Johann Lee
jcl354@cornell.edu | San Jose, CA | Github | LinkedIn

EDUCATION

Cornell University Ithaca, New York
Bachelor, Computer Science *December 2024*

- **GPA**: 3.97 **Honors**: TA for **Grad** Machine Learning, Ex-President of Cornell Data Science, Teradata Analytics Challenge 1st Place

WORK EXPERIENCE

Adobe San Jose, California
Data Science Engineer *Feb 2025 - Present*

- Targeting and recommendation systems for Adobe Acrobat & GenAI subscriber growth.
- ML for causal inference: conditional ATE models for optimal treatment assignment.

Cornell University Ithaca, New York
Researcher *Feb 2023 - Present*

- Currently investigating intra-GPU caching to speed up data center LLM inference & curriculum training multi-hop retrieval LLMs.

Adobe San Jose, California
Data Scientist Intern *May 2024 - Aug 2024*

- Developed end-to-end and **productionized** subscription likelihood **prediction model**, enabling targeted discounts and pop-ups for **4.5M Acrobat users**. Implemented product variants and A/B tests (estimated **$1M** ARR increase), launched 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
Research Engineer *Feb 2024 - Dec 2024*

- Developed **classifiers** to tag research paper submissions categories for Cornell's arXiv platform (**4M** monthly active users).
- Fine-tuned LLMs to encode text corpuses for document **search** (3% improvement in first search result compared to ElasticSearch).
- Improved ROME's (search algorithm) fact editing algorithm to utilize caching, decreasing average **query response time** by 20%.

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.

PUBLICATIONS

PhantomWiki: Generating Reasoning and Retrieval Datasets On-Demand (ICML 2025)
*A.Gong, C. Wan, K. Stankeviciute, A.Kabra, **J.Lee**, R. Thesmar, J. Klenke, C. Gomes, and K. Q. Weinberger*

- A synthetic dataset generation pipeline for multi-step LLM reasoning across multiple data-sources to address data contamination.
- Implemented Agentic and RAG LLMs for evaluation, built knowledge-graph to dataset generation pipeline (PyTorch, vLLM, HF).

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

- Surveyed and analyzed benchmarks for evaluating bias and hate of LLMs, identifying systemic weaknesses and scaling issues.

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: C++, Java, Python, SQL, C, Bash, Shell, OCaml, JavaScript / TypeScript, HTML, CSS, PHP
Frameworks and Cloud: Pytorch, Tensorflow, Azure, AWS, Spring Boot, Flask, Django, React, Vue, D3.js, Scikit-learn, Pandas
Tools and Database: Spark, Docker, Databricks, Airflow, MySQL, DynamoDB, MongoDB, Cassandra, Git, GitHub, Linux