Dhruvil Patel

Applied AI Engineer
+91 9510573790 | Pune, India | dhruvil7694@gmail.com | LinkedIn | GitHub | Portfolio | X | Medium | Substack

PROFESSIONAL SUMMARY

Applied AI Engineer with ~1.5 years of professional experience building production AI systems with LLMs, including RAG pipelines, NL-to-SQL platforms, and document intelligence workflows used to automate data extraction and reporting. Designed multi-model systems that route between providers based on cost, latency, and task complexity, and implemented validation layers to improve reliability and reduce hallucinations. Experienced with Azure AI services (Azure OpenAI Service, Azure AI Search) and other cloud-based AI infrastructure. Focused on making AI systems usable in real workflows by handling scaling, failure cases, and long-running processing rather than one-off model outputs.

SKILLS

- **Applied AI & LLM Engineering**: Multi-Agent Systems, RAG Pipelines, Agentic Workflows, Prompt & System Design, LLM Evaluation, Hallucination Control, OpenAI API, Azure OpenAI Service, Azure AI Search, LangGraph, LangChain.
- **Cloud Platforms**: Azure (Azure OpenAI Service, Azure AI Search, Azure AI Studio), AWS (S3, SageMaker, EC2), Docker-based deployments on cloud infrastructure. Experienced in building and deploying AI workloads on Azure cloud services.
- **AI Infrastructure**: Async Processing, Concurrent Pipelines, Background Jobs, Token & Cost Optimization, Model Routing (OpenAI, Anthropic, Ollama, vLLM), Streaming APIs (SSE).
- **Backend & Data**: Python, FastAPI, REST APIs, PostgreSQL, FAISS, Qdrant, External API Integration.
- **Machine Learning**: Transformers (Hugging Face), LoRA/QLoRA, NLP Systems, Feature Engineering.
- **MLOps & Deployment**: Docker, GitHub Actions, MLflow, Weights & Biases, AWS (S3, SageMaker), Model Versioning.

PROFESSIONAL EXPERIENCE & AI SYSTEMS

AI Engineer | 1POINT1, Pune **Sept 2025 – Present**

- Built NL→SQL platform with schema-aware guardrails enabling non-technical sales and ops teams to query PostgreSQL, MySQL, and MSSQL via natural language — eliminating analyst dependency for enterprise reporting and accelerating forecasting workflows.
- Designed hybrid AI document intelligence system extracting 80+ structured fields from enterprise bidding documents using rule-based parsing + selective RAG; reduced contract review cycle time and improved deal visibility.
- Automated large-scale document workflows via parallel PDF splitting + Google Drive pipeline, processing 1,000+ page files in under 2 minutes vs. 1–2 hours manually — freeing sales teams from administrative overhead.
- Delivered computer vision POC for automobile defect classification achieving 85%+ accuracy across 20 job categories under varied real-world conditions.
- Designed high-performance concurrent processing pipelines for enterprise automation, optimizing latency and throughput across large unstructured datasets.

AI/ML Engineer | Cyber Security Umbrella **Dec 2024 – Sept 2025**

- Deployed GenAI real-time assistant handling 82,000+ cybersecurity scenarios using RAG + LangChain + Gemini API — 40% faster incident triage vs. manual lookup, on lightweight infrastructure with no external API dependency for core inference.
- Built multi-model compliance system using LoRA/QLoRA fine-tuned LLMs achieving ~95% operational accuracy in regulatory mapping and gap reasoning — directly comparable to AI-driven sales qualification and competitive intelligence workflows.
- Designed SOC analytics pipeline aggregating data from 6+ security tools with real-time ingestion, anomaly detection, and automated threat monitoring; demonstrates full-stack AI forecasting architecture applicable to sales performance monitoring.
- Led cross-functional team of 5 engineers using agile ML workflows and automated testing, accelerating delivery by 25%; regularly communicated complex AI capabilities to non-technical executive stakeholders.
- Deployed scalable solution on AWS SageMaker with auto-scaling FastAPI endpoints handling 1,000+ concurrent requests.

AI/ML Researcher | SVNIT (NIT Surat) **May 2024 – Aug 2024**

- Designed CNN-LSTM hybrid architecture for EEG-based depression detection achieving 90% accuracy — 15% improvement over state-of-the-art baselines.
- Engineered distributed TensorFlow training pipeline on AWS EC2 cluster, reducing training time from 12 hours to 7 hours through parallel processing and memory optimization.

VOLUNTEER / RESEARCH EXPERIENCE

AI/ML Researcher | P P Savani University **Dec 2023 – May 2024**

Unpaid research internship

- Built end-to-end NLP depression detection system analyzing 20,000+ social media posts with 88.10% accuracy using ensemble methods and advanced feature engineering.
- Published peer-reviewed research in ICICC 2024 (Springer LNNS); presented novel classification methodology at international conference.

PROJECTS

AI Research Automation Platform BohrAI

Multi-Agent Research System · CLI + API · LLM Orchestration · Workflow Automation

- Built an AI-powered research system that automates end-to-end workflows including literature discovery, evidence validation, and structured report generation using multi-agent pipelines.
- Designed a modular architecture combining a CLI tool and a lightweight API service, enabling both local execution and hosted workflows with real-time progress streaming (SSE).
- Implemented multi-agent orchestration using a central runtime with specialized sub-agents for literature retrieval, reasoning validation, and citation integrity.
- Developed a hybrid retrieval pipeline integrating ArXiv, Semantic Scholar, and PubMed with rule-based validation to ensure evidence-backed outputs and reduce hallucinations.
- Engineered scalable execution with parallel processing, concurrency control, disk-backed state management, and dynamic model routing across LLM providers to optimize cost, latency, and reliability.

Enterprise File Governance & AI Assistant Platform

AI Systems · FastAPI · PostgreSQL · Redis · Multi-Agent Workflows · SSE

- Built an enterprise platform for monitoring, controlling, and governing sensitive file operations across Windows endpoints, integrating real-time telemetry, quarantine workflows, and approval-based deletion systems.
- Designed a distributed architecture with a FastAPI backend, PostgreSQL as system of record, Redis-based rate limiting, and a Windows agent with local buffering for reliable event ingestion and offline resilience.
- Developed an AI-powered operator assistant using LLM tool-calling with structured workflows, enabling natural language interaction over live system data with strict RBAC and human-in-the-loop approval for critical actions.
- Implemented streaming APIs (SSE) for real-time chat, event tracking, and system updates, supporting long-running operations with continuous feedback to users.
- Engineered a command execution pipeline between backend and agents with polling, retry logic, and state tracking to ensure reliable remote operations under network constraints.
- Built policy-driven automation (auto-delete scheduler, rule engine) to enforce governance rules while maintaining auditability and control over system actions.

EDUCATION

Bachelor of Information Technology (CGPA: 8.5/10, **Gold Medallist**), P P Savani University | 2021 – 2025

Author of 3 peer-reviewed AI/ML research papers, including ICICC 2024 (Springer LNNS) on large-scale social media depression detection using NLP and ML.