

SHIVAM CHAND KAUSHIK

Agentic AI Engineer | Multi-Agent Systems & GenAI | LangGraph · Python · FastAPI

Bangalore, India · +91-8310644891 · shivamiitj9@gmail.com · LinkedIn · GitHub

Professional Summary

Agentic AI engineer with 9 years of semiconductor systems software experience (C++, SECS/GEM, smart factory), transitioning into **multi-agent LLM orchestration, RAG pipelines, and production-grade AI automation**. **M.Tech AI, IIT Jodhpur**. Published on arXiv (cs.CV, **2604.13236**) — work on autonomous semiconductor Failure Analysis **highlighted by Andre Zayarni, Co-founder of Qdrant**, validating real-world applicability at chip manufacturing scale. Built four-agent **LangGraph** pipelines (LLaVA-1.6, DINOv2, Qdrant, FastAPI) compressing several hours of expert FA work into **48-second automated reports**.

Core Competencies

Agentic AI & GenAI	ML & Computer Vision	Software & Cloud
Multi-Agent Orchestration (LangGraph)	PyTorch, TensorFlow, torchvision	Python (2 years)
LangChain, RAG, Vector Databases	CNNs, ViTs, DINOv2	C++ (9 years production)
LLaVA, Qwen-VL, QLoRA/PEFT	XGBoost, scikit-learn, SHAP	FastAPI
Autonomous Workflow Automation	Computer Vision, Anomaly Detection	Git/CI-CD
Qdrant, FAISS, ChromaDB	Streamlit, ReportLab (AI dashboards)	AWS (Bedrock, ECS, ECR, S3)

Work Experience

Software Analyst — Semiconductor Assembly & Smart Factory AI

Kulicke & Soffa (K&S), Netherlands & Singapore

Sept 2020 — Oct 2025

via ASM Technologies / K&S is a global leader in semiconductor assembly equipment for advanced packaging.

- Built three **production agentic AI systems** grounded in real K&S equipment data: (1) an **LLM-powered SECS/GEM Protocol Assistant** (RAG, FAISS) enabling natural-language queries over SEMI standards, alarm codes, and equipment specs; (2) an **ML Wire Bond Quality Prediction** agent with Bayesian process optimization on RapidPro/ATPP sensor parameters—reducing predicted defect rate by **40%**; (3) a **RAG-based Customer Support agent** over K&S product documentation (ChromaDB, sentence-transformer embeddings).
- Designed the **Smart Factory AI data backbone**: integrated **MQTT** into the Machine Configuration Tool (MCT) for live machine-to-cloud streaming of high-frequency bond-process telemetry across 6 machine variants (iFlex-T2/T4/H1/CX, Luminex, Pixalux)—providing the real-time sensor feed consumed by AI inference and production analytics pipelines.
- Engineered the **C++ real-time control and data-acquisition layer** (RapidPro, ATPP wire bonding machines) that generates ground-truth process telemetry (bond force, ultrasonic power, temperature, loop height) powering ML defect prediction and Bayesian parameter optimization.
- Hardened the **observability and reliability** of the AI data pipeline through compiler warning upgrades (W3→W4), PVS-Studio static analysis, IWYU, and Boost→STL migration—ensuring data integrity for downstream ML workloads.

Senior Software Engineer — IC Inspection & Machine Vision

Quest Global Technologies (client: KLA/ICOS)

July 2018 — Sept 2020

Bangalore / KLA is the world's leading semiconductor process control and inspection equipment company.

- Built a Python-based hardware simulator for the ICOS T800 Gantry System—a TCP/IP server modeling machine behavior for offline testing and digital-twin development.

- Developed data logging and analytics pipelines for production metrics: throughput tracking, taper parameter monitoring, and per-head performance analysis.
- Implemented **SECS/GEM interface for Intel**—the semiconductor industry standard protocol for equipment-to-host communication in automated fabs.
- Engineered granular process control: independent per-head yield-loss diagnostics for YZ1/YZ2/YZ4 pick & place heads.

Software Engineer — Aerospace & Defense Systems

AxisCades Aerospace & Technology

May 2016 — June 2018

Bangalore

- Developed real-time 2D signal visualization for a Bird Detection Radar system using C++/Qt.
- **Led a team of 5 engineers** and commissioned ACRT across **37 units of the Indian Army and Indian Air Force**.

