Himanshu Yadav

+91 93183 38519 | radhekrishna8267@gmail.com | linkedin.com/in/himanshuy26 | portfolio/himanshu.github.io

Summary

Computer Science student interested in applied AI and ML systems. I build end-to-end NLP and LLM applications including vector retrieval pipelines, agent workflows, and structured inference systems. My work focuses on the engineering side of machine learning - designing robust systems that integrate models with real infrastructure.

Education

Galgotias University Gautam Buddha Nagar
B.Tech in Computer Science and Engineering *2023 – 2027*

Technical Skills

Languages: Python, SQL
AI / ML Systems: NLP, LLM Applications, Embeddings, Vector Retrieval, Multi-Agent Systems, Prompt Engineering, Model Evaluation
Frameworks & Libraries: Transformers, Pytorch, Scikit-learn, FastAPI, Flask, Pydantic, SQLAlchemy
Backend & Infrastructure: REST APIs, Async Processing, PostgreSQL, Vector Databases, APScheduler
Tools & Platforms: Git, GitHub, Render, Neon, Docker

Projects

MeetSync - AI Meeting Minutes Generator | Live Demo | *Python, Flask, NLP* February 2026

- Built an AI meeting minutes system converting audio into structured decisions, tasks, and deadlines.
- Processed 30-minute meetings in ∼35s using Deepgram transcription and Gemini summarization pipeline.
- Reduced response latency to <1s using asynchronous background threads for LLM processing.
- Ensured structured output reliability using Pydantic validation and PostgreSQL-backed schema enforcement.

Memoria - Persistent Memory for AI Agents | Project Link | *Vector Databases* March 2026

- Architected persistent memory using Endee Vector DB and 384-dim embeddings to extend AI context.
- Optimized semantic retrieval with cosine similarity (0.45) and recency weighting to filter irrelevant memories.
- Integrated vector memory with Llama 3.3-70B (Groq) for context-aware responses via memory-injected prompts.
- Engineered memory lifecycle operations including a semantic delete (`forget:`) for vector management.

DuelMind - Multi-Agent LLM Deliberation Framework | Project Link | *FastAPI, LLM Agents* Present

- Developed an orchestration framework for Llama-3.3 and Qwen3 to conduct autonomous turn-based discussions.
- Implemented structured argument extraction and conflict detection using Pydantic-enforced schemas.
- Designed an evaluation loop with consensus scoring, reflection cycles and 3-model synthesis via Kimi K2.
- Built a FastAPI backend with a live JS visualization dashboard for real-time consensus tracking.

Hackathons & Achievements

- **Hack The Box Global Hackathon:** Selected as one of **4 teams** representing Galgotias University for the global cybersecurity hackathon hosted by YNOV Campus, Paris (April 2026).
- **Hackaccino 3.0 National Hackathon:** Developed a data-driven women's safety platform predicting area-wise risk levels from historical crime datasets; secured **17th place out of 194 teams**.
- **NASA Space Apps Challenge:** Designed and built an interactive educational web platform in a team to simplify exoplanet science for ages 5–15, translating exoplanet concepts into engaging visual explanations.

Certifications

- Microsoft AI Azure Virtual Internship - Microsoft & Eduskills
- Python Programming - HackerRank
- Data Science & Analytics - Pregrad