

KARTHIK GUDURU

AI Solutions Architect | Agentic Systems | Generative AI | Computer Vision

Hyderabad, India | +91 9550946546 | gudur.karthik9@gmail.com

linkedin.com/in/karthik-guduru | medium.com/@Karthik__Guduru | karthikgudurup.framer.website

SUMMARY

AI Solutions Architect with 4+ years architecting and delivering Generative AI, Agentic AI, and Computer Vision systems for Fortune 500 clients across financial services, media, education, and energy. Proven record of taking solutions from concept to production, building autonomous agents, real-time computer vision (CV) pipelines, on-device LLM inference stacks, and RAG applications. Generated over \$335K in direct client revenue. Recognized as a LinkedIn Top Voice in AI and panelist on the Meta Reality Labs Research Panel.

Core Competencies: Machine Learning • Deep Learning • Natural Language Processing (NLP) • Computer Vision • Generative AI • AI Solution Architecture • Multi-Agent Systems • Model Deployment • RAG

TECHNICAL SKILLS

- **LLM & Inference:** LLM fine-tuning, prompt engineering, model evaluation, vLLM, NVIDIA NeMo, Triton Inference Server, CUDA optimization, LangChain, LlamaIndex, Ollama, model quantization (8-bit/4-bit), Gemini 2B, DeepSeek-7B, Qwen-1.5-7B, GPT-3.5/4
- **Agentic AI & RAG:** Multi-agent systems, agentic orchestration, Microsoft AutoGen, Model Context Protocol (MCP), Google ADK, Vertex AI (Enterprise AI), n8n, RAG pipelines, FAISS, vector databases, GitHub Copilot
- **Computer Vision:** YOLO-11n, RF-DETR, TorchServe, Roboflow, PyTorch, TensorFlow, OpenCV, FAISS similarity search
- **Generative AI & NLP:** Natural Language Processing (NLP), Stable Diffusion, Flux, ComfyUI, LoRA fine-tuning, image generation pipelines, Luma AI, Whisper API
- **MLOps & Infrastructure:** MLOps, model deployment, model serving, Databricks, Delta Lake, MLflow, Docker, TorchServe, Azure ML Studio, GPU optimization
- **Languages & Platforms:** Python, Streamlit, REST APIs, Jira API/JQL, Unity 3D, Arduino

EXPERIENCE

COGNIZANT TECHNOLOGY SOLUTIONS

Gen AI Solutions Architect, Associate

Hyderabad, India

Apr 2025 to Present

Google, Alphabet Ecosystem (Agentic AI & Operations)

- Streamline operations across YouTube Music, Google Artist Authentication, and Google TV, standardizing workflows and reducing manual operational overhead across teams.
- Architect an orchestration agent to handle customer support across a range of Alphabet products, routing queries, coordinating sub-agents, and resolving issues end-to-end within the Google ecosystem.
- Build multi-agent workflows on Google ADK and Vertex AI (Enterprise AI), leveraging native tooling for orchestration, grounding, and scalable deployment across Alphabet product surfaces.

NCR Voyix (Agentic AI, Automation & MLOps)

- Designed an autonomous QA agent with Microsoft AutoGen and Playwright that compressed testing cycles from 2 weeks to 3 days, a 70% reduction, by traversing test flows, validating outputs, and surfacing failures without human intervention.
- Built an Audit Bot on the Jira API that autonomously inspects every ticket for completeness, acceptance criteria quality, and metadata hygiene, eliminating manual sprint review overhead.
- Developed a Feature Guide Generator that queries Jira via JQL and synthesizes ticket data through an LLM pipeline to auto-produce formatted release documentation, saving hours of manual effort per sprint.
- Engineered end-to-end ML pipelines on Databricks (Delta Lake and MLflow) for Gen AI workload management, enabling experiment tracking and reproducible model versioning across the team.
- Trained teams on GitHub Copilot workflows, enabling organization-wide AI-augmented development and measurably improving code quality and delivery velocity.

