António Cruz

Senior engineer and technology leader. A career across telecom, SaaS, biometrics, identity and passenger information systems. Hands-on with AI and Machine Learning systems, from architecture to deployment.

Lisbon, Portugal logus2k@gmail.com logus2k.com GitHub LinkedIn

SUMMARY

Senior engineer and technology leader building production systems across telecom, SaaS, biometrics, identity and passenger information systems. Currently hands-on in AI and machine learning: ISCTE Postgrad in Applied Machine Learning, plus a personal portfolio of MLOps, agentic and deep-learning systems, all running on a self-hosted, fully on-premises stack of open-source models at logus2k.com. Building, not just managing.

CORE SKILLS

Concepts	Machine Learning, Deep Learning, GenAI, RAG (Vector + Graph), LLMs, Transformers, GANs, Reinforcement Learning, MLOps, Computer Vision, NLP, Time Series, Model Context Protocol (MCP), tool and function calling, agent skills (trigger-based auto-injection), OpenAI-compatible APIs.
Stack	Python, PyTorch, TensorFlow, Keras, Hugging Face Transformers, llama.cpp, FastAPI, Socket.IO, Docker, Docker Compose, CUDA, MLflow, Airflow, DVC, Hydra, ChromaDB, ArcadeDB, Whisper, Kokoro, Anthropic Claude.
On-Premises AI Stack	Gemma 4 E4B via llama.cpp, Whisper Large v3 Turbo, Kokoro TTS, bge-m3 embedder, bge-reranker-v2-m3 cross-encoder, ArcadeDB graph database, ChromaDB vector database. No cloud-AI API dependencies.
Infrastructure	Linux, WSL2, Docker and Docker Compose, nginx reverse proxy with TLS, Let's Encrypt, dynamic DNS (NoIP), systemd auto-start, ~15+ container production stack at logus2k.com.
Leadership & Delivery	Engineering leadership, product management, software architecture, technical strategy, customer-facing roles, partnership and standards work.

SELECTED PROJECTS

Live demos, source code and detailed reports for everything below are linked from logus2k.com.

noted: collaborative MLOps platform with on-premises GenAI

Live

Self-hosted MLOps platform integrating MLflow, Airflow, DVC, Hydra, MinIO, ChromaDB and ArcadeDB into a single collaborative web interface. Real-time multi-user notebook editing, multi-runtime Python 3.10–3.14 (including free-threaded variants), kernel management, virtual environments, model registry and serving.

Native LLM tool-calling via 24+ MCP-style tool schemas (read and write tiers with approval flow for write actions); trigger-based skill registry auto-injects curated instructions into the prompt; per-Domain capability scoping (each Knowledge Domain pins its own tools and skills). GraphRAG (entity extraction, community detection, sameAs identity merging, denormalized chunk grounding) and Vector RAG (bge-m3 + cross-encoder reranking) over per-Domain knowledge bases.

Inference runs locally on Gemma 4 E4B via llama.cpp; the LLM router also supports Anthropic Claude as an optional backend.

FastAPI · Socket.IO · Python 3.10–3.14 · CUDA · Docker Compose · MLflow · Airflow · DVC · Hydra · ChromaDB · ArcadeDB · llama.cpp · bge-m3.

Demo video

Agent stack: privacy-first local AI orchestration

Three coordinated services that form a fully on-premises AI agent platform:

- **agent_server:** LLM control plane with an OpenAI-compatible `/v1/chat/completions` streaming facade. Local models behave as drop-in replacements for cloud APIs, hot-swappable per request via named presets (chat, judge, graph-answer, etc.).
- **STT:** Whisper Large v3 Turbo + Silero Voice Activity Detection over Socket.IO; real-time streaming transcription with multi-client support.
- **TTS:** Kokoro multi-language (English, Japanese, Mandarin, Spanish, French, Hindi, Italian, Portuguese) with WebSocket audio streaming, advanced sentence splitting via spaCy.

