Jonny Li

Toronto, Canada
jonny.li@alumni.utoronto.ca linkedin.com/in/jonnylii github.com/jonnyli1125

Experience

SoundHound AI **Toronto, Canada**

Senior Machine Learning Engineer *Aug 2025 – Present*

- Developing in-house real-time API service supporting custom ASR+LLM+TTS stacks or multimodal LLMs, achieving **10× cost savings** and **3x lower latency** over external APIs, enhancing the voice agent product experience.
- Fine-tuning LLMs for function calling with SFT and RL (GRPO, DPO), achieving optimal accuracy with smaller models, improved inference speed, and significant cost savings.
- Creating the first large-scale NLU simulation and evaluation framework with automated user–agent interactions, establishing reproducible benchmarks where no standard evaluation previously existed.

Machine Learning Engineer II *Aug 2023 – Jul 2025*

- Led the design and deployment of an ASR error-correction LLM that achieved **+90% accuracy improvement** in entity recognition using a custom loss function and knowledge distillation from larger teacher models.
- Primary author of distributed LLM training infrastructure (DeepSpeed, Ray, Kubernetes) supporting ZeRO-3, long-context training, and multi-node experiments; adopted by research teams to accelerate LLM experimentation.
- Conducted in-house multimodal audio LLM research: implemented speech adapters/tokenizers from recent papers and developed high-throughput audio pipelines with S3 integration and lazy transforms to remove I/O bottlenecks.

Software Engineer *Oct 2021 – Jul 2023*

- Developed automated hyperparameter optimization pipelines for ASR, delivering **+30% accuracy improvement**.
- Engineered Spark-based ETL pipelines processing **10s of TBs** of text data, achieving **2× throughput** improvements and enabling scaled training experiments.
- Authored end-to-end MLOps pipelines with well-designed API and integrations across Docker/Kubernetes, Spark, and internal tooling, reducing time spent on experiment iteration cycles by **5×**.

Amazon **Vancouver, Canada**

Software Engineer Intern *Jun 2020 – Aug 2020*

- Built Python dependency graph analyzer to remove redundant configs, optimizing backend search engine efficiency.

Mitsucari **Tokyo, Japan**

Software Engineer Intern *Sep 2018 – Aug 2019*

- Delivered full-stack web features with Rails, PostgreSQL, Heroku and improved UI with jQuery, Bootstrap, and SASS.

Projects

CUDA Vector Search Engine | *C++, CUDA, Python*

- Implemented GPU-native vector search with custom CUDA kernels, achieving **50× better latency** vs baseline.

Japanese Grammar Correction BERT | *Keras, Tensorflow*

- Trained and evaluated grammar correction models that surpassed prior state-of-the-art by **+10% accuracy**.
- Built large-scale synthetic data generation pipeline with streaming/chunked processing.

Open Source Contributor | *PyTorch, DeepSpeed, HF Transformers*

- Resolved DeepSpeed ZeRO-3 integration bug in HF Transformers for audio models (e.g., Wav2Vec2).

Technical Skills

Languages: Python, C++, Java, JavaScript
Frameworks & Infra: PyTorch, TensorFlow, Ray, DeepSpeed, Kubernetes, Docker, Spark, Hadoop
ML/AI: LLM post-training, reinforcement learning, distributed training, inference optimization, MLOps pipeline design, ASR, large-scale ETL

Education

University of Toronto **Toronto, Canada**

Honours Bachelor of Science in Computer Science & Linguistics *Sep 2015 – Jun 2021*