Manish Tiwari

Applied AI Engineer | LLM Applications, Agents, Guardrails & Evaluation

+91 9779405863 | manish.tiwari.09@zohomail.in | LinkedIn | GitHub | Portfolio | Blogs

Summary

Final-year B.Tech CS student building production LLM applications: agent services, retrieval pipelines, structured outputs, and the evaluation that tells you whether they work. Currently shipping a FastAPI planner service with server-side guardrails and a 58-test suite. 42 merged open-source PRs; three packages on PyPI.

Experience

DALM Labs | *AI Engineer (Contract)* — Remote | Oct 2026 – Present

- Built the AI planner behind a CPaaS audience studio: a FastAPI service turning a domain, objective and prompt into an audience recommendation with per-signal reasoning and a budget plan, with Sarvam and Claude behind a provider-agnostic interface and the tenant resolved from the caller's JWT.
- Enforced LLM guardrails: signals outside the tenant's resolved vocabulary dropped server-side, budget channels renormalised to **100%**, Pydantic schema validation with one corrective retry, and a static fallback when the upstream platform was unavailable. Wrote the **58-test** suite as acceptance criteria and fixed three production bugs that surfaced only against a live model.

Smartivity Labs Pvt. Ltd. | *Automation Intern* — New Delhi | June 2026 – Aug 2026

- Built a Slack-integrated project platform used by **100+ designers and managers** across **200+ projects**, replacing a 100-user spreadsheet: Slack OAuth (bot install + OIDC login with PKCE), per-project channels, timezone-aware reminders, and an approval workflow that auto-advances project state.
- Automated certificate generation and PPT/PDF → Excel extraction, cutting processing time by **over 90%**: ~1 minute to **10–12 seconds** per certificate, and a 2–3 hour manual workflow down to a few minutes.

Projects

SafeRun AI, Guardrails and Sandbox for AI Agents [PyPI] [GitHub] [Live] — May 2026

- Docker sandbox executing untrusted AI-generated Python with CPU, memory and timeout limits; an AST scanner and YAML policy engine classify each execution into **4 risk levels** before anything runs.
- Published as **saferun-ai** on PyPI and shipped as an **MCP server** exposing scan, execute and history tools to any MCP client, with audit logging and LLM explanations.

Bharat Benefits, Voice-First RAG Assistant [Github] — 2026

- End-to-end STT → retrieval → LLM → TTS pipeline over Indian government welfare scheme documents, using Sarvam speech models with ChromaDB retrieval.
- Answers are generated only from retrieved scheme text, so the assistant cannot invent an eligibility rule or a benefit amount.

Laya on a GTX 1650, LLM Inference Benchmark [GitHub] [Blog] — Sep 2026

- Benchmarked a **421M**-parameter LLM on 4GB consumer hardware with CUDA-synchronised timing and p50/p95 reporting: found latency **15–22x** slower than published T4 figures, a **3.8x** per-question cost drop from batching, and diagnosed a batch that exceeded VRAM from mean latency at 2.4x median.

Technical Skills

LLM applications: agents and tool calling, RAG, prompt engineering, structured outputs (Pydantic), evaluation harnesses, guardrails, LLM APIs (Claude, Sarvam, OpenAI)

ML: PyTorch, HuggingFace, Transformers, LoRA fine-tuning, scikit-learn

Backend: FastAPI, REST APIs, Docker, PostgreSQL, SQLite, ChromaDB, GitHub Actions, Linux

Languages: Python (primary), Rust (learning)

Education

IK Gujral Punjab Technical University — Jalandhar, Punjab, India

B.Tech Computer Science & Engineering; CGPA 8.0 — July 2023 – June 2027

Open Source & Achievements

- **42 merged pull requests** across 10+ organisations (Hugging Face, Moss, CHAOSS/Augur): SDK improvements, automated tests, developer tooling and AI application features.
- Trained **Trimurti-LM**, a 4.2M-parameter multilingual GPT (EN/HI/PA), from scratch to perplexity **3.32** in **2.38 hours** on a 4GB GTX 1650. Published Ollama libraries for Sarvam 30B and 105B, **100+ community pulls**.
- Won the Industrial Automation hackathon for **EnergiX Copilot**; fine-tuned Mistral-7B and Qwen-14B with LoRA at the AMD Hackathon, IIT Delhi. Also published **mesh** (GPU capacity planner for heterogeneous fleets) and **LocalMind** (local LLM hardware advisor) on PyPI.