KEHAN QI

Ph.D. Student | Eligible for CPT/OPT | kehan.qi@stonybrook.edu

EDUCATION

Stony Brook University — PhD Student
Biomedical Informatics — *Aug 2024 - present*

University of Chinese Academy of Sciences — Graduate Study
Master of Engineering in Computer Technology — *Sept 2018 - July 2021*

Zhejiang University — Undergraduate Student
Bachelor of Engineering in Measurement Control Technology and Instruments — *Sept 2013 - July 2017*

IN-CAMPUS PROJECTS

Conditional Probability Flow-based MRI Reconstruction — Research Project
Project Main Member – Methodology, Experiments and Writing — *Jul 2025 – Sept 2025*

- **Objective**: Conditional probability flow for MRI reconstruction in k-space.
- **Progress**: Paper submitted to ICLR 2026.

VLM-based Patholology Image Processing — Research Project
Project Member – Baselines, Ablative Studies and Writing — *July 2025 – Aug 2025*

- **Objective**: LLaVA with learnable tokens to compress giant number of tokens for VLM tasks.
- **Link**: [Preprint PDF]

LLM-based Paper Review BOT — Personal Project (in progress)
Project Lead – System Design and Implementation — *May 2025 – Present*

- **Objective**: Build an end-to-end LLM based paper review BOT deployed on HF and AWS; lead engineering aspects including LLM one-time inference, Orchestrator, Frontend, backend, monitoring, CI/CD.
- **Progress**: Implemented MVP and V1.
- **Useful Links**: [PRD] [System Design]

WORK EXPERIENCE

Stori — Full-time Employee, Hangzhou, China
Data Engineer — *Apr 20 2023 - July 31 2024*

- **Low-latency ML Inference System for Risk Control**: Designed and deployed a real-time ML inference system for transaction-level risk control. Used AWS DMS + Flink + Kinesis + Lambda + SageMaker for cross-account model invocation with 1-second average latency. Optimized inference pipeline for throughput and latency.
- **Real-time Data Infrastructure**: Built real-time data pipelines supporting ML model invocation, data monitoring, and downstream query API integration using Flink, Kinesis, Lambda, DynamoDB, and Elasticsearch. Served as backend for real-time financial indicators and risk flag triggering.
- **Team Leadership and Standards**: Established internal coding and deployment standards, CI/CD pipeline, and AWS CDK infrastructure templates. Mentored two junior engineers and led weekly sprint planning and code reviews.

Amazon — Full-time Employee, Beijing, China
Software Development Engineer — *Aug 02 2021 - Feb 10 2023*

- **Applied ML System Engineering**: Designed and implemented an automated pipeline for weekly product classification updates using ML models deployed on AWS SageMaker. Integrated Lambda, SNS, S3, and DynamoDB to support scalable, production-level ML inference and ingestion with tens of millions of products.
- **Impact Analysis via Distributed Processing**: Built large-scale PySpark pipelines to evaluate financial impacts of updated classification models. Analyzed billions of records to compute fee deltas pre- and post-deployment across multiple dimensions (product, seller, category). Applied Spark job optimization (e.g., executor tuning, broadcast disabling, RDD reuse) to reduce runtime to within 20 minutes.
- **Future Fee Prediction System**: Developed inference-based fee projection system utilizing classification results. Performed batch processing on 1.5B+ records with AWS Glue and Redshift, and optimized TPS throttling to support SageMaker-based fee computation. Enabled daily updates within a 24-hour SLA.

Tencent — Intern, Shenzhen, China
Research Intern — *June 18 2020 - Sept 04 2020*

- **Registrated medical image quality analysis**: a) Detect landmarks from registrated CT images. b) Train a neural network to predict registrated image quality score, with lanmarks and registrated image as input. c) A Chinese patent produced.

Papers and Publications

- **Kehan Qi**, Saumya Gupta, Qingqiao Hu, Weimin Lyu, and Chao Chen*. "Unrolled Networks Are Conditional Probability Flow ODEs in MRI Reconstruction". (ICLR 2026 under review)
- Weimin Lyu, Qingqiao Hu, **Kehan Qi**, Zhan Shi, Wentao Huang, Saumya Gupta, and Chao Chen*. "Efficient Whole Slide Pathology VQA via Token Compression." arXiv preprint arXiv:2507.14497 (2025).
- **Kehan Qi**, Hao Yang, Cheng Li, Zaiyi Liu, Meiyun Wang, Qiegen Liu, and Shanshan Wang*. "X-Net: Brain Stroke Lesion Segmentation Based on Depthwise Separable Convolution and Long-range Dependencies". MICCAI 2019.
- Hao Yang, Weijian Huang, **Kehan Qi**, Cheng Li, Xinfeng Liu, Meiyun Wang, Hairong Zheng, and Shanshan Wang* "CLCI-Net: Cross-Level Fusion and Context Inference Networks for Lesion Segmentation of Chronic Stroke". MICCAI 2019.

Skills

- **Data Processing Techniques**: Spark, Flink, Hive, MySQL, No-SQL
- **Amazon Web Service (AWS) Skills**: Glue, EMR, Lambda Function, SQS, Managed Service for Apache Flink, API Gateway, VPC, DMS, S3, SageMaker
- **Deep Learning Techniques**: Diffusion Model, In-context Learning, Visual Language Model