Annie Cai

ajcai@uwaterloo.ca | anniecai.com | github.com/aanxniee | in/anniecai2004

Education

University of Waterloo *Waterloo, ON*
B.ASc in Systems Design Engineering - GPA: 4.00/4.00, 95% *2022 - 2026*

- *Awards*: 4x Term Dean's List, Faculty of Engineering Scholarship, President's Distinction
- *Relevant Courses*: Data Structures, Algorithms, Computer Networks, Distributed Systems, AI Algorithms

Skills

Languages: Python, Go, TypeScript, JavaScript, C++, SQL, GraphQL, HTML, CSS
Technologies: Apache Spark, Apache Kafka, React, Flask, Node.js, Next.js, Express.js, Django
Tools: Git, PostgreSQL, MongoDB, AWS, GCP, Kubernetes, Docker, Redis, gRPC, Bash,

Work Experience

Datadog *New York, NY*
Software Engineer Intern *Jan. 2024 - Present*

- Engineered a **Kafka-based streaming solution** that eliminates the network overhead of polling, handling **100k+ messages per second**. Implemented partitioning and dynamic bucketing strategies, resulting in **millisecond-latency** processing across enterprise-scale distributed systems
- Optimized data processing pipelines by leveraging concurrency control mechanisms in **Go**, including mutexes, semaphores, and worker pools, leading to a **2x CPU utilization drop**
- Spearheaded the development of a **distributed multilayer caching system**, using hashing algorithms and LRU eviction strategies to improve data access speeds and cache hit rates

Newfront (YC W18) *San Francisco, CA*
Software Engineer Intern *May. 2024 - Aug. 2024*

- Built a real-time messaging system using **Websockets** for an AI chatbot, enabling bidirectional communication for **8000+ users**. Integrated pub/sub for efficient message routing
- Refactored synchronous I/O-bound tasks within **FastAPI** endpoints to asynchronous operations, allowing concurrent task execution and **reducing latency and server response times by 3x**
- Proposed and led text chunking and hybrid search experimentations with **Python**, **Pinecone** and **LangChain**, resulting in **11% reduction** in query response times and improved retrieval accuracy
- Conducted A/B testing of various combinations of vector embeddings and LLM models. Identified configurations that **reduced computational load by 28%** while maintaining output quality

Royal Bank of Canada *Toronto, ON*
Software Engineer Intern *Sep. 2023 - Dec. 2023*

- Improved ETL pipeline reliability and ingestion speed by refactoring a monolithic **Python** pipeline into modular pipelines, enabling parallel and independent execution and enhancing workflow efficiency
- Achieved a **40% cost reduction** in storage expenses by **migrating 8GB+** of data to **AWS S3 buckets**
- Scaled database retrieval job during high load periods using **Redis** and **Apache Spark**, leveraging resource tuning and caching strategies to **lower job time from hours to under 5 minutes**

Projects

UW Blueprint
Directed a cross-functional student team of 9 through the entire software lifecycle—from ideation to launch—developing an scalable pet management platform for a nonprofit, enhancing operational efficiency and positively impacting over 2,000 community members.