Anant Patankar

ML Engineer | Distributed Systems & Production ML | Scalable AI Pipelines, MLOps

E-Mail: patankar.anant123@gmail.com **|| Phone:** +91-9834881001

LinkedIn URL: linkedin.com/in/anant-patankar
Website: https://anantpatankar.com/

Profile Summary

Results-driven ML/AI Engineer with 5.9+ years building production ML and agentic AI systems at scale. Expert in distributed architectures (Celery/Redis/Cloud Run), async Python, and multi-agent orchestration (Google ADK), deploying containerized AI pipelines designed for 100K+ users. Deep expertise in LLM integration, RAG, vector databases, and MLOps across GCP/Azure/AWS delivering measurable business impact.

Technical Skills:

- **Programming & Advanced Python:** Python, Async/Await, AsyncIO, Regex, FastAPI, Git, Bash, Type Hints
- **Distributed Systems & Infrastructure:** Celery, Redis, RabbitMQ, Docker, Kubernetes, Cloud Run, CI/CD, Microservices
- **LLMs & Agentic AI:** Google ADK, Multi-agent orchestration, Gemini, Claude, HuggingFace, OpenAI API, Azure OpenAI, LangChain, LlamaIndex, RAG Systems, Reasoning & Tool Use, MCP Protocol
- **ML/DL Frameworks:** PyTorch, TensorFlow, scikit-learn, NumPy, Pandas
- **Vector Databases & Search:** FAISS, HNSWLIB, Qdrant, Elasticsearch, Vertex AI Search (Discovery Engine), Hybrid Search, Semantic Embeddings
- **Data Engineering & Pipelines:** SQL/SQLglot, Feature Engineering, ETL Pipelines, PySpark, Data Validation, Schema Management
- **MLOps & Cloud:** MLflow, Model Deployment, Inference Optimization, GCP (Vertex AI, Cloud Run, Firestore, GCS, Discovery Engine), Azure (OpenAI, CosmosDB), AWS (S3, SageMaker, Lambda)
- **ML Algorithms:** Classification, Clustering, Random Forest, SVM, Neural Networks (LSTM, CNN, Transformers), Time Series, NLP

Professional Experience

Xevyte Technologies Pvt. Ltd. (Client: Holcim), Bengaluru, Karnataka

AI Engineer (April 2026 – Present)

- **Project 1: FLAME — Recommendation Engine for Gamified HSE Crowd-Labelling:**
  - Architected the AI backend of FLAME, a gamified crowd-labelling platform for real HSE image annotation through four mini-games, designed for 100K+ concurrent users and 50K+ new images/day
  - Built a Google ADK multi-agent recommendation engine (Gemini 2.5 Flash) with fully LLM-driven reasoning over structured Firestore tools - no hardcoded routing - executing weighted topic selection, per-player deduplication, and label persistence
  - Engineered a prefetch queue with pre-signed GCS URLs and a FastAPI/Cloud Run Refill Service, achieving sub-second question-serving latency at target scale
  - Integrated Vertex AI Search as the semantic retrieval layer to avoid Firestore index explosion; implemented zero-redeployment configurability via Firestore config with a TTL-cached in-memory ConfigLoader
  - Led technical direction for a junior engineer; authored architecture sign-off documents and presented vector-DB trade-off analysis (Firestore findNearest vs. Vertex AI Search vs. Qdrant) to senior stakeholders
  - **Technologies:** Python 3.13, Google ADK, Vertex AI (Gemini 2.5 Flash, Vertex AI Search), Firestore, Cloud Run, GCS, FastAPI
- **Project 2: AI-Powered HSE Training Content Generation System:**
  - Designed a 9-agent Google ADK orchestration pipeline with personas spanning document ingestion, question generation, quality review, tone adaptation, and gap analysis
  - Built a RAG pipeline over a curated HSE document library (PDF, web URLs, multi-format) using Vertex AI and Gemini, with grounded generation and mandatory source traceability
  - Shipped an admin-facing review workflow (approve/reject/modify, tone adjustment, AI-generated explanations) plus a Gap Analysis module detecting topic-coverage gaps; deployed on Cloud Run with Firestore and GCS
  - **Technologies:** Python, GCP (Vertex AI, Google ADK, Cloud Run, Firestore, GCS), Gemini, RAG

Netlink Software Private Limited, Bhopal, Madhya Pradesh

Software Engineer (July 2025 - April 2026)

- **Project 1: Lumenore Analytics MCP Server:**
  - Architected an MCP server using async Python (asyncio/aiohttp) supporting 100+ simultaneous connections with sub-50ms latency through non-blocking I/O and connection pooling
  - Designed distributed request handling with JWT authentication middleware and API proxying across 7 analytics endpoints with automated retry logic and circuit-breaker patterns
  - Implemented SSE streaming for real-time data delivery with connection lifecycle management and fault-tolerant error recovery across network disruptions
  - **GitHub:** https://github.com/Lumenore-Platform/lumenore-mcp
  - **Technologies:** Python 3.13, FastMCP, MCP Protocol, aiohttp, asyncio, JWT, SSE
- **Project 2: Lumenore AI Customer Support Chatbot:**

