Shabd Sahni

+91 8448224980 | shabdsahni2005@gmail.com | linkedin.com/in/shabdsahni

Profile

Electrical + CS Engineering undergraduate (Rank 1, CGPA 9.39/10) specializing in **Large Language Models**, **Multimodal AI**, and **Edge-Optimized Inference**. Experienced in *fine-tuning and deploying* large-scale LLMs on accelerators, integrating ML into *high-performance systems*, and developing award-winning *AI applications for social good*.

Passion Areas: cutting-edge AI/ML, particularly Large Language Models, adaptable high-performance scalable inference, brain computer interfaces and human-centered AI.

Education

Dayalbagh Educational Institute, Agra *Bachelor of Technology* *2023 – 2027*
Major: **Electrical Engineering** Minor: **Computer Science** **Rank: 1/328** (Engineering) CGPA: 9.391/10
Co-Curriculars: Social Service (NSS), Debate Team, Sitar (Indian Classical), Local *sustainable farming community*

Apeejay School, Noida *High School Diploma* *2009 – 2023*
Co-Curriculars: **President - Code Club** *2023*, **President - Robotics Club** *2022*,
Head Boy - Student Council *2018, 2015*

Work Experience

LLM Infra Intern – Gatespeed, Pleasanton CA (Intel AI Partner Alliance Member) [*Aug - October, 2025*]
Multilingual RAG & High-Performance LLM Deployment

- Built **multilingual RAG** with semantic retrieval using **Qdrant**.
- Deployed optimized inference of **DeepSeek R1 Llama 70B** on **Intel Gaudi2 cluster (8×98GB HPUs)** with hybrid RTX (2x5070ti + 5090ti) acceleration via **vLLM/Ollama**.
- Fine-tuned **DeepSeek-R1-Distill-Llama-70B** on Gaudi2, achieving **73.77% accuracy, 2.72 perplexity** in 52min with **16.4M trainable parameters**; optimized GPT-2 to **26.95 samples/sec**.
- Enhanced inference with HPUgraph, KV caching, containerized scaling, and ZeRO Stage 3, sustaining **94.1GB peak memory** per HPU (bfloat16).
- Cut **e2e latency from 261s → 46s**; reached **1199 tokens/s output**, **6691 tokens/s total throughput** at **256 req/s concurrency**.
- **Tech Stack:** Intel Gaudi2 HPU, vLLM, Ollama, Qdrant DB, DeepSeek LLMs, LoRA/PEFT, GGUF quantization, Docker, Habana Frameworks, DeepSpeed ZeRO, multilingual NLP.

Machine Learning Trainee – Cadence Design Systems, Noida [*June - August, 2025*]
EDA Tool Optimization via ML – Fortune 500 Semiconductor Company

- Applied **ML-based predictive filtering** to speed up ECO optimizer runtime on **Samsung, Qualcomm and Renesas 3nm chips** with no QoR degradation. Filtered 65% wasteful evaluations on avg.
- Developed and Integrated feature engineering and training pipeline on ML models (Random Forest, LightGBM, XGBoost) into **Cadence C/C++** code for **Tempus** and **Certus** signoff tools.
- Automated log parsing, profiling, and data preparation using Bash, CSH and Tcl/Tk.
- **Tech Stack:** Python, scikit-learn, LightGBM, XGBoost, C/C++, Tcl/Tk, Shell, NumPy, Pandas, Cadence EDA tools.

Intern – MindLab, Indian Institute of Technology, Delhi [*May - June, 2025*]
LLM & Multimodal AI for Neuroscience & Therapeutic Gaming

- Prototyped **adaptive NPC dialogue systems** using fine-tuned LLMs (DeepSeek-V3-70B, Mistral-8B) with **OCEAN personality modeling** and vector-based memory.
- Implemented **weighted temporal graphs** for multi-agent knowledge diffusion with personality-filtered propagation and distributed vector databases.
- Built **FastAPI middleware** around **OpenRouter** with vanilla JS/HTML test UI, integrating GQ-6 metrics for therapeutic RL optimization targeting gratitude learning.
- **Tech Stack:** FastAPI, OpenRouter API, Ollama, DeepSeek-V3-70B, Mistral-8B, Python, Qdrant, temporal graph algorithms, vanilla JS/HTML.

