

SARATH PAVAN KONDAPALLI

AI Engineer | Agentic AI | Generative AI | Large Language Models | Machine Learning
Bangalore, India | Open to Remote | +91-9390320844 | sarathpavan88@gmail.com | linkedin.com/in/sarathpavankondapalli

PROFESSIONAL SUMMARY

AI Engineer building Generative AI and Agentic AI systems, with a focus on multi-agent orchestration (LangGraph, LangChain), Retrieval-Augmented Generation (RAG), and LLM/SLM inference optimization across CPU, GPU, and AI accelerator infrastructure. Fine-tunes LLMs and computer vision models (LoRA, QLoRA, PEFT), builds scalable REST APIs and microservices, and takes AI systems from architecture through deployment. Has one enterprise workload running in production at 475,200 requests/month and three additional systems developed, validated, and benchmarked at production-grade concurrency. Moved four enterprise workloads off commercial APIs onto self-hosted open-source models, cutting inference costs 83 to 96% with no loss in performance. Works across AWS, GCP, Azure, and on-premise environments. Represented Infosys at the India AI Summit 2025.

TECHNICAL SKILLS

- **Agentic AI & GenAI:** LangGraph, LangChain, LlamaIndex, CrewAI, PydanticAI | Multi-Agent Orchestration, AI Agents, ReAct Loops, MCP (Model Context Protocol) Servers, Function Calling, Tool Calling, Agent Workflows, Workflow Automation, Context & Memory Management, Session Management, Prompt Engineering, Structured Outputs, Conversational AI, Chatbot Development, Human-in-the-Loop
- **Machine Learning & Deep Learning:** PyTorch, Hugging Face (Transformers, PEFT), scikit-learn, pandas, NumPy | Fine-Tuning (LoRA, QLoRA, SFT), LLM/SLM Training, Model Training, Model Evaluation, Computer Vision, Object Detection, Neural Networks, NLP, Deep Learning
- **Inference & Optimization:** vLLM, SGLang, llama.cpp, IPEX-LLM, OpenVINO, ONNX | GGUF, GPTQ Quantization | Tensor Parallelism, Continuous Batching, KV Cache Optimization | Latency, Throughput & Cost Optimization | LLM Evaluation, LLM-as-a-Judge, Hallucination Mitigation
- **RAG & Retrieval:** Retrieval-Augmented Generation (RAG), Vector Databases, Semantic Search, Hybrid Search | FAISS, ChromaDB, Pinecone | Nomic, bge-base-en-v1.5, all-MiniLM Embeddings | Document Ingestion, Chunking Strategies, Indexing, Retrieval Optimization, Document Intelligence, OCR (Tesseract, PyMuPDF), Data Preprocessing
- **Models:** Llama 3.x/4, Qwen 2.5/3, Gemma 2/MoE, Mistral, DeepSeek, Phi, GPT-4o, Claude, Gemini, YOLO | Multi-Modal, Mixture of Experts (MoE), TTS/STT | 4B to 70B Parameters
- **MLOps & Deployment:** MLOps, LLMops, LangSmith, MLflow | Docker, Kubernetes, CI/CD, Git | Model Deployment, Model Serving, Monitoring, Logging, Observability, Load Testing, Performance Benchmarking, Scalable AI Systems
- **Cloud & Infrastructure:** AWS (Bedrock), Azure (Azure OpenAI), GCP (Vertex AI) | Cloud-Native Architecture, On-Premise Deployment | Intel Xeon 6 (AMX), Intel Gaudi 3, NVIDIA H100 | Linux
- **Backend & Data:** Python, FastAPI, REST APIs, API Development, Microservices, Backend Development, Async Programming | SQL, PostgreSQL, MongoDB, Cosmos DB, Redis, Firestore, Neo4j (Graph DB) | AWS Boto3, ServiceNow REST, JWT / OAuth2 | n8n, Bash/Shell, Streamlit | AI Coding Tools: Claude Code, Cursor, OpenAI Codex, Gemini CLI
- **Security & Governance:** AI Governance, Responsible AI, Guardrails, Access Control (RBAC), Keycloak, Identity & Access Management (IAM), Data Privacy, Compliance

