

Zhiyuan(Allen) Song

+1 (434) 466-0096 | realoulasong@gmail.com | www.zhiyuansong.com

EDUCATION

University of Pennsylvania

Philadelphia, PA

M.S. in Computer and Information Science (CIS) - AI Track

Aug. 2025 – May 2027

- Coursework: Advanced Machine Learning, Computer Vision, Artificial Intelligence, Software Systems.

University of Virginia

Charlottesville, VA

B.A., Double Major in Computer Science and Applied Statistics

Aug. 2021 – Dec. 2024

- GPA: 3.91/4.00 | Dean's List, NSF Fund | TA: CS 4774 Machine Learning, CS 2120 Discrete Math & Theory.

PUBLICATIONS

Z. Song, et al., "Learning When to Think While Listening in Large Audio-Language Models," *Under review*, **NeurIPS 2026**. arXiv [↗](#) Code [↗](#)

Z. Song, Z. Mei, L. Li, and N. Nguyen, "SL4HOI: Supervised Learning for Predicting Health Opportunity Index across States," in **Knowledge Discovery and Data Mining (KDD)** Conference, 2024, Barcelona. [↗](#)

PROFESSIONAL & RESEARCH EXPERIENCES

Millennium Management LLC

Jun. 2026 – Aug. 2026

AI Engineer Intern - Cloud & AI/ML Infrastructure

New York, NY

- Designed and shipped an in-house **multi-agent code-review platform** (*Go*, *PostgreSQL*, *Kubernetes*) for a **6,700+**-employee firm: a **planner-routed pipeline** of specialist review agents grounded in **pre-indexed repo RAG**, with synthesis and **LLM-as-judge** adjudication, exposed via three surfaces—**MCP tool**, GitHub PR comments, and **pre-commit** review.
- Engineered the production **agent harness**: DB-leased execution with retry/failure recovery, **delta-only re-review**, bidirectional GitHub sync, **tiered model routing** (lightweight scans → frontier models on critical paths) with **budget-aware spend gates**, and per-specialist **token-cost & latency** tracking.
- Built a React/TypeScript **live review dashboard** streaming each specialist's reasoning trace in real time (SSE), with **multi-turn follow-up Q&A** anchored to findings and a full finding lifecycle (dismiss / acknowledge / auto-resolve on fix); deployed to production as the team's first milestone.

University of Pennsylvania

Aug. 2025 – Present

Deep Learning Researcher — Advisor: Prof. Jiatao Gu

Philadelphia, PA

- Post-trained **Qwen2.5-Omni-7B** (7B omni-modal LLM) with a **text-LLM-style SFT** → **RL** post-training stack (recipe transferable to pure text LLMs), learning a streaming **wait-think-answer controller** that decides under partial audio when to **wait**, emit a compact **think** update, or commit to an **answer**.
- Built a two-stage post-training pipeline (**SFT** with **LoRA** via **MS-Swift** on 73K traces → from-scratch **DAPO** trainer over 37K verifiable rollouts) handling live streaming rollouts, action parsing, dynamic resampling, KL regularization, and in-runtime adapter updates; scaled with **DeepSpeed ZeRO-3** on 4×**B200** GPUs.
- Designed a **6-reward** trajectory objective—answer correctness, format gate, residual-latency, update timing, thought quality, chain consistency—lifting row-weighted accuracy from **67.6%** → **70.3%** on a 6-task spoken-reasoning benchmark and cutting post-endpoint final-think length by **14%**.

Legal Manus

May 2025 – Aug. 2025

Machine Learning Engineer

Ningbo, China

- Led a team of 5 to build a **retrieval-augmented generation (RAG)** pipeline (*FastAPI*, *LangChain*, *ChromaDB*) for U.S. & Chinese legal Q&A grounded in official legal documents; containerized with Docker and shipped CI/CD to **Google Cloud Run**, later migrated to **AWS EC2** with Sentry error tracking.
- Built a vector database of **600K+** legal chunks via OpenAI text-embedding-3-small, then applied **BGE reranker** (CrossEncoder) to lift Top-3 retrieval accuracy from **68%** to **98%** on a 200-question test set.

TECHNICAL SKILLS

Programming Languages: Python, C/C++, Java, Go, JavaScript/TypeScript, SQL, R, Bash, Assembly

Large Language Model (LLM) Training & Post-training: PyTorch, Transformer, Distributed Training (DeepSpeed ZeRO-3, Megatron-LM), MS-Swift, verl, LoRA/PEFT, Supervised Fine-Tuning (SFT), Reinforcement Learning (RLHF/RLAIF, GRPO/DAPO), FlashAttention, CUDA, Mixed Precision (BF16/FP16), Wandb

Agentic Systems, Inference & Multimodal: Multi-Agent Orchestration, LLM-as-Judge, Agent Evaluation & Benchmarking, MCP, Skills, Temporal, LangChain/LangGraph, LangSmith, Logfire, vLLM (*vllm-metal contributor*), SGLang, llama.cpp, KV Cache, HuggingFace Transformers, ModelScope, Whisper, VAD/ASR/TTS

Cloud & Backend Infra: AWS (EC2, Bedrock, S3), GCP (Cloud Run), Azure, Slurm, Linux, Docker, Kubernetes (Helm), CI/CD (GitHub Actions), Git, gRPC, FastAPI, Next.js/React, PostgreSQL, MySQL, ChromaDB, Redis