# Derek Deming

**Email**: derekdeming17@gmail.com
**Website:** http://derekdeming.io/ **Phone:** 760-525-0871
**Github**: https://github.com/derekdeming **LinkedIn**: https://www.linkedin.com/in/derek-deming/

## Education

PhD in Computational Chemical Biophysics (dropped out)
MS in Computational Chemistry – *Emphasis in Machine Learning*

- PNNL-WSU Distinguished Graduate Research Program Fellow
- Research Assistant: Radioactive Material and Engineering Fellowship @ Department of Energy (DoE)
- Research Gate: https://www.researchgate.net/profile/Derek-Deming-2

BA in Biology and Chemistry – *Emphasis in Chemical Biology and Machine Learning Research*

- Bioinformatics Research Fellow at the Orthopedic Surgery Specialty Clinic
- Distinguished Presidential Scholar Undergraduate Research Fellow

## Skills

**Languages:** Proficient in C, C#, C++, Python, R, Golang (GO), JavaScript, TypeScript, KQL, SQL, SCOPE, TLC, YAML, and Bash
**Compiler Expertise:** LLVM, MLIR, TensorRT, CUDA, ONNX Runtime, TVM
**Optimization Techniques**: FP16/INT8 quantization, kernel fusion, loop unrolling, memory reuse with OrtValue, model pruning, quantization, efficient attention mechanisms, sparse training
**Frameworks & Tools:** AutoML, Express, DrizzleORM, ONNX, MLflow, Langchain, LlamaIndex, DeepSpeed, Triton, PySpark, Spark SQL, React.js, .NET, TensorRT, TVM, Docker, Kubernetes
**Cloud & Distributed Systems**: AWS, Azure, high-performance computing, container orchestration with Kubernetes, EKS

**Machine Learning Architectures & Techniques:**

- *Deep Learning Frameworks:* PyTorch, TensorFlow, CUDA, Unsloth, HuggingFace Transformers, Keras, Prophet
- *Regression & Ensemble Methods:* Expertise in **Linear Regression**, **Random Forests**, and **XGBoost** for predictive modeling
- *Convolutional Neural Networks (CNNs):* Experience with models like **YOLO** for object detection and image processing tasks as well as OCR and segmentation using other CV models
- *Recurrent Neural Networks (RNNs) & LSTMs:* Implemented for time-series forecasting and sequential data analysis.
- *Transformers & Large Language Models:* Worked with transformer-based language and vision models such as **BERT** (Google)**, GPT-2** (OpenAI)**, GPT-3** (OpenAI), **LLaMA** (META), **RETNet** (MSFT), etc., for NLP and computer vision applications.
- *Techniques:* LLM finetuning, transfer learning, knowledge distillation, pruning (filter, channel, layer pruning), low rank factorization, quantization, retrieval-augmented generation (RAG), sparse training, efficient attention mechanisms, model compression

**Cloud & Distributed Systems**

- *Platforms* & *Tools*: Extensive experience with AWS, Azure, Snowflake, Databricks, high-performance computing
- *Certifications*: Certified AWS Databricks Platform Architect and in Databricks Lakehouse Platform Fundamentals
- *Containerization* & *Orchestration*: Docker and Kubernetes

## Experience

**Software Engineer, Machine Learning** (Security Research Org), Microsoft – Boston, MA Dec 2023 – Present

- I am part of the Microsoft Defender of Office (MDO) Security Research Org at MSFT specifically working on building out the machine learning capabilities and scalable infrastructure of the product. I am also part of the Sonar Machine Learning (Sonar ML) team which is the full detonation platform we built for detonating threat

vectors in real time. One project in particular that I was a part of was building an internal small language model (SLM) for Business Compromised Emails (BEC), Spam and Phish detection. For this project we trained, fine-tuned, and optimized perception DNNs in FP16/INT8 precision, enhancing model efficiency while reducing computational overhead for production deployments and leveraged NVIDIA's TensorRT and CUDA extensively to optimize neural networks for speed and accuracy improvements, ensuring compatibility with GPU architectures used internally.

