LIKITH VISHAL BODDEDA

likithvishal.b@gmail.com — +1(408)752-6185
LinkedIn — GitHub — Portfolio
Richardson, TX

Technical Skills

Python · C++ · JavaScript · TypeScript · SQL · Docker · Kubernetes · AWS (Lambda, S3, EventBridge) · Terraform · PyTorch · TensorFlow · LangGraph · RAG Pipelines · Splunk · Grafana · FastAPI · React · CI/CD · Pytest

Professional Experience

CISCO — *Software Engineer (SRE/AI)* Richardson, TX — Nov 2024 – Present

FROC v2 (LLM-Powered Incident Response): Deployed self-hosted Llama 3.2 on air-gapped infrastructure. Built 6-source RAG pipeline (PagerDuty, Splunk, metrics, alerts). Designed intent-detection engine for dynamic SPL query construction with 7-rule anti-hallucination framework. Integrated Langfuse observability. Developed 40+ Pytest test suites. Containerized stack (FastAPI, React 18+TypeScript, PostgreSQL) with GitHub Actions CI/CD. Impact: 70% MTTR reduction, 100% automated Splunk interpretation.

- Architected and deployed production self-hosted LLM system (Meta Llama 3.2 3B, Q4_K_M quantized, ˜2GB VRAM) via Ollama on air-gapped Cisco infrastructure with zero data egress; built custom MCP-compatible inference wrapper enabling seamless LLM service integration across multiple downstream agent workflows and internal tooling
- Built 6-source real-time RAG pipeline ingesting PagerDuty incidents, hot datacenter statistics, on-call rotation schedules, datacenter health metrics, and Splunk Cloud logs — context assembled fresh per inference request to eliminate stale data issues; engineered sophisticated intent detection engine extracting datacenter names, check names, and time ranges to dynamically construct validated SPL queries
- Designed custom anti-hallucination prompt framework (7 production-hardened rules) with injection-safe query validation and structured output enforcement; integrated self-hosted Langfuse for full LLM observability tracking tokens, latency distributions, prompt-response pairs, and cost attribution
- Containerized full-stack application (FastAPI backend, React 18 + TypeScript frontend, Langfuse telemetry, PostgreSQL persistence) across 4 Docker containers with automated GitHub Actions CI/CD pipeline for continuous deployment
- Implemented MCP Server integration enabling FROC v2 to expose incident investigation capabilities and alert context as structured tools consumable by external agentic systems; designed protocol-compliant tool schemas for alert correlation, root cause analysis, and cross-system incident linking

JENAI (Agentic Jenkins Automation Across 6 Enterprise Systems): Architected LangGraph state machine (9 nodes) orchestrating Jenkins, GitHub, Splunk, SignalFx, Confluence, Webex, PagerDuty. Transformed 30–45 min manual process into autonomous 3-tap workflow. Integrated LangSmith tracing. Implemented 7-step failure recovery with intelligent retry logic and Redis checkpointing. Orchestrated 10+ parallel Jenkins jobs with cross-system coordination. Impact: 95% SRE engagement reduction, <3min recovery time.

- Designed production agentic system using LangGraph state machine architecture (9 nodes with conditional edges and cycles) transforming a 30–45 minute manual Out-of-Rotation process spanning 6 enterprise systems (Jenkins, GitHub, Splunk/SignalFx, Confluence, Webex, PagerDuty) into fully autonomous 3-tap workflow
- Built MCP-compliant tool layer exposing Jenkins job queries, GitHub PR filtering, and incident escalation as standardized protocol endpoints; enabled seamless agent decision-making by abstracting enterprise system complexity into composable, type-safe tool abstractions with automatic retry and timeout handling
- Integrated LangSmith as comprehensive evaluation and tracing platform — providing full visibility into every LLM call, agent routing decision, token usage breakdown, and latency measurement across entire state machine execution for continuous accuracy monitoring and prompt quality improvement
- Implemented parallel tracking of 10+ Jenkins child jobs with real-time status aggregation, human-in-the-loop approval gates for critical decisions, and 7-step failure recovery sub-workflow with Redis checkpointing ensuring full crash resilience and resumable execution

Infrastructure-as-Code (Terraform & Detector-as-Code): Architected Terraform alert infrastructure with burn-rate SLOs and GitOps workflow (PR → review → CI/CD → prod). Codified Splunk dashboards and Grafana alerts as infrastructure. Provisioned AWS pipeline (Lambda, S3, EventBridge, IAM). Developed Pytest validation suites. Impact: Eliminated drift, weeks → hours onboarding, 100% config traceability.

