

Abhiram Sharma

LLM Infrastructure | Agentic AI Platforms | Machine Learning Engineering
sharmaabhiram1809@gmail.com • +91 79760 90859
linkedin.com/in/sharmaabhiram • github.com/abhiram1809 • abhiram1809.github.io

Experience

Panasonic Avionics

Machine Learning Engineer II (promoted from MLE I, Jun 2026)

Hybrid

Jul 2025 – Present

- Optimized custom MiniMax-M3 inference on NVIDIA RTX PRO 6000 GPUs, reaching 160 tokens/second decode at batch size 1.
- Implemented and optimized an Ideogram forward pass for vLLM-Omni, cutting image generation from roughly 2 minutes in the reference implementation to 24 seconds.
- Architected an end-to-end LLM routing service for coding agents (Claude Code, Cursor, Kilo Code), improving cost utilization 4x through request classification and model-tier routing.
- Founded and scaled the organization's LLM inference platform serving open-source and closed-source models behind a unified, OpenAI-compatible endpoint that works across LLM SDKs, increasing engineering productivity 60%.
- Built an agentic AI platform that standardized development and deployment of production-grade AI agents, reducing agent creation time from days to under 10 minutes.
- Architected a Kubernetes controller platform (Operators, CRDs) for organization-wide LLM deployments, with automated lifecycle management, budget controls, and centralized observability, cutting debugging and incident resolution time 90%.
- Mentored an engineering intern in AWS, AI-driven engineering practices, and independent production troubleshooting.

Calsoft, Inc.

Data Scientist, Consultant for Panasonic Avionics

Pune, Maharashtra

Jan 2025 – Jul 2025

- Built LLM input/output guardrails with open-source and closed-source tooling, achieving an 88% safer AI system.
- Engineered a WAF-fortified vLLM inference pipeline serving dense models at 80 tokens/second for proprietary coding workflows.

Softsensor.AI

Associate Data Scientist

Jaipur, Rajasthan

Sep 2023 – Jan 2025

- Built Text2SQL and unstructured MITRE query workflows for a cyber security chatbot, surfacing attack-pattern insights.
- Led unstructured-to-structured extraction for claims letters, emails, and SRRs with interlinked document relationships, cutting extraction cost up to 85%.
- Built end-to-end document translation software preserving tables and diagrams across document formats.
- Fine-tuned open-source LLMs with LoRA and QLoRA, and drove R&D across topic modeling, ColBERT reranking, knowledge-graph retrieval, and multimodal data ingestion pipelines.

Blackcoffer Consulting

Associate Data Scientist

Remote

Mar 2022 – Feb 2023

- Built a computer vision solution automating vehicle detection and localization in high-resolution satellite imagery.

Projects & Open Source

- vLLM on Vulkan (merged upstream):** Served Qwen3.5-2B on a legacy RX 560X at 316.60 aggregate decode tokens/second, and merged a 2.55x bit-exact CPU embedding-decode speedup plus Vulkan portability fixes covering device-feature gating, memory barriers, and FP32 GEMM fallback (ericcurtin/vllm-vulkan#101).
- vLLM MiniMax-M3 on NVFP4/Blackwell** (PR in review): Root-caused and fixed four parser bugs that garbled output, misrouted streaming reasoning tokens, and dropped tool calls on RTX PRO 6000 (vllm-project/vllm#47001).
- AnyBench (open source):** Turns repository history into coding tasks, verifies each against the historical fix in Docker, and reports how candidate models perform on your own codebase. github.com/abhiram1809/any-bench
- Merged Open-Source Contributions:** *LlamaIndex* added Google Sheets ingestion into pandas DataFrames (#11640) and embeddings for custom knowledge-graph triplets (#10629). *HuggingFace lighteval* documented the hosted_vllm provider prefix needed to reach a locally served vLLM endpoint (#1060).
- Writing:** Six engineering notes on inference internals (paged attention, Vulkan kernels, legacy-GPU bring-up, benchmark design) at abhiram1809.github.io/blogs, plus commissioned Zilliz posts on LLM hallucinations and Milvus vector search.

Technical Skills

Languages: Python, SQL, Rust, JavaScript, C++, Java, Bash

AI & Machine Learning: Large Language Models (LLMs), Agentic AI, Retrieval-Augmented Generation (RAG), Fine-Tuning (LoRA, QLoRA, PEFT), Quantization, Distributed Training, Model Evaluation, Prompt Engineering, Natural Language Processing (NLP), Computer Vision, Multimodal AI, Knowledge Graphs, Text2SQL, AI Guardrails (NeMo Guardrails)

Inference & Serving: vLLM, SGLang, Triton Inference Server, llama.cpp, Ray, CUDA, Vulkan, GPU/CPU Inference, Inference Optimization

Infrastructure & MLOps: Kubernetes (Operators, CRDs), ArgoCD, Docker, CI/CD, Prometheus, Grafana, LLM Observability, Git, Linux

Cloud & Data: AWS, Microsoft Azure, PostgreSQL, pgvector, Qdrant, Redis, Kafka, Pandas, NumPy, Data Pipelines, ETL

Frameworks & Tools: PyTorch, HuggingFace Transformers, LlamaIndex, LangChain, LangGraph, CrewAI, LiteLLM, Unsloth, FastAPI, gRPC

Education