Devansh Vishwakarma

AI Engineer | AI Agents · LLM Evaluation · Production AI Pipelines · RAG · LLM Fine-tuning
Noida, India | vdevwish12312@gmail.com | +91 75259 36363 | github.com/dv2202 | linkedin.com/in/devansh-v | huggingface.co/rambo2202

SUMMARY

AI Engineer with **2.7+ years** shipping **LLM-backed features to production** in TypeScript and Python. At Forest AI I build and own **tool-calling AI agents** run in production by **AXA Group** and Alteos, plus the **evaluation harnesses and release gates** that decide what ships; my evals raised agent accuracy from **78% to 95%**. I also **fine-tune** small LLMs with LoRA and QLoRA.

SKILLS

AI Agents & Pipelines: AI agents, tool calling, deterministic workflows, multi-step AI pipelines, workflow orchestration (LangGraph), Vercel AI SDK, Mastra, LangChain, LLM APIs (Gemini, OpenAI, Anthropic), multimodal / document understanding (OCR, vision), RAG, retrieval systems, hybrid search, Pinecone, FAISS, pgvector
Evals, Safety & Reliability: evaluation harnesses, evaluation datasets and metrics, release gates, LLM-as-a-judge, backtesting and replay, regression suites, red-teaming, prompt-injection testing, idempotent / retryable / resumable pipelines, prompt and dataset versioning, observability (Langfuse)
Fine-tuning: LoRA, QLoRA, PEFT, 4-bit quantization, PyTorch, Hugging Face Transformers, MLX, precision / recall / F1
Engineering: TypeScript, Python, SQL, backend services (Node.js, FastAPI, PostgreSQL, MongoDB), Docker, AWS (SNS, Lambda), GCP Cloud Run, CI/CD, React, Next.js, Playwright

WORK EXPERIENCE

Forest AI — AI Engineer | Aug 2025 – Present

- Built and shipped **tool-calling AI agents** to **production** (Vercel AI SDK, Gemini) for AXA Group and Alteos insurance claims across WhatsApp, Telegram and web, with **provider fallbacks** for model failures.
- Designed the email agent as a **deterministic LangGraph workflow**: a **multi-step pipeline** that is **idempotent, retryable and resumable after failure** via checkpointing.
- Designed the **evaluation harness** and backtesting framework with versioned **evaluation datasets** (regression suites, benchmarks, adversarial sets) and task / tool-call metrics; raised **agent accuracy 78% → 95%** and cut manual testing ~70%.
- Used eval results as **release gates**, blocking releases on system-prompt leakage under **prompt-injection** turns and on **latency regressions** traced via Langfuse.
- Made model behaviour **reproducible** by **versioning** prompts (Langfuse) and eval datasets (config) and **replaying** past conversations to reconstruct decisions.
- Built an agent-simulation engine (Mastra, ElevenLabs) monitoring **8 production agents** with 50 persona conversations/day, scored by **LLM-as-a-judge** to catch **outputs that vary for the same input**.
- Built a **multimodal document-understanding pipeline** (OCR, vision models, deterministic rule checks) for per-claim risk scores that **fails closed** on timeouts or outages, and a **RAG retrieval system** on Pinecone at ~60 ms p50.
- **Fine-tuned** Gemma 3 1B with **LoRA** for 13-intent JSON routing: **macro-F1 0.455 → 0.962**, valid JSON 12.5% → 100%, 2.2x lower latency.
- **Owned agents after launch**: **observability**, AWS SNS alerting for **production issues**, GCP Cloud Run deploys, CI gates and Playwright end-to-end tests.

Userology — Software Engineer | Aug 2024 – Jul 2025

- Debugged **production issues** in an **LLM-backed** user-research analysis pipeline, raising **accuracy to 95%** by tracing data-quality failures from raw transcripts to model outputs.
- Rebuilt the campaign-creation flow (Next.js), improving recruitment efficiency **~30%**, and cut re-renders **~40%** by migrating state to Zustand.

Infitout Financial Technology — Frontend Developer | Mar 2024 – Aug 2024

- Built the frontend of 1lyoptions, a real-time options-trading terminal on REST APIs, and a strategy-builder module reused across products.

PROJECTS & OPEN SOURCE

BotWebAI, Talk to Website | Next.js, LangChain, FAISS, pgvector, Gemini | Live

- Shipped a multi-tenant SaaS turning any website or PDF into an embeddable **RAG pipeline**, with **hybrid vector + keyword retrieval** (reciprocal rank fusion) and **idempotent** content-hash sync.
- Protected ingestion against **adversarial input through system design**: SSRF-safe URL fetching and HMAC-signed visitor tokens.

ClauseFinder | Qwen2.5-1.5B, QLoRA, MLX | GitHub

- **Fine-tuned** Qwen2.5-1.5B with **QLoRA** (4-bit) on 1,626 CUAD contracts; held-out **F1 0.664 → 0.924** (beats 3-shot at 0.740), **hallucinated quotes 21.7% → 0%**.

peerBench (arXiv:2510.07575): built dataset upload and auto-generated evaluation jobs for an open-source LLM benchmarking platform. **Electron:** docs-site fix, PR #771 merged.

EDUCATION

Jaypee Institute of Information Technology — B.Tech, Computer Science | Aug 2021 – May 2025