Jiyao Pu

+44 7907 400115 | rtpujiyao@gmail.com | LinkedIn | Portfolio

Professional Summary

Machine Learning Engineer focused on building and operating machine learning systems, with a **PhD** in Machine Learning from Durham University. Experienced in designing, standardising, and deploying end-to-end ML pipelines, covering data processing, feature engineering, model training, evaluation, deployment, monitoring, and lifecycle management in cloud environments. Strong background in machine learning engineering practices, including scalable infrastructure, ML operations, and reliability-focused system design. Hands-on experience working with Python, Spark-based data pipelines, containerised deployment, and cloud-native ML platforms.

Skills

- **Machine Learning Engineering:** ML pipelines (training, evaluation, deployment, monitoring), feature engineering, model serving, lifecycle management, production ML best practices.
- **Data & Pipelines:** Python, Apache Spark, PySpark, distributed data processing, ETL/ELT pipelines, batch analytics, data validation and quality checks.
- **MLOps & Infrastructure:** Docker, Kubernetes, Git, CI/CD, MLflow (experiment tracking), model/data versioning, deployment automation, model testing, monitoring and alerting, performance, reliability optimisation, and Databricks-based workflows.
- **Cloud Platforms:** AWS, Google Cloud; Azure(fundamentals).
- **Software Engineering:** API-based services, modular system design, debugging and troubleshooting, reliability-focused engineering practices.
- **Applied ML:** NLP and LLM-based systems, evaluation-driven optimisation, accuracy/latency/cost trade-offs in real-world deployment.

Employment

- **Application Solutions Engineer** *Jul 2016 – Mar 2019*
  China Mobile Communications Group Yunnan Co., Ltd
  - **Intelligent Monitoring (Computer Vision, Edge/Cloud Deployment):** Improved and deployed production CV features (fall detection, object detection, abnormal-sound alerts, day/night adaptation) using YOLOv3 / SSD with lightweight backbones for in-home monitoring. Built and owned an end-to-end iteration loop covering data collection guidance, failure-case analysis, model training, deployment and monitoring. Reduced false alarms and improved detection accuracy from **81.4%** to **87.2%** on field-like validation data. Deployed a prototype on Alibaba Cloud for API-based testing and integrated performance logging to support ongoing monitoring.
  - **Smart TV Box (Search, Ranking, Personalisation):** Designed, deployed and refined search and personalisation services for Smart TV, including keyword retrieval, document ranking, and targeted content delivery. Built modular retrieval pipelines using GBDT with custom tokenisation and stopword handling for high-frequency terms and incorporated Word2Vec embedding features for semantic relevance. Exposed scoring models as scalable REST APIs with in-memory caching and feature stores, enabling iterative online tuning. Improved VTR by **26.3%** through systematic offline evaluation and iterative optimisation, and strengthened production reliability via performance profiling, structured debugging and logging for ongoing monitoring.
  - **Internal Training & Assessment Platform (Data Pipelines, ML Lifecycle & Deployment):** Designed, built and deployed a prototype training management and automated assessment system on **Google Cloud** with a Python backend and **PostgreSQL**. Developed scalable data pipelines to ingest, preprocess and aggregate interaction data and extract behavioural features. Trained lightweight predictive models (e.g., logistic regression / decision trees with scikit-learn) for answer correctness classification and anomaly detection, and established repeatable workflows for model retraining and version control.

Projects

- **AI Talent Bench** *Feb 2025 – Jun 2025*
  NLP, LLM, Multi-agent, Generation pipelines, AWS, Web technologies
  - **Task:** Build a scalable platform to evaluate AI candidates by automatically generating realistic, domain-specific tasks from job descriptions.
  - **Challenge:** Translating unstructured job descriptions into standardised, evaluation-ready task specifications (datasets, objectives, metrics) while ensuring reproducibility, low latency, and reliability under rapid iteration.
  - **Solution:** Designed and implemented a server-based ML system on AWS to generate structured task from job descriptions. Built a standardised, modular pipeline covering dataset selection and synthesis, metric assignment, and LLM-based task generation with schema validation. Exposed the pipeline via RESTful APIs to support integration and iterative tuning, and added logging and metadata tracking to enable debugging, traceability, and repeatable generation workflows. Optimised the pipeline for latency and robustness through input validation and reusable components.

