SIDDHARTH RAVIKUMAR
AI PRODUCT SPECIALIST | Full-Stack AI Architect

Email: siddharth.ravikumar4521@gmail.com | **Mobile:** +971503658915 |
LinkedIn: https://www.linkedin.com/in/siddharth-ravikumar-17262a50/ | **Address:** Abu Dhabi, UAE
Github: https://github.com/SiddharthRavikumar1989

PROFESSIONAL SUMMARY

AI Product Specialist and Full-Stack Data Scientist with 14+ years of experience delivering AI-driven solutions across NLP, Generative AI, Statistical Modeling, and Geospatial Analysis. Strong people and product leadership experience, including owning AI roadmaps, leading cross-functional teams, and managing end-to-end model lifecycles from concept to production. Proven ability to align AI initiatives with business strategy, translate complex requirements into scalable, production-grade systems, and deliver measurable business impact.

CORE SKILLS

Languages: Python, Java, SQL

AI/ML: PyTorch, TensorFlow, scikit-learn, Hugging Face Transformers, PyTorch Lightning, Whisper

AI Specialization: NLP, Generative AI, Quantitative Analysis, Automatic Speech Recognition

LLMs & GenAI: Fine-tuning (LoRA/QLoRA), Agentic RAGs, Multi-agent Systems, Open-source LLMs, Prompt Engineering, Langgraph, Langchain, Unsloth

Data & Analytics: PySpark, Pandas, NumPy, Dask, DuckDB, Apache Spark, Apache Kafka

MLOps & Deployment: MLflow, Airflow, Docker, Kubernetes, CI/CD Pipelines

Architecture & Design: Distributed Systems, Microservices, Event-Driven Architecture, Data Pipeline Design, Plugin/Extensibility Patterns, API Design

Software Engineering: Python (Advanced), FastAPI, Django

Cloud Platforms: GCP, AWS

PROFESSIONAL EXPERIENCE

AI Product Specialist | SAAL AI January 2025 – Present | Abu Dhabi

Domain Specialized Knowledge-Aware RAG Chatbot

- Designed and implemented scalable Multimodal RAG pipelines using vector databases, embedding models, and prompt orchestration to power knowledge-aware chatbots and virtual assistants for complex Structural Engineering Documents.
- Evolved the RAG architecture into an Agentic RAG framework by introducing task-aware agents, dynamic retrieval strategies, and reasoning loops, enabling deeper document understanding, multi-step reasoning, and higher answer accuracy for complex engineering use cases.
- Designed and implemented enterprise ingestion & ETL pipelines (Airflow, cloud-native orchestration) for document chunking, metadata enrichment, embedding generation, and continuous knowledge-base updates.
- Built reusable standardized RAG/agentic architecture, defining retrieval schemas, memory strategies, hybrid search routing, evaluation frameworks, and company-wide best practices for production AI systems.
- Engineered advanced OCR pipelines for RAG ingestion, converting structured/unstructured documents (Engineering diagrams and Merged tables) into high-fidelity, chunkable representations optimized for retrieval.

Speech Intelligence & Multilingual ASR Systems

- Architected and deployed speech intelligence systems including audio enhancement, ASR, and speaker diarization for Arabic and Persian languages.
- Adopted state-of-the-art noise removal strategies (bandwidth filters, layered noise reduction) and speech enhancement for clear transcription.
- Fine-tuned Whisper and NVIDIA Conformer Models with military jargon-specific voice samples.

Multi-Agent Quantitative Research & Trading Intelligence Platform

- Designed and built a multi-agent AI system acting as a quantitative analyst, orchestrating News, Fundamental, Technical, and Algo-Trading agents to generate unified investment insights with self-correction and validation loops.
- Implemented an agentic orchestrator (LangGraph) to coordinate analysis, aggregate agent reports, apply feedback loops, and produce final trade recommendations.
- Developed an Algo Trading Agent leveraging quantitative strategies including Modern Portfolio Theory (MPT) and rule-based trading algorithms.
- Built a multi-turn conversational interface with Text-to-SQL capabilities for querying trade history, positions, and performance metrics from structured databases.
- Integrated real-time and near-real-time market data from sources such as Yahoo Finance and Financial Modeling Prep (FMP).
- Led experimentation with advanced AI architectures for edge deployment in extreme warfare settings, incorporating noise-resilient sensing and symbolic reasoning to improve threat assessment accuracy.

Automatic Defense Action Trigger with World Foundation Models

- Engineered world foundation models tailored to defense autonomy applications — focused on multimodal training across challenging physical conditions to support reliable detection and adaptive action execution.

