

# 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**  
*Bachelor's, Computer Science*

**September 2025 - June 2027**

**Arizona State University**  
*Bachelor's, Computer Science*

**January 2024 - May 2025**  
*GPA: 4.0*

## PROFESSIONAL EXPERIENCE

**bilibili**  
*Software Developer*

**Shanghai, China**  
*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.**  
*Data Developer*

**Shanghai, China**  
*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 Django 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**  
*Research Assistant*

**Tempe, AZ, USA**  
*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**  
*Research Assistant*

**Tempe, AZ, USA**  
*June 2025 - August 2025*

- Engineered a full-stack educational chatbot using React and Python Django, 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**  
*Committee & Officer*

**Tempe, AZ, USA**  
*January 2025 - May 2025*

## SKILLS

**Programming Languages:** C/C++, Python, JavaScript, TypeScript, SQL, Java, React, C#, HTML/CSS  
**Frameworks & Middleware:** Django, Spring Boot, RabbitMQ, Kafka, MyBatis  
**Operating System & Tools:** Linux, Git/GitHub, Agile, Docker, GDB, Valgrind, Maven, Jenkins  
**Databases & DevOps:** PostgreSQL, MySQL, Redis, Hadoop, Spark, AWS, Azure