- Within the first 3 months of being at MSFT, I built and deployed a proprietary near real-time inference computer vision model with aides in detecting and decoding malicious QR codes in messages. The model is scaled out touching over half a billion messages on a daily basis saving the company over 25 million dollars in COGs processing messages with third party software.
- I am the creator and primary contributor to a new ONXX Model Predictor library built in C# and .NET which serves as a CPU inference compute engine for our models. This library allows the security models we build to be consumed by organizations across MSFT (inside and outside my org) as a Nuget package and offers near real time inference (sub 30 milliseconds) time with extensive memory optimization through OrtValue reuse.
- Due to MSFT Security First Initiative (SFI), there has been a mass infrastructure lockdown to ensure our systems are "SFI compliant". I am the lead engineer for my team and much of the work associated with SFI has been around proper authentication, storage account cloud security, network isolation of our virtual machines (VMs) and much more.

**Open-Source ML Software Engineer,** Independent Contractor Nov 2022 – Present

- I strongly believe in contributing and working with open source software when it comes to building solutions for enterprises, especially in-house solutions. One thing I have worked on is implementing RLHF (Reinforcement Learning from Human Feedback) pipelines in Langchain and LlamaIndex and Langraph within a startup called Cartha which develops personalized therapy solutions with AI. As part of this work, I developed custom retrievers and query transformers to route data to the proper places so make sure the query was being served appropriately.
- Applied advanced optimization techniques such as model pruning, quantization, and kernel fusion to deep learning models, enhancing their suitability for low-latency applications on edge devices. Implemented RAG systems with NVIDIA acceleration frameworks, utilizing hybrid retrieval methods for better performance, leading to more efficient memory utilization during inference.
- Another project I have worked on is optimizing RAG systems using hybrid vector + sparse retrieval and reciprocal rank fusion via advanced retrieval systems. This work included implementing concepts like Hypothetical Document Embeddings (HyDE) for zero-shot dense retrieval, ColBERT for late-interaction dense retrieval and Chain-of-Thought (CoT) for enhanced reasoning of multi-step queries. I also lead in the developing custom tokenizers and embedding models for specialized security datasets
- Other projects have included finetuning and deploying the latest edge open-source language models (Phi 2, Phi 3, and Phi 3.5, Llama 1/2/3(3.2), and Mistral 8x7B. One library in particular that I find very useful for finetuning language models is Unlsoth as it allows you to unlock the GPUs that are otherwise locked up during the finetuning process. I have applied a variety of AI training and post-training processing techniques such as PEFT methods (LoRA, QLoRA) on Phi-3, Llama 3, and Mistral-7B for efficient adaptation. So that we could use the models on edge devices via Ollama, I leveraged GPTQ and GGUF quantization for optimized inference.

**Data Platform ML Software Engineer,** Securian Financial – REMOTE in Utah May 2023 – Dec 2023

- ML-Focused Infrastructure Design & MLOps: Designed, built, and managed scalable AI infrastructure, leveraging MLOps best practices to implement ML solutions into production. Emphasized continuous integration (CI), continuous testing (CT), and continuous deployment (CD), utilizing GitHub Actions for workflow automation. Developed and maintained Infrastructure as Code (IaC) using Terraform and CloudFormation to provide consistent environments for data science workflows within AWS.
- Data & ML Architecture: Architected data pipelines for high-throughput processing using **AWS Glue**, **Apache Spark**, and **S3** to support both real-time and batch processing. Created ETL workflows with **AWS Step Functions** to automate data ingestion from financial datasets, optimizing downstream ML training. Collaborated with data scientists to build and deploy credit risk models in **Amazon SageMaker**, leveraging **SageMaker Pipelines** for

parameter tuning, and using **Spot Instances** to reduce costs by up to 60%.

