Nitinkumar Patel

South Elgin, IL | 484-447-7008 | npatel121.py@gmail.com

linkedin.com/in/nitinkumar-patel | github.com/nitinkumar-patel | NitinMPatel.com

Building systems that give teams the stability to stop worrying and keep moving forward.

PROFESSIONAL SUMMARY

Staff AI/LLM Engineer with 12+ years building high-scale distributed systems and production Generative AI (GenAI) platforms. Built and led teams of 5+ engineers to ship Retrieval-Augmented Generation (RAG) pipelines, multi-agent workflows, and full-stack SaaS products achieving **95% accuracy, 20% cost reduction, and 99.99% availability**. Driven by a belief that stable, well-architected systems free teams to move fast and think big, not worry about what might break. *Actively seeking Staff AI/LLM Engineer roles (Remote-first).*

TECHNICAL SKILLS

- **AI & ML Engineering:** Deep Learning & ML Principles, Fine-tuning (Hugging Face), Agentic Workflow Orchestration (LangGraph, CrewAI), Multi-Modal AI, LLM Evaluation (RAGAS, LangSmith), Vector DBs (pgvector, Pinecone), MLOps, Prompt Engineering, LangChain, OpenAI, Llama 3, RAG Pipelines, Semantic Caching, n8n.
- **Engineering Excellence:** System Design, Code Reviews, Software Architecture Decision Records, Tech Debt Management, CI/CD, Observability (OpenTelemetry, Prometheus, Grafana)
- **Backend Architecture:** Python, FastAPI, Django, Flask, Event-Driven Microservices, REST APIs
- **Frontend Development:** TypeScript, React, Next.js, Vue.js, Angular, Redux, Server-Side Rendering (SSR), Tailwind CSS, Mantine UI
- **Data & Analytics:** PostgreSQL, Redis, RabbitMQ, Kafka, Pandas, Scikit-learn, ETL Pipelines
- **Cloud & DevOps:** AWS (Lambda, EC2, S3), GCP, Terraform (IaC), Docker, Kubernetes, CI/CD, GitHub Actions, Pytest, AI-Assisted Development (Cursor, Copilot, Claude)

PROFESSIONAL EXPERIENCE

Early-Stage AI SaaS Startup (Stealth) | AI Solutions Architect — Nov 2025 – Present

- **Agentic AI :** Architected complex **Agentic AI** systems utilizing **LangGraph, n8n and CrewAI** for advanced **workflow orchestration**, implementing tool-calling and output guardrails to automate high-stakes content pipelines - improving local search rankings by 30% for pilot users.
- **LLM Workflows:** Engineered a multi-modal n8n agentic workflow to process structured data and unstructured visual/image assets for automated GMB content generation, reducing manual content creation effort by 70%.
- **RAG Pipeline & Vector Search:** Architecting production-grade **RAG** pipelines using vector embeddings and semantic search to power AI-driven local SEO recommendations, handling chunking strategy, embedding model selection, and retrieval quality tuning.
- **Evaluation & Observability:** Building the end-to-end **LLM Lifecycle** from data curation and **fine-tuning** via **Hugging Face** to production **evaluation** and observability using **LangSmith, RAGAS, and OpenTelemetry,** reducing inference costs by 20%.
- **Full-Stack Product Delivery:** Delivering AI-powered storefronts using **Next.js, FastAPI, and Tailwind CSS** and targeting sub-1s page loads, contributing end-to-end across **React UI, Python** microservices, and cloud inference endpoints.
- **Automated Data Pipelines:** Building scalable Python and n8n **ETL** pipelines via Google APIs to analyze local market trends, feeding structured consumer behavior data into AI recommendation workflows.
- **Platform Architecture & Infrastructure:** Defining greenfield platform architecture - **pgvector/Pinecone**, **AWS/Terraform, and CI/CD automation** - using open standards and modern **GenAI** tooling with no legacy constraints.

NOPSEC INC. | NY – Remote (Cybersecurity SaaS) — Nov 2021 – Nov 2025

Engineering Lead (Hands-on)

- **GenAI Strategy:** Applied Deep learning principles to fine-tune open-source LLMs (Llama, Mistral) using **Hugging Face** transformers, optimizing for domain-specific vulnerability intelligence. Defined technical strategy for a multi-stage **RAG** pipeline using **LangChain**, automate vulnerability intelligence for 500k+ records; achieved a **20% reduction in inference costs** via semantic caching and reduced p99 response latency by **600ms**.
- **Product Leadership:** Directed the full-stack delivery of a flagship Dashboard using **React, FastAPI, and Django**. Leveraged FastAPI for high-performance AI endpoints and Django for robust core platform management, improving user decision-making speed and decreasing initial page load by **40%**.
- **Intelligent Automation:** Engineered an **AI-powered automation framework** utilizing **pgvector** for context-aware CVE prioritization, achieving **95% accuracy** and **increasing SOC team triage capacity by 40%** without additional headcount.

