

# JONNY LI

Toronto, Canada

✉ [jonny.li@alumni.utoronto.ca](mailto:jonny.li@alumni.utoronto.ca)

in [linkedin.com/in/jonnyli](https://www.linkedin.com/in/jonnyli)

github.com/jonnyli1125

## Experience

---

### SoundHound AI

Toronto, Canada

Senior Machine Learning Engineer

Aug 2025 – Present

- Led the development of LLM post-training pipeline (SFT, RL, GRPO, DPO) for function calling and other tasks, using **DeepSpeed, Ray, and Kubernetes**, supporting long-context sequence parallelism and FSDP/ZeRO-3.
- Created real-time voice agent LLM service (ASR + LLM + TTS) with sub-second latency, serving production traffic at scale, reducing costs by **10x**, and delivering **20% higher accuracy** in function calling LLMs.
- Developed reproducible evaluation frameworks for LLMs, including **large-scale simulation and benchmarking**, enabling scientific teams to compare post-training strategies and meaningful feedback for product teams.
- Collaborated cross-functionally with research scientists, product engineers, and infrastructure engineers to integrate experimental LLM methods into production systems.

Machine Learning Engineer II

Aug 2023 – Jul 2025

- Led the research and deployment of an ASR error correction LLM that achieved **+90% accuracy improvement** in spoken entity name recognition; applied modeling techniques like synthetic data generation, multi-task loss, and knowledge distillation.
- Implemented multimodal speech LLMs from scratch and developed distributed multimodal LLM training infra with high-throughput streaming audio pipelines, S3 integration, and lazy transforms overcome I/O bottlenecks.
- Developed automated hyperparameter optimization pipelines, delivering **+30% accuracy improvement** in production ASR model.

Software Engineer

Oct 2021 – Jul 2023

- Engineered Spark-based ETL pipelines processing **10s of TBs** of text data, achieving **2x throughput** improvements and enabling large-scale training experiments.
- Authored end-to-end MLOps pipelines with test-driven development, clean code, and well-designed API; implemented integrations across Docker/Kubernetes, Spark, and internal tooling, reducing time spent on experiment cycles by **5x**.

### 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 **50x better latency** vs baseline.

### Japanese Grammar Correction BERT | Keras, Tensorflow

- Implemented a grammar correction LLM paper and applied to a different language to achieve **+10% accuracy over previous state-of-the-art**.
- Built large-scale synthetic data generation pipeline with streaming/chunked processing.

### Open Source Contributor | PyTorch, DeepSpeed, Hugging Face Transformers

- Caught/resolved several distributed training pipeline bugs in open source libraries, e.g. fixing the DeepSpeed ZeRO-3 integration for audio models such as Wav2Vec2, SFT Trainer bugs in TRL, etc.

## 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, LLM inference optimization, MLOps pipeline design, mulitmodal LLMs, ASR, large-scale ETL

## Education

---

### University of Toronto

Toronto, Canada

Honours Bachelor of Science in Computer Science & Linguistics

Sep 2015 – Jun 2021