Siddharth Vohra

MACHINE LEARNING ENGINEER & RESEARCHER

siddvoh@gmail.com |www.linkedin.com/in/siddvoh |www.siddvoh.com | +1 (858) 349-3962

INDUSTRY EXPERIENCE

Amazon Web Services — August 2022 – Present

Machine Learning Engineer, AWS Generative AI Innovation Center (September 2025 - Present) — Pittsburgh, PA

- Work as MLE for the new AI native org, working to design new strategies to work directly with AWS customers to deliver products and projects that are built keeping agentic AI at the centre
- Work in an inter-disciplinary small team consisting of Scientists, Researchers and Engineers to develop comprehensive processes and pipelines to reduce time taken from ideation to MVPs and prototypes
- Building an automated security & compliance tool for AWS architects and external clients to use while developing POCs to reduce manual security assessment times by at least 50%

Software Development Engineer II, AWS AI (April 2024 - September 2025) — Seattle, WA

- Specialized in maintaining, enhancing, and improving Amazon Transcribe, AWS Bedrock Data Automation and Amazon Translate within the AWS Bedrock Generative AI Services organization by leveraging cutting-edge AWS AI/ML models and technologies
- Developed and collaborated on large-scale, advanced automatic speech recognition (ASR) models, significantly enhancing speech-to-text capabilities for up to 30% faster and 2x more accurate transcription services for multi-language use-cases
- Designed, prototyped and built a new service architecture to completely revamp audio and video tasks for Agentic AI use-cases via Bedrock Data Automation
- Serve as tech lead for AWS Translate, supporting custom feature requests, security mitigations, and continuous service enhancements
- Led the designing, implementation, and and delivery of a pipeline to serve speaker diarization models for Bedrock Data Automation
- Led a 3-person team to expand Custom Language Models, facilitating the integration of domain-specific knowledge to transcripts
- Led development and deployment of next-generation multilingual identification model, implementing comprehensive testing suites with 100+ test cases to ensure accurate transcription for customers
- Collaborated with over 30 enterprise customers to architect bespoke solutions, deliver feature requests, and resolve engineering and model-related issues across Amazon Transcribe and Amazon Translate
- Optimized deployment and hosting of models on AWS SageMaker, creating an efficient and scalable model serving pipeline
- Improved and optimized scalable pipelines on AWS infrastructure, ensuring high availability and low latency for translation services

Tech stack: Java (Back-end), Python (Back-end), PyTorch, TensorFlow & AWS AI/ML tools suite

Software Development Engineer I, AWS Lambda (August 2022 - March 2024) — Seattle, WA

- Contributed to a diverse portfolio of products in the AWS Lambda organization as part of the AWS Elastic Beanstalk & App Runner team, collaborating across 5+ teams, blending various research methodologies for internal, open-source, and customer-driven projects
- Led the cross-team integration of AWS WAF into Copilot CLI (Go) for App Runner, ensuring secure apps for thousands of AWS users
- Architected and engineered an automated testing suite that simulates every potential user interaction with the App Runner console to ensure a 99.9% up-time, securing stability and scalability within budget constraints
- Led the redevelopment of a versatile internal tool, reducing issue resolution time by 85% through systematic evaluation
- Improved codebase, worked with product managers and designers to fix multiple bugs and remove numerous customer-facing pain-points for App Runner & Elastic Beanstalk console experience, leading to a 18% increase in service adoption using the console
- Improved team-owned product backend infrastructure by migrating to AWS CDK, enabling efficient deployment methodologies
- Played a pivotal role in the launching App Runner into 3 new regions and catalyzing a multi-million dollar revenue increase for AWS
- Employed AWS ecosystem technologies (EC2, CDK, WAF, CloudWatch, S3, DynamoDB, Lambda) to deliver high-quality solutions

Tech stack: Go (Back-end), Java (Back-end), Python (Back-end), TypeScript, Node.js, JavaScript (Front-end), React.js & AWS tools

Teradata — July 2021 – September 2021

Software Engineer Intern — San Diego, CA

- Devised and implemented strategies to streamline storage and management of large objects in the TeraCloud system architecture
- Collaborated with TeraCloud team to restructure storage for large objects within and across Teradata database systems, resulting in 50% more efficient & swift computation for stakeholders
- Engineered sophisticated wrapper functions for use across Teradata systems, ensuring enhanced security, efficiency, and scalability
- Enhanced system reliability and scalability, supporting seamless SQL integration and management of large datasets