AI/ML Projects

1. SemiFA: An Agentic Multi-Modal Framework for Autonomous Semiconductor Failure Analysis Report Generation [GitHub](#) | [HuggingFace Dataset](#) | [arXiv:2604.13236](#)

- Built a four-agent LangGraph pipeline (Defect Descriptor, Root Cause Analyzer, Severity Classifier, Recipe Advisor) that fuses DINOv2 visual embeddings, SECS/GEM-format equipment telemetry, and a Qdrant historical defect retrieval database into a unified LLaVA-1.6 context—generating structured Failure Analysis reports in 48 seconds, replacing a 2–4 hour manual process.
- Trained a DINOv2 + MLP defect classifier on SemiFA-930, a self-constructed 930-image dataset spanning 9 semiconductor defect classes (SEM, optical, wafer map); achieved 92.1% accuracy / 0.917 Macro F1 with only 214K trainable parameters vs. 23.5M for ResNet-50 (82.9%).
- Publicly released SemiFA-930 dataset on HuggingFace; paper published on arXiv (cs.CV) — [arXiv:2604.13236](#).
- **Stack:** LLaVA-1.6, DINOv2, QLoRA/PEFT, LangGraph, SECS/GEM, Qdrant, FastAPI, ReportLab, Streamlit
- **Industry recognition:** Research featured by **Andre Zayarni, Co-founder of Qdrant**, on LinkedIn (2,120 impressions, 54 reactions, 7 reposts) — publicly validating the system’s industrial applicability for autonomous semiconductor FA.

2. Wafer Defect Detection & Classification (WM-811K) [GitHub](#)

- Built an end-to-end deep learning pipeline for wafer defect pattern classification on the WM-811K dataset (811K+ wafer maps, 9 defect classes) using ResNet18 and EfficientNet-B0, with a ConvAutoEncoder for anomaly detection on unseen patterns. Achieved 97.83% overall test accuracy and 91.95% defect-class accuracy on a severely class-imbalanced dataset (103,201 normal vs 104 Near-Full samples). Improved Scratch F1 from 0.48 → 0.820 by moving from 64×64 to 128×128 resolution, adding FocalLoss($\gamma=2$) and synthetic scratch augmentation; diagnosed critical training bugs across normalization, BatchNorm domain shift, loss double-suppression, and `torch.compile` checkpoint corruption.
- Trained a two-stage classifier (binary Stage 1 → 8-class Stage 2) for comparison: Stage 1 achieved 98.45% binary accuracy but error propagation reduced combined accuracy to 96.74%, confirming single-stage as the correct production model. Applied 8-view Test-Time Augmentation (+0.33% overall, stabilises low-support classes) and built a v2+v3 ensemble classifier with per-class confidence thresholding and human-review routing for safety-critical predictions.
- **Stack:** PyTorch, torchvision, scikit-learn, FastAPI

3. LLM-Powered SECS/GEM Protocol Assistant (RAG) (*Applied at Kulicke & Soffa*)

- Built a RAG-powered LLM assistant for the SECS/GEM semiconductor protocol, enabling natural language queries over SEMI standards, equipment specifications, and alarm codes. Fine-tuned retrieval on SECS/GEM documentation with FAISS vector store.
- RAG pipeline indexes SECS/GEM standards, equipment manuals, and alarm code definitions into FAISS; retrieves top- k relevant chunks at query time to ground LLM responses in accurate, protocol-specific context.

- Deployed as an interactive web application for equipment and process engineers in semiconductor fabs.
- **Stack:** HuggingFace Transformers, LangChain, RAG, FAISS, Streamlit

4. Customer Support Chatbot for K&S Equipment Clients (RAG + ChromaDB) *(Applied at Kulicke & Soffa)*

- Built a RAG-based customer support chatbot for K&S semiconductor equipment clients, enabling engineers to query machine manuals, troubleshooting guides, spare-parts catalogues, and service bulletins in natural language—reducing average ticket resolution time.
- RAG pipeline ingests and chunks K&S product documentation (RapidPro, ATPP, iFlex series) into **ChromaDB** vector store with sentence-transformer embeddings; retrieves top-*k* semantically relevant passages to augment LLM context and prevent hallucination on equipment-specific queries.
- Implemented conversation memory and session history to support multi-turn diagnostic dialogues; added confidence-score filtering to route low-certainty queries to human support agents.
- **Stack:** LangChain, RAG, ChromaDB, Sentence-Transformers, HuggingFace Transformers, FastAPI, Streamlit

