Mohammed Omar **Khan**

Platform Engineer

Abu Dhabi, UAE

(+971) 56 286 1951 | omar@itsmokha.com | www.itsmokha.com | itsmokha | m-omarkhan

Summary

Systems engineer with experience across infrastructure, data engineering and ML, currently running production platforms inside an air-gapped UAE government environment: OpenShift, Nutanix, Linux/Solaris servers and a centralised Prometheus observability stack. Outside work, I build and operate a self-hosted Kubernetes platform that serves open-weight LLMs through vLLM behind an OpenAI-compatible API, with Prometheus/Grafana monitoring, persistent storage and GitOps-style tooling. Focused on production LLM inference: latency, throughput, GPU utilisation and reliability.

Projects

Local GenAI Stack - LLM Serving, RAG, and Agentic Orchestration

vLLM, TGI, Ollama, Kubernetes, LangChain, LangGraph, Chroma

- Deployed Qwen, DeepSeek and LLaMA-family open-weight models fully offline on a self-managed Kubernetes cluster, benchmarking vLLM, Hugging Face TGI and Ollama as inference runtimes.
- Exposed an OpenAI-compatible inference API through vLLM and configured tool calling with automatic tool choice and a Hermes-style parser for agentic applications.
- Built LangGraph orchestration workflows with stateful multi-step execution, tool selection, conditional routing and handoff patterns.
- Benchmarked GGUF, GPTQ and AWQ quantized models across vLLM and Ollama, measuring TTFT, tokens per second, latency and GPU memory usage to select deployment configurations for constrained GPU resources.
- Built a LangChain RAG pipeline over a Markdown knowledge base using local embeddings and Chroma, tuning chunking, retrieval top-K and prompt construction against context-window constraints.
- Tuned vLLM serving parameters including max context length, GPU memory utilisation and concurrent request limits to balance KV cache capacity against throughput.

Self-Hosted Kubernetes Platform

Kubernetes, Proxmox, Helm, Prometheus, Grafana, Loki, Tailscale

- Built and operate a 3-node kubeadm Kubernetes cluster across Proxmox and bare-metal nodes, scheduling GPU inference workloads alongside file sync, document management and Forgejo.
- Designed the storage layer with dynamic PV provisioning through CSIs, CloudNativePG for stateful Postgres workloads, and backups.
- Implemented cluster networking and access control, along with Tailscale to manage remote connectivity to the cluster applications
- Deployed full-stack observability with kube-prometheus-stack, Grafana, Loki and Alloy for metrics and logs across cluster and application workloads.

Advertisement Detection (Final-Year Capstone)

PyTorch, TensorFlow, Scikit-learn, NLP

- Personally implemented and trained classifiers on a 1M+ row URL dataset assembled from multiple sources
- Engineered NLP-based features from URL structure to detect ad links.
- Compared neural networks, XGBoost, SVMs, LightGBM and Random Forests to find the best-performing model and reached 80% classification accuracy.

Sentiment Analysis - Amazon Reviews (227k) and Twitter (10k+)

TensorFlow, Scikit-learn, NLP, Pandas

- Two end-to-end NLP pipelines: ETL, text preprocessing, feature engineering, model training, evaluation and visualisation.
- Achieved 90% classification accuracy on Amazon and 92% on Twitter by combining neural networks with classical methods (XGBoost, SVMs, Random Forests) and hyperparameter tuning.

vscan - Open-source Python security scanner

Python, REST APIs, PyPi, Github CI/CD pipelines

- Project lead on a Python-based virus scanner that uses different APIs to scan files and websites for malicious content.
- Built a GitHub Actions CI/CD pipeline that runs tests on every pull request and automatically builds and publishes versioned releases to PyPi on tagged commits.

Experience

Midis Group
Abu Dhabi, UAE

Systems Engineer - Deployed onsite to a UAE government agency
March 2024 - Present

- Operate as the onsite technical point of contact within a strictly air-gapped government environment, translating operational requirements into deployed solutions.
- Designed, deployed and operate centralized Elasticsearch + Kibana observability used by stakeholders for operational visibility and troubleshooting.
- Used N8N to automate workflows to improve efficiency in the workplace by at least 20%.
- Administer Nutanix and Red Hat OpenShift platforms, mixed Linux/Solaris servers, GPU-equipped infrastructure and a local Elastic Cloud Enterprise environment.
- Author Python and Bash automation for operational data and reporting workflows, and use Ansible for infrastructure automation.
- Coordinate with vendor teams and debug issues across OS, application, network and platform layers

MDS TSS (Midis Group)
UAE

Big Data Intern
May 2023 - March 2024

- Completed Hadoop training and contributed to customer big-data engagements (cluster deployment exercises, pipeline tuning).
- Authored Python automation for customer data-mining and analysis workflows, where I produced operational reports on ticket throughput and SLA performance for stakeholders.

Heriot-Watt University
Dubai, UAE

Teaching Assistant
Sept. 2022 - May 2023

- Mentored 50+ Computer Science students in Python, machine-learning algorithms, data visualisation, Jupyter and ML tooling through regular lab sessions.
- Assisted faculty across undergraduate Computer Science modules, including AI/ML-focused courses.

Skills

Agentic AI & LLMs	LangGraph, LangChain, tool-calling / Hermes parser, MCP, RAG, Chroma vector store, local embeddings, Qwen / DeepSeek / LLaMA
LLM Serving	vLLM, TGI, Ollama, OpenAI-compatible APIs, SSE streaming, continuous batching, PagedAttention, KV cache, AWQ, GPTQ, GGUF
ML Frameworks	PyTorch, TensorFlow, Keras, Scikit-learn, XGBoost, LightGBM, SVMs, NLP, hyperparameter tuning
Software & APIs	Python, Flask, FastAPI, REST APIs, microservices, PyPi packaging, Git
Infra & Kubernetes	Bare-metal Kubernetes, Helm, CSI storage, Docker, Proxmox, Red Hat OpenShift, Nutanix, Ansible, OpenShift Pipelines, NVIDIA GPU Operator
Observability	Prometheus, Grafana, Loki, Alloy, kube-prometheus-stack, Elasticsearch, Kibana
Operating Systems	Linux (RHEL, Rocky, Ubuntu, Debian), Solaris, bash scripting
Programming	Python, SQL, Bash, C, C++, Java, JavaScript
Languages	English, Hindi, Urdu

Education

Heriot-Watt University
Dubai, UAE

B. Sc.(Hons) Computer Science with Artificial Intelligence
Sep. 2020 - June 2023

- Upper Second Class Honours (2:1) - Awarded the Deputy Principal's Award for academic excellence.
- AI-focused coursework: neural networks, supervised and unsupervised machine learning, applied text analysis (NLP), and computer vision foundations.
- Final-year capstone in ML-based advertisement detection.

Certificates

2024 **Data Science A-Z,** SuperDataScience
2023 **Azure Data Fundamentals,** Microsoft
2023 **Career Essentials in Data Analysis,** Microsoft
2021 **Certified Junior Associate – Java,** Oracle