Tom Pollak

Bristol, UK
tompollak1000@gmail.com
github.com/tom-pollak
tom-pollak.github.io
(+44) 77400 54268

EXPERIENCE & EDUCATION

Graphcore April 2025 – Present
Machine Learning Engineer – Applied AI *Bristol, UK*

- At Graphcore we are focused on building next-generation AI accelerators, and developing the open-source ecosystem to allow for more heterogeneous compute.
- Contributing to open-source ML infrastructure: Fixed a PyTorch pipeline parallelism deadlock bug when using Gloo that affected distributed training (#152938).
- Developing mixed-precision pre-training infrastructure. I'm currently working on an experiment targeting a 1B model with up to 1T tokens.

Cisco June 2023 – April 2025
Machine Learning Engineer – Camera Intelligence Team *London / Remote, UK*

- Developed computer vision models and pipeline, serving over 4 million networks globally.
- Designed and implemented high-performance C++ inference engine and firmware for edge devices (10K+ LOC).
- Built distributed k-NN search system across mesh network of cameras, enabling real-time search & retrieval that scales to thousands of devices per network with no hit to the backend.
- Technical lead of a team of 6 engineers managing firmware, model training, inference optimization, and architecture; product featured at Cisco Live 2025.
- 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

Structured Generation for LLMs 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" betting opportunities
- Implemented adapted 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++, Cuda C, Mojo (learning).
ML	PyTorch, Triton, Slurm, TorchTitan, vLLM, Modular MAX, Faiss, Huggingface.