Rajath Rao

US Citizen

🌐 https://jarhatz.github.io/rajath-rao/

☎ +1 (408)-300-4564
✉ rajathrao1001@gmail.com
https://www.linkedin.com/in/rajath-rao/
https://github.com/Jarhatz
https://scholar.google.com/citations

EDUCATION

M.S. Applied Mathematics & Data Science at Stony Brook University, *New York,* **GPA: 3.9** Aug 2023 - Dec 2024

B.S. Computer Science & Engineering at University of California - Irvine, *California* Sep 2019 - Mar 2023

PAPERS

[1] **Rao, R.**, Ganesan, A., Kjell, O., Luby, J., Raghavan, A., Feltman, S., Ringwald, W., Boyd R., Luft, B., Ruggero, C., Ryant, N., Kotov, R., & Schwartz, H. (2025). WhiSPA: Semantically and Psychologically Aligned Whisper with Self-Supervised Contrastive and Student-Teacher Learning. ***Published at ACL 2025.*** **LINK**

[2] Gu, Z., **Rao, R.**, Nillson, A., Eijsbroek, V., Kjell, O. (2025). Probed Language-Based Assessments Using Spoken Language. ***Accepted at IEEE ProComm 2025.***

[3] Varadarajan, V., Lahnala, A., Ganesan, A., Dey, G., Mangalik, S., Bucur, A.M., Soni, N., **Rao, R.**, Lanning, K., Vallejo, I., Flek, L., Schwartz, H., Welch, C., & Boyd, R. (2024). Archetypes and Entropy: Theory-Driven Extraction of Evidence for Suicide Risk. ***Published at ACL 2024.*** **LINK**

[4] **Rao, R**., Zhou, L., Samaras, D. (2024). BrainDiffusion: Reconstructing Visual Semantics from Non-Invasive Neural Activity Readings. **Preprint**.

RESEARCH EXPERIENCE

Graduate Research Assistant, Stony Brook University, *Stony Brook, NY* Aug 2023 - *Present*

Human Language Analysis Beings (HLAB) *Advised by Dr. H. Andrew Schwartz*

- Developed WhiSPA, a speech encoder, to address the semantic and psychological representation gap between speech-based LMs and text-based LMs for psychological and mental health assessments.
- Proposed novel cross-modal alignment objective to fuse acoustic/visual/lexical modalities for rich representations for downstream tasks.
- Integrated psycholinguistic and neuroscientific insights with existing foundation model architectures for enhanced emotional assessments.
- Conducted in-depth quantitative analyses to interpret how models integrate multimodal cues, revealing representation patterns.
- **Utilized:** *PyTorch, NLP/CV, HuggingFace, CUDA, NCCL, NLTK/DLATK, SQL, Linux, Academic Writing*

PROFESSIONAL EXPERIENCE

Platform Architecture ML Engineer Intern, SiMa.ai, *San Jose, CA* Jan 2025 - *Present*

- (Graph Surgery) Translated unsupported operators from FP32 transformer architectures to INT8 to run on MLA.
- Conducted distillation research for PaliGemma and LLAVA to reduce **TTFT < 1** second and delivered POC to Porsche. **4B parameter reduction** with QLORA. **60% speed-up** in throughput.
- Ported Pi-Zero Vision/Language/Action (VLA) model to BF16 on MLA for high frequency robotic arm control.
- Leveraged quantization schemas for retaining model accuracy during compilation for MLA inference workloads.
- Collaborated with cross-functional (hardware/software) teams to integrate AI solutions into the SoC.
- Developed end-to-end inference pipeline for real-time anomaly detection for Mercedes Benz vehicle cameras.
- **Utilized:** *Python, C++, PyTorch Lightning, TFLite, ONNX, TensorBoard, Quantization, Linux*

AI / Deep Learning Engineer Intern, Intel, *Hillsboro, OR* May 2024 - Aug 2024

- Profiled and enabled Microsoft's state-of-the-art SLM, Phi3, to run on Intel Habana's Gaudi2 hardware accelerators
  - Implemented graph compilation (lazy execution) for 40% increase in inference throughput
  - Developed memory efficient optimization of **reusing KV cache** matrices for up to 80% HPU utilization
- Conducted experiments and highlighted performance regression on Intel Habana's transformers with DeepSpeed
- Benchmarked and optimized *Llama, Mixtral, StarCoder*, etc. with 15% higher GPU utilization on Gaudi2 using profiler kernel trace
- Leveraged vLLM for continuous batching on quantized models running in mixed precision for greater throughput.
- **Utilized:** *Python, C++, PyTorch, HuggingFace, TensorBoard, DeepSpeed, CUDA, SYCL, H/NCCL, Linux*

