# Yuyang Wang

Santa Clara, CA | yw8396@gmail.com | 914-830-8396 | yuyangwang.org | github.com/echo-cool | linkedin.com/in/yuyangwang-yw

## SUMMARY

Software engineer with 2+ years of professional **Python** development on production **LLM agent and inference platforms** and large-scale data systems. Shipped an **AI-powered data analysis platform** end-to-end with **vLLM** serving and a **133-case evaluation harness**, sped up billion-record analysis **5x**, and own production monitoring for customer deployments in two countries.

## EDUCATION

**Cornell University** **Aug 2023 – May 2024**
*M.Eng. in Computer Science | GPA 4.0/4.0 | Noel Croucher Graduate Fellowship* *New York, NY*

**University College Dublin (UCD) / Beijing University of Technology (BJUT)** **Sep 2019 – Jul 2023**
*Dual degree in Software Engineering: B.Eng. (BJUT), B.Sc. (UCD) | GPA 4.01/4.20* *Beijing, China*

## EXPERIENCE

**Polaris Wireless** **Jun 2024 – Present**
*Software Engineer* *Santa Clara, CA*

- Shipped an **AI-powered data analysis platform** end-to-end in **Python** (FastAPI, LangGraph, PostgreSQL, React): a streaming **LLM agent runtime** with disconnection-safe resumable runs, checkpointed state, MCP tools, JWT/RBAC access controls, human-in-the-loop approval, reusable agent skills, and an **nsjail sandbox** for model-generated code; agents run governed text-to-SQL across **five database engines** (ClickHouse, DuckDB, SQL Server, MySQL, PostgreSQL).
- Deploy and operate the platform's **self-hosted vLLM inference** (Docker, GPU): own model deployment and upgrades, tune **KV-cache** memory, batching, and context length for throughput and latency, and debug the serving path in production.
- Built a **133-case end-to-end evaluation and regression harness** (Python) that grades each agent answer in four layers (tool routing, arguments, results, rendered charts/maps) against a ClickHouse ground-truth oracle; drove accuracy to **97%** (129/133) and cut suite runtime **57%** by fixing the tool-call and grounding defects it exposed.
- Cut co-presence (meeting detection) analysis, which scores which subscribers met a target, from **2 minutes to under 30 seconds (5x)** across **billions of mobility records** by redesigning the **ClickHouse pipeline**; derived and documented its statistical confidence model.
- Built and own production **KPI monitoring** for customer deployments in **two countries** (Python/Flask, Grafana, Prometheus, Loki, Docker): **100+ KPIs** across 48 dashboards, 15 alert rules, scheduled 15-minute analysis jobs, a web command portal, and automated monthly reports.
- Developed natural-language **workflow agents** for the AIGIS product: intent classification, fuzzy resolution of target and geofence names, natural-language date parsing, parameter validation, and generation of structured API requests.
- **Technical lead** for the analysis platform: set its architecture and drove delivery with an offshore engineering team; **mentored several interns** and conducted **technical interviews for 10+ engineering candidates** across multiple teams.

**Cornell Tech / National Zoning Atlas** **Aug 2023 – Dec 2023**
*Graduate Research Assistant (Software Research Engineer)* *New York, NY*

- Built an LLM pipeline for **structured extraction and RAG** over U.S. zoning codes (OpenAI API, AWS Textract OCR) with DVC-tracked data and automated evaluation, cutting manual review time for the Atlas team.

**University College Dublin** **Sep 2020 – Mar 2023**
*Teaching Assistant, School of Computer Science* *Beijing–Dublin International College*

- Ran labs and tutorials, graded projects, and held office hours for **six undergraduate modules**: Data Structures & Algorithms, Operating Systems, Computer Networks, Object-Oriented Design, Intro to Program Construction, and Computer Graphics; mentored students on debugging, Git workflow, and code quality.

## SELECTED PROJECTS

- **Compile N Run** (compilenrun.com, 2024 – present): Built an open-source programming tutorial platform covering **70+ topics**; reached **265K active users in 2026 YTD** and **300+ GitHub stars** across its English and Chinese editions.
- **MiniTorch**: Implemented a PyTorch-like deep-learning framework with autograd, Numba/CUDA kernels, and conv layers.

## SKILLS

| | |
|---|---|
| **LLM / ML Infra** | vLLM serving & tuning, KFServing, agent runtimes, MCP, SSE streaming, text-to-SQL, RAG, evals, PyTorch |
| **Languages** | Python, Java, SQL, TypeScript/JavaScript, C/C++, Swift, Bash |
| **Data** | ClickHouse, DuckDB, PostgreSQL, MySQL, SQL Server, Redis, Hadoop/Spark, Kafka, geospatial analytics |
| **Platform** | Docker, Kubernetes, Grafana, Prometheus, Loki, Flask, FastAPI, React/Next.js, AWS, CI/CD, Linux |
| **Mobile/Embedded** | Android (Java), iOS (Swift), ESP32 firmware (C/C++) |

## PUBLICATIONS

- **Y. Wang**, Y. Hao, A. Cong. "Harnessing Machine Learning for Discerning AI-Generated Synthetic Images." *arXiv* 2401.07358, 2024.
- **Y. Wang**. "Heterogeneous Seismic Waves Pattern Recognition in Oil Exploration with Spectrum Imaging." *IEEE ICCIA*, 2022.