- Architected Terraform-based alert infrastructure with codified trigger conditions, burn-rate SLOs, and full GitOps workflow (PR submission → automated review → CI/CD validation → production deployment) ensuring reproducible and auditable infrastructure changes
- Codified Splunk Cloud dashboards and Grafana alerting rules as version-controlled infrastructure; provisioned complete AWS event-driven pipeline (Lambda functions, S3 storage, EventBridge scheduling, IAM policies) as declarative Terraform modules
- Developed comprehensive Pytest validation suites for infrastructure correctness; eliminated configuration drift entirely; reduced new engineer onboarding from weeks to hours; achieved 100% configuration traceability across all monitoring infrastructure

Network Observability Migration: Migrated 230 synthetic tests (110 TCP + 120 DNS) across 5 DC types. Built dashboards (Splunk, SignalFx, Grafana). Orchestrated Docker deployments to 20+ datacenters via Ansible/Kubernetes.

- Led comprehensive migration of 230 synthetic monitoring tests from Catchpoint to ThousandEyes (110 TCP connectivity tests + 120 DNS resolution tests) across 5 distinct datacenter architecture types (DCv1/SIG, DCv2/SSE Edge, vEDC/SSE AWS, vEDC_CN/SSE China, Regional Internet), maintaining zero observability gaps during transition
- Built and maintained production monitoring dashboards across Splunk Cloud, SignalFx, and Grafana; managed FR-Monitor Docker container deployments via Ansible playbooks and Kubernetes manifests to 20+ geographically distributed datacenters
- Engineered AWS Lambda aggregation pipeline (15-minute intervals) normalizing heterogeneous alerts from 3 monitoring systems (ThousandEyes, Catchpoint, internal probes) into unified S3 schema for historical trend analysis and capacity planning

Site Reliability Engineering — Oncall & Incident Management:

- Architected full-stack FROC v2 dashboard (FastAPI + React) integrating PagerDuty, Splunk Observability, Splunk Cloud, Catchpoint, and Grafana — delivering unified incident streaming, DC-level heatmaps, and AI-powered investigation summaries
- Operated distributed monitoring infrastructure spanning 6 production repositories, paging first-responder engineers across U.S., EU, APAC, Middle East, and China regions via zeustest probe failures; maintained 99.9% alert accuracy

Symbiosis Technologies — *Data Scientist Intern* Vishakapatnam (Remote) — Sep 2020 – Mar 2021

NLP-Based Ticket Classification & Production ML Pipeline: Engineered NLP pipeline (scikit-learn, NLTK, spaCy) for multi-label categorization. Achieved 85% accuracy via stratified k-fold ($k = 5$). Built Flask microservice deployed to Heroku with CI/CD. Impact: 30% triage reduction, 30% resolution time improvement, 100% automated routing.

- Engineered production NLP pipeline for automated ticket categorization using scikit-learn, NLTK, spaCy; implemented multi-label classification with binary relevance approach handling 15+ overlapping support categories with imbalanced class distributions
- Achieved 85% micro-averaged accuracy with stratified k-fold cross-validation ($k = 5$); evaluated ensemble algorithms (SVM with RBF kernel, Gaussian NB, logistic regression with TF-IDF vectorization); optimized feature selection via information gain reducing dimensionality 40% while improving F1 by 15%
- Built production ML pipeline using Flask microservice architecture, containerized with Docker, deployed to Heroku with CI/CD for automated model retraining; implemented comprehensive inference monitoring (latency tracking, data drift detection) maintaining 99% uptime
- Reduced false positive rate to 8% through systematic misclassification analysis, confidence thresholding, and ensemble voting; implemented Platt scaling for probability calibration enabling human-in-the-loop review for borderline cases

---

The Climber — *Digital Marketing Specialist* Remote — Jan 2019 – Nov 2019

SaaS Growth & Marketing Automation: Content strategy & keyword research (SEMrush), 40% MoM organic growth, 50% ranking improvements. CRO testing (25–30% conversion lift). Email automation (18–22% CTR). Analytics dashboards (CAC/LTV, 40% ROI improvement).

