Jeffrey Wang

925-915-7267 | Jeffreywang204@gmail.com | github.com/jjwang8 | Fremont, CA

Research Interests

I am interested in architectural mechanisms that enable genuine reasoning in neural networks, where models perform internal structured computation through search, verification, and planning instead of pattern matching. I focus on compact recurrent "reasoning cores" and learned world models that support recursive deliberation in latent space, generating and refining hypotheses, simulating outcomes, and backtracking before emitting outputs. In practice, I instantiate these ideas in activation-free polynomial networks and transformer-style backbones, studying how primitive design choices such as multiplicative interactions, normalization, and recurrence shape what structures models can efficiently learn and how these architectures interact with noisy preference feedback in RLHF-style alignment.

Education

University of Wisconsin–Madison Madison, WI

B.S. Computer Science (Honors), B.S. Data Science *Aug. 2022 – May 2026*

- **GPA: 4.0/4.0**; Dean's List (all semesters).

Selected Coursework

- **Mathematics / Theory:** Mathematical Foundations of Machine Learning (Graduate level course), Intro to Learning Theory
- **Artificial Intelligence / ML:** Honors Introduction to AI, Matrix Methods in ML, Comp Vision for Deep Learning, Big Data Systems

Research Experience

UW–Madison, PI: Grigorios Chrysos Oct. 2024 – Present

Undergraduate Research Assistant

- **Activation-free Polynomial Networks (PNs).** Designing CNN and attention-equivalent PNs built on multiplicative interactions (Hadamard products) rather than nonlinear activations.
- *CNN-equivalent PN:* Two branches with different dilation/kernel sizes; multiply branch outputs to capture cross-receptive-field interactions.
- *Attention-equivalent PN:* Modified self-attention replacing softmax with a 4th-degree polynomial and depthwise convs for Q/K/V to inject spatial inductive bias.
- **Results (ongoing):** Attention Variants: 6.5M-parameter PN ($\approx$ 1.2 GFLOPs) reaches 80.7% top-1; a 26M-parameter PN ($\approx$ 4.9 GFLOPs) reaches 84.3% top-1. Both models surpass prior polynomial networks by $\sim$ 2% at comparable sizes while using less than half the FLOPs and outperform activation-based baselines of similar design trained under the same recipe.
- CNN Variants: 6.4M-parameter backbone ($\approx$ 1.2 GFLOPs) reaches 80.2% top-1; a 26M-parameter backbone ($\approx$ 4.8 GFLOPs) reaches 83.9% top-1.
- **Neural Architecture Search:** Built a DARTS variant with an entropy-like penalty over edge operations to converge to a single op (instead of softmax mixtures); observed $\sim$15% relative error reduction over popular NAS variants on reference tasks.
- Presented early PN results (CIFAR-10) at the ECE Undergraduate Research Symposium, UW–Madison (2025).

UW–Madison, PI: Yixuan (Sharon) Li Jan. 2024 – Sept. 2024

Undergraduate Research Assistant

- Co-authored an **ICML 2025 Position Paper** on data-centric AI alignment; assisted with writing multiple sections, running/evaluating experiments on sources of noise in RLHF preference data, and analyzing reward-model strategies for cleaning datasets.
- Explored a noise-robust DPO variant using Generalized Binary Cross-Entropy (GBCE) in place of log-sigmoid; obtained modest accuracy/margin gains and improved stability vs. label-smoothing-style approaches.

Publications

- Min-Hsuan Yeh, **Jeffrey Wang**, Xuefeng Du, Seongheon Park, Leitian Tao, Shawn Im, Yixuan Li. *Challenges and Future Directions of Data-Centric AI Alignment*. Proceedings of the 42nd International Conference on Machine Learning (**ICML 2025**), Position Paper Track.
- *In preparation:* First-author manuscript on activation-free **Polynomial Networks** for efficient vision

Industry Experience

Google Sept. 2025 – Present
CS Capstone Project *Madison, WI*

- Developed a quantum-inspired solution to the join reordering problem using QAOA, formulating SQL query optimization as a QUBO problem and implementing the pipeline in Qiskit
- Achieved **15% faster** join orders compared to classical greedy heuristics on queries with 3-5 tables.
- Engineered SQL-to-QUBO transformation logic to convert relational query plans into quadratic optimization problems suitable for quantum annealing, enabling quantum algorithm application to real-world database workloads.

Amazon May 2025 – Aug. 2025
Software Development Engineer Intern *San Diego, CA*

- Delivered a reusable AWS CDK construct to onboard any DynamoDB table with standardized schemas and reconciliation/quality checks, cutting data-integration time from weeks to minutes.
- Built a high-throughput DynamoDB CDC → Apache Iceberg/S3 ingestion system (∼**100K events/s**) with PySpark/EMR and Kinesis; enabled real-time, ML-ready feature stores and analytics for **AI model training** at scale.
- Hardened reliability and security with CloudFormation/IAM and automated validation; shipped components in TypeScript, Java, and PySpark.

Correkt AI Aug. 2023 – Aug. 2024
Founding Machine Learning Engineer *Santa Barbara, CA*

- Co-built and launched a multimodal AI search engine serving **10K+ monthly users**; contributions helped secure funding from CNSI, Crescent Fund, and SBAA.
- Fine-tuned LLaMA-3-70B with RLHF; designed synthetic-data pipelines (LLM-judge ranking/modification) to curate millions of preference pairs; achieved **+10%** preference rate vs. competitor AIs.
- Implemented web-scale retrieval with Elasticsearch + distributed crawling, indexing **200M+** pages for real-time querying.

2Sigma School May 2023 – Aug. 2023
Software Engineer Intern *Remote*

- Fine-tuned LLM agents and NLP pipelines for personalized feedback, skill assessment, and question generation, cutting teacher prep time by **30%** and boosting student satisfaction by **23%**.
- Created automated hallucination detection by executing generated code, delivering **100%** error-free AP CS content.

Honors & Awards

- Dean's List (all semesters), University of Wisconsin–Madison
- USACO Gold (2021)
- ICPC Mid-Central USA Regional: Top 26 (team)

Technical Skills

- **Machine Learning:** Neural Networks, Transformers, RLHF, Computer Vision, NLP, RAG
- **Frameworks:** PyTorch, Hugging Face, DeepSpeed, TensorFlow, scikit-learn, Pandas, NumPy, PySpark
- **Cloud / Data:** AWS, GCP, DynamoDB, Kinesis, Apache Iceberg, EMR, Distributed Systems
- **Backend:** Django, Flask, MongoDB

- **Languages:** Python, C/C++, Java, TypeScript, JavaScript, SQL
- **DevOps:** Docker, Git, Linux, Nginx, CI/CD, Weights&Biases