

Piotr Gryko

Senior AI / ML Engineer — Production ML Systems · Computer Vision · Edge Inference & ML Evaluation

piotr.gryko@gmail.com · github.com/pgryko · gitlab.com/pgryko · Warsaw, Poland · Open to relocation

Physicist-turned-engineer (MSci Physics UCL, PhD Imperial College): 12+ years in engineering, 5+ in applied ML, 3+ shipping deep-learning systems that survive production — GPU inference fleets, evaluation infrastructure, data engines, and latency-bound edge computer vision. Contractor via my consultancy DevaLogic: Hypodossier is my primary engagement, alongside a counter-UAS programme for a defense-tech client fed by my own drone dataset engine. I own systems end-to-end — model serving, distributed pipelines, security hardening, and the measurement discipline (confidence intervals, cost evals, reproducible runs) that separates shipped ML from demos.

Core skills

- **ML engineering:** PyTorch, transformers, VLMs (Gemini, Qwen2-VL, SigLIP 2), YOLO/detection & tracking, diffusion models, embeddings, LayoutLM, scikit-learn; rigorous model evaluation — golden datasets, FROC/Pd/FAR curves, calibration, bootstrap and exact confidence intervals, cost-per-inference analysis; experiment tracking & feature stores (MLflow, Feast)
- **Systems & performance:** GPU inference optimization (CUDA, TensorRT), NVIDIA Jetson edge pipelines, C/C++, HPC signal processing, async/distributed architectures (RabbitMQ incl. mTLS, Hatchet durable execution), memory-leak and throughput forensics
- **Platform & infra:** Python, PostgreSQL/pgvector, Docker, Kubernetes/ArgoCD GitOps, Pulumi IaC, supply-chain security (Trivy, OpenVEX, SBOM), AWS/GCP; Django/FastAPI/React product layers

Experience

Hypodossier — R&D Software Engineer (contract via DevaLogic) | May 2023 – Present

AI document-intelligence platform for the Swiss mortgage-banking industry. ~7,000 commits across 34 services over 3 years; largest single contributor to the core product monorepo.

- Built and operated a fleet of GPU inference microservices over RabbitMQ: LayoutLM extraction, EfficientNet page classification, Qwen2-VL visual document embeddings, and page-stream segmentation running at <6.25 ms/page — with reproducible CUDA/PyTorch/TensorRT base images for the whole ML fleet.
- Sole author of the internal ML labeling and evaluation platform: golden-dataset lifecycle with versioned snapshot releases, per-class F1 and calibration dashboards, KNN mislabel detection over pgvector embeddings — then used it to drive a shadow-mode → primary rollout of a new production page classifier with divergence logging and vote aggregation.
- Delivered measured performance wins: redesigned a vision-language embedding pipeline for a 25% speedup by decoupling preprocessing from GPU inference; diagnosed production memory leaks and RabbitMQ channel-death stalls with A/B soak-test harnesses; added cost-control guards preventing silent commercial-OCR billing on cache misses.
- Led the org-wide container-security program: standardized Trivy + pip-audit + OpenVEX CI gates and runtime hardening across 15 image-building services, driving scans to zero active findings; rolled out fleet-wide RabbitMQ mTLS with fail-closed design and over-the-wire integration tests.
- Built supporting services end-to-end: a national building-registry search service (3.3M buildings, Meilisearch, daily immutable rebuilds), multi-engine antivirus scanning, PDF/A conversion, and bank-tenant integration APIs on the core Django/React product.

DevaLogic — Founder & Principal ML Engineer | Jan 2023 – Present

My consultancy — the contracting vehicle for the Hypodossier engagement above, plus the following client and product work:

Counter-UAS Computer Vision — Software & AI Lead (client engagement) | Jun 2026 – Present

Ongoing R&D: real-time detection of very small (sub-pixel-to-few-pixel) aerial targets in compressed video on Jetson-class edge hardware — v1 detector in development, evaluation harness built first.

- Architected a classical-CV detection pipeline (video-only stabilization, CFAR-style adaptive thresholding, velocity-hypothesis track-before-detect) under a hard end-to-end latency budget; authored the technical research plan from the small-target detection literature.
- Built the evaluation harness before the detector: contract-driven evaluator (135 tests, golden fixtures) producing probability-of-detection, FROC, and false-alarm-rate metrics with proper confidence intervals, plus hashed-config run logs making every published number reconstructible.

Drone Detection Dataset Engine — own product, sole engineer | Jul 2025 – Present

Designed, built, and operate solo a production ML data engine converting publicly posted video into annotated computer-vision training datasets — running unattended in production.

- Automated pipeline: multi-source ingestion (220k+ videos, 1.5 TB) → vision-language-model batch triage (118k+ analyses) with compilation detection and frame-accurate cutting → embedding-based keyframe selection → segmentation-model-assisted annotation in CVAT → YOLO/COCO/VOC export.
- Built a model-migration evaluation harness (400-clip labelled fixture, bootstrap confidence intervals, measured cost) that disproved a “cost-neutral” migration assumption by 2.6× and caught an accuracy collapse in AI-generated-content detection (0.875 → 0.375) before it contaminated training data.
- Production operations as a one-person team: durable workflow execution across CPU/GPU/ingest workers, infrastructure-as-code deployment with zero public ports, tiered backup verification with weekly restore drills, 2,000+ tests in CI.
- Shipped the commercial layer: per-client rate cards over seven billing metrics, freeze-at-ship invoice snapshots, source-stripped client dataset bundles.

Earlier consulting projects | Jan 2023 – Nov 2025

- Document anonymization with diffusion models (VAE/UNet inpainting per the DiffUTE paper) — presented at EuroPython 2025 and PyCon Lithuania 2025.
- Property-suitability ranking model for a UK energy-retrofit client (2024, two-person team): EPC open data joined to survey outcomes for 13,000 properties, 173 engineered features incl. NLP over free-text building descriptions; Feast feature store + MLflow experiment tracking, SMOTE, Optuna-tuned XGBoost — F1 0.78.
- Full-stack RAG system with self-hosted LLMs (async Django REST API with streaming responses, React/TypeScript frontend, vector DB).

Qogita — Senior Software Engineer | Mar 2022 – May 2023

Helped scale the engineering behind a B2B marketplace doing ~\$20M/month: payment-system migration (Square→Wise) with zero outages, two-way Snowflake–Django data sync, high-throughput ingestion pipelines.

Unipart Digital — R&D Software Engineer | Oct 2015 – Jan 2022

Multidisciplinary R&D for a large logistics group: classical ML (demand forecasting, telemetry anomaly detection; scikit-learn, pandas), web systems (Django/React), devops (Docker, Ansible, Elastic), embedded C++ (Arduino, ESP32); led teams of 3–4 engineers.

ION Geophysical — R&D Software Engineer | Jun 2013 – Oct 2015

Signal-processing tools for seismic imaging on high-performance computing clusters (C++); integrated Python/NumPy into the C++ processing system.

Education

- **Imperial College London** — PhD, Self-Assembling Biomaterials / Nanotechnology, 2007–2012
- **University College London** — MSci Physics, First Class Honours, 2003–2007

Talks & Community

- “Anonymization of Sensitive Information in Financial Documents” — EuroPython 2025, PyCon Lithuania 2025, Data Science Summit ML 2025
- “Building and Scaling an AI Startup with Async Django”; “Scaling, Refactoring and Fixing a Django MVP for Production” — PyCon Lithuania 2024
- Founder of Code & Coffee Warsaw tech mentoring meetup