

ILEANA RUGINA

(617)655 4815 | irugina1.github.io | ileana.rugina.2@gmail.com

WORK AND RESEARCH EXPERIENCE

Software Engineer at Emerald Innovations

Cambridge, MA

Oct. 2023 - Jun. 2025

Developed end-to-end ML systems that extract clinically relevant biomarkers from RF signals.

- *Cut inference cost and latency:* shifted ML workloads to spot instances, refined the job scheduler, and re-indexed / re-designed SQL schemas; drove additional savings by migrating to multi-tenant clusters and consolidating networking cloud infrastructure. Reduced cloud costs by 70% and some latencies from hours to minutes.
- *Managed asynchronous uploads:* rolled out a manifest service that tracks data uploaded by IoT sensor files.
- *Other Responsibilities:* production deployments, ops for IOT devices

Machine Learning Software Engineer at Gridspace

Los Angeles, CA

Sep. 2021 - Oct. 2023

Developed speech and language AI systems at conversational AI startup (Stanford/SRI Speech Labs collaboration)

- *Speech AI:* Developed and deployed a speech-synthesis model with controllable prosody via variational embeddings. Improved a dual-mode Conformer with joint CTC / RNN-T decoding, eliminating $\propto 85\%$ of hallucinations.
- *LLM Inference:* Implemented speculative sampling, cutting encoder-decoder summarization latency by up to 40%.
- *Dialog Systems:* Improved planning algorithm and integrated vector embeddings in dialog system.
- *Platform:* Coordinated data-collection with in-house and third-party teams for audio transcription and LLM instruction tuning. Optimized distributed systems and anomaly detection to boost analytics reliability and accuracy.
- *Other Responsibilities:* production deployments, mentored interns, polished demos, recruiting.

MEng Research Assistant at MIT Soljačić Lab

Jul. 2019 - Jun. 2021

- Designed a data-informed task-agnostic attention pruning method for transformer models. Evaluated performance on various models (autoregressive, autoencoder, or seq-to-seq transformers) and application areas (language modelling, translation, natural language understanding, question answering). Used sparse GPU kernels to lower memory footprint by 30% and increase inference speed by 10%.
- Defined a few-shot multi-task conditional image generation benchmark by leveraging structure in a large-scale storm event dataset. Improved performance of conditional GANs using meta-learning algorithms and contrastive pretraining.

Research Intern at Celixir

Stratford-upon-Avon, UK

Jun. 2018 - Aug. 2018

- *Overview:* regenerative medicine founded by Nobel Laureate Professor Sir Martin Evans
- Performed cell image analysis (segmentation, feature extraction) to predict cell culture health.
- Estimated feature importance to design future experiments and reduce number of assays.
- Collaborated with biologists to incorporate expert priors for Bayesian inference with MCMC simulations.

Research Intern at Shell Technology Centre Bangalore

Bangalore, India

Jun. 2017 - Aug. 2017

- Skeletonized 3D voxel grids using either their distance transforms or thinning algorithms.
- Implemented and evaluated heuristics for graph search algorithms to speed up numerical simulations.

EDUCATION

Massachusetts Institute of Technology, Cambridge MA

MEng. in EECS (5.0/5.0 GPA) , B.S. in EECS and Physics

Sep. 2015 - May 2021

Coursework: Algorithms, Machine Learning, Statistics, Software, Systems, Quantum Physics, General Relativity

SKILLS AND INTERESTS

- Python: ML (numpy, pytorch, jax, sklearn, pymc) and web (asyncio, tornado, FastAPI, Django)
- ML: NLP; speech recognition and synthesis; data-efficient (contrastive, few-shot) learning; bayesian learning
- Cloud & infra: GCP, AWS, docker, kubernetes, redis, rabbitmq, Prometheus, Grafana, terraform, nginx

Interests: Efficient and Interpretable ML

ACADEMIC ACHIEVEMENTS

Peer Reviewed Publications:

- Adib Hasan, **Ileana Rugina**, Alex Wang; *Pruning for Protection: Increasing Jailbreak Resistance in Aligned LLMs Without Fine-Tuning* ; EMNLP 2024 Workshop BlackBoxNLP
- **Ileana Rugina**, Rumen Dangovski, Li Jing, Preslav Nakov, Marin Soljačić; *Data-Informed Global Sparseness in Attention Mechanisms for Deep Neural Networks*; LREC-Coling 2024
- **Ileana Rugina**, Rumen Dangovski, Mark Veillette, Pooya Khorrami, Brian Cheung, Olga Simek, Marin Soljačić; *Meta-Learning and Self-Supervised Pretraining for Storm Event Imagery Translation*; IEEE High Performance Extreme Computing Conference 2023; earlier version presented at ICLR AI for Earth and Space Science Workshop 2022
- Pooya Khorrami, Olga Simek, Brian Cheung, Mark Veillette, Rumen Dangovski, **Ileana Rugina**, Marin Soljačić, Pulkit Agrawal ; *Adapting Deep Learning Models to New Meteorological Contexts Using Transfer Learning*; IEEE International Conference on Big Data 2021

Silver Medal - Asian Physics Olympiad; **Bronze Medal** - International Physics Olympiad

2015

CLASS/PERSONAL PROJECTS

- SWE: implemented mapreduce, subset of raft consensus in go; networked multiplayer Pinball game in java.
- ML Theory: Last Iterate Convergence of EG Methods for Variationally Coherent Min-Max Problems.