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*

- Led modeling efforts for voice agent LLMs, implementing simulated environments for large-scale synthetic data generation, systematic error analysis loops, and online RL, leading to equal/better tool use and instruction following performance than closed models for application domains at **90% less cost**.
- Led development of SFT/RL post-training pipeline for LLMs up to 120B params on Ray/Kubernetes with integrations for PEFT, FSDP/ZeRO-3, and long context sequence parallelism; unblocked key experiments by debugging GPU out-of-memory issues due to integration errors between HF Transformers, PEFT, DeepSpeed, and Liger-Kernel.
- Implemented multimodal speech LLMs from research papers with speech encoders/projectors/tokenizers, and experimented with custom encoders adapted from in-house ASR models to deliver initial proof-of-concept of multimodal instruction following LLMs.
- Built an OpenAI Realtime API clone from scratch to support in-house ASR/LLM/TTS models and speech LLMs in voice agent applications, designing a robust async-iterator streaming architecture and resolving complex coroutine, buffering, and cancellation issues for reliable and responsive behaviour at sub-second latencies.

Machine Learning Engineer II *Aug 2023 – Jul 2025*

- Led LLM-based ASR entity name correction project from initial research to production, combining detailed data/error analysis, synthetic data generation, knowledge distillation, and architecture/loss function refinements to achieve **+90% accuracy on entity names** while staying under 300ms latency per request.
- Built Bayesian optimization-based hyperparameter tuning pipeline for tuning dataset mixture weights, delivering **+30% accuracy improvement** in production ASR model.

Software Engineer *Oct 2021 – Jul 2023*

- Engineered large-scale Spark ETL pipeline handling tens of TBs, debugged cluster out-of-memory issues by analyzing logical/physical plans of Spark SQL queries and implementing salting/sharding to resolve data skew across partitions, achieving **200% throughput** of previous pipeline, significantly reducing experiment iteration time.
- Authored end-to-end MLOps pipelines, designed clear and extensible APIs, resolved complex Docker dependency issues, and used Docker Compose/Kubernetes to stand up test environments for integration/end-to-end testing, which reduced manual work and experiment iteration time by **500%**.

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 (Github)

Open Source Contributor | *Hugging Face Transformers, TRL, PEFT* **2024 - Present**

- Contributed several distributed training bug fixes to open source libraries, e.g. caught/debugged GPU out-of-memory issues due to memory leaks and fixed several DeepSpeed ZeRO-3 integration bugs.

CUDA Vector Search Engine | *C++, CUDA, Python* **2024**

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

Japanese Grammar Correction LLM | *Keras, Tensorflow* **2023**

- Implemented/reproduced a grammar correction BERT model from a research paper and built a large-scale synthetic data generation pipeline with streaming/chunked processing to achieve **+10% accuracy over previous state-of-the-art**.

Education

University of Toronto **Toronto, Canada**

Bachelors of Science in Computer Science & Linguistics (Honours) *Sep 2015 – Jun 2021*

University of Tokyo **Tokyo, Japan**

Foreign Exchange *Sep 2018 – Aug 2019*