

Lennox (Siyan) Fu

631-627-9272 | lenoplus42@berkeley.edu | [linkedin.com/in/lennoxfu/](https://www.linkedin.com/in/lennoxfu/) | github.com/Lenoplus42

EDUCATION

University of California, Berkeley

Bachelor of Arts in Computer Science, GPA: 3.7/4.0

Awards: USACO Platinum Full Mark (Top 1%)

Berkeley, CA

Expected May 2027

EXPERIENCE

TikTok (ByteDance)

San Jose, CA

Software Development Engineering Intern

May 2026 – Aug 2026

- Developed a **coding agent harness** for **platform-wide migration** of recommendation pipelines from legacy JSON configurations to a Python DSL, reducing per-pipeline migration time from **2–3 days of manual work to hours** of automated generation and verification
- Raised migration pass rates from **~20% to ~80%** on evaluated configurations through **automated verification and agent self-correction**, with passing migrations requiring no manual code edits
- Built an **agent benchmark and verifier** comparing retrieval-index metadata to detect structural mismatches, guide **agent self-correction** and improve migration correctness
- Implemented **vector-search integration** over internal code and DSL documentation, enabling the agent to retrieve relevant APIs and examples during translation

SGLang

Remote

Open Source Contributor

Jul 2026 – Present

- Profiled and optimized **text-to-speech inference** in SGLang-Omni using **CUDA Graphs**, reducing **first-audio-chunk latency by 54–59%** and increasing **request throughput by 15–21%** in paired H200 benchmarks at concurrency 16
- Standardizing **GPU assignment** across SGLang-Omni stage factories to enable **process-level replication** across GPUs, covering **32 model topologies**
- Corrected **prefill FLOPs accounting** for cached prefixes and batched requests, eliminating 10–100× **underestimation** in long-context workloads and restoring meaningful GPU throughput reporting
- Separated **readiness and liveness probes**, preventing premature traffic routing during server warmup and documenting Kubernetes health-check configuration

Berkeley Artificial Intelligence Research (BAIR)

Berkeley, CA

Undergraduate Researcher

Sep 2024 – Dec 2024

- Contributed data-loading and benchmarking software to **Robo-DM**, an open-source robotics data management toolkit; co-authored the **Robo-DM CoRL 2024 Workshop paper**
- Engineered **multiprocessing** data loading pipelines with chunked decoding, reducing dataset load times from **5 minutes to under 30 seconds** for ML training workloads
- Developed frame-level **random slicing strategies** for efficient trajectory sampling, cutting average data retrieval latency by **70%**

1stCollab (YC W23)

San Francisco, CA

Software Development Engineering Intern

May 2024 – Aug 2024

- Built a **Python/Playwright ETL pipeline** across 8+ platforms, supplying creator data for **ML training and campaign matching** that powered advertiser recommendations and automated email outreach
- Used multithreaded Docker workers on **Compute Engine** to scrape into **Cloud Storage**, dispatch extraction through **Pub/Sub**, and parse and normalize creator data into **Cloud SQL/PostgreSQL**
- Processed **2K–10K creators/day** from a corpus of **2.5M+ profiles**, achieving **90% recall and 95% precision** against the company's manually curated reference database

TECHNICAL SKILLS

Languages: Python, C/C++, Java, Go, SQL, HTML, CSS

Technologies: AWS (Lambda, SQS, RDS, S3, ElastiCache, DynamoDB, CloudWatch), Docker, Redis, Kafka, HBase, MySQL, PostgreSQL, Spring Boot, FastAPI, Selenium, LangChain, PyTorch, SGLang, vLLM

Concepts: Distributed Systems, Microservices, LLM Inference, Operating Systems, CI/CD, Agile, Networking, Caching, Performance Tuning, Cloud Computing, Serverless, Consensus, DevOps