- Cloud & Real-time Data Streaming: Deployed scalable ML workloads using **AWS Lambda** for event-driven processing and **EKS** for container orchestration. Automated training and inference pipelines using **Apache Airflow** for end-to-end scheduling and monitoring. Implemented model versioning with **MLflow**, capturing metadata for model training and evaluation metrics. Managed codebases through **GitHub**, integrating branch protection rules, **GitHub Actions** for automated testing, and code review best practices to maintain quality control. Built a real-time data ingestion system using **Apache Kafka** integrated with **Amazon Kinesis** to provide continuous data flow for ML models. Implemented a stream processing application to generate insights from incoming customer behavior data, feeding it into production-grade fraud detection models.

- Containerization & Microservices: Developed and containerized ML training and inference services using **Docker**, and deployed to **Amazon EKS** for scalability and resource efficiency. Created **Helm charts** to manage complex deployments involving multiple services and dependencies, facilitating rapid scaling during financial end-of-quarter reporting. Configured **horizontal pod autoscaling (HPA)** to manage spikes in model inference load, optimizing resource utilization across multiple microservices. Used **Prometheus** and **Grafana** for real-time monitoring of containerized ML services to maintain SLA compliance and system health visibility.

**Founding ML Software Engineer,** Thera AI – REMOTE in Utah Jan 2023 – July 2023

- Founded a startup company trying to build tools for bioinformatics researchers. The focus was on the practical application and composability of LLMs to build transformative applications in the space of biology and high-throughput experimentation.
- Built out end to end retrieval augmented generation (RAG) pipelines utilizing data framework Llama Index to utilize personalized data with LLMs.
- Designed and implemented machine learning pipelines to support the embedding of biological data for LLMs to interact with. Utilized LLMs to generate novel biological sequences and predict potential functional implications, providing a powerful tool for experimental design and hypothesis testing.
- Implemented advanced features such as callback handlers for the collection of nested run objects, enabling more sophisticated evaluation and testing procedures for LLM-based applications.
- Built websites around implementing LLM and Gen AI technology into the frontend and backend. The tech stack used for this ranges from: React.js, Next.js 13, MongoDB, PostgreSQL, Tailwind, Prisma, OpenAI APIs, Langchain APIs, Llama Index APIs and much more. The MVP was developed on Streamlit.
- Collaborated in the development of AI-powered chatbots for biological research, providing natural language processing capabilities that can answer complex queries, summarize research findings, and facilitate knowledge sharing within the scientific community.

**ML Data Scientist II (MLOps Specialist),** Swire Coca Cola – Salt Lake City, UT Aug 2022 – June 2023

- Overview: Started and grew a team of data scientists and machine learning engineers at a startup initiative inside Coca Cola to handle supply chain issues.
- Machine Learning & Deep Learning: Built and implemented advanced machine learning models for diverse business applications, including sales demand forecasting via cutting-edge deep learning techniques (at the time). We performed a ton of time-series analysis accounting for extreme seasonality due the dataset we were working with. I also designed and developed the inference infrastructure for a trade promotion optimization application using genetic algorithms and Streamlit.
- MLOps & Data Engineering: Developed and streamlined CI/CD pipelines via Jenkins and Azure DevOps for seamless ML model deployment. Used Kubernetes for orchestrating containerized applications and Git for version control. Enhanced security and monitoring practices for robust ML application reliability.
- Real-world Business Solutions: Led projects addressing key business challenges such as **sales demand forecast** using time series analysis with deep learning models, **customer churn prediction** and intervention strategies using NLP analysis of feedback, **customer segmentation** through NLP-based clustering, and **supply chain optimization** using reinforcement learning.
- Technology Evaluation & Implementation: Identified and incorporated new software technologies to improve performance, maintainability, and reliability of ML systems. Tools included MLOps life cycles, Streamlit, MLflow,

Delta Lakes, Docker, Cloud development, Snowpark, Snowflake database, Databricks, and Azure.

