

Tom Pollak

Bristol, UK

tompollak1000@gmail.com

github.com/tom-pollak

tom-pollak.github.io

(+44) 77400 54268

EXPERIENCE

Graphcore

Machine Learning Engineer – Applied AI

April 2025 – Present

Bristol, UK

- At Graphcore we build next-gen accelerators, and develop the ecosystem for more heterogeneous compute.
- Writing Triton kernels targeting our hardware. Authored fused performant kernels for MoE, Flash Attention, RoPe. Performance profiles help inform Triton compiler and PyTorch teams.
- Presented workshop paper on Bayesian inference: [Variational Entropy Search is Just 1D Regression](#) at NeurIPS.
- Helping develop pre-training infrastructure, working on load-balancing large MoE models.
- Contributed to PyTorch: PyTorch PP deadlock bug when using Gloo ([#152938](#)), fix SDPA MATH backend reference implementation: ([#163508](#)).

Cisco Meraki

Machine Learning Engineer – Camera Intelligence Team

June 2023 – April 2025

London / Remote, UK

- At Cisco, I focused on building cross-camera tracking over the 2 years, which was just [released in Beta](#).
- Technical lead of a team of 6 engineers personally managing firmware, model training, inference optimization and architecture; product presented at Cisco Live 2025.
- Designed and implemented firmware for high-performance C++ inference engine and scalable distributed k-NN search system across mesh network of cameras (10K+ LOC).
- This enabled real-time search & retrieval that scales to thousands of devices per network with no hit to the backend.
- Created multimodal dataset (>200K objects with a mix of synthetic and human labelled annotations) and fine-tuned CLIP-based models for zero-shot object retrieval.

University of York

BEng. Computer Science – First Class with Honours

June 2023

PROJECTS

On-Policy quantization

December 2025

GPUMODE NVFP4 GEMM Competiton

November 2025

- Fastest Triton implementation in the first two competitions, third ongoing. [Annotated NVFP4 GEMM](#).

Parscale Cross-attention

August 2025

- Extension to PARSCALE enabling data-dependent communication between parallel replicas via cross-attention.

Diffusion LLM replication

July 2025

Blender Copilot

January 2025

- Blender plugin for generating 3D meshes from text prompts. Modal for inference, FastAPI / FastHTML backend.

Structured Generation for LLMs with RLVR

March 2025

- Structured generation and tool use with auto-generated GBNF grammars and Pydantic validation for RLVR.

Interpretability Research

Aug 2024 – Jan 2025

- Trained SAEs on ARC-AGI like puzzles. Contributed to SAELens activation caching (#321, #367).

Claudette Pydantic

July 2024

- Extended Claudette with structured outputs via tool use.

SKILLS

Languages

Python, C++.

ML

PyTorch, Triton, TorchTitan, Faiss, Slurm, Kubernetes.