Diogo Antunes

Research Scientist | Applied AI Scientist | Research Engineer

Agent Runtimes | World Models | Evaluation & ML Infrastructure

Zurich Metropolitan Area | Open to U.S. relocation | U.S. work authorization required

contact@diogoantunes.ai | +351 939 911 077 | diogoantunes.ai

LinkedIn | GitHub | Hugging Face

Research engineer building governed AI systems across transformer training and evaluation, multimodal and world-model research, uncertainty and calibration, secure agent runtimes, and distributed deployment. Founded Vaisys and secured pre-seed funding; delivered **50B+ token** data pipelines, a **1.044B-parameter** decoder, a thesis-reported **93.1% held-out four-label SAST score**, and a **290-commit Rust agent kernel**. Strongest in under-specified problems requiring rigorous evaluation, explicit state, and hands-on systems delivery, following a scientific path from falsifiable hypotheses through controlled experiments to reproducible evidence.

Professional Experience

Vaisys | Applied AI Scientist; Founder & CEO — **Jan 2023 – Present**

Applied AI Scientist: Jan 2023–Present | Founder & CEO: Mar 2025–Present

- Founded the company and secured **pre-seed funding**; lead technical strategy, product architecture, research direction, and investor-facing execution for private, governed enterprise agents.
- Building **Harmony**, a model-agnostic enterprise AI foundation: knowledge-graph/vector retrieval, role-specific agents, MCP service discovery and tool use, RBAC/SSO, human approval gates, audit trails, and on-premises/private-cloud deployment with replaceable model providers.
- Curated and validated corpora spanning **50B+ tokens** for specialized transformer pretraining/domain adaptation; designed provenance, deduplication, quality/coverage controls, and evaluation across **memory, retrieval, tool execution, and escalation**.
- Co-authored a **31-page Large Action Model feasibility study** (May 2025): a code-interpreter-coupled model with persistent external memory, iterative execution/self-correction, phase-dependent decoding, sandboxed tools, and verifiable reward/safety loops; the architecture informed later agent-runtime work.

DefensePoint | AI Tech Lead Consultant; ML Engineer / Data Scientist — **Nov 2024 – Oct 2025**

ML Engineer / Data Scientist: Nov 2024–May 2025 | AI Tech Lead Consultant: May–Oct 2025 | Braga / Lisbon, Portugal

- Core engineer of **Adaptive MFA** (open source since Q4 2026, Apache-2.0): an on-premise risk engine for Keycloak that scores every login against **19 configurable signals** (impossible travel, VPN/proxy/Tor, device-fingerprint novelty, behavioural and time anomalies) with local ML inference (K-Modes, DBSCAN, time profiling) and returns a risk level (1–4) driving step-up MFA.
- Built the **FastAPI decision service** (PostgreSQL, Redis), the **Keycloak SPI integration**, and per-realm policy across a Python/Java/TypeScript stack; deterministic, auditor-traceable scoring with no external model, GPU, or outbound request, so it runs air-gapped.
- Outside this project, led client AI consulting independently end to end: technical discovery, prototyping, evaluation criteria, and delivery through to handoff.

Checkmarx | Software & Machine Learning Engineer — **Mar 2024 – Oct 2024**

Braga, Portugal

- Independently designed and implemented an end-to-end SAST query-selection system: local-LLM annotation, custom hierarchical tokenizer, masked-language-model pretraining, Longformer-style multi-label fine-tuning, REST inference, web UI, scripts, and automated tests.
- Processed **60,000 open-source projects** (**2.6 TiB** source corpus; 600 MiB structural representations) and ran **228 hours** of masked-language-model pretraining plus 12 hours of downstream fine-tuning on 24 GB VRAM.
- Reported **91.9% validation** and **93.1% held-out accuracy**; evaluated on separate 600-example sets for four-label classification, documented the complete pipeline in a 115-page technical thesis.

Research, Open Source & Venture Systems

Splendor Kernel | Founder & Principal Engineer — Docs | GitHub

Jan 2026 – Present | Rust, Python/PyO3, TypeScript, SQLite, OpenAPI, Docker

- Created an OS-like substrate for persistent, governable agent loops rather than another prompt framework: **290 public commits, 5 implementation milestones** (0.01–0.05), a stable 0.1 primitive compatibility contract, **28 runnable reference examples**, and a dedicated public documentation site.
- Implemented scheduler/resumable loops, SQLite state graphs and snapshots, integrity-chained traces/replay, a verifier-chain action gateway, typed inbox/outbox messaging, scoped delegation, per-agent isolation, daemon/CLI/OpenAPI, Python and TypeScript clients, CI, and multi-architecture Docker artifacts.
- Extended the runtime into resident/fleet and governance primitives (distributed identity, signed work orders, placement, state handoff, telemetry, approvals, escalation, circuit breakers, policy TTL/revocation) and physical-edge contracts (device profiles, offline policy caches, reconnect trace sync, bounded robotics adapters, safety verifiers, and simulation harnesses).
- Systems thesis: robotics and autonomous agents already need enforced identity, explicit state, permissioned side effects, failure containment, and replayable evidence; Splendor separates fast-moving model/Python space from Rust-enforced system invariants before actions reach tools or hardware.

