JOE DAVISON

(385) 338-0930 · joeddav@gmail.com
LinkedIn · GitHub · Hugging Face · Google Scholar

SUMMARY

Machine learning researcher and engineer with a decade of experience training models, running research, and building production LLM systems. Work spans early language-model prompting research, widely-used Hugging Face Transformers and Datasets tooling, distributed training infrastructure for generative Transformers, and harnesses & evaluation for agentic systems. Recent independent work keeps a direct line to post-training questions, including RLVR-style verifiable feedback and synthetic data generation.

Focus areas: research tooling & infrastructure · RL/post-training · trace-based evaluation · durable orchestration · context engineering · code generation · distributed systems · latent space modeling

EXPERIENCE

BambooHR
Draper, UT · Sep. 2024 – Present

Staff AI/ML Engineer · Senior Data Scientist

- Designed and built shared infrastructure for production AI agents, including durable runtimes, persistent event streaming, tool-execution tracing, observability, eval workflows, and reliability patterns used across BambooHR's AI platform.
- Built evaluation harness and failure-analysis loop for Ask BambooHR, turning ambiguous model/tool behavior into measurable regressions, debugging workflows, and eval-gated release criteria.
- Re-architected agentic orchestration and tool-use flows to reduce median response time by 60% while improving performance on internal evaluation suites.
- Led major release of Ask BambooHR as AI Platform tech lead, coordinating product, platform, and model behavior work while expanding model/tool coverage and analytical capabilities.

Enveda Biosciences
Lehi, UT · Sep. 2021 – Dec. 2022

Senior Machine Learning Scientist

- Built internal ML research and training infrastructure on PyTorch DDP and Hugging Face Transformers, enabling reproducible distributed training workflows for generative Transformer research.
- Trained multi-GPU Transformer models for noisy sequence-to-structure prediction from MS/MS spectra, improving state-of-the-art accuracy by 20% through custom tokenization, model iteration, and training pipeline design.
- Established practical workflows for experiment tracking, data preparation, model iteration, and distributed training across a growing ML research organization.
- Owned research roadmap and mentored 6 interns and 2 FTEs while partnering with chemists, biologists, and data scientists to convert ambiguous scientific needs into reliable training and evaluation systems.

Hugging Face
Brooklyn, NY · Feb. 2020 – Jul. 2021

Research Engineer, Science Team

- Built and released open-source Transformers and Datasets tooling used by the ML research community, contributing to two EMNLP Best Demo Award-winning libraries.
- Implemented state-of-the-art zero-shot classification tools across the Transformers library and Hugging Face API, enabling classification in 100 languages without supervised task-specific training.
- Developed an unsupervised task-adaptation and self-training workflow that delivered up to 100x latency reduction for zero-shot classification, combining algorithmic changes with systems-oriented performance improvements.
- Trained and released 4 public fine-tuned models with more than 15M combined downloads on the Hugging Face Hub.

IBM Research / MIT-IBM Watson AI Lab
Cambridge, MA · Jun. 2019 – Sep. 2019

Research Intern

- Created a novel variational autoencoder variant for isolating latent structure across differing data distributions, implemented in TensorFlow Probability and published at the NeurIPS Bayesian Deep Learning Workshop.
- Ran 200+ model training jobs on Slurm-managed research compute to validate model behavior, debug failures, and iterate quickly under constrained compute budgets.

EARLIER ML AND SOFTWARE ENGINEERING EXPERIENCE

Pluralsight · Data Scientist Intern — Boston, MA · May 2018 – Aug. 2018
Improved search and recommendation systems by developing a universal embedding model across 5 educational content formats; prototyped evaluations for Doc2Vec, FastText, LDA, and TF-IDF.

Zeff · Machine Learning Engineer, Part-time — Orem, UT · Sep. 2017 – Mar. 2018
Developed a FaceNet-style triplet embedding model trained on a proprietary dataset of more than 30M unique face images; created an open-source toolkit for intelligent indicator-variable creation.

Microsoft · Software Engineer Intern, ML — Cambridge, MA · May – Aug. 2017
Expanded CNTK by creating R bindings to the Python interface; released MMLSpark components for scalable natural-language processing through distributed Spark pipelines.

Qualtrics · Software Engineer, Part-time — Provo, UT · Aug. 2016 – May 2017
Developed a Redis-backed microservice for logging and maintenance of an internal export utility.

Instructure · Software Engineer Intern — Salt Lake City, UT · May – Aug. 2016
Converted the Pages module in the Canvas app from Objective-C to Swift while updating the interface with modern styling.

SELECTED PUBLICATIONS

- **Joe Davison**, Joshua Feldman, and Alexander Rush. **Commonsense Knowledge Mining from Pretrained Models**. EMNLP Oral, 2019. 420+ citations.
  Demonstrated how pretrained language models encode extractable commonsense knowledge, contributing to early understanding of model capabilities before instruction-tuned LLMs.
- **Joe Davison**, Kristen Severson, and Soumya Ghosh. **Cross-Population Variational Autoencoders**. NeurIPS Bayesian Deep Learning Workshop, 2019.
  Developed a probabilistic representation-learning method for separating shared and population-specific latent structure across differing data distributions.
- Thomas Wolf et al. **Transformers: State-of-the-Art Natural Language Processing**. EMNLP Demo Track, 2020. Best Demo Award; 14,000+ citations.
  Contributed to the core open-source library that standardized practical Transformer model development, fine-tuning, and deployment workflows across the ML research community.
- Quentin Lhoest et al. **Datasets: A Community Library for Natural Language Processing**. EMNLP Demo Track, 2021. Best Demo Award; 900+ citations.
  Helped establish reusable data loading, sharing, and benchmarking infrastructure for reproducible NLP and model-evaluation workflows.
- Thomas Butler et al. **MS2Mol: A Transformer Model for Illuminating Dark Chemical Space from Mass Spectra**. ChemRxiv, 2023. 35+ citations.
  Applied generative Transformer modeling to noisy scientific sequence-to-structure prediction, improving molecular identification workflows from mass spectra.

EDUCATION

University of Utah · Doctoral Studies, Computing - Artificial Intelligence — Salt Lake City, UT · Aug. 2023 – Aug. 2024
Harvard University · M.S., Data Science — Cambridge, MA · Aug. 2018 – Dec. 2019
Brigham Young University · B.S., Computer Science — Provo, UT · Jan. 2015 – Apr. 2018

SKILLS

Accelerated compute: PyTorch DDP, ZeRO, Slurm, CUDA, Apple MPS, RunPod, AWS, GCP

Frameworks: Hugging Face TRL, Hugging Face Transformers, Hugging Face Datasets, Accelerate, PyTorch, JAX, TensorFlow/Keras, TensorFlow Probability, PyMC3, scikit-learn, MXNet, CNTK, MMLSpark

Research tooling: Weights & Biases, PyTorch Lightning, Hydra

Engineering: Python, TypeScript, Clojure, Java, Kubernetes, Temporal, distributed systems, microservices

Machine learning: Transformers, deep learning, zero-shot classification, distillation, self-training, generative models, VAEs, latent-space models, CNNs, LSTMs, Bayesian models, statistical machine learning, uncertainty calibration