

Shivam Singh

AI / ML Engineer · LLM Systems · Agentic AI · Production Backend

Mathura, UP, India · +91-9760112964 · shivam.singh.koyad2006@gmail.com

linkedin.com/in/shivam-singh-a58364384 · github.com/shivamsingh-007

PROFESSIONAL SUMMARY

Computer Science undergraduate (GLA University, CGPA 9/10) specializing in AI/ML and LLM engineering. Builder of production-grade agentic systems, RAG pipelines, and AI security platforms using Python, PyTorch, FastAPI, LangChain, and state-machine architectures. Recognized at TechNavya expo and certified by Anthropic. Passionate about evidence-grounded AI, hallucination prevention, and high-throughput backend engineering.

EDUCATION

GLA University, Mathura — B.Tech Computer Science Engineering

Aug 2025 – Jun 2029 | CGPA: 9.0 / 10.0

CORE COMPETENCIES

Languages: Python, Java (OOP), SQL, R (basic)

AI / ML / LLM: GPT-4o, Phi-2/3, Llama 3.1, Mistral, Zephyr-7B, LoRA/QLoRA fine-tuning, Prompt & Context Engineering, RAG (BM25 + FAISS/ChromaDB), HyDE, Multi-Query, Embeddings

Agentic Frameworks: LangChain, LangGraph, MCP SDK, ReAct loop, Tool Registry, Sub-agent Orchestration, LLM-as-Judge, State Machines

Backend / Systems: FastAPI, Uvicorn, LiteLLM, asyncio/aiohttp, REST API design, Docker, GitHub CI, YAML

Observability & Testing: OpenTelemetry, LangSmith, Structured logging, pytest (628+ tests), mypy, ruff

Vector Databases: ChromaDB, FAISS, Supabase/pgvector, SQLite (encrypted)

Security: AES-256-GCM, Argon2, Scapy/PCAP, DNS/DGA detection, Ransomware & beaconing detection

PROJECTS

Scout — Modular Monolith Agent Runtime

Nov 2025 – Present · Python, FastAPI, State Machines, BM25, OpenTelemetry, Supabase/PostgreSQL, pytest

- Built production-grade single-agent research backend achieving 628 passing tests, 0 mypy errors, and 0 ruff errors across planner, verifier, tools, guardrails, and state management.
- Implemented deterministic loop controller with typed state machines, evidence pointers, and structured logs — treating hallucination as a structural engineering problem.
- Achieved multi-provider LLM abstraction (GPT-4o, OpenAI, Anthropic, Ollama, mock) validated against local Llama 3.1:8b with 100% prompt parse success and 70% citation validity.
- Engineered BM25 evidence scoring (replacing Jaccard) with claim-to-evidence validation (4 support levels) and integrated OpenTelemetry tracing and PostgreSQL persistence.

SHARP — LLM Agent Harness Engine

May 2026 · Python, FastAPI, LangChain, MCP SDK, YAML, asyncio, OpenTelemetry

- Engineered a 4,601-LOC production harness reducing pipeline errors by 30% and increasing reliability 25% via a Detect→Prompt→Execute→Validate→Respond loop with typed transitions.
- Achieved sub-10ms pipeline latency (5.7ms excluding LLM calls), 730K circuit breaker checks/sec, and 80K MCP tool conversions/sec across 218 passing tests.
- Designed programmatic validation gates for safety, lint-based prechecks, and confidence-weighted memory.

Veritas — Truth-Seeking RAG Pipeline

Mar 2026 · Python, LangGraph, LangChain, BM25, FAISS, ChromaDB, HyDE

- Built production-ready RAG pipeline with LangGraph orchestration, intelligent query routing, and hybrid retrieval (BM25 + vector) to prevent hallucination at the architectural level.
- Implemented 5-iteration validation loop (groundedness, relevance, hallucination risk, completeness) with HyDE and Multi-Query query transformation.

RepoBugFix — Agentic Bug-Fix Harness

Apr 2026 · Python, Hybrid RAG, FastAPI, pytest, Validation gates

- Built production-minded bug-fix agent harness with hybrid RAG, programmatic hard gates, and full retry loop — 82/82 tests passing with lint-trigger rejection.
- Designed trigger layer (GitHub issue / CI failure), scoped-context layer, and validation gates approving fixes with evidence and human review status.

CyberSentryAI — AI Cyber Threat Detection

Jan 2026 · Python, FastAPI, LangChain, Red-Team AI, JSONL audit logs

- Developed AI-powered cybersecurity platform detecting phishing, scams, deepfakes, and suspicious URLs across text, email, image, and web inputs, with Red-Team AI for attacker-intent reasoning.
- Engineered Cyber DNA Fingerprinting for similarity/lineage tracking and unified FastAPI backend with async processing.

GestureFlow — Real-Time Gesture Recognition

Feb 2026 · Python, OpenCV, MediaPipe, Deep Learning

- Built real-time computer vision system achieving ~5μs latency (200× faster than real time) at 30 FPS with 44 production tests.

OPEN SOURCE EXPERIENCE & CONTRIBUTIONS

github.com/shivamsingh-007 — 12 merged PRs, 13 open across AI/ML and systems OSS projects (verified).

Active Open-Source Contributor

Ray (ray-project), OpenSRE (Tracer-Cloud), T3MP3ST (elder-plinius), RISC-V Unified DB (riscv), EvalAI (Cloud-CV), ZenML (zenml-io)

- Merged PR #64567 into ray-project/ray (AI compute engine) — hide GPU/GRAM dashboard columns when no GPUs are present.
- Merged PRs #4388 + #4442 into Tracer-Cloud/opensre (LLM SRE agent toolkit) — JWT auth cleanup and gap-note fixes.
- Merged PR #56 into elder-plinius/T3MP3ST (autonomous LLM red-teaming platform) — portable War Room preflight command.
- Merged PR #1983 into riscv/riscv-unified-db — C.NOP encoding with don't-care wildcard fields for immediate ops.
- Open PRs under active review in Cloud-CV/EvalAI, kubernetes-sigs/headlamp, avantifellows/quiz-frontend, zenml-io/zenml, pymc-devs/pymc, json-schema-org, velero-io/velero.

CERTIFICATIONS & ACHIEVEMENTS

- TechNavya Expo — Participant and Exhibitor; recognized for innovation in Agentic AI and LLM engineering.
- Anthropic — AI Safety & Claude Engineering courses covering safety principles and responsible LLM deployment.
- ThinkQbator — Certificate for building an intelligent smart assistant/chatbot.
- GLA University — Academic Excellence, 9.0/10.0 CGPA.