Vaibhav Lalwani

AI & Full-Stack Engineer · MSc Advanced Data Science & AI, University of Liverpool

Liverpool, UK · +44 7544 537 860 · vaibhavlalwani26969@gmail.com · vaibhavlalwani.vercel.app · github.com/vaibhav4046 · linkedin.com/in/vaibhav-lalwani

Open to: Full-time | Part-time | Contract | Remote | Hybrid | On-site

PROFILE

AI engineer building production language-model systems end-to-end — inference, retrieval, evaluation harnesses and the front ends that ship them. Comfortable across the model–system boundary. Author of an open-access preprint on on-controller transformer inference (NEXUS, 2026). Two years freelancing 20+ LLM-powered apps for paying clients on Claude, OpenAI and LangChain.

PUBLICATION

Lalwani, V. (2026). *NEXUS: On-Controller Transformer Inference and Speculative Edge Execution for Console-Wired Latency Parity in Cloud Gaming.* Zenodo. DOI 10.5281/zenodo.20059414 · ResearchGate. Extends the Outatime line (MobiSys 2015) with on-controller transformer inference and speculative edge execution.

SKILLS

LANGUAGES	Python · TypeScript · JavaScript (ES6+) · SQL · HTML · CSS
AI / ML	OpenAI · Anthropic / Claude · Gemini · Groq · LangChain · LangSmith · RAG · structured outputs · MCP servers · Whisper STT · PyTorch · transformer inference
BACKEND	FastAPI · Node.js / Express · Postgres · pgvector · Redis · BullMQ · Supabase (RLS) · vector search
FRONTEND	React 19 · Next.js 15 / 16 · Tailwind CSS · Vite · REST + GraphQL · accessibility
DEVOPS	Docker · GitHub Actions · Vercel · GCP Cloud Run · AWS · Sentry · PostHog · Jest / Vitest · Playwright · Puppeteer

EXPERIENCE

Full-Stack / AI Engineer Intern · Meta Solution Technologies — Apr 2026 – Present
Remote, UK

- Building an AI-guided admissions platform on Next.js 15 + React 19 + TypeScript. Chat-style guidance flow on OpenAI with structured outputs, served from a Python FastAPI service against Postgres + pgvector.
- Wrote auth end-to-end (Google OAuth, magic-link, PKCE) on Supabase, plus a BullMQ + Redis async queue driving document checks, generation jobs and email triggers.
- Shipped a 10-language i18n layer and a dual-theme CSS system so non-engineering teammates roll out copy and translation updates without engineering involvement.

AI Engineer · Freelance / Self-employed — May 2024 – Jan 2026
Remote

- Designed and shipped 20+ LLM-powered applications for paying clients across SaaS, e-commerce and consulting. Most pair a React or Next.js front end with a Python service running Claude or OpenAI APIs, structured outputs and a RAG retrieval layer.
- Built longer-running AI agents for lead generation, customer support and data processing. Each project shipped with a README, a walkthrough video and a written hand-off doc.

Software Engineer Intern · Recruit Pilot — Jan 2026 – Apr 2026
Recruitment Technology · Remote

- Built React + TypeScript UI features against a REST API. Code review and weekly sprint cadence alongside more senior engineers.

SELECTED PROJECTS

apex — autonomous job application engine. CLI driving real Chrome via Playwright, generates a tailored 1-page resume per job (Puppeteer markdown → ATS-safe PDF), fills custom LinkedIn Easy Apply questions with an LLM (profile mapping → Q+A cache → free-LLM fallback), submits autonomously until LinkedIn's daily cap. Free LLMs only. github.com/vaibhav4046/apex

Praxon — open-source AI agent platform. A Claude Cowork alternative on Next.js 16 + React 19 + TypeScript. Multi-LLM router across free providers (Groq, Cerebras, Gemini, Ollama) with auto-fallback; MCP-native tool layer; 3-schema Postgres + RLS for tenant isolation; cloud-deployable on Vercel + Supabase. Live · Code

Cogniloop — Socratic study tool. Locked-prompt evaluator grading free-form student answers 0–3 with explanation; tracks concept mastery (weak → mastered) across sessions. Single-prompt design — no agent loop, no retrieval — keeps median latency under one second on Groq Llama 3.3 70B. Edge runtime, Next.js 16, KaTeX. Live · Code

MCP Marketplace — registry of 800+ Model Context Protocol servers. Daily auto-sync from Glama and the official MCP repo, normalised tool schemas, one-line install snippets for Claude Desktop, Cursor and Claude Code. Next.js 15 RSC, Cmd-K palette, dynamic OG cards. Live · Code

EDUCATION

University of Liverpool — Jan 2026 – Jan 2027
MSc Advanced Data Science & Artificial Intelligence

Christ University, Bengaluru — 2022 – 2025
Bachelor of Computer Applications · CGPA 8.7 / 10