Ishaan Chaturvedi

AI Engineer — Production LLM & Agentic Systems

London, UK · ishaan27jobs@gmail.com · +44 7377 667721 · LinkedIn · GitHub · nullsutra.com

PROFILE

AI Engineer specialising in production LLM and agentic systems, with **7+ years** across enterprise machine learning and independent product building. I design, ship, and evaluate end-to-end AI inside regulated financial environments — self-hosted LLM deployment under GDPR, document-intelligence and agentic pipelines, and enterprise-wide AI use-case discovery — focused on the parts that decide whether AI products survive production: **evaluation, observability, structured outputs, and cost control**. Independently, I architect and operate **Alter**, a self-hosted production AI "second brain," and design generative-media tooling (**Spiral**). Fluent across the modern stack — Claude, GPT, local models, agentic frameworks, RAG/retrieval — and AI-native build tools (Cursor, Claude Code) applied with full specification discipline.

EXPERIENCE

PRA Group · London, UK — Oct 2022 – Present

AI Engineer — Global AI Tiger Team — 2025 – Present

- Selected onto PRA's cross-organisation AI Tiger Team (reporting to an SVP) to identify and ship LLM and agentic AI use cases across US and international operations on an Azure stack — Azure AI Foundry, Microsoft Fabric, Google ADK, Microsoft Agent Framework, and Copilot Studio.
- Deployed a **self-hosted, multi-model LLM stack** on 4× NVIDIA Tesla T4 GPUs to keep regulated data in-environment under GDPR — routing tasks to purpose-specific models (a fine-tuned Mistral for summarisation, DeepSeek-R1-Distill-Qwen for reasoning) — owning model selection, fine-tuning, prompt design, and output evaluation.
- Led a company-wide AI discovery initiative across every department, surfacing **15 high-value use cases** each quantified with effort–value scoring to drive leadership prioritisation and shape the delivery roadmap.
- Built an LLM + OCR pipeline over court documents that extracts and classifies legal failure reasons into a tiered taxonomy, enabling trend analysis by date and geography and proactive intervention.
- Developed agentic LLM pipelines for affluence detection — cross-referencing communications, bureau, and internal signals to surface inconsistencies in token-payer behaviour — and full customer-journey summarisation for frontline staff.
- Built an LLM pipeline to extract, translate, and map foreign-language seller files to the internal data model, automating a previously manual onboarding process.

Data Scientist — Decision Science & AI — Oct 2022 – 2025

- **Portfolio valuation & forecasting:** Led end-to-end pricing of non-performing loan portfolios using KNN peer selection, pay-curve generation, and calibrated recovery models (breakage, legal vs. voluntary) — supporting 5 of 8 acquisitions now performing at ~98% of projected value.
- **Monte Carlo simulation:** Developed a state-based simulation using a classifier for account state transitions and a regressor for payment amounts; benchmarked XGBoost, Random Forest, and a Hugging Face Time-Series Transformer.
- **Book reforecasting:** Designed bespoke forecasting models across 5 lines of business (voluntary, legal, DMC, DCA, insolvency) with backtesting validation — now a standard monthly production process.
- **Collections decision science:** Built WoE-based scorecards (avg AUC-ROC 0.74) for propensity-to-pay, legal referral, and settlement; segmented customers into 8 hierarchical clusters; optimised dialler contact limits per segment; and constructed multi-dimensional customer-journey time series for next-best-action analysis.

ARCON TechSolutions · Mumbai, India — Jan 2018 – Aug 2021

Data Scientist

- Built near-real-time anomaly-detection models for a privileged-access-management product, moving threat detection from static rules toward predictive, behaviour-based scoring of access events.
- Developed an online hierarchical-clustering system for user-behaviour analytics — profiling users into behavioural cohorts and flagging anomalous deviations to inform access-elevation and restriction rules.
- Led a two-person team to design and ship fraud-detection analytics tooling (drill-down investigation views), later adopted across the wider product suite.

