Gaoxiang Luo

gluo0401@gmail.com | 267-366-0327

EDUCATION

UNIVERSITY OF MINNESOTA
Ph.D. in Computer Science
2024-x | Minneapolis, MN

UNIVERSITY OF PENNSYLVANIA
M.S. in Computer & Information Science
2022-2024 | Philadelphia, PA
Ph.D. leaving with a master degree

UNIVERSITY OF MINNESOTA
B.S. in Computer Science
Minor in Math
2018-2021 | Minneapolis, MN
Graduated with High Distinction

LINKS

Google Scholar: z.umn.edu/gl-scholar
Homepage: gaoxiangluo.github.io
LinkedIn: z.umn.edu/gl-linkedin
GitHub: z.umn.edu/gl-git

RESEARCH INTEREST

Diffusion/Flow Matching
LLM Uncertainty
Conformal/Selective Prediction
Bayesian Optimization
AI for Healthcare

HONORS/AWARDS

2025 AgentDS Overall 1st Place
2025 DSI-MnDRIVE PhD Fund
2024 OpenAI Researcher Access Program
2021 Google CSRMP[a] Scholar
2021 UROP Scholarship x 2
2021 Maximillian Lando Scholarship

SERVICE

SYMP. | Organizer
MMLS'24 (Web Chair)

CONF. | Reviewer
CVPR'24(4), UAI'24(5), MICCAI'24(4)
AAAI'25(6), CVPR'25(3), ICLR'26(1)
CVPR'26(1)

COMP. | Judge
OpenAI Custom GPT Hackathon in Twin Cities (Sep. 2025)

[a]CS Research Mentorship Program

PUBLICATION

ACADEMIC PAPER

[1] Center AI Safety, Scale AI, HLE Contributors Consortium. *A benchmark of expert-level academic questions to assess AI capabilities*. Nature 2026. [https://www.nature.com/articles/s41586-025-09962-4] [Data]

[2] Gaoxiang Luo, Aryan Deshwal. *COM-BOM: Bayesian Exemplar Search for Efficiently Exploring the Accuracy-Calibration Pareto Frontier*. EMNLP 2025. [https://aclanthology.org/2025.emnlp-main.1027] [Code]

[3] Birra R Taha, Gaoxiang Luo, Anant Naik, Luke Sabal, Ju Sun, et al. *Automated Ventricular Segmentation in Pediatric Hydrocephalus: How Close Are We?* Journal of Neurosurgery: Pediatrics. [https://doi.org/10.3171/2025.2.PEDS24590]

[4] Zhanming Chen, Minghe Lu, Minzhu Zhao, Gaoxiang Luo, et al. *Empowering Farming Communities Through Information Tracking: A Design Approach to Crop Planning and Management*. CHI 2025 LBW. [https://dl.acm.org/doi/10.1145/3706599.3719713]

[5] Xue Wang, Gaoxiang Luo. *MetaMate: Large Language Model to the Rescue of Automated Data Extraction for Educational Systematic Reviews and Meta-analyses*. SREE 2024. [https://metamate.online]

[6] Haowen Lai, Gaoxiang Luo, Yifei Liu, Mingmin Zhao. *Enabling Visual Recognition at Radio Frequency*. MobiCom 2024. Best Demo Award. [https://dl.acm.org/doi/10.1145/3636534.3649369] [Project] [Code] [Data]

[7] Le Peng, Gaoxiang Luo, Sicheng Zhou, Jiandong Chen, Ziyue Xu, Rui Zhang, Ju Sun. *An In-Depth Evaluation of Federated Learning on Biomedical Natural Language Processing*. npj Digital Medicine. [https://www.nature.com/articles/s41746-024-01126-4] [Code]

[8] Le Peng, Hengyue Liang, Gaoxiang Luo, Taihui Li, Ju Sun[a]. *Rethink Transfer Learning in Medical Image Classification*. BMVC 2023 (Oral). [https://arxiv.org/abs/2106.05152] [Code]

[9] John Burns, Zachary Zaiman, Jack Vanschaik, Gaoxiang Luo, et al. *Ability of Artificial Intelligence to Identify Self-Reported Race in Chest X-Ray Using Pixel Intensity Counts*. Journal of Medical Imaging (JMI). [https://doi.org/10.1117/1.JMI.10.6.061106]

[10] Le Peng, Gaoxiang Luo, Andrew Walker, Ju Sun, Christopher J Tignanelli, et al. *Evaluation of Federated Learning Variations for COVID-19 Diagnosis Using Chest Radiographs from 42 US and European Hospitals*. Journal of the American Medical Informatics Association (JAMIA). [https://doi.org/10.1093/jamia/ocac188]

