

Siddharth Bayapureddy

+91 7075286119 | siddharthbayapureddy@gmail.com | siddharth-bayapureddy.vercel.app |
linkedin.com/in/siddharth-bayapureddy | github.com/SiddharthBayapureddy

SUMMARY

Machine learning engineer with hands-on experience building end-to-end MLOps pipelines, generative AI systems, and LLM-powered agentic applications. Proficient in PyTorch, AWS SageMaker, retrieval-augmented generation (RAG), and few-shot learning. Seeking AI/ML engineering or applied research internships where strong fundamentals meet production-grade thinking.

EDUCATION

BITS Pilani, Hyderabad Campus

Bachelor of Engineering in Computer Science (CGPA: 9.1/10)

Hyderabad, TS

Aug. 2024 – Jul. 2028

EXPERIENCE

Research Intern, Mechanistic Interpretability

Genquery Pvt Ltd — Under Prof. Dhruv Kumar

Hyderabad, TS

May 2026 – Jul. 2026

- Built a logit lens pipeline on GPT-2-small to trace intermediate-layer predictions and analyze how representations evolve across transformer blocks
- Ran activation patching experiments to localize causal circuits driving specific model behaviors, isolating attention heads and MLP layers responsible for target completions

PROJECTS

Vigil | *Python, XGBoost, AWS Lambda, Step Functions, SageMaker, Docker, Mistral*

Jun. 2026 – Present

- Engineered an autonomous MLOps pipeline for aircraft engine predictive maintenance (NASA CMAPSS FD001); PSI/KS drift detection triggers a SageMaker-backed self-healing retraining loop, eliminating all manual intervention
- Boosted predictive accuracy by 58% (R^2 : 0.57 $\rightarrow$ 0.90, MAE: 9.3 cycles) via 112-feature rolling window engineering; orchestrated zero-downtime champion/challenger promotion via AWS Step Functions (SageMaker retraining $\rightarrow$ validation $\rightarrow$ live swap) with Mistral LLM drift narration

Stellwerk | *Rust, Tokio, Redis, Prometheus, LLM Inference Gateway*

Jun. 2026 – Present

- Built a high-performance multi-provider LLM inference gateway in Rust, routing requests across providers with automatic failover and load balancing
- Implemented Redis-backed semantic caching, token-bucket rate limiting, and circuit breakers to cut redundant API costs and harden reliability, with Prometheus metrics for real-time observability

CLUBS & LEADERSHIP

ACM – GenAI & ML

BITS Pilani, Hyderabad Campus

Sep. 2025 – Present

- Built a real-time face anonymizer using flow matching, achieving lower-latency inference than diffusion-based baselines by reducing required sampling steps
- Developed a resume screening tool that maps student profiles to PS-II and placement trends, generating data-driven, actionable feedback

Lynq – Agentic AI

BITS Pilani, Hyderabad Campus

Sep. 2025 – Present

- Built an agentic expense-validation system for campus events that cross-checks submitted bills against verified purchases and flags discrepancies automatically
- Integrated Crawl4AI to scrape real-time market prices and compare them against submitted receipts, automating pricing-mismatch detection

TECHNICAL SKILLS

Languages: Python, Java, C, C++, SQL

ML & AI: PyTorch, scikit-learn, XGBoost, Hugging Face Transformers, LangChain, LangGraph, LiteLLM, Stable Diffusion, LoRA, NumPy, Pandas, Matplotlib

MLOps & Cloud: Amazon Web Services (AWS SageMaker, Lambda, Step Functions), Docker, Git, GitHub, Jupyter Notebooks

Backend & Developer Tools: FastAPI, Typer, Rich, Postman, Crawl4AI

Databases: PostgreSQL, ChromaDB, Supabase