Sam Havens

Machine Learning Researcher, Engineer, & Manager | Portland, Oregon | linkedin.com/in/samhavens | samhavens@gmail.com | 818.590.0484

Experience across the stack, specializing in NLP research, engineering, and management.

WORK EXPERIENCE

Databricks, Staff Research Scientist

Jul 2023 - Current

- Led development of DBRX-Instruct, recognized as the leading open-source model upon release.
- Implemented advanced post-training techniques that improved scores on benchmarks including programming, mathematics, and tool-use.
- Guided multiple research projects in RLHF, synthetic data generation, RAG, and PEFT.

MosaicML, Research Scientist

Sep 2022 - Jul 2023

- Led the development of chat/instruction-tuned variants of MPT-7B and MPT-30B, enhancing usability for various downstream applications.
- MosaicBERT: Developed a BERT-style encoder architecture and training recipe optimized for fast pretraining. Incorporated FlashAttention, Attention with Linear Biases (ALiBi), Gated Linear Units (GLU), dynamic padding removal, and low precision LayerNorm.
- LIMIT: Investigated the impact of small, high-quality instruction fine-tuning datasets on Large Language Models. Demonstrated that subsets of 1k-6k samples were sufficient for performance on both NLP benchmarks and model-based evaluation.

Writer, Director of NLP Engineering

Sep 2020 - Sep 2022

Writer is an AI writing assistant used by brands like Twitter, Intuit, and Accenture. The NLP team at the time used a microservice architecture based on Kubernetes, FastAPI, HuggingFace Transformers, NVIDIA Triton, and ONNX.

- Responsible for NLP from research to operations, including >25 microservices
- Trained an encoder/decoder Grammar Error Correction model using novel synthetic data techniques, which outperformed an open-source baseline by 130%
- Used NVIDIA Triton and ONNX to serve a character-based transformer spelling correction model while keeping inference latencies below 300ms at 50 req/s

Qordoba, Director of Data Science

Feb 2019 - Sep 2020

- Reduced mean service latency from >1.5s to <300ms.
- Grew team from 2 to 6, while improving onboarding effectiveness (time to first commit: from weeks to < 1 day).
- Implemented classification and seq2seq models in spaCy, Flair, and Marian with aggressive latency requirements.
- Responsible for all ML Ops. Made models available using modern async/await Python, and Docker/Kubernetes/PubSub, with some help from Bash and Jenkins.

Carlabs, Chief Technology Officer

May 2016 - Jan 2019

- Created a suite of tools for automotive OEMs and dealers to manage chatbots for their brand on web, chat, and voice platforms.
- Used Docker/Kubernetes to operate services written in Node.js, Elixir, and Python, and models made in FastText and TensorFlow.
- Engineering team grew from 4 to 16 during my tenure. Established a strong engineering culture of testing, code reviews, pair programming, and mentorship.

Carlabs, Software Engineer

Mar 2015 - May 2016

- Created a car comparison shopping tool using React, ES6, Webpack, and MaterialUI.
- Built a conversational agent with a Node.js/Docker backend and NLP services in Python using FastText, NLTK, and Gensim.

Topanga Mountain School, Board Member and Teacher

Aug 2007 - Current

- Full-time math and science teacher from 2007-2014.
- Current board member.

Revmaker, Software Engineer

Apr 2014 - Mar 2015

- Implemented a probabilistic lead-scoring model and built an MVC lead generation web app with a LAMP backend and d3.js + jQuery frontend

EDUCATION

California State University, Northridge, Master of Science Mathematics

2011 - 2013

- Bianchi Outstanding Graduate Student research award, and IRIS Fellowship recipient.

UC Santa Barbara, Bachelors Physics

2002 - 2006

- LEAPS Fellow