- Developed content strategy targeting B2B personas; keyword research (SEMrush) driving 40% MoM organic growth and 50% ranking improvements across 15+ keywords; implemented CRO testing achieving 25–30% conversion improvement
- Built email nurture sequences with behavioral triggers (18–22% CTR vs 2–3% baseline); analytics dashboards tracking CAC/LTV with 40% ROI improvement; delivered 40% QoQ pipeline growth, 30% CAC reduction
- Impact: 40% QoQ pipeline growth, 30% CAC reduction, 35% from organic.

VDC/NU-IDEA Program — *Entrepreneurship & Technology Intern* GITAM × Northeastern — Feb 2020 – Oct 2020

EyeCan (Hardware-Software Health-Tech Product)

- Built real-time computer vision pipeline on Raspberry Pi achieving 20 FPS with <75ms end-to-end latency using OpenCV; integrated HC-SR04 ultrasonic and infrared proximity sensors (±4cm accuracy) for distance and posture tracking
- Implemented multi-modal alert system (distance warnings, posture angle detection, break reminders) reducing daily screen exposure; conducted 25+ customer discovery interviews achieving 76% problem-solution validation
- Piloted with 12 beta users demonstrating 47% eye strain improvement and 68% posture compliance; developed monetization roadmap including B2B licensing strategy. Stack: Raspberry Pi, Python, OpenCV, C, HC-SR04, IoT.

Projects & Portfolio

SpeakLine/VoiceComment (Voice-to-Code Commenting Tool, 2024): Production AI/ML developer tool enabling voice-driven inline code comments. Architected modular system with pluggable audio backends (LocalAudioRecorder, WhisperTranscriber, OpenAITranscriber). Built language-specific AST parsers supporting 8+ languages (Python, JavaScript, TypeScript, Go, Rust, Java, C#, Ruby). Configurable silence detection, indentation-aware comment insertion. CLI (Typer), Python API, IDE integration (VS Code, Vim, Jupyter). Production-ready: type hints, error handling, pytest coverage. Open-source MIT, GitHub with CI/CD.

AstroFusion: Multi-Messenger Black Hole Detection (2024): AI system analyzing gravitational waves (LIGO), neutrinos (IceCube), and cosmic rays (Pierre Auger). Transformer-based cross-modal attention with physics-informed encoding. LLM-powered RAG pipeline (10,000+ papers) for novelty verification. Targeting arXiv publication.

Vision-Language Image Captioning: Fine-tuned Salesforce BLIP (223M params) on Flickr8k, 70+ epochs, 15–20% BLEU improvement. Mixed-precision training, cosine annealing LR. Deployed on Streamlit Cloud.

Natural Language to SQL via Fine-Tuned LLMs (2023): Fine-tuned Llama-2-7B using LoRA achieving 99.4% parameter reduction (14GB → 3.5GB with 4-bit NF4 quantization). 72% exact match accuracy, 85% BLEU score on 5K sample evaluation set. Sub-second inference on consumer GPU (RTX 3060, 6GB VRAM).

Research Publications

Vision for Eyes: Deep Learning-Based Assistive System for Visually Impaired Jan 2022
Springer - Lecture Notes in Networks and Systems, Vol. 340 (ICIoTCT 2021) — DOI: 10.1007/978-3-030-94507-7_17

Comparative Study of Classification Algorithms Over Images Using ML and TensorFlow Jul 2021
Springer - Algorithms for Intelligent Systems (SCIS 2021) — DOI: 10.1007/978-981-16-1866-6_20

Education

Westcliff University Irvine, California — Mar 2022 – Mar 2024
Master of Science – Computer Science GPA: 3.81/4.0

GITAM University Vishakapatnam, India — Jun 2017 – Apr 2021
Bachelor of Technology – Computer Science