Xinyu Wang

linkedin.com/in/xinyuwang1209

University of Connecticut, Storrs, Connecticut
xinyuwang1209@gmail.com — +1 (860) 634-2830

SUMMARY

- PhD researcher in machine learning with hands-on experience in LLMs, generative AI, diffusion models, retrieval-augmented generation (RAG), language agents, agentic control, agent evaluation and safety, scalable machine learning systems.
- **3 years** of experience supporting a mission-critical real-time ML system for utility operations, with work spanning workflow automation, production validation, incident response, and stakeholder-facing delivery.
- **ML Systems / Engineering**: Python, C++, PyTorch, asynchronous CPU–GPU pipelines, large-scale evaluation harnesses, backend services, monitoring, experiment automation, and production support.

EDUCATION

University of Connecticut, Storrs, CT — Sep 2019 – Expected Aug 2026
Ph.D., *Computer Science and Engineering* — GPA: 3.97/4.00

University of Connecticut, Storrs, CT — Sep 2015 – May 2019
B.S.E. *Computer Science and Engineering* (Honors Program) — GPA: 3.81/4.00
B.A. *Mathematics* (Dual Degree) — Dean's List, New England Scholar

WORK EXPERIENCE

UConn Outage Prediction Model (OPM) at Eversource Energy Center — Storrs, CT
Machine Learning Systems Engineer (Graduate Technician) — May 2023 – Present

- Architected and largely rewrote a legacy ML operations stack into a fully automated system that ingests external data every 6 hours, transforms it into model-ready features, and triggers forecasting workflows through a custom web interface for internal teams and external collaborators.
- Built the end-to-end orchestration layer that executes trained ML models, returns results to stakeholders, and supports timely operational decisions through automated data processing, scheduling, and result delivery.
- Scaled typical event runs to 96-core parallel jobs with optional A100 GPU usage, and implemented adaptive scheduling that supports multiple concurrent events while automatically switching between full parallel execution and priority-based submission under constrained resources.
- Built monitoring, error tracking, performance logging, automatic retry, resumable execution, and failure alerting into the pipeline, enabling fully automated normal operation and faster recovery when infrastructure or data-source issues occur.
- Collaborated with internal researchers and external utility partners to deliver timely model outputs under real-world operational deadlines, including support during high-pressure weather events.
- Provided 24/7 on-call support during severe weather events, debugging failures and hot-patching workflows under strict delivery windows for a mission-critical ML system.

RESEARCH EXPERIENCE

Laboratory of Machine Learning and Health Informatics — Storrs, CT
Graduate Research Assistant — May 2017 – Present

- **LLM Agents, Agentic Control & Safety**
  - Developed learning and evaluation methods for long-horizon language agents that decide when to ask, delegate, verify, act, or escalate; relabeled exact counterfactual action values under the learner's own continuation policy and improved audited-seed success from 6.2% to 37.8%, utility from -0.237 to 1.051, and decision regret from 0.323 to 0.109.
  - Built and audited a benchmark and guard framework for tool-using coding agents, using route, provenance, and capability checks to control high-risk actions; strongest safe-family evaluations achieved 6/6 task success with zero unauthorized effects and zero route misfires.

- **Foundation Models & Model Adaptation**
  - Developed a data-free performance enhancer for model merging that can be applied on top of broadly used task-vector-style merging methods rather than a single merge recipe.
  - Improved average benchmark performance to 86.1 across 7 NLP tasks and 8 vision tasks without training, task data, or test-time tuning, demonstrating strong generalization for foundation-model adaptation.
  - Designed strong baseline comparisons under strict no-data constraints, emphasizing reproducible evaluation and consistent gains across all evaluated data-free baselines.
- **Sequence Modeling, Decoding & Search**
  - Developed structure-aware decoding and search methods for autoregressive sequence models, addressing repeated outputs and hidden structural collapse in constrained generation settings.
  - Engineered a distributed asynchronous generation and evaluation framework for autoregressive language models, with token-level value prediction and human-in-the-loop steering; scaled candidate generation to 1B+ samples in 6 days on 8 V100 GPUs.
  - Designed data processing and selection pipelines to expand coverage of novel outputs while reducing redundant generations, enabling broader exploration over large structured candidate spaces.
- **Representation Learning & Sequence-Level Optimization**
  - Developed a contrastive alignment method to detect semantically equivalent latent states across divergent autoregressive trajectories, improving knowledge consistency and downstream task performance.
  - Designed an RL correction mechanism for constrained decoding by estimating token-level invalidity and using Hellinger-distance-guided anchoring to improve search efficiency, stability, and output quality.
- **Multimodal Modeling & Applied Collaboration**
  - Built multi-view representation learning models for large-scale MRI and multimodal health data, integrating imaging, demographic, and auxiliary biomarkers for downstream prediction tasks.
  - Applied sequence-modeling and generative methods in interdisciplinary workflows with domain collaborators, including downstream candidate selection and evaluation in structured scientific settings.

SELECTED PUBLICATIONS

- X. Wang, K. Deng, F. Dou, J. Bi, J. Lu. "*DIAL: Data-Free Interference-Aware Layer-Scaling via Diagnostic Composition for Task-Vector Model Merging*", submitted to 2026 European Conference on Computer Vision (ECCV).
- X. Wang, F. Dou, J. Bi, M. Song "*SIGMA: Structure-Invariant Generative Molecular Alignment for Chemical Language Model via Autoregressive Contrastive Learning*" submitted to 2026 International Conference on Machine Learning (ICML).
- X. Wang, J. Bi, M. Song. "*Leveraging Partial SMILES Validation Scheme for Enhanced Drug Design in Reinforcement Learning Frameworks*", submitted to 2026 AAAI Conference on Artificial Intelligence.
- T. Zhu, F. Dou, X. Wang, J. Lu, J. Bi. "*Polyhedron attention module: learning adaptive-order interactions*". Advances in Neural Information Processing Systems 36 (NeurIPS), 9213–9225, 2023.
- S. Sahoo, C. Shende, M. Z. Hossain, P. Patel, Y. Niu, X. Wang, S. Ware, J. Bi. "*Cross-platform Prediction of Depression Treatment Outcome Using Location Sensory Data on Smartphones*". 2025 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), 2025.
- Z. Gao, X. Wang, B. Blumenfeld Gaines, X. Shi, J. Bi, M. Song. "*Fragment-based deep molecular generation using hierarchical chemical graph representation and multi-resolution graph variational autoencoder*". Molecular Informatics, 42(5), 2200215, 2023.
- J. Lu, J. Sun, X. Wang, H. R. Kranzler, J. Gelernter, J. Bi. "*Inferring phenotypes from substance use via collaborative matrix completion*". BMC Systems Biology, 12, 15–27, 2018.
- J. Lu, J. Sun, X. Wang, H. R. Kranzler, J. Gelernter, J. Bi. "*Collaborative phenotype inference from comorbid substance use disorders and genotypes*". 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2017.