XUKUN LIU

+1 2244176836 | e: xukunliu2025@u.northwestern.edu | Github | Personal Website

EDUCATION

Northwestern University — Evanston, United States
Master of Computer Science
GPA: 3.90 — Sept 2023 – June 2025

Southern University of Science and Technology(SUSTech) — Shenzhen, China
Bachelor of Engineering in Computer Science and Technology — Sept 2019 – June 2023
GPA: 3.71/89.37

SELECTED AWARDS

Outstanding Graduate of Southern University of Science and Technology (SUSTech). (May 2024)
Outstanding Graduate of the Computer Science Department at Southern University of Science and Technology (SUSTech). (May 2024)
Bronze Medal in 2020 China Collegiate Programming Contest, Mianyang Site. (Oct 2020)
Bronze medal in the 2020 ICPC Asia Nanjing Regional Contest. (Dec 2020)

PUBLICATIONS

1. **XukunLiu**, BowenLie, RuqiZhang, Dongkuan Xu. *Adaptive Draft-Verification for Efficient Large Language Model Decoding*, Accepted at the 39th Annual AAAI Conference on Artificial Intelligence (AAAI 2025)
2. Benyamin Tabarsi, Aditya Basarkar, **Xukun Liu**, Dongkuan Xu, Tiffany. *BarnesMerryQuery: A Trustworthy LLM-Powered Tool Providing Personalized Support for Educators and Students*. Accepted at the 39th Annual AAAI Conference on Artificial Intelligence (AAAI 2025)
3. Dong Shu, Haoran Zhao, **Xukun Liu**, David Demeter, et al. *LawLLM: Law Large Language Model for the US Legal System.* Accepted at the 33rd ACM International Conference on Information and Knowledge Management (CIKM 2024)
4. BinfengXu, **XukunLiu**, et al. *Gentopia. AI: A Collaborative Platform for Tool-Augmented LLMs*, Accepted at The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)
5. **XukunLiu**, ZhiyuanPeng, DK Xu. *ToolNet: Connecting Large Language Models With Massive Tools*, Submitted to 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024)
6. **X. Liu**, *The Utilities of Evolutionary Multi-objective Optimization for Neural Architecture Search –An Empirical Perspective*, Accepted at the 17th International Conference on Bio-inspired Computing: Theories and Applications (BIC-TA 2022)
7. **XukunLiu**, Haoze Lv, Chi Wang, et al. *DyESP: Accelerating Hyperparameter-Architecture Search via Dynamic Exploration and Space Pruning*, Submitted to The 18th European Conference on Computer Vision (ECCV 2024)

RESEARCH EXPERIENCES

Scaling Law for Genome Foundation Models — Evanston, IL, USA
Key Contributor — Oct 2024 - Present

- Objective: Investigate and define scaling laws for genome foundation models to optimize computational efficiency and performance for genomic data.
- Designed experiments using a 3.9 TB genome dataset and implemented optimized training setups for models ranging from 50M to 30B parameters. Developed a surrogate model to predict scaling laws and integrated carbon footprint tracking for sustainable AI practices.
- Enabled efficient resource allocation and model size adjustments under computational constraints, improving FLOP utilization and minimizing environmental impact during high-performance GPU training.

Magics.AI: An Open-Source LLM Platform for Academia — Evanston, IL, USA
Group Leader — Oct 2024 - Present

- Objective: Build an open-source platform to make large language models (LLMs) accessible and affordable for academia, enabling scalable AI model development and integration.
- Created a distributed system leveraging idle computational power in universities to facilitate large-scale AI model training and inference. Developed a Python SDK, user-friendly front-end, and CLI tools to lower technical barriers for academic researchers. This system accelerates LLM training and inference, enabling deployment on low-resource machines, thereby expanding accessibility to advanced AI technologies.
- Significantly reduced costs and infrastructure requirements for academic users, fostering a collaborative environment for scalable AI development and experimentation.

MerryQuery: A Trustworthy LLM-Powered Tool Providing Personalized Support for Educators and Students
Raleigh, NC, USA

Key Member Sept 2024 - Present

- Objective: Develop a trustworthy AI educational assistant that supports personalized, course-specific learning while aligning with educators' pedagogical goals.
- Built MerryQuery, an LLM-powered platform using Retrieval-Augmented Generation (RAG) to deliver contextually relevant responses with source citations. Integrated multimodal data processing for complex PDF documents using advanced OCR pipelines and vector embeddings.
- Achieved significant adoption and positive feedback in NCSU, positioning MerryQuery as a robust, reliable alternative to general-purpose tools in educational contexts.

Adaptive Draft-Verification for Efficient Large Language Model Decoding West Lafayette, IN, USA

Group Leader Feb 2024 - Present

- Objective: Enhance the efficiency and speed of Large Language Model (LLM) decoding for real-time applications, reducing latency and computational demands.
- Developed ADED (Adaptive Draft-Verification for Efficient LLM Decoding), a novel methodology that accelerates LLM decoding without fine-tuning. Designed an adaptive draft-verification process using a tri-gram matrix-based LLM representation for dynamic output approximation. Created a Monte Carlo Tree Search (MCTS)-inspired draft maker to balance exploration and exploitation in generating high-quality drafts.
- Achieved up to a 2.5X speedup in latency and a 20% improvement in acceptance rates over existing methods, demonstrating significant efficiency gains in decoding while maintaining high accuracy.

