Josiah Putman

☎ 603-892-3104 ✉ joshikatsu@gmail.com linkedin.com/josiahputman github.com/katsutoshii

Education

Dartmouth College Class of 2020
Bachelor of Arts in Computer Science and Japanese, summa cum laude *Hanover, NH*

- Coursework focus: Robotics, Artificial intelligence, Deep learning, NLP. GPA 3.98 / 4.00

Research Interests

ML-based climate forecasting, natural language processing, deep recommendation systems, approximate optimal control

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 $O(10\text{B})$ 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.

Morehouse College - Google-in-residence September 2023 – January 2024
Guest Lecturer and Teaching Assistant *Remote*

- Expanded access to high-quality computer science education in HBCUs by volunteering as a guest lecturer and teaching assistant for introductory CS courses as part of the Google-in-residence program.

UpTime Solutions September 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#.

Dartmouth Robotics Lab June 2017 – August 2019
Undergraduate Researcher *Seattle, Washington*

- Research towards regression-based approximation algorithms for motion planning using Python, C/C++, Julia. Project lead for integration with Open Motion Planning Library.

Dartmouth Department of Computer Science June 2017 – August 2019
Teaching assistant *Seattle, Washington*

- Teaching assistant for CS10 (Intro to OOP), CS30 (Discrete Math), CS31 (Algorithms).
- Led section tutorials, graded problem sets and exams, hosted general office hours.

Projects and Publications

PLRC* For Motion Planning | *IROS 2020* researchgate.net

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

Regression-based Motion Planning | *Undergraduate thesis* digitalcommons.dartmouth.edu

- Researched and developed two novel approaches to motion planning that utilize function approximations to reduce the memory cost of planning with bounded increases in path cost. Advised by Dr. Devin Balkcom.

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).

Planar MaxCut | *Approximation algorithm research article* digitalcommons.dartmouth.edu

- Researched and developed two novel approaches to motion planning that utilize function approximations to reduce the memory cost of planning with bounded increases in path cost.

Honors and Awards

Google SAGE Hackathon Winner 2024
Google, Search Ads Org-wide Hackathon

Pray Modern Language Prize in Japanese 2020
Dartmouth Department of Asian Societies, Cultures, and Languages

Phi Beta Kappa 2020
Member, Dartmouth College Chapter

Neukom Research Scholarship 2018
The Neukom Institute for Computational Science

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