

Yanze (David) Wu

NYU computer science major, mathematics minor

Expected graduation: May 2027

yanzewu88@gmail.com | [Website](#) | [GitHub](#) | [LinkedIn](#)

I build and test ML infrastructure, with a focus on inference reliability, reproducible evaluation, and backend systems. I am interested in Fall 2027 PhD study in ML systems and in engineering work that turns failures into reproducible cases and focused fixes.

Current systems

CloudTune. An early ML infrastructure project for supervised fine-tuning on customer-operated infrastructure, failure recovery, and portable signed execution receipts. The [public research preview](#) includes an illustrative receipt-verification demo; the application and source remain private. Production qualification and customer validation remain open.

AgentCI Guard. An experimental TypeScript static analyzer for GitHub Actions workflows that run AI agents. It models workflow permissions and reachability and reports diagnostics with explicit analysis limits. [Source](#) / [v0.6.2](#).

Selected merged contributions

Hugging Face Transformers. Fixed half-precision `torch.compile` failures in DETR-family position embeddings and added a compiled-dtype regression test. [PR #47238](#).

marimo. Restored interpreter completion after Jedi static-analysis failures, with regression tests. [PR #10100](#).

Home Assistant. Improved [Tuya startup recovery](#), [LG ThinQ service error handling](#), and [LinkPlay discovery](#).

Selected publications

Co-author, IEEE ICC 2025. Urban Outdoor Propagation Measurements and Channel Models at 6.75 GHz FR1(C) and 16.95 GHz FR3 Upper Mid-Band Spectrum for 5G and 6G. [DOI](#) / [Open paper](#).

Co-author, IEEE ICC 2025. Upper Mid-Band Channel Measurements and Characterization at 6.75 GHz FR1(C) and 16.95 GHz FR3 in an Indoor Factory Scenario. [DOI](#) / [Open paper](#).

Prior experience

NYU WIRELESS research and teaching support; multimodal generation workflows at ArchAI; Java and Spring Boot backend engineering at Nanjing HuiJin Tech.

Research direction

How should inference changes be evaluated when latency, throughput, application output requirements, and run-to-run variation all matter? I am seeking supervised research opportunities involving controlled experiments, failure-case analysis, and reproducible evaluation.