

ISHAN SHARMA

Applied AI Engineer · Forward Deployed Engineer

ishan@bloortech.com | ishan.website | [LinkedIn](#) | [GitHub](#)

SUMMARY

Full-stack and AI engineer who ships production systems end to end and verifies they work. Built and led a multi-tenant AI agent platform from zero to live client deployments, and earlier shipped to 500K+ users at Amazon. Strong across backend, frontend, and applied AI, with the customer-facing judgment to take a vague brief to a deployed outcome.

EXPERIENCE

Founder & Lead Engineer, Bloort

2025 – Present

AI agent platform with RAG roots, built and led from zero to production.

- Shipped a multi-tenant AI agent platform end to end on live client sites (RAG chatbot and agent sandbox over per-client Pinecone namespaces); cut client onboarding from a multi-day cycle to a ~5-minute flow that auto-brands each bot from its site
- Built an evaluation harness scoring retrieval on versioned golden datasets (hit-rate, MRR, recall@k) and rejected a hybrid retriever that scored worse than dense search on an already-clean corpus
- Gave the agent guardrails, a failure-mode taxonomy, tool calls over the Model Context Protocol (MCP), and outbound VAPI voice / Twilio WhatsApp actions; an agent-trajectory eval caught a hallucinated tool argument (an invented appointment date), and the fix took the 6-scenario suite from 5/6 to 6/6
- Hardened it for production: layered prompt-injection defense (LLM guard plus regex fallback), SSRF and email/phone PII filtering, a 0.88-cosine semantic cache that cut LLM cost and latency, retry-with-backoff, timeout races, and dry-run gating on irreversible actions
- Built the first version solo, then brought in and led two part-time engineers; owned the full cycle: architecture, delivery, and direct customer conversations, iterating on production feedback

Highlights: *Next.js, TypeScript, LangChain, OpenAI, Pinecone, MCP, VAPI, Twilio, Firebase, Railway, RAG, multi-tenancy, prompt-injection defense, agent + retrieval evals*

Full-Stack Engineer, Moonvalley (VC-backed)

2024

Generative video. A startup building its own proprietary AI video model.

- Built a generative-AI canvas application with real-time image generation and editing; cut time-to-market roughly 30% with a modular backend

Highlights: *Vue.js, FastAPI, Python, Stable Diffusion, Replicate, Docker, AWS*

Software Development Engineer, Amazon.com, Inc.

2022 – 2023

Consumer hardware. Amazon Kids+, an all-in-one service for kid-friendly content on Kindle e-readers.

- Shipped landing-page and authentication features for Amazon Kids+ on Kindle to 500K+ global users (React Native, TypeScript, C++), contributing to a ~20% revenue increase
- Drove a ~15% increase in successful sign-ins by integrating accessibility-compliant enhancements into the Kindle Kids+ authentication flow, and led A/B testing that improved user retention ~10%
- Improved software reliability ~25% with robust testing frameworks and reduced UX issue latency ~10% via AWS CloudWatch and SQS monitoring

Highlights: *React Native, TypeScript, C++, Java, AWS, CloudWatch, SNS, SQS, Jest, TDD, Agile, Jira*

PROJECTS

HANDSCENES

handscenes.vercel.app

Real-time hand- and body-tracked apps, running fully in the browser.

- Built Sign School, a Duolingo-style ASL fingerspelling trainer that reads your handshape live from MediaPipe hand landmarks, plus hand- and body-tracked scenes; all deps vendored behind a strict CSP (three.js, WebGL2, GLSL, p5)

SKILLS

Applied AI: RAG pipelines, agent orchestration, evaluation harnesses, vector databases, prompt engineering, LangChain, OpenAI APIs, Pinecone, Firecrawl

Full-stack: TypeScript, Python, Next.js, React, Node, FastAPI, Firebase, PostgreSQL

Infra & delivery: AWS, Vercel, Railway, Docker, CI/CD, production observability

EDUCATION

University of Toronto, Computer Science (B.Sc., Honours)

2017 – 2022