

Josiah Putman

📞 603-892-3104 ✉ joshikatsu@gmail.com [in linkedin.com/josiahputman](https://www.linkedin.com/josiahputman) github.com/katsutoshi

EDUCATION

Dartmouth College

Class of 2020

Bachelor of Arts in Computer Science and Japanese Language

Hanover, NH

- Coursework focus: Robotics, Artificial intelligence, Deep learning, NLP.
- Awards: Phi Beta Kappa, summa cum laude, Neukom Scholar ([NICS](#)).

GPA 3.98 / 4.00

EXPERIENCE

Google

September 2020 – Current

Senior Software Engineer

Seattle, Washington

- Lead developer of an LLM-based query targeting system used across Local Search Ads (LSA), accounting for over \$1.3B in annual revenue. Utilized supervised fine-tuning and knowledge distillation to create a two-tower embedding model for low-latency serving. Managed cross-organization collaboration to integrate the system with various products in LSA, driving +\$200M total growth in ARR for local, services, and travel ads.
- Fine-tuned SoTA Gemini models with RL to build a highly customizable query relevance classifier, generating an additional \$120M in ARR for services ads while maintaining high user interaction rates.
- Built TPU-accelerated vector retrieval pipelines for classifying $O(10B)$ queries in under 1 hour.

Urban Systems Lab - ClimateIQ

May 2024 – May 2025

Machine Learning Fellow

Seattle, Washington

- Designed and trained custom ConvLSTM architectures in JAX and TensorFlow for flood forecasting and atmospheric predictions in urban centers with less than 5% of the compute cost of standard physics-based simulations (event.newschool.edu/climateiq).
- Led research and exploration of different problem formulations, model architectures, and resource optimizations, enabling 100× training throughput and 90% reduction in RMSE.

UpTime Solutions

May 2019 – August 2020

Machine Learning Engineer

Hanover, New Hampshire

- Engineered ML models and data processing frameworks for bearing-fault detection in Keras and TensorFlow. Led development of Python API used for data aggregation, serving, and low-latency analysis pipelines.

Microsoft

May 2019 – August 2019

Software Engineer Intern

Seattle, Washington

- Developed cloud-scale WebSocket Server for MS Graph WebHook notifications in C#.

PROJECTS AND PUBLICATIONS

PLRC* For Motion Planning | *Published in IROS 2020*

[researchgate.net](https://www.researchgate.net)

- Lead developer (Julia, C++) and researcher for piecewise-linear regression complexes for approximately optimal motion planning.

Kataru | *YAML Based Dialogue Engine*

kataru-lang.github.io

- Developed high-performance dialogue engine for simplifying writing dialogue for story-driven games in Rust, supporting JS/WASM targets and Unity. Built comprehensive developer tooling through a VS Code extension.

WASM Galaxy Simulation | *Physics Simulation on the Web*

galaxy-sim.github.io

- Developed WASM-deployed Rust implementation of Barnes-Hut algorithm for scalable galaxy simulation ([barnes-hut-rs](#)).

TECHNICAL SKILLS

Languages: Rust, C/C++, C#, Python, Java, Go, Kotlin, SQL, GLSL, CSS

Technologies: CUDA, JAX, Keras, Tensorflow, PyTorch, ROS

Concepts: Machine Learning, Generative AI, Compilers, Data Analytics, Neural Networks, HPC, Game Development, Shaders, GPU