Keplen / ValueAI | Founder & Research Lead SLWM | Natural Noise

Jan 2026 – Present | World models, multimodal learning, uncertainty, scientific evaluation

- Leading **Signal-Latent World Model (SLWM)** research: modality-specific text/code, audio, and visual/video codecs feed a shared latent field with tensor contract Z [B,T,D], a signal-world processor, prediction/uncertainty heads, and a policy/commit gate with valid no-op behavior.
- Built evidence-first experimental infrastructure: exact **124,439,808-parameter GPT-2-small baseline** and **125,489,152-parameter vanilla multimodal comparator**, T1 text/code and T2 audio/visual pipeline mechanics, registered hypotheses/baselines/ablations/failure cases, **74 passing tests**, and an explicitly non-claiming **700M+ fit-check profile**.
- Established research claim gates: match data, parameter, and compute budgets; report unsupported-claim and contradiction rates, usefulness/accuracy, abstention/no-op rate, and calibration; treat smoke/foundation tests as mechanics evidence only.
- Designed **Natural Noise**, a matched GPT-2 Large B0/N0 experiment isolating bounded, non-trainable Q/K/V/O head-channel perturbations. The primary N0 path uses **K=8 independent full-prefix Monte Carlo passes** without KV cache and reports entropy, mutual information, cross-pass disagreement, and JS divergence; the scientific hypothesis remains preregistered pending matched multi-seed runs.

Swivver | Founder & Systems Architect Product / technical deck

Jan 2026 – Present | Go, Temporal, PostgreSQL, Redis, MinIO, event-driven services

- Built a trust-aware media object: each canonical Story, Article, and Swivve package carries short video, written context, atomic claims, sources and evidence, factuality state, correction history, provenance, rights, and machine-readable structure across reader, writer, publisher, and AI surfaces.
- Built verification agents that evaluate atomic claims across multiple sources, attach evidence and conflict/uncertainty state, update labels such as Verified, Developing, Disputed, and Corrected, and preserve the correction timeline as the story evolves.
- Orchestrated the full lifecycle: ingest to claim extraction to retrieval/verification to factuality report to policy/human gate to versioned publish to continuous recheck, correction propagation, and notification. Claim/evidence outcomes and state transitions become trust signals for ranking, monitoring, and AI retrieval.
- Engineered separate Go API/worker/migration binaries with clean bounded contexts, PostgreSQL/Redis/MinIO, Temporal workflows, transactional outbox, durable idempotency, RBAC/object authorization, typed SQL, versioned OpenAPI, architecture tests, and citation-enforced grounded-AI contracts.

Selected Model & Systems Work

TA Model 1B / TA Model V2 | Model training & validation-first decision systems Model | Data

Public repositories: Feb 2026 (1B) | Jun 2026 (V2)

- Built and released a **1.044B-parameter LLaMA-family decoder** (20 layers/20 heads, hidden size 2,000, context 1,860) with a custom **1,024-token vocabulary** for structured OHLCVT, price-change fragments, trade/volume buckets, and technical-indicator windows.
- Implemented timestamp normalization, missing-interval repair, contiguous grouping, and **160-step fine-tuning windows with stride 10**; published source, model, and dataset. V2 extends the work into an asset-agnostic probabilistic decision/risk platform with lineage, simulation, walk-forward, paper-trading, and promotion gates rather than unsupported prediction claims.

Housify | Multimodal inference platform GitHub | Case study

Public repository activity: Oct 2023–Jan 2024 | archived

- Built distributed orchestration across LLaVA, Stable Diffusion, and Pix2Pix with Python/PyTorch and Kotlin services, JWT/rate limiting, Nginx, PostgreSQL, and Redis; adapted across vision-language, diffusion, mobile, API, and infrastructure layers in one system.

Selected Research & Technical Writing

Scan Optimization with AI July 19, 2024

115-page applied-ML thesis covering data engineering, tokenizer design, transformer pretraining, multi-label evaluation, secure inference services, testing, and deployment.

Large Action Model Viability Study May 2025

31-page co-authored architecture and feasibility study for persistent, code-executing agents with external memory, adaptive inference, tool control, RL with verifiable rewards, and sandboxed safety boundaries.

Open Research Documentation 2026–Present

Splendor runtime and specification docs (Jan 2026–Present); SLWM hypotheses, architecture, training/evaluation plans, and experiment registry (May 2026–Present); Natural Noise preregistration, ablation matrix, regression accounting, and failure-case reports (July 2026–Present).

Education & Technical Profile

B.S. in Software Engineering, Instituto Politécnico do Cávado e do Ave (IPCA) 2024 | Portugal

Coursework: Algorithms, linear algebra, statistics, calculus, computer vision, natural-language processing, and artificial intelligence

Research: Transformers, tokenizer design, pretraining/fine-tuning, world models, multimodal representations, uncertainty/calibration, controlled baselines, ablations, failure analysis, model/agent evaluation

Programming: Python, Rust, Go, C/C++, CUDA, TypeScript, Kotlin; PyTorch, Transformers, Hugging Face, scikit-learn

Systems: Agent runtimes, MCP/tool use, state/replay, distributed messaging, REST/OpenAPI, PostgreSQL/SQLite/Redis, Docker/Kubernetes, Nginx, CI/CD, observability, IAM/RBAC/SSO, audit and data governance

Communication: Public research and architecture writing for an **8,300+ follower** professional audience

Credentials: NVIDIA DLI: CUDA C/C++, Deep Learning, Accelerated Data Science, Satellite Imagery; Google Advanced Data Analytics; Checkmarx Security Champion

Languages: Portuguese and English — native or bilingual proficiency