Keshav Bihani

+91 94069-54553 | bihanikeshav@gmail.com | github.com/bihanikeshav | meownikov.xyz | linkedin.com/in/bihanikeshav

3+ years shipping production AI systems — from RAG assistants to real-time inference to agentic pipelines.
Fascinated by systems at scale: why they break and how to keep them from breaking.

Skills

LLM Systems: RAG, Agentic Pipelines, Text-to-SQL, Prompt Engineering, Evaluation, Vector Search (FAISS, ChromaDB)
ML & Inference: PyTorch, TensorRT, Triton Inference Server, YOLOv8, SlowFast, real-time model serving
Backend & Data: Python, FastAPI, Node.js, Go, PostgreSQL, MongoDB, Redis, Docker, Azure

Experience

AI Engineer Jan 2023 – Present
Arcturus Business Solutions

- **Vaani-AI**: RAG safety assistant deployed across 2 industrial plants, **10,000+ active users**. Vision+LLM pipeline for automated work-zone compliance, grounded in plant-specific SOPs.
- **SmartDoc**: LLM pipeline that turns unstructured tender documents into structured SQL — hallucination-resistant risk scoring and automated multi-vendor bid comparison for enterprise procurement.
- Built a **4-stage agentic pipeline** for a fintech client: natural language to SQL over 3,000+ Indian equities. FAISS entity resolution, 19-rule preprocessing with phased LLM routing — simple lookups skip the LLM entirely (zero cost), complex queries go through Haiku then Sonnet.
- Microservices video analytics platform (Go, Node.js, Triton, TensorRT) — **140% throughput**, 44% latency reduction across 100+ camera streams.
- Fire detection at scale: SlowFast temporal pipeline handling **100+ concurrent streams**, <1% false positives, 20× throughput over frame-based baselines.
- Hyperspectral classifier: **2.1s** → **180ms** inference (12× speedup) — 224-band NIR, 7 plastic types, 92% production accuracy.
- Crane safety monitoring: YOLOv8 + monocular depth (Depth Anything V2) → 3D geometry reconstruction, self-calibrating — **1–7% height error** in production.
- Drone-based safety monitoring for Odisha DISCOM: RTMP real-time video ingestion pipeline for live SOP compliance assessment from aerial feeds.

Technical Co-Founder Oct 2023 – Apr 2024
Zoop.Buzz

- Anonymous social platform for research institute students — **3,500 users**, 21% D7 retention, DAU/WAU >50%. Owned product, full-stack (React, Node.js, MongoDB), and growth.

Projects

Promptry — LLM Regression Testing | *Python, SQLite, sentence-transformers* 2026
promptry.meownikov.xyz | PyPI

- Prompt versioning via SHA-256 dedup, eval suites with semantic + LLM-as-judge assertions, drift detection over score history, and 25+ safety templates. Local-first, zero infrastructure, one-line integration. 101 tests across 11 files.

Doer — AI Productivity App | *React Native, Node.js, LLM APIs, GitHub API* 2026
doerapp.in

- Syncs with calendar and GitHub to break goals into adaptive micro-tasks based on available time. Startup team mode for task delegation, blocker resolution, and repo analytics.

Pitch Perfect — Vocal Analysis Platform | *Python, React, FFT, Transformers* 2025
pitch.meownikov.xyz

- Transformer-based vocal separation, spectrogram classifiers for technique detection, FFT spectral comparison for pitch scoring, and emotion analysis via musical scale features + lyrical sentiment.

Automated Twitch Highlights | *Node.js, Python, LSTM* 2023

- LSTM pipeline that identifies high-engagement moments in Twitch streams from chat frequency signals, then classifies them (funny, exciting, surprising) using chat-content sequence models.

Education

Indian Institute of Science Education & Research (IISER), Bhopal Aug 2017 – Jun 2022
BS-MS in Electrical Engineering & Computer Science, Minor in Physics