

Milan Shah

Senior Engineering Manager (Head of Personalization & Applied Science Platforms)

📞 +1-646-647-4178 ✉️ mshahsmv@gmail.com 🔗 www.linkedin.com/in/milan-shah-lin/ 📍 NYC

SUMMARY

Innovative engineering executive with 16+ years of experience, currently leading Amazon's Intent Personalization org of 70+ engineers and applied scientists. My teams deliver real-time large-scale personalized recommendations, and systems that power Search, Homepage, Cart, and Detail Page experiences for **500M+ customers globally**, driving high-impact customer engagement, contributing **\$12B+ in ARR** in 2025 and accelerating Amazon's topline growth.

I define and drive the multi-year technical vision for customer understanding and personalized recommendation engine at scale, leveraging ANN-based retrieval, transformer-based ranker and NLP to deliver nuanced product and intent understanding. My influence extends across **150+ engineers, scientists, and stakeholders** in Search, Ads, and Home, aligning cross-org investments to scale innovation in personalization and advertising.

I excel in building high-performance AI organizations that combine cutting-edge innovation with operational excellence. I am passionate about mentoring engineers and managers, fostering collaborative environments where transformative AI solutions can flourish and drive significant business impact.

EXPERIENCE

Senior Engineering Manager

Amazon 05/2021 - Present NYC

- Lead a **75+ person Personalization Intent organization** (engineering and applied science), setting strategy and delivering AI-powered shopping experiences across Amazon.
- Define the multi-year vision and roadmap for Amazon's recommendation and customer intent platforms, contributing over **\$12B in annual attributed revenue (2025, +120% YoY)**.
- Lead a high-performing organization of Principal Engineers, Applied Scientists, and Engineering Managers, driving technical strategy, organizational scaling, and operational excellence.
- Own Amazon's real-time customer intent platform (**500K RPS / 45B daily transactions**) powering personalization across Search, Homepage, Cart, and Detail Pages.
- Lead strategy, development, and adoption of an AI agent platform enabling autonomous engineering workflows through persistent memory, multi-agent orchestration, and secure integrations across Slack, Quip, Microsoft 365, AWS tools, and other internal systems. Scaled adoption across Amazon Personalization and expanded the platform for multi-organization use, accelerating AI-driven automation and engineering productivity.
- Led development of LLM-generated customer profiles that fused deep learning behavioral signals with semantic customer understanding to power last-mile ranking, improving personalization quality and recommendation relevance.
- Lead Amazon's next-generation AI recommendation platform, leveraging customer intent, embeddings, retrieval, and GenAI to deliver personalized recommendations and explainable shopping guidance across major customer experiences.
- Built an NLP/LLM-powered product intelligence platform that standardizes product understanding and catalog attributes, enabling Search, Rufus, and p13n systems at scale.
- Established a new strategic organization developing ML models and distributed infrastructure for Amazon's prompt-based conversational shopping experiences.
- Drive multi-year investment planning, organizational strategy, and operational planning for the Personalization organization, partnering with senior leadership on long-term priorities.
- Lead cross-organizational execution across Search, Home, Ads, and Personalization, aligning engineering, science, product, and business leaders to deliver AI innovation at scale.

Engineering Manager

Valtech (Formerly Kin+Carta) 02/2019 - 03/2021 NYC

- Transitioned from Principal engineer to an Engineering Manager in Q120.
- Led 15 global engineers; led the implementation for a recommender engine using Google Vertex AI and custom embeddings.
- Built a multi-tenant real-time content personalization engine with support for multi-language and A/B testing infrastructure.
- Built event-driven Microservices. Orchestrated and managed Kubernetes (K8s) clusters across various environments to ensure high availability, seamless scaling, and failover capabilities; optimized pod scheduling and resource allocation, resulting in a 40% reduction in downtime.
- Implemented gRPC-based microservices for low-latency communication, reducing request processing time by 50%.

Principal Software Engineer

JPMorgan Chase & Co. 12/2017 - 01/2019 NYC

- Developed distributed pricing engines, personetics engine, deposit engine and real-time alerts for the Digital Wealth team.

Senior Software Engineer

Lightspeed Systems 04/2015 - 12/2017 NYC

- Led backend development of an AI-powered student web filtering system, serving over 10 million students globally.
- Introduced real-time content classification using NLP and ensemble-based URL scoring, cutting review delays by 70%.

CTO

Propagate 01/2012 - 02/2015 NYC

TECHNICAL SKILLS

🔗 Languages

Python, TypeScript

🧠 ML/AI

TensorFlow, PyTorch, Deep Learning, XGBoost, SageMaker, Bedrock, Transformers, RL, CUDA

★ Infra

AWS, Kafka, Spark, Flink, OpenSearch, CDK, Terraform, Docker, Kubernetes (K8s)

🗄️ Databases

DynamoDB, Cassandra, PostgreSQL/SQL, OpenSearch, Pinecone

👤 Leadership

Org design, strategic planning, OKRs, hiring & performance management, P&L

NOTABLE PROJECTS & INITIATIVES

👥 Cross-Amazon Personalization Guild

- Co-founded internal guild that created shared libraries, best practices, and hiring rubrics across Amazon AI teams.

⚙️ Real-Time Behavioral Personalization Layer

- Built near-real-time streaming pipeline (Kinesis + Flink + DynamoDB) that enabled session-aware ranking across Amazon.

EDUCATION

🎓 Bachelor of Science in Computer Science

Rutgers University

EXPERIENCE

Co-Founder & CTO

Olter Inc 01/2010 - 01/2012 NYC