Ripunjay Kashyap

AI / ML Engineer
ripun.j.kashyap@gmail.com | +91 8940621897 | Bangalore, Karnataka, India
linkedin.com/in/ripunjay-kashyap | github.com/ripunjkashyap-a11y | ripunjay.vercel.app

PROFILE

AI Engineer & IEEE-published researcher specialising in production agentic systems - LangGraph multi-agent pipelines with deterministic guardrails that prevent LLMs from hallucinating numerics, rules, or physics constraints. Multiple systems shipped: hybrid RAG at 0.907 faithfulness, XGBoost at R^2=0.98 RUL, and a FastMCP server making a 6-stage ML pipeline callable by LLM agents.

EDUCATION

B.E. Computer Science Engineering 2021 - 2025
Sathyabama Institute of Science and Technology, Chennai, India
Best Project Award - Excellence Day 2025
AR/VR based Campus Navigation System (IEEE ICISS 2025)

TECHNICAL SKILLS

ML / AI:	Python · PyTorch · TensorFlow/Keras · XGBoost · Scikit-Learn · SHAP · Librosa · HTDemucs
LLM & Agents:	LangGraph · LangChain · RAG · RAGAS · ChromaDB · Qdrant · HuggingFace · Groq · Gemini
Backend & MLOps:	FastAPI · Flask · Pydantic v2 · Docker · Podman · MCP (FastMCP) · LangSmith
Frontend:	React · Next.js · Tailwind CSS · Vite · Flutter · Mapbox GL JS

PUBLICATIONS & RESEARCH

AR/VR based Campus Navigation System (CNS) DOI: 10.1109/ICISS63372.2025.11076255
IEEE Xplore · ICISS 2025 (7th Int'l Conf. on Intelligent Sustainable Systems) | March 2025

- Co-authored. Designed AR/VR campus navigation (Flutter + Mapbox GL API + AR Core/Unity); manually collected all campus POIs with a Garmin GPS device; deployed as in-house university system.
- Web POC (Next.js + Mapbox GL JS v3): 4 navigation modes on one map instance via pitch/zoom/bearing changes - no re-renders. POC live: https://campus-navigation-poc.vercel.app

PROJECTS

Zenic - AI Health & Nutrition Assistant [GitHub]
Python · LangGraph · Groq (Llama 3.3 70B) · ChromaDB · Qdrant · SentenceTransformers · RAGAS

- Hybrid RAG (BM25 + vector + cross-encoder reranking) over 10,201-chunk corpus; scored **0.907** RAGAS faithfulness / **0.799** context precision with Gemma 4 31B as LLM-as-judge.
- LangGraph StateGraph routes 6 intent classes; deterministic math nodes (BMR/TDEE) fully isolated from LLM - zero hallucinated numerics. **53** automated tests: unit, integration, node-sequence, RAGAS.

SHPS v2.0 - Structural Health Prediction System [GitHub] [Demo]
XGBoost · Keras/LSTM · Flask · SHAP · Pydantic v2 · Docker · ReportLab

- XGBoost quantile regression (Q0.025/0.5/0.975) for calibrated 95% CI on Remaining Useful Life; **R^2=0.98** (MAE 1.58 yrs) on RUL, R^2=0.89 on health score. Deployed on HuggingFace Spaces.
- SHAP background compressed 10k rows -> 100 KMeans centroids; attribution latency under **50 ms** per request. Pydantic v2 physics layer rejects cross-field impossible sensor inputs before any model runs.

SoundReverse - AI Mastering Analyst [GitHub] [Demo]
LangGraph · Gemini API · FastAPI · React/Vite · Tailwind CSS · LangSmith · FPDF2

- LangGraph Analyst-Critic loop reverse-engineers EQ/compression from audio fingerprints; rules.yaml owns all numeric values - LLM writes only reason strings, preventing hallucinated audio settings.
- LangSmith observability surfaces a public shareable trace URL per run. React + FastAPI frontend on Vercel with 5 pre-computed SignalSignatures enabling sub-second agentic pipeline demos.

Audio Stem Splitter - Custom MCP Server [GitHub]
FastMCP · PyTorch · HTDemucs · CLAP · Librosa · FFmpeg · asyncio · Docker

- FastMCP server wraps a **6-stage** ML pipeline (yt-dlp -> FFmpeg -> HTDemucs source separation -> Librosa MIR -> CLAP **512-dim** embedding) callable by LLM agents via MCP stdio protocol.
- Idempotent per-stage checkpointing and asyncio.Lock serialisation handle **10-15 min** CPU inference latency; CLAP falls back to librosa embeddings maintaining identical 512-dim output schema.