

Anil Sahith Vallepu

+91 8143400946 ✉ anilsahithvallepu@gmail.com

in linkedin.com/in/anil-sahith github.com/sahit1011 anil-portfolio-chi.vercel.app

PROFESSIONAL SUMMARY

Software Engineer with 2+ years of experience building production-scale AI systems, distributed backend infrastructure, and low-latency ML inference pipelines. Experienced in developing LLM applications, multimodal AI systems, GPU inference optimization, and cloud-native services using Python and Rust. Proven track record of architecting scalable systems processing millions of requests across Kubernetes, cloud GPUs, and edge devices.

Seeking roles: Software Engineer, Forward Deployed Engineer, AI/ML Engineer, Data Scientist.

SKILLS

- **Languages & Tools:** Rust, Python, C++, JavaScript, TypeScript, SQL, Git, Linux, Zed, Cursor, Postman, Figma.
- **AI/ML:** Hugging Face, TensorFlow, PyTorch, CUDA, ONNX, TensorRT, vLLM, SGLang, OpenCV, Pipecat, STT, TTS, LLMs, VLMs, RAG, MCP, LoRA/QLoRA, DPO, GRPO.
- **Backend:** Next.js, Node.js, Flask, FastAPI, WebSockets, ZeroMQ, RabbitMQ, Kafka, Redis Streams, TensorZero, LiteLLM.
- **Databases:** MySQL, PostgreSQL, MongoDB, DynamoDB, Redis, Pinecone, Qdrant, ChromaDB.
- **Cloud & Infra:** Git, Linux, Docker, Kubernetes, AWS, GCP, Terraform, ArgoCD, Prometheus, Grafana, Temporal, CubeAPM, Modal, Supabase, Pipeshift.

EXPERIENCE

- **Arrowhead AI** Mar 2026 - Current
SDE - 2 Bangalore, Karnataka
 - Built and deployed 20+ production client integrations as Rust plugins compiled to WebAssembly(Extism), connecting the AI voice platform to partner CRMs and APIs in real time during live calls.
 - Architected a platform-wide circuit breaker driven by PromQL metrics, exposing REST APIs, MCP tools, and CLI utilities that improved production resiliency and accelerated incident response during live campaigns.
 - Contributed to building in-house self-hosted STT, TTS, and LLM models, supporting dataset generation, model fine-tuning, and evaluations; powering most of the production campaigns.
- **Matters.AI** Dec 2025 - Mar 2026
Software Engineer - AI/ML Bangalore, Karnataka
 - Designed and deployed a distributed AI inference platform on GKE, routing OCR and NER workloads across Modal and self-hosted GPU backends using Redis Streams, sustaining up to 10k RPS ingest and 3K RPS downstream with health-aware routing, rate limiting, and retry/DLQ, at 0.001% failure across 1M+ documents.
 - Deployed GPU inference services using ONNX Runtime, TensorRT, Flash Attention and vLLM, reducing p99 inference latency below 100 ms for NER models and below 500 ms for OCR workloads.
- **Noccarc Robotics Pvt. Ltd.** June 2024 - May 2025
AI/ML Engineer Pune, Maharashtra
 - Implemented an end-to-end real-time ECG Arrhythmia Detection pipeline for edge patient monitors, integrating signal preprocessing, deep learning inference, alarm generation, and frontend visualization to detect 20+ cardiac abnormalities reducing clinical response time in ICU environments.
 - Designed a lightweight ResNet + CBAM + GRU hybrid neural network optimized for embedded deployment, achieving sub-300 ms inference while utilizing less than 12% CPU on ARM-based patient monitoring hardware.

PROJECTS

- **CryptAI:** Built an autonomous crypto futures trading system using a multi-agent architecture where specialized AI agents analyze markets, identify swing/scalp setups, validate risk, execute real-time trades via WebSockets, store trade history, with a live dashboard for portfolio and agent activity monitoring.
Tech stack: Python, Next.js, FastAPI, Redis, PostgreSQL, LangGraph, Groq LLM API, Docker, Render, Vercel.
- **StudyArc:** Developed an adaptive learning platform with a multi-agent orchestration system for personalised NCLEX exam preparation; programmed an intelligent scheduling algorithm for dynamic study calendar generation based on performance metrics and integrated a contextual AI chatbot tutor.
Tech stack: Next.js, TypeScript, Chart.js, Node.js, MongoDB, OpenRouter LLM API, Ollama SLMs.

EDUCATION

- **National Institute of Technology (NIT) Warangal.** Warangal, India
Bachelor of Technology in Electrical and Electronics Engineering. May 2024

ACHIEVEMENTS

- Secured 99.3rd percentile in JEE Main (India) - Top 0.7% among 1 million+ candidates
- Recognised with Top 100 Projects at NASA AMES Space Settlement Contest, San Juan, Puerto Rico (2017)
- Solved 300+ Data Structures and Algorithms problems (Medium-Hard difficulty) on LeetCode