

JAGAN SIVAKUMARAN

AI Engineer | Generative AI, LLM & Applied ML Systems

jagansivakumaran@gmail.com | 9499926325 | [LinkedIn](#) | [Portfolio](#)

PROFESSIONAL SUMMARY

AI Engineer with 1.5+ years of experience designing and deploying production-grade enterprise AI systems across Generative AI, LLMs, NLP, and Computer Vision. Built agentic architectures on AWS Strands, LangChain/LangGraph, and Amazon Bedrock that reached 95%+ orchestration accuracy, 98% NLP-to-SQL accuracy, and 100% RAG accuracy at sub-1200ms latency for 50 concurrent users. Experienced across the full AI lifecycle: data preparation, fine-tuning (SFT, RLHF), RAG/agent orchestration, API development, and cloud deployment, including a GPU-to-AWS Trainium/Inferentia migration, CV-based defect detection (96% accuracy), and real-time translation. Led teams of 10+ engineers, completed 40+ code reviews, and shipped 100+ production APIs on schedule.

TECHNICAL SKILLS

AI / Generative AI: Deep Learning, Machine Learning, NLP, LLMs, Generative AI, Computer Vision, Fine-Tuning, SFT, RLHF, Model Evaluation, Prompt Engineering, Embeddings, RAG, NLP-to-SQL, Real-Time Inference

LLM / Agentic AI: LangChain, LangGraph, AWS Strands, Pydantic, LangSmith, AWS AgentCore, AgentCore Memory, Agent Orchestration, Tool Calling

AWS & Cloud: Bedrock, SageMaker AI, Knowledge Bases, OpenSearch, Trainium, Inferentia, Neuron, EC2, S3, SQS, Lambda, EKS, ECS, EventBridge

Data & Backend: PostgreSQL, MySQL, MongoDB, Redis, SQL, Vector Databases, Python, FastAPI, REST APIs, WebSockets

Frontend & DevOps: React, JavaScript, TypeScript, HTML/CSS, Docker, Kubernetes, CI/CD, Kubeflow

Databricks & Tools: Model Registry, Model Serving, Notebooks, Postman, API Testing, Code Reviews

PROFESSIONAL EXPERIENCE

AI Engineer

Enterprise Generative AI, LLM & Applied ML Systems

- Designed agentic AI architectures using AWS Strands, LangChain, and LangGraph, integrating RAG, NLP-to-SQL, and computer-vision pipelines into production apps for conversational analytics, audio generation, compliance inspection, translation, and manufacturing QC.
- Built 100+ production REST APIs with Python and FastAPI to support AI agent orchestration and model inference, and owned these systems end-to-end, from data preparation and fine-tuning through cloud deployment and optimization.
- Led teams of 10+ engineers and completed 40+ code reviews, shaping architecture decisions and keeping delivery on schedule across AI initiatives.

AI PROJECTS

RedeAPP: Agentic Conversational Analytics | AWS Strands, Bedrock, RAG, NLP-to-SQL, AgentCore

- Architected an agentic system on AWS Strands and Amazon Bedrock letting users query historical PostgreSQL messaging/scheduling data via natural language, routing each request to a RAG tool (Titan V1 embeddings, AWS Knowledge Bases vector store, refreshed on a rolling two-day cycle) or an NLP-to-SQL tool that converts date/structured queries into executable SQL.
- Added short-term and long-term AWS AgentCore Memory on AgentCore Runtime, then tightened prompts and cut redundant LLM calls to reach 95%+ orchestration, 98% NLP-to-SQL, and 100% RAG accuracy at sub-1200ms latency for 50 concurrent users.

SoundryAI: GPU-to-Trainium/Inferentia Migration | Trn2.48xlarge, Inf2.24xlarge, AWS Neuron

- Migrated a multi-model audio-generation training pipeline from GPU to AWS Trainium on a Trn2.48xlarge instance using tensor parallelism across 8 NeuronCores, resolving OOM and gradient-explosion issues via workload tuning and gradient clipping, across a pipeline using HTDemucs for source separation, an RVQ tokenizer to compress ~48kHz audio into codebooks, and FlowVAE decoding feeding an autoregressive audio-token generation model.
- Deployed inference on AWS Inferentia Inf2.24xlarge with 8-core tensor parallelism on Amazon EKS, generating a 3-minute song in roughly 2 minutes, and built training/inference UIs plus an event-driven PostgreSQL notification queue for concurrent requests.

Solar Installation Compliance & QC Platform | RF-DETR, OCR, LangChain, LangGraph, RAG

- Trained an RF-DETR computer-vision model with integrated OCR to detect installation defects and compliance issues in solar photos, automating a QC process that previously took hours to days.
- Built an agentic workflow in LangChain and LangGraph with a RAG tool retrieving NEC and state compliance rules to explain each pass/fail result, escalating low-confidence cases to a human reviewer with retake instructions and automated lender-folder organization across Dev, UAT, and Production on AWS.

Real-Time Multilingual Translation Platform | Translate Gemma 27B, Bedrock Sonnet, RLHF, vLLM, Redis, EKS

- Benchmarked translation models on NVIDIA B200 GPUs across 76 languages, selected Translate Gemma 27B, and built a routing architecture, a ~1.5TB Redis cache feeding language/model matching that falls back to Gemma or Amazon Bedrock Sonnet, for low-latency serving.
- Implemented an RLHF pipeline using Claude Opus 4.5-generated training data, a custom reward function, and vLLM-optimized, weight-parallelized training to raise accuracy to roughly 95%, then deployed to Amazon EKS for real-time inference at ~1500ms P90 latency against an original sub-1200ms P90 target.

AI-Powered PCB Defect Detection & QC System | Computer Vision, SFT, Pydantic, Vector DB

- Curated a training dataset from external sources, client-provided data, and augmentation, then fine-tuned a computer-vision model via SFT to detect 10 defect classes across a line producing roughly 20,000 boards a day on 10 production lines, optimizing batch size, epoch count, and training strategy to raise accuracy to roughly 96%.
- Built a Pydantic-based agent that performs vector retrieval against a PostgreSQL vector database of known defects to generate operator recommendations, cutting manual inspection effort.

CERTIFICATIONS & EDUCATION

Databricks Certified Machine Learning Professional | Issued Jan 2026 | Expires Jan 2028

Saveetha Engineering College | B.E., Computer Science | 2021-2025 | Chennai, India