

Pranit Sehgal

A I P R O D U C T E N G I N E E R

London, UK • +44 7920 835639 • pranitsehgal@gmail.com • pranitsehgal.com • linkedin.com/in/pranitsehgal • github.com/pranit1812

SUMMARY

AI product engineer with four years taking LLM products from first model to shipped, production software. At Nothing, rewrote the widget builder agent behind Essential Apps and **cut cost per widget 10x**, took it to **open beta** and on to **about 75,000 users**, then joined the AI OS team to build generative UI and on-device AI for the company's next operating system. Before Nothing, founding engineer at a US startup: first product from **zero to paying customers in six months**, then promoted to senior in four. Seeking AI product roles in London, with visa sponsorship.

SKILLS

AI and ML	LLM and agentic systems, tool use and function calling (Claude Agent SDK, MCP), retrieval-augmented generation over large document sets, evaluation harnesses and benchmarking, on-device and edge inference (Gemma, TensorFlow Lite), multimodal image and video generation, prompt engineering and prompt caching, machine learning research, generative AI, PyTorch
Engineering	Python, TypeScript, React, Node.js, Go, Kotlin and Android, Electron, Three.js and WebGL, React Native, backend services and REST APIs
Platform	AWS, GCP, Vertex AI, cloud deployment, CI/CD pipelines, Git, Docker
Working	Cross-functional work with designers and product managers, technical direction and ownership, mentoring, hiring and interviewing, live customer demos

EXPERIENCE

Nothing

London, UK

Senior Engineer, AI OS Team

JUN 2026 TO DATE

- Build generative UI and on-device agents for Nothing's next OS: on-device memory and the model's template layer.
- Prototype on-device image generation and local models including Gemma: offline, low latency, no data off device.
- Interview graduate engineering candidates for the team, reviewing their take-home assessments and running one of the later interview rounds. Ran the company's student hackathon with the team and built its judging tools.

Senior Engineer, Essential Apps Team

JAN 2026 TO JUN 2026

- Rewrote the widget builder agent from three chained agents into one orchestrator with a wide tool surface. Plain-English request in, working homescreen widget out. Shipped to **open beta**, now **75,000 users**.
- **Cut cost per widget 10x** through model routing, prompt caching and variable thinking budgets, each change validated on a benchmarking harness that scores every generated widget on rendered output and on code.
- Instrumented the agent end to end: **tool-call accuracy, latency and cost per widget**, with stress-test scripts and round-robin routing across regions to hold throughput under load.
- Added web search, image attachment and user images to widgets, plus **per-user and per-chat memory**.
- Built image and video generation for widgets: agent-authored keyframes with video interpolation between them, producing animated homescreen widgets **new on Android**, used in leadership and investor demos.

HyperWater.ai

US, remote

Founding AI Engineer

NOV 2024 TO DEC 2025

- First engineer at HyperWater, an AI bidding network for commercial construction. Took it from prototype to production and set technical direction: Python for inference, Go for the service layer. **First revenue in six months**.
- Shipped the **retrieval** pipeline behind bid recommendations: four extraction backends with escalation on failure, drawings and tables preserved as first-class chunks rather than flattened text, and dense and BM25 results fused so every answer cites the exact sheet. **Cost fell 85%** by analysing drawings only on demand.
- Promoted to senior in four months, then led **eight developers** while still writing code.
- Fed contractor corrections back into the recommender so later bid calls reflected them.
- Ran the contractor demos that converted into the company's first revenue.

RoomieHub

US, part-time contract

Software Developer

JUN 2024 TO JUN 2025

- Contract engineer on a roommate-matching app past **10,000 users and £150k ad revenue** in eight months.
- Built in-app AI messaging on self-hosted LLaMA models rather than a paid API, lifting engagement **45%**, and stood up the team's first CI/CD pipelines, cutting deployment time 70%.

Earlier roles

Mumbai, India

Data and automation engineering

2020 TO 2022

- Kalpataru Power Transmission: company-wide support chatbot cutting resolution time **30%**.
- Metropolitan Stock Exchange of India: price forecasting and automated market data pipelines.
- The Starter Labs: ad-spend models worth **20%** better return; segmentation lifting lifetime value 10%.

SELECTED PROJECTS

Mindscape

Three.js, Claude Agent SDK, node-pty

A walkable 3D world where every terminal is a live shell. An AI agent rebuilds the world on request; a companion smartwatch app approves its actions, keeping long runs supervised away from the desk.

Notebook

React, Excalidraw, MCP

An infinite canvas where the model answers by drawing: diagrams, ink notes, comments pinned to sketches. An arrow drawn at anything makes it context. Claude Agent SDK over a custom canvas server.

Aesthetic Camera

TypeScript, Claude

Photograph an object and the interface restyles to match. A deterministic k-means pass in OKLab extracts the palette, one model call turns it into design tokens, and a contrast solver holds every text pair above WCAG 4.5:1.

Glyph Agent

Kotlin, GlyphMatrix SDK, Vertex AI

A described mood becomes LED matrix animations, sound EQ profiles and matching camera colour presets on a Nothing phone, generated in a single pass on the public GlyphMatrix SDK.

Foresight for Fires

PyTorch, Three.js, NVIDIA London hackathon

Android dispatch client shipped in two days at NVIDIA Hack London: Kotlin WebView over a self-contained risk map, voice assistant and station picker, against the team's ward-level risk model.

bitchat-geo

Swift, Kotlin, agent loop

Geo mesh mapping added to the iOS and Android forks of open-source bitchat, driven by a self-written agent loop running one task per fresh context. Every iteration gated on the full Swift suite staying green across 1,622 tests.

Construction RAG

Python, Whoosh, PaddleOCR, github.com/pranit1812/Pranit-RAG

Question answering over construction bid packages. Every file routes through one of four extraction backends with escalation on failure, and dense vectors are fused with BM25 by reciprocal rank so each answer cites its exact sheet.

Rhythm Password

TensorFlow Lite, Arduino

A PIN tapped in rhythm, where the timing between presses is the credential. **92% accuracy**, tempo-invariant so the same rhythm unlocks whether tapped fast or slow, running offline on a stamp-sized chip.

Signal Grid

Electron, node-pty

A spatial desktop where live terminals sit on an infinite canvas, so a dozen jobs are watched at once.

Thought-space

Agents, Stable Diffusion

Daily voice notes organised by an agent into a searchable archive, illustrated with locally generated images.

EDUCATION AND PUBLICATION

Arizona State University, MS in Computer Engineering, GPA 3.57/4.0

2022 TO 2024

Peer-reviewed conference paper: Gradient Boosting Approach for Sentiment Analysis in Job Recommendation Systems, published in the proceedings of the IEEE Bombay Section Signature Conference.

BEYOND WORK

Former **professional Counter-Strike and Valorant team captain**: strategy, preparation and calling plays for a five-person team in live matches. Regular hackathon competitor, most recently at NVIDIA London 2026.