Sudeep Aryan Gaddameedi

+353 89 961 3030 | sudeeparyang@gmail.com | Portfolio | LinkedIn | GitHub |
Dublin, D01PC92

Summary

AI Engineer and tech polymath with 3 years of experience in Generative AI, ML model development, and cloud deployment. Skilled in Python, RAG systems, and MLOps, with publications and awards in applied AI research. Expert in building scalable RAG systems and autonomous agents, delivering high-impact solutions that reduced operational costs by 50% and secured $1M in contracts Passionate about building scalable AI solutions, mentoring teams, and applying cross-disciplinary knowledge to solve complex problems.

Education

National College of Ireland — Dublin, Ireland
MSc in Artificial Intelligence — *2025 – 2026*

Experience

National College of Ireland — Sep 2025 – Present

Research Support Supervised by Prof. Mohit Garg and Dr. Arghir Nicolae Moldovan

- AI for Autonomous Vehicles (under Mohit Garg): Integrating Transformer-based architectures into self-driving systems using the A2A protocol for V2V and V2I communication. Simulated real-time scenarios in SUMO with offline SLMs and MCP tools to enable predictive decision-making and intelligent vehicle interaction.
- AI in Cybersecurity (under Arghir Nicolae Moldovan): Building an intelligent vulnerability detection framework by combining CodeBERT and Transformer models with traditional ML and rule-based methods to proactively identify system threats on the Diveservul platform.

Soliton Technologies — Jan 2023 – Aug 2025

Senior Project Engineer

- Strategic Leadership & Revenue Growth : Secured a **$1M contract** by leading a 3-member engineering team to rapidly architect and deliver an AI-enabled Copilot POC, directing technical roadmapping and cross-functional execution to exceed client expectations.
- System Architecture & Optimization**:** Engineered a scalable, event-driven AI test generation system that reduced generation time by **94% (60 mins to 4 mins)** and cut operational costs by **50%** through concurrent multi-stage requests, predictive output modeling, and advanced prompt caching.
- MLOps & Developer Efficiency**:** Increased developer productivity by **80%** by architecting an automated CI/CD and MLOps ecosystem—utilizing DsPy, that parallelized RAG evaluations and streamlined prompt optimization.

Project Engineer

Agent Vina – Soliton AI Assistant

- Developed an autonomous support agent automating complex workflows across HR and IT (e.g., skill mapping, travel booking); reduced support ticket volume by **35%** and integrated seamlessly with internal APIs for 300+ employees.

Client Project: Texas Instruments – DataSheet Validator (POC)

- High-Performance Validation System (Texas Instruments): Architected the "DataSheet Validator" platform to convert massive test procedures into executable programs. Deployed via **Docker & Kubernetes**, achieving a **96% latency reduction (300s to 10s)** through distributed caching and concurrent processing.
- Cloud-Scale Evaluation Pipelines**:** Built robust monitoring workflows to track retrieval accuracy (RAG) and generation quality over time, ensuring continuous model improvement and reducing overall validation timelines by **80%**.

Internship

Client Project : Intel – Device Vision Platform (POC)

- Predictive Analytics & Tooling: Developed an ML-powered GUI tool using regression models to predict device performance; reduced manual validation effort by **90%** and eliminated domain-specific dependencies for validation engineers.
- Modern Data Engineering (Azure Fabric): Led the migration of legacy data pipelines to **Microsoft Azure Fabric** (OneLake, Data Factory), implementing automated ETL processes that enabled real-time analytics and proactive alerting.

Projects

Serverless Accessibility Auditor (SaaS) Present

- **System Architecture**: Architected a scalable **Node.js serverless backend** on AWS (Lambda, API Gateway, SQS) to automate WCAG 2.1 compliance auditing for high-volume concurrent requests.
- **Hybrid Analysis & Outcome**: Built a dual-layer pipeline combining **Axe-core** static checks and **GPT-4o** semantic analysis to detect complex violations, reducing false positives by ∼40%.
- **Distributed Scraping**: Engineered a fault-tolerant scraping system using **Puppeteer** on AWS Lambda to render dynamic JS-heavy content with automatic retry logic and proxy rotation.
- **Tech Stack**: Node.js, TypeScript, AWS Lambda, DynamoDB, Puppeteer, OpenAI API, Serverless Framework.

