

Jai Kumar Meena

AI/ML Engineer — Agentic Systems, Tool-Use & Multi-Agent Orchestration (MCP, A2A), LLMs/SLMs, Retrieval, Inference & Edge

New Delhi, India | +91 8800682416 | jk422331@gmail.com | github.com/mannuking | linkedin.com/in/jai-kumar-meena | jaimeena.com

Summary

I build agentic AI systems in production: tool-using agents over retrieval and structured outputs, with memory that persists across sessions and the serving latency underneath it. My own project, CVC, is a cognitive-version-control substrate for AI agents — published on PyPI as **tm-ai** and now at v3.8.19. Before that I shipped an enterprise agentic-RAG system that cut analyst review time ~40% and P95 latency 1.6s→700ms, and a production edge-inference app with 10,000+ scans processed. I work end-to-end: model, runtime, and the product surface.

Experience

Co-founder, AI/ML Engineering

lvl360 (train360) · 2026 – Present

- Built **train360**, a multi-tenant enterprise learning platform that turns raw documents into interactive courses: Next.js 15 + React 18 + TypeScript on Firebase, with a Google Genkit / Gemini AI pipeline (PDF → OCR → structured content) behind it.
- Designed the AI content pipeline end to end — OCR extraction, automatic course generation, SCORM playback for legacy content, and a manager compliance dashboard.
- Shipped AI avatars, voice cloning, MCQ/evaluation generation and multi-team chat workspaces; real-time Firestore listeners with company-level data isolation across 50+ organisations.

Data Architecture Design Specialist (Contract)

US-based Enterprise AI Startup · Remote · Apr 2025 – Present

- Designed an agentic RAG system with tool-use and function-calling chains for compliance gap analysis, orchestrated via LangChain with MCP-style tool interface definitions and structured outputs — **~40% less analyst review time**.
- Implemented vector search (FAISS / Chroma) with hybrid retrieval and reranking (SentenceTransformers); added an eval harness covering faithfulness, groundedness and latency SLAs.
- Built FastAPI microservices with streaming tokens, retry/backoff, circuit breakers and Redis caching — **P95 latency 1.6s → 700ms**.
- Integrated Gemini 2.0 Flash for rapid tool-use and added a small local SLM (Phi-3 / Llama-3.2 3B) for on-device redaction and cost control, with automatic model routing by task and latency budget.

AI Developer (Part-Time)

IrisIdea TechSolutions · Bangalore (Hybrid) · Apr 2024 – Present

- Lead on **LeafAid**, a plant-disease diagnosis app: an ensemble CNN (EfficientNet + ViT + ConvNeXt, late fusion) at 98–99% top-1 across 39 disease classes, quantised and distilled to a <10 MB TensorFlow Lite model for sub-2-second offline inference on \$80 phones.
- Built the “Hibiscus” care chatbot on Gemini tool-use for diagnosis, care schedules and treatment lookups, with safety filters and disclaimers — **20,000+ sessions at a 99.6% crash-free rate**.
- Set up Crashlytics telemetry, custom events, A/B prompts and staged Play Store rollouts; 10,000+ scans processed, 90% model-size reduction.

AI & Full-Stack Developer (Contract)

Luxury Rydes · Remote · May 2025 – July 2025

- Delivered tri-platform apps (iOS / Android / Web) in 12 weeks on Flutter + Firebase, provisioning Cloud Functions for realtime pricing, driver ETA and surge logic.
- Built a support agent with retrieval over policy/FAQ and tools for booking lookup, refund quotes and route estimation, with live handoff to a human via the ticket API — **~35% lower median resolution time**.
- Added Maps APIs, geo-indexed queries and anomaly detection for fraudulent trips using heuristic and embedding features.

AI Model Trainer (Freelance)

Outlier · San Francisco Bay Area (Remote) · Sep 2024 – Nov 2024

- Authored high-difficulty code and mathematics prompts and applied RLHF-style preference ranking, specialising in prompt / program-of-thought for tool-former-style usage.
- Evaluated chain-of-thought suppression against structured rationales — **+3–5pp exact-match on held-out tasks**.

