Biyyala Saathvik

+91 6305425040 saathvikbiyyala@gmail.com linkedin.com/in/saathvik-biyyala github.com/bsaathvik19

Education & Certifications

VNR Vignana Jyothi Institute of Engineering and Technology (VNRVJIET) 2021 – 2025
B.Tech in Electronics & Communication Engineering (CGPA: 8.39 / 10.0) *Hyderabad, India*

- **Minor:** Artificial Intelligence & Machine Learning (CGPA: 8.30 / 10.0)
- **Certifications:** Claude Certified Associate – Foundations, Data Structures & Algorithms (SmartInterviews), Programming in Java (NPTEL Elite), Introduction to IoT (NPTEL Elite + Gold)

Professional Experience

Tata Consultancy Services (TCS Prime – BFSI CTO Unit) Jul 2025 – Present
AI Systems Engineer — Agentic AI, SLM Engineering & Enterprise AI Architecture *Chennai, India*

- **Empathetic Financial Guidance Agent (TCS Blitz 2026 Winner) :** Architected and deployed a production LangGraph multi-agent system with Human-in-the-Loop (HITL) orchestration in 3 months for a Tier-1 UK bank following a 9-month client impasse, earning the Best Executive AI-Led Service Delivery Award.
- **Financial Crime Investigation Assistant :** Engineered an autonomous 9-agent AML orchestration pipeline automating end-to-end case lifecycles, cross-source typology analysis, automated Suspicious Activity Report (SAR) drafting, and rubric-based iterative refinement.
- **Clicks2Converse & Digital Onboarding :** Engineered an intent-driven conversational product discovery and advisor scheduling agent for a major Australian financial institution, collapsing 20+ manual navigation clicks into a single conversational turn (95% step reduction).
- **ACE Framework Core Architecture & Analyst Briefings :** Core architect of the unit's flagship Adaptive Conversational Experience (ACE) framework (delivered 4 of 10 production ready solutions); conducted technical briefings for C-suite bank executives and global industry analysts, securing direct citation in Celent's Flash Report (May 2026).
- **Legacy Code Modernization SLM :** Leading development of a domain-adapted SLM for legacy code modernization, overseeing curation of a 7B-token corpus, synthetic data generation, continued pre-training, and supervised fine-tuning for bidirectional code engineering across 8 languages.
- **MIRA & Unified Asset Aggregation :** Implemented an agentic mortgage consultant integrating domain-aware RAG, dynamic financial visualizations, and CRM workflows; built a Unified Asset Aggregation MVP consolidating real-time holdings and analytics across heterogeneous wealth systems.

Systems & AI Engineering Projects

Proteus – Self-Hosted Agentic AI Workspace | *FastAPI, MCP, ChromaDB, ONNX, Docker, SQLite, PWA* Proteus-ai

- Solo-architected and shipped a self-hosted agentic AI workspace with ∼60 FastAPI route modules, a PWA frontend, and 4 native execution sandboxes.
- Engineered a RAG-gated agent execution loop using semantic top-k tool retrieval to reduce token overhead and improve tool selection across large MCP registries.
- Built a hybrid RAG and persistent memory pipeline using ChromaDB, ONNX embeddings, and BM25; shipped a CLI with ∼20 subcommands and native installers for Linux, macOS, and Windows.

Real-Time Traffic Sign Distance Estimation | *Depth-pro, YOLOv8, TensorFlow, Python, OpenCV* 2023 – 2024

- Engineered a real-time computer vision pipeline pairing YOLOv8 object detection with Depth-pro zero-shot monocular depth estimation to calculate metric distance to road signs in live video feeds.
- Benchmarked spatial bounding box localization against dense relative depth maps to achieve low-latency hazard detection for autonomous driving edge hardware.

Honors, Awards & Recognition

Winner – Best Executive AI-Led Service Delivery Award (TCS Blitz 2026): Awarded for the 3-month architectural turnaround and production delivery of the LangGraph Empathetic Financial Guidance Agent after a 9-month client impasse.

External Industry Analyst Citation (May 6, 2026): Core ACE framework engineering work cited in Celent's Flash Report, *"AI is Enhancing Enterprise Software, Not Replacing It"*, following sole technical briefings to Gartner, Everest Group, and Celent.

Technical Skills

Programming & Data: Python, SQL, Java, Bash, NumPy, Pandas, Pydantic
LLM & SLM Engineering: MLX, Hugging Face Transformers, Continued Pre-Training (CPT), Supervised Fine-Tuning (SFT), QLoRA, LLM Inference, OpenAI / Gemini / Claude APIs
Agentic AI & RAG: LangGraph, LangChain, Multi-Agent Systems, Agent2Agent (A2A), Model Context Protocol (MCP), Tool Calling, Agent Orchestration, Agentic RAG, Hybrid Search, ChromaDB
ML, Deep Learning & CV: TensorFlow, Neural Networks, Transformers, ONNX, OpenCV, YOLOv8
Cloud & AI Systems: Google Cloud Platform (GCP), Vertex AI, FastAPI, Docker, SQLite, SQLAlchemy, Git, GitHub
LLM Evaluation & Reliability: Arize Phoenix, LLM Evaluation, Guardrails, Human-in-the-Loop (HITL)