All inference on-prem. No external dependencies.

Python · FastAPI · Socket.IO · Hugging Face Transformers · llama.cpp · Whisper · Kokoro.

Autonomous warehouse robotics: voice + vision + LLM, end to end

End-to-end voice-controlled YouBot in Webots. The pipeline is: voice input, STT, quantized LLM (SmolLM2-135M via llama.cpp), structured JSON command, robot action, ResNet-18 box-damage classification, and TTS confirmation. Three-mode state machine (standby, patrol, goto) with aisle-aware path planning, dual-camera inspection, real-time inventory updates broadcast to connected clients.

Four independent Docker services communicating over Socket.IO. Entirely on-prem, no cloud calls in the loop.

PyTorch · ResNet-18 · Webots · llama.cpp · FastAPI · Socket.IO.

Demo: voice control Demo: autonomous patrol

scipredictor: multi-label classification across 148 research domains

Live

Multi-label classifier for academic papers across 148 research domains. Independent per-class probability output (no softmax constraint) so an interdisciplinary paper can simultaneously belong to multiple fields, mirroring real research practice instead of forcing a single label. Companion case study using SciBERT for arXiv primary-subject prediction with hierarchical-domain analysis and confidence-based triage of ambiguous cases for human review.

PyTorch · Hugging Face Transformers · SciBERT.

GAN comparative study + GAN-vs-Human game

Live

Conditional MNIST generators benchmarked across four loss strategies (BCE, LSGAN, Hinge, WGAN-GP) with identical architectures and FID and KID evaluation. The best-performing model powers a real-time browser game scoring human-drawn vs generated digits via a trained MNIST classifier. Three difficulty levels with time pressure (2.0 s, 1.0 s, 0.5 s).

PyTorch · WGAN-GP · FID and KID · WebSockets.

Multi-step temperature forecasting: GRU vs Transformer

Multivariate-input, univariate-output forecasting study on the Jena Climate dataset (10-minute resolution, 2009–2016). Predicts 24-hour air temperature using GRU and Transformer architectures under controlled conditions: temporal resampling, multi-step windowing, multi-horizon evaluation, hyperparameter exploration. Six input features (temperature, pressure, humidity, wind speed, max wind speed, wind direction), each paired with timestamps.

PyTorch · Keras · time-series preprocessing.

Report

DQN vs PPO on LunarLander-v3: value-based vs policy-based RL

Comparative deep-RL study contrasting Deep Q-Network and Proximal Policy Optimization in the LunarLander-v3 environment. Identical environment settings and multiple random seeds; analysis combines learning curves, sample efficiency, cross-seed stability and qualitative landing-behavior evaluation from recorded simulations. Covers DQN's overestimation bias, ε-greedy exploration, replay buffer dynamics; PPO's clipping, GAE, and entropy-driven exploration.

PyTorch · Gymnasium · DQN · PPO · GAE.

Report + annexes

FEM execution-strategy benchmark: CPU vs GPU vs JIT

Live

Same Finite Element Method problem implemented across multiple execution backends (sequential CPU, multi-process and shared-memory CPU parallelism, Numba JIT, Numba CUDA, CuPy with custom raw kernels) under identical numerical formulation, discretisation, boundary conditions and solver. Numerical equivalence preserved across all paths, enabling apples-to-apples comparison of execution behaviour, performance and scalability. Containerised with platform-specific helpers and automatic CPU and GPU fallback.

Python · NumPy · Numba · CUDA · CuPy · Docker.

---

doco + docbro: document tooling

doco · docbro

- **doco:** Self-hosted notebook to styled-Word converter, 5-step wizard, real-time WebSocket progress. Configurable fonts, layout, tables, code styling, image handling, optional HTML and Markdown exports.
- **docbro:** Static client-side documentation browser. Split-pane tree navigation, tabbed switching, KaTeX math, Mermaid diagrams, syntax highlighting, image lightbox, deep linking, scroll-synced TOC.

