Sushil Kumar Das

sushilkumardas07@gmail.com | +91-9348255351
GitHub – github.com/sushildas100
LinkedIn – www.linkedin.com/in/sushil-kumar-das-911087a7

AI Engineer with over 1 year of experience in designing and deploying **agentic AI systems**, **LLM orchestration**, and **context-aware architectures**. Specialized in **Model Context Protocol (MCP)**, **Retrieval-Augmented Generation (RAG)**, and **NLP pipelines** using Rasa, Llama-2, and DistilBERT. Demonstrated expertise in reducing latency by 40%, improving accuracy by 35% (F1-score: 92%), and minimizing hallucinations via LoRA fine-tuning. Proficient in Python, Java, and MLOps. Proven expertise in document intelligence, chatbot development, and ML-driven analytics for real-world applications.

PROFESSIONAL EXPERIENCE

INVENIO BUSINESS SOLUTIONS, **Hyderabad, India**
Senior Associate Consultant – AI *June 2024–Present*

- **Agentic AI & MCP Systems:**
  - Designed **stateful agents** using **Model Context Protocol (MCP)** for multi-turn dialogue management in a Discord chatbot, achieving 95% intent recognition accuracy with Rasa's DIETClassifier.
  - Implemented **hierarchical agent architecture** (orchestrator → specialist agents) for dynamic intent routing and session persistence, reducing context loss by 30%.
  - Integrated **semantic search** with FAISS indexing for knowledge retrieval, enabling context-aware responses in real-time (150ms average latency).
- **Hybrid RAG Solutions**:
  - Engineered a **Resume Parser** using **DistilBERT + Llama-2** for unstructured PDF/DOCX data extraction, achieving 92% F1-score on entity recognition.
  - Optimized RAG pipeline with **RLHF-based ranking** and sliding-window chunking, reducing retrieval latency by 40% on AWS EC2 g4dn.2xlarge (NVIDIA T4 GPU).
  - Built a **document sensing system** with **Sentence-BERT** and LangChain for querying unstructured data, improving query accuracy by 35%.
- **Tax and Revenue Chatbot:**
  - Developed a **Rasa-based chatbot** with custom NLU (DIETClassifier), slots, and action.py for automating tax workflows, processing 10,000+ GST records with 98% slot-filling accuracy.
  - Incorporated **BERT embeddings** for slot filling and anomaly detection in revenue streams using Random Forest and Z-score (false positives reduced by 25%).
  - Developed Rule-based anomaly detection in revenue streams and automated action.py workflows for GST classification.
- **MLOps & Optimization**:
  - Reduced web scraping complexity from $O(n^2)$ to $O(n \log n)$ using depth-optimized BFS/DFS, deployed on GPU-accelerated EC2 instances (g4dn.2xl larges).
  - Deployed **Ollama** for local LLM serving with NGINX load balancing (200 RPM) and Kubernetes for high availability (99.9% uptime).
  - Implemented **Prometheus** and **Grafana** for real-time monitoring, reducing system downtime by 15%.

INVENIO BUSINESS SOLUTIONS **Hyderabad, India**
Java Full Stack Intern *January 2024–March 2024*

- Learned software engineering (UML, data flow diagrams), LINUX commands, operating systems, DBMS, Docker, MySQL, and advance Java (JDBC, OOP, REST API, Apache Tomcat)
- Built a project with HTML, CSS and JavaScript for the frontend, and using Java as backend. Managed endpoints with Apache Tomcat and Postman, maintaining a singleton design pattern.
- Used POST to store data and GET to retrieve data in MySQL, ensuring smooth frontend-backend integration.

INTERNPE PVT. LTD. **Odisha, India**
Web Development Intern *April 2023–May 2023*

Created a responsive weather website using HTML, CSS, JavaScript and Rapid API, enabling real-time weather updates for user-input locations.

PROJECTS

Discord AI Agent with MCP

- Built a **stateful chatbot** using **Model Context Protocol (MCP)** for session-aware dialogues, integrating Rasa for intent recognition (95% accuracy).
- Orchestrated multi-agent system (NLU → DB query → response generation)
- Tech: Python, Rasa, AWS Lambda, Redis

Antaryami AI | https://antaryami-ai.netlify.app

- Developed a full-stack chatbot with **Google Gemini Pro API** and fallback to **Llama-2 via Ollama** for offline processing.
- Implemented Google OAuth V2 and Redis for session management, achieving 99.8% uptime and 200ms response time.
- Tech: React, Redux Toolkit, Node.js, MongoDB, NGINX

Tax Analytics Agent

- Created a **Rasa-based agent** for GST rule retrieval, using **BERT embeddings** for slot filling and Random Forest for anomaly detection (25% fewer false positives).
- Integrated **PowerBI** for real-time dashboards, visualizing tax compliance metrics for 5,000+ records.
- Tech: Rasa, Python, PowerBI, MySQL

Document Sensing Pipeline

- Built a pipeline for unstructured data (PDFs/DOCX) using **DistilBERT**, **Tesseract OCR**, and **FAISS** for semantic search, achieving 90% query accuracy.
- Optimized with **LangChain** and **Celery** for asynchronous processing, deployed on AWS ECS.
- Tech: Python, FastAPI, Sentence-BERT, Redis

TECHNICAL SKILLS

Category	Skills
AI Frameworks	MCP, RAG, LLMs (Llama-2, Gemini, Ollama), Agentic AI, NLP (Rasa, spaCy, NLTK), Transformers (DistilBERT, Sentence-BERT), SVM, LCM
Programming Languages	Python (PyTorch, LangChain, FastAPI), Java (Spring Boot), JavaScript (Node.js, React), SQL, C
Data Management	MongoDB, MySQL, Redis, REST API, PowerBI, FAISS
Web Development	React, Redux Toolkit, HTML, CSS, Node.js
Other	Web Scraping (BFS/DFS), Manual Robotics, IoT, MS Office (Word, Excel, PowerPoint), Salesforce, TFS Project Management

Languages: Fluent in English, Hindi and Odia

EDUCATION

GIET UNIVERSITY — **Odisha, India**
B.Tech, Computer Science and Engineering (AI/ML) — *2020–2024*

TSG GURUKUL — **Odisha, India**
12th Science in Physics, Chemistry & Mathematics — *2018–2020*

CERTIFICATIONS

- Salesforce Certified AI Associate
- Salesforce Certified Data Cloud Consultant
- Accenture Forage: Advanced Engineering (AI Track)
- NPTEL: The Joy of Computing with Python
- NPTEL: Cloud Computing
- MCA PL: MEAN Full Stack
- Basics of C by IIT Bombay

ACHIEVEMENTS

- Achieved 92% F1-score in hybrid RAG document parsing for ResumeParser ai.
- Implemented Antaryami as full fletched ai application. URL – https://antaryami-ai.netlify.app.
- Technical Lead: Students Association of Robotics Science (SARS) and Android/IoT Club, GIETU.
- Winner: Intercollege Robo-Race, Silicon Institute of Technology.