Shounak Ray

Stanford, CA • shounak@stanford.edu
shounakray.github.io • linkedin.com/in/rayshounak • github.com/ShounakRay • shounakray.github.io/googlescholar

Check out my website above, it has a lot more than I could fit here!

Education

Stanford University **Stanford, CA**
Core: Bachelor in Computer Science (BSc. CS) + Masters in Computer Science (MSCS) *2021 – Pres.*
Focus: Artificial Intelligence and Computer Systems
Graduate study: fully-funded, working w. Eran Bendavid & Carlos Guestrin on agentic systems to automate research
Cool classes: operating systems, parallel computing, ML/NLP/CV/RL, self-improving agents, applied physics
Some tools: PyTorch, Perfetto, Kubernetes, Python, C++/C, Cuda, Jax, S3/EC2/GCP, React, Typescript, Vercel

Work Experience

Baseten **San Francisco, CA**
Software Engineer Intern – Model Performance, Kernels *June – Sept.'24*

- **Profiled** H100/B200 video-gen. inference; wrote scripts to analyze kernels and map hot ones back to PyTorch code, pinpointing bottlenecks and guiding hardware-aware optimizations and fusion opportunities.
- **Reduced cold start times** by over 4 min. and compilation latencies by 30s via specific torch.compile + caching optms.
- **Created b10-transfer**: an environment-aware PyPI package that manages distributed caching of both torch.compile artifacts and model weights, minimizing start-up times by 75-to-95% (dep. on model). **Published** on company website.
- **Built and deployed** a node warming system in Kubernetes that minimizes cold-start latency by prefetching model weights to nodes with hardware-aware filtering, disk space monitoring, and rollback/forwarding support.
- Used: Perfetto, Docker, Kubernetes, PyTorch (torch compile + torch profiling), hf-transfer/hf-xet/hf-api

Stealth Startup **Palo Alto, CA**
Software Engineer – AI *June – Sept. '24*

- **Engineered** end-to-end agentic and precision-critical LLM pipelines, embedders, clusterers, frontend developed to find insights and risk in massive set of documents.
- **Developed** abstractions capable of extracting knowledge from any structured database (not RAG).
- **Founded** AIxGood initiative; developed full-stack system that extracts knowledge, embeds, clusters, and identifies bias across 50K documents for a **criminal justice case**.
- Used: Huggingface Transformers, Python, React, Typescript, FastAPI, Vercel, Tailwind CSS, Next.JS

Intelligent Systems Laboratory (SISL). **Stanford, CA**
Research Asst., Aero. Astro. Dept. *May '23 – May '24*

- **Architected** framework to generate OOD driving datasets utilizing semi-sup. latent diffusion embeddings. "Hacked" Waymo's open-source Waymax package to adapt to our research purposes (deep-RL on social vecs.).

Changing Cities Research Lab (CCRL) **Stanford, CA**
Research Asst., Sociology Dept. *Sept. '22 – Jun '25*

- **Engineered** end-to-end CV pipeline (data preprocessing to modeling) involving CNNs and ViTs to parse 400GB+ Google Street View images and track gentrification. Achieved SOTA Pr./Re./etc.

Used: Jax, PyTorch, Python, AWS S3, EC2, GCP, WandB, ViTs, CNNs, latent diffusion

Relativity-Text IQ (acquired) **Seattle, WA**
Product Management Intern – AI Team *June – Sept. '22*

- **Conducted** market sizing & competitive analysis for self-proposed product, secured leadership support
- **Prototyped** NLP clustering pipeline (able to rapidly cluster terabytes of documents for M&A due diligence and litigation events) and product UI, set to save millions of dollars (est.) in operational costs
- **Utilized** AutoML to reduce project costs by 20%, helped secure prospects (e.g. AstraZeneca) for pilots
- Used: Figma, Python, Confluence, AutoML, scikit-learn

White Whale Analytics **Calgary, AB**
Data Science Intern *July '20 – Aug '21*

- **Secured** multiple long-term energy and hospitality contracts and maintained strong client relationships
- **Programmed** scalable energy-optimization algorithms for energy clients, resulting in est. 30% savings
- **Developed** network graph, anomaly-detection, and ML solutions for diverse Canadian clients
- Used: Python, H2O.ai, scikit-learn, changepoint detection, statistical modeling, feature engineering

Milestones

Ex-teaching assistant for CS 41 at Stanford Univ.
Developing Context-Adaptive AI/RL Framework
Developing Emotive ASL Translation System
More cool projects: **check them out**!

Awards and Recognition

Team Canada: Regeneron-ISEF
TreeHacks, Winner: VMWare Award
TreeHacks, Winner: Education Grand Award
Canada-Wide Sci. Fair, Grand Award Winner