Pearson OnVUE (Real-Time Computer Vision Proctoring)

- Containerized a YOLO-11n model with TorchServe for real-time exam proctoring at scale and fine-tuned an RF-DETR detector in Roboflow, leading data cleaning and augmentation across 10,000+ images, achieving sub-50ms detection latency per frame without cloud GPU dependency.

On-Device LLM Inference & NVIDIA Agentic AI

- Architected a private on-device LLM inference stack, quantizing and serving Gemini-2B, DeepSeek-7B, and Qwen-1.5-7B on an 8GB GPU via LangChain and vLLM, achieving sub-400ms responses while eliminating cloud costs for regulated enterprise environments.
- Delivered two production-grade NVIDIA agentic AI blueprints using NeMo, Triton Inference Server, and CUDA-optimized pipelines, reducing inference latency and enabling robust agentic decision flows.

TATA CONSULTANCY SERVICES (TCS)

AI Engineer, Research & Innovation (CTO Office)

Hyderabad, India

Jun 2022 to Mar 2025

TCS Media Lab (Generative AI Icon Pipeline)

- Built a zero-cost icon generation workflow using Flux, ComfyUI, and a locally trained LoRA model, delivering 60% faster output, 40% faster generation, and 30% lower compute versus commercial tools.

Munch Museum, Oslo (AI-Powered Visitor Experience)

- Built a FAISS-powered artwork similarity search engine across 50 artists achieving sub-1-second retrieval, and produced 30 generative AI image sets that expanded accessibility by 50% by removing AR hardware requirements.
- Shipped a GPT-3.5 Turbo and Whisper API voice chatbot as an AI art mediator, improving visitor engagement by 20%.

S&P Global, Lloyds Bank & Southern California Edison

- Built a generative AI-personalized 360-degree VR onboarding platform for S&P Global, achieving 20% faster time-to-productivity and generating \$85,000 in direct project revenue.
- Led delivery of a modular VR collaboration platform for Lloyds Bank with real-time multiplayer functionality, generating \$250,000 in revenue.
- Built an IoT-integrated VR application for Southern California Edison enabling real-time operational data access, improving field efficiency by 30% and enabling AI-driven fault prediction.

CXO Engagements (Qiagen, Verizon, TotalEnergies, Elisa)

- Presented AI, digital twin, and agentic AI adoption strategies to CTO, CMO, and COO teams across 4 global enterprises, and represented TCS at Elecrama 2023 and IMC 2023, evaluating 800+ AI/IoT startup partnerships.

AWARDS & RECOGNITION

- **LinkedIn Top Voice, AI:** Recognized by the LinkedIn editorial team for contributions to Artificial Intelligence.
- **Meta Reality Labs Research Panel:** Active panelist on Meta's closed AI research panel, Reality Labs division.
- **Luma AI Creator Program:** Selected for Luma AI's invite-only creator program (Nov 2024).
- **TCS Innovista Award:** Recognized for a 5G-enabled Digital Twin in the Piloted Technologies category.
- **IEEE Chairperson:** Served as chapter chairperson for 2 consecutive years, organizing multiple industry events.
- **Guinness World Records:** Spearheaded team participation in a Generative AI Hackathon record attempt.

EDUCATION

MUFFAKHAM JAH COLLEGE OF ENGINEERING & TECHNOLOGY

Bachelor of Engineering (B.E.)

Hyderabad, India

2018 to 2022

CERTIFICATIONS

- Databricks Certified AI Agent Developer, Databricks (2025)
- AI Advisor, Generative AI Technical, NVIDIA (2024)
- Career Essentials in Generative AI, Microsoft & LinkedIn (2024)
- Machine Learning for All, University of London (2023)
- AI for Everyone, DeepLearning.AI (2023)
- Claude 101, Anthropic Academy (2025)
- Claude Code 101, Anthropic Academy (2025)
- AI Fluency: Framework & Foundations, Anthropic Academy (2025)
- Introduction to Model Context Protocol (MCP), Anthropic Academy (2025)