Yucheng Shi

HomePage | LinkedIn | Github

Email: yucheng.shi@uga.edu
Mobile: +1-706-765-5574

Summary

My research focuses on the exciting fields of **Responsible AI**, specializing in the development of **interpretable and reliable** AI systems. My research projects include foundation model **post-training** (instruction fine-tuning, PPO/DPO training), multi-modal **synthetic data** generation, **RAG**, and foundation model **interpretability**.

Education

- **University of Georgia**
  Ph.D. in Computer Science (Advisor: Ninghao Liu) *Jan 2022 - Present*
- **North China Electric Power University**
  B.Eng. and M.S. in Renewable Energy Science and Engineering *Sep 2014 - Jun 2021*

Experience

- **Harvard Medical School**
  Research Intern (Mentor: Xiang Li) *May 2024 - Sept 2024*
  - Developed MGH Radiology LLM by further pre-training a **LLaMA-70B** on **6.5M+** radiology reports with **DeepSpeed** accelerators, achieved **93%** improvement in ROUGE compared to original LLaMA model.
  - Proposed a RAG system that decomposes complex medical questions into search-engine-friendly **synthetic queries** for improved retrieval, enhancing LLaMA-8B's accuracy by **11%** on USMLE dataset.

Publications (Full List)

- *"Enhancing Cognition and Explainability of Multimodal Foundation Models with Self-Synthesized Data."*
  – **Yucheng Shi**, Quanzheng Li, Jin Sun, Xiang Li, Ninghao Liu.
  • *International Conference on Learning Representations* (**ICLR**), 2025.
- *"ECHOPulse: ECG Controlled Echocardio-gram Video Generation."*
  – Yiwei Li, Sekeun Kim, Zihao Wu, Hanqi Jiang, Yi Pan, Pengfei Jin, Sifan Song, **Yucheng Shi**, Xiaowei Yu, Tianze Yang, Tianming Liu, Quanzheng Li, Xiang Li
  • *International Conference on Learning Representations* (**ICLR**), 2025.
- *"MQuAKE-Remastered: Multi-Hop Knowledge Editing Can Only Be Advanced with Reliable Evaluations."*
  – Shaochen Zhong, Yifan Lu, Lize Shao, Bhargav Bhushanam, Xiaocong Du, Yixin Wan, **Yucheng Shi**, Daochen Zha, Yiwei Wang, Ninghao Liu, Kaixiong Zhou, Shuai Xu, Kai-Wei Chang, Louis Feng, Vipin Chaudhary, Xia Hu.
  • *International Conference on Learning Representations* (**ICLR**), 2025.
- *"Quantifying Multilingual Performance of Large Language Models Across Languages."*
  – Zihao Li, **Yucheng Shi**, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, Mengnan Du.
  • *Association for the Advancement of Artificial Intelligence* (**AAAI**), 2025.
- *"Retrieval-enhanced Knowledge Editing for Multi-hop Question Answering in Language Models."*
  – **Yucheng Shi**, Qiaoyu Tan, Xuansheng Wu, Shaochen Zhong, Kaixiong Zhou, Ninghao Liu.
  • *The Conference on Information and Knowledge Management* (**CIKM**), 2024.
- *"MKRAG: Medical Knowledge Retrieval Augmented Generation for Medical Question Answering."*
  – **Yucheng Shi**, Shaochen Xu, Tianze Yang, Zhengliang Liu, Tianming Liu, Quanzheng Li, Xiang Li, Ninghao Liu.
  • *American Medical Informatics Association Annual Symposium* (**AMIA**), 2024,
  • ***Distinguished Paper Award***.

- *"Usable Interpretability for Large Language Models."*
  – **Yucheng Shi**, Haiyan Zhao, Fan Yang, Xuansheng Wu, Mengnan Du, Ninghao Liu.
  • IEEE International Conference on Healthcare Informatics *(**IEEE ICHI**)*, Tutorial, 2024.

- *"Could Small Language Models Serve as Recommenders? Towards Data-centric Cold-Start Recommendation."*
  – Xuansheng Wu, Huachi Zhou, **Yucheng Shi**, Wenlin Yao, Xiao Huang, Ninghao Liu.
  • *The Web Conference* (**WWW**), 2024.

- *"Automated Natural Language Explanation of Deep Visual Neurons with Large Models."*
  – Chenxu Zhao, Wei Qian, **Yucheng Shi**, Mengdi Huai, Ninghao Liu.
  • *Association for the Advancement of Artificial Intelligence* (**AAAI**), Student abstract, 2024.

- *"Chatgraph: Interpretable Text Classification by Converting Chatgpt Knowledge to Graphs."*
  – **Yucheng Shi***, Hehuan Ma*, Wenliang Zhong*, Qiaoyu Tan, Gengchen Mai, Xiang Li, Tianming Liu, Junzhou Huang.
  • *International Conference on Data Mining* (**ICDM**), Data Mining Workshops, 2023.

- *"Black-box Backdoor Defense via Zero-shot Image Purification."*
  – **Yucheng Shi**, Mengnan Du, Xuansheng Wu, Zihan Guan, Jin Sun, Ninghao Liu.
  • *Conference on Neural Information Processing Systems* (**NeurIPS**), 2023.