- **Infrastructure Evolution:** Led the strategic migration of a legacy Django monolith to an **event-driven architecture** using **FastAPI** and **RabbitMQ**. Deployed via **Terraform** on **AWS Lambda**, resulting in **99.99% availability** and a **30% boost in developer velocity**.
- **Process & Culture:** Introduced **automated testing & CI/CD standards** that reduced deployment failures by 30% and improved team sprint velocity by 25%.
- **Team Leadership:** Mentored a cross-functional team of **5+ engineers and designers**, building a culture of psychological safety and technical clarity through rigorous code review standards and architecture workshops - resulting in **2 internal promotions to Senior level** within 12 months.
- **Developer Velocity:** Spearheaded adoption of AI-Assisted Development **(Cursor/Copilot/Claude)** across the SDLC and overhauled the onboarding process, reducing ramp-up time from 3 weeks to 4 days.

ZORO TOOLS INC. | Buffalo Grove, IL **Apr 2019 – Nov 2021**
Python Full Stack Developer

- **High-Scale Software Architecture:** Architected a **high-concurrency data layer** and RESTful APIs for a **3M+ page catalog** using **Django and Flask**. Slashed p99 latency by **55%** via **Redis** caching and **PostgreSQL** query optimization, successfully sustaining a **300% YoY traffic surge** (10k+ concurrent users).
- **Growth Engineering & UI:** Engineered SEO optimization algorithms using **Pandas and Scikit-learn**, served via **Flask** microservices to a **Vue.js** frontend. This software architecture drove a **6.3% daily increase** in organic traffic and a **1.3% lift in daily revenue**.
- **System Reliability & Dashboards:** Engineered **Django-based SCM microservices** to automate inventory updates across **4M+ SKUs**, reducing data discrepancy errors by **92%**. Developed a **real-time monitoring dashboard** in **Vue.js and Vuex** that eliminated manual reconciliation for the procurement team.

NOKIA | Naperville, IL **Apr 2018 – Apr 2019**
Python Full Stack Developer

- **Security & Compliance:** Engineered a centralized security management platform for **20,000+ users** ensuring **100% compliance** with **NSA and EU GDPR** standards. Developed the backend in **Django** and a high-performance frontend in **Angular** to provide real-time visibility into cross-border compliance health.
- **Infrastructure Automation:** Architected automated workflows for **Verizon Wireless** infrastructure using Python and shell scripting, increasing user provisioning efficiency by **80%**. Maintained a **0% failure rate** across consecutive federal security audits through rigorous automated validation and **Pytest-driven unit testing**.
- **System Integration:** Built and documented internal **RESTful APIs** that bridged **legacy telecom databases** with modern security auditing tools, enabling seamless, **real-time data synchronization** between US-based infrastructure and EU regulatory reporting layers.

MOTOROLA MOBILITY LLC | Chicago, IL **Jun 2015 – Apr 2018**
Python Developer

- **Scale Engineering:** Architected and owned the Logcat-Dashboard using **Python and Django** to aggregate and visualize real-time crash telemetry from **millions of mobile devices** – enabling firmware regression detection that directly **prevented potential mass-market recalls**.
- **Data Pipeline Optimization:** Engineered an automated **ETL pipeline** to parse **terabytes of raw log data**, cutting the 'field crash detection' to 'actionable developer visibility' interval by **95%** - shifting resolution time from days to minutes.

CERTIFICATIONS & LEARNING

- **Completed:** Proficient AI Engineer: Builder/Coder/Leadership/Core Track/Agentic Track/MLOps Track, AI Engineer Core Track: LLM Engineering/RAG/QLoRA/Agents, AI Engineer Agentic Track: The Complete Agent & MCP, AI Engineer Production Track: Deploy LLMs & Agents at Scale, AI Leader: GenAI and Agentic AI , AI Coder: Complete Claude Code & Coding Agents, AI Builder: Create Agents - Voice Agents & Automations in n8n
- **In Progress**: Andrej Karpathy's Neural Networks: Zero to Hero
- **Projects**: **MarketMind** (github.com/nitinkumar-patel/ai-market-mind) - Agentic multi-tool research assistant built with LangGraph + OpenAI tool-calling; demonstrates autonomous web research, summarization/report
- **Nitin's Digital twin GenAI Chatbot**, **Multi Locations Route Planner**

EDUCATION

- **M.S. in Information Technology Management** | Campbellsville University, KY | 2021
- **M.S. in Computer Engineering** | Illinois Institute of Technology, Chicago, IL | 2016
- **B.E. in Electronics & Communication Engineering** | Gujarat Technological University, India | 2013