

Alexander Oh

Research Scientist Intern

New York, NY | [LinkedIn](#) | [GitHub](#)

AI researcher with a Master's in Artificial Intelligence from Duke, focused on model evaluation, benchmark design, and self-supervised representation learning. Has built retrieval benchmarks, adversarial evaluation suites, and production document parsing pipelines.

EXPERIENCE

Research Scientist Intern

Jun 2026 — Sep 2026

Petra Labs

New York, NY

Wrote three research papers on a statistical evaluation framework for Generative Engine Optimization, measuring how prompt variations change LLM citations and the brands named in responses. Advised operators on prompt-map optimization for customer visibility in AI search systems. Built statistical diagnostics that flag significant shifts in citation data for strategy and ops teams.

AI Capstone Engineer

Dec 2025 — Jun 2026

Huma.ai

Durham, NC

Architected a LangGraph deep research agent that generates medical affairs documents, saving MSLs 24 to 72 hours of manual research per document. Built a 60-question RAG benchmark over 14 clinical PDFs with a 0.461 baseline that surfaced specific retrieval failures. Created an LLM-as-Judge compliance benchmark grounded in FDA guidance, which showed a 4.5/10 improvement in agent performance. Deployed a table parsing pipeline using UniTable and LayoutLMv3, evaluated with TEDS across more than 10K tables from PubMed research PDFs.

AI Research Scientist

Sep 2025 — May 2026

Duke Natural Language Processing Lab

Durham, NC

Showed that semantically guided view selection during ImageNet-1K self-supervised pretraining improves transfer across seven downstream datasets, with the largest gains in low-data settings (+7.6% at 1-shot). Co-led ArXivDoc, an 8,210-document scientific retrieval benchmark comparing document-as-image retrieval with text-based methods. Designed the annotation protocol that narrowed 4,059 candidate queries to 547 evaluation queries. Combined InfoNCE with LeJEPa's SIGReg regularization for self-supervised RL, cutting training time 24% on JaxGCRL Humanoid and AntMaze. Built an adversarial evaluation suite for CLIP under PGD attacks that exposed embedding instability and vision-language failure modes under distribution shift.

AI Graduate Research Scientist

Sep 2025 — Jun 2026

Duke University

Durham, NC

Duke Natural Language Processing Research Lab

JEPa Models, Reinforcement Learning, Information Retrieval

Multimodal AI, Distributed Data Parallel (DDP), Slurm, Adversarial AI (PGD attacks),

Associate

Mar 2025 — Jun 2025

Riviera Partners

Los Angeles Metropolitan Area

Pitched opportunities to candidates and evaluated them for technical positions within public and VC-backed companies. Placed engineers and scientists in research-heavy roles in HPC, MLOPS, CFD Simulation, and IT.

Data Engineer Analyst Intern

Jun 2023 — Sep 2023

Maroon Invest Global

London, England, United Kingdom

Designed AWS data pipelines to support the unique needs of fund research and early-stage portfolio companies. Created dashboards in Plotly Dash and Tableau to visualize metrics for investment committee presentations. Conducted extensive market research and due diligence to identify investment opportunities.

Product Management Intern

Jun 2023 — Aug 2023

Adventuretown Toy Emporium

Los Angeles

Analyzed data on social media, website traffic, and competitors' portfolios to construct a strategic marketing plan. Networked and negotiated promotion deals with 7 influencers, achieving a 25% increase in social media activity. Redesigned website to improve user experience, resulting in a 23% increase in web traffic.

Teaching Assistant

Aug 2022 — May 2024

Middlebury College

Middlebury, Vermont, United States

Led weekly drop-in tutoring to support student learning for Computer Architecture and Systems Programming courses. Guided students through assignments and reinforced core computer architecture concepts including circuits simulation, assembly-level debugging, and low-level memory. Collaborated with the professor to relay student progress, identify areas of difficulty, and provide course feedback.

Information Technology Lead

Jun 2022 — Aug 2022

Roughing It Day Camp

Lafayette, California, United States

Student Shadow

Kaiser Permanente

Shadowed family medicine physicians in day to day appointments, meetings, and procedures.

Jul 2021 — Aug 2021

Pleasanton, California, United States

Summer Camp Counselor

Tyler Hill Camp

Jun 2021 — Jul 2021

Tyler Hill, Pennsylvania, United States

Lifeguard

ACALANES UNION HIGH SCHOOL DISTRICT

Jun 2017 — Aug 2018

Moraga, California, United States

EDUCATION

Duke University

Master of Engineering - MEng in Artificial Intelligence

GPA: 3.74

Aug 2025 — Aug 2026

Middlebury College

Bachelor of Arts - BA in Computer Science

Magna Cum Laude, College Scholar's List, NESCAC Winter All-Academic Team

Aug 2020 — May 2024

PROJECTS

LeDEEP: Depth Perception from JEPA

Ever see a Waymo and wonder why it needs those bulky LiDAR sensors? While modern autonomous systems rely on expensive, specialized hardware, humans navigate the world using monocular RGB input alone. This motivated our development of LeDEEP, a vision-only framework that uses LeJEPA's self-supervised learning to extract 3D structure from simple camera frames. By leveraging stable SSL objectives to produce rich latent spaces, we've built a resource-efficient alternative that shows that robotics don't need a thousand-dollar "hat" to perceive depth.

