Yucheng Shi

HomePage | LinkedIn | Github | Google Scholar

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

Summary

Ph.D. Candidate in Computer Science with expertise in **Large Language Models (LLMs), Large Multi-modal Models (LMMs), and Trustworthy Machine Learning**. Specialized in developing **interpretable and responsible** AI systems, with extensive experience in foundation model **post-training** (instruction fine-tuning, DPO/GRPO training), multi-modal **synthetic data** generation, **RAG**, and foundation model **interpretability**. Published ML research at top-tier conferences (ICLR, NeurIPS, WWW, CIKM, AAAI, ECML-PKDD, ICDM, AMIA).

Education

- **University of Georgia**
  Ph.D. in Computer Science (Advisor: Ninghao Liu) — *Jan 2022 - Dec 2025 (Expected)*
- **North China Electric Power University**
  B.Eng. and M.S. in Renewable Energy Science and Engineering — *Sep 2014 - Jun 2021*

Experience

- **Tencent AI Lab (Seattle)**
  Research Scientist Intern (Mentor: Wenhao Yu) — *May 2025 - Aug 2025*
- **Harvard Medical School**
  Student Researcher (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 **16%** on MedMCQA dataset.

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 **visual reasoning 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/GRPO** on Qwen2.5-VL for better multi-image understanding and reasoning.
- **Advanced RAG Systems [CIKM2024, AMIA2024 , arXiv2025]:**
  - 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, ICML2025, 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.

First-authored and Co-first-authored Publications (Full List)

Multi-modal Models: [1, 2, 9, 15, 19]; **RAG:** [3, 4, 5, 16]; **LLMs:** [6, 7, 17, 18]; **Trustworthy AI**: [8, 9, 11, 12, 13].

1 *“Towards Trustworthy GUI Agents: A Survey."*
– Yucheng Shi, Wenhao Yu, Wenlin Yao, Wenhu Chen, Ninghao Liu.
● *(arXiv)*, 2025.

2. *“CORTEX: Concept-Oriented Token Explanation in Vector-Quantized Generative Model."*
– Tianze Yang*, Yucheng Shi*, Mengnan Du, Xuansheng Wu, Qiaoyu Tan, Jin Sun, Ninghao Liu.
● (**ICML**), *International Conference on Machine Learning*, 2025.

3. *“Enhancing Cognition and Explainability of Multimodal Foundation Models with Self-Synthesized Data."*
– Yucheng Shi, Quanzheng Li, Jin Sun, Xiang Li, Ninghao Liu.
● (**ICLR**), *International Conference on Learning Representations*, 2025.

4. *“SearchRAG: Can Search Engines Be Helpful for LLM-based Medical Question Answering?"*
– Yucheng Shi, Tianze Yang, Canyu Chen, Quanzheng Li, Tianming Liu, Xiang Li, Ninghao Liu.
● *(Under review*), 2025.

5. *“Retrieval-enhanced Knowledge Editing for Multi-hop Question Answering in Language Models."*
– Yucheng Shi, Qiaoyu Tan, Xuansheng Wu, Shaochen Zhong, Kaixiong Zhou, Ninghao Liu.
● (**CIKM**), *The Conference on Information and Knowledge Management* , 2024.

6. *“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.
● (**AMIA**), *American Medical Informatics Association Annual Symposium*, 2024,
⋆ ***Distinguished Paper Award***.

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

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

9. *“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.

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

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

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

13. *“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.
● (**ICDMW**), *International Conference on Data Mining*, Data Mining Workshops, 2023.

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

Other Co-authored Papers

15. *“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
• (**ICLR**), *International Conference on Learning Representations*, 2025.

16. *“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.
• (**ICLR**), *International Conference on Learning Representations*, 2025.

17. *“Quantifying Multilingual Performance of Large Language Models Across Languages."*
– Zihao Li, **Yucheng Shi**, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, Mengnan Du.
• (**AAAI**), *Association for the Advancement of Artificial Intelligence* , 2025.

18. *“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.
• (**WWW**), *The Web Conference*, 2024.

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

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

- Dissertation Completion Award Assistantship 2025-2026.
- 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).