EcoAgentRAN: Intent-Driven Agentic AI for Energy Efficiency Optimisation in O-RAN

Abdelaziz Salama[1], Mohammed M. H. Qazzaz[1], Yejing Fan[1], Maryam Hafeez[1], Syed Ali Zaidi[1]
1 School of Electronic and Electrical Engineering, University of Leeds, Leeds, UK

MOTIVATION

- **Mobile networks face a growing energy crisis** — and the radio access network is the main consumer, accounting for **over 70%** of total network energy.
- **The opportunity:** traffic varies sharply by time and place, leaving capacity idle yet still burning power — sleeping underutilised RUs can cut RAN energy by **20–40%.**

Why existing approaches fall short:

- **Rule-based scheduling** — static thresholds cannot react to dynamic, event-driven demand.
- **Learning-based control** — requires retraining for every environment and is blind to real-world context: concerts, weather, incidents.
- **The missing piece:** LLM agents can interpret operator intent and reason over live context — but their probabilistic outputs are unsafe for live networks without deterministic guardrails.

Our Solution

EcoAgentRAN — a multi-agent AI framework deployed as rApps in the Non-RT RIC: *Intent & Scenario* • *Traffic Forecasting* • *Energy Efficiency.*

Agents reason, tools act — each agent calls purpose-built tools: a GNN that scores spatial coverage redundancy and adaptive LSTMs that forecast traffic.

Safety by design — agents act autonomously, yet every RU sleep decision is verified to guarantee zero network outage before execution.

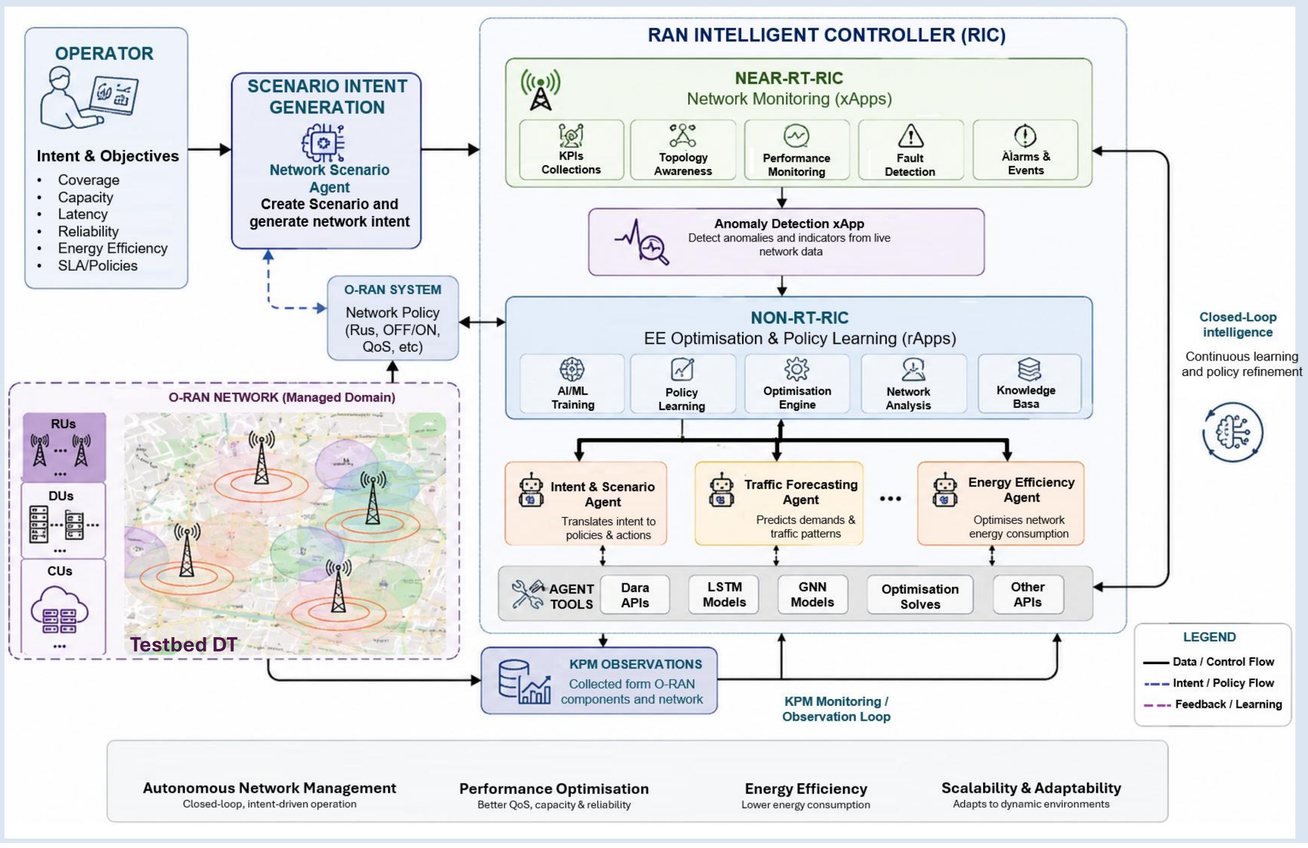
Always watching — an independent anomaly-detection xApp on the Near-RT RIC raises out-of-cycle risk alerts in real time.

