checks into GitHub Actions including unit tests and static checks, improved code review efficiency and prevented regressions before merge.

PROJECTS/COMPETITIONS

Steam Game Recommendation System

- Built an AWS EMR PySpark pipeline processing 41M+ interactions into curated S3 Parquet datasets, enforced schemas and null and duplicate checks, published versioned snapshots for reproducibility.
- Produced feature tables with stable schemas and refresh conventions, implemented an offline evaluation harness using HR@K and NDCG@K with sliced comparisons for baseline QA. Deployed a containerized FastAPI service with Docker Compose serving precomputed Top-N results from S3, exposed metrics for RPS, p95 latency, and error rate, enabled SQL QA via Athena external tables.

LLaRA++ — Cold-Start Music Recommendation System

- Developed a cold-start hybrid recommender using DistilBERT embeddings and an MLP projector with LLaRA, trained on 1M playlists and 20GB metadata, implemented leakage-safe evaluation using randomized splitting and cross-matching.
- Built a cloud-native MLOps stack on TACC (Terraform, Kubernetes, Ansible, ArgoCD) with persistent object storage while integrating MLflow + MinIO for experiment tracking and artifact versioning.
- Optimized inference with ONNX quantization and Triton serving, benchmarked CPU and CUDA and TensorRT and OpenVINO providers, sustained over 5K requests per second with sub-millisecond latency on A100 and MI100 under controlled benchmarks.
- Implemented CI CD for multi-stage model promotion using GitHub Actions and ArgoCD, supported automated rollback based on MLflow metrics, monitored latency and errors via Prometheus and Grafana dashboards.

Mixture of Semantic Modalities for Recommendation (MSMRec)

- Built a RAG pipeline to ground metadata generation in retrieved evidence, reduced hallucinations and improved consistency of generated features.
- Designed modality-specific encoders and a Mixture-of-Experts architecture to capture user preferences across plot, cast, visuals, and release attributes, enabled interpretable analysis of modality attention over time.
- Evaluated on MovieLens 1M and 100K, improved HR@100 by 1.6% over SASRec, performed modality utilization analysis across user groups.

PUBLICATIONS

- First Author, Siqi Xu, Sentiment Analysis for Individual Business Decision Making based on Hotel Reviews, (IEREK) The 5th International Conference on Future Smart Cities
- Co-first Author, Siqi Xu, Propensity Score Matching on Discrete Treatment: Beijing PM2.5 Case Study, ICMVA 2022: The 5th International Conference on Machine Vision and Applications (February 2022, Pages 93–98)