

Syed Muhammad Sarim Ahsan

+92 323 703 4858 | sarim.ahsan101@gmail.com | [GitHub](#) | [LinkedIn](#) | [HuggingFace](#) | [Medium](#) | [Portfolio](#)

RESEARCH SUMMARY

AI Research Engineer and Computer Engineering undergraduate (GPA 3.50/4.00, SSUET) working at the intersection of LLM interpretability, preference alignment, model distillation, and GPU inference systems. Independent research spans residual-stream linear probing for hallucination prediction, DPO vs. PPO preference optimization, hidden-state distillation, and speculative decoding. Open-source maintainer with 15+ public model checkpoints, 3 datasets, and reproducible code repositories on HuggingFace and GitHub, alongside a technical publication series demystifying transformer internals for practitioners.

EDUCATION

Sir Syed University of Engineering & Technology

B.S. in Computer Engineering · GPA: 3.50 / 4.00 · 6th Semester

Beaconhouse School System

Higher Secondary School · Computer Science · A Grade

Karachi, Pakistan

2023 – Expected Oct 2027

Karachi, Pakistan

2009 – 2021

INDUSTRY EXPERIENCE

Software Engineer

Global Integration System (GIS)

- Engineered and maintain [Portalersm](#), a production student recruitment platform serving 600+ active users across the UK.
- Architected scalable full-stack features, Node.js backend services, secure role-based access control (RBAC), Docker containerization, and zero-downtime CI/CD deployment pipelines using GitHub Actions.
- Translated machine learning systems into high-availability production web infrastructure.

2025 – Present

London, UK (Remote)

AI Research Engineer & Open-Source Contributor

Independent Research

- Conducted independent empirical research on LLM residual stream activations, preference alignment loss dynamics (DPO vs. PPO), hidden-state knowledge distillation, and speculative decoding algorithms.
- Authored and released 15+ fine-tuned checkpoints, 3 public datasets, and benchmark suites on HuggingFace ([sarimahsan101](#)) and GitHub.

2024 – Present

Karachi, Pakistan

Machine Learning Engineer Intern

FlyRank

- Developed applied machine learning solutions in a fast-paced engineering environment, contributing to production-oriented ML systems, model training, evaluation metrics, and scalable AI workflows.

Jun 2026 – Jul 2026

Remote

Machine Learning Engineer Intern

Evello Pathways

- Designed end-to-end machine learning pipelines across regression, classification, clustering, recommendation systems, time-series forecasting, and computer vision using PyTorch, Scikit-learn, FastAPI, Docker, MLflow, and YOLOv8.

May 2026

Remote

IT Intern

IFFCO Pakistan

- Audited and documented over 50 enterprise hardware assets, maintaining infrastructure records, inventory management systems, and enterprise technical support operations.

2025

Karachi, Pakistan

RESEARCH PAPERS & MANUSCRIPTS

Cross-VM Domain Shift in HPC-Based Intrusion Detection

Manuscript Completed – Under Conference Review

- Conducted a novel empirical study investigating cross-virtual-machine domain shift in Hardware Performance Counter (HPC)-based intrusion detection systems in cloud environments.
- Curated and released **hpc-dataset** (60,000 labeled HPC samples across virtual machine environments) and systematically evaluated **18** machine learning and deep learning models (LSTM, BiLSTM, Random Forest, SVM).
- Demonstrated significant performance degradation under unseen cross-VM deployment conditions, proving that single-VM baselines substantially overestimate real-world IDS generalization and highlighting the necessity for domain-robust security models.

Activation Function & Normalization Layer Interaction Dynamics | [Github](#) Empirical Research Paper – Manuscript Completed

- Executed a comprehensive ablation study analyzing interaction dynamics between non-linear activation functions (**GELU**, **ReLU²**, **SwiGLU**) and layer normalization methods (**LayerNorm**, **RMSNorm**) in decoder-only language models.
- Evaluated architectural pairings across loss convergence speed, gradient variance, training stability, and downstream perplexity, proving how SwiGLU paired with RMSNorm stabilizes optimization and yields superior perplexity scaling.

