

ABBURI SAI KARTHIK

AI / GenAI Engineer | Python Backend Developer

9347981113 | abburisaikarthik@gmail.com | [linkedin.com/in/abburisaikarthik](https://www.linkedin.com/in/abburisaikarthik) | github.com/abburisaikarthik

PROFESSIONAL SUMMARY

AI / GenAI Engineer and Python Backend Developer with 2+ years of experience building production RAG systems, GPU/HPC platforms, secure Django and Flask microservices, and Kubernetes infrastructure. Delivers measurable outcomes across retrieval quality, GPU utilization, API security, observability, and cloud-native execution.

TECHNICAL SKILLS

AI / GenAI: RAG, LangGraph, LangChain, vLLM, Ollama, Qdrant, Memgraph, PyTorch, TensorFlow, Kubeflow
Python Backend: Python, Django, FastAPI, Flask, REST APIs, microservices, PostgreSQL, SQL, Redis, RabbitMQ, Celery
Platform & Cloud: Kubernetes, Docker, AWS, GPU/HPC, CUDA, MLPerf, Prometheus, Grafana, CI/CD, Ansible, Git

EXPERIENCE

- Software Developer** | CorespanSystems, Hyderabad

May 2024 - Present

 - Engineered a centralized Django reverse proxy for 15+ microservices and frontend traffic, with MFA, JWT session management, middleware-based RBAC, and zero-downtime schema migrations across 3+ environments.
 - Built a production RAG assistant with LangGraph, Memgraph graph retrieval, Qdrant vector search, and Kubeflow ingestion/evaluation pipelines; improved retrieval precision by 25% against a golden dataset.
 - Automated AI/ML HPC benchmarking with Django, vLLM, and MLPerf across 10+ training and inference workloads, capturing TTFT, TFLOPs, and GPU utilization.
 - Orchestrated RabbitMQ job queues and PostgreSQL persistence with Grafana/Prometheus monitoring; enabled Kubernetes-based multi-GPU, multi-node workloads and improved GPU utilization by 30%.
 - Developed a high-concurrency Flask telemetry microservice streaming real-time metrics with sub-200 ms latency across 10+ concurrent monitoring sessions.
- AWS Cloud Intern** | Swabhav Tech Labs, Remote

Oct-Dec 2021

 - Built a serverless application using AWS Lambda, API Gateway, and S3; used EC2, IAM, and DynamoDB for secure cloud infrastructure and deployment workflows.

PROJECTS

- rag-drift-monitor** | Python, FastAPI, Qdrant, Celery, Redis

2025

 - Built a FastAPI service for scheduled RAG evaluation using precision/recall, Z-score, and CUSUM drift detection; added Celery scheduling, webhook/email alerts, reports, CLI tooling, and pluggable vector backends.
- django-migration-auditor** | Python, Django, PostgreSQL

2025

 - Published a fail-safe Django package that audits apply/unapply operations, schema diffs, executor identity, duration, and failures; generates rollback SQL, reporting commands, and a read-only admin interface.

CERTIFICATIONS

AWS Certified Cloud Practitioner | [Credential](#)

EDUCATION & PUBLICATIONS

B.Tech, Computer Science and Artificial Intelligence - Amrita Vishwa Vidyapeetham | CGPA 8.7/10 | 2020-2024
3 IEEE publications covering PySpark/AWS review analytics, CNN-LSTM and Vision-GPT image captioning, and yoga pose recognition (2023-2024).