Vedaant Jain

•Email: vvjain3@illinois.edu • Mobile: +1-447-902-2107
linkedin.com/in/vedaant-jain/ github.com/Vedaant-J

Education

University of Illinois at Urbana Champaign May 2026
Bachelor of Science: Computer Science & Mathematics, GPA: 3.95/4.00; Dean's List

Skills & Coursework

C/C++, Python, Go, TypeScript, Linux, Git, Docker, CI/CD, AWS, MongoDB, PyTorch, SQL, CUDA, PostgreSQL LLVM, Rust, GCP, DeepSpeed, Firebase, NoSQL, React

Distributed Systems, ML Compilers, Computer Security, Computer Architecture, Machine Learning, Programming Languages

Publications & Preprints

Jain V, Alves Feitosa F, Kreiman G (2024). Is AI fun? HumorDB: a curated dataset and benchmark to investigate graphical humor.**ICCV 2025** arXiv: 2406.13564.

Chaudhary I, Jain V, Singh G (2024). Decoding Intelligence: A Framework for Certifying Knowledge Comprehension in LLMs **SeT LLM@ICLR 2024**, arXiv: 2402.15929 [cs.AI].

Work Experience

Kumo.AI, Software Engineering Intern May 2025 - Aug 2025

- **Scoped and built** a deterministic end-to-end, full-stack code generation system using **compiler principles** to automatically translate Kumo UI entities into Python SDK code, **reducing manual authoring time by over 90%**.
- Designed and executed **120+ experiments** and reduced search space by **75% for 10 hyperparameters** for Graph Transformers (GT).
- Developed **synthetic datasets and trained custom GNN models** outperforming baselines by an average of **10%** for pre-sales engagements with enterprise clients including **Cisco, ThredUp, and eBay**.
- Enabled large file upload feature supporting > **1GB local uploads**, resolving critical pain points for 5 enterprise customers. Delivered customer-facing tutorial notebooks for **Kumo-RFM product release**.

Kreiman Lab - Harvard Medical School, Computer Vision Researcher June 2023 - June 2024

- Created **3500+ sized novel dataset** across **3 humor detection tasks** (relevance, creativity, impact), designing custom **annotation platform supported with AWS (S3, Lambda)** to gather **60k+ human responses** for data analysis and human humor baseline.Implemented pipelines using **DeepSpeed** and **PyTorch** for distributed fine-tuning of multimodal models on multi-GPU clusters.

Metaphor Data Inc., Software Engineering (ML) Intern May 2024-Aug 2024

- Built **CI/CD** testing using **TypeScript/Jest** with synthetic data generation for LLM/RAG pipelines, and boosted performance **20%**.
- Pioneered a **CLI tool** with automated **GraphQL** query generation and **5+ API endpoints** using **AWS, MongoDB** for batch transactions increasing efficiency by **10x**.

Focal Lab - UIUC, Machine Learning Researcher Aug 2023 - Present

- Implemented **probabilistic certification framework** for LLM knowledge comprehension via Knowledge Graphs using **PyTorch/Huggingface**, evaluating **8 models** on few-shot medical QA via distributed inference on supercomputing clusters.
- Identified **adversarial input distributions** causing **15%** performance drop in models like Gemini-1.5-Pro with **realistic noise**.
- Designed **domain-specific programming language (DSL)** for image distribution specification using scene graphs, with an **optimization framework** for realistic adversarial scene generation.

Disruption Lab-UIUC, Head of AI/ML Aug 2022 - May 2024

- Led over **100 students** on more than 10 AI projects (e.g., LLM RAG crawler for 1000+ sites, accurate hardware-counter based malware detection, Unity virtual classrooms) for clients like AMD; deployed 5+ ML models on **AWS/GCP** and built full-stack apps with **React/Next.js and PostgreSQL**.

Projects

- Developed & optimized Matrix Multiplication and Convolution kernels in **CUDA and NKI**, achieving 1000x performance improvement over baseline implementations on **GPU and NPU architectures on NVIDIA and AWS Trainium chips**.
- Developed **LLM persona simulators** that adopt Reddit users' styles from their comment histories; compared next-comment vs masked-fill prompting with **sentiment similarity analyses**.
- Integrated **Curriculum Learning** with language models and RL for embodied planning, **improving performance by 10%**.
- Implemented a thread based LLM chat interface for desktop using **Tauri (Rust), React, SQLite**.
- Built **iOS app** aggregating thrifting deals from various websites, with **Firebase, AWS, SQL, S3**, reducing shopping time by **20%**.
- Engineered android **mobile-smart watch ECG** system using **ML and signal processing, React Native** for music recommendations
- Engineered **distributed file system** in **Go** enabling real-time file operations with gRPC communication.
- Built **indoor wireless localization** system using **IMU sensors**, **wireless signal processing**, and **Kalman filters**.