SAMEET ASADULLAH

+923194598889 ◇ Islamabad, Pakistan
sameetassadullah744@gmail.com ◇ linkedin.com/in/sameet-asadullah ◇ github.com/SameetAsadullah

EXECUTIVE SUMMARY

AI/ML Engineer focused on building and deploying production-grade systems across **LLMs, RAG, semantic retrieval, generative AI, NLP, and computer vision**. Delivered scalable products used by millions, improved retrieval relevance by **35–45%**, and worked on infrastructure supporting up to **20M daily requests**. Hands-on across the full lifecycle of systems, from backend APIs and model serving to vector search, observability, and deployment, using **Python, FastAPI, Docker, Kubernetes, Triton, AWS, GCP, Elasticsearch, and PostgreSQL**. Brings a practical, product-minded approach to engineering with a strong focus on performance and reliability.

EDUCATION

The University of Adelaide
Master of Artificial Intelligence and Machine Learning

National University of Computer and Emerging Sciences
Bachelor of Computer Science

EXPERIENCE

AI Engineer Jul 2024 - Jan 2026
Stealth Startup Sydney, Australia

- Designed and rolled out a **vector-based search engine** with **Elasticsearch and transformer embeddings**, making product search far more accurate and user-friendly (around **35–45%** improvement in relevance).
- Integrated an **LLM-driven recipe feature** powered by **RAG**, where users enter a dish name and instantly get a recipe, with all required ingredients semantically matched to supermarket products and added straight to the cart.
- Built **Web Scrapers for Coles, Aldi, IGA, and Woolworths** that keep thousands of product listings up to date in real time, ensuring complete and fresh product catalogs.
- Created a **product normalization pipeline** that cleaned up messy product names and generated embeddings, which improved both search ranking and product grouping.

Machine Learning Engineer Feb 2025 - Jul 2025
Add Life Technologies Adelaide, Australia

- Optimized real-time body tracking application by integrating **MediaPipe** with Unity and fine-tuning performance for mobile deployment, resulting in a **40%** reduction in processing latency on mid-range devices.
- Led the end-to-end development of a cross-platform AI-powered mobile app, by building backend APIs with **FastAPI** and front-end in **Flutter**, enabling seamless real-time video stream handling.
- Improved system reliability and scalability by conducting stress tests, resolving memory leaks, and implementing asynchronous data handling, which stabilized performance under **10+ concurrent video streams**.

Machine Learning Engineer Aug 2022 - Jan 2024
Vyro Wyoming, United States

- Developed a bespoke serving architecture for *ImagineArt*, the second most popular AI Art Generator in the US, using **FastAPI** and **Docker**. It supports over 30 ML models, efficiently manages millions of daily requests with peak performance of **20 million** requests per day and a **99.5%** uptime, all deployed on a **Kubernetes** cluster.
- Deployed machine learning models for *Phototune* (**10M+ downloads**). Leveraged **Triton Infrastructure**, optimizing for an impressive user response time of only **2-3 seconds**.

- Implemented a **Docker-based Serverless Architecture** for the *AvatarMe* app using **Runpod** and **AWS (ECS, S3)**, enabling runtime ML model training for creating Avatars and **reducing response time from hours to 15 minutes**.
- Designed **CI (Continuous Integration) GitHub Actions** pipelines for feature integration testing, achieving **97%** test coverage. This robust testing strategy resulted in a **95%** reduction in production errors, ensuring high reliability and efficiency in deployment processes.
- Engineered a **Python** based **SDK** to host any **Stable Diffusion** workflow in production, resulting in a **80%** reduction in feature hosting through enhanced reusability.

PROJECTS

ApplyGraph: Built a **session-based agentic AI job copilot** using **FastAPI, LangGraph, Streamlit, OpenAI/Gemini APIs, PostgreSQL + pgvector, and SQLAlchemy** to analyze job fit, tailor resume content, draft outreach, and persist **semantic memory** across multi-session chat threads. Implemented **LLM routing, guardrails, SSE-based workflow streaming, custom evaluation harnesses, human feedback capture, and OpenTelemetry/Prometheus/Grafana observability** for production-style **LLMOps** and real-time AI-assisted career workflows. *GitHub Repository*

MergeWise: An AI powered pull request reviewer combining **RAG** with **FAISS** based context retrieval, **OpenAI LLM** reasoning and **GitHub Checks**. Supports both inline execution and **Celery/Redis** queue processing, built with a **FastAPI** backend, structured logging and fallback mechanisms. Fully production ready and hosted on **Render**. *GitHub Repository*

AutomateIt: An advanced Home Automation System leveraging **Deep Learning**, including **CNN (Convolutional Neural Network)**, **React Native**, and **MongoDB**, boasting over **85% accuracy** in recognizing and responding to **Urdu Voice Commands**. This versatile system is adept at interfacing with various household appliances, ensuring a user-friendly and efficient home management experience. *GitHub Repository*

Temporal Financial Forecasting: A deep learning project for **time series forecasting** of financial market data using **RNNs, GRUs, and LSTMs**. Implemented a robust pipeline for multi-horizon prediction, hyperparameter optimization, and model evaluation, achieving superior performance with GRU models (e.g., RMSE ≈ 0.024). The system predicts multiple features such as **Open, High, Low, Close, and Volume**, providing accurate insights for financial trend analysis. *GitHub Repository*

CNN Benchmark Suite: A modular, **PyTorch**-based deep learning framework for benchmarking multiple **CNN architectures** (e.g., **ResNet-18**, **MobileNetV2**, **GoogleNet**, **AlexNet**) across classification tasks. Supports **grid search** over optimizers (**SGD**, **Adam**), learning rates, and batch sizes, with **data augmentation**, stratified validation splits, automated metric logging, and model selection. Designed for scalable, reproducible experiments and real-world transferability. *GitHub Repository*

Clinical Risk Prediction: An end-to-end **binary classification pipeline** built with **PyTorch** for **clinical risk detection** from patient diagnostics. Features include **data preprocessing**, class imbalance correction using **SMOTE**, optimizer benchmarking (**SGD** vs **Adam**), and comprehensive performance visualization (**ROC-AUC**, **confusion matrix**). Built to align with **healthcare AI best practices** and deliver interpretable model evaluation. *GitHub Repository*

TECHNICAL SKILLS

Languages	Python, C++, Java, SQL
Databases	PostgreSQL, MySQL, SQLite, MongoDB, Redis/KeyDB, Firebase, Cloudflare KV
Vector Databases	Elasticsearch, Pinecone, FAISS
Machine Learning	PyTorch, TensorFlow, Scikit-learn, OpenCV, MediaPipe, Transformers, VertexAI
GenAI	GANs, Stable Diffusion, LLMs(BERT, GPT), Retrieval Augmented Generation (RAG)
MLOps	Docker, Kubernetes, FastAPI, BentoML, Triton Inference Server, DVC, GitHub, Linux
Cloud	AWS (S3, Lambda, ECS, EKS, Sagemaker), GCP, Runpod, LambdaLabs, DigitalOcean

References Available on Request