

# Ariel N. Lee

AI Researcher & Engineer

@ [ariellee@bu.edu](mailto:ariellee@bu.edu) | [LinkedIn](#) | [GitHub](#) | [Website](#) | [Greater Boston, MA](#)

## EDUCATION

---

**M.Sc., Boston University (BU)** **GPA: 3.71** Boston, MA  
Electrical & Computer Engineering - Data Analytics Specialization *Sep 2020 – May 2023*  
**Activities:** Out in STEM; Graduate Women in Science & Engineering

**B.Sc., University of California, Los Angeles (UCLA)** **GPA: 3.45** Los Angeles, CA  
Microbiology, Immunology, & Molecular Genetics (MIMG) *Sep 2011 – Jun 2015*

## CURRENT POSITIONS

---

**[Raive](#), Founding Research Scientist, Multimodal Models** *Sep 2023 – Present*  
Generative multimedia foundation models with IP attribution. Applied experience in large-scale multimedia dataset collection/filtering, pretraining, post training, and efficient lowrank model personalization.

**[Data Provenance Initiative](#), Lead** *Mar 2024 – Present*  
Recent work featured by the [New York Times](#): analysis of 14,000+ web domains to understand evolving access restrictions in AI and improve transparency, documentation, and informed use of data.

## PUBLICATIONS

---

**[[Paper](#), Under Submission]**  
Shayne Longpre, ... (23 authors), **Ariel N. Lee**, ... (15 authors), Stella Biderman, Alex Pentland, Sara Hooker, Jad Kabbara. “Bridging the Data Provenance Gap Across Text, Speech, and Video” (2024)

**[[Paper](#), NeurIPS Workshop 2024]**  
Clément Bonnet, **Ariel N. Lee**, Franck Wertel, Antoine Tamano, Tanguy Cizain, Pablo Ducru. “From Text to Pose to Image: Improving Diffusion Model Control and Quality” *NeurIPS Workshop on Compositional Learning: Perspectives, Methods, and Paths Forward (2024)*

**[[Paper](#), NeurIPS 2024]**  
Shayne Longpre, Robert Mahari, **Ariel N. Lee**, ... (45 authors), Sara Hooker, Jad Kabbara, Sandy Pentland. “Consent in Crisis: The Rapid Decline of the AI Data Commons” *NeurIPS Datasets and Benchmarks Track (2024)*

**[[Paper](#), NeurIPS Workshop 2023]**  
**Ariel N. Lee**, Cole J. Hunter, Nataniel Ruiz. “Platypus: Quick, Cheap, and Powerful Refinement of LLMs” *NeurIPS Workshop on Instruction Tuning and Instruction Following (2023)*

**[[Paper](#), arXiv 2023]**  
**Ariel N. Lee**, Sarah Adel Bargal, Janavi Kasera, Stan Sclaroff, Kate Saenko, Nataniel Ruiz. “Hardwiring ViT Patch Selectivity into CNNs using Patch Mixing” *preprint arXiv:2306.17848 (2023)*

## PRESENTATIONS

---

### [[Research Presentation](#), 2024]

Women in AI & Robotics

### [[Research Presentation](#), 2024]

AI Tinkerers x Human Feedback Foundation

### [[Guest Lecturer](#), 2023]

Hong Kong University of Science and Technology

LLMOps, Prof. Sung Kim

## EXPERIENCE

---

### **Platypus LLMs, garage-bAInd**

Boston, MA

*Co-lead Researcher, Open Source Large Language Models*

*May 2023 – Aug 2023*

- [Platypus models and dataset](#) have **1M+ downloads** on HuggingFace. Our best model, tuned on the Llama architecture, was the global leader in post trained open-source LLMs at the time of release and for two months after.
- Researched low-cost and efficient ways to refine domain-specific SOTA LLMs using LoRA and refined datasets with **Cole J. Hunter** and **Dr. Nataniel Ruiz**.

### **Boston University, AI4ALL**

Boston, MA

*Researcher, Program Coordinator*

*May 2022 – May 2023*

- Conducted research with **Dr. Nataniel Ruiz**, **Prof. Sarah Adel Bargal**, and **Prof. Kate Saenko** to study patch selectivity in modern convnets and ViTs. Worked on counterfactual simulation and testing of neural nets.
- Co-led AI4ALL summer program at BU to teach a diverse group of high schoolers about AI.

