

ANIRUDH SARVE

Mumbai, Maharashtra, India | +91 93225 81894 | anisarve2006@gmail.com | anirudhsarve.netlify.app
github.com/anisarve2006 | linkedin.com/in/anirudhsarve

PROFESSIONAL SUMMARY

Generative AI and Agentic Systems Engineer specializing in autonomous multi-agent orchestration, deterministic RAG pipelines, and Model Context Protocol (MCP) integrations. Experienced in eliminating LLM hallucinations via vector-chunk hash verification, building air-gapped multimodal pipelines, and parsing high-stakes tabular financial documents through coordinate-bounded Document AI. Skilled in bridging cutting-edge foundation models with robust production software engineering.

EDUCATION

Vidyalankar Institute of Technology (VIT), Mumbai — B.E. in Computer Engineering

Expected May 2028

Cumulative GPA: 8.90 / 10.0 | Affiliated with University of Mumbai

Relevant Coursework: Natural Language Processing, Machine Learning, Artificial Intelligence, Database Management Systems, Data Structures & Algorithms, Operating Systems, Information Retrieval.

Sonopant Dandekar Shikshan Mandali — HSC (Class XII, PCMB)

May 2024

Score: 92.60% | Distinction in Mathematics, Physics & Computer Science

TECHNICAL SKILLS

Generative AI & LLMs: OpenAI APIs, Google Gemini, Anthropic Claude, LangChain, LlamaIndex, Hugging Face Transformers, Faster-Whisper, Prompt-Conditioned Decoding, RAG Architecture, Context Optimization.

Agentic Systems & Protocols: Multi-Agent DAG Orchestration, Model Context Protocol (MCP), Task Decomposition, Self-Critique Loops, Tool Calling / Function Calling, Zod Validation, Self-Healing JSON Parsers.

Vector Search & Retrieval: Qdrant (HNSW Indexing & Payload Filtering), pgvector, Dense Embeddings (Text & Multimodal), SQLite Vector BLOBs, Perceptual Hashing (pHash), Cosine Similarity.

Document AI & Full-Stack: PyMuPDF, pdfplumber, Coordinate-Bounded Tabular Parsing, Python, TypeScript, Next.js 16, React, FastAPI, Node.js, Express, Docker, PostgreSQL, MongoDB, Git.

PROFESSIONAL EXPERIENCE

NxtChaptr — Software Developer (AI Integration & Edge Systems)

Nov 2024 – Aug 2025

- Integrated **Google Gemini Vision** multimodal APIs into merchant listing workflows, automating brand detection, garment condition categorization, and SEO copy generation to slash listing creation time from 8 minutes to under 30 seconds.
- Deployed client-side WebAssembly background matting (U²-Net) processing images in under **3.5 seconds** directly in-browser, eliminating cloud GPU processing fees while feeding standardized clean cutouts into multimodal classification pipelines.
- Architected resilient backend state management using Upstash Redis atomic locks (SET NX EX) and Ably WebSockets to ensure high-concurrency event drops maintain absolute transactional integrity.

Exposure: Multimodal LLMs, Google Gemini Vision, Edge WASM, TypeScript, Node.js, Next.js, Redis, MongoDB.

Wroomie — Software Developer (Applied AI & Backend)

Nov 2023 – Jan 2024

- Built multi-attribute compatibility matching systems evaluating prospective flatmates across 4 lifestyle axes (circadian rhythms, hygiene standards, dietary constraints, social preferences), achieving a **94% match satisfaction rate**.
- Integrated Google Gemini to generate automated personalized room listings and summarize flatmate preferences.

Exposure: Generative Copywriting, Clustering Algorithms, Node.js, Express, MongoDB.

FEATURED TECHNICAL PROJECTS

Autonomous Multi-Agent Directed Acyclic Graph (DAG) for Federated Literature Synthesis

Technologies: Python, FastAPI, Next.js, Qdrant HNSW, SSE

- Engineered an autonomous multi-agent research platform decomposing academic literature discovery into an isolated agent swarm: Intent Analyzer, Literature Scraper, Vector Ingester, Claim Verifier, and Section Writer.
- Overcame context-window saturation and "lost-in-the-middle" degradation by federating literature retrieval across ArXiv, PubMed, and OpenAlex APIs, vectorizing thousands of full-text papers into Qdrant with dense embeddings.
- Implemented a deterministic vector-chunk hash verification pass that anchors every generated claim directly to source chunk hashes in Qdrant, **eliminating hallucinated citations by mathematical construction**.
- Built an interactive dual-pane Next.js workspace streaming live agent telemetry via Server-Sent Events (SSE) with interactive source inspection tags.

