Rikhil Nellimarla

+91 7386175224 | nrikhil@gmail.com | linkedin.com/in/rikhil-nellimarla | github.com/Rikhil-Nell | remiscus.me

Career Objective

AI and Backend Developer with a background in ECE, specializing in building real-time, agentic systems. Experienced in designing scalable APIs, hybrid memory architectures, and event-driven AI pipelines. Proficient across the full stack of modern AI, including LLMs, multimodal systems, edge robotics and deployment and lifecycle of these systems.

Experience

AI Voice Engineer — Virtual (US-based)

TargetDial — AI voice automation & CRM orchestration for SMBs and real estate — *Jan 2026*

- Shipped production AI voice agents handling 500–600 calls/day for a single client, covering inbound lead qualification, outbound follow-ups, call transfers, and human handoff logic across SIP/Asterisk/FreePBX stacks.
- Designed end-to-end call workflows spanning voice → intent detection → CRM actions → calendar booking, integrating GoHighLevel APIs, n8n automations, and custom webhook-based control planes.
- Deployed self-hosted FreePBX infrastructure to route calls from LiveKit and Retell, building a robust call logging pipeline with per-lead attribution and downstream analytics.

AI Engineer — Hyderabad, India

Indominus Labs — AI solutions provider for enterprise clients like DirecTV — *May 2025 – Present*

- Pioneered a SIP trunking integration to bridge LiveKit's dynamic IPs with legacy telephony providers (VICIDIAL), architecting a user/password auth solution to overcome a critical infrastructure blocker.
- Engineered a scalable data pipeline on a bare-metal GH200 to process over 1 million customer call logs, handling transcription (Whisper v3), PII sanitization, and multi-agent performance analysis.
- Developed and deployed real-time, outbound voice agents with tool-use (web search) and intelligent call handoff logic, orchestrated via a custom FastAPI endpoint.

AI Engineer — Virtual

Stealth VC Project — *June 2025 – Present*

- Developed and deployed a custom turn-detection model for Hindi, overcoming a core limitation in the LiveKit SDK to enable natural, low-latency conversations for non-English voice agents.
- Architected a scalable telephony infrastructure using an Asterisk server as middleware to route calls from LiveKit to various endpoints (e.g., WhatsApp), ensuring static IP compatibility.

Founding AI Engineer — Virtual

Clink — Pre-seed stage cafe analytics & marketing startup — *May 2025 – Present*

- Built a full-stack AI backend in FastAPI with a multi-agent architecture using Pydantic AI, orchestrating 8+ specialized agents (analysis, chat, forecasting, image generation) with lazy-loaded caching and structured output validation.
- Engineered an end-to-end business intelligence pipeline performing RFM segmentation via K-Means clustering, customer churn detection, and co-occurrence analysis to auto-generate targeted coupon strategies for 80+ onboarded cafes.
- Designed a template-driven offer generation system with 8 coupon variants (winback, stamp cards, happy hours, combos), integrating LLM-powered ROI forecasting and AI image generation (GPT-5 + Gemini 3) for marketing assets.
- Implemented production infrastructure including PostgreSQL with asyncpg connection pooling, Redis caching, S3 integration for assets, and full observability via Logfire instrumentation across all async services.

Technical Lead — Hyderabad, India

Open Source Community (Mozilla), VIT Amaravati — *May 2025 – Present*

- Led and mentored a community of 100+ student developers in open-source best practices, version control, and collaborative software development.
- Organized and conducted technical workshops on topics such as WebRTC, API Design, and Backend Development, fostering technical growth for members.
- Oversaw the development and contribution to 6 open-source projects, setting the technical roadmap and performing code reviews to ensure quality.

Projects

Traction: AI Startup Ideation & Pitch Deck Generator | *FastAPI, Pydantic-AI, PostgreSQL, OAuth* Feb 2026

- Identified a gap in founder tooling for structuring early-stage startup ideas into investor-ready formats; built and launched a platform at hackathon and actively developing toward full release.
- Engineered an AI-driven SEO pipeline using Pydantic-AI to auto-generate machine-readable `llm.txt` endpoints, making startup profiles discoverable by AI agents and search algorithms.
- Built a single-URL pitch system aggregating AI-generated full-screen pitch decks, investor summaries, and structured startup specs into one shareable page.

Kinesys-CRM: AI-Powered CRM Platform | *FastAPI, Vue, WebSockets, LiveKit, Google OAuth* Jan 2026

- Identified that early-stage B2B founders managing 200–400 leads were running cold outreach manually with no lightweight CRM option; built a minimal AI-powered CRM with voice agents, automated follow-ups, and calendar integration to fill that gap. Actively building toward full release.
- Implemented real-time WebSocket communication, LiveKit voice agent integration, and Google OAuth with Calendar sync in an async FastAPI backend with a Vue frontend.

IntelliPost: AI-Powered Postal Mail Sorting | *FastAPI, OpenAI Vision, Cloudflare R2, 3rd Place @ Post-a-thon* Jan 2026

- Developed a mobile backend that automates postal mail sorting by extracting address information from envelope images using OpenAI Vision API and Pydantic AI for structured outputs.
- Integrated India Post's Pincode API for smart routing to sorting centers, with Cloudflare R2 for secure image storage and async background processing for scalable mail extraction.

Lexalytics: Legislative Sentiment Analysis | *FastAPI, GPT-4.1, PostgreSQL, Alembic, SIH Project* Dec 2025

- Built a backend for the Ministry of Corporate Affairs to analyze sentiment in stakeholder comments on legislative drafts, featuring PDF upload, AI-powered summarization, and batch CSV processing.
- Implemented JWT authentication, async PostgreSQL operations with SQLModel, and automated sentiment analysis pipelines using GPT-4.1 and GPT-5-nano.