Python · FastAPI · WebSockets · pandoc · KaTeX · Mermaid.

Chest X-Ray pneumonia classifier: binary medical-image CNN

Live

Convolutional neural network for binary classification of chest X-ray images (pneumonia vs. normal). Full training pipeline with data augmentation, validation, and evaluation by accuracy, precision, recall and F1.

Python · PyTorch · CNN · medical imaging.

OpenCV Lab: interactive image-processing sandbox

Live

Web-based experimentation environment for classical computer-vision techniques: grayscale and colour-space conversions, resizing, rotation, Gaussian and median blurring, Canny and Sobel edge detection, binary and Otsu thresholding. High-performance aiohttp backend with non-blocking I/O; Socket.IO for real-time browser interaction.

Python · OpenCV · aiohttp · Socket.IO · WebSockets.

Home-lab production stack: the platform behind the portfolio

Live

Personal infrastructure hosting the entire portfolio. ~15+ Docker containers across noted's MLOps stack, the agent stack (STT, TTS, orchestrator), the application gallery and the web frontend. Linux on WSL2, nginx TLS facade with Let's Encrypt, dynamic-DNS public address. Services auto-start at host boot, the site is available even when the developer is signed out.

Fully on-premises AI stack (Whisper, Kokoro, Gemma, bge-m3, bge-reranker, ArcadeDB, ChromaDB), no cloud-AI API dependency.

WSL2 · Docker Compose · nginx · Let's Encrypt · NoIP · systemd.

PROFESSIONAL EXPERIENCE

Hitachi Rail · Portugal

2023 – 2025

Head of Product Development

- Head of Discipline Passenger Mobility at Hitachi Rail GTS PT.
- Owned product development of APIS 8, APIS 9 and Audio Dispatcher, Hitachi's flagship products for Advanced Passenger Information Systems (real-time multi-channel display and audio across rail networks).
- Hands-on development of processes and dashboards for projects, capacity planning, finance-integrated reporting and quality performance metrics.
- Successfully recovered multiple legacy projects while driving the previous product generation to a positive end-of-life outcome.

TECH5 · Portugal

2022 – 2023

VP Engineering

- Led Engineering and Product teams building biometric identity products (face, fingerprint, iris).
- Owned product roadmaps, development and delivery.
- Implemented SDLC, roadmap practice and knowledge management from the ground up.
- Hiring, management, performance assessment.
- Hands-on guidance on product management processes and practices.

Vision-Box · Portugal

2016 – 2020

Chief Digital Officer · 2019 – 2020

- Owned technology strategy, innovation and adoption; advised the C-suite and Board on industry direction.
- Managed architecture and development (external and internal teams) of Vision-Box's Biometric Check-In CUSS Kiosk, a computer-vision-based biometric pipeline running on embedded hardware.
- Represented Vision-Box as a Strategic Partner of IATA; heavily contributed to several IATA Passenger Experience standards, particularly IATA One ID: Technology, Privacy and Identity Management.

Chief Technology Officer · 2016 – 2019

- Led Software and Hardware Engineering teams.
- Market-leading hardware and software product lines: eGates, kiosks, portable devices, backend platform and customer-facing apps; computer-vision-based face biometric recognition deployed in airports and border posts worldwide.
- Architected and led ground-up development of **Vision-Box Orchestra Platform**: IDaaS biometric-recognition platform that became an industry reference for Seamless Travel in Aviation and Border Control.
- Decisively contributed to closing new multi-million-euro contracts.
- Represented the company in, or supervised projects across, over 20 countries.
- Subject-matter expert at IATA One ID Technical and Privacy groups.