5. Wire Bond Quality Prediction & Process Optimization *(Applied at Kulicke & Soffa)*

- Developed ML-based wire bond quality prediction using process data (bond force, ultrasonic power, temperature, loop height) modeled on K&S RapidPro and ATPP machine parameters. Classified bond pass/fail with root cause analysis and applied Bayesian optimization to identify optimal bonding parameters, reducing predicted defect rate by 40%.
- **Stack:** scikit-learn, Optuna, matplotlib

6. E-Commerce RAG Customer Support Chatbot — Production-Deployed on AWS

- Built a production-ready e-commerce customer service chatbot using RAG, answering natural-language queries over company policy documents (returns, shipping, warranty) with source-attributed, citation-grounded responses and zero hallucination on out-of-scope questions.
- Implemented a three-mode retrieval engine: **AWS Bedrock Knowledge Base** (vector similarity), **BM25** keyword search, and a **hybrid mode** that alpha-weights and normalises scores from both—configurable at runtime via the Streamlit UI sidebar.
- Built conversation memory preserving 6-turn history and a regex-based user context extractor (name, order number, email) that personalises responses across the session; conversations exportable to JSON for audit trails.
- Deployed end-to-end on **AWS**: Docker image stored in **ECR**; Streamlit app served via **ECS Fargate** (serverless containers); knowledge base documents in **S3**; secrets managed in **SSM Parameter Store**; logs streamed to **CloudWatch**; IAM roles scoped to least-privilege Bedrock and S3 access.
- **Stack:** Python, LangChain, AWS Bedrock, ECS Fargate, ECR, S3, SSM Parameter Store, CloudWatch, BM25, boto3, Streamlit, Docker

7. GraphRAG: Knowledge Graph-Augmented Retrieval for Semiconductor Equipment Intelligence

- Built a **GraphRAG pipeline** over K&S equipment manuals, SECS/GEM standards, and internal troubleshooting documentation—extracting entities (Equipment Models, Components, Alarm Codes, Failure Modes, Process Parameters) and typed relationships (TRIGGERS, CAUSES, REMEDIATED_BY, DEPENDS_ON) using **spaCy NER + LLM-assisted relation extraction** into a **Neo4j** property graph.
- Designed a **hybrid retrieval engine**: vector similarity (Qdrant) for dense chunk retrieval paired with Cypher-based **multi-hop graph traversal** for structured reasoning—e.g., “Which bonding parameters are linked to Alarm 0x4F3 across iFlex-T4 and RapidPro?” resolved as a 3-hop Cypher query across Equipment → Alarm → Parameter nodes.
- Outperformed flat-RAG by **31%** on multi-hop diagnostic queries (human-evaluated benchmark of 140 domain-expert Q&A pairs); LLM-generated Cypher queries achieved 87% structural correctness on held-out graph queries.
- **Stack:** Neo4j, Cypher, spaCy, LangChain, LangGraph, Qdrant, Claude API, FastAPI, Docker