XGBoost from Scratch

XGBoost is a foundational tool for data scientists, driving some of the most accurate and scalable machine learning solutions across diverse industries. Its tree-based approach delivers efficient, interpretable, and high-performing models, especially suited for structured, tabular datasets.

By building XGBoost from scratch, I deepen my understanding of the model and similar algorithms, enabling me to fine-tune performance, interpret results more clearly, and develop future solutions for real-world challenges.

This project deepened my understanding of gradient boosting mechanics, optimization strategies, and the inner workings of scalable ML systems.

Advanced Operating Systems Labs

In my operating systems coursework, I gained hands-on experience working directly with kernel-level code to deepen my understanding of core OS concepts. I explored file system internals by navigating inodes, block and fragment addresses, and reimplemented the functionality of Unix commands like find and cat by manually calculating file system structure addresses. Additionally, I experimented with virtual memory management by creating and analyzing custom memory mappings, handling page faults, and observing page swapping behavior using DTrace and C. These projects strengthened my skills in low-level systems programming, memory diagnostics, and performance analysis—key competencies for systems, infrastructure, or performance engineering roles.

ArXivDoc Paper

Document-as-Image Representations Fall Short for Scientific Retrieval

<https://arxiv.org/abs/2604.18508>

Computer Network Simulation

Understanding the fundamental protocols that underpin the internet is crucial for anyone working in networking, cybersecurity, or systems engineering. Protocols like TCP/IP, Ethernet, ARP, and ICMP form the core of modern digital communication, supporting everything from basic web access to large-scale cloud infrastructure. A deep understanding of how data is transmitted, routed, and managed across networks enables the design of more efficient, secure, and scalable systems. In this project, I implemented a network interface using FreeBSD's virtual networking packages and replicated the behavior of key protocols including IP, Ethernet, ARP, and ICMP. This hands-on exploration provided valuable insight into how these protocols operate in practice and how the internet functions at a foundational level.

Cybersecurity Data Science Pipeline

Built an end-to-end local data pipeline to ingest, process, and analyze ~20GB of publicly available security data from the CVE, CWE, and KEV APIs. Designed the pipeline in multiple stages using pandas with multithreading and manual parallelization, then optimized performance with polars for efficient memory management and finally with dask for scalable, multi-core processing. Focused on maximizing local MacBook hardware before transitioning to cloud resources. Performed exploratory data analysis (EDA), primarily on high-dimensional textual data, to extract insights and guide future modeling approaches. This project has strengthened my skills in data engineering, performance optimization, and iterative problem framing—laying the groundwork for future large-scale security analytics applications.

Sports Betting Web App

I collaborated on the development of a sports betting web application that enables users to place and track wagers with persistent account management across sessions. I built a local endpoint API to manage user balances and betting history and contributed to frontend feature design and implementation, with a strong focus on improving the user experience through iterative feedback. I designed the Firebase database architecture and deployed Google Cloud Functions to enable real-time updates within the app. This project strengthened my skills in full-stack development, cloud deployment, and performance optimization.

This full-stack web application allows users to place and track sports wagers with persistent, secure Google account functionality to manage balances and betting history across sessions.

SKILLS

Reinforcement Learning · Model Evaluation · Benchmark Design · Statistical Analysis · Ablation Studies · Robustness and Adversarial Evaluation · Information Retrieval · Representation Learning · Multimodal Learning · Distributed Training · PyTorch DDP · Slurm · GCP Vertex AI · Model Serving · Robotics · World Models · Multimodal AI · Agentic Systems · Model Robustness and Safety · Benchmarking · Agentic AI Development · Retrieval-Augmented Generation (RAG) · Model Context Protocol (MCP) · Self-Supervised Learning · Artificial Intelligence (AI) · PyTorch · CUDA · Computer Vision · Multi-agent Systems · Natural Language Processing (NLP) · Responsible AI · Large Language Models (LLM) · Exploratory Data Analysis · Parallel Processing · BERT (Language Model) · Statistical Modeling · XGBoost · Machine Learning Algorithms · Data Pipelines · Cloud Computing · Business Intelligence (BI) · Search Engine Optimization (SEO) · Due Diligence · Network Design · Back-End Web Development · Product Life Cycle · Google Cloud Platform (GCP) · Agile & Waterfall Methodologies · REST APIs · Linux Kernel · Operating System Security · Dtrace · Memory Management · Portfolio Management · DCF Valuation · Sensitivity Analysis · React Native · Investment Management · Full-Stack Development · Operating Systems · Financial Modeling · Amazon Web Services (AWS) · Pandas (Software) · Data Mining · Product Management · PikToChart