Ethan Chen

New York, NY | 217-200-2324 | jc3766@cornell.edu
www.blue-mirror.com | www.github.com/ethanchen143

EDUCATION

Cornell University, Cornell Tech New York, NY
MEng in Computer Science May 2026
Relevant Coursework: Machine Learning Engineering, ML Hardware and Systems**,** Reinforcement Learning, Generative Models

University of Illinois at Urbana Champaign Urbana, IL
BS in Computer Science and Music, GPA: 3.74 / 4.00 May 2025
Relevant Coursework: Applied Machine Learning, Database Systems, Computer Systems, NLP, Deep Learning for Computer Vision

PROFESSIONAL EXPERIENCE

Cornell Tech – S4AI, Researcher, New York, NY Sep. 2025 - Present

- Collaborating with S4AI group to study the design and implementation of energy-efficient LLM serving system.
- Setting up design space exploration strategies for LLM serving frameworks (vLLM, SGLang) to evaluate energy, throughput, and latency under different scheduler configurations (continuous-batching, chunk-prefill), router policies (cache-aware, least-load), P-D disaggregation vs co-located serving and hybrid parallelism strategy.
- Developing advanced caching policies to tail latency under multi-SLO settings.

Bluffingface, Founder, Chicago, IL Summer 2025

- Built an online poker platform with immersive 3D gameplay (Three.js), raising $100,000 in angel investment.
- Deployed scalable backend on AWS EC2 with redundancy and failover, supporting 800+ concurrent players.

CreatoRain, Software Engineer, Urbana, IL Jun. 2024 – Jul. 2025

- Launched creatorain.com, an influencer matching platform using Pinecone for vector match and React for frontend.
- Designed a database of 2M+ US influencers with PostgreSQL and reached 50k+ influencers for clients.
- Implemented semantic retrieval using Pinecone, optimized index configuration and metadata filters to reduce latency at scale.
- Optimized client chatbot for Nutr (Shark Tank company) with RAG, integrated embeddings, indexing, and retrieval augmented prompts to reduce hallucination and cost.

SELECTED PROJECTS

Algophony (Data Engineering, Deep Learning) Jan. 2024 – May. 2024
Class Project, AI Music Generation

- Trained generative music models in PyTorch, implementing CVAE and diffusion architectures.
- Engineered 500+ GB of audio training data using NumPy, Pandas, and Python audio libraries.

Loopbop (Web Dev, Graph Database) Spring 2025
Capstone Project, 3D Music Genre Explorer

- Created loopbop.com, a music genre explorer, with Three.js and React, blending real-time gameplay and 3D exploration.
- Implemented graph database with Neo4j, enabling fast traversal of 30k+ songs and associated artists, genres, albums, etc.

AltCredit (Machine Learning) Fall 2025
HackIllinois '25 Project, ML credit scoring system to expand loan access for undocumented individuals.

- Prototyped a fast backend with FastAPI to efficiently parse official documents and process user data.
- Built a data pipeline using PySpark, updating the classification model in real time.

TECHNICAL SKILLS

Coding Language: Python, C/C++, JavaScript/TypeScript, SQL, Java, Swift, Scala
Framework and Tool: PyTorch, Numba, Triton, SGLang, GCP, AWS, Neo4j, PostgreSQL, MongoDB, Redis, PySpark, React, vLLM