Sophie Shih

New York, NY | sophie59595@gmail.com | LinkedIn | sophieshih.com

Education

Duke University | *Computer Science and Neuroscience* | 4.0/4.0 — Expected: May 2030
Relevant Coursework: Data Structures & Algorithms; *Upcoming:* Intro to Computer Systems, Proof-Based Linear Algebra, Applied Product Management | Involvements: Catalyst Tech Society, DTech, The Chronicle, StreetMed Dance Team, Club Tennis

Stuyvesant High School | GPA: 4.0/4.0 || 97 — Jun. 2026
Classes: AP Computer Science, AP Physics C Mech, AP Physics C E&M, AP Calculus BC, AP Chemistry, AP Mandarin, AP US Government, AP Macroeconomics, AP European History

Technical Skills

Languages: Python, Java, JavaScript, TypeScript, SQL, HTML/CSS, Racket, NetLogo
Frameworks & Libraries: Next.js, React, Streamlit, LangChain, Pydantic AI, FAISS, scikit-learn, pandas
Tools & Platforms: Git/GitHub, Docker, PostgreSQL, Neo4j, Vercel, OpenAI & Claude APIs
Areas: Full-Stack Web Development, LLM Agents & Retrieval-Augmented Generation, Machine Learning, Data Pipelines

Experience

Software Engineering Intern, AI Research — Jun. 2025 – Sep. 2025
Mass General Brigham, PI: Dr. Synho Do — *Remote / Boston, MA*
- Built Neurobiome Navigator, an interactive app that teaches patients and students how the gut and oral microbiome relate to Parkinson's disease (demo)
- Engineered a retrieval-augmented AI agent (Pydantic AI, LangChain, OpenAI API) that answers natural-language questions by querying the MINERVA knowledge graph (66,000+ microbe–disease associations) and running semantic search over research papers
- Built interactive visualizations backed by knowledge-graph queries for microbe and food data, with user-controlled sorting and views
- Documented the system in a medRxiv preprint on using AI to turn complex biomedical findings into educational tools

Machine Learning Research Intern — Jul. 2024 – Jun. 2025
Weill Cornell Medicine, Department of Radiology (PI: Dr. Martin Prince) — *New York, NY*
- Contributed to TraceOrg, a public web tool serving a 3D U-Net/transformer model that measures kidney, liver, and cyst volumes (0.97 Dice); co-author on the JASN 2026 paper
- Produced training and validation data for the segmentation model by annotating 500+ abdominal and pelvic MRIs of kidney disease (ADPKD) patients
- Prototyped a few-shot LLM prompting workflow that generates patient-friendly explanations of incidental MRI findings (RSNA abstract)
- Evaluated segmentation algorithms across MRI sequences (T1, T2, SSFP) on 3D volumetric data

AI Engineering Intern — Jun. 2024 – Jul. 2024
MD.ai — *New York, NY*
- Engineered an LLM-agent framework that generates patient-friendly letters; co-authored an arXiv paper and won 2nd place at the SIIM Hackathon (slides / video)
- Tested features in MD.ai's production reporting software, including LLM-based automated radiology report generation
- Curated medical question-answering data and presented weekly reports on LLM applications in healthcare

Code Coach — Nov. 2024 – Jun. 2025
TheCoderSchool — *New York, NY*
- Taught 5 students (ages 7–10) 1-on-1 in Scratch, adapting lessons to each student's skill level and interests
- Designed 15+ custom coding projects and challenges teaching core concepts (loops, conditionals, variables), including interactive physics games (gravity, acceleration, collisions)

Projects

College App Atlas · Live site · GitHub | *Next.js, TypeScript, PostgreSQL, Claude & OpenAI APIs* — Apr. 2026
- Built and launched a full-stack web app where students brainstorm, draft, and track college applications and essays
- Designed a PostgreSQL database of 5,700+ colleges built from IPEDS and College Scorecard data, with essay prompts verified against official sites
- Integrated Claude and OpenAI APIs into an "idea garden" feature where students grow ideas into essay threads and get AI-assisted college research tailored to their interests

The New York Census Line · Live site | *Python, JavaScript, Census API, NYC Open Data* — Oct. 2026
- Built and deployed an interactive "subway map" website exploring housing, immigration, and income across NYC's 55 neighborhoods
- Wrote Python data processing that combines 2020–2024 Census API, NYC Open Data, and BLS data, computing inflation-adjusted trends and correlations (e.g. rent burden vs. income, $r = -0.79$)

Online Purchase Intent Prediction · GitHub · Slides | *Python, scikit-learn, pandas · Catalyst Tech Society, team of 4* — Fall 2026
- Built a Random Forest classifier predicting whether online shopping sessions end in a purchase, trained on 12,000+ sessions (0.92 ROC-AUC, 90% accuracy)
- Implemented a leakage-safe preprocessing pipeline (deduplication, one-hot encoding, stratified 80/20 split) and analyzed precision–recall trade-offs under class imbalance (15.6% buyers)

Awards

SIIM (Society for Imaging Informatics in Medicine) Hackathon — 2nd Place, 2024 (slides / video) and 2025 (slides / video)