Selected Projects

CVC — Cognitive Version Control for AI Agents

tm-ai on PyPI · github.com/mannuking/cvc · 944 commits · v3.8.19

- Built a substrate that versions an AI agent's **cognitive state** rather than its source: every conversation turn, decision and tool call persisted as an immutable, content-addressed Merkle DAG in SQLite plus a content-addressable blob store.
- Ships as a Click + Rich CLI, a FastAPI gateway, and an **MCP server** (stdio/SSE) so agents and harnesses can commit, branch and restore their own reasoning.
- Layered semantic recall over ChromaDB, LangGraph orchestration, Pydantic v2 models, zstandard compression, and multi-provider support (Anthropic, OpenAI, Gemini, Vertex, Ollama). 1,707 Python files, 147 test files.

AstroSwarm — 51-agent simulation on CVC

Multi-agent orchestration experiment · 2026

- Ran a 51-agent space-crisis simulation (1 Mission Controller, 5 Captains, 45 Specialists) where **every agent's reasoning is persisted as a real CognitiveCommit** and all 51 share one hive-mind memory.
- **51/51 agents activated, 4/4 experiments passed, 5/5 integrity checks, 128 hive-mind commits, 172.5s runtime** on a 1.2 MB Merkle DAG.

Fine Tune — real-time voice-transform engine

Audio plugin & offline engine · 2026

- Built a voice-identity transformation engine (not pitch correction): a torchcrepe F0 detector at 100 fps driving **RAVE timbre re-synthesis** through IRCAM pre-trained models, with provenance sidecars.
- Measured on Mac Mini M4 (arm64): CREPE RTF **0.48**, RAVE **sol_ordinario_fast RTF 0.08** — 12× faster than realtime; Core ML export under 5% CPU per stereo channel.
- Ships as a JUCE 8 AUv3 and VST3 plugin for Logic Pro, Ableton, Reaper and FL Studio.

Smart AI Agent IDE

Python · Streamlit · Google Gemini AI · NetworkX ·

- Built a hierarchical graph planner with a memory and tool registry (HTTP, SQL, file, search) supporting MCP-compatible JSON tool schemas and function-call I/O.
- Added task-decomposition heuristics, reflection and retries — **+25–40% task success over vanilla single-step prompting**.
- Implemented run logs, traces, and per-subtask cost/latency budgets with pluggable retrievers and safety rules.

Education

M.Tech by Research — Computer Science & Engineering · Delhi Technological University, New Delhi · 2026 – Present

Research on agent memory, multi-agent coordination and speech-emotion modelling. Supervisor: Dr. Rahul Katarya (CSE, DTU).

B.Tech — Computer Science & Engineering · Delhi Technological University, New Delhi · 2025 · CGPA 7.0

Senior Secondary (CBSE XII) · Kendriya Vidyalaya, New Delhi · 2021 · 76.6%

Technical Skills

Agentic AI	LangChain, LangGraph, LlamaIndex, AutoGen, CrewAI, MCP (Model Context Protocol), A2A patterns, tool-use / function calling, multi-agent orchestration, planning, reflection, evals
Models & Inference	Gemini 1.5/2.0, Claude 3.5 Sonnet, GPT-4o, Llama-3, Phi-3, SentenceTransformers, RAVE / torchcrepe, TensorFlow Lite, Core ML, quantization & distillation
Retrieval & Memory	RAG, hybrid retrieval + reranking, vector DBs (Chroma, FAISS, Pinecone), content-addressed storage, Merkle DAG, semantic recall, prompt & audit versioning
Backend & Infra	Python (FastAPI, Flask, Pydantic v2, Click), Node.js, Express, SQLite, PostgreSQL, Redis, MongoDB, Firestore, Docker, Cloud Run, Firebase Functions, GCP, Auth (JWT, NextAuth, Firebase Auth), circuit breakers, retry/backoff, caching
Frontend & Product	React, Next.js 15, TypeScript, Tailwind CSS, Shadcn UI, Framer Motion, Three.js / React Three Fiber, Streamlit, SCORM, real-time Firestore listeners
Mobile & Edge	Flutter (Dart), Android (Kotlin/Java), on-device CNN inference, Flutter Cloud Functions, Crashlytics, staged Play Store rollouts

Achievements

- **Amazon ML Hackathon** — top 50 of 78,000 participants.
- **Smart India Hackathon 2024** — finalist, all-India, from 86,000 participants.
- **Smart India Hackathon 2022** — competed among 100 teams.
- **Vihaan007 Hackathon** — top 25% of 390 teams, advanced to the final round.
- **Science Olympiad** — 1 Gold and 2 Bronze medals.