Anmol Sharma

AI Engineer

anmolsharma.dev@gmail.com +91-84128-80194 Jaipur, India github.com/anmolsharma152

linkedin.com/in/anmolsharma152 anmolsharma152.vercel.app

PROFESSIONAL SUMMARY

AI Systems Engineer specializing in production-grade agentic pipelines, real-time inference optimization, and neuro-symbolic RAG architectures. Achieves sub-100ms voice interaction latency on CPU-only hardware via ONNX Runtime and Rust-based VAD. Builds LangGraph-driven multi-agent systems with LangSmith tracing and RAGAS evaluation harness for measurable quality regression testing. Graduate of the Minor in AI & Data Science from CCE, IIT Mandi (CGPA 8.44/10).

PROFESSIONAL EXPERIENCE

Independent Research & Development, *AI Systems Engineer* 06/2024 – Present

Self-Directed / Open Source

- Engineered nexus-bot, a LangGraph + MCP + PostgreSQL stateful autonomous agent with persistent tool registry and multi-turn episodic memory.
- Built MedPal AI, a neuro-symbolic Dual-RAG clinical decision support system routing queries through dense FAISS retrieval and a logic-gated medical execution path — 40% hallucination reduction on domain-specific queries.
- Developed a fully local emotion-aware voice assistant achieving sub-100ms barge-in latency on a dual-core i5 via Rust VAD processor (WASM/ONNX), Faster-Whisper STT, and Kokoro-ONNX TTS — zero cloud dependency.
- Architected CodexEngine v3, a multi-LLM agentic RAG with LangSmith distributed tracing and RAGAS evaluation harness for cross-version golden dataset regression testing.

Teleperformance, *Technical Support Executive* 07/2023 – 04/2024 | Jaipur, India

- Maintained CSAT 4.8/5 diagnosing complex software issues under strict SLA compliance for enterprise Japanese clients.
- Executed SQL-based root-cause analysis, translating user-reported failures into actionable technical requirements to improve tool reliability.

PROJECTS

Sarvam-1 Fine-Tuning , *Python · QLoRA · PEFT · SFTTrainer · vLLM · FastAPI · W&B*

- Fine-tuned sarvamai/sarvam-1 (3B) on a 10K multilingual corpus (Hindi/Hinglish/English) using QLoRA (4-bit NF4, LoRA r=16, α=32) targeting q/k/v/o attention projections — completion-only loss masking.
- FastAPI SSE streaming server with dual engine: vLLM async (GPU) and HF Transformers fallback (CPU); evaluation pipeline benchmarks Role Classification and Salary Bucket accuracy against base model.

CodexEngine , *Python · LangGraph · LangSmith · FAISS · PostgreSQL*

- Multi-LLM agentic RAG with LangGraph state machine, LangSmith distributed tracing, and RAGAS evaluation harness for cross-version golden dataset regression testing.

Emotion Aware Voice Assistant , *Rust · Python · ONNX Runtime · Silero VAD · Faster-Whisper · Kokoro*

- Fully local CPU pipeline: Rust VAD processor (WASM/ONNX) → Faster-Whisper STT → Groq LLM → Kokoro-ONNX TTS; sub-100ms barge-in latency on dual-core i5, no discrete GPU.
- Interrupt-driven real-time speech with zero cloud inference dependency.

SKILLS

Languages & Tools
Python, TypeScript, Rust, SQL, Bash

Infrastructure
PostgreSQL, Docker, FastAPI, Ollama, Unsloth, Git

AI / LLMOps
LangGraph, LangChain, LangSmith, MCP, Groq API, RAG, Multi-Agent Systems, Prompt Engineering

PEFT / Fine-tuning
QLoRA, LoRA, SFTTrainer, vLLM, W&B, bitsandbytes

ML & Edge Inference
PyTorch, ONNX Runtime, Silero VAD, Faster-Whisper, Kokoro-ONNX, FAISS, ChromaDB, NLP

EDUCATION

Minor in Data Science and Artificial Intelligence, 05/2025 – 06/2026
Indian Institute of Technology (IIT), Mandi
Mathematical Foundations → Classical ML → Deep Learning (GANs, VAEs, Transformers, RL, Computer Vision)
Grade: CGPA 8.44 / 10

Bachelor of Computer Applications (BCA), *Jaipur National University* 06/2019 – 05/2022
Grade: 77.4% — First Division

HONORS

NTSE Scholar