VISHNU ANILKUMAR

Lead Machine Learning Engineer · Production GenAI · LLM Agents & MCP · MLOps
Bengaluru, India · +91 8943372700 · vishnuanilkumar122@gmail.com
linkedin.com/in/vishnuverse · github.com/vishnuverse

SUMMARY

Lead ML Engineer (8 years) owning the AI Python backend for two production GenAI platforms at Pixis — a 25-model creative-generation pipeline (SDXL, ControlNet, custom Flux fine-tunes, UGC video with talking-human avatars) and a multi-tenant competitor-intelligence service that ships 20+ LLM-driven analysis reports per snapshot and exposes 11 tools over an MCP server to AI agents. Built unified provider abstraction routing across OpenAI, Anthropic, Gemini, FAL, Replicate, and Baseten with webhook-driven async fan-out. Combine deep diffusion + LLM/agent expertise with hard MLOps discipline (Kubeflow, GPU autoscaling on EKS, drift monitoring, release gates) and cross-team technical leadership.

CORE EXPERTISE

Generative AI & Diffusion — Stable Diffusion XL, ControlNet, custom Flux fine-tunes, multi-stage diffusion pipelines (LaMa inpainting, harmonization, upscaling), UGC video generation with avatars and voice cloning (ElevenLabs), Hugging Face Diffusers and Transformers.

LLMs, RAG & Agents — Multi-provider orchestration (OpenAI GPT-4o / GPT-4o-mini / o3, Anthropic Claude, Google Gemini, Whisper-1), RAG with Chroma, LangChain, MCP (Model Context Protocol) servers — built and shipped two in production (FastMCP, 11+ tools, integrated with Claude Desktop and Cursor), agentic workflows, prompt engineering, evaluation framework design, context management, prompt caching implementation, and token optimization strategies.

ML Architecture & System Design — Async microservices (FastAPI + asyncpg + SQLAlchemy 2), distributed task orchestration (ARQ, Redis, RabbitMQ), webhook-driven model serving (Baseten, FAL, Replicate), multi-tenant data isolation, credit-aware execution, self-healing job pipelines, dual real-time + batch systems.

MLOps & Production Lifecycle — Kubeflow on GCP AI Platform, GPU autoscaling on AWS EKS, model / dataset / code versioning, custom drift monitoring, release gates and rollback strategy, New Relic APM, structured logging with task-scoped contextvars.

Cloud & Data Infrastructure — AWS (EKS, EC2, S3, CloudFront), GCP (AI Platform, BigQuery, Kubeflow, Vertex AI), Docker, Kubernetes, PostgreSQL, MongoDB, Redis, Chroma vector DB, PySpark.

ML / DL Stack — Python, PyTorch, TensorFlow / Keras, Scikit-learn, OpenCV, XGBoost, BERT, YOLO, CNN architectures, Pandas, NumPy.

PROFESSIONAL EXPERIENCE

Lead Machine Learning Engineer — Pixis

Bengaluru, India · Nov 2022 – Present

Own the AI Python backend across two production GenAI platforms serving enterprise advertisers — Adroom (creative generation) and Competitor Insights (competitive intelligence with AI-agent surface).

Adroom Creative Playground — multi-modal GenAI platform

- **Built AdRoom, an AI-powered creative automation platform from inception to $2M+ ARR;** led end-to-end product development including roadmap, design, sprints, and a 4-person team through v1 launch.
- Architected and operate a creative-generation pipeline orchestrating **25+ Baseten-hosted models** (SDXL, ControlNet, custom Flux fine-tunes, LaMa inpainting, harmonization, upscaling) plus FAL, Replicate, OpenAI, Anthropic, and Google Gemini APIs — supporting controlled product placement, free-form generation, template cloning, and full ad generation across **43+ API surfaces and 38+ services**.
- Built **UGC video ad pipeline** (talking-human avatars + ElevenLabs voice cloning + post-processing with vision-based scene understanding) that automated video creative production previously bottlenecked on actors and studios.
- Designed **unified provider-abstraction layer** with per-provider webhook callbacks, async fan-out via ARQ workers, and bulk-execution batches with parent-child task aggregation — handles thousands of creative jobs per execution batch with task-level retry and reliability. Stack: FastAPI, ARQ, Redis, RabbitMQ, PostgreSQL (asyncpg + SQLAlchemy 2), AWS EKS + S3 + CloudFront, New Relic APM.
- Drove production LLM cost & latency optimization through **context management, prompt caching implementation, and token optimization strategies** across the multi-provider orchestration layer.