LLM Interpretability, Alignment & Distillation

- Hidden-State Hallucination Probing** | [Github](#) | [HuggingFace](#) | [Dataset](#) 2026
- Trained linear probes on intermediate residual stream activations across Qwen2.5 (1.5B, 3B, 7B) to predict factual hallucination prior to token generation, outperforming output-confidence baselines (+0.101 AUC gain at 7B).
 - Localized factual retrieval signals primarily to middle transformer layers (layers 14–22) and published the 15k-sample probe dataset and trained probes on HuggingFace for reproducibility.
- Hidden-State & Logit Knowledge Distillation** | [Github](#) | [HuggingFace](#) 2026
- Built a memory-efficient knowledge distillation framework combining KL-divergence logit matching with projection-based hidden-state alignment transferring representations from Qwen2.5-3B-Instruct to Qwen2.5-0.5B-Instruct.
 - Executed the training pipeline on Kaggle dual-T4 GPUs using 4-bit quantization and LoRA adapters, releasing distilled student weights and LoRA variants on HuggingFace.
- Preference Optimization for LM Detoxification** | [HuggingFace](#) 2026
- Empirical comparison evaluating Direct Preference Optimization (DPO) against Proximal Policy Optimization (PPO) for non-destructive toxicity reduction across GPT-2 Medium (355M) and Qwen2.5-0.5B.
 - Demonstrated that DPO achieves an 84.2% reduction in toxic outputs with $3.2\times$ faster training convergence and superior perplexity preservation compared to RLHF PPO baselines.
- Speculative Decoding for LLM Inference Acceleration** | [Github](#) 2026
- End-to-end Python implementation and CLI benchmark suite for speculative decoding with rejection sampling (reproducing Leviathan et al. & Chen et al.), evaluating lookahead k -value sweeps on Qwen2.5 draft-target pairs.

Transformer Architectures & Low-Level Systems (From Scratch)

- Nova-141M: Modern Decoder Transformer** | [Github](#) | [HuggingFace](#) 2026
- Designed and implemented a 141M-parameter decoder-only transformer from scratch in PyTorch incorporating Grouped-Query Attention (GQA), RoPE, RMSNorm, and SwiGLU; trained end-to-end on TinyStories (2.1B tokens).
- GPT-2 Inference Engine** — Pure NumPy Implementation | [Github](#) 2026
- Re-implemented GPT-2 (124M) inference with zero PyTorch in the forward pass — handwritten attention, LayerNorm, SwiGLU, and nucleus sampling in pure NumPy loading pre-trained HuggingFace weights directly.
- Mini-Qwen** — Autoregressive LM from Scratch | [Github](#) | [HuggingFace](#) 2026
- 56M-parameter Qwen-inspired model in PyTorch — GQA, RoPE, SwiGLU, RMSNorm, a custom AdamW optimizer, and top- k /top- p sampling; trained end-to-end on TinyStories.
- Custom CUDA Kernels** | [Github](#) 2026 – Present
- Handwritten CUDA C++ kernels analyzing GPU thread hierarchy and memory bandwidth: tiled matrix multiplication with shared memory ($\sim 2.8\times$ speedup over naive baseline) and a fused softmax kernel.
- MyOS** — 32-bit x86 Operating System Kernel | [Github](#) 2025
- Developed a 32-bit x86 operating system in C and Assembly featuring a custom bootloader, protected mode initialization, interrupt service routines (ISRs), physical memory management, device drivers, and an interactive shell.

AI Systems, Agentic RAG & Production Platforms

- ArXiv Lens** — RAG Citation Knowledge Graph Assistant | [Github](#) 2026
- RAG research platform that ingests ArXiv paper PDFs, constructs citation knowledge graphs, and enables conversational paper analysis via FastAPI, LangChain, HuggingFace, and React.
- MeezanGPT** — Corrective RAG (CRAG) Chatbot | [Github](#) 2026
- Domain-specific banking assistant incorporating Corrective Retrieval-Augmented Generation (CRAG) with an active retrieval evaluator layer to eliminate hallucinations on financial domain queries.
- FlowForge AI** — Sales Automation Multi-Agent SaaS | [Github](#) 2026
- Enterprise sales automation platform using multi-agent LLM workflows to generate personalized outreach campaigns, score leads, and recommend objection-handling strategies using React, FastAPI, and Docker.
- Chronicle** — Developer Intelligence Knowledge Graph | [Github](#) 2026
- Developer intelligence tool transforming GitHub commits, pull requests, and issues into a Neo4j knowledge graph, enabling natural language exploration of repository commit history.
- ArgoFarm** — Multimodal Agricultural Decision Platform | [Github](#) 2026
- AI decision support platform for farmers offering crop disease diagnosis, weather intelligence, and actionable recommendations using computer vision and multimodal foundation models.