8. SemiKG: Semiconductor Research Knowledge Graph with LLM-Powered Cypher Reasoning [GitHub](#)

- Constructed a **semiconductor research knowledge graph** by ingesting 3,200+ papers from the **Semantic Scholar API** (queries: wafer defect detection, IC failure analysis, semiconductor inspection)—extracting structured entities (Defect Type, Material, Detection Method, Dataset, Equipment, Metric) and typed relationships (DETECTS, USES_DATASET, ACHIEVES_ACCURACY, IMPROVES_ON, CITES) via **LLM structured-output extraction** (Claude API, JSON schema-constrained), yielding a **Neo4j** graph of 24,000+ nodes and 61,000+ relationships.
- Designed a 5-entity domain ontology grounded in semiconductor CV literature; automated entity resolution (fuzzy deduplication + embedding similarity) reduced duplicate nodes by 34%, ensuring graph coherence at scale.
- Developed a **LangGraph agent with Cypher tool-use**: natural-language queries are schema-aware—the agent generates, validates, and executes Cypher against Neo4j, then synthesises grounded answers—enabling multi-hop questions such as “Which detection methods achieve >95% accuracy on WM-811K and what backbone architectures do they use?” without hallucination.
- **Benchmarked** against flat-RAG (Qdrant) on 80 domain Q&A pairs: KG agent achieved **+28% answer correctness** on multi-hop questions and **zero hallucinated citations** vs. 11% for flat-RAG; Cypher generation accuracy 89% on held-out query set.
- **Stack**: Neo4j, Cypher, Semantic Scholar API, Claude API, LangGraph, LangChain, spaCy, FastAPI, Streamlit, Docker

9. SemiAgent Hub: MCP Server + A2A Orchestrated Multi-Agent System for Semiconductor Intelligence [GitHub](#)

- Built a **Model Context Protocol (MCP) server** exposing 7 semiconductor domain tools (Neo4j Cypher query, Qdrant semantic search, SECS/GEM alarm lookup, DINOv2 defect classifier, process parameter retrieval, SemiFA report trigger, wafer map annotator)—making the entire semiconductor intelligence stack accessible to any MCP-compatible client (Claude Desktop, Cursor, custom agents) via a single standardised interface.
- Architected four **Agent-to-Agent (A2A)** protocol agents—each publishing an **AgentCard** declaring capabilities, input/output schemas, and skills per the A2A spec: *Defect Analyst* (image + telemetry intake), *Root Cause Reasoner* (KG causal traversal), *Process Optimizer* (parameter recommendation), and *Report Composer* (structured output)—enabling dynamic agent discovery and task delegation without hardcoded orchestration logic.
- Demonstrated **cross-agent delegation**: Defect Analyst receives a wafer image, classifies it via MCP DINOv2 tool, then posts a structured A2A task to Root Cause Reasoner; the Reasoner queries the Neo4j KG via MCP Cypher tool and returns a causal chain; Process Optimizer receives the chain and responds with Bayesian-optimized parameter corrections—full pipeline completing in under 12 seconds end-to-end.
- MCP server and all four A2A agents containerised and orchestrated with **Docker Compose**; verified MCP compatibility with Claude Desktop; A2A task routing tested with Google’s **a2a-sdk** reference client.
- **Stack**: MCP Python SDK, A2A SDK (Google), LangGraph, Neo4j, Qdrant, DINOv2, Claude API, FastAPI, Docker Compose

Publications

SemiFA: An Agentic Multi-Modal Framework for Autonomous Semiconductor Failure Analysis Report Generation

Shivam Chand Kaushik

2026

Published on arXiv (cs.CV) — [arXiv:2604.13236](#)

Proposes a four-agent LangGraph pipeline fusing DINOv2 visual embeddings, SECS/GEM telemetry, and LLaVA-1.6 for automated semiconductor Failure Analysis report generation; includes the publicly released SemiFA-930 dataset (930 SEM/optical/wafer-map images, 9 defect classes) on HuggingFace.

Highlighted by Andre Zayarni (Co-founder, Qdrant) on LinkedIn — **post reached 2,120 impressions**, recognising the system’s real-world impact in chip manufacturing fabs.

Education

Indian Institute of Technology (IIT), Jodhpur
Master of Technology (M.Tech) — Artificial Intelligence

June 2024 — May 2026
CGPA: 8.46/10

JSS Academy of Technical Education, Noida
Bachelor of Technology (B.Tech) — Information Technology

June 2011 — May 2015

Achievements

- **Research featured by Qdrant Co-founder (2026)** — SemiFA paper ([arXiv:2604.13236](#)) highlighted by **Andre Zayarni**, Co-founder of Qdrant, as a real-world application of vector search in chip manufacturing fabs.
- **Bright Mind Award** — ASM Technologies Ltd. (2021), recognizing exceptional technical contribution in semiconductor software engineering
- **779th Rank** — International Mathematics Olympiad (IMO), demonstrating strong analytical and quantitative aptitude