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*

- Building prediction, targeting, and recommendation systems for Adobe Acrobat & AI Assistant to drive subscriber growth.
- Productionized machine learning models to optimally target audiences for marketing promos and campaigns, earning $600k ARR.
- Developed ML model that predicts subscription purchase group, earning $400k GNARR by personalizing offers.

Cornell University Ithaca, New York

Researcher *Feb 2023 - Present*

- Current: Intra-GPU memory offloading between multiple GPUs to speed up data center LLM inference (PI: Prof. Rachee Singh)
- Current: Training small LLMs' complex reasoning capabilities through synthetic datasets (PI: Prof. Kilian Weinberger)

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

Document Processing Webapp For Unstructured Data *Feb 2025 - Present*

- Built full stack webapp (React, Java, Python, PostgreSQL, AWS) with 30+ API endpoints for this document processing system.
- Implemented secure billing, role-based access controls, async processing workflows, rate limiting, caching to handle enterprises.

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