# Xinyang Wu

## Software Engineer - AI Infrastructure

Zürich, Switzerland | Swiss Work Permit B | +41 76 527 85 77
xinyangwuethz@gmail.com | xinyangwuethz.github.io | github.com/XinyangWuEthz

Software engineer with an ETH Zürich MSc, building production Python services and internally deployed LLM agents for scientific instruments. Hands-on with Go, Kubernetes, Slurm, PyTorch DDP and vLLM, with a focus on reliability, evaluation and cost efficiency. Motivated to support robotics research with AI infrastructure for learning from real-world data.

### Professional Experience

**Miltenyi Biotec | Software Engineer** Sep 2024 - Present

Zürich, Switzerland | Full-time since Aug 2025

- **Proposed and owned the core scientific AI assistant**, from Python/FastAPI APIs and metadata-aware RAG to streamed, source-linked React interfaces for instrument workflows.
- Built internally deployed **agentic tool-use pipelines** for interpreting completed experiments, using persisted findings, bounded read-only tools and shared query budgets to keep explanations tied to recorded results.
- Engineered reliable streaming and asynchronous services with owner-scoped sessions, per-session locks, byte-bounded SSE parsing and conditional retries that never replay partially delivered answers.
- Built deterministic **LLM evaluation** gates sharing production retrieval and prompt code; refactored duplicate extraction into a canonical corpus and integrated artifact freshness checks into CI/CD.
- Implemented **OpenTelemetry observability** for request latency, time to first token, retries, tokens and cost, with bounded metric labels; added stage-level tracing and structured logging to support failure diagnosis.
- Led a month-long internal pilot, gathering domain-expert feedback to support rollout to additional devices; aligned shared evaluation interfaces and technical decisions with engineering reviewers.
- Authored GitLab CI and Ansible configuration changes; personally deployed a tagged **Docker release to an internal instrument**, checking device availability before deployment.

### LLM Training and Evaluation Infrastructure

**Collaborative Research Project** 2026 - Present

GTC 2027 research project in progress | CSCS GPU workloads

- Post-trained a **9B LLM with LoRA adapters**, comparing DPO, GRPO and rejection-sampling SFT on a single 4×GH200 node under Slurm; used torchrun/PyTorch DDP, gradient accumulation and restart from complete checkpoints.
- Built multi-step pipelines for **automated data labeling**, quality checks and resumable sampling using vLLM, including 4-GPU tensor-parallel inference for a 122B teacher; generated **310,608 student samples in 19m46s**.
- Built exact-logprob evaluation with prompt fingerprints, adapter SHA-256, engine versions and code commits; validated vLLM scoring against Hugging Face on identical weights (**rank correlation 0.9997 for SFT**).
- Diagnosed distributed reference-cache and adapter-composition failures; repaired parent-first PEFT merging, reducing maximum logit discrepancy from **0.42 to 1e-7 in a CPU reproduction**.

## Engineering Projects

**Go Services and Kubernetes Deployment** 2026

Local infrastructure lab

- Built a **Go HTTP/gRPC health service** with goroutine fan-out, context timeouts, RWMutex-protected snapshots and Prometheus metrics; verified eight tests with the race detector and packaged a static, distroless non-root container.
- Deployed FastAPI services on a **local single-node Kubernetes/kind cluster with Helm**: probes, service networking, Ingress, ConfigMaps/Secrets, resource limits and horizontal autoscaling (HPA); verified smoke tests and exercised image-pull, crash-loop and selector failures.

**HolmesGPT Incident Investigation**

Engineering project

- Built a read-only **HolmesGPT AI-assisted debugging tool** correlating runtime logs, deployment-matched source code and internal runbooks; implemented an audited gateway with device-level authorization, query limits and evidence-linked engineering handoffs.
- Developed an **incident-replay evaluation suite** comparing root-cause accuracy, unsupported claims and investigation latency against a RAG-only baseline.

## Earlier Experience

**Bosch | Machine Learning Engineer Intern** Mar - Aug 2024

Stuttgart, Germany

- Built dataset versioning and automated data-preparation tooling on a multi-GPU Linux cluster for reproducible Python/PyTorch computer-vision workflows.
- Developed Gaussian Splatting and spatiotemporal **3D video reconstruction** with GPU-memory optimization; named first inventor on U.S. patent application 19/402,148.

**ETH Zürich Computer Vision Lab | Research Assistant** Mar - Sep 2023

- Developed a controllable Stable Diffusion/ControlNet pipeline; trained adapters on **2,500 images using one 24 GB RTX 3090**, evaluated failure modes and implemented structure-preserving output fusion.

## Education

**ETH Zürich | MSc** Sep 2022 - Jun 2025

Information Technology and Electrical Engineering

**Sun Yat-sen University | BEng** Sep 2018 - Jun 2022

Biomedical Engineering | GPA 3.8/4.0

## Technical Skills

**Software engineering:** Python, Go, SQL, TypeScript/JavaScript; FastAPI, React, HTTP/gRPC, asynchronous services; modular design, testing, refactoring, Git/GitHub.

**Infrastructure and DevOps:** Linux, Docker, Kubernetes, Helm, Slurm, Ansible, GitLab CI, GitHub Actions; cluster orchestration, workload scheduling and container deployment.

**ML infrastructure:** PyTorch, DDP, TRL, PEFT, Hugging Face, vLLM, Weights & Biases; distributed training, inference serving, data pipelines, checkpoint recovery and evaluation.

**Reliability and security:** OpenTelemetry, Prometheus; observability, structured logging, secrets management, access control, service networking, incident investigation and cost monitoring.

**Data and perception:** Video and sensor time-series data, 3D reconstruction, Gaussian Splatting, Stable Diffusion, ControlNet; NumPy and pandas.

**Languages:** English C1; German B1; Mandarin native; Japanese JLPT N1.