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** over OpenAI API and 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.
- Building the first in-house NLU simulation framework with a simulated user agent, enabling automated, reproducible evaluation and comparison of NLU systems where no prior method existed.

Machine Learning Engineer II *Aug 2023 – Jul 2025*

- Led the design and deployment of an ASR error correction LLM, achieving **+90% accuracy gain** in entity recognition, and designed the core training strategy with a custom loss function and knowledge distillation from a large teacher LLM.
- Primary author of distributed LLM training infra (DeepSpeed, Ray, Kubernetes), supporting long-context training and ZeRO-3, enabling large-scale multi-node experiments and adopted by multiple research teams to accelerate progress.
- Conducted in-house multimodal audio LLM research, implemented papers for speech adapters/tokenizers, and built high-throughput audio data pipelines with S3 integration and lazy transforms to overcome I/O bottlenecks.

Software Engineer *Oct 2021 – Jul 2023*

- Built a hyperparameter tuning pipeline for ASR models that **boosted accuracy by +30%**.
- Engineered Spark-based ETL for **10s of TBs** of text data, replacing MapReduce and delivering **2× faster** processing; managed Spark/Hadoop clusters and optimized HDFS I/O.
- Authored production-grade end-to-end MLOps pipeline with well-designed Python API and integrations across Docker/Kubernetes, Spark, and internal tooling. Automated workflows and reduced time on manual tasks by **5×**.

Amazon **Vancouver, Canada**

Software Engineer Intern *Jun 2020 – Aug 2020*

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

Mitsucari **Tokyo, Japan**

Software Engineer Intern *Sep 2018 – Aug 2019*

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

Projects

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

- Built GPU-native vector search from scratch with **50× speedup** using vector indexes and parallelization.

Japanese Grammar Correction BERT | *Keras, Tensorflow*

- Achieved **+10% over previous state-of-the-art** model on Japanese GEC benchmark.
- Built synthetic data pipeline with streaming/chunked processing for datasets larger than memory.

Open-source Contributor | *PyTorch, DeepSpeed, HF Transformers*

- Fixed a bug in Hugging Face Transformers for training Wav2Vec2 and other audio models with DeepSpeed ZeRO-3.

Technical Skills

Languages: Python, C++, Java, JavaScript
Frameworks/Tools: PyTorch, Tensorflow, Ray, DeepSpeed, Accelerate, Kubernetes, Docker, Spark, Airflow, Hadoop
Concepts: LLM post-training, supervised fine-tuning, reinforcement learning, synthetic data generation, distributed ML training infrastructure, LLM inference optimization, MLOps pipeline design, ETL at scale, ASR

Education

University of Toronto **Toronto, Canada**

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