

Alakh Sharma

Principal ML Engineer – LLMs & Agents, Generative Video, Ad Tech

📍 Pune, Maharashtra ✉ alakhsharmacs@gmail.com ☎ +91 77372 63588 in alakhsharma 🌐 alakhsharmaa

Professional Summary

Principal ML Engineer (M.Tech, IISc Bangalore) with 10 years of experience building production ML systems across LLMs & agents, generative video, and ad tech. Lead a team of 5 data scientists shipping multi-agent platforms and fine-tuned SLMs (**70% of conversations resolved unescalated**, **2× lower serving cost**), distributed diffusion-video inference (**47% latency reduction**), and real-time bidding optimization at 50B+ requests/day (**10%+ revenue uplift**, **40% low-value traffic filtered**).

Skills

Languages: Python, SQL, Java

LLMs: Fine-tuning (QLoRA, PEFT), LLM Distillation; Familiar: RLHF, PPO, DPO, GRPO

Agentic AI: RAG Architectures, Multi-Agent Systems, LangChain, LangGraph, LangFuse

Machine Learning: CatBoost, XGBoost, Multi-Armed Bandits, A/B Testing, Survival Analysis, Statistics

Deep Learning & CV: PyTorch, Transformers, Diffusion Models, YOLO, MediaPipe, OpenCV

Distributed & Serving: DDP, FSDP, DeepSpeed Ulysses, Multi-GPU Inference, vLLM, ONNX

Cloud & Data: AWS, GCP (AutoML, BigQuery), RunPod, Docker, PySpark, PostgreSQL, PGVector

Experience

Principal Software Engineer – Data Science

Talentica Software Pvt. Ltd.

Pune, Maharashtra

Oct 2019 – Present

- **[Classical ML] Real-Time Bidding – Bid Filtering & Floor-Price Optimization**
 - Architected a real-time floor-pricing engine using a Multi-Armed Bandit layer over CatBoost base-price regression, serving **50B+ bid requests/day** across **100+ bidders**.
 - Built a CatBoost response-probability filter at **1ms p99** latency with a dynamic threshold capping revenue loss at **~1%**, eliminating **40% of low-value traffic** and cutting infrastructure cost by **30%**.
 - Delivered **10%+ revenue uplift**, validated via closed-loop A/B testing and continuous recalibration.
- **[Computer Vision] Expression & Pose Transfer – Talking-Head Animation (Image-to-Video)**
 - Architected an end-to-end GPU video synthesis pipeline pairing a landmark detection model with a diffusion model conditioned on body pose and facial features for expression transfer.
 - Trained an earlier talking-head model from scratch with multi-GPU training on **8 GPUs** (~10 hrs of video).
 - Deployed a **10B+ parameter** diffusion model sharded across 4 GPUs and cut per-clip latency **47%** (170s → 90s) by generating at half frame rate with frame interpolation.
- **[SLM Fine-Tuning] Resume Intelligence – CV Parsing & JD Relevance Scoring**
 - Replaced a frontier-LLM system with two self-hosted SLMs (QLoRA/PEFT) for data privacy, distilled from human-validated frontier-LLM labels and served with **vLLM** on AWS at **2×** lower cost.
 - **Parsing:** fine-tuned **Qwen3-0.6B** to extract structured CV fields at **86%/83%** precision/recall.
 - **Relevance scoring:** fine-tuned **Qwen3-1.7B** for explained JD–resume fit scoring at **84%/88%** precision/recall.
- **[Generative AI] Multi-Agent AI System – Support Automation**
 - Architected a production multi-agent support system (**LangGraph**) serving 500K users/month, with an intent router delegating to FAQ, recommendation, and ticketing agents, resolving **70%** of conversations unescalated.
 - Built structured memory, state management, failure recovery, and evaluation/observability (LangFuse).

- **[Generative AI] Scalable Multimodal RAG & NL2SQL Assistant**
 - Architected an enterprise NL2SQL assistant with schema linking, query validation, and self-correction, reaching **87%** execution accuracy and cutting manual SQL effort by **70%**.
 - Designed a multimodal RAG framework over text, image, video, and transcribed audio, using dense retrieval with metadata filtering and BGE-reranker-v2 reranking on **PGVector** over ~1M chunks at **92%** recall@15.
- **[Computer Vision] Real-Time Document Alignment & ID-Masking System**
 - Designed and deployed a real-time document pipeline (YOLOv8, MobileNet, OpenCV) for orientation correction and sensitive-ID masking, serving a production API at sub-**200 ms** end-to-end latency.
 - Achieved **4° RMSE** orientation error and **99.5% recall** on identity-field detection, prioritizing recall to minimize PII leakage.
- **Additional ML Systems**
 - Built an automated forecasting pipeline on **GCP (AutoML, BigQuery)** over geographic, census, and demographic data, with KS-test-based drift monitoring triggering automated retraining, achieving **0.90 R^2** .
 - Applied survival analysis (Cox PH) to model time-to-engagement and optimize email send timing, lifting open rate by **1.5%**.
 - Built a multi-object tracking system combining **YOLO** detection with **DeepSORT** for identity-persistent tracking across video streams at **20 FPS**.

Software Engineer 2

Citrix R&D India Pvt. Ltd.

Bangalore, Karnataka

July 2018 – Oct 2019

- Built anomaly detection models (Isolation Forest) on server response telemetry, reducing false positives by **30%**.

Software Engineer

Soft Solutions

Jaipur, Rajasthan

July 2013 – July 2015

- Led backend development (Java, MySQL) and on-premise deployment of school and inventory management systems for small businesses.

Education

Indian Institute of Science (IISc), Bangalore

M.Tech in Computer Science and Automation (Machine Learning)

July 2016 – June 2018

- Studied multi-class imbalance; benchmarked SMOTE resampling and cost-sensitive learning on macro-F1.
- Implemented hard-debiasing (Bolukbasi et al.) to reduce gender bias in word embeddings.

Rajasthan Technical University, Kota

B.Tech in Computer Science Engineering

July 2009 – June 2013

Achievements & Technical Writing

- **AIR 45, GATE 2016** – Computer Science
- **Technical Writing**
 - Adopting AutoML: Let's Do a Reality Check – *VentureBeat*
 - Beyond Bots: Rethinking AI Support with a Hybrid AI Architecture – *KDnuggets*
 - Optimizing AI Performance: Contextual Input vs. Fine-Tuning in LLMs – *Analytics Insight*