[11] Majid Farhadloo, Carl Molnar, Gaoxiang Luo, Yan Li, Shashi Shekhar, et al. *SAMCNet: Towards a Spatially Explainable AI Approach for Classifying MxIF Oncology Data*. KDD 2022. [https://dl.acm.org/doi/10.1145/3534678.3539168]

[a]CV Reference Contact: Prof. Ju Sun [www.sunju.org]

SKILLS

Python • C/C++ • SQL • LaTeX
React • Next.js • TypeScript • FastAPI
PyTorch • JAX • Scikit-Learn
Docker • K8s • Anyscale • Terraform
MLflow • Wandb • Opik • Grafana
Snowflake • Qdrant • S3 (Vectors)
Diffusion/Flow-Matching • LLM/VLM
vLLM/SGLang • RAG • Triton
verl • LLaMA-Factory • Ray

PATENTS

2024 US18/829858 (Zscaler)
2024 US63/751820 (UPenn)
2022 US18/101620 (Cisco)
2022 US17/828582 (Cisco)

TEACHING & TUTORING

FIFE-PENN CS ACADEMY
K-8 Coding Instructor
Sep. 2022 - Dec. 2022 | Philadelphia, PA

• Teaching 2-5th grade students coding in Scratch in the after-school program.

UNIVERSITY OF MINNESOTA
Teaching Assistant[a]

Fall 2024	CSCI 5525 ML Theory
Fall 2020	CSCI 2011 Disc. Math
Spring 2021	CSCI 2011 Disc. Math
Fall 2021	CSCI 2033 Lin. Alg.

Library Peer Tutor
Sep. 2020 - Dec. 2020 | Minneapolis, MN

• Tutor single-and-multivariable calculus, linear algebra, intro physics, intro stats, and some programming in Python and C.

SOFTWARE

2024 MetaMate
Supported by Synthesis Comp. (YC S24)

[a]This is a measure by Spoken English Test for Teaching Assistants [SETTA].

RESEARCH

ENDEAVOR.AI | Research Engineer
June 2025 - August 2025 | San Francisco, CA

- Post-trained multi-turn tool-call LLM agents with GRPO and search tools for product-catalog matching (a shared problem across our customers) on high-quality rollouts synthesized from frontier LLMs. (Tool: verl, Ray, MCP)
- Serving post-trained personalized LLM agents at scale for our customers, reducing API costs significantly. (Stack: vLLM production stack, Terraform, AWS EKS)
- Built FastAPI microservices backend with React frontend featuring async OCR/VLM document processing, optimized RAG pipeline (parsing, chunking, embedding, indexing, retrieval, reranking), and LLM agentic search with browser-use and code interpreter; architected with SQS workers, retry mechanisms, and CI/CD automation for production-scale conversational document Q&A. (Tool: Prisma, Postgres, ChromaDB, SageMaker, Node.js, Docker)

ZSCALER | Student Researcher
January 2025 - May 2025 | San Jose, CA

- Built **one** data plane for both structural **tabular data** (e.g., Snowflake) via text-to-SQL and **unstructured data** (e.g., S3, Google Drive) via RAG, ready for downstream LLM apps (e.g., Sales, Support Cases, Finance, Lead Gen).
- Formulated and solved black-box LLM optimization problem of relevant **table catalog search for text-to-SQL backend**, improving retrieval accuracy by 44.62%.
- Developed **inference-time scaling** method for text-2-SQL with principal **uncertainty estimate**, improving answer correctness by 21.52% and reducing SLA to 5s. (Host: Dr. James Zhu) (Tech: Anyscale, Qdrant, Snowflake)

ZSCALER | Machine Learning Engineering Intern
May 2024 - August 2024 | San Jose, CA

- Developed a **multi-agent LLM framework** for enterprise data analysis that integrates **RAG-based QA**, **text-to-SQL**, and **text-to-visualization**, enabling highly accurate, efficient and dynamic data exploration; filed a patent (US18829858).
- In **churn analysis**, our framework reduced costly reliance on BI tools, moved past dashboards with finite permutations, and ensured **complete coverage** for smaller accounts lacking dedicated data support.
(Host: Dr. James Zhu) (Tech: AutoGen, Snowflake, AWS, DBT)

XYLO AI | Advisor
October 2023 - April 2024 | Remote

- Led R&D on **data/label generation** via LLM to train ML regression models for communication quadrant analyses, which was added to **the heart of beta product**. (CTO: Roger Lam) (Tools: Kubernetes, Azure ML, Langchain)
- Collaborated extensively with **cross-functional teams**, including product and engineering to prioritize R&D projects based on **strategic importance**, and incorporate R&D discoveries into feature development.

