

Sai Teja Vaidya

Applied AI & GenAI Engineer | Forward Deployed Engineer
Bangkok, Thailand | vaidyasaiteja143@gmail.com | +66-0825674212
[Linkedin](#) | [Blog](#) | [Github](#) | [x.com](#)

Summary

Applied AI and GenAI Engineer with a **Forward Deployed Engineer (FDE)** mindset and a track record of turning ambiguous problems into working AI systems. Ships LLM- and agent-powered workflows, computer vision pipelines, and edge-optimized models end-to-end under real latency, cost, and accuracy constraints. Thesis work delivered an AI Teacher Agent that cut annotation time by over 80%, demonstrating direct problem ownership from stakeholder need and experimentation through deployment and measurable impact.

Technical Skills

AI/GenAI & Agents	LangChain, LangGraph, RAG, AI Agent Orchestration, Prompt Engineering, LL-MOps, Multi-Agent Systems
Computer Vision	YOLOv8/v11, Mask R-CNN, SAM, Instance/Semantic Segmentation, OpenCV, Object Detection
ML/DL Frameworks	PyTorch, HuggingFace, ONNX, Scikit-learn
Cloud & MLOps	AWS (SageMaker, Bedrock), Model Deployment, Quantization, Edge Inference
Geospatial AI	UAV Imagery, Spatial Intelligence, Geospatial Data Pipelines, Multi-GSD Analysis
Languages/Tools	Python, TypeScript/JavaScript, LaTeX, Git/GitHub, Jupyter, Docker

Experience

AI/ML Trainee – Computer Vision & GenAI Platforms	Apr 2025 – Jan 2026
<i>AI Brain Lab (Contract)</i>	<i>Bangkok, Thailand (Hybrid)</i>

- Collaborated with AI research and engineering teams on computer vision tasks and applied machine learning algorithms for production platforms.
- Designed intelligent platforms transforming complex operational workflows into data-driven insights, improving customer service outcomes.
- Contributed to AI-driven workflow automation solutions, enhancing operational efficiency and end-user experience.

AI Computer Vision Researcher	Aug 2023 – Jan 2026
<i>Asian Institute of Technology (AIT)</i>	<i>Khlong Luang, Thailand</i>

- Led “AI Agent-Orchestrated Instance Segmentation”: detection and canopy size estimation of oil palm trees using UAV imagery across 8 resolution levels (0.03–0.20m GSD).
- Deployed best-performing model weights as an AI Teacher Agent for automatic annotation, reducing manual labelling time by **>80%**.
- Benchmarked 6 models (YOLOv8, YOLOv11, Mask R-CNN, Hybrid-SAM) achieving mean IoU of 0.77 and F1 of 0.789 at optimal 0.04m GSD.
- Supervised by Dr. Sarawut Ninsawat, Prof. Nitin Tripathi, and Dr. Sanit Arunplod; published Q1 journal paper and successfully defended thesis.

Staff Project Assistant	Aug 2025 – Dec 2025
<i>Asian Institute of Technology (Contract)</i>	<i>Bangkok, Thailand</i>

- Planned and executed internal/external events for an international graduate community, managing hybrid Zoom setups for technical operations.
- Produced multimedia content and website updates, aligning messaging with program objectives and stakeholder needs.

Student Project Assistant

Asian Institute of Technology (Part-time)

Aug 2024 – Jun 2025

Pathum Thani, Thailand

- Coordinated academic programs and multicultural seminars/technical talks, boosting student engagement across the AIT_SDP program.
- Designed information-flow systems that improved communication with a diverse international student body.

Geospatial AI Analyst

Asian Institute of Technology

Aug 2024 – Dec 2024

Pathum Thani, Thailand

- Conducted a literature review of AI-assisted COVID-19 trend analysis and spatial hotspot mapping methods (2019–2022).
- Synthesized four forecasting and four spatial hotspot-detection methodologies into a decision-intelligence comparison of predictive performance.