- **Result:** Delivered a production-ready prototype that reduced task generation time from weeks to **<1 minute**. The system was externally recognised by the **UCL CDI Impact Accelerator**, with the team recommended for Cohort 7, validating its practical impact and scalability.

- **Digital Twin Dreamscape** *Jun 2024 - Dec 2024*
  Multi-agent system, Real-time ML inference pipelines, Python, Reinforcement learning, Computer Vision, Unity
  - **Task:** Build a real-time machine learning system enabling bidirectional synchronisation between a physical robotic platform and its digital twin for autonomous control and experimentation.
  - **Challenge:** Designing a reliable, low-latency ML system that integrates heterogeneous models (computer vision and reinforcement learning) while maintaining real-time synchronisation and predictable performance.
  - **Solution:** Designed a modular ML architecture separating perception, decision-making, and control. Implemented real-time vision inference using **YOLOv4** and trained **multi-agent PPO** policies for autonomous control. Built standardised pipelines for sensor ingestion, model inference, and state synchronisation between Raspberry Pi and a Unity digital twin. Added structured logging and latency profiling to optimise real-time performance and system stability, and exposed inference and control via APIs for reproducible experiments and remote operation.
  - **Result:** Achieved stable real-time synchronisation with **<200ms** end-to-end latency and **>95%** target-tracking accuracy under continuous operation. Enabled reliable multi-model inference and control in a live system, demonstrating a production-style ML workflow for real-time deployment, monitoring, and iterative optimisation.

- **EyeGaze Smart Wheelchair** *Jul 2023 - Dec 2023*
  Signal processing pipeline, OpenVINO, OpenCV, Python, Real-time inference, System integration
  - **Task:** Build a reliable real-time machine learning system to enable hands-free wheelchair navigation using eye-gaze signals, suitable for continuous operation in safety-critical scenarios.
  - **Challenge:** Converting noisy, high-frequency gaze signals into stable and low-latency control commands while ensuring robustness, predictable behaviour, and user safety under real-world conditions.
  - **Solution:** Designed a signal processing pipeline separating gaze detection, feature extraction, decision logic, and control execution. Implemented eye detection and tracking using **OpenVINO** for low-latency inference. Applied GaussianBlur and temporal filtering to stabilise gaze features under noisy input. Built a rule-augmented inference layer mapping gaze signals to control actions with configurable thresholds and safety constraints. Added structured logging and latency monitoring to profile inference performance, detect failure cases, and iteratively optimise accuracy and responsiveness.
  - **Result:** Achieved **>92%** gaze-command recognition accuracy with **<150ms** end-to-end latency under continuous operation. Delivered a stable ML system integrating real-time inference, monitoring, and iterative refinement.

Education

- **Durham University** *Mar 2021 - Jun 2025*
  ***PhD** in Machine Learning* Durham, UK

- **Newcastle University** *Sep 2019 - Sep 2020*
  ***MSc** in Computer Science* Newcastle, UK
  - Grade: **76.6%**

- **University of Electronic Science and Technology of China** *Sep 2012 - Jun 2016*
  ***BSc** in Electronic Science and Technology* Chengdu, China

Publications

C=Conference, J=Journal, T=Thesis

[J.1] Pu, J., Duan, H., Zhao, J. and Long, Y., 2023. Rules for Expectation: Learning to Generate Rules via Social Environment Modelling. IEEE Transactions on Circuits and Systems for Video Technology.

[C.1] Gao, R., Wan, F., Organisciak, D., Pu, J., Duan, H., Zhang, P., Hou, X. and Long, Y., 2023. Privacy-enhanced zero-shot learning via data-free knowledge transfer. In 2023 IEEE International Conference on Multimedia and Expo (ICME) (pp. 432-437). IEEE.

[T.1] Pu, J., 2025. Hybrid Intelligence in Evolving Games: Automated Rule Design, Strategy Evolution, and Evaluation Optimisation for Intelligent Societies. PhD thesis, Durham University.