Banghao Chi

410 N Lincoln Ave Apt 2332, Yugo Urbana Illinois, Urbana, IL, 61801
M.S CS at University of Illinois Urbana-Champaign

Email : banghao2@illinois.edu
Mobile : +1 2173286124
Profolio: biboyqg.github.io

Education

- **University of Illinois Urbana-Champaign (UIUC)** — Urbana, IL. U.S.
  Master of Science in Computer Science (GPA: -/4.0) — *Fall 2025 – Present*
  - **Core Modules:** Natural Language Processing, Database Systems, ML for Bioinformatics
- **University of Illinois Urbana-Champaign (UIUC)** — Urbana, IL. U.S.
  Bachelor of Science in Mathematics (GPA: 3.86/4.0) — *Fall 2023 – Fall 2025*
  - **Core Modules:** Artificial Intelligence, Internet of Things, Computer Architecture, System Programming
- **Xi'an Jiaotong Liverpool University (XJTLU)** — Suzhou, Jiangsu, China
  Major in Computer Science (GPA: 3.92/4.0) — *Fall 2021 – Spring 2023*
  - **Core Modules:** Database Systems, Algorithms, Statistics and Probability, Calculus, Linear Algebra

Publications

[1] Hanling Wang, **Banghao Chi***, Yufei Wu, et. al. LLMarking: An Adaptive Automatic Short Answer Grading Using Large Language Models. *Association for Computing Machinery Learning@Scale* (*ACM L@S*), 2025. [Paper], [Code], [Poster]

[2] **Banghao Chi***. Research Advanced in the Object Detection Based on Deep Learning. *International Conference on Applied Physics and Computing* (*ICAPC*), 2022. [Paper]

Research Experience

- **Dynamic and Static Precision Quantization for High-Efficiency 3D Object Detection**
  University of Illinois Urbana-Champaign, advised by Minjia Zhang, [Code] — *Mar.2024 - Dec.2024*
  - **Dynamic and static post-training quantization:** Proposed dynamic and static post-training quantization (PTQ) techniques to optimize the 3D object detection algorithm (i.e., CenterPoint), reducing inference time and computational complexity by 35% while only sacrificing 1% of accuracy;
  - **Progressive quantization:** Proposed maintaining 16-bit activations while progressively quantizing other operators and customized a quantization strategy for Sparse 3D convolutions, achieving a balance between precision and efficiency;
  - **Sensitivity analysis:** Conducted quantization sensitivity analysis to pinpoint efficiency-critical variables, enhancing interpretability and allowing precise model tuning to minimize accuracy impact;
  - **SmoothQuant for extreme outlier resolution:** Applied SmoothQuant to solve extreme outliers issue, and therefore recover the accuracy loss caused by direction PTQ.
- **LLMs-based Knowledge Agent**
  University of Illinois Urbana-Champaign, advised by Kevin Chang, [Code] — *Aug.2024 - Dec.2024*
  - **Hierarchical Assessment Framework:** Developed a novel hierarchical framework that mirrors human information retrieval process, incorporating sophisticated prompts to guide LLMs through sequential steps of understanding, concept extraction, and feedback generation, ensuring systematic and comprehensive assessment;
  - **Structured outputs from LLMs:** Innovatively integrated Finite State Machines(FSM) within the generation process of LLMs to achieve structured outputs from LLMs, enabling improved database operation performance;
  - **Model Finetuning and Evaluation:** Deployed and tested on up to 41 different kinds of LLMs with the number of parameters ranging from 2B to 110B. Tested the stability with both proprietary and open-weight models and the accuracy with custom metrics and finetune the models to stablize the outputs.
- **LLMarking: An Auto Marking System using Large Language Model**
  Xi'an Jiaotong Liverpool University, advised by Xiaohui Zhu, [Project page], [Code] — *Mar.2024 - Sep.2024*
  - **Pipeline Construction:** Focused on streamlining Automatic Short Answer Grading (ASAG) pipeline with Large Language Models (LLMs), involving custom dataset and metrics construction, prompt engineering and supervised-finetuning of LLMs;
  - **Model Implementation and Deployment:** Integrated with PagedAttention to achieve high throughput and is capable of giving feedback on 150 student's answers within 3 minutes concurrently;
  - **Model Verification:** Deployed and tested on up to 41 different kinds of LLMs with the number of parameters ranging from 2B to 110B (The best model is able to achieve F1 score at a high of 90.5% and 86.1% on computer science and finance datasets respectively);
  - **Dynamic System:** Designed and deployed a dynamic system which iteratively update the shots within the prompt with better representitives to achieve better accuracy.
- **IoT-Enabled Intelligent Self-Driving Car Prototype Design**
  University of Illinois Urbana-Champaign [Project page], [Code] — *Nov.2023 - Mar.2024*
  - **Integrated Raspberry Pi and ultrasonic sensors:** Itegrated Raspberry Pi for intelligent control IoT-based autonomous car, and using ultrasonic sensors for advanced mapping and spatial awareness;

