SOHEN PATEL

Toronto, ON, Canada | (416) 666-8954 | sohenpatel.work@gmail.com | linkedin.com/in/sohen-patel | github.com/sohenpatel22 | Portfolio

TECHNICAL SKILLS

Languages & Data: Python, SQL (MySQL, PostgreSQL), Pandas, NumPy, Statsmodels, Matplotlib, Seaborn, Plotly, Power BI
ML & Deep Learning: scikit-learn, PyTorch, Computer Vision (MediaPipe, OpenCV), Model Evaluation (F1, NDCG, RAGAS, DeepEval)
LLMs: LangChain, LangGraph, Langfuse, RAG (hybrid search, Qdrant), DSPy, Claude/OpenAI APIs, Pydantic, Streamlit, Hugging Face
MLOps: Git, GitHub Actions, Docker, FastAPI, MLflow, DVC, pytest, Prometheus/Grafana, ONNX, AWS (S3, EC2, SageMaker, Bedrock)

EXPERIENCE

AI Research Intern Jul. 2026 - Sep. 2026
Ontario Health *Toronto, ON, Canada*

- Built an enterprise Ontario Health News Agent in Microsoft Copilot Studio end-to-end, deployed to 5,000+ employees org-wide
- Created a 259-item human-verified gold-standard dataset for 3 healthcare tasks; built an end-to-end LLM evaluation framework for 5 models (GPT-5, Claude, Gemini, MS Copilot, Perplexity) across 13+ metrics using RAGAS, DeepEval, and LLM-as-a-Judge
- Conducted systematic prompt-engineering experiments across 4 strategies, finding Chain-of-Thought prompting on a 5-year dataset increased agent F1 score from 0.25 to 0.42 (+68%) and reduced hallucination rate by 51% (78% to 27%); logged runs to Langfuse

AI Engineering Project Team Lead May 2026 - Jun. 2026
Deloitte (Sponsored Practicum) *Toronto, ON, Canada*

- Built a 6-agent system that turns a client brief into consulting deliverables (WBS, RACI, RAID, weekly status report) with a human-in-the-loop revision loop, plus Deloitte template sync for DOCX/XLSX exports, backed by 129 automated tests
- Added an LLM routing layer with fallback across 4 backends (Claude, GPT, Gemini, DeepSeek) and prompt-cache warm-up to cut repeated token cost on agent calls
- Built a 13-rule Quality Gates engine (graph-based WBS cycle detection, RAID/RACI cross-artifact validation)

Graduate Engineer Trainee Aug. 2024 - Aug. 2025
ArcelorMittal Nippon Steel India (AM/NS) *Surat, India*

- Used SAP datasets and project logs for real-time monitoring of compliance metrics and operational workflows, generating actionable insights that supported executive-level decision-making across 12 cross-functional departments
- Designed interactive Power BI dashboards to track project KPIs, resource allocation, and progress trends for data-driven planning across a large-scale plant expansion project from 7 to 14 million metric tons per annum (MMTPA)
- Identified workflow bottlenecks and execution gaps, contributing to a 16.7% reduction in project timelines

PROJECTS

Multi-Agent Legal Judgment Prediction | *LangChain, Claude API, Kanon 2, Qdrant, RAG, Langfuse* *MEng Research Project* | Write-up

- Extended a LangChain multi-agent system with a Claude pipeline predicting Federal Court of Appeal outcomes from the lower decision: a grounds agent, appellant and respondent agents briefing with date-filtered precedents, and a judge agent
- Rebuilt a flawed benchmark (38% of cases mismatched their judgments, 52/68 saw later precedents) as a 398-case dataset from open court data, LLM-labelled and blind re-checked (156/158); a 5.7K-decision library lifted authority coverage 37% to 67.5%
- Ran 15+ cost-capped experiments (Batch API, Langfuse): judge AUC 0.72 (CI 0.66-0.77), +0.08 over a metadata baseline

ASL Fingerspelling Recognition | *PyTorch, MediaPipe, OpenCV, MLflow, DVC, Docker, GitHub Actions, FastAPI* Live Demo | GitHub

- Built a video-to-text ASL system: MediaPipe landmarks feed a Conformer encoder + Transformer decoder (27.8M params, beam search) on 115K+ Kaggle sequences; cut test CER 0.334 to **0.307** on 14 unseen signers (13 improved) after 10+ experiments
- Traced errors to missed hands (r=-0.96); a flag catches 98.6% of failures and leaves **0.059** CER on accepted clips (vs 0.307)
- Added MLflow registry, DVC, 140+ tests, GitHub Actions CI/CD, FastAPI + Prometheus/Grafana alerts, Docker, ONNX (1.6x faster)

Market Research Agent: Agentic RAG over SEC Filings | *LangGraph, pgvector, PyTorch, Langfuse, FastAPI, Docker* Live Demo | GitHub

- Built a LangGraph agent over SEC filings, prices and forecasts: hybrid pgvector + full-text search with reranking lifted NDCG@6 from 0.44 to **0.56**; a judge model verifies answers and retries weak ones; 98% cited, 100% out-of-scope refused, $0.001/query
- 99-question golden set (RAGAS faithfulness 0.95, DeepEval CI gate); DeepSeek 14x cheaper than Sonnet 5.5 at similar faithfulness
- PyTorch LSTM beats HAR-RV/GARCH/ARIMA (QLIKE 0.202 vs 0.223, DM p=0.001); monthly CI retrain; MCP server, HF deploy

EDUCATION

University of Toronto Sep. 2025 - Expected Dec. 2026
Master of Engineering | Emphasis: Data Analytics and Machine Learning | GPA: 3.81/4.0 *Toronto, ON*

- Coursework: Deep Learning, Foundations of Data Analytics and Machine Learning, Data Science Methods and Statistical Learning, AI in Finance, Introduction to ML | MEng Research Project: Building Multi-Agent AI Systems for Legal Applications

L.D. College of Engineering, Gujarat Technological University Jul. 2021 - Jul. 2024
Bachelor of Engineering in Mechanical Engineering | CGPA: 8.40/10 *Ahmedabad, India*

AWARDS, ACHIEVEMENTS & ACTIVITIES

- Awarded Aga Khan Foundation International Scholarship, fully covering graduate study costs at the University of Toronto
- Teaching Assistant for MIE354 Business Process Engineering and MIE1628 Cloud-Based Data Analytics at University of Toronto
- Filed 2 provisional patents for ISRO research project on deployable space telescope systems (IN:202421088072, IN:202421087180)
- Achieved All-India Rank (AIR) 3465 in GATE XE 2023, a national-level engineering examination