Siddhant Pathak

siddhantpathak.applications@gmail.com | linkedin.com/in/siddhantpathakk | github.com/siddhantpathakk | Singapore

PROFESSIONAL SUMMARY

AI Engineer with production experience building LLM-powered systems, AI agents, and secure API infrastructure for enterprise applications. Designed a centralised LLM orchestration gateway integrating 30+ models with failover, governance, and streaming support. Re-architected a recommendation system to serverless with ~90% cost reduction and hardened AI service boundaries with least-privilege API governance. Experienced in building production AI systems with strong cost optimization, latency reduction, and reliability focus.

CORE COMPETENCIES

Production AI Systems | LLM Orchestration & Agents | Financial AI Workflows | RAG & Information Retrieval | API Governance & Security | Cost & Latency Optimization | Python / PyTorch / AWS | CI/CD & MLOps

WORK EXPERIENCE

Singapore Airlines

Aug 2024 - Present

Data Scientist

- Designed centralised LLM gateway integrating 30+ frontier and open-source models via LiteLLM proxy with streaming, large-payload, and session support; 10x improvement in average response time
- Re-architected Flight Recommender from SageMaker to serverless Lambda + GPT-4.1 with structured tool-calling; ~90% infra cost reduction, latency 6s to 4s, 70% fewer business-logic errors
- Hardened AI service boundaries with reverse proxy, least-privilege API key governance, model-provider failover; reduced sev1/2 incidents across production AI systems
- Engineered production-grade RAG pipeline for B2B vendor smart search with real-time document ingestion, dense vector retrieval, and re-ranking; reduced query latency from 40s+ to under 10s
- Built end-to-end deployment pipelines for open-source LLMs on AWS GPU instances and on-prem clusters, enabling private air-gapped inference
- Embedded GenAI across SDLC -- LLM-assisted coding, automated test generation, intelligent deployment pipelines -- cutting dev cycle time by 20%

Asurion

Jan - Apr 2024

Data Science & GenAI Engineer Intern

- Built ML predictive models for supply chain optimisation and customer segmentation driving business automation
- Engineered GPT-4 + LangChain chatbot with real-time information retrieval for customer-facing workflows

Glance, InMobi

May - Aug 2023

SDE Intern

- Built CI/CD pipeline for ML workflows using Scikit-learn and TensorFlow
- Developed federated learning server using Flower framework for privacy-preserving model training

Panasonic R&D

May - Aug 2022

AI Algorithm Engineer Intern

- Built real-time monitoring system with Flask, PaddleOCR, and Tesseract-OCR
- Optimised edge inference pipeline, reducing latency by 30%

MSD

Aug - Dec 2022

Data Analytics Consultant Apprentice

- Led team of 4, built analytics tool integrating clinical data sources with Power BI dashboards
- Developed Python ETL + AWS Lambda + S3 data pipeline for automated reporting

PROJECTS

Icarus AI -- GenAI Job Recommendation

3rd Place, Deep Learning Week

Built AI agent-driven recommendation engine using LLMs for semantic matching with automated workflow orchestration and cost-optimized inference.

LLM, AI Agents, RAG, Python

TAGON -- GNN Sequential Recommendation

President Research Scholar

Developed graph neural network-based recommendation system achieving 20% improvement over baselines. Applied production ML optimization techniques for model serving.

PyTorch, GNN, Recommendation, Production ML

Expected Goal Model for SG Football

Built predictive ML model achieving 99.58% accuracy on 1M-point dataset. Applied rigorous data pipeline engineering and model validation for production-grade predictions.

Python, Scikit-learn, Data Pipeline, ML

EDUCATION

BEng Computer Science (Data Science & AI) -- NTU Singapore 2020 - 2024

CGPA 4.27/5.0, Honours (Distinction)

SKILLS

GenAI & LLMs: LiteLLM, vLLM, SGLang, MLX, LangChain, LangGraph, Google ADK, RAG, Structured Tool-Calling

ML & DL: PyTorch, TensorFlow, Scikit-learn, GNNs, Flower **MLOps:** CI/CD, DataDog, Splunk, Prometheus, GitHub Actions

Cloud: AWS (Lambda, SageMaker, EC2, S3, ELB), GCP, On-Prem GPU **Languages:** Python, SQL, R, C++, Java, Bash