DEEPAK KARKALA

Senior ML Engineer | Founding AI Engineer, Vavi Labs

dkarkala01@gmail.com | +91 9480401270 | Mangalore, India | Open to Remote (Global)

linkedin.com/in/deepak-karkala | github.com/deepak-karkala | deepakkarkala.com

SUMMARY

Senior ML Engineer and Founding AI Engineer at Vavi Labs with 6+ years of experience in ML/AI (2018-present), shipping production ML systems delivering measurable business impact across e-commerce and IoT (12% higher marketing ROI, 5% conversion uplift, 4% search-to-purchase improvement, 20% reduction in maintenance callouts). At Vavi Labs, building Arbiter for governance of architectural decisions, BatSwing for computer-vision based cricket coaching reports, Creative Collab OS for creative AI workflows with better judgement, AI/ML interview preparation platform, Agentic-AI plugins for coding agents. Experience across the full ML lifecycle - problem framing, feature engineering, training, pipelines, evaluation, deployment and monitoring. Specialized in cost-effective, moderate-scale MLOps with minimal infrastructure footprint. Holder of 1 granted US patent.

SKILLS

ML/AI: Machine Learning, MLOps, Computer Vision, Time Series Forecasting, LLMs, RAG, Fine-tuning, Anomaly Detection
Frameworks: Python, PyTorch, XGBoost, scikit-learn, LangGraph, LangChain, FastAPI, SQL, Spark, Claude Code, Codex
Cloud/Infra: AWS, SageMaker, CI/CD, A/B Testing, Terraform, Docker, vLLM, Pinecone, Git
MLOps: Monitoring, Data Pipelines, Drift Detection, Calibration, Model Serving, Evaluation, Feature Engineering, SLOs

WORK EXPERIENCE

Founding AI Engineer **Feb 2026 - Present**
Vavi Labs *Mangalore, India*

- **Arbiter:** Building and beta testing a CLI-first decision-intelligence platform that scores architecture choices using GitHub/Jira/PagerDuty context, tenant-isolated audit trails, deterministic + LLM confidence scoring.
- **BatSwing:** Built a PWA + MediaPipe cricket-vision pipeline that converts batting clips into branded parent-facing reports; currently processing a few dozen reports weekly from one academy while tuning the pipeline.
- **Creative Collab OS:** Built a beta-stage creative AI studio using LangGraph/FastAPI workflows, direction checkpoints, reusable taste profiles, multimodal artifact generation, and long-lived project context for stand-up, songs, comics, sitcoms, and memes.
- **Tech Abstractions (techabstractions.com)**: Built an AI/ML interview prep platform focused on production reasoning, system design, LLM infrastructure, and applied practice; grew waitlist to a few dozen users.
- **Dev tools and explainers:** Created plugins for coding agents (**Agentic AI Skills**, **Production MLOps Skills**), and illustrated technical explainers on **LLM Inference**, **Coding Agent**, RLHF, Statistics for MLOps.
- **Fine-tuning projects:** Fine-tuned small language models for **Sitcom Scriptwriter** and **Feynman style Kannada Physics Tutor** built datasets, SFT/RFT-style training workflows, evaluation harnesses for domain-specific generation and tutoring.

Senior ML Engineer (Contract) **Aug 2022 - Jan 2026**
Confidential Client - Mid-sized European E-commerce Marketplace (75K SKUs, 50K DAU, 2.5K orders/day) *Remote*

- Led end-to-end delivery of **4 production ML systems** driving **12% higher marketing ROI**, **+5% conversion uplift**, and **+4% search-to-purchase** across millions of sessions; owned cost-bounded, minimal-footprint AWS MLOps.
- **RAG Search (Hybrid Retrieval, Ranking):** Led design of **BM25 + dense vectors** with re-ranking to define latency/error SLOs and cost targets upfront. Built evaluation framework (offline metrics, CI/CD regression tests, load envelopes) including scalable LLM-driven approach to create **golden dataset** for ranking evaluation. Architected for scale (**3M queries/month**) and delivered **p99 latency under 500 ms** at **EUR 0.001 per query**, improving **search-to-purchase by 4%**.
- **Review Summarization (LoRA, vLLM):** Fine-tuned **Mistral-7B** on **100K+ reviews**; cut LLM cost **60% vs GPT-4 baseline** while maintaining quality. Partnered with product and legal to define governance standards (model cards, versioning, bias-aware retrieval, PII redaction); productionized batch RAG with automated quality checks (hallucination/toxicity/relevance).
- **CLV Prediction (Batch MLOps, XGBoost):** Built segmented CLV models (early-stage vs established); improved **RMSE by 15-20%**, **Gini 0.47 to 0.62** for budget-constrained CRM, enabling **12% higher marketing ROI**. Diagnosed feedback loop where top-decile targeting reinforced bias; led cross-functional exploration budget experiment (10% random allocation), retrained on less-biased data for better high-potential discovery. Shipped weekly pipeline scoring **300K+ profiles** with segment monitoring.
- **Real-Time Purchase Intent (Low-latency, Streaming):** Deployed sub-second scoring for **60K+ daily predictions** (Feast + ElastiCache) with **Spark Structured Streaming** features (**50K+ events/hour**) and class-weighted LightGBM (20:1 imbalance);