UNIVERSITY OF MINNESOTA | Research Professional I
February 2024 - May 2024 | Remote

- Contributed to **large language model (LLM)** experiments in demonstrating language models (e.g., GPT-2) trained with federated learning significantly outperform LLMs that have **1,000x** more parameters (e.g., GPT-4, PaLM2 Unicorn, Claude3 Opus) with few-shot prompting in biomedical information extraction tasks. (See publication [7]) (Skills: LLMOps)

UNIVERSITY OF PENNSYLVANIA | CIS PhD Research Fellow
September 2022 - August 2023 | Philadelphia, PA

- Achieved **LiDAR-comparable** 3D range imaging (median absolute error=3.4cm) using radio-frequency (RF) signals, **for the first time**, enabling RF-based visual recognition tasks such as object detection and semantic segmentation. (Tools: detectron2, MMsegmentation) (Advisor: Prof. Mingmin Zhao)
  - Designed a novel **2D approach to tackle 3D learning**, significantly reducing memory footprint and FLOPs, accelerating inference speed by 5.4x.
  - Directed the creation of the **first public dataset** featuring 11,033 paired RF/LiDAR data points across 12 buildings with indoor semantic and detection labels, thereby facilitating future research in RF imaging. (See publication [6])

COURSEWORK

UMN

Data Struct. and Alg. (CSCI 4041)
Operation System (CSCI 4061)
Software Engineering (CSCI 5801)

———————————————————

Applied Linear Algebra (MATH 4242)
Artificial Intelligence (CSCI 4511W)
Data Mining (CSCI 5523)
Machine Learning (CSCI 5525)
Spatial Data Science (CSCI 5715)
ML Theory (CSCI 8980)
Deep Learning (CSCI 8980)
Generative AI (STAT 8931)
LLM Systems (CSCI 8980)
RL and BayesOpt (CSCI 8980)

UPENN

Software System (CIS 5050)
Networked System (CIS 5530)
Big Data Analytics (CIS 5450)

———————————————————

Machine Learning (CIS 5200)
Analysis of Algorithms (CIS 5020)
Theory of Computation (CIS 5110)

CISCO RESEARCH | AI/ML Research Intern

Feb 2022 – July 2022 | Remote

- Contributed to a novel **scalable federated learning (FL) system** by implementing abstraction topologies to simply FL deployment; filed 2 patents (US17828582 and US18101620) enhancing model convergence and communication efficiency. The system is now open-sourced at **Project Flame**.
  (Host: Dr. Myungjin Lee) [Slide] (Stacks: K8s, Docker, MLflow, Git)

CISCO RESEARCH | Research Fellow

May 2021 – May 2022 | Remote

- Facilitated cross-functional collaboration between tech and medical teams to develop an automated **multi-modal** rib fracture detection system on chest X-rays for real-time deployment. This work is featured on two undergraduate research conferences NCRC 2022 and NCUR 2022.
  (Skills: data analysis, project management, communication)
  (Advisor: Prof. Christopher Tignanelli & Prof. Ju Sun)

UNIVERSITY OF MINNESOTA | Undergrad Research Assistant

August 2020 – May 2022 | Minneapolis, MN

- Proposed a novel **truncated transfer learning** (TL) method in medical imaging classification under data-poor regimes, that consistently leads to superior performance compared to other TL strategies. Our method is **versatile** for various deep neural networks and **adaptable** to other tasks (e.g., segmentation). (See publication [8]) (Tools: PyTorch, NumPy) (Advisor: Prof. Ju Sun)

- Improved AUPRC of chest x-ray COVID-19 classification by 9% compared to local training, by implementing real-data federated learning with several partner institutes. The case study is featured in the white paper – *Federated Learning for Healthcare Using NVIDIA Clara* (See publication [10])
  (Tools: NVFlare, MONAI) (Advisor: Prof. Ju Sun)

- Outperformed current state-of-the-art point set classifiers in terms of accuracy and F1-score on our tumor cell datasets by designing a novel **S**patial-interaction **A**ware **M**ulti-**C**ategory deep neural **Net**work (**SAMCNet**), contributing **location representation and point pair attention layers** for multi-categorical point set classification. See Publication [11]. (Advisor: Prof. Shashi Shekhar)

MEDIA

[1] *Scientists work on 'superhuman' vision systems for robots*. In BBC. Feb. 13, 2025.

[2] *Interning with Cisco Research*. In Cisco Emerging Tech & Incubation (ET&I) Blog. Aug. 12, 2022.

[3] *First gen student chosen for Google mentorship program*. In UMN CS&E Department News. Nov. 19, 2021.

[4] *CSpotlight: Experiencing research as an undergrad*. In UMN Department of Computer Science & Engineering spotlight program. May 12, 2021.

[5] *Application of Artificial Intelligence to Help Fight COVID-19* In Minnesota Undergraduate Research & Academic Journal (MURAJ), Vol.4 No.3, 2021.