VISU-X: Agentic Humanoid Robot Backend | *FastAPI, Groq, Whisper, WebRTC, DeepFace, Supabase, RPi5*

- Inherited a stalled humanoid robot project in its final month after budget was exhausted, independently architecting and shipping a multimodal real-time backend combining vision transformers and LLMs on RPi5 edge hardware under a hard demo deadline.
- Solved concurrent STT/TTS and computer vision on constrained hardware, achieving stable real-time performance with Whisper, PyAudio, WebRTC VAD, and DeepFace running simultaneously.
- Implemented cross-session facial recognition with persistent hybrid memory via pgvector, enabling personalized user recognition — project was subsequently presented to the Chief Minister of AP at VLaunchPad.

Multi-Agentic Graph RAG | *PyPDF, WebSearch, DuckDuckGo, Pydantic AI, Streamlit, Neo4j*

- Developed a multi-agent Retrieval Augmented Generation (RAG) system with a graph-based approach to knowledge retrieval and synthesis using Neo4j.
- Implemented data ingestion pipelines capable of parsing and structuring information from websites, Markdown, DOCX, and PDF files using PyPDF and web search capabilities.
- Built an interactive frontend using Streamlit with deep research functionality powered by DuckDuckGo search and Pydantic AI to query the multi-document knowledge base and visualize agentic reasoning paths.

MemoryWeave: AI-Powered Storytelling | *YOLO World, Groq LLMs, Flask, Winner @ HackSRM 2024*

- Developed a multi-agent system that converts real-time object detection streams into structured, AI-generated stories and interactive timelines.
- Achieved ¡100ms latency via async processing and caching for dynamic, prompt-aware event filtering from YOLOv8 and OpenPose detections.

Converso: Voice-Based Conversational RAG System | *ESP32-S3, IoT*

- Developed an IoT device using an ESP32-S3 to record, transcribe, and store conversations as structured documents for a RAG system.
- Integrated a contextual chatbot that retrieves information from past discussions to enable seamless, voice-based interactions.

Dungeons & Fallacies: Logic-Driven RPG | *Multi-Agent Systems, LLMs*

- Created a unique text-based RPG where combat is framed as a logical debate, using AI agents to simulate turn-based arguments against fallacy-themed mobs.
- Designed an AI judge to evaluate the soundness of player arguments against a mob's "Null Hypothesis," determining outcomes based on logic.

Pixy: Self-Hosted Discord Bot | *Groq, Supabase, pgvector, NGINX*

- Built and deployed a self-hosted, AI-powered Discord bot for a 100+ member tech server, running on a private web server behind an NGINX reverse proxy.
- Implemented persistent, user-specific memory using Supabase with pgvector, allowing for isolated and personalized conversations.

Terminal Chatbot | *Groq, Textual*

- Built an intelligent terminal-based chatbot with tools to execute bash scripts and perform various automation tasks directly from the command line.
- Integrated Groq LLMs with Textual for a rich terminal user interface, enabling seamless interaction between natural language commands and system automation.

Technical Skills

Programming Languages: Python, Java, C, C++, Go, SQL, JavaScript/TypeScript
Machine Learning & AI: PyTorch, Transformers, Scikit-Learn, Spacy, Unsloth, DeepStream, Voiceprint DSP (FFT), Computer Vision, NLP, RAG, LLMOps
Agentic & Generative AI Frameworks: LangChain, LangGraph, Pydantic, Pydantic_AI, LiveKit, Retell
Backend & Telephony: FastAPI, REST APIs, WebRTC, WebSockets, Asterisk, FreePBX, SIP Trunking, RTP/NAT Traversal
CRM & Automation: GoHighLevel (GHL), n8n, Webhooks, Google Calendar API, OAuth 2.0
Databases & Storage: PostgreSQL, MySQL, Supabase, Redis, Pinecone, Weaviate, Cloudflare R2
Frontend: Vue.js, Streamlit, Textual (TUI)
Cloud & DevOps: AWS (IAM, S3, EC2, ECS, SES, SQS, Bedrock), Docker, NGINX, GitHub Actions, VPS/Linux, Logfire
Core Competencies: Voice AI Systems, Telephony Infrastructure, CRM Integration, Backend Architecture, Agentic Systems, Generative AI, AI/ML/DL

Education

Vellore Institute of Technology, Amaravati — Amaravati, AP
Pursuing B.Tech in Electronics and Communication Engineering - 5th semester; GPA: 8.46/10.0 — *2023 – 2027*

Achievements and Activities

- Selected participant in Y Combinator Startup School, engaging in structured startup validation, founder-market fit exploration, and early-stage product development with global founder cohorts.
- Event Director for Recon 2026 (VIT-AP's flagship tech fest), leading end-to-end planning and execution of a multi-day event involving cross-functional teams, sponsorship coordination, speaker onboarding, and large-scale participant management.
- Presented the VISU humanoid robot project to the Chief Minister of AP, Mr. Chandrababu Naidu, at VLaunchPad.
- Winner of HackSRM 2024 (Open Innovation track) for the MemoryWeave project.
- Secured 3rd place in Post-a-thon conducted in VIT-AP, and got to showcase my mobile app for automating pincode mapping to Union Minister Shri Chandra Sekhar Pemmasani.
- Delivered technical talks on agent-based architectures and LLM security to 100+ attendees per session at venues including Null Vijayawada, NIT Warangal, and SRM University.
- Led the backend and on-site tech team for the AP CID home guard exams.
- Led multiple paper reading sessions on the architecture of the Pixy Discord bot and related AI systems.