# RAVI RACHAKONDA

*Staff Machine Learning Engineer · Foundation Models & AI Infrastructure*

+1 (904) 610-0261 · ravikiran2508@gmail.com · LinkedIn · Portfolio

## SUMMARY

Staff ML Engineer with 17+ years architecting and shipping AI systems at Amazon scale — from founding engineer on Alexa to key technical leader on Amazon Nova. Currently titled Senior MLE at Amazon while operating at Staff scope: setting technical direction across 4+ orgs, defining standards adopted by 100+ teams, and owning architecture decisions that directly shape Amazon's most strategic AI products. Specializes in cross-org technical leadership for foundation model pre-training, data infrastructure, and MLOps without requiring formal authority. Co-inventor on 3 granted US patents.

## SCALE AT A GLANCE

| | |
|---|---|
| **Data Infrastructure** | ► **Several hundred machines**<br>► **17 trillion tokens**<br>► **Petabyte-scale multimodal datasets** |
| **Production Systems** | ► **1M+ image generation requests / month**<br>► **10,000+ TPS at peak**<br>► **99.9% availability** |
| **Org Leadership** | ► **100+ internal & enterprise teams served**<br>► **4+ orgs aligned**<br>► **12-person team led directly** |

## EXPERIENCE

### Senior Machine Learning Engineer, AGI Data Scaling · Amazon AGI

*Jan 2026 – Present | Sunnyvale, CA*

- **Cross-org tech lead** for data infrastructure strategy powering Amazon Nova foundation models and the Nova Forge platform — the standardized data preparation platform for foundation model customization across internal and enterprise teams on Amazon Bedrock.
- **Defined the Nova Forge SDK architecture** for foundation model customization, setting the technical standard adopted across all customization workloads; drove adoption across 100+ internal and enterprise teams, directly accelerating Bedrock's enterprise customer onboarding velocity.
- **Established training data integrity processes** for Amazon Nova foundation models that are now standardized org-wide, directly impacting model quality for Amazon's highest-revenue AI product line.
- **Leading strategic customer migration** from manual, ad hoc data prep workflows to Nova Forge's managed end-to-end platform — reducing customer time-to-first-trained-model and expanding Bedrock's addressable enterprise market.
- **Technical bridge between enterprise customers and platform teams** — translating customer requirements into platform roadmap decisions that shape Nova Forge's competitive positioning against OpenAI and Google fine-tuning offerings.

### Senior Research Engineer, Nova Pre & Post Training · Amazon AGI

*2024 – Jan 2026 | Sunnyvale, CA*

- **Cross-org technical lead** for Amazon Nova — Amazon's flagship multimodal foundation model family, now serving enterprise customers on Bedrock and directly competing with GPT-4o and Gemini in the enterprise AI market.
- **Defined org-wide multimodal (vision) pre-training strategy** — aligned data, pre-training, and evaluation teams across org boundaries; authored training protocols adopted as the canonical standard, accelerating Nova's path to production by eliminating duplicated cross-team effort.
- **Standardized the vision-language evaluation framework** across the org, giving leadership a consistent, trusted signal on model quality that directly informed Nova's launch readiness decisions.

- **Drove org-wide adoption of NVIDIA NeMo 2.0, MegatronCore, and MegatronEnergon** as the canonical large-scale training stack — consolidated tooling across teams, reducing engineering overhead and unblocking faster iteration cycles.
- **Reduced training cluster MTTR by 13× (40 min → 3 min)** by redesigning the Energon dataloader — directly improving GPU cluster utilization and reducing wasted compute cost across petabyte-scale training runs.
- Designed and delivered the real-time Avatar demo at **AWS re:Invent 2024** — a flagship public demonstration of Nova's multimodal capabilities that drove enterprise awareness and Bedrock pipeline generation.

### Senior Software Engineer / Tech Lead, FireTV Background & Creative AI · Amazon

*2022 – 2024 | Sunnyvale, CA*

- **Founded the Image Generation initiative** spanning two org boundaries (Alexa AI + FireTV) — shipped voice-driven background generation to millions of FireTV devices, creating a differentiated premium feature that directly supported FireTV's competitive positioning against Roku and Apple TV.
- **Delivered the model productized as Amazon Bedrock Titan Image Generation (Nova)** — now processing 1M+ image generation requests/month in production, establishing Amazon's presence in the generative AI image market and contributing to Bedrock's enterprise revenue growth.
- **Defined technical roadmap and architectural standards** for an end-to-end MLOps pipeline with fully automated CI/CD, reducing model deployment cycle time and setting engineering norms adopted across the org.
- **Founded Creative AI for Amazon Kids** — 0→1 system generating personalized animated stories; ranked top 3 among all Amazon Kids Alexa skills at launch, driving measurable engagement uplift in the Kids subscription tier.

### Senior Software Engineer / Tech Lead, Alexa Astro · Amazon Lab126

*Aug 2018 – 2022 | Sunnyvale, CA*

- **Founded and led the Personality program for Amazon Astro** — Amazon's first consumer home robot — defining technical vision for expressive robot behavior; shipped to ~500 devices at launch, establishing the foundation for Astro's differentiated user experience and repeat engagement.
- **Defined engineering standards and system architecture** for the Personality system; patterns adopted org-wide, reducing duplicated engineering effort and accelerating feature velocity across the Astro org.

### Software Development Engineer / Tech Lead, Alexa Platform · Amazon Lab126

*Sep 2012 – Aug 2018 | United States*

- **Founding engineer on Amazon Alexa** — defined engineering patterns across Weather, Sports, Movies, Notifications, and Spatial Perception verticals during Alexa's critical growth phase; patterns adopted org-wide as Alexa scaled to hundreds of millions of devices globally.
- **Tech lead for Alexa Echo Spatial Perception (2017–18)** — co-authored patents that became foundational Alexa IP, enabling multi-device arbitration that became a core Alexa differentiator.
- **Tech lead for Alexa Notifications (2016–17)** — architected the notification system enabling first- and third-party skills to reach users across all Alexa devices, unlocking a new engagement and monetization channel for the Alexa ecosystem.
- **Scaled Alexa platform services to 10,000+ TPS** at 99.9% availability during peak holiday traffic, directly supporting Amazon's ability to sustain Alexa's market leadership during its highest-growth years.

## TECHNICAL SKILLS

| | |
|---|---|
| **ML & AI** | PyTorch, NVIDIA NeMo 2.0, MegatronCore, MegatronLM, MegatronEnergon, Triton Inference Server, Diffusion Transformers, Multimodal LLMs, Vision-Language Models |
| **Distributed Systems** | Ray, Apache Spark, EMR, Distributed Training, Data Parallelism, Fault-Tolerant Pipelines at petabyte scale |
| **MLOps & Cloud** | SageMaker, SageMaker HyperPod, AWS Batch, ECR, S3, Lambda, DynamoDB, EC2, CloudWatch, CI/CD |
| **Languages** | Python, Java, JavaScript, C++ (intermediate), Perl, Shell |
| **Frameworks** | FastAPI, React, Jupyter, Alexa Skills Kit |

## PATENTS

- **Context-Based Device Arbitration — Awarded Mar 2022 |** US #10546583 · US #11289087
- **Device Selection from Audio Data — Awarded Jun 2020 |** US #10685669

## EDUCATION

**B.Tech, Information Technology** · Indian Institute of Information Technology, Prayagraj, India · 2003 – 2007