

Rayhan Khilji

Founder, AI Engineer

London, United Kingdom | [LinkedIn](#) | [GitHub](#) | [Website](#)

AI engineer and technical founder building AI-native products from zero to production. Currently scaling Lumi, an AI tutor for UK GCSE and A-level students.

EXPERIENCE

Founder Aug 2026 — Present
TryLumi Greater London

Built and launched an AI tutor for UK GCSE and A-level students featuring an interactive whiteboard, exam board adaptation, and concept re-explanation. Grew to 249 users with 40+ hours of learning time in three weeks, saving families under £2,500 versus private tutoring. Own product, marketing, and user acquisition; running user calls and setting 48-hour build priorities with co-founder. Opening schools distribution channel with paid pilot in discussion.

Forward Deployed Engineer (Contract) Jul 2026 — Sep 2026
algosoup London Area, United Kingdom

Worked directly with 3 clients across manufacturing, quant finance, and law to identify inefficiencies and build custom AI and agentic solutions. Owned full lifecycle from discovery and prototyping through deployment, reducing client cost and time across multiple projects.

AI Engineer Intern Jul 2026 — Aug 2026
MAGIC AI London, UK

Built AI-powered session analysis and worked on computer vision edge cases for smart mirror applications. Designed personalized workout generation driven by user data. Pitched CEO a new B2B sales funnel for developers building apartment complexes.

Social Media Marketing Intern Jun 2026 — Present
Bleep64 Greater London

Rebuilt company website and shipped AI workflows automating UGC generation and video editing. Planned content calendar across LinkedIn, Instagram, and TikTok.

Intern Apr 2026 — May 2026
QinetiQ Target Systems - United Kingdom London Area, United Kingdom

Co-Founder Mar 2026 — May 2026
Optic London, UK

Architected and shipped an AI-driven user reputation engine from zero to launch in four weeks. Built full-stack asynchronous data ingestion pipeline scraping and parsing large-scale digital footprints. Designed Supabase/Postgres schema for risk categorization and behavioral metrics, and built RAG chatbot vectorizing user posts for real-time profile analysis.

Future Leader Sep 2025 — Aug 2026
Future Leaders UK London Area, United Kingdom

Weekly online calls with experts from different fields
Visited parliament to speak to Wes Streeting
launch workshop at the Business Design Centre in London

Founder + FDE Apr 2025 — Jun 2026
Toasted Beans Creative Studio London, UK

Founded and ran a studio delivering 50+ websites and software products, owning engineering, sales, QA, and client delivery. Recruited, trained, and managed a 12-person sales and engineering team, building sales process from scratch. Grew revenue multiple month-over-month to just under \$14,000 over the year.

Prop Trader (funded evaluation account) Nov 2024 — Nov 2025
Topstep Remote

Traded funded evaluation account with \$750K in buying power across passed evaluations, taking roughly \$3,000 in payouts with 45% win rate and ~3.5 profit factor under strict daily stop-loss and drawdown limits. Built and iterated trading automation on real-time market data, technical indicators, and macroeconomic events.

EDUCATION

Bae HQ Jun 2026 — Sep 2026
Startups

2nd place in National Finals!

3x member

Built Voluble (<https://www.voluble.works>)

Built Lumi (<https://www.trylumi.co>)
Building sidequest (<https://www.sidequest.cyou>)

Ilford County High School

Sep 2025 — Jun 2027

A Levels in Economics, Politics, Mathematics

AS 2026: ADU. GCSEs (2025): AAAAAAABCC across Maths, English Lit/Lang, Geography, French, Biology, Chemistry, Physics, Design & Technology, Philosophy, Psychology.

Ilford County High School

Sep 2020 — Jun 2025

GCSE's

Activities and societies: Language Captain Y10, Y11

Represented by Standout

PROJECTS

TensorVM

TensorVM is a small neural network runtime built from scratch in Rust to understand and implement the systems underneath modern ML frameworks.