Research & Development Director · 2016

MEO (Altice) · Portugal 2004 – 2016

SBD Team Lead

- Led the API Management team (for PT and SAPO) and the GIS and Maps Software teams.
- Architected and built **Service Delivery Broker**: Portugal Telecom's API management platform delivering 40M requests per day to 60+ web, mobile and IPTV apps. Finalist of two Business & Innovation Awards.
- Designed systems for high availability, scalability, security, modularity and consistency.
- Expanded Portugal Telecom platform to Oi (Rio de Janeiro and São Paulo, Brazil).
- TM Forum Software Enabled Services Team Co-Chair. **Outstanding Contributor Award** for industry-adoption.

Former Professional Experience

Senior Software Engineer
ParaRede
04.2004 – 09.2004

Senior Software Engineer
Sybase
02.2004 – 04.2004

Senior Software Engineer
Espírito Santo Informática
08.2003 – 02.2004

Senior Software Engineer
Millennium BCP
06.2003 – 07.2003

Software Engineer, Team Leader
Methodus
08.2001 – 05.2003

Senior Software Engineer
Link Consulting
01.2001 – 07.2001

Logistics Manager
Hikma Pharmaceuticals
02.2000 – 12.2000

Software Engineer, Captain
Portuguese Army, Division of Justice and Discipline
11.1996 – 02.2000

Software Engineer, Lieutenant
Portuguese Army, General Staff
09.1993 – 10.1996

Software Engineer, Lieutenant
Portuguese Army, Centre of Computer Science
09.1993 – 10.1996

Software Engineer

Portugal Telecom, SAPO
07.1989 – 07.1992

EDUCATION

ISCTE University Institute of Lisbon Oct 2025 – Jul 2026

Postgraduate in Applied Machine Learning

Coursework spanning Deep Learning, Generative AI, Reinforcement Learning, Time Series, Computer Vision, NLP and MLOps. Concrete deliverables include the multi-step temperature forecasting study (GRU vs Transformer on Jena Climate), DQN and PPO comparison on LunarLander-v3, SciBERT-based arXiv subject classification, and the conditional-GAN studies, all detailed at logus2k.com.

Coimbra and Lisbon Universities

Law Degree Frequency

LANGUAGES

Portuguese: Native **English:** Fluent

CERTIFICATIONS

AI, Machine Learning, Data

- Azure AI Fundamentals AI-901 (Microsoft) · Verify
- Building Agentic AI Applications with a Problem-First Approach (Maven)
- Deep Learning Specialization (DeepLearning.AI)
- Machine Learning (Stanford University)
- Data Analyst Professional (Google)
- Essential Math for Machine Learning, Python Edition (LinkedIn Learning)
- Systems Thinking (LinkedIn Learning)

Architecture, Cloud, Engineering

- TOGAF 9.1, Level 2 (The Open Group)
- SOA Professional (SOA Systems)
- Software & Cloud Architecture (LinkedIn Learning)
- Azure Fundamentals (Microsoft) · Verify
- Software Development Professional (Forino School of New Technologies)
- Microsoft Certified Application Developer (Microsoft)

Project Management and Agile

- Project Management Professional (Google)
- Professional Scrum Master I, Ken Schwaber (Scrum.org) · Verify
- Professional Scrum Developer I (Scrum.org) · Verify
- Scrum Master and Product Owner (Scrum Alliance)

SELECTED ACHIEVEMENTS

> **Vision-Box Orchestra Platform:** Led ground-up development of Vision-Box's IDaaS biometric-recognition platform; became an industry reference for Seamless Travel in Aviation and Border Control.

> **Service Delivery Broker (Portugal Telecom, SAPO):** Architected and built the API management platform; finalist of two Business & Innovation Awards; delivered 40M requests per day to 60+ web, mobile and IPTV applications.

> **TM Forum Outstanding Contributor Award:** *"For outstanding performance and lasting contribution on the Software Enabled Services work. António has been a driving force behind several catalysts in this area, lead in industry adoption and spoke on behalf of the TM Forum several times."*

"Learning never exhausts the mind." (Leonardo da Vinci)