- Designed distributed message processing handling 10,000+ daily messages with sub-200ms P95 latency using Celery worker pools, Redis pub/sub, and PostgreSQL with connection pooling
- Architected hybrid search combining vector similarity (Qdrant) and cross-encoder reranking achieving 85%+ retrieval accuracy with semantic embeddings and custom scoring
- Implemented intelligent escalation with sentiment analysis (HuggingFace transformers), PagerDuty integration, multi-tenant row-level security, and WebSocket-based real-time updates for 1000+ concurrent connections
- **Technologies:** FastAPI, Azure OpenAI, Qdrant, PostgreSQL, LangChain, sentence-transformers, Docker, WebSockets

- **Project 3: Agentic MCP Client & Orchestrator:**
  - Architected microservices orchestration platform managing multiple MCP servers with a RESTful API supporting 4 transport protocols (HTTP, WebSocket, SSE, stdio) and RBAC/ABAC access control
  - Designed parallel execution engine with Celery achieving 60% latency reduction through intelligent workload distribution and automatic retry with exponential backoff
  - Built LLM-powered query decomposition with regex-based intent classification (95%+ accuracy) and SQLAlchemy persistence for concurrent transactional integrity
  - **Technologies:** Python, FastAPI, Celery, Redis, SQLAlchemy, PostgreSQL, aiohttp, WebSockets, LLM integration, RBAC

Affine Analytics Pvt. Ltd., Bangalore, Karnataka
Senior Associate Data Scientist (July 2024-June 2025)
Associate Data Scientist (June 2022-June 2024)

- **Project 1: Code Analysis and Transformation AI System**
  - Built scalable code analysis pipeline processing 500K+ lines across 10+ languages using HuggingFace transformers with custom tokenization, regex-based dependency-graph construction, and security-vulnerability detection (92%+ accuracy)
  - Developed FastAPI-based API with rate limiting, authentication, and Azure CosmosDB caching, reducing redundant analysis by 70% with sub-500ms response times
  - **Technologies:** Azure OpenAI, LangChain, FastAPI, Docker, Azure CosmosDB, HuggingFace

- **Project 2: Enterprise Knowledge Q&A CHATBOT**
  - Architected production RAG system with ChromaDb vector indexing and Elasticsearch hybrid search, implementing semantic embeddings with custom relevance scoring and 20% token-cost reduction through prompt engineering
  - Built monitoring tracking P50/P95/P99 latency with an A/B testing framework comparing retrieval strategies (dense vs. hybrid) with statistical significance testing
  - **Technologies:** Python, PyTorch, Transformers, HuggingFace, LangChain, ChromaDb, Elasticsearch

- **Project 3: Indian Rural Development Bank- 30+ Data Science Use Cases**
  - Designed and deployed 30+ ML/DL solutions (regression, time series, NLP, OCR) processing 10M+ records using PySpark and Vertica SQL with query optimization for sub-second response times
  - Built Elasticsearch search infrastructure with NLP preprocessing and end-to-end ETL pipelines for credit forecasting and CRAR aggregation with incremental processing
  - **Technologies:** Python, PySpark, MLflow, Elasticsearch, Ray.IO, Vertica SQL, PowerBI

RPDS Innovations Pvt. Ltd., Pune, Maharashtra
Junior Data Scientist (Mar 2021-May 2022) | Data Science Intern (Jan 2021-Mar 2021)

- Built time-series forecasting models (RF, SVM, LSTM) for chemical reactor optimization with feature engineering and anomaly detection. Python, scikit-learn, PyTorch.

Miscos Technologies Private Limited, Pune, Maharashtra
Python Developer (Aug 2019-Feb 2020)

- Built ML forecasting pipeline (ARIMA, RF, LSTM) achieving 79% accuracy and a real-time object-detection system processing 15 FPS video feeds with TensorFlow. Python, scikit-learn, TensorFlow, OpenCV.

Personal Projects

- **Project: MLOps MCP Server - Open-source (PyPI: mlops-mcp-server)**
  - **GitHub: github.com/anant-patankar/mlops-mcp-server)**
  - Built and published an open-source MCP server (v0.1.0, MIT) giving AI assistants direct tool-based access to MLOps workflows - experiment tracking, model registry, dataset management, pipeline orchestration, and lineage tracing - wrapping MLflow, DVC, and Git
  - Architected a two-tier tool registry (always-on core tools + 15 on-demand domain modules) to keep the agent context window lean; implemented drift detection (KS-test), DAG cycle detection (Kahn's algorithm), and BFS lineage tracing with Mermaid visualizations
  - **Technologies:** Python 3.11+, FastMCP, MCP Protocol, MLflow, DVC, Git, Pandera/SciPy

Education

- **M.Sc. Forensic Science (2017)**, Dr. Hari Singh Gour Central University, Sagar
- **B.Sc. Forensic Science (2015)**, Government Institute of Forensic Science, Aurangabad, Maharashtra

Awards

- 2x Game Changer of the Month (Sept & Dec 2023) - Enterprise Knowledge Q&A Chatbot and Credit Forecasting System