Technical Skills

Artificial Intelligence: Generative AI (LLMs, SLMs, VLMs), RAG & GraphRAG, Agentic Workflows, Prompt Engineering, Fine-tuning (LoRA, QLoRA), Natural Language Processing (NLP), Semantic Search
Agentic Orchestration : Model Context Protocol (MCP), Agent-to-Agent (A2A), Google ADK, LangChain, LlamaIndex
Machine Learning: CNNs, Transformers, PyTorch, TensorFlow, Scikit-learn
AI Ops &Evaluation : MLflow, Arize Phoenix, Trulens, Ragas, Dspy, PromptFlow
Programming Languages: Python, JavaScript, TypeScript ,C
Cloud Platforms: (Azure OpenAI, AI Search, Cognitive Services, VMs), AWS (EC2, S3, Lambda)
Data & Knowledge Engineering : SQL (PostgreSQL, SQLite), NoSQL (MongoDB), Knowledge Graphs (Neo4j), Vector DBs (Qdrant, Weaviate, Pinecone, Milvus), Microsoft Fabric
DevOps Tools: Docker, Kubernetes, CI/CD (GitHub Actions, Azure Pipelines), LGTM Stack (Grafana, Loki, Tempo, Prometheus)
Web Technologies: Next.js, React, Angular, TypeScript, Tailwind CSS, HTML, Node.js, Polymer
Developer Tools: VS Code, Github Copilot, N8N, Firebase Studio, Vercel, Google AI Studio, Postman, Microsoft Copilot Studio

Publications

RAG HUB — A Comprehensive Guide to Build an AI Architecture (Best Paper Award)
- Proposed a question-driven framework for simplifying AI architecture design using dynamic filtering and decision-tree logic. The approach reduced development complexity and improved scalability for enterprise AI systems.

CNN-Based Curved Path Detection for Autonomous Rover
- Designed an autonomous rover using Raspberry Pi 3B+ and reinforcement learning for real-time lane-following and obstacle avoidance. Integrated CNN models, Pi Camera, and ultrasonic sensors for adaptive navigation.

GRASP: A Scalable RAG-Enhanced Multi-LLM Architecture for Intention-Aware Autonomous Vehicle Decision-Making (Under Final Review)
- Research focuses on integrating multi-agent learning with communication-aware control to enhance collaborative decision-making in connected autonomous vehicles.

Comparing the Capabilities of Ensemble Learning Algorithms and SAST Tools for Effective Code-Based Vulnerability Detection (Under Final Review,)
- Evaluates the performance of ensemble learning models against static analysis tools for early vulnerability detection. Aims to build a hybrid AI-assisted security framework for code-level risk identification.

Achievements

Best Paper Award (2025) – Received at the IEEE International AI Conference, recognized among 1000+ global submissions including PHd for introducing the question-driven RAG HUB framework that streamlines AI development and reduces the learning curve for developers.
Innovator of the Year – Soliton – Honored for pioneering cutting-edge AI techniques, mentoring junior engineers, and driving transformative innovations across multiple projects.
Google Cloud Agentic AI Day 2025 Hackathon – Shortlisted from 57,000+ participants for developing AI Money, a privacy-first personal finance agent with real-time financial simulations and family-focused planning.
Breaking Barriers for Agentic Networks – AWS (2025) – Shortlisted and invited to the Amazon Office in Dublin for presenting research on agentic intelligence frameworks and autonomous collaboration in distributed AI systems.

Certifications

Microsoft Azure
Neural Networks and Deep Learning
AWS
Machine Learning
Full Stack Web Development
Google Professional Workspace Administrator