Abhishek Saxena

AI/ML Solution Architect | 9+ yrs | GenAI, RAG & LLM Platforms | Self-Hosted AI on Azure

Bangalore, India | abhishek2024@gmail.com | +91 96859 20902 | linkedin.com/in/abhishek-saxena-ab20965b

PROFESSIONAL SUMMARY

AI/ML Solution Architect with 9+ years designing and building production AI products that automate slow, manual business processes. Owns the full lifecycle, from problem framing and architecture through build and production operations, and is strong on the architectural calls that matter: self-hosting open-source models for cost and data control, cost vs. quality trade-offs, and how much autonomy to give a system. Deep travel and hospitality experience, having built a portfolio of AI platforms for a 10,000+ campsite ecosystem (Eurocampings and Suncamp) across generative AI, NLP, computer vision, and recommendation.

SELECTED ACHIEVEMENTS

- Sole AI/ML architect and builder behind six production AI products delivered end to end for a 10,000+ campsite travel platform, covering generative AI, NLP, computer vision, and recommendation.
- Standardized a self-hosted, open-source AI platform on Azure AKS (vLLM, Qdrant, Ray Serve) reused across the AI products, using compact open-source models that run locally during development and on modest cloud GPUs in production, keeping full data ownership, cost control, and governance of the model layer.
- Cut LLM-serving GPU cost with tiered model routing that serves most traffic on lightweight models and keeps larger models for complex queries, and reduced moderation compute by handling deterministic cases before involving any AI.
- Set up cross-product MLOps and governance practices (evaluation, distributed tracing, model and prompt versioning, canary rollout) that moved early PoCs and notebook pipelines into reliable, monitored production.

TECHNICAL SKILLS

Generative AI & LLMs: RAG, agentic orchestration, multilingual content generation, guided/structured (JSON) output, self-hosted model serving with vLLM, tiered model routing, LoRA/QLoRA fine-tuning; Qwen3 (4B/8B), Llama 3.1 (8B), Mistral-7B, Qwen3Guard

Retrieval & Recommendation: Qdrant (HNSW, hybrid dense+sparse, RRF), embeddings (BGE-M3, Qwen3-Embedding), cross-encoder reranking, TF-IDF and weighted ranking, learning-to-rank, ONNX scoring, BlueConic

ML & NLP: Multilingual transformers (mmBERT, XLM-RoBERTa, RoBERTa, BERT), hierarchical and product-aware classification, content moderation, fastText; scikit-learn, NLTK

Computer Vision: YOLO (v8, v11, v26), BoT-SORT tracking, EfficientNetV2, SSD/VGG-16, OpenCV/RTSP, ONNX Runtime

MLOps & Observability: MLflow, OpenTelemetry, Prometheus, Grafana, Jaeger, Azure Monitor/App Insights; offline and online evaluation, A/B and shadow testing, drift monitoring, model/prompt/corpus versioning, canary rollout

Cloud & Platform: Azure AKS, GPU nodes (T4, A10), Ray Serve, FastAPI, Redis and Redis Streams, Kafka, MongoDB, MySQL, Docker, Helm, Terraform, ACR, Azure Key Vault, GitHub Actions, Azure DevOps

Languages & Frameworks: Python, PyTorch, TensorFlow / TF Serving, ONNX, FastAPI, Flask

PROFESSIONAL EXPERIENCE

Eastern Enterprise | Pune, India (Remote) | 2020 to Present

AI/ML Solution Architect *(promoted from Data Scientist)*

Client: ACSI (Netherlands), a European travel and hospitality company running a 10,000+ campsite ecosystem.

Worked within ACSI's software product development team of around 40 people, organized into 8 to 9 sub-teams covering mobile apps, websites (Eurocampings, Suncamp, and other ACSI portals), CRM, and integrations. As the only AI/ML person on the team, designed and built every AI component end to end and worked with each sub-team to integrate AI into their products.

Personalization & Recommendation Platform*: personalized campsite recommendations across email, the website, and real-time scoring for Eurocampings and Suncamp.*

- Designed and shipped a four-channel recommendation product covering lifecycle email campaigns, on-site personalization, real-time scoring, and alternative-campsite suggestions for non-bookable requests, replacing manual, generic handoffs with automated personalized targeting that improved campaign engagement and booking conversion.
- Built a hybrid ranking approach that combines content similarity with weighted business rules across 10+ dimensions, and unified the offline and ONNX real-time scoring paths through a shared feature and weight contract so online and offline results stay consistent.

  Stack: *Python, scikit-learn, TF-IDF with weighted reranking, ONNX Runtime, BlueConic AI Workbench, Azure.*

Multilingual Support Email Automation*: automated triage and draft replies for thousands of multilingual support emails a day.*

- Built a five-stage decision engine (rules, multilingual classification, API enrichment, grounded drafting, then human review) that routes each email automatically and drafts agent-ready replies, cutting handling time and misroutes.
- Chose a layered design over a single end-to-end LLM so routing stays accurate, facts come from authoritative APIs, and the LLM only handles phrasing instead of inventing business data.

