Linxi Wu

linxiw@unc.edu | (984) 261-7719 | Chapel Hill, NC
https://opal.ils.unc.edu/ linxiw/Website/

Education

University of North Carolina at Chapel Hill Aug 2022 – May 2026 (Expected)
B.S. in Computer Science, B.S. in Mathematics, Minor in Statistics **GPA:** 3.924/4.0 (Dean's List)

- Relevant Coursework: Transformer Models, Deep Learning, Machine Learning, WebGL Graphics, Algorithm, Data Structure, Computer Organization, Real Analysis, Linear Algebra, Differential Equations, Optimization, Probability

Technical Skills

Programming & Infrastructure: Python, C++, Java, Go, MongoDB, HTML/CSS
Language Modeling: LLaMA2/3, Qwen2/2.5, GPT-2, LLaVa
Training: LoRA, QLoRA, SFT, instruction tuning, PPO, DPO
LLM Inference & Optimization: vLLM, GPTQ, prompt caching, TensorRT, Nginx
Reasoning & Control: Token routing, PPO, long-horizon planning

Research Experiences

MedReason-Multimodal: Biomedical VQA Reasoning May 2025 - Present
VLAA Lab in UCSC with Prof. Yuying Zhou

- Developed a structured data generation pipeline for multimodal biomedical reasoning tasks, synthesizing **image + question + reasoning + answer** samples grounded on knowledge graphs.
- Designed high-quality prompts to retrieve reasoning paths and simulate stepwise diagnostic thinking using **Qwen2.5-VL** and **GPT-4**-style models.
- Implemented quality and difficulty filtering modules, enabling dataset curation across medical VQA benchmarks such as **VQA-RAD, SLAKE, and PathVQA**.
- Led full-stack training of vision-language models (e.g., **LLaVA-Med**) with reasoning-augmented supervision to improve performance on complex biomedical tasks involving image-text fusion.

CITER: Collaborative Inference Framework Aug–Dec 2024
AIMING Lab in UNC with Prof. Huaxiu Yao

- Identified the limitation of myopic routing in multi-LLM collaboration, where locally optimal decisions at each token lead to long-term performance degradation in complex reasoning tasks.
- Co-developed **CITER**, a collaborative inference framework that learns a token-level routing policy between small and large LLMs using reinforcement learning to optimize long-horizon rewards.
- Achieved **+17% accuracy gain** on GSM8k and **+23% over myopic baselines** on MATH at equal inference cost; tuning led to **30% cost savings** vs. Co-LLM and RouteLLM.
- Demonstrated that long-term token-level planning is critical to collaborative generation, providing a scalable solution to efficient multi-model inference—an open challenge in controllable LLM deployment.

Internship Experiences

Temporal Information Extraction with LLaMA Jun–Aug 2024
China Unicom AI Center

- Recognized that fuzzy and context-dependent temporal expressions in customer service dialogues hinder reliable downstream automation such as form completion and scheduling.

- Designed and implemented an end-to-end pipeline using **LLaMA3 (8B, 70B), LLaMA2-13B**, fine-tuned via **LoRA/QLoRA** on A100-80G*4 GPUs; incorporated a **GPT-4–based automatic evaluator** for label correction and generalization analysis.
- Improved vertical-domain F1 from **77.6% to 88.5%**, and general-purpose F1 from **90.6% to 92.4%**; boost throughput by **30%**, supporting the real-time analysis of 10K+ China Mobile dialogues.

LLM Inference Deployment Dec 2023–Jan 2024
Xi'an Xinfang Electronic Technology Company

- Tasked with improving latency and throughput for a document intelligence product based on LLMs; Enhanced system scalability and stability for **production-level document processing** pipelines.
- Deployed a **Qwen2-72B-GPTQ-Int4** model using **vLLM** for enterprise text extraction and summarization, balancing speed and output fidelity.
- Achieved **20% speedup** by analyzing prompt-to-decode ratios and implementing token prefill caching.
- Integrated multi-machine deployment with **Nginx load balancing**, improving throughput **+70%** over legacy llama.cpp backend.