

Sainyam Kapoor

hello@sainyam.me

linkedin.com/in/sainyamkapoor

sainyam.me

Senior AI Leader driving frontier foundation model development - from evaluation and benchmarking to model delivery at scale, with deep model-level debugging.

Experience

Senior Manager (AI) **Amazon Nova - AGI Foundations** **Jun 2023 - Present**
(EM → Sr. Manager, Sep'25)

- Led end-to-end evaluation for Amazon Nova and Titan across pre-training, post-training, release, and runtime
 - Measured agentic, multimodal, reasoning, generation, and response quality signals such as repetition and self-doubt across public and internal benchmarks
 - Launched **Amazon Nova 1** at **Re:Invent 2024** (Andy Jassy keynote) - best price performance delivering frontier knowledge at a fraction of the cost
 - Led **Nova 2** benchmarking at **Re:Invent 2025** (Matt Garman keynote) - partnered with 6+ external and 15+ internal customers to represent real use-cases to track as part of model building
- Co-created **OneClickEval** - cut manual eval time by 100s hrs/wk to zero and reduced model-feedback loop from 2 weeks to **1.5 days**. Adopted across all AGI-F teams (~10 verticals, 300+ scientists) and 20+ external orgs. Supports 500+ metrics, 300+ eval datasets with 125+ monthly commits
- Led the pre-production inference team to support model evaluation, enabling rapid experiments with new architectures, extended reasoning, multimodal inputs, and tool use across vLLM, SGLang, TensorRT, and Neuron
- Led Nova / Titan Model Delivery into **Bedrock Production**: Directed 30+ SDEs across tiger teams bridging science specs and production stacks
- Reduced Post-Training + Release cycle time from 3 weeks to **3 days (5x faster)**. Aligned multiple orgs on common model spec, standardized evals, and complex agentic benchmarking

Engineering Manager **Amazon - Alexa NU** **Nov 2019 - Jun 2023**
(SDE → Sr. SDE, Dec'21 → EM, Jun'22)

- Led delivery of Alexa's Context-Aware NLU Stack, enabling intent prediction improvement of over 5% in English and **14%** in multiple languages (FR, HI, ES, JP, etc.)
- Deployed large-transformer models with **sub-100ms latency** at Alexa scale across 100M+ active devices, billions of daily utterances
- Reduced ML-to-prod launch time from 3-4 weeks to 3 days via ECS service with automated safeguards and A/B rollouts
- 2 patents filed, 1 granted

AI Research Resident **Facebook AI Research (FAIR)** **Oct 2018 - Sep 2019**

- Built graph embedding models using BERT representations to explore entity-content relationships and social influence at Facebook scale. Improved content integrity detection by **6%**

Software Development Engineer **Amazon - Alexa Info Federation** **Aug 2017 - Sep 2018**

- Improved Alexa's knowledge-answering accuracy +4% by paraphrasing utterances before querying across 50M daily utterances; containerized ML-to-prod A/B rollout pipeline

Global Alpha Researcher **Trexquant Investment LP** **Jan 2017 - Aug 2017**

- Building trading signals for medium frequency trading.

Education

M.Tech & B.Tech in Computer Science & Engineering
IIT (BHU), Varanasi

May 2017
8.5 CGPA

Skills

AI/ML: LLMs, Model Evaluation, Post-Training, Test Time Scaling, Agentic AI, NLU/NLP, Transformers
Infra: Python, PyTorch, vLLM, SGLang, TensorRT-LLM, AWS Neuron/Bedrock, ECS, K8S