Rajesh Yarra

github.com/Rajesh9998 | ✉ rajesh.yarra1241@gmail.com | linkedin.com/in/rajesh-yarra-850b40254 | ☎ +91 7093670671 | 🌐 rajesh-yarra.vercel.app

Summary

Generative AI Engineer specializing in Large Language Models (LLMs), Retrieval Augmented Generation (RAG), Agents, and intelligent systems. Passionate about building innovative AI-powered applications that address real-world challenges. Proven ability in leveraging diverse AI models (GPT, Gemini, Claude, Llama) and creating multi-agent architectures. Strong foundation in computer science principles and software development.

Education

Andhra Loyola Institute of Engineering and Technology — **Vijayawada, India**
B.Tech in Information Technology; CGPA: 8.01 — *2021 - 2025*

Sri Chaitanya Junior College — **Mangalagiri, India**
Class XII; Marks: 865/1000 — *2021*

Nirmala EM High School — **Atmakur, India**
Class X; GPA: 9.3 — *2019*

Skills

Programming Languages: C, C++, Java, Python, R
Generative AI: LLMs, Retrieval Augmented Generation (RAG), Agents, LangChain, Graph RAG
Vector Database: Pinecone
Frameworks/Libraries: LangChain, Agno, CrewAI, Streamlit, Next.js, React, Mem0, Supabase
Course Work: Problem Solving, Data Structures and Algorithms, OOPS, Database and Management Systems, Operating Systems, Computer Networks, Cyber Security, Generative AI, Artificial Intelligence & Machine Learning
Soft Skills: Leadership, Communication, Adaptability, Time Management
Languages: English (Full Professional Proficiency), Hindi (Full Professional Proficiency), Telugu (Native Proficiency)

Experience

Supraja Technologies — **May 2023 - Jul 2023**
Ethical Hacking & Cyber Security Intern

- Performed Assessment Tasks Like conducting XSS attacks, SQL Injections and vulnerability scans on websites
- Performed Penetration testing on vulnerable websites using tools like Nmap, Nessus, and Burp Suite
- Worked on a final project creating a Python application called Audio Steganography to embed/hide images and text in audio without external detection
- Tech Used: Python, LSB Algorithm, Fernet & tkinter libraries

Research Experience

Pentester's Copilot: Context Aware, Agent Powered, Pentest Perfected

- Introduced a novel framework leveraging LLMs to enhance skills and efficiency of cybersecurity professionals performing offensive security tasks
- Implemented a persistent memory layer using Mem0 and Supabase enabling personalized, context-aware interactions across sessions
- Built on the ReAct (Reasoning and Action) framework with LangChain orchestrating autonomous planning and tool interaction
- Leveraged state-of-the-art LLMs including Meta Llama 3.1-405b, Google Gemini 2.0 Flash, and DeepSeek R1 for generating recommendations, crafting exploit payloads, and analyzing target environments

LLMPatronus: Leveraging Large Language Models for Advanced Vulnerability Analysis

- Research exploring the potential of LLMs in identifying vulnerabilities while addressing limitations such as hallucinations and limited context length
- Proposed a robust AI-driven approach using innovative methodologies combining Retrieval-Augmented Generation (RAG) and Mixture-of-Agents (MoA)
- Focused on mitigating limitations of traditional static and dynamic analysis tools that suffer from high false positive rates
- Leveraged strengths of LLMs while addressing their weaknesses to provide dependable and efficient AI-powered solutions for software security

Projects

LLMPatronus — Sep 2024

GitHub

- Python-based application using a multi-agent LLM system to identify vulnerabilities in Android apps
- Combined models like GPT-4o-mini, Claude-3-haiku, Qwen-2-75B, and Gemini-1.5 pro to create a Mixture of Agents (MoA) approach
- Implemented Retrieval Augmented Generation (RAG) and Pinecone Vector Database for efficient data storage and retrieval
- Accurately detected 87.5% of predefined vulnerabilities by integrating with Google AI Studio API
- Tech Used: Python, LLMs (Gemini 1.5 Pro, Gemini 1.5 Flash, OpenAI-o1-preview, Meta Llama-3.1-405b & 70b, DBRX-Instruct, Qwen2-75b, ChatGPT-4o-mini, Claude-3-haiku), Pinecone API (RAG), Mixture of Agents

Business QA Bot — Aug 2024

GitHub

- Developed an AI bot leveraging RAG and Firecrawl API for efficient data collection and Gemini-1.5-Flash LLM for complex business queries
- Integrated Groq LPU™ to ensure rapid and accurate responses to user inquiries
- Tested on OpenAI's website, demonstrating adaptability to various business domains
- Designed for applications including customer support and internal knowledge management
- Tech Used: Python, RAG, Firecrawl API, Gemini-1.5-Flash, Groq LPU™ AI

Pentester's Copilot — Apr 2024

GitHub

- Built an AI-powered assistant for cybersecurity professionals leveraging LLMs, multi-modal capabilities, and a persistent memory layer
- Implemented interactive guidance and autonomous task execution for penetration testing workflows
- Tech Used: Next.js, React.js, FastAPI, Supabase, Mem0

Achievements

- Participated in Andhra Pradesh State Skill Competition 2024 organized by APSSDC and made it to the Finals of State Level Round, demonstrating skills in Windows Hardening
- Participated in Capture The Flag Hackathon 2023 organized by Supraja Technologies and solved 8 problems related to Packet Capture Challenges and Cryptography, excelling as Best Student of the Hackathon
- Certified in C programming by NPTEL (July 2023) with 79% score
- Won 1st Place in a Technical Quiz Competition (Sep 2023) conducted in college, competing against many senior participant teams
- Ranked 13th in Accenture Innovation Challenge 2024 among many teams from India with the LLMPatronos project

Certifications

- Microsoft's Build an Azure AI Vision Solution (Feb 2024): Demonstrated proficiency in creating computer vision solutions using Azure AI Vision, including analyzing images, creating custom models, and implementing vision-based AI applications
- Microsoft's Azure AI Intelligent Document Processing Solution (Feb 2024): Acquired skills in developing document intelligence solutions, including programmatic data analysis in forms, creating custom models, and extracting key-value pairs from documents using Azure AI services
- Tata Consultancy Service's Cybersecurity Analyst Simulation (Jan 2024): Acquired expertise in AIM Principles, Cybersecurity best practices, and strategic alignment with business objectives; delivered comprehensive documentation and presentations, showcasing the ability to communicate complex technical concepts effectively