

Dhruv Patel

B.Tech. Mechanical Engineering | SVNIT Surat
+91-7383308394 | dhruvvcodess@gmail.com
LinkedIn | GitHub | X | importdhruv.in

Summary

AI-focused software engineer passionate about building robust, scalable, and reliable AI systems. Deeply interested in the intersection of **LLM evaluation, execution isolation, and reproducible benchmarking** to advance autonomous agents. Skilled in designing real-time evaluation pipelines, secure backend architectures, and full-stack applications. Eager to tackle complex engineering challenges that bridge cutting-edge AI research with production-ready software.

Experience

ADOFLABS, *Applied Researcher*

Jul 2026 – Sep 2026

- Researched and prototyped a **real-time evaluation layer for AI voice agents** that scores phone-call turns during live conversations, tracks sentiment and escalation risk, and surfaces potentially at-risk interactions before a call ends.
- Built a concurrent evaluation pipeline using **Pipecat, Twilio, Groq (STT/LLM/TTS), and WebSockets**, running evaluation alongside response generation; **achieved approximately 3-second scoring latency** during real phone-call tests.
- Designed a modular `score_turn()` contract returning an LLM-based sentiment score in $[-1, +1]$, escalation classification, and confidence. Defined evaluation categories including human-agent requests, repeated complaints, and neutral interactions, with prompt rules to distinguish explicit escalation requests from repeated frustration.
- Implemented a WebSocket dashboard for per-turn scores and escalation alerts; wrote unit and integration tests against live Groq endpoints and a real WebSocket server. Defined four key metrics: sentiment trajectory, goal-completion probability, escalation triggers, and time-to-recovery.

ADOFLABS, *Software Engineer*

Jan 2026 – Jul 2026

- Engineered a full-stack **Employee Intelligence Platform** using **Node.js, Express.js, PostgreSQL, Prisma, Next.js, React, and TypeScript**, separating routes, services, repositories, business engines, middleware, and background jobs.
- Developed employee-scoring, flight-risk, and pay-fairness engines with configurable weighted scoring, normalization, percentile analysis, risk thresholds, keyword detection, and severity classification.
- Implemented **JWT/RBAC, company-level access isolation, Zod validation, rate limiting, bcrypt, and AES-256-GCM encryption**. Added tamper-evident SHA-256 hash-chained audit logs, versioned configuration, approval workflows, and scheduled background jobs.

Projects

Dynamion — Long-Horizon Coding-Agent Evaluation Harness

2026

Open-source Python framework for reproducible repository-level benchmarks

GitHub

- Designed a **Python evaluation harness** for coding agents solving multi-step repository tasks. Combines hidden executable tests and anchored LLM rubrics to generate structured run records, partial-credit scores, execution trajectories, patches, and leaderboard aggregates.
- Implemented strict task specifications using **Pydantic** validation and JSON Schema generation; developed CLI workflows for **validate, list, schema, run, and bench**, plus a dependency-free static HTML leaderboard.
- Integrated **Aider** as the primary agent runner, designing a configurable execution architecture for extensible agent integration and run metadata.
- Built local and Docker execution environments with **network-disabled grading containers and read-only workspaces**. Implemented digest-pinned image support, per-environment Python resolution, and setup commands baked into derived images.
- Developed hidden-test grading with explicit source-to-destination mapping, per-test partial credit, pytest-result parsing, collection/infrastructure-failure detection, and a **nine-category failure taxonomy** separating agent errors from grading and environment failures.
- Implemented LLM-rubric grading with anchored criteria, self-consistency sampling, malformed-response handling, and judge-independence checks. Added provenance and contamination metadata to improve benchmark traceability and result interpretation.
- Authored **five benchmark task suites** (HTTP retry/backoff, idempotent payment ledgers, behaviour-preserving typed refactoring, Decimal billing arithmetic, event-bus concurrency) with **145 grader test cases** covering atomicity, edge cases, compatibility, synchronization, and API preservation.
- Added a task-design handbook, grader-design notes, CI checks, isolation tests, and a scheduled benchmark workflow. **Resolved** and regression-tested grading defects involving incorrect hidden-test paths, false zero scores, parser miscounts, and infrastructure failures misclassified as agent failures.

Education

Sardar Vallabhbhai National Institute of Technology (SVNIT), Surat

B.Tech., Mechanical Engineering
Aug 2023 – Present

Technical Skills

Programming & Testing	Python, TypeScript, JavaScript; Pydantic, pytest, JSON Schema, type validation, CLI design, automated testing
AI Evaluation & Agents	Coding-agent evaluation, LLM-as-judge, rubric design, hidden-test grading, partial-credit evaluation, failure taxonomy, patch and trajectory analysis, benchmark integrity; Aider runner integration
Benchmark Infra & CI	Docker, hermetic grading, read-only workspaces, network isolation, reproducible testing, CI, test-result parsing, failure classification
Backend & Full-Stack	Node.js, Express.js, PostgreSQL, Prisma, Next.js, React, REST APIs, WebSockets, JWT/RBAC, Zod, background jobs
Voice AI	Pipecat, Twilio, Groq, speech-to-text, text-to-speech, concurrent evaluation, real-time dashboards