Bohan Yao

Email: s1104@cs.washington.edu — *LinkedIn* — *Website* — *GitHub*

Education

University of Washington Rising Senior, Bachelor of Science in Computer Science & Bachelor of Science in Mathematical Statistics **GPA: 4.0/4.0**

Relevant Coursework Natural Language Processing, Deep Learning, Machine Learning, Data Structures & Algorithms

Research Specialization

LLM Reasoning, Multi-Agent Systems, Tool-Augmented LLMs

Technical Skills

- **NLP** LLM Finetuning, PEFT, Multi-Agent Systems, Tool Use, Data Curation, xBERT models
- **Machine Learning** Deep Learning, Computer Vision, Deep RL, LLM Alignment, Statistical Learning
- **ML/Data Science Libraries** HF Transformers, HF Accelerate, HF TRL, Unsloth, LangChain, LlamaIndex, SentenceTransformers, NLTK, vLLM, Triton, PyTorch, TensorFlow, Keras, NumPy, FAISS, SciPy, Scikit-learn, Pandas, Matplotlib
- **Frameworks** MCP, React.js, REST APIs, Django, FastAPI
- **Programming Languages** Python, Java, C, C++, TypeScript, R, Julia, Bash, LaTeX

Work & Research Experience

Machine Learning Research Scientist Part Time @ ServiceNow September 2025 – Present

- Working on designing agentic systems capable of reasoning across multiple modalities.

Machine Learning Research Scientist Intern @ ServiceNow June 2025 – September 2025

- Developed **Agentic Reasoning Module**, a novel framework for automatic generation of multi-agent systems optimized for solving multi-step math, science, and commonsense reasoning tasks. Achieves 10.6% higher performance compared to previous state-of-the-art baselines across several challenging reasoning benchmarks.
- Published work at The 5th Workshop on Mathematical Reasoning and AI @ NeurIPS 2025.

Independent Research with Dr. Vikas Yadav October 2024 – June 2025

- Developed **Multi-TAG** framework for scaling the inference time compute of tool-augmented LLMs on complex math reasoning tasks. Achieves 13.7% higher performance compared to previous state-of-the-art baselines on challenging math benchmarks.
- Published work at EMNLP Findings 2025 and The 5th Workshop on Mathematical Reasoning and AI @ NeurIPS 2025.

Machine Learning Engineer Intern @ ServiceNow June 2024 – September 2024

- Invented novel LLM instruction finetuning technique that upcycles dense Transformer models into sparse Mixture-of-Expert models, then merges back to original dense architecture. Achieved state of the art results across three model families on HumanEval, MBPP, and internal benchmarks, improving performance by up to 14.7%, 5.5%, and 3.6%, respectively, with no increase in inference time compute cost.
- Gave an oral presentation about the work at ServiceNow AI conference.

Undergraduate Researcher @ Noah's ARK Lab December 2023 – Present

- Working on developing a novel system that for the first time, enables automatic documentation of linguistic features of low resource English dialects.

Publications

ARM: Discovering Agentic Reasoning Modules for Generalizable Multi-Agent Systems [arXiv]

Bohan Yao*, Shiva Krishna Reddy Malay, Vikas Yadav

[The 5th Workshop on Mathematical Reasoning and AI @ NeurIPS (2025)]

Diverse Multi-tool Aggregation with Large Language Models for Enhanced Math Reasoning [arXiv]

Bohan Yao*, Vikas Yadav

[EMNLP Findings (2025) & **The 5th Workshop on Mathematical Reasoning and AI @ NeurIPS** (2025)]