Alok Upadhyay

Seattle, WA | alok0412@gmail.com | linkedin.com/in/alok˜upadhyay | alok.ai

ML Engineering Leader with 12+ years at Amazon/AWS. Manager of Managers leading ∼20 person orgs building large-scale Recommendation Systems, Generative AI pipelines (3B+ messages/month), and multi-modal ML products (VoiceID, VisualID, BLE) (31B+ inferences/month). 4 granted US patents. 4 peer-reviewed publications including ICLR 2026 Workshops. Proven track record shipping zero-to-one ML products on consumer devices serving millions of users, with measurable business impact on engagement, retention, and operational efficiency.

Technical Skills

ML & AI: PyTorch, TensorFlow, SageMaker, Bedrock, HuggingFace Transformers, LangChain, agentic AI systems, parameter-efficient fine-tuning (LoRA, PEFT), RAG pipelines, prompt engineering, LLM evaluation & safety, inference optimization, distributed systems for ML, model serving at scale
Recommendation Systems: SASRec, BERT4Rec, Two-Tower DNNs, LightGBM, collaborative filtering, retrieval & ranking, real-time personalization at billions-scale
Multi-Modal ML: voice biometrics, facial recognition, BLE proximity, behavioral signal fusion, online/offline learning, human-in-the-loop ground truth pipelines
Infrastructure: AWS (EC2, DynamoDB, SQS, SNS, IAM, Lambda, Step Functions, GovCloud), Kubernetes, vLLM, CUDA, GPU clusters, ElasticSearch/OpenSearch
Data & MLOps: batch & streaming ETL, feature stores, A/B testing, CI/CD for ML, model monitoring & observability, HIPAA/FIPS compliance
Languages: Python, Java, C/C++, SQL

Experience

Software Development Manager – Prime Video Personalization & Discovery Jul 2024 – Present
Amazon *Seattle, WA*

- Lead the Outbound Message Generation & Engagement organization (∼20 engineers, SDMs, and TPMs), owning customer enrollment, engagement, and retention through personalized touchpoints
- Architected and deployed an autonomous AI agent on AWS Bedrock for content QA – performs multi-step reasoning over text, images, and links to validate 500M+ assets/month across Email, Push, and WhatsApp, eliminating manual vendor review
- Direct a massive-scale content generation pipeline delivering 3B+ personalized messages monthly, leveraging intricate transformer recommendation architectures for retrieval and ranking with time-decay, anti-noise filtering, and churn prevention to maximize CTR, DSD, Hours Watched.
- Established the technical roadmap transitioning the org from legacy systems to a GenAI-first architecture; drove high-velocity experimentation cycles (A/B tests, rapid model iteration) reducing inference cost by 17% while maintaining strict latency SLAs

Software Development Manager – Ambient Recognition & Authentication Mar 2020 – Jul 2024
Amazon AGI *Seattle, WA*

- Led a cross-functional team of ∼14–18 SDEs, SDMs, Applied Scientists, and TPMs; managed the "Manager of Managers" layer for Authentication/Authorization/Presence teams
- Launched the Multimodal Recognition Engine fusing VoiceID, VisualID, PhoneID, and usage patterns to infer user identity, with a cloud-edge synchronization protocol reconciling results in under 8ms
- Delivered VisualID (Facial Recognition) and PhoneID (Bluetooth Proximity) for Alexa personalization, coordinating across 23+ external science, engineering, and CX teams
- Built the ML training pipeline for Continuous Improving Multimodal Recognition – an online learning system refining models via user feedback loops integrated with LLM-powered Alexa+ experiences
- Devised the Authentication Confidence Levels (ACL) security standard at Amazon – a 6-tier scoring scheme adopted company-wide and by external experience builders

Senior Software Engineer – Ambient Recognition & Authentication Nov 2018 – Mar 2020
Amazon AGI *Seattle, WA*

- Founding Engineer for Ambient Recognition & Authentication; architected high-stakes identity and security features resulting in multiple granted US Patents
- Architected "Limit Access" on Alexa – a multi-factor authentication system (Voice PIN + VoiceID) achieving HIPAA compliance for Amazon Pharmacy, and 3P pharmacy partners
- Designed the Person Recognition Identity API for 3rd-party developers, enabling secure personalization of Alexa interactions by 1P and 3P skill builders

- Invented the Cross-Modal Automated Ground Truth scheme (patented) using BLE proximity history to ground-truth voice prints, which brought in 100M+ monthly labels, enabling rapid expansion of the multi-modal person recognition science modeling program

