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** (instruction fine-tuning, PPO/DPO training), multi-modal **synthetic data** generation, **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*
  - 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.

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 **visual entity 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 and Explainability of Multimodal Foundation Models with Self-Synthesized Data, **[Under Review]**, 2024
  2. CORTEX: Concept-Oriented Token Explanation in Vector-Quantized Generative Model, **[Under Review]**, 2024
  3. MGH Radiology Llama: A Llama 3 70B Model for Radiology, **[arXiv]**, 2024
  4. Usable Interpretability for Large Language Models, **[ICHI]**, Tutorial, 2024
  5. Retrieval-enhanced Knowledge Editing for Multi-hop Question Answering in Language Models, **[CIKM]**, 2024
  6. MKRAG: Medical Knowledge Retrieval Augmented Generation for Medical Question Answering, **[AMIA]**, 2024
  7. Usable XAI: 10 Strategies Towards Exploiting Explainability in the LLM Era, **[Under Review]**, 2024
  8. Chatgraph: Interpretable Text Classification by Converting Chatgpt Knowledge to Graphs, **[ICDM]**,workshop,2023
  9. Black-box Backdoor Defense via Zero-shot Image Purification, **[NeurIPS]**, 2023
  10. GiGaMAE: Generalizable Graph Mask Autoencoder via Collaborative Latent Space Reconstruction, **[CIKM]**, 2023
  11. ENGAGE: Explanation Guided Data Augmentation for Graph Representation Learning, **[ECML-PKDD]**, 2023
  12. Interpretation of Time-Series Deep Models: A Survey, **[Arxiv]**, 2023
  13. Expected output calculation based on inverse distance weighting and its application in anomaly detection of distributed photovoltaic power stations, **[JCP]**, 2020

- **Other Co-authored Papers**
  14. Could Small Language Models Serve as Recommenders? Towards Data-centric Cold-Start Recommendation, **[WWW]**, 2024
  15. Leveraging Large Language Models with Chain-of-Thought and Prompt Engineering for Traffic Crash Severity Analysis and Inference, **[Computers]**, 2024
  16. Automated Natural Language Explanation of Deep Visual Neurons with Large Models, **[AAAI]**, Student Abstract, 2024
  17. Quantifying Multilingual Performance of Large Language Models Across Languages, **[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

- **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).