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

WORK EXPERIENCE

Adobe — San Jose, California

Data Scientist — *Feb 2025 - Present*

- Targeting and recommendation systems for Adobe Acrobat & GenAI subscriber growth.

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 **$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

Research Assistant — *Feb 2024 - Dec 2024*

- 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%.

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.

PUBLICATIONS

PhantomWiki: Generating Reasoning and Retrieval Datasets On-Demand (ICLR 2025 DATA-FM)

A.Gong, C. Wan, K. Stankeviciute, A.Kabra, ***J.Lee****, R. Thesmar, J. Klenke, C. Gomes, and K. Q. Weinberger*

- We introduce a synthetic dataset pipeline for multi-step LLM reasoning across multiple data-sources. By generating our dataset on demand, we mitigate the issue of data contamination. We also scale reasoning and retrieval separately, disentangling performance.
- Implemented CoT 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*

- We survey and analyze 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