**Research Scientist** University of California, Irvine – Irvine, CA Aug 2020 – Sep 2022

- Applied statistical and machine learning techniques, including deep learning, to create scalable simulations for systems of interest. Employed deep learning models such as CNNs for the analysis of molecular structures, and unsupervised learning techniques for clustering and dimensionality reduction. Analyzed and understood large amounts of data for specific conditions and worked closely with collaborators to optimize the complexity of the simulations.
- Implemented atomic-scale molecular dynamics and multi-conformational Monte Carlo simulations, as well as machine learning techniques to simulate protein structures and optimized conformations of the protein structures. This required in-depth statistical analysis as well as dimensionality reduction analysis such as, KNN, regression, clustering, SVMs. The data analysis was performed in both Python and R scripting.
- Used a multiscale molecular simulation approach to gain atomic-level insight into the interprotein interactions that stabilize concentrated solutions of wild-type γ-crystallins and lead to the formation of aggregates in solutions of their cataract-related mutants. Utilized deep learning techniques for protein structure prediction and classification, enabling a better understanding of the underlying mechanisms.
- Translated statistical simulation analysis results to experimentalist collaborators to verify results and discuss further simulations and types of analyses that needed to be completed, including the potential integration of advanced machine learning methods for the interpretation of research findings and the identification of novel therapeutic targets.

**Research Scientist,** Department of Energy & Washington State University – Pullman, WA June 2018 – May 2020

- Spearheaded multiple research projects in collaboration with Pacific Northwest National Laboratory to develop a data pipeline between WSU and National Laboratory for continuous research analytics in our computational models.
- Applied various statistical methods in computational design of Metal-Organic Frameworks (MOFs), including regression analysis, Principal Component Analysis (PCA), cluster analysis, machine learning algorithms, Bayesian optimization, Monte Carlo simulations, molecular dynamics simulations, genetic algorithms, and artificial neural networks (ANNs) for property prediction and optimization.
- Utilized these techniques to effectively explore the vast MOF design space, identify structure-property relationships, and guide experimental synthesis efforts towards optimal material designs.
- Deployed genetic algorithms as an optimization technique based on the principles of natural selection and genetics, used to search for optimal MOF structures by evolving a population of candidate materials through selection, crossover, and mutation operations.
- Exploited synthetic crystallographic techniques to understand MOFs as extrapolating agents in solid-solvent phase extractions to improve radioactive waste separations. Leveraged Monte Carlo simulations paired with experimental data to design the most ‘optimized’ material.
- Mentored undergraduate students pursuing research by teaching them laboratory techniques, conceptual understanding of research tactics, and leading them on projects they found interesting, including the application of advanced machine learning techniques in their research.

**Undergraduate Research Assistant,** Concordia University, Irvine – Irvine, CA June 2015 – May 2018

- Developed a machine learning program using python to determine the difficulty of a college course based on previous grades and the current grade distribution.
- Gained experience with experimental molecular and biomolecular techniques (i.e., DNA isolation, PCR, sub-cloning, microbial transformation, solution/media preparation, aseptic techniques) as well as computational protein modeling and statistical models (utilized python and R).
- Synthesized, purified, and spectroscopically characterized chromium transition metal complexes with acetylacetone, chloride and bromide ligands. Complexes were synthesized using controlled conditions.

- Determined degradation pathway of Sphingomonas bacterium of antibiotic resistant bacteria through Spectroscopic Techniques. Compared the localization of human and yeast copper-zinc superoxide dismutase (SOD1) in Saccharomyces cerevisiae.

## Projects + Hackathons

**ONNX ML Predictor** Microsoft Internal

- I developed an ONNX ML Predictor library built in C# and .NET which serves as a CPU inference compute engine for our models. This library allowed us to modularly add and remove models as new ones were added to production as well as load in over 40+ models asynchronously. The library essentially maps inputs to outputs of a specified scenario and model. So we are able to have computer vision models as well as text based models mapped through the Predictor serving a near real time inference of sub 30 ms for image detection models.

**Mechanistic Interpretability of GPT 2** Github: ML Interpretability

- This work was inspired by Chris Olah (Anthropic) and built on top of foundational work done by Neel Nanda. I looked into the emergent properties of positional embeddings in GPT-2 using TransformerLens (developed by Neel Nanda) and custom probing classifiers. Implemented data generation pipelines, logistic regression and neural network probes, and performed layer-wise analysis and attention head ablation studies. Demonstrated the ability to predict word positions from residual streams and identified specific layers and attention heads crucial for position encoding. This project contributes to LLM interpretability by providing insights into how positional information is encoded and processed within the model architecture.

