Asmit Dash

asmitdash44@gmail.com | +91-9619288756 | Mumbai, India | linkedin.com/in/asmitdash | github.com/asmitdash

EDUCATION

Bachelor of Technology, Computer Engineering *Jun 2026*

University of Mumbai *Mumbai, India*

- CGPA: 9.2 / 10. Honours in Artificial Intelligence and Machine Learning.

SKILLS

AI Product and Solutions: Enterprise AI Use-Case Discovery, Requirement Framing, Stakeholder Alignment, ROI / Cost Modeling, Solution Architecture, Vendor and Model Selection, GTM and Client Demos, Roadmapping.

AI Systems and Architecture: End-to-End LLM Pipelines, RAG, GraphRAG, Agentic Workflows, Multi-Model Orchestration, Evaluation Design, Governance and Compliance, Latency / Cost Trade-offs.

LLMs and Platforms: GPT-4/5, Claude, Gemini, Llama; LangChain, LangGraph; Vector DBs (pgvector, FAISS, Pinecone); AWS Bedrock, SageMaker, Lambda, S3, Cognito, CloudWatch.

Delivery and Tools: Cross-functional Team Lead, Executive Communication, Python, SQL, TypeScript, Git, Linux, Docker.

EXPERIENCE

AI Systems Architect, Intern (AI Automation) *Jun 2025 - Jul 2026*

Deloitte *Mumbai, India*

- Owned end-to-end delivery of the Volkswagen document-verification automation for Deloitte India: framed the business case, defined success metrics with the client, and architected an LLM-driven pipeline over 50,000+ documents/month; cleared a 4-5 month manual backlog held by 20-25 reviewers within one calendar month (~4x throughput lift, >60% turnaround improvement).
- Made model-tier, vendor, and cost / latency trade-offs across the stack (OCR, LLM extraction, government API verification, audit) and translated them into an AWS deployment (S3, Lambda, SageMaker) the operations team could run without engineering support.
- Led a 5-member cross-functional pod through requirement gathering, prompt-strategy standards, weekly stakeholder demos, and client sign-off; primary technical point-of-contact to a Partner-tier sponsor.

Machine Learning Engineer, Intern *Sep 2024 - Jan 2025*

Ocufox Technologies Pvt Ltd *Mumbai, India*

- Shipped a production captcha recognition service end-to-end (CRNN + ViT, ONNX serving, AWS deployment, CI/CD, CloudWatch) at sub-100ms latency; owned data strategy, serving, and observability decisions with the founding team.

PROJECTS

Lex-RAG: Production-Grade Retrieval-Augmented Generation Platform *Next.js, Supabase, AWS Bedrock, Gemini*

- Designed the full RAG stack for a legal / regulated-corpus use case: contextual retrieval, hybrid BM25 + pgvector via Reciprocal Rank Fusion, BGE cross-encoder reranking, and an agentic multi-hop planner with tool-use over corpus and live web.
- Layered GraphRAG (entity / edge extraction, Louvain communities, LLM summaries) and a Self-RAG / CRAG critique step; explicit trade-offs on chunking, top-k, and model tier to hit the latency / fidelity profile a paying customer would accept.

HCR-CAF: Multimodal Emotion Recognition with Evidential MoE *PyTorch, RAVDESS, MFCC + MTCNN, Streamlit*

- Designed Synergy-Bottleneck Mixture-of-Experts with Evidential Routing over an HCR-CAF cross-attention backbone under a four-term composite loss (CE + IB + synergy hinge + InfoNCE).
- Outperforms prior MoE-emotion baselines (MiSTER-E, GEM-FI, SynIB, CAG-MoE) by uniquely combining evidential routing, synergy maximization, and information bottleneck; real-time-safe single forward pass.

OpenSource: Memory and Retrieval Infrastructure for Agents *Python, sqlite-vec, MCP, Chroma*

- **replay-rag**: reasoning-trace memory for o1 / R1-style agents; cuts output tokens 40-70% on repeated patterns. **palimpsest**: contradiction-aware agent memory with provenance-preserving forgetting via a lineage DAG. **chronograph**: temporally-aware RAG with as-of queries, current-truth retrieval, and stale-fact detection.

LEADERSHIP

Chairperson *Mar 2023 - Oct 2025*

Code AI Club *Mumbai, India*

- Founded and led Code AI Club from zero; set technical direction, ran ML / AI workshops and internship drives, and partnered with IIT Bombay and Deccan College Pune on joint NLP workshops and projects.