

VATSA JOSHI

AI Engineer · Multi-Agent Systems · Production LLM Platforms

Bengaluru, India | +91 97375 57799 | vatsajoshi2@gmail.com

linkedin.com/in/vatsa-joshi | github.com/Vatsa10 | vatsa.online

SUMMARY

AI engineer with 2+ years shipping production LLM systems: multi-agent orchestration, tool-calling agents with human-in-the-loop approvals, durable agent runtimes, grounded hybrid RAG, LLM eval harnesses and cost-engineered inference. Own agentic products used by 2,000+ users; built billing-integrated (Razorpay) SaaS modules; cut legal-NER extraction cost 25x over an 8M-document corpus at 96% zero-failure.

EXPERIENCE

AI Engineer — Optivus Technologies

Jun 2026 – Present

AI products & services startup; first engineer after the founders · Remote

- **Veritas** (multi-tenant marketing SaaS, 2,000 users): own the product end to end (Next.js, TypeScript, Node.js, MongoDB, OpenAI); built the AEO, Ads and Local SEO modules and the chatbot-native rebuild.
- **Vera agent**: architected a tool-calling agent (OpenAI Agents SDK) that routes queries to live SEO, AEO and Local data, verifies every stated number against tool outputs, and compiles multi-step requests into a single human-approval plan card with upfront cost; ~15 paise per query, English and Hinglish.
- **Agent runtime**: built a durable plan executor on MongoDB compare-and-set with idempotent steps, automatic retries, stop and resume-from-failed-step; LLM eval harness gates releases (plan evals 15/15; 48 routing tests, 0 ungrounded answers).
- **AEO module**: tracks how ChatGPT, Perplexity, Gemini and Google AI Overviews answer buyer queries; visibility score (breadth × depth), competitor share of voice, cited sources, answer churn, content gaps and claim checks that flag wrong AI statements.
- **Ads & Local modules**: Meta and Google Ads analytics, AI-generated creatives with pre-publish previews, and a paid add-on on **Razorpay billing**; Google Business Profile health scores, geogrid rank scans, review-theme analysis and approval-gated AI replies.
- **Aethonix (multi-agent company brain)**: built a multi-tenant WhatsApp connector turning chats, voice notes, images and documents into speaker-aware episodes for hybrid RAG (FastAPI, Celery, Postgres, OpenSearch). Architected summarizer, nightly memory-reconciliation (decisions, commitments, cross-source identity linking) and cited tool-calling chat agents; content-hash caching, batched budget-capped resumable runs, tuned on a 60-question eval.
- **FlowFin & Canary**: built agentic workflows for FlowFin (finance approvals, procurement, invoice automation) and Canary (AI lead capture and qualification).

Founding Engineer (Contract) — Starborn

Jun 2026 – Present

US behavioral-science parenting startup · Remote

- Built the core platform (Next.js, TypeScript, Supabase, Claude): behavior tracking, Standard Celeration Chart-style log-scale analytics, co-parent accounts, and a 12-screen assessment flow ending in a scored verdict; reviewed by a behavior scientist.
- **Minibooks**: multi-agent book engine (BAML) — Outline → Research → Writer ⇌ Critic → Assembly — producing cited 25–30 page books; research agents autonomously call search and claim-verification tools, with a deterministic filter dropping untraceable citations; live Author Studio.
- **Cognitive**: per-child agents converting behavior logs into functional-analysis profiles; planner → writer → critic course generator with per-claim citation checks; clinical moderation agent grounded in AAP and BACB standards gating publication.
- Guardrailed LLM features: voice assistant with a deterministic output validator, RAG over a 23-book / 14k-passages corpus, and a case chat at ~\$0.0001/message with on-device answers, fixed safety responses and a live-model eval suite.

Founding AI Engineer — Maaya AI

Apr 2026 – Jun 2026

- Owned the agent memory and retrieval layer over meetings, emails and tasks; built recall evals (precision/recall@k, LLM-as-judge, regression gates) and LoRA-tuned embedding and reranker models.

Software Engineer Trainee (AI) — Promact (BharatLaw.ai)

Jan 2026 – May 2026

- Cut extraction cost 25x (₹1 → ~4 paise/doc) across 8M judgments by collapsing a 7-stage LLM pipeline into 2 calls; 96% document-level zero-failure on a 4M-row eval set.
- Built multi-agent legal systems (drafting, judge profiling, litigation strategy) used by 50–60 law firms, with DB-grounded extraction eliminating hallucinated fields.

AI Engineer (Research) — Arizona State University × Navrachana University

Jun 2025 – Jul 2026

- Grounded health assistant: QLoRA fine-tuning (Unsloth, NF4), groundedness and hallucination evals, prompt-injection red-teaming and clinician escalation.

Internships: Hiraeth Tech (Full Stack, 2025) · Produgal AI (Deep Learning, 2025) · Sparks To Ideas (AI/ML, 2025)

SKILLS

Agents: OpenAI Agents SDK, LangGraph, BAML, CrewAI, FastMCP, tool-calling, multi-agent orchestration, human-in-the-loop, durable execution

LLM Systems: Hybrid RAG, GraphRAG, evals (LLM-as-judge, regression gates), guardrails, caching & cost optimization

Training: PyTorch, HF PEFT/TRL, LoRA/QLoRA, Unsloth, DeepSpeed ZeRO

Infra & Data: Python, TypeScript, Rust, FastAPI, Next.js, Node.js, Celery, Postgres, MongoDB, OpenSearch, Qdrant, Redis, Docker, AWS

PROJECTS, PUBLICATIONS & AWARDS

- Hindi2Guj-Qwen3.5 (DeepSpeed distributed fine-tune) · Qwen3-VL-2B encoder isolation · Rustformer (transformer in Rust)
- IEEE Xplore: EfficientNet-B7 + Bi-LSTM Deepfake Detection (DOI 10.1109/CCIC68129.2026.11486188)
- 2nd Place, Tinkerthon 3.0 +SSIP Gujarat grant · Top 10, Odoon × NMIT Hackathon

EDUCATION

B.Tech, Computer Science — Navrachana University, Vadodara

2022 – 2026