Yucheng Shi

HomePage | LinkedIn | Github

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

Summary

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

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*
  ◦ Led the development of MGH Radiology **LLaMA-70B**, which is fine-tuned on over **6.5 million** radiology reports, achieving a **93%** improvement in ROUGE scores compared to baseline models.
  ◦ Developed a RAG system using **synthetic queries** to decompose complex medical questions for precise content retrieval, improving LLaMA-3-8B's accuracy by **11%** on the USMLE benchmark.

Research Topics

- **Large Foundation Model Post-training [arxiv2024a1, arxiv2024a2]:**
  ◦ 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, arxiv2024a3, ICDM2023, arxiv2024a4, 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.

Selected Publications (Full List)

Multi-modal Models: [1,2,16]; **LLMs:** [3, 4, 7, 8, 14]; **RAG:** [5,6]; **Trustworthy AI**: [9, 10, 11, 12].

- **First-authored and Co-first-authored Papers**

  1. Enhancing Cognition of Multimodal Foundation Models, **[Under Review]**, 2024
  2. CORTEX: Concept-Oriented Token Explanation for LMMs, **[Under Review]**, 2024
  3. MGH Radiology Llama: A Llama 3 70B Model, **[arXiv]**, 2024
  4. Usable Interpretability for LLMs, **[ICHI]**, Tutorial, 2024
  5. Retrieval-enhanced Knowledge Editing for Multi-hop QA, **[CIKM]**, 2024
  6. MKRAG: Medical Knowledge RAG, **[AMIA]**, 2024
  7. Usable XAI: Strategies in the LLM Era, **[Under Review]**, 2024
  8. Chatgraph: Interpretable Text Classification, **[ICDM]**, Workshop, 2023
  9. Black-box Backdoor Defense via Zero-shot Image Purification, **[NeurIPS]**, 2023
  10. GiGaMAE: Generalizable Graph Masked Autoencoder, **[CIKM]**, 2023
  11. ENGAGE: Explanation Guided Data Augmentation, **[ECML-PKDD]**, 2023
  12. Interpretation of Time-Series Deep Models: A Survey, **[Arxiv]**, 2023
  13. Anomaly Detection for PV power stations, **[JCR]**, 2020

- **Other Co-authored Papers**

  14. Could Small Language Models Serve as Recommenders?, **[WWW]**, 2024
  15. LLMs for Traffic Crash Analysis, **[Computers]**, 2024
  16. Automated Explanation of Deep Visual Neurons, **[AAAI]**, Student Abstract, 2024
  17. Quantifying Multilingual Performance of LLMs, **[Arxiv]**, 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

- 370+ citations on Google Scholar.
- NeurIPS 2023 Scholar Award.
- China National Scholarship (2020).
- Pacemaker to Graduate Student (top 0.8%) (2020).
- First-class Scholarships (2019, 2020).