- *"GiGaMAE: Generalizable Graph Masked Autoencoder via Collaborative Latent Space Reconstruction."*
  – **Yucheng Shi**, Yushun Dong, Qiaoyu Tan, Jundong Li, Ninghao Liu.
  • *Conference on Information and Knowledge Management* (**CIKM**), 2023.

- *"ENGAGE: Explanation Guided Data Augmentation for Graph Representation Learning."*
  – **Yucheng Shi**, Kaixiong Zhou, Ninghao Liu.
  • *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (**ECML-PKDD**), 2023.

- *"Expected output calculation based on inverse distance weighting and its application in anomaly detection of distributed photovoltaic power stations."*
  – **Yucheng Shi**, Weiguo He, Jian Zhao, Aoyu Hu, Jingna Pan, Haizheng Wang, Honglu Zhu.
  • *Journal of Cleaner Production* **(JCP) (IF=11.1)**, 2020.

Preprints

- *"CORTEX: Concept-Oriented Token Explanation in Vector-Quantized Generative Model."*
  – Tianze Yang*, **Yucheng Shi***, Mengnan Du, Xuansheng Wu, Qiaoyu Tan, Jin Sun, Ninghao Liu.
  • *(**under review**)*, 2024.

- *"MGH Radiology Llama: A Llama 3 70B Model for Radiology."*
  – **Yucheng Shi**, Peng Shu, Zhengliang Liu, Zihao Wu, Quanzheng Li, Xiang Li.
  • *(**arXiv**)*, 2024.

- *"Usable XAI: 10 Strategies Towards Exploiting Explainability in the LLM Era."*
  – Xuansheng Wu*, Haiyan Zhao*, Yaochen Zhu*, **Yucheng Shi***, Fan Yang, Tianming Liu, Xiaoming Zhai, Wenlin Yao, Jundong Li, Mengnan Du, Ninghao Liu.
  • *(**arXiv**)*, 2024.

- *"Interpretation of Time-Series Deep Models: A Survey."*
  – Ziqi Zhao*, **Yucheng Shi***, Shushan Wu*, Fan Yang, Wenzhan Song, Ninghao Liu.
  • *(**arXiv**)*, 2023.

Selected Projects

- **Large Foundation Model Post-training [ICLR2025, arxiv2024a1]:**
  - Designed a novel **multi-modal data-synthesis** pipeline for **LLaVA**, incorporating **rejection sampling** to generate high-quality interpretable training data, significantly improving the model's expert-level **object identification and explanation** capabilities on benchmarks from multiple domains.
  - Built medical domain-specific LLM using LLaMA-3-70B with **ZeRO-3 Offload** techniques.
  - Currently advancing **DPO/KTO** on LLaVA models using model internal states for better alignment.
- **Advanced RAG Systems [CIKM2024, AMIA2024]:**
  - Proposed a novel RAG system for **multi-hop model editing** by next fact prediction on a knowledge graph containing **over 5 million facts**, achieving SOTA performance on the MQUAKE benchmark.
  - Designed a **dense retrieval**-based medical RAG, improving **8%** in medical QA accuracy with Vicuna.
- **Trustworthy AI Framework [NIPS2023, arxiv2024a2, ICDM2023, arxiv2024a3, arxiv2023, AAAI2024]:**
  - Designed a backdoor attack defense strategy using zero-shot purification with **diffusion models**.
  - Developed a novel interpretability framework for **VQ-GAN** that identifies concept-specific visual token combinations, enabling transparent analysis and targeted **image editing** capabilities.
  - Proposed a post-hoc explanation framework leveraging foundation models for **automated semantic interpretation** of neural network neurons, enabling **scalable** analysis without human intervention.
  - Built interpretation pipelines to explain LLMs and LMMs decisions at token/feature level.
- **Graph Self-supervised Learning [CIKM2023, ECML-PKDD2023]:**
  - Developed novel GNNs combining **contrastive learning** with explanation-guided augmentation.
  - Designed generalizable **graph masked autoencoder** supporting multi-task learning such as node classification/clustering and link prediction tasks.

Technical Skills

- **Programming:** Python, PyTorch, JAX, Shell Scripting, MySQL.
- **LLMs/LMMs Development:** Transformers, PEFT, TRL, vLLM, Flash Attention.
- **ML Infrastructure:** Linux, Git, Docker, Slurm, Distributed Training (DeepSpeed, FSDP, Accelerate).

Activities

- Talk at Harvard Medical School AIxMed Seminar (Aug 2023)
  –Topic: LLMs editing with external knowledge graphs for medical QA.
- Talk at Harvard Medical School AIxMed Seminar (Oct 2024)
  –Topic: Self-synthesized data can help improve cognition and explainability of LMMs.
- Reviewers at top ML conferences and journals (NeurIPS, ICLR, WWW, AISTAT, IEEE TNNLS).

Awards

- AMIA 2024 Distinguished Paper Award.
- NeurIPS 2023 Scholar Award.
- China National Scholarship (2020).
- Pacemaker to Graduate Student (top 0.8%) (2020).
- First-class Scholarships (2019, 2020).