Selected Applied AI / Agentic Projects

AI Sales Enablement Speech Coach

Apr 2026 – Present

Tauri, React, Whisper (local inference)

- Built a real-time teleprompter and rehearsal coach ingesting scripts/slides via OCR with local speech transcription for instant pacing and filler-word feedback.
- Designed a latency-budgeted feedback engine: fast local pass for immediate flags, slower deep pass for structural/tone analysis.

Compounding Institutional Memory AI Agent

Mar 2026 – Present

NetworkX, Vector Store

- Built a self-organizing memory graph (episodes + abstractions as nodes) that reduces agent run cost over time instead of scaling context tokens.
- Implemented a reflection agent merging recurring patterns into reusable abstractions; benchmarked cost-per-task, retrieval hit rate, and accuracy delta across a 100–200 task stream.

Cost-Governed Ops Triage Agent

Jan 2026 – Present

LangGraph, Qwen2.5-3B, API Fronting

- Built a LangGraph-based agentic router that sends each support ticket to the cheapest capable model tier while holding accuracy steady.
- Added early-exit and guardrails with a “return-to-control” path, benchmarking cost-per-ticket reduction and p95 latency vs. an always-expensive baseline.

AI Thesis Agent Kit (Open Source)

Jun 2025 – Apr 2026

Multi-Agent Orchestration, LLM Guardrails

- Designed a 9-agent research writing system (interrogator, writer, reviewer, citation checker, and more) enforcing 6 immutable writing laws to eliminate hallucination.
- Built a 3-layer architecture and 90% confidence gate blocking drafting until research foundations are solid; used to support Q1 paper acceptance and thesis defense.

Qualcomm AI Hub LPCVC 2026 – Track 1

Mar 2026 – Apr 2026

ONNX, Edge Inference, XR2 Gen 2

- Reproduced the LPCVC 2026 image-to-text retrieval baseline for XR2 Gen 2 (proxy edge device): ONNX export, model compilation/profiling, dataset upload.
- Achieved a validated baseline of Recall@10 = 0.7299 with 29.3ms image encoder and 5.4ms text encoder latency; resolved key Python/SSL/dataset reproducibility blockers.

Oil Palm Instance Segmentation (Master's Thesis)

Jan 2025 – Jan 2026

YOLOv8/v11, Mask R-CNN, SAM

- Delivered accurate oil palm crown delineation and canopy size estimation from UAV RGB imagery across 8 GSD levels (0.03–0.20m) for the \$67B global oil palm industry.

- Compared 6 segmentation models on precision, recall, F1, IoU, and crown geometry error; agent-in-the-loop annotation cut labelling time by 80%.

Education

Master's Degree, Information and Communications Technology Jan 2023 – Jan 2026
Asian Institute of Technology GPA: 3.0/4

- Thesis: “Detection and Canopy Size Estimation of Oil Palm Trees Using YOLOv8, YOLOv11, and SAM on Multi-GSD UAV Imagery”

Bachelor of Technology, Computer Science Apr 2019 – Nov 2022
JNTUH College of Engineering, Hyderabad GPA: 8.8/10

- Built “College Buddy App” – a peer-to-peer campus marketplace for sustainable textbook exchange; Student Volunteer & Coordinator, QUEST Technical Fest.

Certifications

- AWS AI Practitioner (AIF-C01) – Micro-Credentials Track, Amazon Web Services (Oct 2025)
- LangChain Academy – Quickstart: LangChain Essentials (Python), LangChain (Jan 2026)
- Kaggle x Google – 5-Day AI Agents: Intensive Vibe Coding Course, Kaggle (Jun 2026)
- Generative AI Mastermind, Outskill (Nov 2025); Generative AI Essentials & Introduction and Applications, IBM (Jul 2025)
- Foundational Mathematics for AI, Johns Hopkins University (Jul 2025); NVIDIA Prompt Engineering and Data Analysis, Whizlabs (Jul 2025)