Sirui (Ray) Li

github.com/LiSiruiRay industry.slray.com blog.slray.com

Education

New York University Sept 2022 – Aug 2026
B.A. in Mathematics & Computer Science *New York, NY*

- **Relevant Coursework:** Algorithms, ML, OS, Compiler Construction (PhD), Abstract Interpretation (PhD), Theory of Computation, Real Variables (PhD), Differential Geometry, Complex Analysis, Real Analysis (Honors), Abstract Algebra (Honor), ODE, Topology (PhD), Computability & Incompleteness, Computer Security, OS
- **Awards:** Dean's List (All Semesters), Courant Scholarship ($160K)

Work Experience

Software Development Engineer Incoming
Tiktok | ***Website*** *San Jose, CA*

- Incoming Full Time, ng 2026

Software Development Engineer May 2025 – Aug 2025
Tiktok | ***Website*** | *Golang, RPC, Python, AIGC, MCP* *San Jose, CA*

- Built **ReAct (Reasoning + Acting)** agent framework for processing the whole lifecycle of chargeback cases, increasing auto case creation from **12% to 75%**, able to process **8k** cases per year.
- Integrated **LangGraph** with multiple **MCPs** (data/file regulation, channel interaction) and **Graph RAG knowledge base**, lowering cost by **20** times, able to auto defense **9000 cases** per year, increased coverage by **4.5%**.
- Developed both HTTP and RPC (Thrift) APIs for seamless backend integration.

Chief AI / Co-Founder May 2025 – Present
Auray | ***Website*** | *Python, Langgraph, TypeScript, AIGC* *San Jose, CA*

- Built **Multi-agent** financial advisor with **Langgraph**, **OpenAI API**. Interviewd by **Y Combinator** (top 1%).
- Integrated **Intent Recognition**, **Market News Process**, user-bank-info reading system, **Web Search**, with human in the loop approval system, increasing accuracy by **52.1%** and increased user trust by **80.9%**.
- Enabled **Streaming** style token level reponse and explicit thinking process, reducing response time by **33.5%** and increasing user satisfaction by **81.2%**.

Software Development Engineer Feb 2025 – Apr 2025
Jupiter Plans (Medical Service Agent) | ***Website*** | *Python, AIGC, MCP, AWS* *New York, NY*

- From 0 to 1, developed a production-grade medical chatbot using **OpenAI GPT-4, function calling**, and a hybrid **retrieval** system (semantic + keyword) backed by PostgreSQL and Qdrant vector database, serving more than **10,000 users**, conneted more than **500 hospitals**.
- Replaced third-party search with in-house service-matching API, enabling fuzzy keyword resolution and improving retrieval accuracy for complex queries by **50%** and lower the cost for **80%**.
- Implemented structured output pipelines, dynamic chatflow routing, and backend integrations to support intelligent service discovery, triage classification, and location-aware results.
- Migrated infrastructure to Dify Cloud, resolved frontend and document duplication bugs, and contributed to open-source enhancements on the Dify GitHub repository.

Software Development Engineer – Trading Infrastructure Jun 2023 – Sep 2023
300K.xyz (High-Frequency Cryptocurrency Exchange) | ***Github*** *Austin, Taxes*

- Designed and implemented a data validation framework (Python) to ensure data integrity for ML pipelines, reducing training errors by validating date continuity, detecting outliers, and enforcing schema compliance across **250+** datasets.
- Architected a real-time prediction system using PyTorch, Docker, and message brokers (STOMP), enabling live inference for crypto swap markets with **less than 500ms** latency; deployed via Jenkins.
- Developed automated configuration workflows to generate dynamic training datasets from top-performing trading pairs, integrating MongoDB for version control and reducing manual setup time by **70%**.
- Built data quality monitoring tools to flag anomalies in live market feeds, improving model accuracy by **15%** through outlier detection in swap price/trade signals.
- Built Python SDK for cryptocurrency trading; supports transaction processing and liquidity management

