# XUKUN LIU

+1 8722392517 | e: xukunliu2025@northwestern.edu

## EDUCATION

**Northwestern University** Evanston, United States
Master of Computer Science Sept 2023 – June 2025
*Related Course: Deep Learning for NLP, Machine Learning, Deep Learning, Conversation AI, Artificial Intelligence Programming*

**Southern University of Science and Technology** Shenzhen, China
Bachelor of Engineering in Computer Science and Technology Sept 2019 – June 2023
*Related Course: Machine Learning and AI, Data Structures and Algorithms, Software Design Methods*

## *WORK EXPERIENCE*

**Huawei Technology** Shenzhen, China
Software Development Engineer June 2022 – July 2022

- Designed a neural network to reconstruct global beam information from local beam measurements.
- Led model design, data processing, and enhancements to model accuracy.
- Applied various classical Graph Neural Network (GNN) methods to formulate the problem for advanced development.
- Developed a state-of-the-art (SOTA) neural network by integrating a co-occurrence matrix with Graph Attention Networks (GAT).

## SELECTED AWARDS

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. BinfengXu, **XukunLiu**, et al. Gentopia. AI: A Collaborative Platform for Tool-Augmented LLMs, *The 2023 Conference on Empirical Methods in Natural Language Processing*
2. *Learning from myself matters: Accelerated LLM Decoding via Monte Carlo Tree Search and Self-evolved Speculation, Under Preparation for Submission*
3. **XukunLiu**,, ZhiyuanPeng, DK Xu. ToolNet: Connecting Large Language Models With Massive Tools
4. **X. Liu**, The Utilities of Evolutionary Multi-objective Optimization for Neural Architecture Search –An Empirical Perspective, *The 17th International Conference on Bio-inspired Computing: Theories and Applications*
5. **XukunLiu**,, Haoze Lv, Chi Wang, et al. DyESP: Accelerating Hyperparameter-Architecture Search via Dynamic Exploration and Space Pruning. *Submitted to ECCV 2024*

## 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*, Spring 2021

## RESEARCH EXPERIENCES

**Accelerated LLM Decoding via Monte Carlo Tree Search and Self-evolved Speculation** Raleigh, NC
Group Leader Feb 2024 - Present

- Pioneered a novel decoding technique that integrates Monte Carlo Tree Search (MCTS) with retrieval-based speculative decoding, enhancing the speed and efficiency of large language model (LLM) text generation.
- Developed a hybrid strategy combining a simplified 3-gram grammar with MCTS to anticipate potential continuations, significantly reducing computational demands during the decoding process.
- Implemented speculative decoding that employs precomputed text segments and probabilistic modeling, minimizing reliance on continuous forward passes and improving memory efficiency.
- The approach has shown viability for practical applications requiring rapid and reliable text generation, offering substantial improvements over traditional autoregressive decoding techniques.

**ToolNet: Connecting Large Language Models With Massive Tools** Raleigh, NC

Group Leader Oct 2023 - Present

- Objective: Enhance the capabilities of Large Language Models (LLMs) to execute higher-level tasks, including following human instructions for proper use of external tools (APIs).
- Developed ToolNet, a plug-and-play framework capable of integrating thousands of tools without performance degradation and maintaining constant token costs.
- Designed a network structure in which each node represents a tool, and weighted edges represent transition probabilities, allowing an LLM to navigate the network by sequentially selecting the next tool from its neighbors until the task is completed.
- Conducted experiments demonstrating ToolNet's ability to handle complex tasks with high efficiency and robustness against tool failures.

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

Key Member June 2023 – Oct 2023

- Objective: Develop a collaborative platform to enhance Large Language Models (LLMs) with tool augmentation capabilities.
- Played a key role in the development of Gentopia, which enables flexible customization of agents via simple configurations, integrating various language models, task formats, prompting modules, and plugins into a unified framework.
- Contributed to the launch of Gentpool, a public platform that facilitates the registration and sharing of user-customized agents, promoting the democratization of artificial intelligence.
- Assisted in designing Gentbench, a component of Gentpool, to evaluate user-customized agents on metrics such as safety, robustness, and efficiency.

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

Raleigh, NC

Group Leader July 2022 – Present

- Developed DyESP, a novel framework that integrates dynamic exploration with space pruning to enhance the efficiency and accuracy of hyperparameter-architecture search (HAS).
- Engineered a meta-scheduler to adapt search strategies across varying spaces, utilizing historical data to dynamically refine exploration and focus on high-potential areas.
- Demonstrated through extensive benchmarks that DyESP outperforms existing methods in speed and stability, optimizing search processes with notable reductions in computational demand.

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

Group Leader May 2022– July 2022

- Developed EvoXBench, an open-source library that consolidates essential technologies for NAS algorithm development, enabling straightforward testing or development through Python or Matlab interfaces.
- Managed data processing, integration, and database construction by compiling NASBench datasets, extracting and curating data using Django’s ORM framework.
- Trained surrogate models and supervised the experimental process.
- https://github.com/EMI-Group/evoxbench

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

Group Leader Sept 2021 – Jan 2022

- Designed an AutoML algorithm optimized for deployment across various devices, focusing on enhancing performance on small and low-power edge devices.
- Deployed and evaluated various neural networks on devices with differing architectures, conducting comprehensive performance analyses and overseeing the design of algorithms and architectures.
- Utilized PyTorch for neural network instantiation and employed Celery for task dispatching and distributed evaluation.

## SELECTED PROJECT EXPERIENCES

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

- Developed a scalable online judge system capable of evaluating code correctness across multiple programming languages.
- Spearheaded the website design, backend architecture, deployment, and development of an evaluation engine utilizing Python's Django framework and Google's nsjail.
- Implemented system deployment using Kubernetes to enable automatic scaling and self-repair features.
- Successfully underwent third-party penetration testing; system now officially adopted by the Computer Science Department at the university, serving over 3,000 students across 17 courses.

**User Profile Webpage Design for SUSTech Library** Shenzhen, China

- Designed a unique memorial page for students on the Southern University of Science and Technology Library's WeChat public account.
- Developed and launched the backend service, enhancing the library's digital interface and user engagement.
- Achieved the highest score among competing teams in similar project categories.
- https://github.com/liuxukun2000/CS330-library

## Additional Information

**Interests**

- NLP, Large Language Model, Multi-modal, Efficient AI

**Technical Skills**

- Programming Languages: Python, Rust, C/C++, JAVA, HTML, JavaScript, SQL, React
- I am a full-stack developer.

**GitHub**

- Homepage: liuxukun2000 (Xukun Liu) (github.com)