Intern (Systems & Networking Engineering) – Omnitech Solutions, San Jose CA [*May 2024 – July 2024*]
High-Performance Networking, Linux/BSD, Virtualization

- Setup NGINX media streaming server on **Intel Xeon Gold CPUs** and high-speed network cards **(e810-cqda2, xxv710)** and ran low-latency media streaming protocols.
- Deployed Linux VMs with KVM/QEMU, virtual networks, and utilized DPDK and custom kernel devive drivers for high-speed packet processing.

Intern (DevOps & Backend Systems) – QuditBrain, Noida [*June 2023 – August 2023*]
Backend APIs, Server-Side Systems

- Managed VPNs (WireGuard), and Linux VMs. Developed Python backend APIs and deployed self-hosted services like Nextcloud.
- Managed AWS infrastructure including EC2, S3, Lambda, DynamoDB, SQS, Route 53, Cognito, and CloudWatch; automated tasks via Boto3 and AWS APIs.
- Built and deployed Python backend services, secured access-control workflows, and integrated VPNs (WireGuard) with AWS IAM–aligned policies.
- Developed monitoring and logging pipelines; improved backend reliability across distributed systems.

Projects and Research

deGuppe | *TOR, Blockchain (hybrid private/public), Python, WebSockets, SQLite*

- Developed a decentralized, peer-run, real-time communication system over **TOR** with **hybrid blockchain storage**.
- Presented at the 47th National Systems Conference organized by the Systems Society of India, winning **Best Poster Presentation** Award.
- Secured **International Soonami Cohort 3 funding** and won **Best Project in Web3/AI for Good** and Third Prize at IITD Tryst Track, Best Live Demo.

Gam-i-yog | *Python, OpenCV, MediaPipe, PyTorch, Decision Trees, K-means Clustering, Multi-modal GenAI*

- **Live Pose Classification** and **Multimodal** Feedback Using Generative AI.
- Developed an ensemble with a dynamically trainable pose classifier for temporal body landmarks, K-means clustering for posture quality analysis and GenAI for personalized feedback.
- Presented at the DSC Conference Winter Session, organized by University of Waterloo, CAU Kiel, Western University, and University of Birmingham, winning **Best Poster Award**.
- Grand finale winners of **National Toycathon Hackathon** and *received funding from the Government of India* for its impact on Yoga practitioners, including accessibility features for injuries and disabilities.

ESG-Financial Performance Research | *Python, Bloomberg Terminal, Statistical Analysis, Data Imputation*

- Conducted empirical analysis exploring **ESG score correlation with firm financial performance** through comprehensive literature review and statistical modeling.
- Gathered ESG data from **Bloomberg terminals** and developed **imputation algorithms for missing ESG data** to enhance corporate accountability metrics.
- Research published in **Springer Cureus** journal through collaboration with **ICASSSD Summer School 2024**.

Abhinandan (War Robot) | *Combat Resilient Robot Design, Mechatronics Engineering, CAD*

- **First Prize at IITD**'s *National* Tryst RoboWars. Designed a non-violent defensive battlebot, with a frugal and disruptive attack resilient design.
- Designed actuator and drivetrain systems with shock-resilient mechanics, impact-damping, and closed-loop motor control for stable maneuvering under collision stress.
- *First junior high school team* to become winners among funded collegiate teams from across the country.

Pehchaan | *Python, PyTorch/TensorFlow, EdgeFace, OpenCV, NumPy, Raspberry Pi, Embedded Linux.*

- Developed an edge-optimized face recognition algorithm, utilizing EdgeFace model (LoRaLin distilled layers)
- Deployed on edge devices like **Raspberry Pi** used in security turnstiles and classroom attendance.

Technical Skills

Programming, Frameworks & Development: Python, C/C++, Rust, Flask, FastAPI, SQL
Systems & Networking: Linux/BSD, Shell scripting, NGINX, Docker/Kubernetes, Grafana/Prometheus, Vector.dev, QEMU/KVM, WireGuard/VPNs, AWS(ec2, Lambda, S3, DynamoDB, RDS, Aurora, Route 53, Cognito, Boto3 and AWS APIs), Wireshark, Nmap, OpenSSL
ML & Computer Vision: TensorFlow, PyTorch, OpenCV, MediaPipe, EdgeFace, K-Means, Ollama, vLLM, Unsloth, multimodal LLMs, world models, vision-language-action models
Blockchain & Web3: Hybrid Blockchain, Proof of History, Reseach/Design of decentralized systems, TOR, TON blockchain, Solana