Cross-Lingual Non-Compositional Figurative NLP Engine & Unicode-Safe Indic Tokenizer (108k Corpus)

Technologies: Python, PyTorch, NLLB-200, Unicode Devanagari NLP

- Engineered an end-to-end NLP framework for translating non-compositional idiomatic expressions and proverbs (मुहावरे व म्हणी) across Hindi, Marathi, and English, eliminating literal word-by-word translation errors.
- Curated a master deduplicated computational linguistics repository of 108,890 unique idioms across 20 languages (removing 40,110 redundant entries with a 26.9% deduplication rate) with balanced multi-task training sets.
- Built a Unicode-aware Devanagari span detector isolating figurative idioms from contextual morphology, feeding a cultural parallel injection engine on top of fine-tuned NLLB-200 checkpoints.
- Benchmarked across multilingual test pairs, achieving a **93.75%–100% Idiom Preservation Rate (IPR)** and reducing the Literal Error Rate (LER) to under 6.25% across Indic-to-English translation directions.

Coordinate-Bounded Document AI & Deterministic Tabular Ledger Extraction Engine

Technologies: PyMuPDF, pdfplumber, LLM Risk Scoring, FastAPI

- Built an intelligent document-AI underwriting engine ingesting multi-page bank statement PDFs (HDFC, SBI, ICICI, Axis) to parse transaction ledgers and compute debt-to-income metrics.
- Replaced error-prone generic OCR engines with native character-vector extraction via PyMuPDF/pdfplumber coupled with explicit (x_0, y_0, x_1, y_1) bounding-box coordinate slicing, achieving **99.8% tabular accuracy** on digital statements.
- Deployed an LLM reasoning agent providing explainable risk assessments and credit metrics tied directly to verified line items, slashing statement review time from 45 minutes to **under 90 seconds (a 10x throughput boost)**.

Air-Gapped Multimodal RAG Pipeline with Prompt-Conditioned Hinglish Beam Decoding

Technologies: Faster-Whisper, Prompt-Conditioned Decoding, SQLite Vector, pHash

- Architected a 100% offline multimodal video intelligence pipeline executing on a single 16 GB RAM laptop with zero external API calls or third-party cloud dependencies for high-security enterprise environments.
- Implemented domain prompt-conditioned beam search decoding in Faster-Whisper, guiding decoder token probabilities to transcribe Hindi-English (Hinglish) code-switched conversational speech without model weight fine-tuning.
- Developed an in-memory SQLite float32 BLOB dot-product vector scanner completing semantic retrieval across 50,000 video segments in **under 12 ms**; combined with perceptual hashing (pHash) to filter out 85% of redundant frames in <1 ms on CPU.

Schema-Validated Model Context Protocol (MCP) Server with Self-Healing JSON Middleware

Technologies: TypeScript, MCP Server, Zod, jsonrepair

- Built an open-source MCP server enabling external AI agents to programmatically create, query, and modify resumes as first-class tool calls over standard I/O.
- Developed strict Zod schema validation pipelines paired with a self-healing middleware (jsonrepair) that automatically intercepts, sanitizes, and repairs malformed agent outputs before database commitment.
- Integrated @react-pdf/renderer with Rust-native canvas bindings to compile vector PDF documents from schema representations in **under 150 ms**.

HONORS & LEADERSHIP

- AI Lead, ISRO Smart India Hackathon (SIH 2026):** Architected temporal-attention neural networks and physics-constrained estimation filters for ISRO problem statement PS 26168.
- Autonomous Systems Member, Indian Rover Challenge (IRC):** Designed state estimation and computer vision pipelines for rover navigation.
- Academic Distinction:** 8.90/10.0 CGPA at VIT Mumbai; 92.60% in HSC Board Examination.