Harold Benoit *(email, website, blog, twitter, scholar)*

WORK EXPERIENCE

Liquid AI
July 2024 - Present

Member of Technical Staff - Research Scientist/Engineer — *Remote*

- *Inference co-lead*
  - Hardware-aware model design and inference - one of the main contributors to the released LFM2 models, which have ~ 50% faster decode and prefill latency, compared to other 1B models like Llama-3.2-1B.
  - Implementation and optimization of our models in PyTorch, ExecuTorch & C++.
  - Hiring manager for the research engineer roles.
- *Quantization lead*
  - Quantized inference support - Implemented quantized CPU Mixture Of Experts (MoE) kernel 2 times faster than the next fastest open-source alternative.
  - Currently leading the quantization effort at Liquid, setting up the research agenda , taking care of the evaluations, and implementation of the best PTQ and QAT schemes. Responsible for the quantization of models delivered to customers.
- *Misc. research & engineering work*
  - Distributed training work (e.g. 15% throughput increase with FP8 training, lots of debugging).
  - Synthetic data pipeline for coding data (SFT & reasoning). Support for multi-language sandboxed execution, unit-test driven feedback & revision.

Swiss Federal Institute of Technology (EPFL)
October 2023 - June 2024

Research Associate — *Lausanne, Switzerland*

- Research and engineering of LLMs (distributed training optimization, evals) on the Swiss AI 10'000 GPUs cluster.
- Worked on multimodal conditional image diffusion models @ VILAB. We proposed a method to generate tailored synthetic training data**,** by introducing two feedback mechanisms to guide the generation: 1) model-based and 2) target domain-based.
- The result is the **TMLR 2025 paper** *"Controlled Training Data Generation with Diffusion Models"*

IBM Research
February 2023 - September 2023

Research Intern — *Zurich, Switzerland*

- **ICLR 2024 first-author paper** - "*Unraveling the Key Components of OOD Generalization via Diversification*"
- Research on synthetic data & algorithmic design for out-of-distribution (OOD) generalization of neural networks

G-Research
July 2022 - September 2022

Quantitative Research Intern — *London, UK*

- Classic quant work. Predict things in financial markets. *Skills*: Scala, Spark, Python, ML, Algorithmic Design

PUBLICATIONS

Controlled Training Data Generation with Diffusion Models — **TMLR, 2025**

T. Yeo*, A. Atanov*, **H. Benoit^**, Aleksandr Alekseev^, Ruchira Ray, Pooya Akhoondi, Amir Zamir

Unraveling the Key Components of OOD Generalization via Diversification — **ICLR, 2024**

H. Benoit*, L. Jiang*, A. Atanov*, O. Kar, M. Rigotti, A. Zamir

SOFTWARE & PROJECTS

- **1st place at Huawei LLM training hackathon** - trained the best performing language model on Slimpajama with 3 hours of compute on a A100 GPU. I won $1.5k and a trip to Huawei HQ in China. (Github)
- **Diffusion Model for Protein Structure Modeling** - trained a protein diffusion model from scratch. (Github)
- I have a blog, where I write down almost everything I learn. I also have a Twitter, where I post semi-often.

EDUCATION

Swiss Federal Institute of Technology (EPFL) **2021 - 2023**

MSc of Data Science / Computer Science, 5.8/6.0 GPA (ranked 3rd in year) *Lausanne, Switzerland*

Skills: Deep Learning, Machine Learning, FSDP, Python, PyTorch, C++, Go, SQL, Git, Linux, Docker