Oliver Grainge

LinkedIn: linkedin.com/in/oliver-grainge
GitHub: github.com/olivergrainge
Website: olivergrainge.github.io
Email: oliver@grainge.me
Mobile: +44 7481 881 153

Professional Summary

- AI Engineer specializing in production ML optimization and deployment. Expert in taking models from research to production with measurable performance improvements across edge and cloud platforms. Proven track record of delivering significant latency and memory reductions through hardware-software co-design, model compression, and custom kernel development. Published 4 papers in top-tier ML venues (IEEE RAL, AAAI) on efficient deep learning systems.

Experience

- **Arm** Remote
  Contract Researcher – Performance Engineering *Feb 2025 – Present*
  - **Developer Education Initiative**: Architecting comprehensive performance engineering curriculum with 6 hands-on tutorials for Arm Total Performance toolkit, enabling developers to optimize C++ workloads on AWS Graviton instances.
  - **Cross-Platform Optimization Demonstrations**: Designing tutorials on top-down performance analysis, memory optimization, library acceleration (Arm Performance Libraries, KleidiAI), and automated porting techniques for cloud and edge deployment.
- **University College London** London, UK
  Research Assistant *Jun 2025 – Nov 2025*
  - **Extreme Low-Bit Model Compression**: Engineered 1.58-bit precision pipeline for Stable Diffusion, achieving 4× memory reduction and 95% quality retention while enabling deployment on consumer GPUs.
  - **High-Performance Kernel Development**: Designed custom CUDA and Triton kernels for bit-packed tensor operations, delivering 30% speedup over PyTorch baseline with full integration into research codebase for reproducible benchmarking.
- **Queensland University of Technology** Brisbane, Australia
  Visiting Researcher *Aug 2024 – Jan 2025*
  - **Real-Time Robotic Vision Systems**: Engineered speculative decoding for vision-language transformers achieving 2.5× inference speedup, enabling sub-100ms latency for robotic navigation applications.
  - **Training Data Efficiency**: Implemented filtering methods for deep metric learning, achieving equivalent localization accuracy with 38% less training data through intelligent sample selection.
- **Arm** Remote
  Contract Researcher – AI Inference Optimization *Nov 2024 – Jan 2025*
  - **Multi-Platform AI Deployment**: Engineered educational demonstrations of AI optimization for Arm Cortex-A (edge) and Neoverse (cloud) platforms. Achieved 40% latency reduction via SIMD/INT8 on mobile devices and 2.1× throughput on cloud instances.
  - **Advanced Mixed-Precision Framework**: Built Hyperopt-based per-layer precision optimizer enabling 2–8-bit configuration search. Demonstrated 22% memory reduction over uniform approaches on GPT models for Arm developer documentation.
- **AI Security Institute** Remote
  Research Fellow *Jan 2024 – Jan 2025*
  - **Large-Scale VLM Evaluation Infrastructure**: Built automated benchmarking framework evaluating 25+ vision-language models across 26k geo-tagged images. Unified API supporting GPT-4V, Claude 3, Gemini, and 15 open-source models achieved 99.9% reliability over 500k+ API calls, reducing evaluation time by 3×.
  - **Privacy Research & Interactive Tools**: Developed privacy-preserving techniques reducing geolocation accuracy by 40%. Launched Gradio-based interactive benchmark on Hugging Face Spaces attracting 5k+ users for human-AI capability comparison research.

Technical Skills

- **Languages & Systems**: Python, C/C++, CUDA, Triton, SQL, CMake, Bash
- **ML Frameworks**: PyTorch, PyTorch Lightning, TensorFlow, ONNX, TensorRT, Hugging Face Transformers, Scikit-Learn
- **Infrastructure & DevOps**: AWS EC2/Graviton, Google Cloud, Docker, SLURM, Git, CI/CD, MLflow, Weights & Biases
- **Specializations**: Model compression (INT8/INT4/ternary/binary), custom CUDA kernels, LLM deployment & fine-tuning, diffusion models, edge AI optimization, performance profiling (perf, NVIDIA Nsight), distributed training, RAG systems
- **Tools & Libraries**: FastAPI, ROS, Llama.cpp, OpenCV, Arm Performance Libraries, pytest, pandas, NumPy

Open Source Projects

- **BitOps – Ternary Matrix Multiplication Library (C++, CUDA, ARM NEON)**: High-performance inference library with multi-backend support (ARM NEON, x86 AVX2, CUDA). Delivers 16× memory reduction via 2-bit weight packing with architecture-specific kernels outperforming PyTorch float32 on edge devices.
- **BitCore – Quantization-Aware Training Toolkit (PyTorch)**: Production-ready framework for training ternary neural networks with drop-in `BitLinear` layers supporting BitNet, TWN, and ParetoQ 1.58-bit schemes. Seamless train-to-deploy mode switching with optional BitOps acceleration, enabling 8× memory savings during inference.
- **BitNet Chat Interface – Interactive LLM Demo (Python, Gradio, CUDA)**: Gradio-based web application demonstrating real-time chat with 1.58-bit BitNet models. Achieved 24× inference speedup (12 vs 0.5 tokens/sec) and 80% memory reduction on ARM M4 using BitOps backend vs PyTorch FP32. Features backend switching, streaming responses, and real-time performance monitoring across 2B parameter models.
- **VSLAM – Stereo Visual SLAM Pipeline (Python, OpenCV, NumPy)**: Production-ready implementation of stereo visual simultaneous localization and mapping in pure Python. KITTI dataset-compatible system featuring feature detection/tracking, stereo matching, motion estimation via PnP/ICP, and bundle adjustment optimization. Comprehensive pytest suite with demos illustrating matching, tracking, and pose estimation components.

Education

- **University of Southampton** — Southampton, UK
  PhD (iPhD) in Machine Intelligence – Thesis: Efficient Resource-Constrained Visual Place Recognition — *Oct 2022 – Current*
- **University of Southampton** — Southampton, UK
  BEng Electronics and Electrical Engineering – First Class Honours (83%) — *Sept 2019 – Jul 2022*

Publications

- **Assessing the Geolocation Capabilities, Limitations and Societal Risks of Generative Vision-Language Models** — AAAI 2025
  First comprehensive benchmark of VLM geolocation capabilities across 4 datasets, revealing privacy risks and capability limitations. — Feb 2025
- **TeTRA-VPR: A Ternary Transformer for Compact Visual Place Recognition** — IEEE RAL
  Novel two-stage compression pipeline reducing memory by 65% and latency by 35% for visual place recognition transformers. — Mar 2025
- **Design Space Exploration of Low-Bit Quantized Visual Place Recognition** — IEEE RAL
  Systematic study establishing deployment guidelines for extreme quantization on embedded devices with 63% latency and 95% memory savings. — Jun 2024
- **Structured Pruning for Efficient Visual Place Recognition** — IEEE RAL
  Channel pruning methodology achieving 21% latency and 16% memory reduction with ¡1% accuracy loss on embedding models. — Aug 2024