Software Development Engineer – Alexa Identity May 2017 – Nov 2018
Amazon – Alexa Identity *Seattle, WA*

- Founding engineer for Alexa Identity; built foundational identity services, human-in-the-loop ground truthing, offline/online ML pipelines, launched Alexa Voice Training & Recognition (Oct 2017), and unsupervised clustering-based - Automatic Voice Recognition (May 2018)
- Established operational excellence mechanisms ensuring low-latency, high-availability voice biometric services at scale

Software Development Engineer – AWS IAM Aug 2016 – May 2017
Amazon Web Services *Seattle, WA*

- Built the Resource Groups Tagging API enabling tag-based access control (TBAC) at 40K+ TPS
- Redesigned Tagging Discovery Services to eliminate single-point-of-failure by regionally self-containing writes, re-indexing, and reads; deployed to AWS GovCloud with air-gapped security protocols

Software Development Engineer – Amazon Home Services Mar 2014 – Aug 2016
Amazon *Hyderabad, India*

- Founding engineer for Amazon Home Services marketplace; designed seller onboarding flows, ASIN search via ElasticSearch, and AWS Step Functions workflows
- Built a hyperlocal seller notification mechanism reducing customer order claim time by 95% (∼12hrs to ∼30min)

Associate Software Engineer – Symantec NetBackup Jul 2013 – Mar 2014
Symantec *Pune, India*

- Developed FIPS-compliant secure authentication for NetBackup, upgraded OpenLDAP across 28 platforms, and mitigated security vulnerabilities via penetration testing and static analysis (Coverity)

Leadership

Org Leadership: Manager of Managers overseeing SDMs, Sr. SDEs, Applied Scientists, and TPMs across teams of 14–20+; drove org design, charter definition, and 3-year technical roadmaps

Science-to-Production: bridging Applied Science and Engineering – translating research prototypes into high-availability production systems serving millions of users

Talent Development: built and mentored high-performing teams, growing ICs to Staff Engineer, Senior Applied Scientist, Senior TPM, and SDM roles

Program Leadership: led company-level flagship programs applying ML/DL to deliver personalized experiences on consumer devices (Echo, Fire TV, mobile), from zero-to-one ideation through global launch

Education

Birla Institute of Technology and Science (BITS), Pilani 2009-2013
Master of Science (Tech.), Information Systems

Patents (4 Granted, 1 Pending)

Input Processing with Profile Context | *US 12,573,408* (link) Granted Mar 2026
User Identification Attribution for Touch Interactions | *US 12,443,687* (link) Granted Oct 2025
Authenticating a User Profile with Devices | *US 12,236,957* (link) Granted Feb 2025
Presence Data Determination and Utilization | *US 11,437,043* (link) Granted Sep 2022
Multi-Modal Person Recognition | *USPTO – Patent Pending* Filed Mar 2023

Publications

Riemannian Geometry of Multimodal Biometric Embedding Spaces. 2026
A. Upadhyay. *Conference of Mathematics of AI* (link)

Are VLM Identity Judgments Logically Consistent? Evaluating Symmetry, CoT, and Transitivity in Person Re-ID. A. Upadhyay. *ICLR 2026 WS on Logical Reasoning of LLMs* (link) 2026

Do LLM Recommenders Obey Preference Axioms? Testing Logical Rationality in LLM-Based Recommendation. A. Upadhyay. *ICLR 2026 WS on Logical Reasoning of LLMs* (link) 2026

A Novel Architecture for Secure Communications in Mobile Systems 2012
A. Upadhyay, J.K. Sahoo, V. Bajpai. *Int'l Conf. Internet Technology & Secured Transactions* (link)

Additional papers under peer review at TMLR, ACM ICMR 2026, CVPR 2026 GenBio Workshop, and ECCV 2026

Technical Peer Review / Program Committee

ACM ICMR 2026 | *International Conference on Multimedia Retrieval – Reviewed 4 papers* 2026
ICLR 2026 Workshop | *Logical Reasoning of LLMs – Reviewed 3 papers* 2026
ICLR 2026 Workshop | *AI in the Wild – Reviewed 7 papers* 2026
ICLR 2026 Workshop | *Multimodal Intelligence – Reviewed 5 papers* 2026
CVPR 2026 Workshop | *Foundational and Generative Models in Biometrics* 2026