- **Real-time image processing and object detection:** Implemented advanced object detection using OpenCV and TensorFlow for real-time image processing, and utilized a modified A* pathfinding algorithm for precise route planning;
- **Full self-driving functionality:** Achieved full self-driving functionality through seamless integration of hardware and software components, successfully demonstrating the vehicle's autonomous capabilities in controlled environments, showcasing efficiency and innovation.

- **IoT-Based Home Security Camera**
  University of Illinois Urbana-Champaign [Project page], [Code] *Aug.2023 - Oct.2023*
  - **Histogram of Oriented Gradients (HOG) for facial recognition:** Integrated Histogram of Oriented Gradients (HOG) for facial recognition and manually incorporated plate number detection, enhancing the system's scalability and recognition accuracy;
  - **Multi-layered security defense:** Developed a web application based on the SpringBoot framework, implementing a multi-layered security defense strategy with SpringSecurity and JWT verification to improve system stability and security;
  - **SSL encryption and CDN integration:** Adopted SSL encryption for both front and back ends to ensure secure data transmission, and integrated a CDN to improve response times and enhance resilience against Distributed Denial of Service (DDOS) attacks.

Working Experience

- **Research Intern**
  Research Intern at National Center for Supercomputing Applications (NCSA) *Fall 2024 – Spring 2025*
  - Presented a structured, modular information retrieval system that combines Finite State Machines (FSMs) with Large Language Models (LLMs) to automatically extract and enhance entity-specific information from the web, using recursive link analysis, dynamic schema generation, and JSON-based structured outputs.

- **Research Intern**
  Research Intern at Supercomputing System AI Lab (SSAIL) *Fall 2024 – Spring 2025*
  - Introduced Q-LiDAR, a training-free quantization framework for 3D LiDAR object detection models that improves inference efficiency without compromising accuracy by combining component-specific techniques like SmoothQConv, channel-wise quantization, and Hessian-guided bit-width allocation.

- **Course Assistant**
  CA for CS 409, University of Illinois Urbana-Champaign *Fall 2024 – Spring 2025*
  - Implemented a fully autonomous grader that can run student's submitted code in a sandbox environment and grade on the code based on the results of the program;
  - Graded MPs of students and attended Q&A and Office Hours to address issues from students.

- **Teaching Assistant**
  TA for Calculus course, Xi'an Jiaotong Liverpool University *Fall 2021 – Summer 2022*
  - Held lectures about Calculus, explicitly illustrating essential knowledge step by step and creating relevant quizzes to better assist students in getting the hang of basic content and structures of Calculus.

Awards & Honors

2025	*University Dean's List(top 20% excellence)*	*U of I at Urbana-Champaign*
2024	*University Dean's List(top 20% excellence)*	*U of I at Urbana-Champaign*
2023	*University Academic Excellence Award(top 1% excellence)*	*Xi'an Jiaotong Liverpool University*
2022	*Summer Undergraduate Research Best Poster Award*	*Xi'an Jiaotong Liverpool University*
2022	*University Academic Achievement Award(top 2% excellence)*	*Xi'an Jiaotong Liverpool University*
2022	*Awarded 2nd Prize in Asia and Pacific Mathematical Contest in Modeling*	*Consortium for MAP*
2021	*Awarded 2nd Prize of FLTRP Cup National English Speaking Contest*	*Foreign Language Research Press*

Skills

- **Programming**: Golang, Python(Pytorch, vLLM, NumPy, Pandas, FastAPI), C++, Java, Javascript, R, SQL, NoSQL
- **Tools**: ONNX, ROS2, Kubernetes, Docker, Nginx, Redis, RabbitMQ, React, Spring, Unreal Engine, bash, R Studio, LaTeX