  Stack: *mmBERT and XLM-RoBERTa, fastText, Ray Serve, vLLM (Llama-3.1-8B), FastAPI, MLflow, OpenTelemetry, Azure AKS.*

Review Moderation & Content Quality*: automated multilingual review moderation across the 10,000+ campsite catalog.*

- Designed a hybrid moderation product that pairs a versioned policy engine with an AI moderation model, auto-actioning clear cases and sending uncertain ones to human reviewers with plain-language explanations, which cut manual workload while keeping human oversight.
- Kept deterministic cases (profanity, PII, competitor references) away from the LLM to reduce compute, and enforced schema-constrained outputs so downstream systems could consume the results reliably.

  Stack: *Deterministic policy engine, Qwen3Guard-Gen-4B and Qwen3-8B via vLLM (guided JSON), FastAPI, MySQL, Prometheus, Grafana, Azure Monitor, AKS.*

AI Knowledge Assistant (RAG)*: document-grounded operational Q&A for staff and field teams.*

- Built a retrieval-augmented assistant that answers from approved documentation with citations and escalates to a human team when confidence is low, using hybrid retrieval and reranking over a large internal document corpus.
- Reduced GPU cost with tiered model routing that serves most queries on a lightweight model and keeps larger models for complex cases, and used a bounded, state-machine orchestrator instead of autonomous agents so behavior stays auditable and predictable.

  Stack: *Qdrant (hybrid retrieval), Qwen3-Embedding with a cross-encoder reranker, vLLM tiered serving (Qwen3-4B for simple queries, Qwen3-8B for complex), Redis, FastAPI, OpenTelemetry, Azure AKS (T4/A10 GPU).*

Multilingual Content Generation*: on-brand campsite descriptions and points-of-interest copy at catalog scale.*

- Built a two-stage generate-then-validate content product that produces brand-consistent copy across 5 brands and 10 languages, with automated brand-compliance checks before editorial sign-off, cutting turnaround from days to minutes.
- Moved prompt governance into external YAML with golden-output regression tests so business teams can steer editorial output without code changes, and ran generation on a self-hosted open-source model with an automated validation pass to keep output on-brand and publishable.

  Stack: *FastAPI, Qwen3-8B (generation) and Qwen3-4B (validation) via vLLM (guided JSON), Redis caching, Azure AKS, GitHub Actions, ACR, Azure Key Vault.*

Smart Table Service Intelligence (Computer Vision)*: real-time table-service monitoring from overhead cameras.*

- Built a streaming perception-to-action product (detection, tracking, state classification, then an action engine) that alerts staff in real time when a table needs service, using temporal smoothing to suppress false alerts and improve service responsiveness and table turnover.
- Hardened a detection proof-of-concept into a governed platform with confidence-gated abstain, drift monitoring, and pinned model artifacts for safe, repeatable rollout.

  Stack: *YOLO26, BoT-SORT, EfficientNetV2, Redis Streams, OpenTelemetry, Azure AKS (T4/A10 GPU), ACR.*

Cleareye | Trivandrum, India | 2019 to 2020

Machine Learning Engineer, *Enterprise Contract Intelligence Platform for LIBOR Transition*

- Built a contract-intelligence platform that pulled benchmark references and fallback language out of thousands of mixed-format contracts (PDF, DOCX, scans) at the clause level, replacing manual legal review with auditable automation at portfolio scale.
- Sent only low-confidence clauses to human reviewers and kept clause-to-source traceability for regulatory audit, structuring the extracted output into amendment-ready records.

  Stack: *Python, NLP clause classification (BERT, RoBERTa), TensorFlow Serving, PDFMiner, Apache Tika, Tesseract OCR, Kafka, MongoDB, Flask, Azure DevOps.*

eInfochips | Bangalore, India | 2016 to 2017

Software Engineer, *Real-time Video Surveillance & Analytics for Bharat Electronics Limited (BEL)*

- Built a real-time vehicle and pedestrian detection system with zone-based intrusion detection for defense perimeter surveillance, integrated into BEL's video management system for automated operator alerts.
- Delivered real-time detection at sub-200ms latency with around 74% mAP on a custom defense dataset, and reduced false alarms with hard-negative mining and confidence-calibrated NMS, replacing manual monitoring of restricted zones.

  Stack: *Python, PyTorch, SSD300 (VGG-16 backbone), OpenCV, RTSP streaming, non-max suppression, Flask REST, NVIDIA GPU.*

EDUCATION & CERTIFICATIONS

B.Tech, Computer Science, Rajiv Gandhi Proudyogiki Vishwavidyalaya (RGPV), 2014

Cloudera Certified Developer for Apache Hadoop (CCDH), Cloudera, 2014