Mahendra Dewangan

+919179473930 | mahendradewangan195@gmail.com | linkedin.com/in/mahendradew | github.com/mahendradew | Bengaluru, India

Summary

System Engineer at TCS, working as an AI Engineer within Applied AI, with hands-on experience building AI agents, RAG pipelines, and full-stack applications. Developed Agentic RAG systems for Cisco, an AI coding agent for NextGen, and retrieval systems that improved information discovery across large unstructured datasets. Experienced with LLMs, vector databases, and AI system design.

Education

Bachelor of Technology (Computer Science) — Jun 2020 – Aug 2024
Vishwavidyalaya Engineering College CGPA: 8.0/10 — *Ambikapur, India*

Experience

System Engineer — Feb 2025 – Present
Tata Consultancy Services (TCS) — *Bengaluru, India*
Client: NextGen – Agentic AI Coding Agent

- Architected an **LLM-powered agentic Coding Harness** for C/C++ codebases by implementing **tool calling** and **tool orchestration** across a full 5-component **MCP stack (server, client, tools, AgentController, Ollama LLM client)** with policy-based patch validation, enabling fully automated build-fix-verify cycles with zero manual intervention.
- Reduced **LLM token consumption** by **35%** (approx.) (10 – 12K to 7 – 8K tokens per call) by auditing and restructuring prompt payloads to eliminate redundant repository context, reducing inference cost and improving agentic loop speed without loss of fix accuracy.
- Enabled feedback-driven iterative error resolution on **LLVM/Clang** (20M+ LOC) by implementing project-aware Ninja builds, incremental compilation, and structured compiler-diagnostics parsing, **reducing rebuild scope** from full-codebase to targeted components.

Client: Cisco – Applied AI / RAG Systems

- Reduced lookup time for network infrastructure data by **30%** (approx.) by replacing manual log/documentation search with an **Agentic RAG** retrieval system, improved with semantic chunking, hybrid search, and reranking.
- Built a RAG-based video search system for a **6-person team**, allowing timestamp navigation across 20–40 min WebEx recordings, eliminating manual scrubbing.
- Implemented semantic and hybrid search with metadata filtering, delivering accurate timestamp retrieval across the recordings and reducing query time from **minutes to seconds**.

Full Stack Developer Apprenticeship — May 2024 – Aug 2024
Buildspace — *Remote*

- Engineered **askpdf**, enabling real-time question answering over PDFs, by designing a **RAG pipeline** (LLM embeddings, Pinecone vector DB, BM25 hybrid search) for accurate, context-grounded answers over documents up to **10 pages.**
- Integrated **streaming responses** via **Vercel AI SDK**, achieving low-latency interaction during Q&A enabling sub-second first-token responses via streaming.

Frontend Developer Internship — Sept 2023 – Nov 2023
HNGx — *Remote*

- Developed a screen recording Chrome extension using JavaScript and browser APIs, supporting 3 recording modes: screen, tab, and window capture.
- Streamlined video data transfer with **Blob objects**, reducing memory usage and boosting backend efficiency by **40%**, while implementing synchronized video and transcript playback to reduce user navigation time by **25%**.

Projects

AuraMaxxing | *Typescript, Next.js, Node.js, MongoDB, PostgreSQL, PineconeDB* — Dec 2024 – Feb 2025

- Shipped a **GenAI platform** at **NosuAI Hackathon** to convert YouTube videos into notes, flashcards, quizzes, and stories, reducing manual content creation effort by **90%** (approx.).
- Transformed transcripts into 4 formats of immersive narrative content with AI-generated visuals and speech synthesis, expanding a single video into multiple learning experiences

SecondBrain | *Typescript, React.js, Node.js, Express, MongoDB, JWT Authentication* — Sept 2024 – Nov 2024

- Designed a **centralized content management** system unifying tweets, YouTube videos, and articles into a single tagged, searchable hub, eliminating the need to switch between platforms.
- Consolidated cross-platform bookmarks into a **unified storage system** enabling structured access, reducing bookmark retrieval to a single-query search.

Certifications

Claude Certified Architect - Foundations	Credly - Aug 2026
Claude Certified Developer - Foundations	Credly - Aug 2026
GitHub Certified - Agentic AI Developer	Microsoft - Jul 2026
Microsoft Certified - Azure AI Engineer Associate	Microsoft - Apr 2026
AWS Certified Developer - Associate	Credly - Apr 2026

Skills

Languages: Python, JavaScript, TypeScript, SQL, Java
Frameworks & Libraries: React.js, Next.js, Node.js, Express.js, TailwindCSS, FastAPI
Database & DevOps: MySQL, PostgreSQL, MongoDB, Pinecone (Vector DB), Git, GitHub, Docker, Linux
GenAI & ML: Large Language Models, RAG, LangChain, LangGraph, LangSmith, Embeddings, Ollama, MCP, Prompt Engineering, Local models (qwen 2.5 27b, qwen 3.6 27b, qwen 3.8 27b)