

## Rajath Rao

+1 408-300-4564 • [rajath.rao@stonybrook.edu](mailto:rajath.rao@stonybrook.edu) • [Website](#) • [LinkedIn](#) • US Citizen

### EDUCATION

- M.S. Data Science** Stony Brook University, New York Aug 2023 - Dec 2024  
❖ GPA: 3.75
- B.S. Computer Science & Engineering** University of California - Irvine, California Sep 2019 - Mar 2023

### PROFESSIONAL EXPERIENCE

- 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 compute graphs compilation for 40% increase in LM inference throughput  
➤ Developed memory efficient optimization of reusing Key/Value cache matrices for 80% HPU utilization  
❖ Conducted experiments and discovered performance regression on Intel Habana's HF transformers library with multi-GPU (DeepSpeed) inferencing  
❖ Benchmarked and optimized Llama3\_70B, Mixtral\_8\_7b customer tickets for 15% higher GPU utilization on Gaudi2 using profiler trace analysis on kernel operations  
❖ Implemented preprocessing collate function for batch tensor truncation leveraging attention masks for faster tokenization and efficient caching during training  
❖ **Utilized:** Python, PyTorch, HuggingFace, TensorBoard, DeepSpeed, SYCL, HCCL, Linux
- Graduate Student Researcher, HLAB at Stony Brook University, Stony Brook, NY** Aug 2023 - Present  
❖ Currently researching multimodal VLMs for mental health illness detection, integrating psycholinguistic and neuroscientific insights for enhanced mental health assessments  
❖ Developed an "Affect Transformer" model grounding Whisper and RoBERTa embeddings to detect emotions in voice, achieving high accuracy in emotion recognition  
❖ Discovered a novel approach for PTSD symptom severity prediction using multimodal data integrating human-interpretable acoustic features with learned transformer embeddings  
❖ **Utilized:** PyTorch, NLP/CV, HuggingFace, CUDA, NCCL, NLTK/DLTK, SQL, Linux, Academic Writing
- HPC Software Engineer Intern, Intel, Santa Clara, CA** Jun 2021 - Apr 2023  
❖ Proposed and developed machine learning models for failure prevention, data analysis, pattern recognition, and automation scripting for Intel's HPC data centers  
❖ Spearheaded development of an ensemble of neural networks for predicting hard drive failures up to 4 months before they occur and the approximate number of days till imminent failures  
➤ Trained ensemble notifies of imminent failures up to 4 months in advance with a 97% accuracy on test environment servers saving terabytes of data loss alerting maintenance teams for hard drive backups  
❖ **Utilized:** Python, Machine Learning, Feature Engineering, Jupyter, Sci-Kit Learn, Flask, MongoDB, Docker

### SKILLS

**Languages:** Python, Java/JavaScript/TypeScript, C/C++/C#, HTML/CSS, R, SQL, LISP, Perl, Verilog, Linux  
**Skills:** Deep Learning, AI/ML, Optimization, Computer Vision, NLP, PyTorch, CUDA, HuggingFace, Data Analysis, AWS, Azure, Google Cloud, MPI, Apache Spark, React/Node.js, Docker, Kubernetes, Git, JIRA, Confluence

### PAPERS

- BrainDiffusion: Generate Images with Your Mind, [Research Paper](#)** Jan 2024 - Jun 2024  
❖ Developed a Masked Autoencoder (MAe) to extract latent feature representations for EEG signals  
❖ Proposed and developed a novel self-supervised contrastive learning framework to align EEG embeddings with CLIP image embeddings  
❖ Validated and fine-tuned Stable Diffusion on CLIP aligned EEG embeddings to generate images corresponding to respective EEG signals  
❖ **Utilized:** Python, PyTorch, HuggingFace, Computer Vision, CUDA, NCCL, [Git](#)
- Thought-to-Text: Reconstructing Semantic Language from Neural Activity, [Research Paper](#)** Feb 2024 - Present

- ❖ 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
- ❖ Proposed a candidate sequence token prediction algorithm using a variant of Beam Search with nucleus sampling
- ❖ **Utilized:** Python, PyTorch, HuggingFace, NLP, CUDA, NCCL, [Git](#)

## PROJECTS

---

### **Real-Time Location Services (RTLS) via Bluetooth (BLE), Consulting**

*Apr 2023 - Aug 2023*

- ❖ IoT solution with BLE anchors/tags for RTLS and resident-monitoring currently in use at [Roseleaf Senior Care](#)
- ❖ Leveraged AWS IoT/DynamoDB to route Received-Signal-Strength-Indicator (RSSI) packets from tags to anchors
- ❖ Developed AWS Lambda functions to triangulate locations using 3-point trilateration based on RSSI-distance values estimated with a logistic regression model
- ❖ Designed a mobile application with React Native for an organized display of resident data and real-time locations
- ❖ **Utilized:** AWS, React/Node.js, Python, IoT, BLE, [Git](#), [Devpost](#)

### **B.S. Capstone: 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
- ❖ Created the Mealy Machine (FSM) structure 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](#)