Ritesh Pallod

Senior Staff ML Engineer

ritesh.pallod7@gmail.com +919763036878 riteshpallod

PROFILE

Machine Learning Architect with deep expertise in building & scaling Generative AI systems for millions of users. Proven track record of slashing inference costs by 93% and driving AI driven significant revenue growth across global markets leading to 20K USD DRR. Skilled in guiding high-performance teams to bridge the gap between complex R&D and production-grade business outcomes.

PROFESSIONAL EXPERIENCE

Glance

Senior Staff MLE 01/2025 – Present | Bangalore

- **Strategic Team Leadership**: Built and led a high-performing ML team, defining the technical roadmap for ML for "Glance AI." Managed hiring, upskilling, and performance, successfully balancing deep R&D initiatives with business-critical delivery (BAU).
- **Massive Cost Reduction (93%) for Image Generation Models:** Spearheaded the "Fractini" initiative, reducing avatar generation costs from 3¢ to <0.21¢ per unit by leveraging TensorRT, quantization, and CPU-bound consumers. Our models comprise of Qwen Image Edit, SDXL, CN, Loras, GANs. Currently working on exporting the models for Triton.
- **Gen AI Production Scale:** Architected a stable Generative stack supporting 1.1M profiles and ~500k DAU, capable of handling millions of users. Gatekept production standards by implementing mandatory OpenTelemetry (OTEL) and strictly enforcing IaC/CI/CD protocols. Unified system observability into a "Concurrency" dashboard, consolidating metrics across GPU resources (Alloy, Redis, Vector Search) to minimize Time-To-Detection (TTD) for production outages.
- **IP Independence & Innovation**: Led the development of "Fractini 2.0" (Internal Virtual TryON & Edit models), eliminating reliance on third-party vendors and enabling a commerce-first, fully automated pipeline with zero human moderation. Virtual TryOn models built on top of CAT VTON & Synthetic Datasets that we generated.

Staff MLE 01/2024 – 12/2025

- **Generation Models**: Part of a 4 member team that built the first Avatar Generation model stack for "Glance AI". Models comprised of combination of SDXL, Flux, Controlnet, LORAs & LLMs
- **Product Launch & Scale**: Led the end-to-end execution of ML models for "Glance AI" (Avatar, Reference Image Generations, Search), scaling the backend to handle 200 QPM on constrained GPU infrastructure
- **Automation of AI Generated Content at Scale:**
  - Spearheaded the integration of Generative AI in creating dynamic video, landing pages, AMP, and quizzes.
  - Achieved 100% automated content generation for English, Spanish, Portuguese, and Hindi markets, and fully automated publishing for News, freeing editorial teams to focus on high-value curation.
  - This included using LLMs like GPT 3, finetuning Stable Diffusion models & building Internal Image Search system.
- **Revenue & Monetization:** Drove the "AIGC Publisher" track to unlock liquidity in key markets, achieving $3k/day revenue in Brazil and fully automating content streams for US Tracfone (20% liquidity auto-generated).
- **Stakeholder Management:** Served as the primary AI counterpart to Product Directors (Glance AI, News units), translating ambiguous business goals into executable technical architecture.

ML Engineer, AI & Platforms 2018 – 12/2023 | Bengaluru, India

- **CMS & Content Curation Efficiency:**
  - Pioneered the use of AI in news content curation to enable a transition from manual curation to automated moderation.
  - **Increased content liquidity by 2x** and operational efficiency by automating summarization, categorization, image selection, and metadata enrichment.
  - Work included using BART, T5 & then GPT 3 for summarization, and building an Image Search system leveraging CLIP, Efficient Net & BGE Large
- **"Fast Cards" Personalization:**
  - Built the initial set of personalization models for an infinite feed of news and sports cards to increase user engagement.
  - Work included requirements for signal gathering, generation of features, and LR model for dense users & popularity for sparse users
- **ML Platform:**
  - Part of the team building an ML Platform on top of GCP. Productionizing GANs, recommendation systems at scale, data processing in PySpark & building ML based internal tools end to end.
  - We got our serving infra to do **10 million predictions per second, 80k QPS within 250 p99 ms.**
  - Our stack comprised of highly performant Python servers with scoring functions written in C++ to bypass GIL and a model controller in Go.
- **Cross-Functional Team Role:**
  - Acted as the key stakeholder liaising between engineering, product management, and editorial teams in a fast-paced, agile environment.

EDUCATION

BTech in Computer Science, *Pune Institute of Computer Technology* 06/2014 – 05/2018 | Pune, India

SKILLS

ML Models & Inference: Models: Logistic Regression, Transformers, CLIP, Stable Diffusion, Controlnet, Lora | **Inference:** Ability to run an ML model at cheap using ML inference optimizations like TensorRT, quantization and infra choices like Compute Classes, preemptable VMs, price to performance VM metrics

Language & Frameworks: Proficient in Python, with frameworks including PyTorch, numpy, pandas, fastapi, PySpark | Working proficiency in Java, Go

Platforms: Experienced in working on GCP & Azure. ML & Data Platforms, relational & non relational databases, Kubernets, storage.

PUBLICATIONS & PATENTS

LookSync: Large-Scale Visual Product Search System for AI-Generated Fashion Looks, *CODS* 12/2025

Product Recommendation System for Avatars. Built in two parts - Candidate Generation with CLIP & CLIP like models and a Ranker with an LLM.