### **Boston University, College of Engineering**

Boston, MA

*Deep Learning Course Grader*

*Jul 2022 – May 2023*

- Completed grading and answered student questions for the Deep Learning graduate course with **Prof. Sarah Adel Bargal** and **Prof. Brian Kulis**.

### **TeachForward & BU Wheelock Educational Policy Center**

Boston, MA

*Data & Process Engineer, MLOps Dev Team*

*Sep 2022 – Dec 2022*

- Developed a feature extraction pipeline to analyze the use of teaching time based on 10,000+ videos of classroom observations.
- Created a simple user interface for client using gradio and Hugging Face spaces. User uploads a video and pipeline returns mp4 files with object and activity detection annotations, among others.

### **eMinutes**

Los Angeles, CA — Boston, MA

*Corporate Paralegal (Remote)*

*Aug 2019 – Mar 2021*

*Manager of Entity Management*

*Oct 2018 – May 2019*

*Corporate Paralegal*

*Apr 2017 – Oct 2018*

- Identified optimization opportunities in the company's web-based document and communication system, in addition to corporate work such as entity formations.

### **Law Offices of Sanford Jossen**

Los Angeles, CA

*Paralegal*

*Oct 2016 – Apr 2017*

*Legal Assistant*

*Oct 2015 – Oct 2016*

- Researched and drafted legal documents, and summarized complex medical records.

## PROJECTS & COMPETITIONS

---

### [Competition, META AI 2023]

Meta AI Video Similarity Challenge – **8/196** overall, **1/42** in AI grad course | [Leaderboard](#)

- Used a pretrained, Self-Supervised Descriptor for Copy Detection model (ResNeXt101) to find similar, manipulated videos in a dataset of 40,000+ videos.

### [Competition, Kaggle 2023]

Leveraging Fine-tuned Models for Prompt Prediction | [Code](#) | [Leaderboard](#)

- Ensemble-based approach for predicting text prompts used to generate Stable Diffusion images.
- Surpassed the performance of traditional image captioning models by employing fine-tuned CLIP and ViT models and using a custom dataset of 105,000 image-prompt pairs.

### [Competition, Computer Vision Course 2022]

Visual Odometry: Mapping Out the Camera Path | [Code](#)

- 3rd place in CS 585 Computer Vision class challenge, focused on estimating the camera path by recovering relative motion between successive frames.

### [Final Project, Deep Learning Course 2022]

Crypto of the Future: Reinforcement Learning | [Code](#)

- DL reinforcement algorithm — proximal policy optimization — to devise an automatically generating strategy for Ethereum transactions.

## UNDERGRADUATE RESEARCH

---

### UCLA Department of MIMG

Los Angeles, CA

*Undergraduate Researcher, Characterization of Novel Bacteriophages*

*Sep 2014 – Jun 2015*

- Worked with **Dr. Giorgia Pirino** to advance phage therapy research in the SEA-PHAGES project by isolating a novel bacteriophage: PH8s.
- Probed potential gene functions via electron microscopy and plaque assays, leading to a fully annotated genome added to the [NCBI GenBank database](#).
- Poster presentation at the UCLA MIMG Symposium on Characterization of Novel Bacteriophage PH8s

### UCLA Department of Psychology

Los Angeles, CA

*Undergraduate Researcher, Directed Research in Medicine*

*Jun 2014 – Aug 2015*

- Conducted research with **Dr. Thomas Minor** for senior project by using learned helplessness to model symptoms of Post-Traumatic Stress Disorder.

## SKILLS

---

**Programming & Technologies:** Python (PyTorch, jax, transformers, diffusers, TensorFlow, NumPy, Pandas, scikit-learn), OpenCV, GCP, Lambda Cloud, RunPod, Git/GitHub, AWS, Hugging Face Hub (spaces, datasets, models)

**ML/AI Techniques:** multimodal pretraining and post training, diffusion, LLM instruction tuning, LoRA tuning, large-scale data collection and refinement, data augmentation, ML pipeline deployment, open-source models and datasets