Johann Lee

(609) 865-5476 | jcl354@cornell.edu | Google Scholar

Education

Cornell University *Ithaca, NY*

Bachelor, Computer Science | GPA: 3.97/4.0 *Jan. 2025*

- **Honors**: TA for the masters' level Machine Learning course, Ex-President of Cornell Data Science, Milstein Tech Scholar, 1st Place Teradata Hackathon

Work Experience

Adobe Inc. *Feb. 2025 – Present*

Machine Learning Engineer, Growth *San Jose, CA*

- Proposed and built agentic analytics platform for analyzing audience segments without coding, increased mean AB test opportunity size by 10%.
- Implemented AI-data-analyst to instantly analyze and interpret AB test results, increasing AB test velocity by 20%.
- Deployed an ML model for optimal audience-segment targeting of Adobe Acrobat's marketing promos and campaigns, increasing CTR by 5%.
- Developed ML model to personalize product bundle recommendations at checkout, increasing Acrobat & AI Assistant's average order size by 4%.

arXiv.org (5M Monthly Active Users) *Feb. 2024 - Dec. 2024*

Machine Learning Engineer Intern, Search & Recommendations *Ithaca, NY*

- Developed ML-in-the-loop data labeling pipeline to classify research paper submissions to this open-access archive, increasing throughput by 80%.
- Built a real-time recommendation system for paper submission categories, decreasing submission errors flagged by moderators by 10%.
- Fine-tuned LLMs for indexing and retrieval in the document search engine, improving top-5 search hits by 3% compared to Elasticsearch.

Adobe Inc. *May 2024 – Aug. 2024*

Data Scientist Intern, Ad Targeting *San Jose, CA*

- Productionized an ML model that targeted discounts and messaging for 4.5M churned Acrobat users, increasing cohort ARR by 7% in AB test.
- Implemented accompanying MLOps pipeline by orchestrating compute and developing model deployment and performance monitoring systems.

Bank of America *Jun. 2023 – Aug. 2023*

Machine Learning Engineer Intern, AI Chatbots *Charlotte, NC*

- Built hallucination evaluation and out-of-distribution detection system for chatbot (40M users), increasing response helpfulness by 20% in internal testing. Business unit estimated $1M call-center cost savings, work featured at the department's monthly Townhall.
- Tuned chatbot training objective to achieve a 3% improvement in the 25 most frequent queries and 20% improvement in the 10 hardest queries.

NextStep HealthTech (Series A Startup) *Jul. 2022 – Aug. 2022*

Software Engineer Intern, Analytics *New York, NY*

- Developed a full-stack real-time data analytics webapp for non-technical stakeholders to surface data-driven "next item" suggestions.
- Implemented a robust data pipeline with schema versioning and fault-tolerance, supporting the app's scalable deployment for this startup.

Publications

[ICLR '26] **Learning From Synthetic Data Improves Multi-hop Reasoning** | *International Conference on Learning Representations*

- We show post-training on synthetic data can induce LLMs to learn generalized reasoning skills in multi-step question answering.
- I identified knowledge composition as the key improvement driver and built experiments showing its persistence across domains.

[ICML '25] **PhantomWiki: Generating Reasoning and Retrieval Datasets On-Demand** | *International Conference on Machine Learning*

- We built a synthetic data pipeline for multi-step LLM reasoning across many documents, which is data-contamination resistant.
- I built the document generation pipeline, trained models to show contamination-resistance, and built agentic and RAG evaluation.

Projects

AI Engineer, Bastion *Aug. 2025 – Feb. 2026*

- Built full-stack AI webapp (50+ API endpoints) with commercial real estate brokers in Atlanta, Chicago, and D.C.: AI agent parses unstructured documents, completes financial modeling, generates marketing materials, then matches suitable real estate buyers based on webscraped data. [Website]
- Tinkered with prompt prefix caching, language model cascades, task-based model routing, and output caching to reduce LLM API costs by 80%.

Machine Learning Engineer, MathSearch *Jan. 2023 – Jun. 2024*

- Led team of 10 to build a cloud-native search engine for math equations, enabling students to search textbooks for complicated formulas. [GitHub]
- Trained vision models to detect equations, fine-tuned image embeddings for similarity search, built the fullstack webapp, and deployed on AWS.

Technical SKills

Languages: Python, Java, C++, JavaScript, SQL, OCaml, C

Frameworks: FastAPI, Spring / Spring Boot, Flask, Django, JUnit, React

AI/ML: PyTorch, HuggingFace, vLLM, LangChain, Scikit-learn, FAISS

Tools: AWS, Postgres, Docker, Spark, Git / GitHub, Airflow, MongoDB

Community: Adobe Community Impact Team - Champion, Synopsys Championship - Judging Team Lead (software)