Evans Eburu.

SOFTWARE ENGINEER · AI-SAFETY RESEARCHER

evanseburu.com · +234 906 134 9498 · evanseburu32@gmail.com · linkedin.com/in/0xjevan
github.com/Johnnyevans32

RESEARCH INTERESTS

AI safety and the evaluation of large language models — deception, detection, and the *monitorability* of machine reasoning; chain-of-thought faithfulness; safety and robustness of agentic multi-agent systems.

EDUCATION

MSc, Applied Information Technology — Ural Federal University (UrFU) 2025 – 2027 (expected)
Institute of Radio Electronics and Information Technology (RTF), Yekaterinburg

BEng, Electrical & Electronics Engineering — Federal University of Technology, Owerri (FUTO), Nigeria 2016 – 2022
Academic credentials verified by World Education Services (WES), 2023

PUBLICATIONS

[1] **MandateBench: Mandate Faithfulness and Pre-Signature Monitorability for Agentic-Payment LLMs. Eburu, E.** Preprint, Zenodo, 2026. doi:10.5281/zenodo.21411376

[2] **Among LLMs: A Cross-Play Benchmark for Deception, Detection, and the Monitorability of Reasoning. Eburu, E.** Preprint, Zenodo, 2026. doi:10.5281/zenodo.21209429

[3] **Discrete Event Simulation of a Municipal Meat Production System: Bottleneck Analysis and Optimization. Eburu, E. C.** & Balungu, D. M. *Sustainable Innovative Development: Design & Management*, 2026. RSCI-indexed (eLIBRARY.ru).

RESEARCH

MandateBench mandatebench.xyz 2026

An open benchmark of whether nine frontier LLMs, acting as payment agents, stay inside Ed25519-signed spending mandates under adversarial pressure, and whether a monitor reading the agent's reasoning can catch a violation *before* the irreversible signature.

- Judge-free ground truth: rule violations are deterministic checks against the signed mandate; intent traps are hand-labelled before any model runs. No LLM judge.
- Separates rule violations a payment gateway already blocks from intent violations only the model can catch; across 17 intent traps in 5 domains, models catch 77% pooled (best 90%, worst 57%) with near-zero false refusals.
- Pre-signature monitorability is only moderate (AUROC 0.62); telling agents to hide their reasoning raises violations by about a third and drops the monitor below chance (AUROC 0.38).
- 9×9 cross-play matrix in which every model attacks every other model's agent in a multi-turn negotiation; open-source harness (MIT) with a live dashboard.

DeceptionBench — "Among LLMs" deceptionbench.xyz 2026

A live, cross-play benchmark measuring how six frontier LLMs (one per major lab) deceive, detect deception, and whether their private reasoning can be monitored to catch them, played as a full impostor-versus-villager matrix in a text social-deduction game.

- Deception and detection co-vary with capability (Spearman $\rho = 0.89$) yet dissociate in magnitude — an elite detector can be an average deceiver.
- A monitor reading *only* a model's private reasoning identifies the deceiver with AUROC = 1.00; intent leaks in 96% of impostor statements.
- Monitorability is robust to a naive reasoning-hiding instruction (AUROC 1.00 → 0.9999) — the leak is structural, not stylistic; metric saturation frames the open adversarial question.
- Built the full evaluation harness solo: TypeScript / NestJS, OpenRouter, MongoDB, and a live public leaderboard.

PROFESSIONAL EXPERIENCE

Founding Engineer — Figo (cross-border neobank) 2026 – present

- Core backend for a cross-border neobank serving 20k+ users across Africa and LatAm: fiat on/off-ramps, multi-chain wallets, virtual card issuing, and a provider-agnostic embedded-finance API.
- Owned reliability-critical paths — idempotent money movement, cross-chain settlement, fee/quote consistency, and balance-ledger validation.

Founding Engineer — Autospend (stablecoin infrastructure) 2025 – present

- Built an **agentic trading system** that autonomously executes on-chain trades from user-defined rules and market analysis, with MPC-secured transaction signing.
- Designed stablecoin wallet infrastructure and a Web3 execution layer (DeFi protocol integration, event listeners, secure relayers).

Founder / Lead Engineer — Jevan Studio 2026 – present

Independent product & engineering studio

Senior Backend Engineer — Bento (Nigeria's leading payroll platform) 2020 – 2024

- Led the v2 → v3 migration of the payroll automation & payments API; architected scalable services (∼30% efficiency gain) and mentored junior engineers.

Founder — Deg X 2022 – 2023

Cross-chain non-custodial DeFi wallet – 12 blockchain networks, 3 protocols; selected to interview with Axelar Network.

Lead Developer — Tales Consulting 2024 – 2025

Sole-built an institutional communication platform with LMS integration via LTI 1.3.

TECHNICAL SKILLS

LANGUAGES	TypeScript, JavaScript, Python, Solidity, Java
BACKEND	NestJS, Node.js, FastAPI, event-driven & service-oriented architecture, REST APIs
DATA	PostgreSQL, MongoDB, Redis
INFRASTRUCTURE	Docker, Kubernetes, AWS, CI/CD, observability
ML / RESEARCH	LLM evaluation, OpenRouter, experimental design, discrete-event simulation (AnyLogic)
WEB3	Smart contracts, Solidity, MPC signing, multi-chain integration

LANGUAGES

English — fluent · Norwegian — basic (in progress)

CERTIFICATIONS

Software Architecture & Software Design (Coursera, 2023) · API Security Architect, API Designer (API Academy, 2023) · MongoDB M320 Data Modeling (2023) · HNG Backend Finalist (2020)