Netflix AI Platform: MLSys 2026

Pooja Maknikar
Engineering Manager, Model Serving Pillar

AI Platform: ML Lifecycle Stages

Data, Observability & Visualization

The foundation every model is built on. AI Datasets, the Feature Store for structured & vector features, and unified AI Observability — logging, drift detection, quality eval across classical ML, LLMs, and agents.

AI Agent, Runtime & Compute

Model development and management, distributed model training, runtime optimization (CUDA, TensorRT, PyTorch), model lifecycle, and the Agent Platform.

100M → 70B Parameter range of models in production

Model Serving

Unified, high-performance serving: live online, near-real-time async, and large-scale batch inference.

Range of models supported: classical ML, generative LLMs, multimodal models etc. 500K+ RPS at peak, sub-ms latency for live inference

AI Developer Enablement

Empowers practitioners to focus on solving business problems by providing stable, reusable, and secure building blocks for GenAI and ML/AI development at scale

AIP at Scale: Powering Business Verticals

Personalization & Recommendations

Title ranking, artwork selection, next-episode recommendation, search, embedding-driven candidate generation, mix of batch precompute and live inference

Growth & Commerce

Pricing, payments & fraud, VPN detection, acquisition, plan upsell- revenue-critical

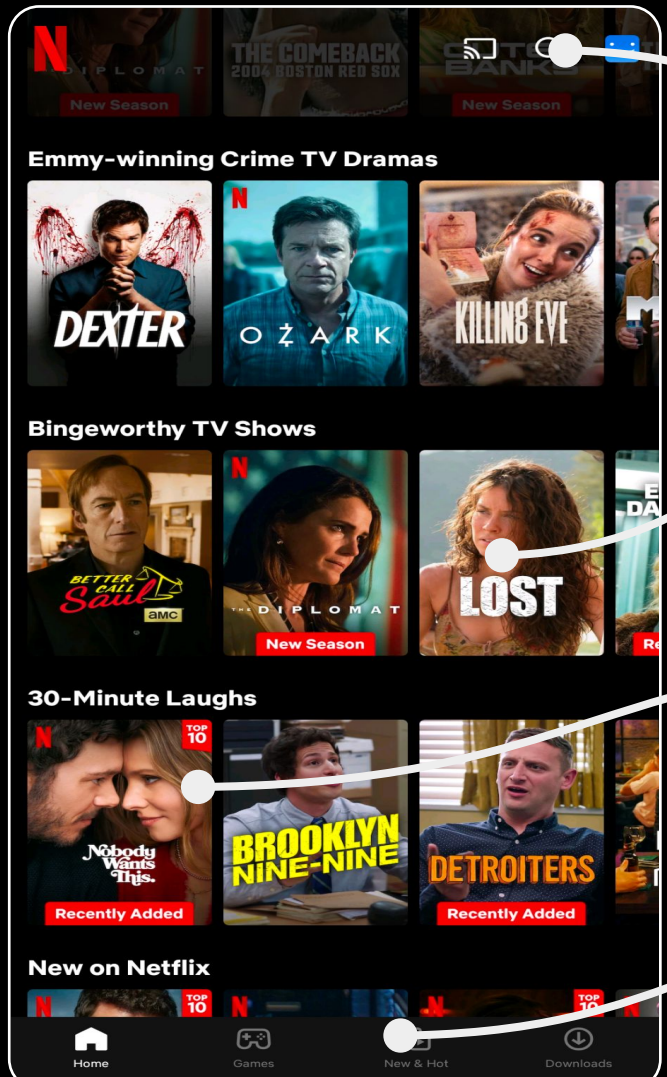
Ads

watch/click prediction, price optimization, ad placement etc - has high RPS and strict latency

Content & Studio

budget forecasting, multimodal tagging, subtitles, dubbing, lip sync - offline batch or near real time use cases

and multiple other business vertical.



Personalized **Search**

Personalized **Assets**

Personalized **Recs**

Personalized **Messages**

We are hiring for multiple positions

STAY CONNECTED
WITH NETFLIX

NETFLIX