PROFESSIONAL EXPERIENCE

AI Engineer | Infosys Limited, Bangalore

Nov 2024 – Present

- Built and deployed a legal contract extraction and document intelligence pipeline to production, processing 475,200 requests/month. Ingests PDF, DOCX, and DOC files via PyMuPDF with Tesseract OCR for scanned documents, chunks at sentence boundaries with parallel stateless processing for deterministic outputs, extracts contractual obligations through an LLM with Pydantic structured outputs, and validates via LLM-as-a-Judge hallucination checks. Built a semantic classification engine using pre-computed Nomic embeddings and cosine similarity ranking to auto-categorize obligations into departmental taxonomies (HR, Legal, Compliance, Finance) with confidence scoring. Replaced Azure OpenAI GPT-4o with self-hosted Qwen3-4B on a 96 vCPU Xeon 6 GCP VM, reaching 97.5% extraction accuracy at 21s per page with 40 concurrent requests validated, reducing inference cost roughly 96%.
- Architected an enterprise IAM orchestration platform built as seven independently deployable services. A master orchestrator agent (ReAct loop, session manager, SSE event bus) coordinates an access advisor sub-agent for planning, diagnosis, and evidence gathering, and an access writer sub-agent for onboarding, offboarding, and

revocation. Integrated Keycloak through a custom IAM MCP server for role-based access control (RBAC) and group management, with ServiceNow as the ticket intake layer, policy lookup RAG on ChromaDB with bge-base-en-v1.5 embeddings, a PostgreSQL-backed access registry, and Redis for session persistence. Served Gemma 4B on Intel Xeon 6 (96 vCPU GCP VM) via vLLM, validated at 100% workflow accuracy, 21.9 tok/s throughput, and 34.71s average agent latency at 30 concurrent requests, with projected monthly inference cost savings of 88.9% versus GPT-4o API baseline.

- Developed a context-aware hospitality assistant chatbot using a 5-agent architecture (1 master orchestrator plus 4 domain child nodes) on LangGraph, with custom templated conversational buffer memory and Firestore-backed long-term state for booking, cab, hotel, and restaurant workflows. Fine-tuned Qwen SLMs using LoRA/QLoRA (Hugging Face PEFT) for domain-specific task performance. Rebuilt the model stack from Azure OpenAI GPT-4o onto Qwen3-4B served via vLLM on Intel Xeon 6 (96 vCPU), which outperformed every other model evaluated on node invocation accuracy, detail extraction, and concurrency handling while running at roughly 17% of the API baseline cost. Validated at 50 requests/minute across multiple tensor parallelism configurations.
- Developed an autonomous AIOps platform that automates AWS incident triage, EC2 lifecycle management, EBS volume inspection and resize, CloudWatch metric retrieval, and dynamic Boto3 execution, with ServiceNow incident creation and SRE trend reporting. Orchestrated the agents on LangChain and LangGraph with Pinecone-backed SOP RAG, FastAPI, and JWT with Azure AD OAuth2, serving Qwen3-4B-Instruct via vLLM on Intel Xeon 6 (96 vCPU, AMX/AVX-512). Validated at 100% tool invocation accuracy and 100% task completion at 21.75 peak tok/s, 1.65s average TTFT, and 14.57s average end-to-end latency, with a projected 12.6x cost advantage over the GPT-4o API baseline.
- Provisioned on-premise Intel Gaudi 3 AI accelerator infrastructure end-to-end and benchmarked it against NVIDIA H100 across the Llama, Qwen, Gemma, Mistral, and DeepSeek model families, measuring TTFT, ITL, throughput, and accuracy across multiple tensor parallelism configurations via vLLM. Gaudi 3 matched H100 inference performance at roughly 50% lower hardware cost, pointing to a meaningful price-performance advantage for enterprise inferencing.
- Fine-tuned, trained, and benchmarked YOLO computer vision models (YOLOv5-L for 33-class P&ID detection, plus YOLOv8n) across the full model lifecycle from data preparation through evaluation and deployment. Optimized CPU inference with OpenVINO and BF16 on Intel Xeon 6 with AMX, cutting end-to-end latency 2.8x versus native PyTorch (51ms single-user, scaling predictably to 100 concurrent users) and identifying a 15-core thread-tuning sweet spot. Separately evaluated edge AI deployment on Intel AI PCs with GGUF-quantized sub-8B models for on-device tooling.

INDUSTRY REPRESENTATION & TECHNICAL LEADERSHIP

- Selected as the sole technical JL3 representative for Infosys at the India AI Summit, a 7-day national expo, presenting Generative AI and Agentic AI solutions alongside senior leadership. Delivered technical presentations across Bangalore, Mysore, and Delhi, and served as a cross-team architecture advisor on solution optimization and multi-hardware platform performance tuning.

CERTIFICATIONS

- Infosys Certified Generative AI Professional: Foundation, Intermediate & Applied (2025)
- Anthropic Hackathon (2025). Built agentic AI solutions using Claude with prompt engineering and tool-use patterns.

EDUCATION

B.Tech in Computer Science & Engineering | SRKR Engineering College, Bhimavaram

2019 – 2023