# Shengze (William) Wang

Computer Science Doctoral Student, advancing to Ph.D. Candidate in 2025 Fall

# TECHNICAL SKILLS

- Academic: Computer Networks & Distributed Systems and their applications in AI Infrastructure (Systems & Networking for AI), AI for Systems & Databases, LLM Inference & Serving Efficiency, VectorDB
- Languages & Databases: C/C++, Rust, Python, eBPF, Redis, Cassandra, DynamoDB, Milvus, Pinecone, HDFS, SQL
- Platforms & Tools: Linux, Network Operating Systems, ROS, HPCs, AWS, Azure, GCP, Ray, Kubernetes, Apache Spark, Kafka, ns-3, P4 Switch, RDMA, CUDA, OpenMP, FPGAs, vLLM, SGLang, llama.cpp, PyTorch, Transformers, Docker, Git
- Highlights: Scalable Key-Value Storage, LLM Inference & ML System Design, Applied AI for Systems & Databases, MLOps & Production Deployment, Advanced Network Protocols & Security, Workload Characterization & Balancing, Content Delivery & High-Performance Networks, ML Stream Processing & Real-Time Analytics, Hashing Algorithms, Edge-Cloud Systems

#### EDUCATION

# University of California, Santa Cruz (UCSC)

San Francisco Bay Area

Homepage: https://shengze.io

Email: shengze@ucsc.edu

Ph.D. in Computer Science and Engineering; Regents Fellowship; BE Dean's Fellowship

2023 - present

Relevant Courses: Adv Computer Networks, Network Security, Computer Architecture, Adv Distributed Systems, Stream Processing, Programming Languages, Adv Machining Learning, Adv Natural Language Processing, Artificial Intelligence, Analysis of Algorithms

#### University of North Texas (UNT)

Dallas - Fort Worth

B.S. in Computer Science; GPA: 4.0; Outstanding Award (Top 1 of Class 2023); President's List 2020 - 2023

Relevant Courses: Algorithms, Machine Learning, Software Engineering, Systems Programming, Database Systems, Computer

Networks, Computer Security, Operating Systems, Probability Models, Linear Algebra, IT Project Management, Technical Writing

# King's College London (KCL)

London, United Kingdom

Visiting Student in Computer Science; Scored: 95/100; JEISE Scholarship

2019

# Working Experience

• Graduate Student Researcher

Baskin School of Engineering, UCSC June. 2023 - present

• NSF REU Research Mentor

The VEC Lab, NSF eCAT Center Jun. 2022 - Aug. 2022

• Undergraduate Research Assistant

Department of Computer Science, UNT Sept. 2021 - May. 2023

• Full-stack Web Engineer

D.S. Creative Sept. 2021 - Jan. 2022

# Selected Projects

• Scalable AI Infrastructure for High-Performance LLM Serving — C/C++, Python 2023 - present
• Qian Lab, https://users.soe.ucsc.edu/~qian/; Center for Research in Systems and Storage (CRSS)

- Designed methods to provide user anonymity and low-overhead encryption for queries/responses in large-scale overlay networks, blending failover resilience with robust privacy guarantees. Developed load balancing mechanisms with distributed key management to support a decentralized and fault-tolerant LLM system.
- Integrate Vector Databases to accelerate retrieval-augmented generation over billions of embeddings, achieving efficient approximate nearest neighbor lookups for knowledge-intensive LLM applications. Investigate novel indexing and query pipelines to enhance retrieval accuracy and throughput.
- Devised a **draft-then-filter** mechanism to generate candidate tokens, then selectively offload low-confidence tokens to the full-scale LLM, achieving up to **2**× **speedup in real-time inference** without sacrificing output quality.
- Implemented **NLL-based confidence scoring** to dynamically filter high-quality drafts, preventing unnecessary requests to the target LLM and thus lowering GPU consumption. Enhanced MLOps with continuous integration and robust monitoring, **streamlining the model lifecycle from draft-model updates to large-scale deployments.**
- Proposed "CALID": a novel inference framework that integrates speculative decoding principles to boost throughput and reduce computational overhead for large language models.
- Proposed **Span-Level Fine-Tuning with unlikelihood training**: a novel approach that leverages annotated unfaithful spans in LLM-generated summaries to **reduce hallucinations and improve factual accuracy.**
- Proposed "GenTorrent": a decentralized overlay network to enhance Large Language Model (LLM) serving scalability and cost-efficiency by harnessing computing resources from distributed contributors. It addresses fundamental challenges in decentralized LLM serving, including overlay network organization, anonymous communication for privacy, efficient overlay forwarding for load balancing and cache reuse, and decentralized verification of model serving quality. GenTorrent aims to democratize AI innovation, significantly reduce serving latency, and improve user privacy, offering a novel approach to future AI deployment.
- Resource Storage and Discovery in Network & Database Systems C/C++, Rust 2023 present Qian Lab, https://users.soe.ucsc.edu/~qian/; NSF Center for Systems and Storage, https://ssrc.us/
  - o Investigate fundamental problems in emerging networks, emphasizing efficient data placement, fault tolerance, and high-throughput designs. (e.g., datacenter networks, CDNs, and quantum networking)
  - Architect and refine critical components—network protocols, routing algorithms, hashing strategies, and load balancers—for **enterprise-scale deployments** (e.g., HPC clusters, IoT networks, programmable switches).
  - Implement and evaluate prototypes using **event-driven simulators** (e.g., ns-3, p2psim) and cloud-based testbeds (AWS, Lambda Labs, CloudLabs, Supercomputers), leveraging asynchronous I/O and concurrency.