optimized feature fetch to achieve **40% p99 latency reduction**. Resolved production calibration breakdown eroding marketing trust; added segment-wise reliability analysis + post-hoc calibration, restoring confidence. Designed end-to-end A/B testing strategy delivering **+5% conversion uplift**.

Co-Founder - Social Impact Ventures **Jan 2021 - Jul 2022**
Spiticart, Rumi Schools *Bangalore, India*

- Built and piloted two social-impact ventures - an artisan-first D2C handicrafts marketplace (spiticart.com) and a digital-first rural education model (rumischools.org) - then transitioned back to ML engineering.

ML Engineer **Jun 2018 - Dec 2020**
eSMART Technologies *Renens, Switzerland | Remote*

- Owned end-to-end IoT ML systems for smart-building operations across 3000 apartments, spanning data quality checks, evaluation, monitoring, and scheduled retraining.
- **Predictive Maintenance (Anomaly Detection + HITL)**: Deployed anomaly detection + alert triage for heating systems; reduced emergency maintenance callouts by **20%**. Diagnosed production alert-fatigue issue causing low adoption; partnered with maintenance teams and leadership to redesign system as a prioritization tool vs. fully automated detector. Evolved approach from unsupervised (residuals + LOF) to supervised as labels grew via technician human-in-the-loop validation; achieved **75% Precision@50** for high-priority alerts and restored stakeholder trust.
- **Energy Forecasting (Smart Energy Advisor)**: Delivered 24-hour ahead energy demand forecasting (XGBoost + weather, lag, rolling windows, holiday features) powering resident-facing recommendations; achieved **under 10% MAPE** and **10 percentage points** increase in solar self-consumption. Partnered with domain experts to design tiered cold-start strategy for newly commissioned buildings (physics-informed heuristics -> archetype models -> individualized forecasts). Implemented walk-forward validation, baselines (ARIMA/Prophet), and drift monitoring.

ML Research Intern **Feb 2017 - Jul 2017**
NEC Labs America *Princeton, NJ, USA*

- **Traffic Surveillance Object Detection**: Fine-tuned YOLOv2 for tandem-motorbike detection in traffic surveillance video; tuned anchor boxes/heads + targeted augmentation to surface candidate "tandem near pedestrian" incident windows for human review.
- **Scene Understanding (Visual Relationship Detection)**: Built a PyTorch subject-predicate-object relationship detector with scene-graph outputs; engineered spatial features and explored graph/message-passing + translation-embedding formulations to improve predicate classification robustness for downstream incident reasoning.

Signal Processing Engineer **Feb 2012 - May 2015**
Signalchip Innovations *Bangalore, India*

- Developed and validated WCDMA uplink receiver algorithms (Path Searcher, RAKE, Multiuser detection) to meet 3GPP NodeB receiver conformance BER under 0.001 across multipath/interference channel conditions via link-level simulations and optimization.

Junior Research Fellow **Jun 2011 - Jan 2012**
Indian Institute of Science (IISc) *Bangalore, India*

- Published first-author paper in **Medical Physics** on DOT measurement-selection optimization; cut data acquisition time by 46% with under 1% reconstruction quality loss, and identified 20% more independent measurements vs. prior singular-value methods.

EDUCATION

MS Communication Systems **Aug 2015 - May 2018**
Ecole Polytechnique Federale de Lausanne (EPFL) *Lausanne, Switzerland* | GPA: 5.25/6.0

- Thesis: Data Analysis and Anomaly Detection in Buildings Using IoT Sensor Data

BE Electronics & Communication Engineering **Aug 2007 - Jun 2011**
RV College of Engineering *Bangalore, India* | GPA: 9.23/10.0

PUBLICATIONS, PATENTS

Patents:

- US9602240 (granted patent): Method and System for Symbol Level Interference Cancellation for Multiuser Detection (2017)
- US20160365991 (application publication): Method and System for Performing Optimized Channel Estimation (2016)

Publication:

- First-author, Medical Physics Journal (2012): Data-resolution matrix optimization for diffuse optical tomography