Sarvesh Swami

Delhi, India | +91-8290323758 | sarveshswami10701@gmail.com | LinkedIn | GitHub

Summary

AI Engineer building production RAG systems, LLM orchestration platforms, and real-time distributed AI applications. Track record of improving retrieval accuracy, cutting inference latency and cost, and scaling multi-tenant systems to thousands of concurrent users.

Technical Skills

Languages: Python, TypeScript, JavaScript, SQL, MATLAB
Generative AI: OpenAI & Azure OpenAI APIs, AWS Bedrock, RAG, Agentic AI, LangGraph, Embeddings, Pinecone, Knowledge Graphs, LLM Caching, Prompt Engineering
ML & Vision: PyTorch, TensorFlow, NLP, CNNs, Recommender Systems, YOLOv8x, BoxMOT
Backend & DevOps: FastAPI, Node.js, REST, gRPC, WebSockets, Microservices, Docker, Kubernetes, CI/CD, MLOps, OAuth2, JWT, RBAC, Twilio Voice, Meta WhatsApp Business, MS Graph
Data & Cloud: PostgreSQL, MySQL, Redis, AWS, Azure (Functions, Fabric), ETL Pipelines, Pandas, NumPy

Experience

Maaze March 2026 – Present
AI Engineer

- Architected multi-channel agentic AI frameworks on AWS Bedrock, including an autonomous Twilio voice agent and a multi-tenant WhatsApp chatbot scaling to 1,000+ concurrent conversations, cutting manual handling time 45%.
- Designed a dynamic routing orchestrator with intent parsing that directs requests across rule-based engines, hybrid ML pipelines, and LLM agents, improving intent-classification accuracy 28%.
- Cut inference costs 30% at scale through token-optimization strategies in the routing engine, using efficient prompt construction and payload batching.
- Advanced RAG architecture by pairing Pinecone vector search with multi-tenant knowledge graphs for entity modeling and structured reasoning, raising retrieval precision 20% and reducing latency 30%.
- Engineered an AI meeting-assistant bot with real-time audio processing that joins Microsoft Teams, Google Meet, and Zoom calls, serving 150+ daily active users at 92% satisfaction.
- Built a real-time computer vision stack for fleet safety: YOLOv8x with BoxMOT cross-frame tracking for unique passenger counting at 30 FPS / 96% accuracy, plus a multimodal collision-prediction model using g-force and delta-velocity anomalies at 92% accuracy.
- Deployed a store-analytics vision system for MOD Pizza generating foot-traffic heatmaps and peak-hour insights, lifting targeted campaign engagement 18%.
- Developed a MATLAB signal-processing algorithm parsing continuous 3-axis gyroscope streams at a 10 ms sampling rate, triggering accident alerts via peak-threshold analysis at 98% detection accuracy.

Crenovent Sep 2025 – Mar 2026
AI Engineer

- Delivered an enterprise agent catalog that converted complex operational workflows into policy-safe, one-click automations executed from a single conversational interface.
- Built trust and compliance frameworks with human-in-the-loop overrides, audit logs, feature-contribution weighting, and automated "evidence packs" for end-to-end traceability.
- Enforced strict tenant isolation and RBAC across hybrid retrieval pipelines, keeping vector and knowledge-graph context partitioned per customer.

Decision Point Apr 2024 – Jan 2025
AI Engineer

- Optimized high-throughput LLM pipelines, reducing end-to-end query latency 20–30% via prompt optimization, asynchronous API layers, and improved request handling.
- Engineered a GPTCache layer that bypassed ∼35% of redundant model calls, materially lowering compute usage and API spend.
- Scaled embedding and vector retrieval pipelines over large unstructured text corpora, increasing semantic search precision and RAG relevance.
- Architected modular serverless ML microservices on Azure Functions for event-driven retraining, surfaced through interactive Adaptive Cards in enterprise chat.

Independent Apr 2022 – Mar 2024
Game Developer – Deadtoy Showdown

- Shipped a production 3D top-down survival shooter in Unity (C#), owning architecture through final build: vector-based movement, raycast shooting, custom physics, and pathfinding enemy AI.

Projects

Enterprise WhatsApp AI Support Assistant | *LLMs, RAG, Meta WhatsApp Cloud API*

- Intelligent support system answering customer queries over a retrieval-augmented knowledge base, delivered natively through WhatsApp.

Person Re-Identification System | *Python, PyTorch, Deep Learning*

- Deep-learning pipeline that tracks individuals across video frames, assigning consistent IDs through appearance feature matching and re-identification.

Vehicle Analytics Chatbot | *NLP, GPS & Telematics, Python*

- Conversational assistant for natural-language querying of telematics data, surfacing live tracking, idle detection, route analysis, and driver-behavior insights in chat.

Lead Intelligence & Business Discovery Platform | *Google Places API, Python*

- AI-driven lead generation platform automating business data collection, enrichment pipelines, and exportable intelligence dashboards.

AI Virtual Try-On Platform | *Computer Vision, Image Pipelines*

- Garment-overlay system using computer vision models and automated image processing to render personalized try-on visualizations.

Education

Maharaja Ganga Singh University 2019 – 2022
Bachelor of Computer Applications *Sardarshahar, Rajasthan*

Certifications

Advanced Learning Algorithms, *Stanford Online* (2025) | Generative AI Language Modeling with Transformers, *IBM* (2025) | Generative AI for Data Science, *IBM* (2025) | Google Cloud Infrastructure, *Google* (2025)