Jacob Platin

(314)-605-4110 | jacobplatin@google.com | jrplatin.github.io | github.com/jrplatin | linkedin.com/in/jacob-platin

EDUCATION AND SKILLS

University of Pennsylvania, School of Engineering and Applied Science — **Philadelphia, PA**
MSE in ***Robotics*** *(Machine Learning Specialty)*, *GPA 3.7/4.0 (Magna Cum Laude)* — Aug 2017 – May 2022

- **Relevant Coursework**: Machine Learning, Network System Design, Computer Vision, Deep Learning

University of Pennsylvania, School of Engineering and Applied Science — **Philadelphia, PA**
BSE, Majors in ***Computer Science*** & ***Economics***, *GPA 3.5/4.0 (Cum Laude)* — Aug 2017 - May 2022

- **Relevant Coursework**: Cloud Computing/Scalability, Econometrics, Game Theory, Data Structures, Software Design

ETH Zurich, Departments of Computer Science and Economics — **Zurich, Switzerland**
Exchange Program, GPA 3.75/4.0 — Sep 2019 - Dec 2019

- **Relevant Coursework**: Computer Architecture, Reliable Artificial Intelligence, Wireless/Mobile Computing

EXPERIENCE

Google | *Software Engineer III (Machine Learning)* | Kirkland, WA — **Feb 2025 – Present**

- Leading weight + activation and KV cache quantization efforts on Google's TPU and GPU OSS vLLM stack to increase end-to-end throughput by 90% while preserving 99% quality
- Modified OSS Pallas kernels to address inference inefficiencies identified through JAX profiling + roofline analysis
- Collaborated internally and externally to implement new models, including DeepSeekV3 and Llama4, and new features, including multi-latent attention and mixture-of-experts, into the stack

Microsoft | *Software Engineer II (Machine Learning)* | Redmond, WA — **Dec 2023 – Feb 2025**

- Integrated internal speech models and LLMs to achieve SOTA multi-modal model performance
- Led two sub-teams dedicated to increasing model size and training speed at minimum compute cost
- Fostered an inclusive and growth-oriented team by organizing paper readouts, relaxation retreats, and learning sessions

Microsoft | *Software Engineer (Machine Learning)* | Redmond, WA — **Aug 2022 – Nov 2023**

- Led efforts on optimizing model size, performance, and throughput for Microsoft's latest speech recognition models by applying state-of-the-art sharding, networking, and architecture-based techniques
- Owned and enhanced the software framework that 200 members of the Azure AI Speech team used to train models

Unity Technologies | *Software Engineer Intern (Robotics)* | Seattle, WA — **May 2021 - Aug 2021**

- Integrated an inverse kinematics solver directly into the Unity engine using linear algebra and robotics principles
- Collaborated with NVIDIA to implement realistic, physics-based joint controllers for robotic simulations
- Engineered a VR experience to define and visualize a robot's workspace in Unity

NVIDIA | *Software Engineer Intern* | Redmond, WA — **Feb 2021 - May 2021**

- Developed and shipped a cloud-based searching solution for game meta-data using Elasticsearch, GraphQL and AWS that improved query latency by 35% and reduced the existing codebase size by 20%

Unity Technologies | *Software Engineer Intern (AI)* | Seattle, WA — **May 2020 - Aug 2020**

- Explored and implemented classical and machine-learning driven robotic manipulation in the Unity engine
- Deployed a more efficient bridge between ROS and Unity that was adopted by over 1000 users

PUBLICATIONS

Microsoft | *Phi-4-Mini Technical Report* | arxiv.org/abs/2503.01743 — **Mar 2025**

- Integrated Microsoft's SOTA multi-modal language model with vLLM to optimize inference performance

LEADERSHIP & MENTORSHIP

TAMID Group | *Mentor* — **Dec 2024 – Present**

- Coach current TAMID students on how best to navigate their undergraduate experience by utilizing personal experience
- Guide multiple students through career exploration while helping them develop necessary skills to prepare for SWE/ML

Global Mentoring Initiative | *Mentor* — **Aug 2023 – Present**

- Provide career and technical mentorship to international students from disadvantaged backgrounds
- Serve as a Google ambassador for the program and help to expand the curriculum to fit students' unique interests