LawLLM: A Multi-Task Large Language Model for the US Legal System Evanston, IL, USA

Key Member Jan 2024 - July 2024

- Objective: Develop a specialized AI-powered tool to enhance legal research and decision-making by addressing challenges in case retrieval, precedent recommendation, and legal judgment prediction.
- Designed and implemented LawLLM, a multi-task LLM fine-tuned on real-life US legal datasets. Developed customized data preprocessing pipelines, including Knowledge Graph (KG) construction and vectorized legal case embeddings, to handle complex legal relationships. Integrated advanced Information Retrieval (IR) techniques and In-Context Learning (ICL) to enhance zero-shot and few-shot performance.
- Delivered superior accuracy and reliability in Similar Case Retrieval (SCR), Precedent Case Recommendation (PCR), and Legal Judgment Prediction (LJP), establishing LawLLM as a powerful legal AI tool.

ToolNet: Connecting Large Language Models with Massive Tools Raleigh, NC, USA

Group Leader Oct 2023 - Present

- Objective: Expand the capabilities of LLMs to perform higher-level tasks by effectively utilizing external tools (APIs) in a scalable manner.
- Developed ToolNet, a plug-and-play framework capable of scaling up to thousands of tools without performance degradation or increased token costs. Designed a network structure where nodes represent tools and weighted edges indicate transition probabilities, enabling the LLM to navigate the network iteratively.
- Demonstrated strong robustness against tool failures and impressive performance in complex tasks, showcasing ToolNet as an innovative approach to augmenting LLM functionality.

Gentopia.AI: A Collaborative Platform for Tool-Augmented LLMs Raleigh, NC, USA

Group Leader June 2023 – Oct 2023

- Objective: Create a collaborative platform for tool-augmented Large Language Models (LLMs).
- Contributed to the development of Gentopia, enabling flexible customization of agents through simple configurations. Integrated various language models, task formats, prompting modules, and plugins into a unified paradigm. Assisted in establishing Gentpool, a public platform for registering and sharing user-customized agents, and contributed to Gentbench, which evaluates agents across safety, robustness, and efficiency.
- Promoted the democratization of AI by simplifying the process of creating, sharing, and evaluating custom AI agents.

Efficient Heterogeneous Bert Redmond, DC, USA

Independent Project, jointly supervised by North Carolina University and Microsoft Research Sept 2022 – Present

- Objective: Establish a more efficient BERT model using Neural Architecture Search (NAS).
- Perfected the training method for supersets and proposed the "Balanced Pareto Sampling" method to improve subnet performance within the same training time. Applied a heterogeneous search space to optimize model architecture.
- Demonstrated improved efficiency and performance over existing methods, advancing research in adaptive and scalable BERT models.

DyESP: Accelerating Hyperparameter-Architecture Search via Dynamic Exploration and Space Pruning

Redmond, DC, USA

Independent Project, jointly supervised by North Carolina University and Microsoft Research — July 2022 – Present.

- Objective: Propose new neural architecture search algorithms for cost-effective architecture exploration.
- Designed meta-learning-based search and pruning algorithms to enable zero-shot exploration of search spaces, significantly reducing search costs. Achieved up to a 1/18 reduction in computational overhead compared to baseline methods.
- Enabled faster and more efficient NAS exploration, contributing to advancements in hyperparameter-architecture search.

EvoXbench, an All-In-One Neural Architecture Search Framework — Shenzhen, China

Research Assistant to Professor Zhichao Lu — May 2022– July 2022

- Objective: Develop an open-source library to facilitate NAS algorithm development and evaluation.
- Processed data integration and constructed databases using Django's ORM framework. Collected and curated existing NASBench datasets and trained surrogate models to oversee experimental processes.
- Created EvoXBench, an open-source library that simplifies NAS algorithm development with Python and MATLAB interfaces.

AutoML Tools Development for Deep Learning on Edge Systems — Shenzhen, China

Group Leader, Advisor: Professor Ran Cheng — Sept 2021 – Jan 2022

- Objective: Design an AutoML algorithm for small, low-power edge devices.
- Deployed and tested neural networks on devices with different architectures. Analyzed results, supervised algorithm development, and implemented architecture designs. Utilized PyTorch to instantiate neural networks and Celery for task distribution and evaluation.
- Delivered a practical AutoML solution for edge devices, enhancing the scalability of deep learning applications in resource-constrained environments.

Selected Project Experiences

Multifunctional and Extensible Online Judge (OJ) System — Shenzhen, China

- Developed a scalable online judge system to evaluate code correctness across multiple programming languages.
- Led the website design, backend construction, deployment, and development of an evaluation engine using Python's Django framework and Google's nsjail.
- Implemented system deployment with Kubernetes for automatic scaling and self-repair capabilities.
- The system passed third-party penetration testing and is now officially used by the Computer Science Department at the university, serving over 2,500 students in 13 courses.

Teaching Assistant Experiences

- Teaching Assistant for *Introduction to Python Programming*. — Fall 2022
- Teaching Assistant for *Principles of Database Systems*. — Spring 2022
- Teaching Assistant for *Computer Organization*. — Spring 2022
- Teaching Assistant for *Introduction to Computer Programming B*. — Spring 2022
- Teaching Assistant for *Data structure and Algorithm Analysis*. — Fall 2021, Fall 2022
- Teaching Assistant for *Introduction to Computer Programming A.* — Spring 2021, Spring 2022

Working Experiences

Research Intern of *Eigent.ai* — Nov 2024 – Present

Co-founder and CTO of *CourtAI* — June 2024 – Dec 2024

Additional Information

Interests

- NLP, Large Language Models, Machine Learning Systems, Efficient AI, Multi-modal Learning, HCI

Technical Skills

- Programming Languages: Python, Rust, C/C++, Java, Nodejs, HTML, JavaScript, SQL, MIPS, Verilog

Open Sources Contributions

- **Camel**
- **Gentopia**
- **Dataease**
- **Evoxbench**