Soumyadeep Dhali

+91 8777890451 22f1001762@ds.study.iitm.ac.in linkedin.com github.com/soumyadeep-git

Education

Indian Institute of Technology, Madras **Jan 2022 – Sep 2025**
B.S. in Data Science and Applications, Minor in Economics *Chennai, Tamil Nadu*

Experience

Assured **Mar 2026 – Present**
AI Engineer - Consultant *Chennai*

- Built an AI voice-ops platform for payer enrollment follow-ups using ElevenLabs ConvAI, Twilio, Next.js, and PostgreSQL/Prisma, enabling automated outbound calls with dynamic provider, group, and health-plan context across 1–3 applications per call.
- Engineered a 1,000+ line post-call processing pipeline that transformed transcripts, webhook events, and tool payloads into structured operational data, capturing 10+ fields including enrollment status, effective date, provider ID, group ID, line of business, turnaround time, welcome-letter state, and IVR navigation traces.
- Designed 5 server-side agent tools across 15 API routes, backed by an 11-model Prisma schema for applications, calls, outcomes, comments, action items, and IVR playbooks to support reliable call logging and analyst review workflows.
- Built AI document-workflow APIs on FastAPI and AWS Lambda with 2 SQS queues, 4 secured API endpoints, DynamoDB-based async job tracking, and schema-guided PDF extraction/autofill pipelines spanning 30+ healthcare data tables and 5 document templates.

Solverminds **Sep 2025 – Feb 2026**
AI Engineer Intern *Chennai*

- Optimized a flagship RAG pipeline for the maritime domain, reducing query latency by over 80% (from 40s to 6–7s) by implementing hybrid search, reranking, and hierarchical chunking with Weaviate/FAISS.
- Built a Computer Vision-based OCR pipeline using image preprocessing and OpenCV techniques for document understanding and structured data extraction.
- Designed structured prompts and guardrails, including structured outputs and prompt engineering workflows, to improve grounding, reliability, and consistency of production RAG systems.
- Built a multilingual NL-to-SQL agent with LangGraph for querying multi-source data and generating visual insights.

Allcognix **May 2025 – Sep 2025**
Generative AI Application Engineer Intern *Remote*

- Engineered a scalable multi-agent backend using FastAPI, Celery, and Redis to automate complex workflows and enable task delegation across collaborative autonomous agents.
- Built and deployed an end-to-end RAG pipeline for enterprise knowledge retrieval, integrating S3 ingestion, vectorization, and context-aware querying over PostgreSQL and MongoDB.
- Designed a plug-and-play MCP framework for agent tool interaction and integrated AI workflows with HubSpot CRM APIs for automated data enrichment and workflow execution.
- Managed secure Docker deployments with Keycloak, Vault, and ELK-based logging and monitoring.

Projects

AgriFlow.ai | *Gen AI, CrewAI, RAG, LangChain, FastAPI, Node.js, React, PostgreSQL* **May 2025**

- Improved multi-source query accuracy by ˜30% by building a hybrid RAG agent combining PostgreSQL retrieval with vector search and reranking.
- Processed 1,000+ pages of PDFs into ˜8,000 semantic chunks with metadata-based filtering.
- Improved retrieval relevance by ˜30% using hybrid search + reranking over baseline vector search.

Real-Time Conversational AI | *LiveKit, Next.js, Deepgram, Groq, Supabase, Docker* **Feb 2025**

- Built a full-stack conversational AI system achieving less than 300ms average end-to-end latency with real-time lip-synced avatar interaction.
- Designed a high-throughput streaming pipeline using LiveKit Agents, enabling more than 95% real-time structured data extraction via Deepgram Nova-2 and ultra-fast LLM inference with Groq.
- Deployed on a scalable serverless architecture (Railway, Vercel) with CI/CD pipelines enabling zero-downtime releases and reducing manual deployment effort by more than 90%.

Signalyze | *LangGraph, FastAPI, Groq, Next.js, SQLite, DuckDuckGo Search* **Mar 2026**

- Built an AI account-intelligence platform that converted 3 input modes into structured outputs for company profiling, tech-stack discovery, leadership mapping, buyer-intent scoring, and recommended actions.
- Designed a 7-stage orchestration pipeline with conditional routing and 5 parallel analysis modules, combining top-3 web-result grounding, IP-based company identification, and website enrichment in a single end-to-end workflow.
- Engineered a scoring and enrichment layer using 22 behavioral page signals, 28 HTML technology signatures, and a 24-hour SQLite caching/history system with concurrent batch execution for repeated account analysis.

Pulse: Empathetic LLM Fine-Tuning | *PyTorch, Qwen 1.5B, QLoRA, SFT, Multi-Task Learning* **Jan 2026**

- Implemented a multi-objective supervised fine-tuning (SFT) pipeline on Qwen 1.5B using QLoRA with auxiliary emotion and support-strategy classification heads.
- Constructed a temperature-sampled ($\alpha = 0.7$) multi-task dataset mixture (GoEmotions, ESConv), achieving stable convergence (loss: 3.13) over 500 steps on NVIDIA T4.
- Designed a composite loss with KL-divergence safety regularization and two-step decoding, achieving 100/100 on EQ-Bench 3 evaluation.

Technical Skills

LLM & Agentic Systems: LangGraph, LangChain, CrewAI, RAG, MCP, LLM Evaluation, Groq
Voice & Real-Time AI: ElevenLabs, Twilio, LiveKit, Deepgram, WebRTC
Computer Vision & Deep Learning: PyTorch, TensorFlow, Scikit-learn, Transformers, OpenCV, Image Processing, CNN, QLoRA
Backend & Databases: FastAPI, PostgreSQL, MongoDB, Redis, Prisma, Node.js, React, Supabase
Cloud & Infrastructure: AWS Lambda, API Gateway, SQS, DynamoDB, Docker, Celery, App Runner, ELK Stack, Keycloak, HashiCorp Vault
Vector Stores: FAISS, Weaviate, ChromaDB

Leadership / Extracurricular

- Top 15/57,000+ in Google Agentic AI Hackathon; finalist at Guinness World Record event (650+ pro teams).
- 15th/12,000+ teams in Puch AI Hackathon; built MCP server onboarding 350+ users in 24 hours.
- Top 10/950+ in Kaggle competition; Best Capstone Project Award (ML course).