Johann Lee

johannchrislee@gmail.com | (609) 865-5476 | San Jose, CA 95110 | Github

EDUCATION

Cornell University — Ithaca, New York

Bachelor, Computer Science — *December 2024*

- **GPA**: 3.96. **Honors**: Ex-President of Cornell Data Science, TA for Grad Machine Learning, Teradata Analytics Challenge 1st Place
- **Courses:** Machine Learning, Computer Vision, Natural Language Processing, Systems, Econometrics, Causal Inference, A/B Tests

WORK EXPERIENCE

Adobe — San Jose, California

Data Scientist — *Jan 2024 - Present*

- Working on Acrobat GenAI https://www.adobe.com/acrobat/generative-ai-pdf.html

Adobe — San Jose, California

Data Scientist Intern — *May 2024 - Aug 2024*

- Developed end-to-end and **productionized** subscription likelihood **prediction model** (XGBoost), enabling targeted discounts and pop-ups for **4.5M Acrobat users**. Implemented product variant A/B tests. Estimated **$500k** ARR increase, launching 2024 Q4.
- Identified **1M** related users with graph algorithms, enabling **recommendations** for engagement and upsell worth **$100k** ARR.
- Orchestrated compute clusters and set up model deployment and performance monitoring systems (Airflow, Databricks).
- Improved product-usage compute logic for 1B Photoshop events, reducing compute time from days to hours (Azure, Spark).

ArXiv — Ithaca, New York

Researcher — *Feb 2024 - Present*

- Developed **classifiers** to tag research paper submissions categories for Cornell's arXiv platform (**4M+** monthly active users).
- Fine-tuned T5 (LLM) to encode text corpuses for document **search** (3% improvement in first search result compared to BM25).
- Trained classifier on query embeddings to memorize the best model version per query, achieving 5% improvement on new queries
- Improved ROME's (open source search model) fact editing algorithm, decreasing **query response time** by 30%.
- Currently researching automating question-answering dataset generation to gauge unbiased LLM performance for queries

Bank of America — Charlotte, North Carolina

Machine Learning Intern (Quantitative Summer Analyst) — *Jun 2023 - Aug 2023*

- Built automated hallucination **evaluation infrastructure** for chatbot with **40M users** and designed **out-of-distribution detection** system for questions, increasing helpfulness by 30%. Business unit estimated **$1M+** savings, work featured at July Townhall.
- Tuned chatbot training objective for a 3% improvement in top 25 customer **queries** and 20% improvement in 10 hardest requests.
- Researched chatbot-hallucination's sensitivity to paraphrasing, eliminating 40% of hallucination while retaining 90% of truth.

NextStep Health Tech — New York City, New York

Software Engineer Intern — *Jul 2022 - Aug 2022*

- Developed a full-stack data analytics and dashboard web app for analyzing health data collected and generating visualizations.
- Designed and implemented an interactive dashboard and an automated **data processing pipeline** to load health metrics, clean and validate data, and execute analysis to generate insights. Defined and integrated **data storage** for these insights.

PUBLICATIONS

[1] J. Lee and D. Lee. Towards Safe and Ethical AI. In *Global Review of AI Community Ethics*, accepted into Vol. 3. No 1 (2025)

[2] A.Gong, C. Wan, J.Lee, R. Thesmar, J. Klenke and K. Q. Weinberger. PhantomWiki: Synthetic Data Generation for Accurately Evaluating Agentic AI Multi-hop Question Answering. *Submitted to ICML (2025)*.

PROJECTS

Web Search Engine For Math Equations (Github, advised by Prof. Kilian Weinberger) — Ithaca, New York

Project Lead — *Aug 2022 - May 2023*

- Trained equation detection model (YOLO), finetuned CNN with contrastive loss for clustered vector embeddings (PyTorch).
- Built NoSQL database (AWS DynamoDB) to store user uploaded PDFs; exploring vectorized search and Redis for query caching.
- Developed REST APIs between backend AWS Lambda endpoints and frontend web app to send user search queries.

TECHNICAL SKILLS

Languages: Python, Java, C++, C, SQL, Bash, Shell, JavaScript, TypeScript, HTML, CSS, PHP

Frameworks and Cloud: Pytorch, Tensorflow, Azure, AWS, Spring Boot, Flask, Django, React, Vue, D3.js

Tools and Database: Spark, Docker, Databricks, Airflow, MySQL, DynamoDB, MongoDB, Cassandra, Git, GitHub, Linux