

Samarth Patel

+91 7043571380 samarthbrijeshpatel@gmail.com

linkedin.com/in/samarthpatel24 github.com/samarthpatel24

Education

Symbiosis Institute of Technology, Pune

Aug 2022 – Jun 2026

Bachelor of Technology in Artificial Intelligence and Machine Learning (CGPA: 7.5 / 10)

Pune, India

- **Relevant Coursework:** Natural Language Processing, Supervised Machine Learning, Deep Learning, Data Structures and Algorithms, Cloud Computing

Experience

AI Engineer (Graduate Engineer Trainee)

Jul 2026 – Present

Futurecraft

Pune, India

- Built core of a production **text-to-SQL** agent that turns non-technical users' plain-English questions into validated SQL over a live PostgreSQL warehouse – schema linking, staged query planning, pre-execution validation; moving hand-written prompts to DSPy pipelines **nearly doubled its execution-match accuracy**.
- Co-owned the **AWS** deploy path for the AI services: rebuilt a broken **GitLab CI/CD** pipeline and provisioned a fresh EKS (Kubernetes) cluster, ECR repos, S3 and IAM (OIDC / IRSA) roles, so services ship to the cloud on every merge with no static keys.
- Built the tool layer behind a production **agentic AI assistant** – a LangGraph agent served over FastAPI to both human and API callers, with tools that reason over live platform data and a vector store, returning grounded, cited answers.

AI Engineer Intern

Dec 2025 – Jun 2026

Futurecraft

Pune, India

- Built the **evaluation harness** and Langfuse tracing so every agent and prompt change is scored before it ships; failure analysis traced most misses to the **data tier**, not the model, redirecting where the team spent its fixes.
- Generated correlated FMCG datasets and added automated **data-quality** checks that caught up to **14.6× row duplication** before it could skew the analytics.
- Built an automated **Playwright** test suite and correlated failures with Kubernetes pod logs, surfacing backend defects and root-causing a load blocker that had been failing under traffic.

Projects

RAG Playground | Python, FAISS, ChromaDB, embeddings

retrieval eval

- Built a config-as-code harness that benchmarks RAG retrieval across chunking strategies, multiple embedding providers and **vector databases** (in-memory / Chroma / FAISS).
- Reported latency, cosine-score distribution and overlap per setup, so vector-database and chunking choices are measured against each other instead of guessed.

Route-Sentinel | Python, FastAPI, AST, GitHub Actions

PyPI

- Built a scanner that flags unprotected **REST API** endpoints across FastAPI, Flask and Django (AST-based) plus Express and Spring, with an optional LLM pass for edge cases.
- Shipped it as a pre-commit hook and **GitHub Actions** PR gate after a real incident exposed unauthed endpoints, so the check runs on every change.

ClearQuill | Python, PyTorch, HuggingFace, FastAPI

on-device

- Fine-tuned **ModernBERT** (PyTorch / HuggingFace) on the RAID corpus for sentence-level AI-text classification, reaching accuracy comparable to top-ranked leaderboard entries.
- Wrapped it in a local-first writing assistant running open-source LLMs fully on-device (llama.cpp GGUF, GPU/VRAM auto-detection, per-task model routing), with a test suite mocked at the ML boundary so it runs on GPU-less CI.

Technical Skills

Languages: Python, SQL, TypeScript

Agentic & Generative AI: LangGraph, MCP (Model Context Protocol), tool calling; RAG, prompt engineering, DSPy; LLM APIs (OpenAI, Anthropic, AWS Bedrock)

Backend & Data: FastAPI, REST APIs, microservices; PostgreSQL / pgvector, ChromaDB, FAISS

Cloud, DevOps & Eval: AWS (EKS, ECR, S3), Docker, Kubernetes, GitLab CI/CD, GitHub Actions; Langfuse tracing, evaluation harnesses

Open Source: active contributor – whichllm (6.5k+ stars), tradingview-mcp