Tech stack: C (Back-end), SQL (Queries) & Teradata database system (Database)

EDUCATION

Carnegie Mellon University **August 2025 - May 2027**
Master of Science, Computer Vision (Robotics Institute) Pittsburgh, PA

University of California, San Diego **September 2019 - June 2022**
Bachelor of Science, Computer Science & Mathematics; Cum Laude, Latin Honors San Diego, CA

- Provost Honors, Eleanor Roosevelt College (All Enrolled Quarters)
- Founding & Principal Member, Machine Learning Club at UC San Diego and Member, Data Science Student Society

RESEARCH EXPERIENCE

Independent Research **June 2020 - Present**

- Conducted independent research with mentors from Hitachi R&D and Thoughtworks, spanning traditional machine learning, large language models (LLMs), and computer vision
- Co-authored a COLING 2025 paper introducing a multiple-choice question benchmark for Hindi and Kannada educational texts and providing LLM-based baseline difficulty-estimation models.
- Co-authored a AIED 2025 submission investigating whether semantic similarity between answer options can predict question difficulty
- Benchmarked Temporal Ensembling semi-supervised models across diverse vision datasets, analysing intra-class variability impacts in collaboration with Hitachi R&D; findings published at COMSYS 2023
- Initiated and led a project to benchmark coding-question difficulty using LLMs, developing a BERT-style model for automatic programming-problem complexity estimation.
- Explored computer-vision techniques for pedestrian-attribute recognition and person re-identification, evaluating multiple network architectures to improve identification accuracy in security applications.
- Built reproducible pipelines to fine-tune GPT, Claude, Llama and Gemma models with LoRA/QLoRA, enabling single-GPU experimentation.

Tech stack: Python (Back-End), PyTorch (ML Library) & Google Colaboratory (Model Training using cloud GPU)

PUBLICATIONS

1. Ravikiran, M., **Vohra, S**., Verma, R., Saluja, R., & Bhavsar, A. (2025). **TEEMIL: Towards Educational MCQ Difficulty Estimation in Indic Languages.** The 31st International Conference on Computational Linguistics (COLING 2025). ACL Anthology. https://aclanthology.org/2025.coling-main.142/.
2. Ravikiran, M., **Vohra, S.**, Nonaka, Y., Kumar, S., Sen, S., Mariyasagayam, N., & Banerjee, K. (2023). **You Reap What You Sow—Revisiting Intra-class Variations and Seed Selection in Temporal Ensembling for Image Classification.** In Proceedings of International Conference on Frontiers in Computing and Systems (pp. 73-82). Springer, Singapore.
   (Manikandan Ravikiran and Siddharth Vohra—Both authors contributed equally. Names are ordered alphabetically)
3. **Vohra, S.,** & Ravikiran, M. (2020, August 21). **Investigating the effect of intraclass variability in temporal ensembling.** arXiv.org. https://arxiv.org/abs/2008.08956

TECHNICAL BLOGS

1. **Vohra, S.** (2023, February 23). **Using WAF with app runner in copilot**. https://aws.github.io/copilot-cli/blogs/apprunner-waf/

CERTIFICATIONS

- **Machine Learning Specialization** by DeepLearning.AI and Stanford Online (2024)
- **Deep Learning Specialization** by DeepLearning.AI (2024)
- **Harvard Business School Credential of Readiness (CORe)** (2023)
- **Amazon Web Services (AWS)** Certified Cloud Practitioner (2020, 2024)
- **Amazon Web Services (AWS)** Certified AI Practitioner (2024)

SKILLS

- **Languages & DevTools:** C, Java, Python, Golang, C++, ARM, Java Script, HTML/CSS, JUnit, Puppeteer, MATLAB, R
- **Libraries & Tools:** AWS, GCP, Azure, React JS, Node JS, Numpy, Pandas, PyTorch, TensorFlow, Git, MySQL, MongoDB, IndexedDB
- **AI/ML Engineering:** AWS Managed AI Suite, AWS SageMaker, Bedrock, Fine-tuning, Quantization, Model Inference pipelines