# Sam Havens

Machine Learning Researcher, Engineer, & Manager | Portland, Oregon

linkedin.com/in/samhavens | samhavens@gmail.com | 818.590.0484 | Google Scholar | DBLP

## WORK EXPERIENCE

**Databricks — Staff Research Scientist**

Jul 2023 – Present

- Led post-training for DBRX / DBRX-Instruct (instruction tuning + preference optimization) and drove cross-team release execution.
- Built the internal agent harness (evaluation, regression testing, tool-use simulation) and productionalized it for multiple teams.
- Developed continual learning and agent memory techniques (e.g., never-ending learning pipelines; memory-augmented judging/evaluation).
- Built and published evaluation & benchmarking for agentic retrieval, long-context context management, tool use, LLM-as-judge, and grounded reasoning (see publications and writing below).

**MosaicML — Research Scientist**

Sep 2022 – Jul 2023

- Led the development of chat/instruction-tuned variants of MPT-7B and MPT-30B, enhancing usability for downstream applications.
- **MosaicBERT**: Developed a BERT-style encoder architecture and training recipe optimized for fast pretraining (FlashAttention, ALiBi, GLU, dynamic padding removal, low precision LayerNorm).
- **LIMIT**: Investigated the impact of small, high-quality instruction fine-tuning datasets on LLMs across traditional 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 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, outperforming an open-source baseline by 130%.
- Served a character-based transformer spelling correction model via Triton/ONNX with 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; productionized models via modern async/await Python, Docker/Kubernetes/PubSub, plus Bash and Jenkins.

**CarLabs — Chief Technology Officer**

May 2016 – Jan 2019

- Created a suite of tools for automotive OEMs and dealers to manage chatbots across web, chat, and voice.
- Used Docker/Kubernetes to operate services written in Node.js, Elixir, and Python; models in FastText and TensorFlow.
- Engineering team grew from 4 to 16 during my tenure; established a 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.

## SELECTED PUBLICATIONS & PREPRINTS

- **FreshStack: Realistic Benchmarks for Agentic Retrieval on Technical Documents.** arXiv:2504.13128 (2025). arXiv
- **Long-context Context Management Performance of Large Language Models.** arXiv:2411.03538 (2024). arXiv
- **LoRA Learns Less and Forgets Less.** TMLR (2024). arXiv | OpenReview
- **MosaicBERT: A Bidirectional Encoder Optimized for Fast Pretraining.** NeurIPS (2023). arXiv | NeurIPS
- **LIMIT: Less Is More for Instruction Tuning Across Evaluation Paradigms.** arXiv:2311.13133 (2023). arXiv

## TECHNICAL WRITING

Full list: Databricks blog author page.

- **MemAlign: Building Better LLM Judges From Human Feedback with Scalable Memory.** (Feb 3, 2026) post
- **Introducing OfficeQA: A Benchmark for End-to-End Grounded Reasoning.** (Dec 9, 2025) post
- **Agent Learning from Human Feedback (ALHF): A Databricks Knowledge Assistant Case Study.** (Aug 4, 2025) post
- **Embedding Model Finetuning for Agentic Retrieval.** (Feb 20, 2025) post
- **The Power of Fine-Tuning on Your Data: Quick Fixing Bugs with LLMs via Never Ending Learning (NEL).** (Apr 8, 2025) post
- **Evaluating Long-context Context Management (OpenAI o1 vs Google Gemini).** (Oct 8, 2024) post
- **Beyond the Leaderboard: Unpacking Function Calling Evaluation.** (Aug 16, 2024) post
- **Long-context Context Management Performance of LLMs.** (Aug 12, 2024) post