Zheyuan Lin

(857)-354-7800| zlin252@emory.edu | LinkedIn | GitHub | Personal Website

Education

Emory University May 2026
Bachelor of Science in Computer Science and Mathematics (3.93 / 4.00 GPA) *Atlanta, Georgia*

Experience

Tencent Shenzhen, China
Software Development Engineer Intern *May 2025 – August 2025*

- Built **SpringBoot A/B** testing traffic allocator serving **500+** experiments; consolidated 5 APIs reducing integration **80%**
- Achieved **70x P99.9 latency** improvement (363ms → 5.2ms) by offloading runtime 4-table joins to backend **Redis** wide tables with 5-minute refresh cycles, reducing SDK memory footprint by **60%** through wide tables containing minimal field sets
- Employed **distributed cache** eliminating configuration update blocking, reducing deployment windows **to <5 minutes**
- Established **Prometheus/Grafana observability** with <1-min detection; built adapters for **.NET/Java** legacy clients; discovered **12 edge cases**; deployed via **Kubernetes** with **Github Action**; resolved Redis pool exhaustion

Invest ATL Atlanta, GA
Software Development Engineer Intern *May 2024 – August 2024*

- Built ML platform to automate affordable housing tax credit applications, eliminating **5 hours** of manual work per application
- Built geospatial scoring engine implementing Georgia QAP compliance rules across 19 amenity types, scraping 3,000+ POI via **Google Maps API** and achieving 88% accuracy through polygon containment tests and Haversine calculations
- Built **RAG** pipeline using **OpenAI GPT-4** and **ChromaDB** vector database to process 124-page manual, achieving **95%** answer accuracy with automated citation generation through **hybrid retrieval**, reducing rule interpretation errors by 60%
- Developed full-stack application using **React** for spatial indexing and interactive heatmap visualization of scoring distributions

XiaoMi Technology Beijing, China
Software Development Engineer Intern *May 2023 – September 2023*

- Developed **Kubernetes** debugging platform serving **200+** backend engineers across **30+** microservices managing **400+** pods in production, reducing infrastructure support tickets by **40%**, saving estimated **40+ hours/month** in manual troubleshooting
- Achieved P95 latency **45ms at 500 req/sec** by optimizing **gRPC** connection pooling and **Redis** caching (94.5% hit rate)
- Built **Go gRPC** microservices platform for workload health analysis, pod management, and monitoring with **JWT/RBAC authentication**, namespace-scoped permissions, and **PostgreSQL audit logging** tracking **10K+** ops/month
- Developed **React console plugin** with real-time workload visualization via **WebSockets** with restart pattern detection
- Architected comprehensive **Chronosphere** monitoring integration with automated alerting and dashboard embedding

Projects

Emory NLP Lab With JenAI | *Python, PyTorch, Flask, React, Docker, LangChain* September 2024 – May 2025

- Fine-tuned **Llama-2-7B** using **QLoRA** (4-bit quantization, rank-32) on Jenkins domain knowledge, achieving **45%** accuracy improvement (ROUGE-L: 0.52→0.75) on technical troubleshooting queries while training only **1.7% of parameters**
- Built **ETL pipeline** scraping **3,200+** Q&A pairs; reduced noise by **40%** through deduplication and quality filtering
- Deployed **LangChain** chatbot with conversational memory; quantized to **GGML** enabling CPU inference at **15 tokens/sec**
- Developed **Flask + React TypeScript** full-stack app for chat interface with **Docker Compose**; achieved one-command setup

Visualized Data Interpretation Bias Study Platform | *React, Python, WebSocket, D3.js* May 2024 – August 2025

- Building research platform on how Socratic prompting influences data interpretation bias, publishing to CHI 2026
- Implemented interaction tracking system capturing **40+** event types (hover, click, filter, encoding changes) on **D3.js** interactive chart types (scatter, dot, bar, line, strip plots) with **millisecond-precision** timestamps and debounced transmission (200ms)
- Designed **Python** analysis pipeline computing attention heatmaps and **chi-square** tests to quantify visualization bias patterns

Technical Skills

Languages: Java, C++, Python, Go, C# JavaScript, SQL, Bash
Frameworks & APIs: Spring Boot, Node.js, React, .NET, FastAPI, RESTful, gRPC, GraphQL, Microservices
Cloud & DevOps: AWS (EC2, S3, Lambda), Azure, Docker, Kubernetes, Terraform, Git, Linux, Nginx
Databases & Systems: PostgreSQL, MongoDB, Redis, Elasticsearch, Kafka, RocketMQ, ZooKeeper, Apache Spark, JUnit
AI/ML & Monitoring: PyTorch, HuggingFace, Scikit-learn, LLM, LangChain, MCP, Azure Open AI, Prometheus, ELK Stack
Technical Skills: Data Structures & Algorithms, OOP, System Design, Computer Networks, Distributed Systems, CI/CD, Agile
Certifications & Awards: AWS Certified Developer-Associate (DVA-C02), 2025 ACM-ICPC Southeast Regional 9th Team