Shawn Presser

(314) 324-1248

shawnpresser@gmail.com
github.com/shawwn

Skills

ML/AI:

Transformer architectures; GANs; diffusion models; VAEs; LLM pretraining & fine-tuning; dataset curation; self-supervised learning; scaling on consumer hardware

Frameworks:

PyTorch, TensorFlow, JAX, Hugging Face Transformers, Weights & Biases

Languages:

Python, C++, Haskell, JavaScript, TypeScript, Bash, Common Lisp, Arc

Infrastructure:

TPU/GPU cluster management; distributed training; Docker

Press Coverage

WIRED Magazine: The Battle Over Books3 Could Change AI Forever

Papers

***The Pile**: An 800GB Dataset of Diverse Text for Language Modeling (arXiv:2101.00027v1)*

Experience

Founding Engineer & ML Researcher 2022 – 2023
Keen Technologies

- Second employee at John Carmack's AGI-focused independent research lab, joining at inception to build research infrastructure from scratch.
- Built scalable training pipelines, experiment orchestration & evaluation tooling; contributed to architectural experiments targeting efficient general intelligence.
- Deep involvement in technical planning, research review, and developer workflows; built-internal tools that accelerated iteration and improved reproducibility.

ML Engineer & Software Engineer 2021 – 2022
Groq

- Designed and optimized AI inference pipelines on Groq's custom LPUs, achieving sub-millisecond latency for LLM-serving.
- Enabled deployment of models such as GPT-J with real-time performance in cloud and on-premises environments.
- Contributed to scaling GroqCloud by integrating model-serving APIs and developing benchmarks comparing LPUs vs. GPUs.
- Collaborated across hardware, compiler, and DevOps teams on v1 LPU documentation and next-gen ASIC planning (v2 on 4nm Samsung process).

ML Researcher & Founder 2019 – 2022
Tensorfork

- Founded the first open-source AI research community, growing it to over 1000 members and coordinating collaborative research efforts across the group.
- Compiled and released foundational training dataset **used in the creation of LLaMA**, **GPT-J**, **BloombergGPT** & other major LLMs.
- Worked directly with Google's TPU Research Cloud to benchmark large TPU pods for swarm training; built TPUnicorn and a real-time dashboard for sharing and monitoring TPU resources.
- Pioneered techniques for running GPT-scale models on consumer hardware, enabling grassroots AI research outside well-funded labs.

Information Security Engineer 2014 – 2016
Matasano Security

- Completed 50+ penetration tests for well-known clients.
- Identified and documented security deficiencies across a wide range of environments using network monitoring and exploitation tooling.

Software Engineer, Market Data Operations 2012 – 2013
Thomson Reuters

- Redesigned a market data routing system in Python and Redis, increasing throughput from ~50 to 1,000+ messages/sec (20x improvement).

Software Engineer, Market Data Operations 2011 – 2012
Scottrade

- Built a concurrent lock-free shared memory ring buffer achieving ~30ns min latency and 11M+ messages/sec throughput on production hardware

Game Programmer — Heroes of Newerth 2010 – 2011
S2 Games

- Reworked memory allocation code in the entire codebase, trapping almost every memory allocation (including STL string and other STL container allocations).
- Implemented various core engine and gameplay improvements

Graphics Programmer — HeroEngine 2005 – 2008
Simutronics

- Implemented a seamless day/night cycle, complete with sunrise, sunset, a starry nighttime skyscape, and procedural cloud coverage, alongside novel weather system.