Bibhanshu Raj

+91 9572487227 | bibhanshuraj@icloud.com | bibhanshuraj | github | portfolio

Work Experience

Makunai Global **Jan 2025 – Present**

AI Engineer & Lead Developer *Noida*

- Built and scaled a **production real-time voice AI platform from prototype** — owned the **VAD → STT → GPT-4o → TTS** worker pipeline built on **LiveKit + Deepgram WebSocket API**, achieving **sub-800ms end-to-end latency** across **100+ concurrent sessions**.
- Designed a **decoupled control/data plane architecture** using **FastAPI** orchestration and horizontally scalable worker processes, improving fault isolation and enabling independent scaling of inference workloads.
- Built and integrated a **Google Calendar tool layer** with OAuth 2.0 and a **multi-tenant RAG pipeline** using **pgvector** (1536-dim cosine search, tenant-isolated retrieval), improving response grounding and reducing hallucination in production queries.
- Contributed to a **VAPI-style integrations layer** (Google Calendar, Slack, WhatsApp, Calendly) with secure credential handling, enabling real-time tool execution directly from voice interactions across **6+ Indian languages** via **Sarvam AI**.
- Core contributor to **MakunFlow**, an in-product AI workflow automation engine — implemented node execution logic for LLM calls, tool routing, and conditional flows in a **DAG pipeline executor**, backed by **Redis Pub/Sub** and **Celery** for asynchronous processing.

SmartLink Holdings **May 2024 – July 2024**

Software Engineer Intern — Backend & Data Systems *Goa, India*

- Developed backend inventory services using **Python** and **PostgreSQL**, supporting **50K+ daily API requests** with high availability across warehouse operations.
- Optimized database performance through indexing, query restructuring, and **Redis caching**, reducing query execution time by **60%** and improving system throughput by **35%**.
- Improved service latency by introducing asynchronous request handling and connection pooling, reducing average API response time by **40%**.

Technical Skills

Languages: Python, C/C++, SQL

AI & ML: PyTorch, TensorFlow, scikit-learn, LangChain, HuggingFace Transformers, OpenAI API, RAGAS

Backend & Infra: FastAPI, WebSockets, Docker, Redis, Celery, Git, GitHub Actions

Databases: PostgreSQL, MongoDB

Projects

Multi-Document RAG Chatbot | *LangChain, GPT-3.5, Python* **Aug 2024 – Dec 2024**

- Built a RAG chatbot over **500+ academic documents** achieving **95% source-grounded accuracy**; implemented multi-turn reasoning and a claim validation pipeline.
- Designed fallback logic to restructure **80%** of indirect queries into structured, answerable inputs, improving response reliability.

AI-Powered Evaluation Bot | *Python, OpenAI API, NeMo Guardrails, RAGAS* **Feb 2024 – May 2024**

- Automated **85%** of grading steps across **200+** submissions, reducing manual review time by **60%** with **95% accuracy**.
- Enforced topic compliance using **NeMo Guardrails** and improved factual consistency with custom **RAGAS**-based evaluation metrics.

Education

BITS Pilani **Expected Graduation: May 2027**

Bachelor of Computer Science and Information Technology *Pilani, India*

- Relevant Coursework: Artificial Intelligence, Operating Systems, Data Structures and Algorithms, Database Systems, Computer Architecture, Data Science