

Joshua Tree Karpel

josh.karpel@gmail.com

jtk.dev

github.com/JoshKarpel

linkedin.com/in/joshkarpel

Also available as [Web](#) or [Markdown](#).

SKILLS

- [Product-minded](#) technical leadership of engineering teams
- Platform, service, and developer tooling engineering, from laptop to production
- Software architecture, from libraries to distributed systems
- Understanding existing code, then refactoring/extracting/migrating/replacing it
- Explaining systems to executives, product managers, and engineers alike
- AI/ML infrastructure, with observability and CI/CD
- Performance and cost optimization, without trading one for the other
- Python primarily, plus Rust and TypeScript in personal projects

WORK EXPERIENCE

[Workday](#)

November 2020 – Present

Principal Machine Learning Engineer

November 2023 – Present

- Served as tech lead of our [Ray](#)-based model-serving system, which we grew into the company-wide platform for model hosting and external LLM access; designed the sharding system that uses [rendezvous hashing](#) to spread tens of thousands of models across several physical clusters behind one logical service.
- Reduced our organization's cloud spend by 40% while improving latency and reliability across services within and on top of the AI platform, by optimizing how they use infrastructure rather than just cutting the capacity they run on.
- Helped design and build an [Argo CD](#)-based deployment system for the AI platform and the ~150 services on our Kubernetes clusters, cutting pull-request-to-production time from eight hours to three minutes, with [Rego](#) policy checks on rendered [Helm](#) and [Kustomize](#) manifests, so bad configuration fails in CI.

Senior Machine Learning Engineer

November 2021 – November 2023

- Led the original design and build of the model-serving system, 50x cheaper than the one it replaced; won a Workday Innovator Award for it in 2023.
- Helped design and build the monorepo that brought our shared inference service framework to dozens of services and hundreds of engineers.
- Migrated a service's model-building and ETL pipelines from our colocated data center to AWS, trading a fixed hardware footprint for elastic capacity; won one of 25 quarterly Outstanding Contributor Awards for the work in Fall 2022.

Machine Learning Engineer

November 2020 – November 2021

- Factored a shareable inference service framework out of an existing service, giving later services a common base to build on.
- Built tooling for defining our alerts and dashboards programmatically, so that observability was reviewed and versioned like the rest of our code.

Center for High-Throughput Computing Summer 2018 – November 2020

Morgridge Postdoctoral Fellow January 2020 – November 2020
Core Computational Technology Group at the [Morgridge Institute for Research](#)

- Worked with researchers to turn their science into automated high-throughput workflows, generalizing one-off solutions into shared tools like [Dask-CHTC](#).
- Modernized how [HTCondor](#) taught and tested itself: rewrote the [Python bindings tutorials](#) for users rather than administrators, and built Ornithology, a [pytest](#)-based alternative to the existing Perl-based testing framework.

Project Assistant Summer 2018 – Fall 2019

- Primary developer of [HTMap](#) and [htcondor.dags](#) (now part of [HTCondor](#) itself), Python packages for describing and running HTCondor workflows.

UW-Madison Physics Department Fall 2014 – Fall 2019

Graduate Research Assistant Summer 2015 – Fall 2019
[Yavuz Research Group](#)

- Modeled [ultrashort intense laser-atom interactions in hydrogen atoms](#) and [non-linear Raman processes in glass microspheres](#), using the Scientific Python stack and high-throughput computing systems to run millions of simulations.

Graduate Teaching Assistant Fall 2014 – Spring 2018

- Taught discussion and laboratory sections ranging from electromagnetism and [optics](#) labs to physics in the arts, and revised the manuals and supplementary materials for the advanced lab courses; won the department's [Joseph R. Dillinger Award for Teaching Excellence](#) in 2018.

SELECTED TALKS

[ALL TALKS](#)

[How Workday Achieved 50x Cheaper Model Serving with Ray Serve](#)
Ray Summit, November 2025

[Concurrency and Parallelism](#)
MadPy, June 2024

[pytest: It's What's For Testing](#)
MadPy, December 2022

[The Python Data Model](#)
MadPy, August 2019

[Rust for Pythonistas](#)
MadPy, September 2018

EDUCATION

Ph.D. in Physics Fall 2014 – Fall 2019
[University of Wisconsin-Madison](#)

B.A. in Physics and Mathematics Fall 2010 – Spring 2014
[University of Colorado-Boulder](#)