SHUNJIE HU

+1 (602) 545-8691 | kylelovescoding@gmail.com | La Jolla, San Diego, CA, USA | linkedin.com/in/kylehu112/

EDUCATION

University of California - San Diego **September 2025 - June 2027**
Bachelor's, Computer Science

Arizona State University **January 2024 - May 2025**
Bachelor's, Computer Science *GPA: 4.0*

PROFESSIONAL EXPERIENCE

bilibili **Shanghai, China**
Software Developer *August 2020 - July 2022*

- Developed robust backend microservices using Java and Spring Boot, significantly enhancing the functionality of the Recommendation Engine while facilitating seamless integration testing across diverse teams.
- Optimized service performance, reducing API latency by 35%, developed and refined real-time user feature pipelines, achieving optimal latency and efficiently processing over 3 billion daily events to enhance user experience.
- Contributed to Agile Scrum processes, promoting teamwork and effective communication to streamline project timelines, code review, prioritize user-centric design and performance goals.

YongHui Superstores Co., Ltd. **Shanghai, China**
Data Developer *September 2018 - August 2020*

- Developed sophisticated data pipelines using Python, leveraging Flink and Spark frameworks to convert extensive E-commerce event streams into strategic insights for decision-making.
- Created a user-friendly monitoring web application with FastAPI and React.js, enhancing the transparency of data pipelines and streamlining the process of error detection across Presto, Spark, and Hive.
- Performed comprehensive analysis and optimization of SQL queries and Spark jobs, leading to improved performance and efficiency through refined query planning and effective parallel processing.

RESEARCH & OUTSIDE EXPERIENCE

ASU FURI | CACTUS data-intensive systems lab **Tempe, AZ, USA**
Research Assistant *October 2024 - May 2025*

- Worked on ML inference inside Meta Velox, a vectorized query execution engine, focusing on operator-level optimization for neural network inference pipelines.
- Implemented C++ inference operators (UDFs) and decomposed fully-connected layers into fused operator executions, reducing intermediate tensor materialization and lowering memory traffic along inference paths.
- Analyzed inference bottlenecks in data-centric ML workloads and optimized memory layout and execution flow to improve end-to-end latency.

LLM-based Q&A Chatbot for Cloud Computing Education **Tempe, AZ, USA**
Research Assistant *June 2025 - August 2025*

- Engineered a full-stack educational chatbot using FastAPI, Node.js, and React integrating the Llama model to provide real-time, domain-specific Q&A support for cloud computing students.
- Optimized the end-to-end RAG pipeline by analyzing tokenization efficiency, improving retrieval accuracy by experimenting with various document chunking strategies, and implementing reranking models.
- Deployed the application using Docker and cloud infrastructure, ensuring scalable performance and low-latency inference for concurrent student users.

ASU's ACM Chapter **Tempe, AZ, USA**
Committee & Officer *January 2025 - May 2025*

SKILLS

Programming Languages: C/C++, Python, SQL, JavaScript, Typescript, Node.js, React, HTML/CSS, Java, C#
Frameworks & Middleware: FastAPI, Node.js, Spring Boot, Kafka, LangGraph, Docker
Operating System & Tools: Linux, Git/GitHub, Agile, Docker, GDB, Valgrind, Maven, Jenkins
Databases & DevOps: PostgreSQL, ChromaDB, Redis, Hadoop, Spark, AWS