The runtime represents models as a tensor intermediate representation (IR), with the same computation graph used for forward execution, reverse-mode automatic differentiation, model serialization, and backend execution. I implemented the core tensor system, computation graph, autodiff engine, backend abstraction, portable model format, and Python API. The core runtime is written in Rust and exposed to Python through PyO3, with a PyTorch-style frontend for defining models and optimizers. TensorVM currently supports CPU execution through a readable reference backend and an optimized backend, where blocked matrix multiplication delivers approximately 4.4x to 6.3x higher throughput on benchmarked matrix sizes.

I also built end-to-end MNIST training and serialization, allowing a trained model to be saved as a portable `.tvm` file, loaded without the original model code, and executed on a different backend while producing identical predictions.

The project is deliberately small and legible rather than production-scale. The goal was to understand how tensor runtimes, computational graphs, automatic differentiation, execution backends, and model serialization work underneath abstractions such as PyTorch and TVM.

Tech: Rust, Python, PyO3, NumPy, ndarray, automatic differentiation, tensor IR, compiler/runtime design, numerical computing, machine learning systems.

Open source : <https://www.github.com/rayhankhilji/tensorvm>

Parkway

An online version of monopoly where you can play against bots of different skill levels. Built custom frameworks to solve edge cases like trading properties, mortgage and remortgaging, jail and free parking.

Engineered to be fast paced, engaging and responsive, so instead of 3 hour long games, each game takes around 25 minutes and is just as fun.

I am still working on the multiplayer mode.

The project is open source.

You can view the code here and contribute:

<https://github.com/rayhankhilji/parkway>

Exposure

Personalised horror game built with realtime ai world model technology, generating the game frame by frame based on refined inputs from user interview powered by openai, reactors technology and veed's interactive models.

Voluble

Built a voice-first language learning platform designed around real conversations rather than flashcards or grammar exercises.

Users practice speaking in realistic scenarios including job interviews, client meetings, ordering food, networking events and everyday conversations. The AI adapts the difficulty, vocabulary and speaking speed in real time based on the user's fluency, confidence and mistakes, while remembering previous sessions to build long-term progress.

Implemented a low-latency conversational pipeline using streaming speech recognition, LLM reasoning and neural text-to-speech to create natural back-and-forth dialogue with minimal delay. Conversation history, vocabulary progress and grammar weaknesses are persisted in PostgreSQL, allowing lessons to build on previous sessions instead of starting from scratch.

Built using TypeScript, React, Node.js, PostgreSQL and Redis with Groq Whisper for transcription, Gemini for conversational reasoning and ElevenLabs for speech synthesis. Additional features include pronunciation analysis, vocabulary revision, contextual grammar feedback, role-specific learning paths and simulated real-world conversations. Focused on helping users become conversationally fluent by simulating situations they are likely to encounter in real life, replacing repetitive language exercises with interactive spoken practice.

Download for free here: <https://voluble.works>

jellybean

jellybean is an MCP server that gives coding agents a structured understanding of large codebases without dumping entire files into context.

It indexes repositories into symbols, imports, files and relationships, then exposes compact tools for repository mapping, symbol outlines, ranked search, dependency tracing, code reading and diagnostics. Results are handle-based and strictly constrained by token budgets, so agents can progressively drill into the exact code they need instead of wasting context on irrelevant source.

Built the indexing and retrieval pipeline in TypeScript, including incremental filesystem watching, parsing, symbol extraction, import graphs, content-derived handles, caching and token-aware result generation.

Benchmarked against repositories including Express, NestJS and VS Code, with Jelly Bean reducing token usage by 97–99.9% on several codebase navigation tasks while keeping tool latency in the millisecond range.

Sotto

Built a macOS desktop application that lets users control AI agents entirely through voice. Instead of answering questions, Sotto executes tasks by launching specialised agents that continue working in the background while the user moves on to something else.

