Xiangyu (Leo) Shi

xiangy.shi@gmail.com | (858) 568-0271 | Menlo Park, CA
linkedin.com/in/xleoshi/ | github.com/xiangyshi | about-leo.dev

---

EDUCATION

Carnegie Mellon University, Silicon Valley
Master of Science, Software Engineering   **August 2025 – December 2026**
- **1st Place**, CMU-SV Hackathon (*"SafeBark" - NLP Positive Content Replacer*)
- Software Engineering Methods, Functional Programming, Applied Data Science

University of California San Diego
Bachelor of Science, Mathematics - Computer Science   **September 2021 – March 2025**
- **Cum Laude**, GPA: 3.9, Provost's Honors
- Data Structures and Algorithms, Object-Oriented Programming, Database Systems, Software Engineering, Computer Systems, Computer Architecture, Design Patterns, Statistical Modeling, Deep Learning, Data Mining

---

PROFESSIONAL EXPERIENCE

J. Craig Venter Institute   **June 2025 – August 2025**
Data Science & Machine Learning Engineer Intern   **La Jolla, CA**
- Developed ML-driven applications for protein structure and stability prediction, integrated into a researcher-facing web platform (Express.js + HTML/JS/CSS) with dockerized ML tools on AWS EC2 and Nginx.
- Designed an orchestration layer with a dynamic job queue for managing tool calls, extendable to Kubernetes environments.
- Automated data ingestion and preprocessing pipelines through Python, Jupyter and Bash scripting, supporting large-scale proteome translation of thousands of viral sequences (~20k amino acids each).

Yunming Technology   **January 2025 – June 2025**
Software Engineer Intern   **Remote**
- Engineered a modular **Retrieval-Augmented Generation (RAG)** chatbot system for Customer Relationship Management (CRM) workflows, integrating query reconstruction, semantic search, and real-time responses via Server-Sent Events (SSE).
- Enhanced document retrieval relevance by **~50%** through TF-IDF re-ranking and custom evaluation metrics (Precision@K, MRR), improving CRM agent knowledge responses.

Human Internet   **September 2024 – June 2025**
Full Stack Engineer Intern   **Remote**
- Transformed the humanID developer console by refactoring backend API (Django/REST, MySQL).
- Boosted SSO reliability by 80% through improved exception handling and JWT-based session flow; dockerized to streamline development across global teams and added support for international phone formats.
- Collaborated with distributed engineering teams to optimize core API performance and ensure seamless system integration.

PROJECTS

SafeBark (CMU-SV Hackathon – 1st Place)   **October 2025**
- Real-time NLP pipeline converting toxic comments to positive tone using OpenAI API and Hugging Face.

Epistasis Bi-Clustering Algorithm (Scripps Research)   **September 2024 – March 2025**
- Reduced pipeline runtime from **16 hours to 10 minutes** through code profiling and algorithmic optimization.

TCDB Classifier Pipeline (UC San Diego)   **July 2024 – September 2025**
- Designed and built data preprocessing and evaluation workflow for large-scale ML classification tasks.

Teacher Analytics Dashboard (STEMz Learning)   **June 2024 – Present**
- Developed app for multi-classroom data visualization and chat features**,** served over **200+ teachers and students.**

TECHNICAL SKILLS

Languages: Python, JavaScript (ES6+), TypeScript, Java, C/C++, F#, Bash, SQL
Frameworks & Libraries: React.js, Next.js, Node.js, Express.js, Flask, Django, LangChain, Socket.IO
Databases & Data Tools: MySQL, PostgreSQL, MongoDB, PyTorch, NumPy, Pandas, SciPy, Diffusion Models
Cloud, DevOps & Other Tools: AWS (EC2, S3, RDS), Azure, Docker, Git/GitHub Actions, Vercel, Render, Linux/Unix, UML, REST API, OpenAI API, Hugging Face, Agile, Scrum, CI/CD, PgVector, Nginx