Co-Founder & Chief AI Jan 2023 – Sep 2023
Capybara AI (AI for Finance) | ***Website*** *New York, NY*

- Secured acceptance into Microsoft Startup Program (**1%** acceptance rate), receiving **$18K** in cloud and AI credits (Azure, OpenAI); also selected for NYU Startup Bootcamp
- Launched financial services app from 0 to 1 providing market data and insights to **9,000+** users worldwide; engineered system to detect stock price fluctuations, retrieve relevant news, and cluster market-moving events
- Developed RESTful APIs for stock data querying, using PostgreSQL (historical storage) and Redis (caching) to reduce query time from minutes to milliseconds (**1,000×** speedup)

Research

Formal Verification Research Jul 2023 – Aug 2025
New York University (Under Prof. Thomas Wies) *New York, NY*

- Investigating formal semantics (operational, denotational) and Hoare logic for concurrency.
- Contributing to the Raven (Link) project for program verification and refinement proofs.
- Systematically studied OCaml, Coq, and lean

AI Research Assistant, AI4Math Jul 2024 – Aug 2024
Peking University (Under Prof. Zaiwen Wen) | ***Github*** *Beijing, China*

- Selected as one of 20 participants nationwide (<1% acceptance rate) for fully-funded mathematical research program valued at $\tilde{\$}$22,200 USD (160,000 RMB); conducted research in machine learning (NLP), category theory, etc.
- Formalized ADMM algorithm in convex analysis using Lean Theorem Prover

Machine Learning Research Assistant Jan 2024 – Aug 2024
NYU Abu Dhabi (Under Prof. Djellel Difallah) | ***Github*** *Abu Dhabi, UAE*

- Explored transformer-based models for non-stationary time series forecasting; built interactive visualization tool for time series data, enabling predictions across multiple benchmark datasets (weather, ETT)
- Developed Python scripts to automate HPC connections and model downloads, reducing manual setup time from approx. 20 minutes per model to near-instant automation, increasing research productivity by 150%

Projects

Compiler Design Project | *OCaml, RISC-V, LLVM, Git, ANTLR, Dataflow Analysis* | ***Github*** Sep 2023

- Engineered a multi-phase compiler in OCaml, implementing lexer/parser generators, intermediate representations (C/ML/Scheme), and code generation for RISC-V assembly.
- Optimized output via SSA-based register allocation and control-flow graph analysis, achieving 22% fewer execution cycles in benchmark tests.
- Collaborated on a performance-tuning module (Agile team of 3) using peephole optimizations and static single assignment, outperforming baseline compilers by 35%.
- Validated correctness with a RISC-V ISA simulator, supporting arithmetic, memory ops, and syscall emulation across 50+ test cases.

EXplain Video Agent | *JavaScript, GPT-3.5, AWS EC2* | ***Github*** Sep 2023

- Developed Chrome extension for transcript analysis using GPT-3.5 and `youtube_transcript_api`
- Implemented binary search with OpenAI API for context-aware video search; built RESTful API backend on AWS EC2

TripGenie Travel AI Agent | *Flask, ChatGPT API* | ***Github*** May 2023

- Developed an AI travel agent in Flask, integrating the ChatGPT API to generate personalized itineraries
- Automated itinerary generation with cost estimation, dynamically adjusting based on user preferences

Meal Pair Platform | *Java, Thymeleaf, MySQL, Spring* | ***Github*** Dec 2022 – Jan 2023

- Created a full-stack social platform for dinner party scheduling with event posting features
- Integrated Google Maps API and adopted MVC architecture to improve scalability and user experience

Competitions & Awards

NYU Hackathon (2022): Awarded "Best First-Time Hacker" out of 800+ participants, winning $200

Technical Skills

Languages: Python, Scala, OCaml, C++, Java, JavaScript, SQL, Lean, Coq
Tools/Tech: Docker, Jenkins, Redis, PostgreSQL, AWS (EC2), Git, Lean
Frameworks: Flask, Django, Time-Series ML (Autoformer), React, Node.js