

# Alok Upadhyay

Seattle, WA | [alok0412@gmail.com](mailto:alok0412@gmail.com) | [linkedin.com/in/alok~upadhyay](https://www.linkedin.com/in/alok~upadhyay)

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. 3 peer-reviewed publications including ICLR 2026. 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.

## 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 operationalized a self-hosted GenAI Auto-QA system on AWS Bedrock using LLMs 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, reducing inference cost by 17% through legacy deprecations, and model optimizations 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, of-line/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)

## LEADERSHIP AND TECHNICAL SKILLS

---

### Leadership & Strategy

- Org Leadership & Strategy: Manager of Managers overseeing SDMs, Sr. SDEs, Applied Scientists, and TPMs across teams of 14–20+; drove org design, charter definition, 3-year technical roadmaps.
- Science-to-Production: bridging Applied Science and Engineering – translating research prototypes into high-availability production systems serving millions of users
- Hire & Develop the Best: built and mentored high-performing teams, growing individual contributors to Staff Engineer, Senior Applied Scientist, Senior TPM, and SDM roles; deeply technical and people-oriented leadership with a focus on career development and bar-raising
- Project Leadership: led company-level flagship cross-cutting programs applying ML/DL to deliver personalized experiences for millions of end users on consumer devices (Echo, Fire TV, mobile), from zero-to-one ideation through global launch

### Machine Learning & AI

- Large-Scale Recommendation Systems (RecSys): SASRec, BERT4Rec, Two-Tower DNNs, LightGBM, retrieval & ranking, collaborative filtering, real-time personalization at billions-scale
- Large Language Models (LLMs): fine-tuning (LoRA, PEFT, QLoRA), RLHF, prompt engineering, RAG pipelines, evaluation frameworks, self-hosted inference
- Generative AI Systems: end-to-end GenAI product design, automated content generation & compliance, multi-modal content understanding (text, image, video)
- Multi-Modal ML: person recognition via voice, face, BLE proximity, and behavioral signal fusion; online and offline learning with human-in-the-loop ground truth pipelines
- Frameworks & Tools: PyTorch, TensorFlow, SageMaker, Bedrock deployment, HuggingFace Transformers, LangChain, FAISS

### Engineering & Infrastructure

- Distributed Systems: high-throughput microservices at 40K+ TPS, event-driven architectures, low-latency (<8ms) cloud-edge systems, real-time ML serving
- AWS: EC2, DynamoDB, SageMaker, Bedrock, SQS, SNS, IAM, Step Functions, Lambda, GovCloud (air-gapped deployments), ElasticSearch/OpenSearch
- Data & ML Pipelines: batch and streaming ETL, feature stores, A/B testing frameworks, CI/CD for ML models, model monitoring & observability
- Security & Compliance: HIPAA-compliant system design, multi-factor biometric authentication, FIPS, zero-trust architecture, privacy-preserving ML
- Languages: Java, Python, C/C++, SQL

## EDUCATION

---

**Birla Institute of Technology and Science (BITS), Pilani**

2009-2013

*Master of Science (Tech.), Information Systems*

## PATENTS (4 GRANTED, 1 PENDING)

---

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

## PUBLICATIONS

---

<b>Are VLM Identity Judgments Logically Consistent? Evaluating Symmetry, CoT, and Transitivity in Person Re-ID.</b> A. Upadhyay. <i>ICLR 2026 WS on Logical Reasoning of LLMs</i> ( <a href="#">link</a> )	2026
<b>Do LLM Recommenders Obey Preference Axioms? Testing Logical Rationality in LLM-Based Recommendation.</b> A. Upadhyay. <i>ICLR 2026 WS on Logical Reasoning of LLMs</i> ( <a href="#">link</a> )	2026
<b>A Novel Architecture for Secure Communications in Mobile Systems</b> A. Upadhyay, J.K. Sahoo, V. Bajpai. <i>Int'l Conf. Internet Technology &amp; Secured Transactions</i> ( <a href="#">link</a> ) <i>Additional papers under peer review at TMLR, ACM ICMR 2026, CVPR 2026 GenBio Workshop, and ECCV 2026</i>	2012

## TECHNICAL PEER REVIEW / PROGRAM COMMITTEE

---

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