Charles L. Wang

charles.w@columbia.edu charleslwang.github.io 860-593-0729 Google Scholar

Education

Columbia University, New York, NY 2022 – 2026
Double B.A. in Computer Science & Mathematics-Statistics

Experience

Columbia University Dec 2024 – Present
Research Assistant, Advised by Micah Goldblum and Julia Hirschberg New York, NY

- Co-trained **Bagel-Zebra-CoT**, fine-tuning a large multimodal reasoning model on a **180K-example** interleaved vision-language dataset and evaluating it across multimodal reasoning benchmarks using multi-GPU `torchrun`, sharded training workflows, and flash-attention-backed setup. (Accepted to ICLR 2026)
- Built GPU-aware ASR training and evaluation infrastructure in **PyTorch/NeMo**, implementing adapter-based PEFT, duration-bucketed **Lhotse** batching, mixed-precision training, gradient accumulation, and automated WER analysis for atypical-speech models.
- Engineered chunked and batched speech / multimodal inference pipelines with FFmpeg audio preprocessing, ONNX / Core ML export, and edge deployment; released open-source datasets, models, and training frameworks on Hugging Face and GitHub, with Zebra-CoT reaching **160,000+ downloads**.

Barclays Investment Bank Jun 2025 – Aug 2025
AI Intern New York, NY

- Developed **production runtime systems** for agentic AI, implementing real-time semantic telemetry, state-machine authorization, and drift detection for secure model execution and governance enforcement.
- Designed agentic AI runtime governance framework that is agnostic with different LLM agent frameworks, including LangChain, Autogen, CrewAI, Google ADK, and implements governance controls that outperform safety metrics in LangSmith, OpenTelemetry, etc.
- Built synthetic scenario generation and concurrent evaluation pipelines to stress-test governance behavior across simulated failures and attack cases, using multithreading / asyncio, schema-constrained JSON logs, and automated result aggregation.
- Accepted to AAAI 2026.

Barclays US Jun 2024 – Aug 2024
Quantitative Analyst Intern Wilmington, DE

- Built Barclays' in-house **CCAR challenger model**, analyzing 20+ years of macroeconomic data across numerous datasets, running extensive multivariable ML experiments, and using **L1/L2 regularization** and related validation techniques to select robust predictors and mitigate overfitting in a **nine-figure** internalization effort.

The Travelers Companies Jun 2023 – Aug 2023
Software Engineer Intern Hartford, CT

- Led re-architecture of internal authentication systems, migrating legacy LDAP flows to OIDC/OAuth with Spring Boot and eliminating single points of failure across internal applications.

Projects

Lumara | ***code***

- Built a deployable LLM refinement platform with a Flask API, recursive judge / critique / refine loops, per-request model routing, and Dockerized frontend-backend integration for iterative output improvement.
- Implemented robust retry / fallback handling, structured iteration scoring, and live inspection of refinement trajectories across GPT-4o and Gemini-based workflows.

Noesis | ***code***

- Built an in-browser visual imagination engine using **WebGPU**, React Three Fiber, and local WebLLM inference to translate natural language into structured scene scripts and real-time particle-scene synthesis.
- Implemented instanced-particle swarm simulation with spatial hashing, behavior integration, schema-validated scene control, and progressive local model initialization for low-latency interactive rendering.

Skills

Languages: Python, C, Rust, Java, JavaScript/TypeScript, Julia, React, LaTeX
ML / Inference: PyTorch, TensorFlow, JAX, Transformers, NeMo, Lhotse, Hugging Face, ONNX, Core ML Tools, Triton, bitsandbytes
Systems / Performance: CUDA, WebGPU, FFmpeg, Linux/Bash, distributed training, mixed precision, multithreading, asyncio, batch processing, parallel computing, Docker
Data / Infra: NumPy, Pandas, PostgreSQL, MySQL, NoSQL, Git, Weights & Biases