Zhongyuan Wang

E wang1504626@gmail.com
linkedin.com/in/zhongyuanwang

Founding ML Scientist specializing in Agentic AI, LLM systems, and ML. 20 patents. 5 enterprise customers across the US, India, and the Gulf.

Work Experiences

2024 – Present Founding ML Scientist, RaptorX.AI, Remote

- Built a production-style Agentic Engineering workflow for bank fraud alerts: Agent Loop, context management, planning, Guardrails, case-store state, Evidence/Report agents, SFT/RLVR/RLHF, evaluation, and runtime/MLOps
- Built criteria-driven multi-agent systems for paper/patent production; as lead of a five-person ML team, owned technical direction, experiment checks, drafting, counsel handoff, and audit logs
- Built the underlying unsupervised fraud engine from zero: GNNs, behavioral autoencoders, explainability, and a 30-type edge-probability model to reduce label dependency
- Achieved 328× candidate reduction (590K → 1,883 high-confidence alerts), 99.9% fraud-type specificity
- Converted PoCs into 5 signed enterprise customers across the US, India, and the Gulf: Payactiv (US), NSDL and Lenskart (India), Omantel and Thawani Pay (Oman); US deployment covers 460K+ users across 6,700+ employers

2021 – 2024 Senior Engineer, AI & Data Platform for Autonomous Driving, Lotus Tech (Geely Group), Hangzhou, China

- Led an in-house video anonymization pipeline from vehicle data ingestion to data-center processing, supporting 80+ TB/day fleet-scale autonomous-driving data
- Trained and deployed YOLOv5-based face/license-plate detection across 10M+ frames, achieving 90%+ recall for privacy-sensitive objects
- Built a Volcano-based Kubernetes GPU job scheduling platform for PyTorch workloads on a local 64-GPU NVIDIA A100 80GB cluster (8 nodes × 8 GPUs), improving utilization and saving ∼RMB 4M/year versus Alibaba Cloud
- Optimized GPU-accelerated batch inference for fleet-scale video processing, reducing latency 30% and improving throughput
- Replaced third-party vendors with an internal ML platform, saving ∼$1M/year while supporting GDPR/PIPL compliance

2020 – 2021 Independent ML Study & Graph-ML Projects

- Self-directed study of graph machine learning: implemented GNN models for node classification and link prediction on public Kaggle datasets

2020 Software Development Engineer, Financial Anti-Fraud, DataVisor, Shanghai, China

- Production Fraud Detection for Top-Tier Commercial Bank:
  - Built an unsupervised loan application fraud model combining GNN and traditional ML, using hundreds of features spanning behavioral, device, third-party credit, and raw transaction data
  - Directly contributed to securing a $500,000 enterprise contract through technical leadership and client demonstration

2019 UCL-DiDi Master Research Program, Intern, DiDi AI Labs, Beijing, China

Supervisor: Prof. Jieping Ye (VP of DiDi Chuxing, IEEE Fellow).
Project: Leveraging Graph Neural Networks for User Travel Intent Prediction.
Selected as one of only 2 projects (out of 30+ candidates) across the entire UCL cohort for the DiDi AI Labs research program.
This work laid the foundation for applying GNNs to financial fraud detection at RaptorX.

Patents & Publications

Patents (inventor): 20 applications spanning graph-based fraud detection and LLM/agent systems, with counsel: 6 filed (US), 4 published (India), 10 in draft. Published (India) App Nos.: 202541024623, 202541045354, 202441079562, 202541032797.

Publications (first author; all under review):

- When the Tool Decides: LLM Agents Defer Blindly to GNN Tools (TMLR; arXiv:2606.14476)
- LLM Features Can Hurt GNNs: Concatenation Interference on Homophilous Graph Benchmarks (TMLR; arXiv:2606.17579)
- How Much Do Deep GNNs Actually Help? Decomposing Depth, Features, and Structure (NeurIPS 2026 E&D)

Technical Skills

Agent / ML Agentic engineering, LLM APIs, function calling/tool use, guardrails, agent evaluation, PyTorch, PyG, GNNs, autoencoders, fraud-ring detection

Data & MLOps Spark / PySpark, Kafka, ClickHouse, Redis, Parquet, Docker, Kubernetes, ArgoCD, CI/CD, FastAPI / gRPC, MLflow

Cloud AWS (S3 / EC2 / ECR), cloud GPU workflows, Linux, Grafana, Prometheus, Loki

Education

2018–2019 University College London, London, UK.
MSc in Data Science and Machine Learning.
Thesis: Graph Convolutional Networks for user behavior prediction (jointly with DiDi AI Labs).
Achieved top 3% in group NLP research project among 300 students.

2015–2018 King's College London, London, UK.
BSc in Computer Science (Artificial Intelligence).