

KESHAV KANDOI

Bengaluru, India | kandoikeshav2004@gmail.com | +91-7903471852 | [LinkedIn](#) | [GitHub](#)

SUMMARY

Generative AI and Full Stack Engineer who built MEMORA, an agent-memory platform using hybrid ANN + BM25 + graph retrieval (92% Recall@5, sub-250 ms p95 latency on 10K+ memories). Also shipped a multi-tenant, OAuth-integrated agentic workflow platform and an offline multimodal RAG assistant using Python, FastAPI, LangGraph, and LLMs.

TECHNICAL SKILLS

Languages: C++, Python, JavaScript, TypeScript

Front-End: HTML, CSS, JavaScript, React, Next.js, React Native, Tailwind, Bootstrap

Back-End: Node.js, Express.js, FastAPI, REST APIs

Databases: MySQL, MongoDB, PostgreSQL, SQLite

AI/ML: Ollama, LangChain, LangGraph, OpenAI API, PyTorch, RAG, Agentic AI, Tool Calling, Fine-tuning, Workflow Orchestration

Core: SQL, LLMs, Vector Databases, Prompt Engineering, High-Level Design (HLD) / System Design

Cloud/DevOps: Docker, AWS, Kubernetes, CI/CD, Redis

Tools: Git, GitHub, Electron, Vite, VS Code

PROJECTS

MEMORA — Persistent Memory Infrastructure for AI Agents

Sep 2026 – Oct 2026

Python · FastAPI · LangChain/LangGraph · MCP · Gemini · Celery · Redis · PostgreSQL · Qdrant · OpenSearch · Neo4j | [Live](#)

- Built persistent memory infrastructure for AI agents with asynchronous extraction, lifecycle management, source attribution, and tenant isolation.
- Designed a multi-tenant architecture where FastAPI/MCP APIs feed Celery + Redis workers that write to 4 purpose-built stores: PostgreSQL (source of truth), Qdrant (vector search), OpenSearch (BM25 recall) and Neo4j (entity graph).
- Engineered hybrid ANN + BM25 + graph retrieval with rank fusion and relevance/recency-based reranking for accurate memory recall.
- Integrated MCP and LangChain/LangGraph for reusable agent memory; Celery and Redis handle distributed ingestion and recovery.
- Achieved 92% Recall@5 with sub-250 ms p95 latency, benchmarked on 10K+ memories from multi-session agent workloads.

WorkForge — Agentic AI Workflow Automation

Jan 2026 – Mar 2026

Python · FastAPI · LangGraph · Gemini · Supabase · OAuth · GitHub/Gmail/Calendar APIs | [Live](#) | [GitHub](#)

- Developed an agentic workflow automation platform using LangGraph and Gemini to turn GitHub, Gmail and Calendar signals into prioritized tasks and automated workflows.
- Architected an event-driven pipeline over 3 integrations (GitHub, Gmail, Calendar): signals are deduplicated, analyzed by a LangGraph agent, approved by a human, then executed, with organization-scoped state in Supabase PostgreSQL.
- Connected OAuth-backed GitHub, Gmail and Calendar access with scheduling and Discord notifications.
- Hardened reliability and security with encrypted OAuth tokens, event deduplication, audit logging, retries and recovery.

OfflineChatBot — Local Multimodal RAG Assistant

Aug 2025 – Sep 2025

Python · LangGraph · Ollama · ChromaDB · HuggingFace · MiniCPM-V · FastAPI · React · Electron | [Live](#) | [GitHub](#)

- Created an offline-first multimodal AI assistant that orchestrates 3 local models (chat, code, vision) with LangGraph, persistent semantic memory and zero cloud dependency.
- Implemented RAG pipelines for 3 document formats (PDF, DOCX, TXT) using HuggingFace embeddings, recursive chunking and ChromaDB vector retrieval, plus MiniCPM-V for local image understanding.
- Packaged the app as an Electron desktop client (FastAPI/React) with automated builds and installers for 2 platforms (macOS, Windows).

EDUCATION

B.E., Information Science & Engineering

2023 – Jun 2027

JSS Academy of Technical Education | CGPA: 8.0/10

CERTIFICATIONS

- Generative AI Professional Certification — Oracle
- Big Data — Infosys Springboard

ACHIEVEMENTS

- Solved 400+ Data Structures & Algorithms problems.
- Selected for a university-wide hackathon among 100+ competing teams.