Tom Pollak

Bristol, UK
tompollak1000@gmail.com

github.com/tom-pollak
tom-pollak.github.io
(+44) 77400 54268

EXPERIENCE

Graphcore April 2025 – Present
Machine Learning Engineer – Applied AI *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 implementenation: (#163508).

Cisco Meraki June 2023 – April 2025
Machine Learning Engineer – Camera Intelligence Team *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 June 2023
BEng. Computer Science – First Class with Honours

PROJECTS

NVFP4 Triton Kernels March 2025
https://github.com/tom-pollak/xverify

- test

Minimal reproduction of Diffusion LLMs March 2025
https://github.com/tom-pollak/dllm/

- foo

Blender Copilot March 2025
https://github.com/tom-pollak/blender-copilot

- test

Structured Generation for LLMs with RLVR March 2025
https://github.com/tom-pollak/xverify

- Developing a library for structured generation and tool use using automatically generated GBNF grammars and Pydantic schema validation for RLVR.

Interpretability Research August 2024 – January 2025
https://github.com/tom-pollak/interpretability-culture

- Investigating features in neural networks trained on ARC-AGI-style 2D grid puzzles
- Trained sparse autoencoders (SAEs), discovering task-specific feature in the models, ablating would degrade performance in a specific task.
- Applying Anthropic's Crosscoders to understand how a model changes throughout training.
- Contributed to the SAELens library: Optimized activation caching with HuggingFace datasets. (PRs #321, #367)

Claudette Pydantic

July 2024

https://github.com/tom-pollak/claudette-pydantic

- Extended the Claudette library with structured outputs via tool use – Example.

NLP Image Retrieval with CLIP & Faiss

September 2022 – June 2023

https://tom-pollak.github.io/clip-index

Algorithmic Trading System – Horse Racing

December 2020 – July 2021

https://github.com/tom-pollak/each-way-matcher

- Developed statistical arbitrage system identifying mispriced ”each-way” bets.
- 3-way Kelly Criterion strategy for optimal stake sizing based on calculated conditional place probabilities.
- Successful with high ROI, but low volume and I got banned from profitable bookmakers.

SKILLS

Languages	Python, C++.
ML	PyTorch, Triton, TorchTitan, Faiss, Slurm, Kubernetes.