

# KEHAN QI

kehan.qi@stonybrook.edu

## EDUCATION

### Stony Brook University

Biomedical Informatics

PhD Student

Aug 25 2025 - present

- **Research Team:** Led by Professor Chao Chen. [homepage]
- **MR Image Reconstruction:** Propose a flow-based MR image reconstruction method and validate on open datasets.
- **OCTA Image Translation:** Propose a flow-based deep learning method for translating OCT volume to OCTA.
- **Pathology Image Classification:** Utilize a Visual Language Model for whole slides pathological image classification. 1 paper as 3rd author submitted to WACV 2025.

### University of Chinese Academy of Sciences

Master of Engineering in Computer Technology

Graduate Study

Sept 01 2018 - June 26 2021

- **Research Team:** Led by Professor Shanshan Wang [homepage]
- **MR Image Reconstruction and Segmentation:** Propose neural networks for MR image reconstruction and segmentation. Produced papers: a pre-print paper as 1st author, 1 MICCAI as 1st author, 1 MICCAI as 3rd author, 1 IEEE Access as 3rd author.
- **MR Image Quality Assessment:** Employ a neural network to assess MR image quality. Produced papers and patents: a pre-print paper as 1st author, a US patent as 3rd author.

### Zhejiang University

Bachelor of Engineering in Measurement Control Technology and Instruments

Undergraduate Student

Sept 01 2013 - July 15 2017

## SELECTED PROJECTS

### LLM-based Paper Review BOT

Project Lead – System Design and Implementation

Personal Project (in progress)

May 2025 – Present

- **Objective:** Build an end-to-end LLM based paper review BOT deployed on HuggingFace Spaces with AWS Lambda and Inference Endpoint integration; use HF Inference Endpoint for inference, AWS Lambda for backend, S3 for document storage, HF Space for frontend and orchestrator (multi-agent use cases) and CloudWatch for monitoring.
- **Current Progress:** Completed problem definition, PRD writing, MVP Implementation, system architecture design, and technical study.
- **Planned Scope:** Lead engineering aspects including LLM one-time inference, Orchestrator, Frontend, backend, monitoring, CI/CD.
- **Reserved Interfaces:** Automated data ingestion (AWS Glue), model retraining (SageMaker), and A/B testing (Step Functions).
- **Tools (Planned):** Hugging Face Inference Endpoint, Hugging Face Space, AWS Lambda, AWS SQS, AWS Cloudwatch, AWS CDK, Github Actions, LangChain, QLoRA
- **Useful Links:** [PRD] [Technical Study] [MVP Design] [MVP Impl] [System Design]

## WORK EXPERIENCE

### Stori

Data Engineer

Full-time Employee, Hangzhou, China

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

Software Development Engineer

Full-time Employee, Beijing, China

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

Research Intern

Intern, Shenzhen, China

June 18 2020 - Sept 04 2020

- **Main Responsibility:** Develop new methods for medical image processing.
- **Registered medical image quality analysis:** a) Detect landmarks from registered CT images. b) Train a neural network to predict registered image quality score, with landmarks and registered image as input. c) A Chinese patent produced.

## PAPERS AND PUBLICATIONS

---

- 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). (WACV 2025 under review)
- Lanting Yang, **Kehan Qi**, Peipei Zhang, Jiaxuan Cheng, Hera Soha, Yun Jun, Haochen Ci, Xianliang Zheng, Bo Wang, Yue Mei, Shihao Chen\*, and Junjie Wang\*. "Diagnosis of Forme Fruste Keratoconus Using Corvis ST Sequences with Digital Image Correlation and Machine Learning." Bioengineering 11.5 (2024): 429.
- Shanshan Wang, Hairong Zheng, **Kehan Qi**, Chuyu Rong, and Xin Liu. "Image data quality evaluation method and apparatus, terminal device, and readable storage medium." U.S. Patent US20240046440A1.
- Dong Wei, **Kehan Qi**, Yuexiang Li, Jiawei Chen, Kai Ma, and Yefeng Zheng. "Image registration quality evaluation model training method, device and computer equipment", Chinese patent CN114519729A.
- **Kehan Qi**, Haoran Li, Chuyu Rong, Yu Gong, Cheng Li, Hairong Zheng, and Shanshan Wang\*. "Blind Image Quality Assessment for MRI with A Deep Three-dimensional content-adaptive Hyper-Network". arXiv preprint arXiv:2107.06888 (2021).
- **Kehan Qi**, Yu Gong, Xinfeng Liu, Xin Liu, Hairong Zheng, and Shanshan Wang\*. "Multi-task MR Imaging with Iterative Teacher Forcing and Re-weighted Deep Learning". arXiv preprint arXiv:2011.13614 (2020).
- **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". Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22. Springer International Publishing, 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". Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22. Springer International Publishing, 2019.
- Xin Liu, Hao Yang, **Kehan Qi**, Pei Dong, Qiegen Liu, Xin Liu, Rongpin Wang\*, and Shanshan Wang\*. "MSDF-Net: Multi-scale deep fusion network for stroke lesion segmentation". IEEE Access 7 (2019): 178486-178495.

## 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