Designed an orchestration layer that routes requests to the appropriate agent, tracks execution state and returns results when tasks are complete. Integrated speech recognition, streaming LLMs and macOS APIs to enable hands-free control of desktop workflows.

Built with Electron, TypeScript and Node.js using Groq Whisper for speech recognition, GPT for reasoning and ElevenLabs for voice synthesis. Supports multi-step task execution, context-aware conversations and integrations with desktop applications and external APIs.

Plaine

Plaine - AI CAD for AE

Forked and extended an open-source CAD software, embedding an AI design agent directly into the modelling workflow instead of treating it as a separate chatbot. Reworked the interaction model so engineers could generate, edit and iterate on conceptual aircraft designs using natural language while retaining full control of the underlying CAD environment. Integrated LLM-driven design generation, parametric editing, conversational refinement and contextual design suggestions directly within the modelling interface. Built support for rapid concept generation, feature editing, version history and engineering specification parsing, enabling aerospace teams to move from requirements to editable CAD concepts in minutes rather than hours.

Demo'd with real aerospace engineers to validate workflows, gather user feedback and refine the interface around early-stage aircraft conceptual design rather than general-purpose mechanical CAD.

Takir

Built a conversational AI tutor that teaches through in realtime an interactive sandbox/canvas.

Students speak naturally while the system responds in real time, drawing diagrams, plotting graphs, annotating equations and generating images on a shared canvas throughout the lesson.

Designed a low-latency realtime architecture that streams voice, reasoning and canvas ops synced perfectly instead of sequentially, allowing lessons to evolve naturally in realtime without large delays. Conversation state and whiteboard state are persisted independently, enabling the tutor to reference anything previously discussed, modify existing diagrams and continue lessons across sessions.

CERTIFICATIONS

Certificate of Completion: AI Fluency for Builders , Anthropic	Jul 2026
Deep Dive: Building Your First Pipeline , Palantir Technologies	Jul 2026
Deep Dive: Data Governance Tools in Foundry , Palantir Technologies	Jul 2026
Introduction to Foundry & AIP for Enterprise Organizations , Palantir Technologies	Jul 2026
Speedrun: Your First AIP Workflow , Palantir Technologies	Jul 2026
Citi - Technology Software Development Job Simulation , Forage	Apr 2026
Commonwealth Bank - Software Engineering Job Simulation , Forage	Apr 2026
Electronic Arts - Software Engineering Job Simulation , Forage	Apr 2026
Hewlett Packard Enterprise - Software Engineering Job Simulation , Forage	Apr 2026
Skyscanner - Software Engineering Job Simulation , Forage	Apr 2026
Walmart USA - Advanced Software Engineering Job Simulation , Forage	Apr 2026
Wells Fargo - Software Engineering Job Simulation , Forage	Apr 2026

SKILLS

Automatic Differentiation · Compiler Construction · Rust (Programming Language) · Abstract Syntax Trees (AST) · Software Architecture · Node.js · Software Project Management · Sales Operations · Leadership · Future Trends · Startups · Sales · Marketing · Swift (Programming Language) · Agentic AI Development · Inter-process Communication · CAD Systems · Three.js · Computational Geometry · LLM Integration · PostgreSQL · LLM Orchestration · Prompt Engineering · Real-Time Operating Systems (RTOS) · Applied AI Engineering · Cramming · Software Development · System Architecture · Retrieval-Augmented Generation (RAG) · Co-creation · Deterministic LLM Programming · Team Management · Cold Calling · Feature-Driven Development (FDD) · Product Management · Object-Oriented Programming (OOP) · Object Oriented Design · Code Review · Game Development · Spring Framework · Data Modeling · TypeScript · Unit Testing · NoSQL · HTTP Client Libraries · State Management · API Testing · Version Control · Python · JavaScript · SQL · HTML · CSS · React · Next.js · Electron · Tailwind CSS · LLM APIs · Agentic systems · Tool calling · Vector search · LangGraph · Evaluation · Supabase · Firebase · Datadog · WebSockets · REST APIs · Git