

# 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*

- Developed high-performance Triton kernels (MXFP4 MoE, Flash Attention, RoPE) for our custom accelerator.
- Develop pre-training infrastructure for large MoE models focusing on load-balancing.
- Presented workshop paper at NeurIPS: [Variational Entropy Search is Just 1D Regression](#).
- Contributed to PyTorch: Pipeline parallelism deadlock fix with Gloo ([#152938](#)), SDPA MATH backend reference implementation fix ([#163508](#)).

### Cisco Meraki

*Machine Learning Engineer – Camera Intelligence Team*

June 2023 – April 2025  
*London / Remote, UK*

- Led cross-camera tracking from idea to Beta release; presented at Cisco Live 2025.
- Technical lead of 6 engineers across firmware, model training, inference optimization, system architecture.
- Designed high-performance C++ inference engine and distributed k-NN search across camera mesh networks, enabling real-time retrieval across thousands of cameras with zero backend load.
- Built multimodal dataset (200K+ objects, synthetic and human-labelled) 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

- Researching LLM quantization via on-policy distillation. FP32 teacher guides MXFP4 student on its own generations rather than static datasets.

### GPUMODE NVFP4 GEMM Competition

November 2025

- Fastest GEMM and GEMV Triton submission, with 4x speedup over baseline. [Annotated NVFP4 GEMM](#).

### Parscale Cross-Attention

August 2025

- Extension to Bytedance's [PARSCALE](#) enabling data-dependent communication between parallel replicas via cross-attention.

### Nano Diffusion LLM

July 2025

- "Nano" training and inference script for diffusion language models that generate text via iterative denoising. Implements [masked diffusion](#) (LLaDA-style) and [Duo](#) (flow-based, self-correcting).

### Blender Copilot

January 2025

- Blender plugin for generating 3D meshes from text prompts. Modal for inference, FastAPI / FastHTML backend.
- [Demo](#).

### xverify – GBNF structured generation

March 2025

- Auto-generated GBNF grammars from Pydantic models. Integrates with RLVR for tool use / structured outputs.

### Algorithmic Trading – Horse Racing

September 2020 – July 2021

- Statistical arbitrage on mispriced "each-way" and "extra-place" bets with three-way Kelly Criterion sizing based on conditional place probabilities. Profitable but low volume until banned.

## SKILLS

### Languages

Python, C++, Triton, CUDA.

### ML

PyTorch, TorchTitan, Faiss, Slurm, Kubernetes, wandb.