Hunter Heidenreich | AI Scientist & Research Engineer

Jersey City, NJ
843-789-0824 • hheiden0@gmail.com • hunter-heidenreich
hunter-heidenreich • hunterheidenreich.com

Professional Summary

AI Scientist & Research Engineer combining industrial expertise in scaling Large Language Models (LLMs) and Vision-Language Models (VLMs) with deep research in machine learning for scientific simulation. Specialized in deploying foundation models on H100 clusters and engineering physics-informed generative models.

Research & Engineering Experience

Roots.ai **Jersey City, NJ**
AI Research Scientist *Feb 2025 – Present*

- Directed the full data lifecycle for a complex, structured extraction project, coordinating a 12-person annotation team to build high-quality ground-truth datasets.
- Personally annotated samples to reverse-engineer domain logic, encoding constraints that drove extraction accuracy from 60% to 95%.
- Engineered parameter-efficient fine-tuning (PEFT) pipelines using Unsloth and Hugging Face TRL to model complex dependencies in semi-structured text, outperforming closed-source models like GPT-4o on extraction tasks while reducing training memory footprint.
- Scaled full-parameter, distributed training of Vision-Language Models (VLMs) on DGX H100 systems, leveraging DeepSpeed ZERO (Stage 2 & 3) to manage massive, terabyte-scale multimodal dataset training jobs.

Roots.ai **Jersey City, NJ**
AI Engineer *Nov 2023 – Feb 2025*

- Led the development of a PEFT framework for sequential data segmentation using large language models (LLMs), achieving state-of-the-art performance on public benchmarks.
- Co-authored and published work in COLING 2025, presenting a novel approach for automated sequential data processing.
- Validated the framework on complex, real-world structured data, demonstrating superior accuracy over baseline methods and investigating key deployment challenges like model calibration.

CSElab, Harvard University **Cambridge, MA**
Graduate Research Assistant *Sept 2021 – Nov 2023*

- Published novel Transformer and RNN architectures for scientific time-series forecasting (dynamical systems), improving short-term prediction accuracy by 7x over baseline.
- Designed generative surrogate models (Transformers, GNNs, VAEs) to accelerate molecular dynamics simulations by orders of magnitude, avoiding computationally expensive physics solvers.
- Refined Encoder-Decoder Transformers for probabilistic forecasting by implementing MDN predictions, capturing the multi-modal nature of stochastic dynamics.
- Built high-fidelity data pipelines using GROMACS and LAMMPS, generating protein dynamics trajectories to train deep learning models on physical systems.

CODED Lab, Drexel University **Philadelphia, PA**
Undergraduate Research Assistant *Jan 2019 – June 2021*

- Spearheaded NLP research leading to publications at EMNLP and AIES, developing novel unsupervised semantic induction algorithms and qualifying universal adversarial triggers in foundation models (GPT-2).

Education

Harvard University **Cambridge, MA**
Master of Science in Computer Science, GPA: 3.94 *Sept 2021 – June 2023*
Academic Track: Completed PhD-level coursework and passed Qualifying Exams before transitioning to industry.
Research: Generative Surrogate Modeling for Molecular Dynamics (CSElab).

Drexel University **Philadelphia, PA**
Bachelor of Science in Computer Science, GPA: 4.00 *Sept 2016 – June 2021*

Technical Skills

- **AI & ML Frameworks:** PyTorch, JAX, DeepSpeed ZERO, Hugging Face (TRL/PEFT), Unsloth, bitsandbytes
- **Generative Models:** Transformers, Diffusion Models, VAEs, Graph Neural Networks (GNNs)
- **Scientific Simulation:** GROMACS, LAMMPS, RDKit, ASE, PySCF, NumPy, Pandas, Scikit-learn
- **Languages & Tools:** Python, C/C++, CUDA, Git, Docker, AWS, Azure ML

Selected Publications

For a complete list of publications, please refer to my Google Scholar profile.

Peer-Reviewed

- **Heidenreich, H.**, Dalvi, R., Verma, N., Getachew, Y. (2025). "Page Stream Segmentation with LLMs: Challenges and Applications in Insurance Document Automation." *Proceedings of COLING 2025*.
- **Heidenreich, H. S.**, Williams, J. R. (2021). "The Earth Is Flat and the Sun Is Not a Star: The Susceptibility of GPT-2 to Universal Adversarial Triggers." *Proceedings of AIES 2021*.
- **Heidenreich, H.**, Williams, J. (2019). "Latent Semantic Network Induction in the Context of Linked Example Senses." *Proceedings of W-NUT 2019, EMNLP 2019*.

Preprints & Working Papers

- **Heidenreich, H. S.**, Vlachas, P. R., Koumoutsakos, P. (2024). "Deconstructing Recurrence, Attention, and Gating: Investigating the transferability of Transformers and GRUs in forecasting of dynamical systems." *arXiv:2410.02654*.
- **Heidenreich, H.**, Dalvi, R., Mukku, R., Verma, N., Pičuljan, N. (2024). "Large Language Models for Page Stream Segmentation." *arXiv:2408.11981*.