- Proposed "LEAD": A novel Distributed Learned Hash Table that embeds machine-learned models within Distributed
  hash table structures to significantly optimize range query performance for distributed networked systems. LEAD
  outperforms existing range-query solutions by demonstrating superior scalability, reduced latency, and robustness
  against network churns. LEAD opens a completely new field for further research on integrating learned models with
  distributed systems.(https://github.com/ShengzeWang/LEAD)
- Vehicular Edge Computing and Connected Autonomous Vehicles Python, ROS 2021 2023

NSF Center for Electric, Connected and Autonomous Technologies, https://ecat.center/, http://veclab.org/

- Profiled hardware resource usage (GPU, CPU, Memory) for real-time object detectors (YOLO, Faster R-CNN, SSD) deployed in ROS-based CAV perception pipelines.
- Investigated model optimizations, including quantization, architectural pruning, and mixed-precision, achieving a measurable trade-off between inference speed and detection accuracy under edge-device constraints.
- Characterized memory contention and identified performance bottlenecks, enabling targeted optimization strategies (e.g., improved scheduling, memory partitioning) that enhanced detection throughput.
- Developed workload models reflecting diverse edge-device configurations (e.g., NVIDIA Jetson, Intel CPUs, Raspberry Pi), facilitating informed resource allocation and adaptive scheduling across heterogeneous deployments.
- Implemented and validated **Vehicle-to-Edge (V2X) communication frameworks** using AWS Edge Services, resulting in reduced latency and improved real-time responsiveness
- False Discovery Rates (FDR) Control in Metaproteomics Search C++, Python

  2021 2023

  Center for Computational Epidemiology and Response Analysis (CeCERA), https://cerl.unt.edu/
  - Addressed systematic FDR biases in metaproteomics pipelines by incorporating **probabilistic modeling and statistical corrections**, reducing false-positive identifications across large proteomic datasets.
  - Proposed "FineFDR": an open-source, fine-grained FDR assessment framework that seamlessly integrates with Comet and Percolator outputs at multiple taxonomic ranks. (https://github.com/Biocomputing-Research-Group/FDR)
  - Implemented the **Expectation-Maximization General-Mixture Model** for clustering proteomic samples based on abundance profiles, substantially enhancing the detection sensitivity for lower-abundance peptides.
  - Benchmarked six FDR control solutions (including Comet, Percolator, and Tailor) on ten diverse datasets, demonstrating notable gains in **precision and increased peptide/protein identifications** compared to state-of-the-art approaches.
- Fatigue Detection for Medical Staffs: Constructed a face-masked data set and developed a CNN-based facial landmark detection model and integrated an LSTM-based PERCLOS (Percentage of Eye Closure) measurement system to detect and quantify medical staff fatigue levels in real-time. (Registered patent: 2020SR1233854)
- DeepEmo.tech: Developed a real-time facial expression recognition web application leveraging Tiny Face Detector and SSD Mobilenet for efficient, on-the-fly inference with a lightweight front-end for live video capture
- Intelligent Traffic Management System: Engineered a reinforcement learning and computer vision—based solution to dynamically regulate traffic signals in urban environments, improving traffic flow. (Registered patent: 2020SR1235776)

# SELECTED PUBLICATIONS

_	GenTorrent: Scaling Large Language Model Serving with An Overley Network	Sep. 2025
•	Submitted to USENIX NSDI 2026. arXiv preprint: https://arxiv.org/abs/2504.20101	Co-primary Author

A Distributed Learned Hash Table (LEAD)

Aug. 2025

2025 IEEE International Conference on Network Protocols (ICNP); IEEE TON submitted First Author & Oral

Vortex: Efficient Decentralized Vector Overlay for Similarity Search and Delivery

2025 IEEE International Conference on Network Protocols (ICNP); \*BEST POSTER AWARD\*

Aug. 2025

First Author

CALID: Collabrative Accelerate LLM Inference with Draft Model with Filter Decoding

Aug. 2025
Submitted to IEEE INFOCOM 2026. Poster at 2024 BayLearn - Machine Learning Symposium (Apple)

Co-Author

Enhancing Faithfulness in Abstractive Summarization via Span-Level Fine-Tuning
Submitted. Poster at 2024 BayLearn - Machine Learning Symposium (Apple)

May. 2025
Co-Author

Distributed Learned Hash Table

Sept. 2024

2024 IEEE International Conference on Network Protocols (ICNP)

First Author & Poster

Perception Workload Characterization and Prediction on the Vehicular Edges
2023 IEEE International Conference on Edge Computing (EDGE)

Jul. 2023
Co-primary Author

Fine-grained Taxonomy-specific False Discovery Rates Control in Metaproteomics

Nov. 2022

2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)

First Author & Oral

Applications of Computer Vision Techniques in Industrial Fields: A Review

Apr. 2021

Journal of Network Security Technology & Application, 2021 (04), ISSN 1009-6833

First Author

# PROFESSIONAL SERVICES

• Reviewer IEEE INFOCOM, IEEE TDSC

• Teaching Assistant CSE 13S: Computer Systems and C Programming - 24 Winter, 25 Winter

• Mentor NSF Research Experiences for Undergraduates (REU) in Vehicular Edge Computing and Security