Andrei Alexandru

AI Engineer

London, UK · alexandruciprian61@gmail.com · andreialexandru.com · linkedin.com/in/andrei-ciprian-alexandru-57407a186 · github.com/Ciucky

SUMMARY

AI Engineer with 4+ years building production AI systems. Currently building LLM pipelines on AWS at Citation Group, and co-founder and engineering lead of Renda. Previously developed RAG-based applications at RLB in London, and worked on Computer Vision and LLM projects at Jaguar Land Rover.

TECHNICAL SKILLS

Languages	Python, TypeScript, JavaScript, C, SQL
AI / ML	LLMs & Prompt Engineering, RAG Systems, LLM Orchestration, Agentic Systems, Structured Output, Computer Vision, Model Fine-tuning, MLOps
Cloud	AWS Bedrock, AWS Lambda, Step Functions, Cloudflare Workers, Google Vertex AI, R2 & KV, Terraform, Azure OpenAI, Azure AI Foundry
Frontend	Next.js, React 19, Vite, Flutter, Tailwind CSS, Framer Motion
Tools	Git, GitHub Actions, Claude Code, pytest, Vitest, Hono, Supabase, Polar

PROFESSIONAL EXPERIENCE

AI Engineer

Apr 2026 to now

Citation Group — Remote

- Main engineer on a multi-agent LLM system (Python, AWS Bedrock, Step Functions, Lambda) that briefs employment-law advisers on 1,000+ cases a day, taken from pilot to full production with every quality gate past target
- Migrated every agent to a newer Claude model, cutting errors in blind review and pipeline failures each by more than half, cost per case by 17% and latency by 40%
- Ran evaluation and quality engineering for the programme: an LLM-as-judge from a different model family calibrated to expert scores, weekly checks of AI output against expert decisions, adversarial test suites, and blind grading of prompt changes against the previous version
- Built an LLM pipeline in under three weeks that turns public health-and-safety data into sales signals; its first overnight run produced signals for 11,000+ leads
- Built an AI referral engine and took it from proof of concept to a production project; its referrals have already led to over £50k in sales
- Built three further LLM proofs of concept, for inbox triage, construction safety plans and ISO readiness; the triage classifier agreed with the team's own decisions 93 to 95% of the time and moved to a production build

Co-founder / Engineering Lead

Sep 2025 to now

Alimio Ltd — London

- Co-founded a two-person UK company and built the technical side of both products
- Renda: SaaS that turns Claude Design exports into decks, carousels and PDFs, past 1000 users on organic acquisition alone
- Alimio: AI food waste app with photo scanning and voice input, published on the App Store and Google Play
- Own the full stack for both: architecture, edge infrastructure, LLM layer, billing and deployment

AI Developer

Mar 2025 to Apr 2026

RLB (Rider Levett Bucknall) — London

- Built RAG-based POC applications for enterprise clients
- Deployed and fine-tuned LLMs on Azure
- Full Azure AI stack (OpenAI, AI Foundry, AI Vision, AI Speech)

Graduate AI Engineer

Sep 2023 to Mar 2025

Jaguar Land Rover

Coventry

- Computer Vision research in the AI department
- Built and evaluated LLM applications for internal tooling
- Contributed to AI Governance frameworks and EU AI Act compliance

Software Intern (Industrial Placement)

Jun 2021 to Aug 2022

Jaguar Land Rover

Coventry

- Used ML to optimize Jaguar I-PACE traction control system
- Wrote C code for traction control algorithms
- Contract extended by 2 months due to performance

PROJECTS

Renda `tryrenda.com`

2026 · Co-founder / Engineering Lead

- 1000+ users and recurring subscribers across multiple countries, all through organic acquisition
- Architected an AI plus runtime split: an LLM emits a small JSON render recipe that a fixed, fully-tested headless-browser runtime executes, so new design patterns need no new code

`TypeScript, Cloudflare Workers, Hono, React 19, Vite, Google Vertex AI, Supabase, Polar`

Alimio `alimio.app`

2025 · Co-founder / Engineering Lead

- Published on App Store and Google Play
- AI photo scanning that auto-detects products and expiry dates

`Flutter, Azure AI Vision, Azure AI Speech, GPT-4, Azure AI Foundry`

No-Numb `github.com`

2026 · Open Source

- 32 stars and 100+ unique cloners on GitHub
- Stop hook intercepts only turns that edited or wrote a file, so pure conversation is never quizzed

`Bash, jq, Claude Code Hooks, Plugin Marketplace, Markdown Skills, JSON`

Nuree AI. Real-time voice AI platform for ADHD focus.
Proactive AI Assistant. A self-hosted agent that decides when to interrupt me.
Open Banking Finance Platform. Aggregates four bank accounts and categorises spending with an LLM.

EDUCATION

BSc Computer Science

First Class Honours

University of Surrey

- **Dissertation:** Explainable AI / Neural Network Interpretability. I wanted to know why a neural network decided what it decided. So I built a way to pull a readable decision tree out of a trained one, and tested it on image classification and on bank fraud.
- Reduced decision tree complexity by up to 70%. Held accuracy within a few percentage points of the original networks. Relevant to XAI, model interpretability and EU AI Act compliance.