

GAOYI WU

Hoboken | +1(201)-283-4980 | criswu20010728@gmail.com | [LinkedIn](#) | [Github](#)

EDUCATION

Stevens Institute of Technology

Master of Science, Computer Science

- **Coursework:** Machine Learning, Deep Learning, NLP, Distributed Systems, Cloud Computing

Present

Hoboken, NJ

Shenzhen University

Bachelor, Logistics Management

Jul 2024

Shenzhen, China

TECHNICAL SKILLS

- **Backend & Engineering:** python, Go, Java, REST/FastAPI, SpringBoot, Kafka, RocketMQ, SQL, Redis
- **ML Frameworks & Tools:** PyTorch, TensorFlow, Hugging Face Transformers, LangChain, MLflow, XGBoost
- **AI & Machine Learning:** LLM Fine-tuning (SFT, LoRA), RAG, RLHF/DPO, Federated Learning, Prompt Engineering, Problem Framing
- **Cloud & DevOps:** Docker, Kubernetes, AWS (ECS, EventBridge), Jenkins (CI/CD), Hadoop, Elasticsearch

PROFESSIONAL EXPERIENCE

Intellisys Lab | *Research Assistant*

Sep 2024 - Present

- **Boosted LLM accuracy by 54%** over centralized baselines by deploying scalable fine-tuning pipelines on Kubernetes, automating edge-data ingestion and scheduled retraining.
- **Orchestrated distributed LLM training** across 100+ edge devices using TensorFlow Federated and Apache Kafka, implementing MLflow to ensure 100% experiment reproducibility and real-time regression monitoring.
- **Slashed backdoor attack success rates from 95% to 35%** on Transformer models by engineering an activation-based anomaly detection system, significantly hardening model security for offline benchmarks.
- **Elevated RAG answer quality by 15%** in federated settings by designing custom evaluation harnesses and privacy-aware metrics, effectively balancing response accuracy with strict document confidentiality.

DHL Express | *AI/ML Engineer Intern - Customer Retention ML Platform*

May 2024 - Aug 2024

- **Improved customer retention targeting precision by 30%** by developing an XGBoost churn-prediction pipeline, utilizing SHAP for model explainability and implementing robust class-imbalance handling.
- **Reduced model reporting latency by 30%** by productionizing automated ML retraining workflows on AWS ECS Fargate, integrating MLflow for continuous drift detection and model version control.
- **Enriched churn predictive signals** by fine-tuning a Transformer model for sentiment analysis (achieving 0.92 F1 score) and deploying it as a highly available Dockerized REST microservice.
- **Hardened ML system reliability** by establishing Jenkins CI/CD pipelines with blue-green deployments, and building SQL/Pandas ETLs to feed real-time self-serve analytics via Power BI.

PROJECTS

[XClaw: Desktop Interface for Open Claw](#)

Feb. 2026 - March 2026

- **Developed an autonomous AI agent orchestration platform** supporting 500+ LLMs, optimizing real-time streaming responses with rich Markdown rendering to enhance human-AI interaction.
- **Expanded agent tool-use capabilities** by integrating 50+ pre-built execution skills (including GitHub CLI, Notion API, and OpenAI image generation), enabling the LLM to successfully execute complex, multi-step real-world actions.
- **Engineered intelligent automation workflows** utilizing cron scheduling to generate customized daily AI briefings, and implemented NLP-driven automated task extraction directly from unstructured chat messages.
- **Enhanced agent accessibility and deployment** by building seamless integrations for WhatsApp, Telegram, and Discord, allowing users to interact with fine-tuned AI workflows across multiple communication channels.

Encrypted Fingerprinting for Reliable LLM Ownership Verification

May 2025 - Aug 2025

First Author; AAAI 2026 Accepted

- **Developed a production-ready post-training watermarking mechanism** for API-only LLMs using an external key-bound encoder and LoRA injection, enabling reliable black-box ownership verification.
- **Achieved a 100% Fingerprint Success Rate** across 12 diverse models, successfully withstanding 10+ adversarial attacks (including unlearning and prompt manipulation).
- **Reduced verification time by 30%** and minimized inference overhead by implementing BLEU-based similarity matching and LoRA-efficient fine-tuning techniques.

LangChain Multi-Agent Auditing & Evaluation Framework

April 2025 - May 2025

- **Accelerated financial audit workflows by 40%** by engineering a real-time multi-agent AI system using LangChain, integrating human-in-the-loop feedback to continuously refine agent behaviors.
- **Attained an 85% alignment rate with human experts** for AI-generated audit reports by developing a task-specific semantic similarity benchmark utilizing advanced BERT embeddings.
- Processed and analyzed unstructured financial data pipelines at scale using custom NLP algorithms in PyTorch.

SUMMARY

Application-focused AI/ML Engineer bridging the gap between complex models and scalable production. Expert in deploying LLMs (SFT/RAG) and distributed ML systems with a strong emphasis on system reliability, evaluation metrics, and real-world business outcomes.