Leisheng Yu

470-696-9610 | ly50@rice.edu | leisheng-yu-2419731a9 | leishengyu.com | Google Scholar

Education

Rice University — Houston, TX
Ph.D. in Computer Science — *Aug 2022 – May 2026*
- Advisor: Dr. Xia "Ben" Hu

Emory University — Atlanta, GA
B.S. in Applied Mathematics (Magna Cum Laude) & B.A. in Computer Science — *Aug 2019 – May 2022*
- GPA: 3.99
- Advisor: Dr. Carl Yang
- Thesis: Deep Learning for EHR-Based Diagnosis Prediction: A General Recipe

Professional Summary

Machine learning engineer/researcher focused on production-scale **personalization** and **resource-constrained deployment** of specialized, trustworthy, and efficient models that cooperate with cloud-hosted **large foundation models**. Specializes in:
Large/Small Language Models: efficient inference, fine-tuning, evaluation, structured reasoning, LLM fusion, agentic LLMs
Applications: recommender systems, digital advertising, time-series analysis, predictive healthcare
Trustworthy AI: explainability, uncertainty calibration, LLM-based explanations, hallucination detection

Experience

DATA Lab at Rice University — Houston, TX
Graduate Research Assistant — *Aug 2022 – Present*
- Developing privacy-preserving device-cloud collaboration for small/large language models such as inference routing
- Designed fair, intervenable, and intrinsically interpretable algorithms to advance personalized healthcare
- Collaborated with UTHealth in individualized sedation level prediction under general anesthesia
- Developed time-series classification algorithms, including LLM-based methods, deployed on edge devices for anomaly detection in the oil & gas (American Innovations) and HVAC (Daikin, Trane) industries

Amazon — Seattle, WA
Applied Scientist Intern — *May 2025 – Aug 2025*
- Mentored by Dr. Peiyao Wang and Dr. Qilin Qi at Prime Video & Amazon MGM Studios
- Reduced the offline–online performance gap for Prime Video recommendation models by 11.59% by building an efficient, training-free, behavior-aware LLM-based reward agent, leveraging vLLM for high-throughput inference

Samsung Electronics America — Mountain View, CA
Machine Learning Research Engineer Intern — *Jun 2024 – Aug 2024*
- Mentored by Dr. Wei-Yen Day and Dr. Rui Chen at Samsung Ads
- Proposed a domain adaptation algorithm addressing delayed feedback that improved the mobile performance product's return on ad spend (RoAS) by 6.46% and generated over $50K in-app purchases
- Ran large-scale hyperparameter tuning & online deployment with CI/CD pipelines, Airflow, and Amazon QuickSight

Samsung Electronics America — Mountain View, CA
Machine Learning Research Engineer Intern — *Jun 2023 – Aug 2023*
- Built the first user response prediction model for mobile performance ads, serving over 20M mobile users
- Developed a Mixture of Experts (MoE) model that increased the impression-to-install (i2i) rate by 15.84%
- Unblocked future applications of pretrained user/ad embeddings for better cold-start mitigation and look-alike targeting

Emory Graph Mining Group — Atlanta, GA
Research Assistant — *Dec 2020 – Sep 2022*
- Collaborated with Emory School of Nursing to build neural networks for diagnosis prediction & drug recommendation
- Conducted research on recurrent nets (vanilla RNN, LSTM, GRU), CNNs, GNNs, and Transformers for mining electronic health record (EHR) data to advance personalized healthcare

Publications

[CHIL 2025] **L. Yu**, Y. Cai, M. Zhang, and X. Hu. "Self-Explaining Hypergraph Neural Networks for Diagnosis Prediction." *AHLI Conference on Health, Inference, and Learning*

[NeurIPS 2025 Workshop] **L. Yu***, Y. Chuang*, G. Wang, L. Zhang, Z. Liu, X. Cai, Y. Sui, V. Braverman, and X. Hu. "Confident or Seek Stronger: Exploring Uncertainty-Based On-device LLM Routing From Benchmarking to Generalization." *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle*