Competitor Insights — multi-tenant LLM analytics service with MCP exposure

- Designed and shipped a workspace-scoped microservice that ingests Meta Ad Library snapshots and produces **20+ LLM-driven strategy reports per snapshot** — using GPT-4o-mini for text reports, GPT-4o vision for opening-frame and creative analysis, OpenAI o-series reasoning models for executive summaries and key takeaways, and Whisper-1 for video-ad audio transcription.
- Built an **MCP server exposing 11 read-only tools** over Streamable HTTP, integrated with Claude Desktop and Cursor — extends the same data backend to AI copilots without duplicating business logic, with API-key auth, context resolution, and SQL-injection guards.

- Implemented **credit-aware execution** (per-report pricing with media-type weighting), recurring sync and report schedules (cron-driven, every 6h), self-healing for stuck jobs with Slack alerting, and Gamma.app-powered PDF / PPTX export for client deliverables. Stack: Python, FastAPI, ARQ, PostgreSQL, Redis, RabbitMQ (Kombu fanout), S3 + CloudFront, ffmpeg, FastMCP.

Cross-platform engineering leadership

- Designed and deployed a **RAG-based LLM agent for pandas data operations** — owned dataset design, training, evaluation, and serving — automating analytical workflows for non-technical users.
- Define the ML technical roadmap and lead cross-team design review of models, datasets, and evaluation patterns; drive MLOps discipline across both products — versioning, drift monitoring, GPU autoscaling, release gates, and rollback strategy on production inference endpoints.

Senior Machine Learning Engineer — Quantiphi Inc.

Bengaluru, India · May 2021 – Nov 2022

- **Dawn Foods (Customer Lifetime Value):** Led the engagement end-to-end as Senior MLE; delivered a production CLTV model with monthly revenue as the North Star metric — covered data onboarding, modeling, and stakeholder sign-off.
- **Dyson UK (Production ML pipeline):** Designed and deployed a Kubeflow pipeline on GCP AI Platform with custom model monitoring, delivering an ensemble-based CLTV model into production.
- Mentored a team of seven MLEs across multiple engagements; advised R&D and product teams in marketing analytics. Recognised as **Employee of the Year** for delivery quality and team leadership.

Machine Learning Engineer — Reflections Info Systems Pvt Ltd

India · Apr 2020 – May 2021

- **Skill-extraction CNN:** Developed and deployed an entity-recognition model for an enterprise hiring platform — reached 82% accuracy and materially improved skill detection and candidate matching.
- **Legal-clause semantic search:** Built a BERT + Elasticsearch search engine enabling faster clause comparison and retrieval across large legal corpora.
- **Signature and stamp-seal detection (YOLOv3):** Computer-vision model automating proofreading of legal documents.

Associate Data Scientist — Techvantage Systems Pvt Ltd

India · Jun 2018 – Mar 2020

- **Resume intelligence platform (client-facing):** Candidate scoring, section identification, content-quality evaluation, and alternative-industry / role prediction using TensorFlow and Random Forest; fine-tuned BERT for role-prediction from experience profiles.
- **Customer-churn prediction:** Achieved 82% accuracy across 22,000 customer records for a financial-services client using XGBoost; supported targeted retention campaigns.
- **Credit-risk analytics:** Probability-of-default modelling from KYC data using Regression, SVM, Random Forest, and XGBoost with SMOTE, anomaly detection, and feature-selection pipelines.

EDUCATION

M.Sc. Computer Science (Machine Intelligence) — Indian Institute of Information Technology and Management, Kerala (CUSAT), India