TECHNICAL SKILLS

Programming Languages: Python, C, C++, CUDA C++, SQL, JavaScript, TypeScript, x86 Assembly
ML & Deep Learning: PyTorch, HuggingFace Transformers, LoRA / QLoRA, DPO, PPO, GRPO, Linear Probing, Knowledge Distillation, Speculative Decoding, Scikit-learn, YOLOv8
LLM Systems & GPU Serving: CUDA C++ (custom kernels), vLLM, AMD MI300X, ROCm, Multi-GPU Training (Kaggle dual-T4), Quantization (4-bit / 8-bit), FlashAttention
Agentic AI & Retrieval: LangGraph, LangChain, RAG, Corrective RAG (CRAG), Knowledge Graphs, FAISS, Vector DBs, Neo4j
AI Infrastructure & MLOps: Docker, FastAPI, Node.js, React, Git, GitHub Actions (CI/CD), MLflow, Linux, MySQL

PUBLICATIONS & TECHNICAL WRITING

- [Nucleus Sampling: Why Your LLM Doesn't Always Pick the "Best" Word](#) *Jul 2026*
- [Temperature \$\neq\$ Creativity: Stop Turning That Dial](#) *Jun 2026*
- [Prompt Caching vs. Context Window — and Why Most Devs Get This Wrong](#) *Jun 2026*
- [The Context Window Lie](#) *Jun 2026*
- [Building a GPT-2 Inference Engine From Scratch](#) *May 2026*
- [Why LLMs Can't Remember: Episodic, Semantic & Procedural Memory Explained](#) *Apr 2026*

OPEN RESEARCH ARTIFACTS (HUGGINGFACE)

- [qwen2.5-7b-hallucination-probe](#) & [qwen2.5-7b-hallucination-probe-dataset](#) — Residual stream hallucination probing weights and 15k feature dataset
- [Qwen2.5-0.5B-HiddenDistilled](#) & [-LoRA variant](#) — Student checkpoints from hidden-state distillation
- [Qwen2.5-0.5B-DPO-detox](#) & [qwen2.5-0.5b-detox-ppo](#) — DPO and PPO detoxification aligned models
- [gpt2-dpo-detox](#) & [ppo-gpt2-imdb](#) — GPT-2 preference optimization benchmarks
- [nova-141m-tinystories](#) & [mini-qwen](#) — Open decoder transformers trained from scratch
- [qwen2.5-7b-thinking-esp](#) & [qwen2.5-3b-cot-gsm8k](#) — Multilingual CoT and GSM8K reasoning models
- [mistral-7b-ai-project-scoper-lora](#) & [ai-project-scoping-instructions](#) — LoRA scoper model & dataset
- [TinyLlama-1.1B-Alpaca-QLoRA](#) & [os-network-intrusion-detector-bilstm](#) — Edge quantization model & BiLSTM security classifier
- [hpc-dataset](#) — 60K-sample hardware performance counter dataset for cross-VM intrusion research

RESEARCH-RELEVANT CERTIFICATIONS

- [Reinforcement Fine-Tuning LLMs with GRPO](#) — DeepLearning.AI *May 2026*
- [Fine-Tuning & RL for LLMs: Intro to Post-Training](#) — DeepLearning.AI *Jun 2026*
- [Transformers in Practice](#) — DeepLearning.AI *May 2026*
- [Foundation: Introduction to LangGraph](#) — LangChain *Mar 2026*
- [LLM Serving & Inference with vLLM and MI300X](#) — AMD AI Academy *Apr 2026*
- [Machine Learning Specialization](#) — Stanford / DeepLearning.AI (Andrew Ng) *Nov 2025*
- [Data Scientist Associate](#) — DataCamp *Jun 2025*