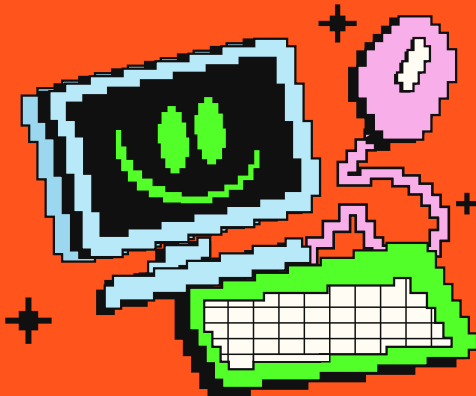


AI PRODUCTS. AGENTS. EVALUATION.

TANMAY SURTE.

AI ENGINEER / FOUNDER @ VIGIL

I build AI tools around real problems, learn what the work demands, and test where my first answer breaks. User feedback and failure cases shape what I build next.



FIRST IDEAS NEED GOOD TESTS.

GITHUB ↗

LINKEDIN ↗

VIGIL ↗

EMAIL ME ↗

NAVI MUMBAI, INDIA

tanmaysurte.ts@gmail.com

+91 93258 55194

01 / THE JOURNEY SO FAR

Learning by shipping.

Forward Deployed Engineer

Navayuvak Entrepreneurs

AUG 2026 - PRESENT / MUMBAI, HYBRID

- Designed a **five-stage platform** for under-25 founders with the Managing Director, including longitudinal assessment, intake, check-ins and routing.
- Specified mobile-first wizard UIs, voice input and WhatsApp-triggered check-ins; resolved an SSL outage caused by a stale AAAA record.

Founder

VIGIL ↗

Vigil / Reputation & AI perception

OCT 2025 - PRESENT / REMOTE

- Built a reputation-monitoring SaaS from scratch with **Next.js, FastAPI and Supabase**; probes ChatGPT, Gemini and Perplexity for brand perception.
- Prioritized crisis detection from PR and GEO agency feedback; launched a free brand checker. NeuroChain, its earlier prototype, placed **top 5 of 5,000 teams** at Mumbai Hacks 2025.

AI Engineer Intern

HaTa Bevtech / Hugg Beverages

MAR 2026 - MAY 2026 / MUMBAI

- Delivered booking and CRM portals for sales and nutritionist teams, with smart routing, payment links and retention analytics.
- Trained a buying-intent classifier with **0.94 ROC-AUC**; tuned thresholds against false-positive cost and integrated it into customer-support workflows.
- Shipped **5+ production automation pipelines**, including Gemini Batch ad classification into a 63-field schema; owned CI/CD on Railway and Vercel.

toolbox.txt

LANGUAGES

Python, TypeScript, JavaScript, SQL

PRODUCT

Next.js, React, FastAPI, Node.js, Supabase

AI & EVALUATION

LLM agents, RAG, PyTorch, scikit-learn, mutation testing

AUTOMATION & INFRA

Playwright, n8n, Docker, GitHub Actions, Vercel

education.txt

B.E. AI & Data Science

Terna Engineering College
University of Mumbai

2023 - 2027 (EXPECTED)

field_notes.txt

Agency feedback changed Vigil's roadmap. False-positive cost shaped my classifier thresholds. When an audit weakened a research result, I changed the claim and rebuilt the pipeline.

Build. Question. Test. Learn.

tireless / agent_qa.exe

The Tireless Hand.

WHAT I BUILT
An LLM mutation agent and **7 deterministic checks** for a drone flight cockpit.

WHAT LOOKED SHAKY
A cockpit could show "connected" while displaying stale data after a lost connection.

HOW I TESTED IT
Compared the screen with independent telemetry; opened sealed keys after blind runs. **No LLM in the test loop.**

WHAT WORKED
Caught **11/13 injected bugs**; zero false alarms across 3 clean/harmless builds. Documented both misses.

[CODE + TEST REPORT ↗](#)

driving_seat / claude_skill

Driving Seat.

WHAT I BUILT
An open-source **Claude Code skill** that defines measurable specs, logs routine choices and routes consequential decisions to a human.

WHAT LOOKED SHAKY
Silent defaults hide risk; asking about every choice invites rubber-stamping.

HOW I TESTED IT
Compared four scenarios with/without the skill, including risky migrations, delegated choices and eval-label edits.

WHAT WORKED
All four reported scenarios met their written checks. **One run each, self-graded**; early evidence, not a benchmark.

[SKILL + SCENARIO NOTES ↗](#)

Reel-MMA.

[VIEW CODE ↗](#)

Built a Python/GPT-4o/ffmpeg pipeline for 9:16 MMA highlights, with **human review**. Paused it when vision cost per reel was too high.

03 / RESEARCH

Can it explain a rule? Can it follow it?

Execution Divergence Metric. Preregistered 900 instances to compare LLM rule explanation with execution across 3 complexity levels, 3 token distributions and 2 model families: Llama 3.1 and Gemini 2.0 Flash.

An annotation audit found **45.5% inter-rater agreement**. I demoted the headline result and rebuilt the pipeline; the original claim could not stand on that measurement.

SUBMITTED TO ACL ROLLING REVIEW / TARGET: BLACKBOXNLP 2026

contributions.log / CHECKED 21 OCT 2026

Sentient #476 ↗	MERGED	Fixed user names in memory; added regression coverage.
StashBase #225 ↗	MERGED	Clarified task results; fixed Windows tests. Merged to the scheduled-tasks feature branch.
CryoSwath #52 ↗	MERGED	Added pixi setup and reproducible install instructions.
AWS Powertools #8617 ↗	OPEN	Proposed consistent parameter-error handling, with regression tests.

ALSO:

[Sentient #486 \(open\) ↗](#)

[Sentient #479 \(draft\) ↗](#)

[imposter #21 \(merged\) ↗](#)