[NeurIPS 2025 Workshop] Y. Chuang, S. Li, J. Yuan, G. Wang, K. Lai, S. Sui, **L. Yu**, S. Ding, C. Chang, Q. Tan, D. Zha, and X. Hu. "LTSM-Bundle: A Toolbox and Benchmark on Large Language Models for Time Series Forecasting." *NeurIPS 2025 Workshop on Recent Advances on Time Series Foundation Models*

[ICDM 2024] **L. Yu**, Y. Cai, L. Chen, M. Zhang, W. Day, L. Li, R. Chen, S. Choi, and X. Hu. "Addressing Delayed Feedback in Conversion Rate Prediction: A Domain Adaptation Approach." *IEEE International Conference on Data Mining*

[ICDM 2023] Y. Tan, Z. Zhou, **L. Yu**, W. Liu, C. Chen, G. Ma, X. Hu, V. S Hertzberg, and C. Yang. "Enhancing Personalized Healthcare via Capturing Disease Severity, Interaction, and Progression." *IEEE International Conference on Data Mining*

[KDD 2022] Y. Tan, C. Kong, **L. Yu**, P. Li, X. Zheng, C. Chen, V. S Hertzberg, and C. Yang. "4SDrug: Symptom-based Set-to-set Small and Safe Drug Recommendation." *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*

** denotes equal contribution.*

Preprints

- **L. Yu**, P. Wang, Q. M. Zeng, R. M. Rustamov, Z. Wen, and Q. Qi. "Behavior-Infused Evidence-First Reasoning: Bridging the Offline-Online Gap in Recommendation." *Submitted to* ICLR 2026
- Y. Wang, X. Han, **L. Yu**, and N. Zou. "Beyond Fairness: Age-Harmless Parkinson's Detection via Voice." arXiv:2309.13292

Honors and Awards

ICDM 2024 NSF Travel Award *Nov 2024*

Rice Graduate Fellowship *Feb 2022*

Phi Beta Kappa: invitation-only honor for interdisciplinary academic excellence at Emory *Mar 2021*

Phi Eta Sigma: invitation-only first-year honor for 4.0 GPA at Emory *Dec 2019*

Dean's list: top 20 percent of all college students *2019, 2020, 2021, 2022*

Academic Activities

Teaching Assistantships

COMP 514 Algorithms, Complexity and Approximations *Fall 2025*

COMP 552 Reinforcement Learning *Fall 2023*

COMP 631 Introduction to Information Retrieval *Spring 2023*

CS 253 Data Structures and Algorithms *Spring 2022*

Service

Conference Reviewer: CIKM 2023, KDD 2024, SDM 2025, ACML 2025, CIKM 2025

Journal Reviewer: ACM Transactions on Intelligent Systems and Technology, IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Cognitive and Developmental Systems, ACM Transactions on Computing for Healthcare, IEEE Transactions on Knowledge and Data Engineering

Invited Talks

Uncertainty Estimation for Large/Small Language Models at Samsung Ads *Feb 2025*

Tabular Data Imputation at Samsung Ads *Oct 2024*

Technical Skills

Languages: Python, C/C++, Java, C#, Bash, HTML, R, Swift
Data & Analytics: SQL, Spark, NumPy, Pandas
ML & LLM Frameworks: PyTorch, TensorFlow, Keras, Hugging Face Transformers, scikit-learn, vLLM, PyTorch Geometric, DGL, NLTK, FAISS, OpenAI API, DeepSpeed, bitsandbytes, FlashAttention-2
MLOps & Cloud: Apache Airflow (MWAA), CI/CD, Git/GitHub, A/B testing, GPU acceleration (CUDA, cuDNN), AWS (EC2, SageMaker, S3, QuickSight), Linux, Conda (Anaconda), LangChain, Agent Development Kit