Tauseef Ahmed Memon

Senior Machine Learning Engineer (L4) — GenAI • Computer Vision • LLMOps • Multimodal AI

Islamabad, Pakistan
+92 335 2121095
tauseefahmed.tam@gmail.com
tauseefml.vercel.app
linkedin.com/in/tauseef-ahmed-memon-15abb5364
Open to: Pakistan (onsite/hybrid) and global remote/relocation

SUMMARY

Machine Learning Engineer (L4) with 4.5+ years delivering production Generative AI and multimodal systems end-to-end: Stable Diffusion/SDXL fine-tuning, PEFT (LoRA/QLoRA) for LLMs, multimodal RAG, and low-latency inference with vLLM/TensorRT on Kubernetes. Led delivery as project/technical lead for two production initiatives (teams of 6 and 3), mentoring engineers and running code reviews to maintain quality and velocity. Strong focus on data quality, experimentation, scalable deployment, and monitoring across social/content platforms, document AI (OCR + key–value extraction), and computer vision.

SELECTED HIGHLIGHTS

- Led teams of **6** and **3** to deliver two production AI systems: (1) social platform fact-checking + moderation pipeline scaling to ∼**5k posts/sec**; (2) custom Stable Diffusion model for game-style character generation.
- Reduced Stable Diffusion inference latency by ∼**40%** using TensorRT quantization.
- Achieved **sub-200 ms** LLM request latency with Kubernetes-based serving and vLLM.
- Improved detection precision by **+12 mAP** using SAM + YOLO multimodal segmentation.
- Accelerated 3D asset creation by **4**× using 3D Gaussian Splatting and PyTorch3D.
- Reduced training-data ETL time by **35%** via automated Airflow pipelines.
- Cut false positives by **25%** using drift/quality monitoring (Evidently AI / NannyML).
- Achieved **93.2%** accuracy for liveness detection in Pakistani CNIC verification (Seldon Core).

TECHNICAL SKILLS

- **Languages:** Python; SQL (basic)
- **ML / Deep Learning:** PyTorch; TensorFlow; Keras; scikit-learn
- **GenAI / Agents / RAG:** Hugging Face Transformers; PEFT (LoRA/QLoRA); LangChain; LangGraph; multimodal RAG; agent/tool orchestration; vLLM; Stable Diffusion/SDXL; TensorRT
- **Search + Enrichment APIs:** Apollo; RocketReach; Hunter.io; Serper (Google Search API); web scraping
- **Computer Vision / 3D:** OpenCV; YOLO; SAM; Detectron2; OpenPose; PyTorch3D
- **MLOps / Deployment:** Docker; Kubernetes; Kubeflow; FastAPI; MLflow; Seldon Core; Terraform
- **Data / ETL:** Pandas; NumPy; Dask; Apache Beam; Airflow; Prefect
- **Monitoring / Quality:** Evidently AI; NannyML
- **Cloud:** AWS; GCP
- **Document AI:** Tesseract; EasyOCR; Textract; PyMuPDF (fitz); key–value extraction
- **Product Integrations:** Chatwoot

EXPERIENCE

Red Buffer Jul 2023 – Present

Machine Learning Engineer (L4)

- Led two delivery teams (**6** and **3**) as technical/project lead; mentored engineers and conducted code reviews to maintain quality, readability, and release velocity.
- Led an end-to-end social platform pipeline for **fact verification** and **content moderation**, designed to handle ∼**5k posts/sec** throughput; integrated LLM-based claim extraction with retrieval-based verification and custom moderation rules.
- Led development of a **custom Stable Diffusion model** fine-tuned for a game art studio to generate **game-style characters**; productionized generation workflows for consistent stylistic output.
- Architected Stable Diffusion/SDXL fine-tuning pipelines; applied TensorRT quantization, reducing inference latency by ∼40%.
- Deployed Kubernetes-based LLM inference endpoints with vLLM, achieving sub-200 ms request latency for LLM queries.
- Designed multimodal segmentation workflows integrating SAM + YOLO, improving detection precision by

+12 mAP.

- Engineered 3D asset generation using 3D Gaussian Splatting and PyTorch3D, accelerating asset creation by 4×; maintained pose consistency using OpenPose/Detectron2.
- Orchestrated end-to-end MLOps for automated invoice processing (RAP) with Terraform-managed IaC; standardized key–value extraction and improved downstream data usability by ∼40%.

Sabhi — Identity for All Jan 2021 – Jul 2023
Machine Learning Engineer

- Deployed canary-released liveness detection models (EfficientNet/YOLO) using Seldon Core, achieving 93.2% accuracy for Pakistani CNIC verification.
- Engineered automated ETL pipelines using Airflow, reducing training data preparation time by 35%.
- Integrated drift/quality monitoring (Evidently AI), cutting false positives by 25%.

SELECTED PROJECTS

- **Social Platform: Fact Checks + Moderation + Recommendation** — Led end-to-end system using LLM-based claim extraction, retrieval-backed verification, and custom moderation rules; designed for ∼5k posts/sec throughput.
- **Game Art Studio: Custom Stable Diffusion Characters** — Fine-tuned custom Stable Diffusion model to generate game-style characters with consistent stylistic output; productionized inference workflow.
- **Sales Pipeline Automation (Multi-agent, LangGraph)** — Built two end-to-end outbound automation systems (team of **3**; led team of **5** to completion) in Python with LangGraph-based multi-agent orchestration. Input company name + URL → generate company report + ICP; discover and enrich best-fit accounts/leads (Apollo, RocketReach, Hunter.io); exclude competitors; generate reports for matched accounts; prioritize prospects; draft personalized outreach emails.
- **Customer Care RAG Chatbot** — Multilingual (English/Tagalog/Taglish) chatbot using LangChain; multimodal inputs; 10-message memory; Chatwoot integration; live-agent escalation for complex/frustrated cases.
- **RAP — Automated Invoice Processing** — OCR-driven pipeline (Tesseract/EasyOCR/Textract) in Python/Pandas; standardization and aggregation to Excel; robust error handling.

EDUCATION

MS, Computer Science (Graphics and Computer Vision) — National University of Sciences and Technology (NUST), 2018–2020
BE, Electrical and Electronics Engineering — Mehran University of Engineering and Technology (MUET), 2014–2017

CERTIFICATIONS

Generative AI Specialization (GenAI)
Machine Learning Engineering for Production (MLOps)

PUBLICATIONS

"Brain Tumor Detection from MRI Images by Fusion of Region Growing and Edge Detection Algorithms" — ESTIRJ