

CHENGZHA ZHOU

(443)416-2805 | ericzcz2025@gmail.com | [linkedin](https://www.linkedin.com/in/ericzcz/) | ericzcz.com | [Github](https://github.com/ericzcz)

EDUCATION

Johns Hopkins University

M.S.E. program in Electrical and Computer Engineering

Baltimore, MD
Aug 2025 - Jun 2027 [Expected]

Zhejiang Gongshang University / University of Sussex

Bachelor of Engineering in Communication Engineering (Cooperative Program)

Zhejiang, China
Sep 2021 - May 2025

TECHNICAL SKILLS

Programming Languages: Java, Python, C++, C#, Unity, Matlab, Arduino

Backend, AI & Data: FastAPI, Spring Boot, Celery | LangChain | PostgreSQL, MySQL, Redis, Milvus

Tools: Docker, Git, Playwright, Prometheus, Grafana

PROFESSIONAL EXPERIENCE

Datalynn

Software Engineer Intern

New York, NY
Jun 2025 - Aug 2025

- Developed a scalable and modular Python crawler with on-demand data updates; deployed as a reusable plugin in staging and production environment; served as an upstream data source supporting data query.
- Built a high-concurrency scraping pipeline with Playwright and asyncio; use asyncio and aiohttp to support concurrent task and reduced end-to-end latency by 60%; processed data with Pandas and stored in MySQL Server (180K rows per day); exposed functions via FastAPI (retrieve, update, delete) for seamless integration.
- Scaled a modular crawling plugin across 5 Celery workers, using asyncio for concurrent I/O and robust logging and error handling to increase throughput from 0.5 GB/hour to 2 GB/hour.

Sunny Optical (Zhejiang) Research Institute Co., LTD.

Software Engineer Intern

Zhejiang, China
Jul 2023 - Aug 2023

- Built C#/Unity VR test tools for see-through, image projection, and pose testing on Android devices.
- Doubled VR application performance on older devices by reducing positional-data capture latency to 100 ms.

PROJECTS

MediAgent – RAG-based Medical AI Assistant

Baltimore, MD

- Built a FastAPI-based ReAct medical agent with intent routing and confidence-driven orchestration across 4 specialized skills for medical QA, web search, medical-record insertion, and field-level retrieval.
- Indexed 72,877 chunks from 11,274 XML files and 16,407 QA documents; improved Recall@5 by 9.4 points and Hit@5 from 84% to 100% using query rewriting, Milvus vector search, BM25 hybrid retrieval, and reranking.
- Designed Redis-backed short-term memory with sliding-window retention and rolling summarization, reducing prompt-memory tokens by 61.9% while preserving multi-turn context.
- Built long-term memory with LLM-based extraction, skip/merge/create update decisions, Redis storage, and Milvus semantic recall for personalized cross-session responses.
- Developed PostgreSQL medical-record storage with 2 normalized tables and 3 custom indexes; async endpoints achieved 4.5× higher throughput under 10 concurrent users.

FlashCart – High-Concurrency E-Commerce Platform

Baltimore, MD

- Implemented Cache-Aside and two-level Caffeine/Redis caching, increasing throughput from 6,314 to 8,233 QPS and reducing P99 latency from 43.88 to 30.31 ms.
- Used Redis Lua and Streams for atomic inventory reservation and async order creation, reaching 3,668 QPS with no overselling or duplicate orders.
- Built idempotent Stream consumers with pending recovery, retries, and dead-letter handling, reducing flash-sale P99 latency from 408.81 to 68.59 ms.