

# James Wong, PhD Candidate, MS CS

**Building future ML/LLMs @ AMD/Lamini AI, Oracle, Intel, Cisco, UMich, NUS**

<https://www.linkedin.com/in/jameshsw>

<https://jameshsw.github.io>

jameshsw@gmail.com

1-510-766-0666

Fremont, CA, US

US Citizen

## SUMMARY

Machine Learning Engineer with deep expertise in AI infrastructure, specializing in LLM fine-tuning and inference scalability. Led development of fine-tuning, classifier, and agentic AI platforms while managing a high-performance GPU data center and optimizing production systems with tools like vLLM, Ray, and Llama Factory.

## EDUCATION

-PhD Candidate, MS, Computer Science, University of Michigan  
-MS, Electrical Engineering, National University of Singapore  
-BS, Electrical Engineering and Computer Science, National University of Malaysia

## SKILLS

-AI/ML: PyTorch, Transformers, LLM, vLLM, Ray, Llama Factory, Slurm, MPI  
-Languages: Python, JavaScript, Java, C  
-MLOps: Kubernetes, Docker, Helm, CI/CD, Ansible  
-Cloud: AWS, Azure, GCP  
-Monitoring: Grafana, Prometheus  
-Databases: SQL, NoSQL  
-Frontend: React, Tailwind  
-Hardware: Verilog, Virtuosio, Xcellium, RTL, CAD, SoC, JAG

## LANGUAGE

English, Chinese

## EXPERIENCE

### **Machine Learning Engineer/Architect, AMD acquisition/Lamini.ai, Menlo Park, CA (2024-Present)**

- Engineered distributed inference and training for LLMs using vLLM, Ray, Kubernetes, and Slurm.
- Developed an LLM platform with memory optimization and reducing hallucinations.
- Managed GPU data center and clouds (AWS, GCP, Azure) for training and inference workloads.
- Worked on LLM platforms for Text2SQL, Factual QA, Classification, RAG, and agentic pipelines.

### **Project Lead, Open Compute Project, Santa Clara, CA, (2020-Present), open-source/part-time**

- Co-led chiplet data format standardization (CDXML), contributing to a faster integration process in heterogeneous systems.
- Developed interoperability standards for 3DIC and chiplet-based systems, improving cross-vendor compatibility and reducing development cycle times.

### **Head of Engineering/CTO, zGlue, Palo Alto, CA (2017-2023)**

- Spearheaded the award-winning "ChipBuilder" EDA tool with automated place-and-route and verification tools, reducing design time by an estimated 40%.
- Launched a chiplet marketplace with an ML-driven recommendation system that achieved a 30% improvement in component selection accuracy.
- Directed a cross-functional engineering team, driving 10+ product launches and overseeing the D2D interfaces program, presenting at industry events.

### **Senior Director of Engineering, MA Labs, San Jose, CA (2012-2017)**

- Grew and managed an engineering team of 50+ members to develop scalable eCommerce platforms supporting millions of daily transactions.
- Oversaw an ERP, WMS, and TMS integration, boosting operational efficiency by 70%.
- Implemented CI/CD pipelines, resulting in 50% reduction in deployment times across multiple product lines.

### **Engineering Manager/Senior Principal Engineer, Oracle. Redwood City, CA (2002-2011)**

- Developed the Advanced Pricing Engine with an in-memory database, adopted by 100+ Fortune 500 clients.
- Launched AI-driven pricing optimizer, resulting in 5-10% revenue increase for clients.
- Spearheaded a transition to cloud-based architecture, attracting 200+ new enterprise clients to Oracle Public Cloud.

### **Senior Software Engineer, Cisco, Santa Clara, CA (2000-2002)**

- Architected cloud-based payment systems with 99% uptime and seamless integration with 30+ global gateways.
- Developed standardized REST APIs, enhancing transaction handling efficiency by 20%.

### **Design Engineer, Intel, Malaysia (1995-1996)**

- Designed microcontroller I/O using RTL and Verilog
- Performed system-level simulation and verification on microcontrollers

## PUBLICATION

<https://scholar.google.com/citations?user=BpdqLXIAAAAJ&hl=en>

- [Banishing LLM Hallucinations Requires Rethinking Generalization. James Wong. et al. 2024](#)
- [Electrical Interfaces Performance Metrics. James Wong, et al. OCP White Paper 2024](#)
- Die-to-Die Chiplet Interface Testing, James Wong, et al. OCP White Paper 2024 · Feb 13, 2024
- [Open Platform for Chiplet Development and Bring-up. James Wong. et al. Chiplet Summit 2024](#)
- Functional Simulation and Verification Workflow for Chiplet-based Systems. James Wong. Chiplet Summit 2024
- [Guide to Integration Workflows for Heterogeneous Chiplet Systems. James Wong. et al. OCP White Paper. 2023](#)
- [Business Analysis of Chiplet-Based Systems and Technology. James Wong. et al. OCP White Paper. 2023](#)
- [Panel: Innovating in the Open Chiplet Economy. James Wong, et al. OCP Global Summit 2023](#)
- [Chiplet Design and Verification Using An Open Standard Markup Language. James Wong. et al. DAC 2023](#)
- [Using a Markup Language in Chiplet-Based Design. James Wong. Chiplet Summit. 2023](#)
- [CDXML - Chiplet Data Exchange Markup Language. James Wong. et al. OCP Global Summit. 2022](#)
- [Design of Heterogeneous Integrated Circuits - Chiplets and Models. James Wong. et al. MEPTEC Report. Fall 2021](#)
- [Proposed Standardization of Chiplet Models for Heterogeneous Integrated. James Wong, et al. OCP White Paper. 2021](#)
- [Proposed Standardization of Heterogeneous Integrated Chiplet Models. James Wong. et al. IEEE International 3D Systems Integration Conference. 2021](#)
- [Oracle Advanced Pricing Engine. Hockshan Wong. et al. User's Guide. 2006](#)
- [Oracle Advanced Pricing Engine Implementation Manual. Hockshan Wong. et al. Oracle. 2006](#)
- Improving distributed control coordination using application semanticsImproving distributed control coordination using application semantics. Hockshan Wong et al. 7th Semi-Annual Technical Advisory Committee Meeting, NFS Engineering Research Center for Reconfigurable Machining Systems, Ann Arbor, Michigan, Feb 13, 2000
- Distributed Control System with a State Observer to Decrease Communication. Hockshan Wong, et al. Japan-USA Symposium on Flexible Automation, Ann Arbor, Michigan, Feb 1, 2000
- [Trading Computation For Bandwidth: State Estimators For Reduced Communication In Distributed Control Systems. Hockshan Wong et al. 2000 Japan-USA Symposium on Flexible Automation](#)
- Personalized Bidding Agents for Online Auctions. Hockshan Wong, et al. Proceedings of The 5th International Conference on the Practical Application of Intelligent Agents and Multi-Agents, 2000
- [Agents Participating in Internet Auctions. Hockshan Wong. et al. Proceedings of the AAAI Workshop on Artificial Intelligence. 1999](#)
- [Agent Service for Online Auctions. Hockshan Wong. et al. AAAI. 1999](#)
- [Robustness monitoring for PID control systems. Hockshan Wong. et al. IECON. 1998](#)