**Real-Time Malicious QR-Code Detection & QR-Code Decoding** Microsoft Internal

- QR code detection is a serious project when it comes to security, especially when QR-codes can be embedded with loads of malicious content in them. I built a proprietary computer vision model which is composed of two main building blocks: a QR detector model trained to detect and segment QR codes and a QR code decoder. The decoder is built using Pyzbar, different image preprocessing techniques that maximize the decoding rate on difficult images.

**Blackbox AI: Interpretability of YOLO Object Detection** Github: YOLO Interpretability

- Basically this project was inspired by an internal project I worked on and I wanted to further investigate the YOLO model series and try to understand the “why” behind its verdicts.
- Layer-wise Relevance Propagation (LRP) is a technique used for explaining decisions of deep neural networks by propagating the prediction backward through the network using purposely designed rules. Researchers developed and analyzed various LRP rules (LRP-0, LRP-ε, LRP-γ) for different network layers, optimizing for both fidelity and understandability of explanations. They applied the Deep Taylor Decomposition framework to theoretically justify LRP rules and implemented efficient LRP algorithms using automatic differentiation in PyTorch, enhancing interpretability for complex AI models.
- This project was an implementation of this research paper: Layerwise Relevance Propagation

**Rapidly** Github: Rapidly

- This is a one stop shop for enterprise knowledge management software. Think of it as Glean or GlueAI but modern, dynamic, quick and reliable. We all know that one person who we turn to for all information in the company dating back 15+ years. Rapidly is the software that enables all employees to be knowledgeable over the entire stack without requiring 15+ years of experience. This knowledge management software deploys LLMs safely and reliably across the enterprise. Youtube [DEMO] Link

**Finetune Phi 3.5 using Unsloth** Github: LLM-stuff

- This project fine-tunes the Phi-3.5-mini-instruct model on various cybersecurity datasets using the Unsloth library. It processes and formats multiple datasets related to MITRE ATT&CK, cybersecurity tactics, and vulnerabilities, preparing them for training. The script implements efficient training techniques, including LoRA and 4-bit

quantization, and includes features for model evaluation, saving, and optionally pushing to the Hugging Face Hub.

- Leveraged Unsloth library for advanced LLM fine-tuning, implementing Low-Rank Adaptation (LoRA) for efficient parameter updates and hardware-specific optimizations for NVIDIA GPUs. Utilized Automatic Mixed Precision (AMP) and gradient accumulation to accelerate training while maintaining accuracy. Employed Unsloth's memory-efficient "Tron" kernels and precise analytical backpropagation, significantly reducing memory usage and computational overhead in LLM training pipelines.

**BioInformatics Research Assistant** Github: BioIDE

- This project was in the early days of ChatGPT so I basically wanted to implement a domain specific personal research assistant. I had done quite a bit of computational biophysics research in the past, especially in grad school so I was building something I wish I had at the time. I noticed early on that ChatGPT was too general when it came to domain specific queries, so I figured if we combined the context of the latest research with the knowledge of ChatGPT then we could speed up the process of reading, discovering, and doing research.

## Research Papers

Li X., Ding G., Hao L., Deming, D. A,[b] and Qiang Zhang (2020). *ACS Appl. Mater. Interfaces* https://doi.org/10.1021/acsami.0c04961

Hao, L., Ding, G., Deming, D. A. and Zhang, Q. (2019), *Eur. J. Org. Chem.* doi:10.1002/ejoc.201901303

Derek Deming et. al (2019) A Facile Method to Introduce Iron Secondary Metal Centers into Metal–Organic Frameworks, *J. Organ. Chem*. doi.org/10.1016/j.jorganchem.2019.06.0

## Research Talks

Microsoft MLADS+ Responsible AI Conference: **Real-Time Malicious QR-Code Detection & QR-Code Decoding**

Microsoft FireCon (focused around Security, Incident Response, Reverse Engineering and more): **How to Make Computer Vision Models Less of a Black Box through Layerwise BackPropagation**