Om Shah

+91 998 792 7199 | omshah.tech@gmail.com | linkedin.com/in/om-shah | Portfolio

Education

MPSTME, NMIMS University — Mumbai, India
Bachelor of Technology in Computer Engineering; ***CGPA: 3.96/4.0*** — *2022 – 2026*

R.N. Podar School — Mumbai, India
12th CBSE: 97% — *2022*

Profile

Computer Engineering student with strong interest in machine learning and AI systems. Experience working with deep learning frameworks, building RAG pipelines, agentic AI systems, and implementing reinforcement learning algorithms for practical applications.

Experience

AI Intern — Dec 2025 – Present
Qpiai — *Bengaluru, India*

- *Multi-Agent Orchestration & Platform Engineering*
- Engineered a complete harness for a generative web application builder, implementing multi-agent orchestration, modular agent composition, and coordinated execution pipelines; architected the full-stack system from scratch with a scaffold template infrastructure (frontend templates, backend/auth scripts, RBAC), pre-cached dependency tarballs to minimize generation latency, containerized PostgreSQL with partitioning, and hot module replacement for real-time user modifications.
- Engineered a scalable multi-agent orchestration framework for automated evaluation, validation, and system optimization across generated applications, incorporating structured I/O abstractions, stateful execution management, and evolving memory architecture (hierarchical memory tiers, cross-session retrieval, persistent state); supported robust agent collaboration, workflow continuity, and efficient infrastructure utilization across the platform stack.
- *Computer Vision Model Evaluation, Training & Recommendation*
- Executed end-to-end model training and dataset curation pipelines across diverse object detection benchmarks; constructed a model registry backed by a vector database schema integrated with MLflow experiment tracking, enabling semantic model retrieval based on dataset characteristics, performance metrics, and architectural features.
- Implemented a multimodal video ingestion and retrieval prototype from a research paper, integrating Vision-Language Models and embedding models for video understanding; explored integration of video analysis tool.

Projects

Forge-OSH: Terminal-First AI Agent Platform | *Rust, Tokio, Ratatui, Petgraph, Rayon*

- Architected a modular, terminal-native AI agent platform in Rust across four layers: a Ratatui/Crossterm-powered TUI for rendering and session management; an async Tokio-based agent core for orchestration, permissions, and event handling; a trait-based multi-provider LLM abstraction supporting Anthropic, Gemini, OpenAI-compatible, and Ollama backends via a unified provider router; and an extensible tool registry with permission enforcement for file, git, shell, web, and code operations.
- Constructed a semantic code graph engine using Petgraph and Rayon for parallel, codebase-aware context packing and dependency analysis; implemented session persistence with full undo support via file history snapshots; and integrated secure OS keyring-based credential management for multi-provider API key storage.

Albot: Advanced Multimodal RAG System | *FastAPI, ArangoDB, Next.js, PyTorch, Whisper*

- Built a multimodal RAG system ingesting text, images (OCR + VLM), audio (Whisper transcription), and video (frame extraction) into an ArangoDB-backed knowledge graph; implemented a hybrid retrieval engine combining vector similarity, graph traversal, and BM25 lexical search with Personalized PageRank and Bayesian weight optimization for adaptive, query-complexity-driven retrieval.
- Designed a layered cognitive memory architecture with working (ephemeral), session (contextual), and semantic (long-term) memory tiers with cross-session knowledge retrieval and namespace scoping; developed an RLM deep research framework that autonomously plans, searches, and synthesizes information using a map-reduce paradigm for extended context processing.

DS-Forge: Data Science Operating System | *Next.js, FastAPI, scikit-learn, Docker, Pandas*

- Engineered a no-code/low-code ML lifecycle platform with a spreadsheet-like manual data grid and 25+ atomic cleaning operations including Winsorization, Z-score outlier removal, and smart column-level recommendations; built a feature engineering suite with 28+ transformations covering PCA, t-SNE, Isomap, polynomial features, and robust scaling with full pipeline preservation for consistent inference-time preprocessing.
- Deployed trained models (Random Forest, Gradient Boosting, SVR, SVM, AdaBoost) as auto-generated REST APIs via an inference engine that automatically reapplies training-time encodings for raw input handling; containerized the full stack via Docker Compose with production images published to Docker Hub.

Agflow: Visual AI Agent Orchestration Platform | *Python, React Flow, LangChain, Supabase*

- Developed an open-source platform for building agentic AI workflows via a node-based visual interface powered by React Flow, integrating a RAG pipeline with Supabase pgvector for automated document chunking and embedding. Features multi-model LLM support (Groq, OpenAI), pre-configured tool nodes for web search and external APIs, custom Python execution via Monaco Editor, and cloud-based flow management with automatic saving.

Knowledge Distillation in LLMs via Jensen-Shannon Divergence | *PyTorch*

- Implemented a Jensen-Shannon Divergence-based knowledge distillation framework for LLM compression, addressing KL Divergence's asymmetry and instability. Distilled SmolLM2-135M to 90M parameters achieving F1 of 0.9125 vs. KL's 0.9028, with 12.5% variance reduction, 3.45% ROUGE improvement, 20% faster convergence, and 23.3% lower final loss, validating JSD's advantages in optimization stability and dark knowledge preservation.

Reinforcement Learning for Portfolio Management | *PyTorch, TensorFlow*

- Implemented and benchmarked DDPG, PPO, and Dynamic Embedding RL algorithms for dynamic asset allocation on Indian equities (NSE, 2012–2023), incorporating proportional transaction costs and market friction. DDPG achieved 723.5% total return (Sharpe 1.78) and DERL with Wasserstein autoencoder state representation achieved 357.8%, both substantially outperforming the 185.4% passive benchmark.

Neural Machine Translation with Transformer Architecture | *PyTorch, spaCy*

- Built a Spanish-to-English NMT system using Transformer architecture with multi-head self-attention and sinusoidal positional encodings; implemented a preprocessing pipeline combining spaCy lemmatization with SentencePiece BPE tokenization (16K vocabulary) for robust OOV handling. Trained on 500K Europarl v7 sentence pairs, achieving validation perplexity of 4.613 and BLEU score of 12.41 with beam search decoding.

Technical Skills

Languages: Python, Rust, JavaScript, TypeScript, HTML, CSS, SQL
AI/ML: PyTorch, TensorFlow, LangChain, Transformers, FAISS, ArangoDB, ChromaDB, OpenAI API, Anthropic API, RAG Pipelines, Agentic Systems, Neural Networks, Reinforcement Learning, Knowledge Distillation
Web: ReactJS, Next.js, Node.js, Express, FastAPI, Flask, Django, MongoDB, Supabase, PostgreSQL
Tools & Infra: Git, Docker, Tokio, Ratatui, Petgraph, Whisper, OpenCV, spaCy, Vercel, Render

Certifications

Quantum Machine Learning | ***IBM*** 2026
Deep Learning Specialization | ***DeepLearning.AI*** 2024
AWS Academy Cloud Foundations | ***Amazon Web Services*** 2024
Google Data Analytics Professional Certificate | ***Google*** 2024
Honours Program | ***Coursera*** 2022 – 2026

Leadership & Volunteering

Technical Executive | ***MPSTME ACM Student Chapter*** 2023 – 2024
Class Representative | ***BTech Computer Engineering*** 2022 – 2025
Computer Tutor | ***Underprivileged Children Education*** 2023