Nilesh Raj

+91-8887522416 | nileshraj1735@gmail.com | LinkedIn | GitHub | Portfolio

Summary

Founder and AI engineer. I build and ship AI systems end to end, from architecture to live deployment: agentic systems, RAG pipelines and fine-tuned transformers. I run AgentNine, a live directory of open-source AI agents, and have shipped CrewAI multi-agent platforms, LangGraph RAG pipelines and BERT classifiers to production. I treat JWT RBAC, rate limiting and observability as seriously as prompt design and model accuracy. NVIDIA-certified in LLM application development.

Experience

Founder and AI Engineer *Jun 2026 – Present*
AgentNine *Varanasi, India*

- Founded and shipped agentnine.pro, a directory of open-source AI agents. Each listing carries its source repository, install steps for Windows, macOS and Linux, the access it requests, and what is still undocumented.
- Built the product on Next.js App Router, React 19, TypeScript and Tailwind CSS, with Supabase Postgres for auth and data, a Vitest suite, and deployment on Vercel.
- Designed the browse experience: category filters, operating-system and API-key facets, Fuse.js full-text search, and side-by-side comparison of three agents without an account.
- Defined the five-step setup verification every listing passes before publication: the exact commit reviewed, the operating systems checked, the install command, a first task actually run, and the date. Grew the catalogue past 110 agents across coding, automation, research and content.
- Lead product direction and technical decisions with a small team across product, engineering and launch.

AI/ML Trainee (Internship) *Feb 2026 – May 2026*
Technosavvys Education Technology *Remote*

- Fine-tuned a BERT classifier on 500+ student course reviews with Hugging Face. It reached 88% accuracy, a 15% lift over the TF-IDF baseline, using learning-rate scheduling and dropout tuning.
- Shipped a Streamlit dashboard backed by SQLite that cut report generation for 10+ mentors across 5+ courses from hours to minutes.
- Rewrote the NLP preprocessing pipeline (text normalisation, feature extraction), which raised baseline accuracy and removed manual data entry from the weekly reporting workflow.
- Worked directly with course mentors to turn raw classifier output into a view non-technical staff could use.

Projects

AgentNine | *Next.js, React, TypeScript, Supabase, Tailwind, Vitest* *Live* | *Code* *2026*

- Open-source AI agent directory, 110+ listings, with faceted search, three-way comparison, and per-listing setup, access and verification detail.

OmniCrew AI | *Python, CrewAI, LangChain, FastAPI, Redis, Docker* *Live* | *Code* *2026*

- Built a multi-agent pipeline on CrewAI with a 3-tier LLM fallback router (DeepSeek, Mistral AI, Gemini), reaching 99.9% task completion on real-time web extraction.
- Added Redis-backed MD5 caching for LLM responses, cutting latency from 2.5s to under 50ms on cache hits at no extra token cost.
- Cut redundant database writes by around 35% with MD5 semantic deduplication and INSERT IGNORE.
- Blocked SSRF by resolving target URLs and rejecting private, local and cloud-metadata IP ranges before any fetch.

LumanGuide | *Python, LangGraph, FAISS, FastAPI, MongoDB, LangSmith* *Live* | *Code* *2026*

- Built an agentic RAG system on a LangGraph state machine with document grading and query routing, cutting retrieval latency by 30% against a naive RAG baseline.
- Added a grade-and-rewrite cycle: an LLM-as-judge node scores retrieval quality and routes poor results to a rewrite node that loops back.
- Built a FAISS vector store with semantic chunking for sub-second retrieval, and moved auth from SQLite to MongoDB Atlas, ending the data-wipe incidents.
- Added prompt-injection guardrails, JWT RBAC and regex secret scanning, redacting credentials before anything reaches the vector store.

Technical Skills

AI and GenAI: LangGraph, CrewAI, LangChain (LCEL), agentic systems, RAG pipelines, prompt engineering, NLP, BERT, transformers, semantic chunking, embedding models, vector search (FAISS), OpenAI API, Mistral AI, DeepSeek, Gemini API, NVIDIA NIM, LangSmith

ML Engineering: PyTorch, Scikit-learn, Hugging Face, Pandas, NumPy, fine-tuning, transfer learning, hyperparameter tuning, feature engineering, model evaluation, sentiment analysis

Product and Web: Next.js, React, TypeScript, Tailwind CSS, Supabase, Streamlit, Vitest, Vercel

Infrastructure: Python, FastAPI, Docker, REST APIs, Redis, AWS (S3, EC2, Lambda), Git, CI/CD, Pydantic

Data and Storage: SQL, MongoDB Atlas, MySQL, PostgreSQL, SQLite, vector databases

Education

United University *Aug 2022 – May 2025*
BCA, Artificial Intelligence (IBM Specialization) *Prayagraj, UP*

Certifications

Building LLM Applications with Prompt Engineering | *NVIDIA Deep Learning Institute* *2025*