HPC Software Engineer Co-Op, Intel, *Santa Clara, CA* Jun 2021 - Apr 2023

- ❖ Proposed and developed models for failure prevention, pattern recognition, data automation for HPC data centers
- ❖ Spearheaded development of an ensemble of neural networks for regressing hard drive failures up to 4 months prior
  - ➢ Trained ensemble notifies of imminent failures with a 92% accuracy on test environment servers
  - ➢ Saved terabytes of potential data loss alerting maintenance teams to backup hard drives
- ❖ **Utilized:** *C++, Python, Machine Learning, REST APIs, MongoDB, Docker, Linux*

SKILLS

Languages: *Python, C/C++/C#, Java/JavaScript/TypeScript, HTML/CSS, Rust, R, SQL, LISP, Perl, Verilog, Linux*
Skills: *PyTorch, CUDA, HuggingFace, NCCL, AWS, Azure, Google Cloud, Apache/Hadoop, MPI, React/Node.js, Docker, Kubernetes, Git, JIRA*
Courses: *Reinforcement Learning, Computer Vision, NLP, ML, Statistical Computing, Probability, Algorithms & Data Structures, Time Series Analysis, Big Data Systems, Operating Systems, Embedded Systems Design, Computer Arch.*

ACADEMIC REPORTS & PERSONAL PROJECTS

BrainDiffusion: Generate Images with Your Mind, *Stony Brook University,* Research Report Feb 2024 - Oct 2024

- ❖ Proposed and developed a novel self-supervised contrastive learning framework to align EEG embeddings with CLIP image embeddings
- ❖ Validated and fine-tuned Stable Diffusion 1.5 on CLIP aligned EEG embeddings to generate images corresponding to respective EEG signals
- ❖ Implemented a Masked Autoencoder (MAe) to extract latent representations for EEG signals (EEG encoder)
- ❖ **Utilized:** Python, *PyTorch, HuggingFace, Computer Vision, CUDA, NCCL, Git*

Thought-to-Text: Generate Text With Your Mind, *Stony Brook University,* Research Report Jan 2024 - Jun 2024

- ❖ Developed an Encoder-Decoder framework using fMRI stimulus responses from 6 subjects as they perceived speech from audio books
- ❖ Spearheaded development of predicting fMRI voxel-wise BOLD features using stimulus matrix from GPT2 logits
- ❖ Implemented a candidate sequence token prediction algorithm with Beam Search and nucleus sampling
- ❖ **Utilized:** Python, *PyTorch, HuggingFace, NLP, CUDA, NCCL, Git*

Agentic RAG Stock Analysis and Price Forecasting, *HackAI Dell & NVIDIA Challenge* Sep 2024 - Oct 2024

- ❖ Awarded **Top 15 Use of Generative AI** at HackAI Dell & NVIDIA Challenge competition
- ❖ Implemented a robust Agentic-RAG framework for portfolio asset management and financial question answering, utilizing knowledge graphs to effectively source and analyze popular stock news and articles
- ❖ Enhanced TimeSeriesTransformer performance for S&P500 stocks forecasting through strategic implementation of automatic mixed precision, optimizing training efficiency without compromising model accuracy
  - ➢ Developed an efficient sliding window algorithm for batching the data during tensor preprocessing
- ❖ Developed a full-stack web application for concurrent users with React (front-end) & Flask REST API (back-end)
- ❖ **Utilized:** *Python, NVIDIA AI Workbench, CUDA, HuggingFace, Langchain, React.js, Docker, Git, Devpost, Video*

Autonomous IoT Shopping Cart with Intelligent Tracking, *UC - Irvine* Oct 2022 - Mar 2023

- ❖ Spearheaded the development of an IoT based autonomous shopping cart with user-following, lane-correction, object-collision, and product-search features
- ❖ Constructed a Mealy Finite State Machine for the autonomous drive-state decision making process of the cart
- ❖ Devised a WiFi triangulation algorithm using 3-point trilateration of Received-Signal-Strength-Indicator (RSSI)
- ❖ **Utilized:** *Arduino, ESP8266, C/C++/C#, Python, Scikit-Learn, Firebase, MQTT Broker, Mealy Machine (FSM), Git*