AI Architect | Ministry of Rural Development June 2023 – January 2025 | Riyadh, Saudi Arabia

Urban Violation Hotspot Detection & Forecasting Platform

- Built and analyzed geospatial datasets using GeoPandas, Shapely, and PostGIS, enabling spatial joins, buffering, and distance-based analytics on large city-scale GIS data.
- Led development of a violation hotspot detection platform using density-based clustering algorithms (DBSCAN/HDBSCAN) to identify high-risk zones from GPS-tagged enforcement data.
- Designed a spatial-temporal analytics model combining time-series analysis, spatial lag features, and historical trend decomposition to track violation patterns over time and forecast emerging hotspots for proactive enforcement planning.

AI-Powered Municipal Compliance & Policy Intelligence Platform

- Built an LLM-powered, multi-channel chatbot integrated with municipal backend systems to automate document queries, compliance checks, and regulatory guidance.
- Implemented policy-aware response generation using embeddings, topic modeling, and LangChain, producing accurate, context-grounded guidance aligned with city regulations.
- Automated policy checklist generation and compliance validation, significantly improving accuracy and reducing manual review effort.
- Developed a curated policy knowledge pipeline for ingesting, normalizing, and maintaining multi-source regulatory content used by city development authorities.

Senior Data Scientist | Onit September 2022 – June 2023 | Remote

AI-Driven Legal Intelligence & Contract Analytics Platform

- Designed and implemented LLM-powered pipelines for legal clause extraction, clause generation, and semantic validation to automate contract workflow analysis.
- Fine-tuned domain-adapted legal LLMs using supervised instruction tuning and contrastive examples, significantly improving clause interpretation accuracy and reducing manual legal review effort by approximately 40%.
- Built a content-aware lawyer recommendation engine leveraging embeddings, semantic similarity search, and metadata ranking (jurisdiction, specialization, precedent relevance) to match cases with appropriate legal experts.
- Achieved approximately 95% clause validation reliability through systematic evaluation, including test-set stratification by clause type, adversarial examples, confidence scoring, and continuous model optimization.

Consultant Data Scientist | Verizon November 2018 – September 2022 | Chennai, India

- Designed and deployed network anomaly detection and churn prediction systems using ensemble learning techniques (Random Forests, Gradient Boosting) and sequence models (RNNs, LSTMs) to capture temporal

usage patterns and early risk signals.

- Built domain-specific NLP pipelines for large-scale conversation analysis, combining BM25-based relevance scoring, association rule mining, and contextual embeddings to extract keyphrases, identify recurring themes, and surface high-signal interactions.
- Implemented multi-level language understanding including sentiment analysis, intent classification, and aspect-based sentiment modeling, enabling fine-grained customer experience insights.

Machine Learning Engineer | UST June 2017 – December 2018

- Designed and implemented an edge analytics infrastructure for real-time IoT datacenter monitoring and anomaly detection.
- Built a distributed data pipeline using AWS services (Spark Streaming, Kinesis, DynamoDB, Elasticsearch, S3, Athena) to improve scalability and reduce operational costs.
- Applied Bayesian inference and deep learning models (RNN, LSTM) for sequence-based anomaly detection.

PERSONAL AI PRODUCTS

Tiny Tamil LLM SaaS (Personal AI Product) 202X – Present

- Built and launched a compact, instruction-tuned LLM fine-tuned on classical Tamil literature; deployed as a tiny SaaS for cultural language generation and Q&A.;
- Tech: PyTorch, Hugging Face, Streamlit.

On-Device AI Image Search App (Android) 202X

- Developed multimodal pipeline with multi-object detection, bounding-box filtering, and complex visual reasoning; packaged as efficient mobile app for in-phone search.
- Tech: TensorFlow Lite / PyTorch Mobile, Android SDK; focused on low-latency edge AI.

RESEARCH & PROOFS OF CONCEPT

Liquid Foundation Models vs. LLMs Evaluation

- Conducted comparative analysis of liquid neural networks and transformer-based LLMs, evaluating reasoning stability, data efficiency, latency, and GPU utilization in low-data and streaming environments.

GPU Architecture & Memory Utilization Benchmarking

- Evaluated DGX Spark GPUs vs. standard NVIDIA A100/H100 setups, benchmarking VRAM utilization, memory bandwidth, throughput, and cost-performance trade-offs across LLM inference, fine-tuning, and multimodal workloads.

EDUCATION

Bachelor of Engineering in Computer Science

Anna University, Chennai 2007–2011