Rob Gilks

Software Engineer & Technical Lead · AI products, applied ML, edge systems

London, UK · rob.gilks@gmail.com · linkedin.com/in/rob-gilks · tre.systems

PROFILE

Software engineer and technical lead with over 25 years of shipping systems that get used, across education, finance, healthcare and media. Most recently led the team behind the APIs powering AI auto-marking of written and spoken English exams at Cambridge University Press & Assessment - a hands-on lead, top committer on the core services while running the platform to 99.98% uptime. Now building AI products end-to-end through my own studio: fine-tuning transformer models, deploying them on serverless GPUs and the edge, and shipping polished web apps around them. Documentation-first: every project ships with architecture docs, decision records and diagrams kept as code.

SKILLS

AI & ML	LLM products & evaluation (Claude, Gemini, Llama; LLM-as-judge on AWS Bedrock), fine-tuning transformers (PyTorch, Hugging Face - DeBERTa, GECToR), NLP for assessment (CEFR scoring, grammatical error correction), speech (Whisper ASR, speaker verification, TTS), in-browser ML (Transformers.js, ONNX, WebGPU), serverless GPU inference (Modal)
Languages	TypeScript / JavaScript, Rust, Python, Clojure, Java, SQL
Web & cloud	Next.js, React, Node, Cloudflare (Workers, Durable Objects, D1, R2), AWS (ECS Fargate, SQS, DynamoDB, Lambda, RDS, S3, Cognito, Amplify), PWAs, WebSockets / SSE, WASM
Platform & ops	Terraform, Docker, GitHub Actions CI/CD, CloudWatch, Sentry, PagerDuty, testing (Vitest, Playwright, clippy-gated Rust)
Practice	System design & architecture, team leadership & mentoring, documentation & diagrams as code (Graphviz, Mermaid, PlantUML, Miro, ADRs, RFCs, runbooks), AI-assisted development workflows

EXPERIENCE

Founder / Software Engineer - Total Reality Engineering — 1998 - present

Independent software studio and consultancy - full-time focus since May 2026.

- Design, build and run a portfolio of live AI products for language learning and assessment (see Selected Projects), deployed on Cloudflare's edge with serverless GPU inference on Modal.
- Fine-tune and deploy transformer models: CEFR essay scoring (DeBERTa-v3 - 99.6% adjacent accuracy), grammatical error correction (GECToR, seq2seq, LanguageTool) and on-device speech recognition (Whisper) - served as independent GPU microservices on Modal.
- Documentation and testing as a habit: 20+ repos share a house style - ARCHITECTURE and SPEC docs, decision records, and Graphviz/Mermaid diagrams rendered and validated in CI so they can never go stale - with Playwright/Vitest suites throughout.
- Author of five production AI/ML engineering training courses, covering serverless edge architecture, fine-tuning transformers, browser-based ML, Python DevOps and ML deployment.

API Team Lead - Cambridge University Press & Assessment — Apr 2025 - May 2026

- Led the Assessment & Research Capabilities (ARC) API team - building and operating the Text and Speech APIs that power automated marking across Cambridge assessments, including Write & Improve, Speak & Improve and Linguaskill.
- A hands-on lead: top committer on the Text API and its orchestration service (Clojure event-driven microservices on AWS - ECS Fargate, SQS, DynamoDB, S3) and a lead contributor to the five-environment Terraform estate.
- Ran the platform to 99.98% uptime over 12 months, with CloudWatch monitoring, PagerDuty incident response and DORA delivery metrics (~1% change failure rate).
- Took AI models from research hand-off to production: transformer and XGBoost automarkers, off-topic and anomaly detectors, and Whisper-based speech recognition with speaker verification for malpractice detection; evaluated marking quality with Claude on AWS Bedrock.
- Built the team's documentation culture - architecture, RFCs, runbooks, incident response, onboarding and a public API documentation site - and Claude-driven tooling for Sentry error triage and backlog hygiene.

Senior Software Engineer - English Language iTutoring (ELiT) Feb 2023 - Apr 2025

- Senior engineer on Write & Improve and Speak & Improve - AI feedback tools that help English learners worldwide improve their writing and speaking, built in Clojure around custom marking models.
- The company and its products were merged into Cambridge University Press & Assessment in 2025.

Staff Engineer - OneStudyTeam Aug 2022 - Dec 2022

- Tech lead on a system enabling clinical research sites to communicate with trial patients via SMS.

Technical Lead - Predira May 2020 - Aug 2022

- Envisioned a customisable trading platform for brokerages with the founder (sister company to Limpid Markets), built the early prototypes, then built up the team and engineering practices.

CTO - Limpid Markets Sep 2018 - Apr 2020

- Set the technical direction for a precious-metals derivatives trading platform and transitioned the development team from London to Ukraine.

Technical Lead - SimPlay Jan 2019 - Apr 2021

- Rebuilt a platform for running collaborative role-play training experiences online, alongside other engagements.

Senior Software Engineer - AKQA, London Nov 2007 - Aug 2018

- Ten years building interactive products and campaigns for Nike, Heineken, Nissan, Fiat, the BBC, Grey Goose and Ford - including Nike Pro Genius, Heineken Star Player and the BBC's Story of Life.

Earlier career 1993 - 2007

- **Software Developer, LTA (2005 - 2007):** TotalTennis club-website platform, live scoring used at the Nottingham and Eastbourne tournaments, Siebel CRM integration.
- **Software Developer, Oxford University Learning Technology Group (2005 - 2006):** Java learning-tools integration and distributed-cognition middleware.
- **Interface Developer, Play Sport New Media (2001 - 2003)** and **Lead Front-end Developer, Pretzel Logic (1999 - 2001):** club-website systems for UK cricket, tennis and football clubs; sites for WA government agencies and HBF health insurance.
- **Electronics Technician → R&D Engineer, Ampac Australia (1993 - 1998):** fire-detection systems - built the Melbourne Airport graphics UI and a Java fire-panel simulator for training.

SELECTED PROJECTS (ALL LIVE)

- **Comprehendo** (comprehendo.net) - adaptive reading coach in 16 languages: AI-generated passages at CEFR level, tap-to-translate, TTS. Next.js on Cloudflare Workers + D1, Gemini. In live testing with an early user cohort.
- **Talata** (talata.app) - AI essay scoring and feedback: a fine-tuned DeBERTa-v3 multi-trait scorer and dual grammar-correction models (GECToR + seq2seq), served as five GPU microservices on Modal behind a Cloudflare edge app.
- **Speako** (speako.tre.systems) - pronunciation coach running entirely in-browser: local Whisper plus a fine-tuned DeBERTa CEFR model via Transformers.js and WebGPU. No server - nothing leaves the device.
- **Galacto** (galacto.org) - black-hole accretion simulation: 130k+ particles on GPU compute shaders, with sound synthesised from the galaxy's motion. Rust, WASM, WebGPU.
- **Rowspire** (rowspire.com) - strategy game with two in-browser AI opponents: a bitboard minimax solver and an AlphaZero-style neural-network MCTS, in Rust/WASM.
- **Delta-V** (delta-v.tre.systems) - turn-based multiplayer space tactics: an event-sourced authoritative server on Cloudflare Durable Objects with WebSocket hibernation and checkpoint recovery.

Full portfolio: tre.systems

EDUCATION

BE Electronic Engineering - The University of Western Australia