Yashaswi Sharma

yashjobs20@gmail.com · +1 (301) 503-8933 · Linkedin · GitHub

AWS-certified Machine Learning Engineer

Education

University of Southern California — Los Angeles, CA
M.S. Computer Science — 08/2025 - 05/2027
Courses: Deep Learning, High-Dimensional Math
Research: Reasoning Economics, Differentiable Digital Signal Processing

University of Maryland — College Park, MD
B.S. Computer Science (Machine Learning) & B.S. Economics (Macro, Game Theory) — 08/2018 - 05/2022
Research: Differentiable Audition, Differentiable Economics, Applied AI in Agriculture

Work Experience

Snovation — Remote–Baltimore, MD
AI Engineer — 12/2024 - 07/2025

- Engineered an application-specific Foundation Model hosting service, maximizing token throughput by 200X (20,000 %) (from 0.34 to 67.1 tok/s) and reducing inference latency by 77% (from 180 sec to 42 sec) while optimizing resource utilization to a CPU-only architecture using 2 VCPUs
- Developed a Domain-Specific Language (DSL) converter for healthcare claims using advanced deep learning and compiler techniques, resulting in a 99%-accurate copilot deployed as an Azure Function
- Reduced sub-system end-to-end latency by 98% by improving data retrieval and processing methods, improving end-user experience through a 40+% reduction in latency

Amazon Web Services — Arlington, VA
Software Development Engineer — 09/2022 - 09/2024

- Designed and deployed scalable inference and feedback systems for AWS Config's first Generative AI feature (**ReInvent 2023**), leveraging Function-as-a-Service (FaaS) architectures, automated testing, and adaptive rate limiting.
- Architected and implemented a core data management system with 99.99% availability & performance, leveraging Day-0 AWS services for state management, message processing, and anti-entropy mechanisms
- Built a data integration platform for other service teams to create configuration-based error-correction and security-compliance systems for end customers, reducing integration time by 90+%
- Automated Java 8 to Java 21 code upgrades using large language models (LLMs), eliminating service latency, security, and availability issues across the platform while reducing developer effort by 75+%
- Formulated security-first processes to reliably deploy Config Manged Rules, reducing developer effort by 80+% and ensuring 90+% regional/partitional feature parity
- Led 200+ critical on-call investigations across diverse systems, including those with major global security implications, ensuring 24/7 system availability, reliability, and security

Skills

ML Performance/Inference:	PyTorch, Tensorflow, Scikit-Learn, XGBoost, CUDA, MPI, OpenMP, JAX/XLA, Quantization, Amazon Comprehend, SageMaker Training Compiler, Inferentia, Neo
MLOps:	AWS SageMaker Pipelines, Feature Store, Model Registry/Monitor, Clarify, Debugger GroundTruth, Amazon Bedrock, CI/CD (CodeBuild, CodeDeploy, CodePipeline)
Distributed Systems:	AWS API Gateway, SNS, SQS, VPC, EC2, Lambda, Step Functions, ECS, Fargate, ECR, EKS/Kubernetes, Docker, gRPC/Protobuf, REST, Azure Serverless Functions
Data Engineering:	AWS Glue (DataBrew, Data Quality), Athena, Redshift, Kinesis (Data Streams, Firehose), Kinesis Flink, DynamoDB, OpenSearch, RDS, Batch, EMR, Lake Formation, Spark
Governance & Security:	AWS IAM, VPC Endpoints/PrivateLink, GuardDuty, CloudTrail, CloudWatch, Config, Security Hub, Control Tower, Cost Explorer, Macie, KMS, SecretsManager
Programming Languages:	Java, Kotlin, Python, Go, Rust, C/C++, OCaml, TypeScript, SQL

Notable Projects

Multi-Agent Reinforcement Learning for Stackelberg Competition — Project Link
PyTorch, OpenAI Gym, Matplotlib, Seaborn
Developed a simulation of an Enhanced Stackelberg Competition using multi-agent heirarchical adversarial reinforcement learning (MAHA-RL), modeling leader-follower dynamics in economic markets

Neural Audio Compression — Project Link
PyTorch, Librosa, Numpy, Scipy, Matplotlib, Seaborn
Developed an Encoder-Decoder model with a custom intermediate compression function trained on 2 separate loss functions with separate backpropagation protocols