Shengzhe Xu

shengzhe.xu21@gmail.com +1 (540)9223367 https://people.cs.vt.edu/shengzx Google Scholar

Education

Ph.D. and M.S. in Computer Science, Virginia Tech, Alexandria, USA. *08/2016-08/2025*
Advisor: Prof. Naren Ramakrishnan. ‹ **We work on Synthetic Data before the GenAI Big Bang**.
Dissertation: synthetic table generation, deep generative models, large language models. *01/2019-05/2025*
Early research focus: automated code migration, system and programming language. *08/2016-12/2018*

B.E. in Telecom.Eng.&Mgmt., Beijing University of Posts and Telecommunications, Beijing, China. *09/2012-06/2016*
Beijing Outstanding Graduate Award, 2016.
Silver Medal, **ACM-ICPC** programming contest Asia Regionals, 2014.

Experience

Washington Post, VT-WaPo Research Collaboration, LLM in Journalism, Washington, DC, USA *06/2025-08/2025*
- Researching on multi-modal retrieval-augmented generation (RAG) models for journalism.

Facebook, Machine Learning Engineer Intern, Ads Core ML team, Menlo Park, USA *05/2020-08/2020*
- Developed an attention model for time-series ads video-clip semantic summarization.

Microsoft Research Aisa, Research Intern, Machine Learning team, Beijing, China *05/2018-08/2018*
Supervisor: Dr. Jiang Bian and Dr. Jia Zhang
- Conducted research on deep reinforcement learning (RL) for shipping dispatch problems.
- Post-internship collaboration on multi-agent RL for wind farm energy optimization.
- Awarded **MSRA Award of Excellence**.

Google, Undergraduate Software Engineer Intern, Input Method Engine team, Beijing, China *07/2015-10/2015*
- Designed and implemented a domain-specific language for Android resource linting.

Tsinghua National Laboratory, Undergraduate Research Intern, Beijing, China *01/2015-06/2015*
Supervisor: Prof. Minlie Huang
- Developed NLP-based keyword extraction model for medical science papers in EN and CN.

Highlighted Research Topics

Synthetic Tabular Data Generation (STG): structural data, complex dependence, joint distribution.
- [NeurIPS 25, under review] Embedding Isotropy as a Trust Indicator for STG with LLMs.
- [KDD 25, under review] Prompt is Mightier than the Example: Knowledge-Guided Prompting for STG.
- [KDD 25, under review] Permutation-Aided Fine-tuning for LLM-based STG.
- [CIKM 25, under review] Tabular Data Valuation for Optimizing Data Provenance Verification.
- [KDD-MLHat 21] STAN: Synthetic Network Traffic Generation with Generative Neural Models.

Large Language Models (LLMs): prompting, agents, retrieval-augmented generation (RAG), fine-tuning.
- [**COLM** 24] Information-Guided Regularization for Fine-tuning Language Models.
- [BigData 24 (**Best Paper Award**)] Data Augmentations to Support Speculative Reasoning in LLMs.
- [**IEEE Network** 24, **Impact Factor**: 6.8] Large Multimodal Foundation Models for 6G Wireless.

Systems and Programming Languages: LLVM, WALA, Intermediate Representation (IR).
- [**ICSE** 25] LLM Code-Patching Agent with RAG for Adaptive Spreadsheet Data Cleaning.
- [ICPC 19] API Migration Recognition and Edits Inference.

Applied Machine Learning Projects

Virginia Tech Wireless Lab: Large Multi-Modal (LMM) foundation model for 6G wireless. *12/2023-05/2025*
City of Roanoke: Multi-modal AI solution for heat resilience infrastructure planning. *10/2022-10/2024*
Bank of New York Mellon: Optimization of securities lending. *07/2023-11/2023*
VT Agriculture Lab: Precision Produce Analysis, predicted treatment outcomes from RS images. *06/2023-12/2023*
Washington Nationals: Real-time forecasting of concession demand. *06/2023-09/2023*
Office of the Director of National Intelligence: Real-time news-based migration forecasting. *12/2022-03/2023*
Commonwealth Cyber Initiative: Synthetic table generation empowering CCI AI testbed. *09/2020-05/2021*

Technical Skills

Programming Languages: Python (proficient), C/C++ (proficient), Java, others.

Machine Learning Libraries: PyTorch, CUDA, Hugging Face Transformers, LangChain, OpenAI api, TensorFlow, Scikit-learn, Deep Graph Library (DGL).

Tools and Platforms: Linux, Docker, Git, LaTeX, SQL, Pandas, NumPy, Matplotlib, Seaborn, Jupyter Notebook.

Core Machine Learning Topics: Optimization in Machine Learning, Graph Machine Learning, Computer Vision, Applied Machine Learning in Security, Advanced Machine Learning, Dynamic Programming, Graph Theory.

Frameworks and Concepts: Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Deep Generative Models, Reinforcement Learning, Programming Language Systems (Compilers, LLVM, IR, etc.).

Full Publications

[**ICSE** 2025] Can an LLM find its way around a Spreadsheet? Cho-Ting Lee, et al., *Shengzhe Xu*, Naren Ramakrishnan.

[**IEEE ICC** 2025] Wireless Knowledge Grounding in Smaller LLMs using Retrieval Augmented Generation and Fine-Tuning. A. Neeser, et al., *Shengzhe Xu*, Walid Saad, Naren Ramakrishnan.

[Neurips 2025, under review] When can isotropy help adapt LLMs' next word prediction to numerical domains? R. Shelim, *Shengzhe Xu*, et al., Walid Saad, Naren Ramakrishnan.

[COLM 2025, under review] Can an LM Induce a Graph? Investigating Memory Drift and Context Length. R. Yousef, et al., *Shengzhe Xu*, Naren Ramakrishnan.

[KDD-Prompt Optimization 2025, under review] The Prompt is Mightier than the Example. *Shengzhe Xu*, Nikhil Muralidhar, Naren Ramakrishnan.

[KDD-SKnow-LLM 2025, under review] Why LLMs Are Bad at Synthetic Table Generation (and what to do about it). *Shengzhe Xu*, et al., Mandar Sharma, Nikhil Muralidhar, Naren Ramakrishnan.

[CIKM 2025, under review] Optimizing Product Provenance Verification using Data Valuation Methods. R. Yousef, et al., *Shengzhe Xu*, Ruoxi Jia, Chang-Tien Lu, Naren Ramakrishnan.

[**COLM** 2024] Information Guided Regularization for Fine-tuning Language Models. Mandar Sharma, et al., *Shengzhe Xu*, Naren Ramakrishnan.

[**IEEE Network** 2024, IF: 6.8] Large Multi-Modal Models (LMMs) as Universal Foundation Models for AI-Native Wireless Systems. *Shengzhe Xu*, et al., Walid Saad, Naren Ramakrishnan.

[IEEE BigData 2024 **Best Paper Award**] Data Augmentations to support Speculative Reasoning in LLMs.R. Yousuf, et al., *S. Xu*, N. Ramakrishnan.

[IEEE BigData 2024] Forecasting Migration Patterns and Land Border Encounters. R. Yousuf, *Shengzhe Xu*, et al.

[ICAIF 2023] ML-assisted Optimization of Securities Lending. A. Prasad, et al., *Shengzhe Xu*, Naren Ramakrishnan.

[KDD-MLHat 2021] STAN: Synthetic Network Traffic Generation with Generative Neural Models. *Shengzhe Xu*, et al., Manish Marwah, Naren Ramakrishnan.

[ICPC 2019] Meditor: inference and application of API migration edits. *Shengzhe Xu*, Ziqi Dong, Na Meng.