Sardor Nodirov

Language AI Systems · Retrieval-Augmented Generation · Efficient On-Device Inference
(207) 616-7032 | nodirov.com | sardor@nodirov.com | linkedin.com/in/sardornodirov

Research

Lapwing – Real-Time Multi-Modal AI System with Language-Driven Reasoning Sep 2025 – Present
Honors Thesis | Colby College, Davis Institute for AI | Live Demo: lapwing.live

- Took a CVPR 2024 paper (MoMask) from offline inference code to a **deployed, interactive AI system** – orchestrating **5 concurrent AI subsystems** (LLM language generation, TTS, generative motion, neural expression, lip sync) into a synchronized WebSocket-streamed experience at 60 FPS with **sub-2s end-to-end latency**.
- Engineered a **five-way parallel orchestration pipeline** (asyncio) firing dual LLM calls for language generation and action planning simultaneously, cascading into downstream generators on completion. Designed the **agentic reasoning layer** with persona-aware language understanding and contextual expression tagging.
- Identified a **redundant CLIP text-conditioning load** in MoMask's dual-transformer architecture via systematic profiling. Sharing a single ViT-B/32 instance **saved ∼400MB VRAM and ∼6s cold-start** – a **memory-constrained inference optimization** applicable to any dual-stage architecture with shared text conditioning.
- Deployed on GPU cloud with **baked model checkpoints into Docker images** (eliminating 6–12s network fetch on cold start), optimized VRAM allocation across A10G + T4 instances, and implemented **dual-mode streaming inference** (progressive vs. chunk-based) to study latency-quality tradeoffs under **computation- and memory-aware constraints**.

Experience

Davis Institute for Artificial Intelligence | Research Poster Sep 2023 – May 2024
AI Research Assistant – Real-Time Conversational Language Agent *Waterville, ME*

- Built a **real-time conversational AI agent** with natural language understanding over Twilio telephony, achieving **sub-2s end-to-end latency** via async FastAPI WebSockets with streaming TTS and Azure Speech STT. Engineered a **barge-in protocol** for natural mid-utterance interruption over raw mulaw/8000Hz audio.
- Designed a **retrieval-augmented generation (RAG) system over 800+ pages** using Scrapy, Qdrant vector store, and Cohere embeddings for **knowledge-augmented semantic retrieval**. Built a **hybrid inference router** (GPT-4 Turbo + Groq/Llama 3) achieving **300+ tok/sec** balancing quality against latency constraints.

Ember AI May 2024 – March 2025
Founding Full-Stack AI Engineer | Early-Stage Startup *San Francisco, CA (Remote)*

- Built the core product end-to-end as founding engineer: a **multi-modal AI pipeline** using Vision Transformers and LLMs to generate personalized experiences with synchronized audio-visual playback. Architected async backend with **Redis** queues handling 30+s generation jobs and a real-time engine achieving **<50ms sync** across distributed clients.

Projects

Cascade – Open-Source Agentic AI Research Platform | GitHub Code Feb 2026

- Built a 5,500-line **multi-agent AI system** integrating 5 APIs (arXiv, Semantic Scholar, GitHub, OpenAI, Anthropic) with **hybrid SQLite + ChromaDB vector storage** and a **RAG pipeline** using OpenAI text-embedding-3-large (3,072-dim) embeddings with cosine HNSW indexing for knowledge-augmented semantic retrieval.
- Engineered **agentic reasoning and planning**: provider-agnostic LLM abstraction with hot-swappable OpenAI/Claude backends, **dual-LLM feedback synthesis**, intelligent token budget management (6K–150K context windows), and **adaptive memory** with automatic dual-write to relational and vector stores and sliding-window session context tracking.

Education

Colby College Waterville, ME
B.A. Honors in Computer Science, AI Concentration *Expected May 2026*

- Presidential Scholar | Program GPA: 3.77 | Dean's List | President of Colby AI Society
- Coursework: Bio-Inspired Neural Networks, Natural Language Processing, Data Analysis & Visualization, Computer Architecture, DSA, AI Ethics

The University of Edinburgh Edinburgh, Scotland
Study Abroad – School of Informatics & Electronics/Electrical Engineering *Jan 2025 – May 2025*

- Coursework: Natural Language Processing, Quantum Computing, **Embedded Systems**, Digital Design

Honors & Awards

Presidential Scholar, Colby College: Awarded to the selective cohort of top incoming students with access to research centers, labs, and project grants.
The Gold Medal for Academic Excellence: National distinction awarded to the top ∼1% of high school graduates nationwide.

Technical Skills

Languages: Python (asyncio, multiprocessing), Java, TypeScript, C
AI / ML: PyTorch, TensorFlow, CLIP, VQ-VAE, Transformer Architectures, RAG, Embeddings (OpenAI, Cohere), LLM Orchestration, Whisper, Model Profiling & Memory Optimization
Retrieval & Data: ChromaDB, Qdrant, SQLite, Redis, Scrapy, Data Cleaning & Visualization
Infrastructure: FastAPI, WebSockets, Docker, Modal (GPU Cloud), Cloudflare R2/Workers, Git, Linux