PROJECTS

Alter — Production AI "Second Brain" Lead / Primary Author

- A self-hosted, production AI "second brain" on a **TypeScript / NestJS microservice architecture**: an LLM recognition-and-orchestration layer classifies unstructured input, routes it across five interaction modes, and auto-compiles it into a per-user knowledge wiki with semantic search.
- **Agentic orchestration / harness:** Built the orchestration layer that runs multi-step, multi-mode LLM workflows — with automatic re-grounding on long-running reasoning, branching, and end-of-thread synthesis — driving recognition, briefing, research, and decision-cascade flows.
- **Prompt evaluation:** Backed every prompt and mode with golden eval suites run as CLI eval runners with per-run cost budgets, treating prompts as tested, versioned components rather than ad-hoc strings.
- **Multi-provider orchestration & structured output:** Designed a provider abstraction (Anthropic / OpenAI / Google / Perplexity) with complexity-based cost routing and prompt caching, enforcing strict structured output via provider tool-use + Zod schemas with a retry-repair flow and safe fallback — never prompt-only JSON.
- **Retrieval (RAG):** Built the retrieval layer — pgvector-backed hybrid (semantic + keyword) search over an auto-compiling wiki, fed by an event-driven reindex pipeline.
- **Production-grade engineering:** Encryption-by-default on all content, CI-enforced privacy invariants, per-user LLM cost caps with graceful degradation, and a full specification / ADR process — built to run in production, not as a prototype.

Spiral — AI Filmmaking Tool Product Design & Design Engineering

- A 0-to-1 product design for AI-assisted filmmaking, architected around the way creative work actually proceeds — as iterative revision loops rather than a linear prompt-to-video pipeline — where the system surfaces its reasoning at each step so the user's judgment compounds.
- **Agentic core:** Designed the central mechanic — a typed dependency graph where edits propagate through three AI-driven cascade modes (auto-propagate, flag-for-review, hard-invalidate), and a "converge" synthesis step that clusters unresolved changes into additions, refinements, and conflicts, computes downstream consequences, and walks the user through resolution.
- **Consistency & cost model:** Introduced "locks" as first-class consistency primitives (style, character, motif, format, voice) that auto-inject into every downstream LLM prompt; defined a clean generation boundary separating free, auto-propagating text operations from explicit, user-triggered paid generation.
- **Design engineering:** Built the design through 7 interactive prototype iterations (dependency-free HTML/CSS/JS with a full design system) and validated the methodology by producing a complete short film inside it.

Generative Media Pipelines

- Directed and produced 2 AI short films end-to-end (story → moodboard → scene-by-scene prompt engineering → consistent keyframes → ElevenLabs original scores); built an automated content pipeline (trending-topic scrape → story → AI video → audio alignment → auto-publish).

SKILLS & TOOLS

AI & LLM Multi-provider LLM orchestration · Agentic harness / multi-step workflows · Structured outputs (Zod, tool-use) · Prompt engineering, evals & cost budgeting · RAG, hybrid & semantic search (pgvector) · LLM fine-tuning & self-hosting · OCR pipelines · LangChain · Anthropic / OpenAI / Google / Perplexity APIs · ElevenLabs

BACKEND & SYSTEMS TypeScript · NestJS · Python · PostgreSQL · Redis / Bull · Microservices & event-driven architecture · SSE streaming · REST APIs · System design / ADRs

MACHINE LEARNING XGBoost · Random Forest · Logistic Regression · KNN · Hierarchical clustering · Monte Carlo simulation · WoE scorecards · Time series · Anomaly detection

MLOPS & CLOUD Azure (AI Foundry, Fabric) · MLflow · Docker · GitLab CI · pytest / Jest / Playwright (TDD) · Experiment tracking & model monitoring

GENERATIVE AI AI video generation · AI music composition · Automated content pipelines · Prompt & storyboard design

REGULATORY GDPR · EU AI Act (regulated-finance AI delivery)

EDUCATION

MSc Artificial Intelligence (Distinction) · Queen Mary University of London 2021 – 2022

Dissertation: long-term planning agent for sparse-reward environments using Monte Carlo Graph Search & Quality-Diversity search — directly relevant to modern agentic systems.

BTech Computer Science & Engineering (Distinction) · Vellore Institute of Technology, India 2013 – 2017