METHODOLOGY

Intent & Scenario Agent — translates plain-English operator intents into complete network deployments: location geocoding, real-map topology, and smart RU placement.

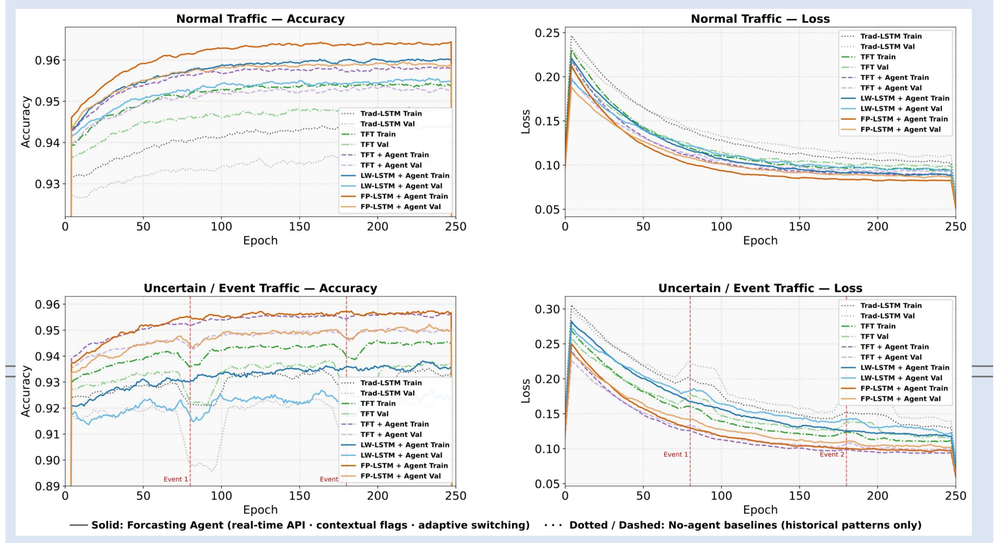
Traffic Forecasting Agent — predicts per-RU load, switching between a Lightweight LSTM (LW-LSTM) and Full-Performance LSTM (FP-LSTM) as traffic conditions change; live context flags block unsafe sleep.

EE Agent — combines GNN coverage-redundancy scores with load forecasts to decide which RUs sleep, actuating decisions via FlexRIC E2 **(Testbed A)** and App Twin ADK **(Testbed B).**



Key Tools:

Adaptive forecasting: switches between lightweight and full performance LSTM models as traffic changes; low cost when stable, full accuracy when uncertain



Forecasting Agent performance against convergence models under normal and event-driven conditions.

GNN spatial reasoning: it learns which RUs are safe to sleep by scoring coverage redundancy across neighbouring cells

Live context awareness: events, weather, and incidents retrieved via real-time APIs and converted by the LLM into sleep-gating flags

EXPERIMENTAL SETUP AND RESULT

Testbed A — Software-defined O-RAN • NVIDIA DGX running OAI 5G, Sionna-based digital twin, and FlexRIC • USRP B210 radio + commercial 5G UE • 1 physical RU extended with virtual RUs

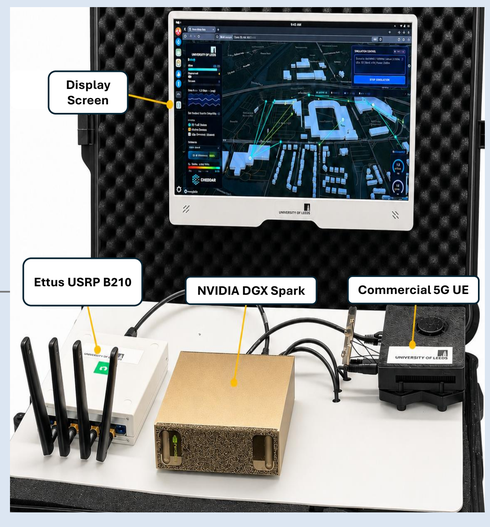
Testbed B — Commercial emulator • VIAVI AI-RSG integrated via App Twin ADK • fully virtual RU deployment • identical Barcelona scenario for direct comparison

Key Achievements

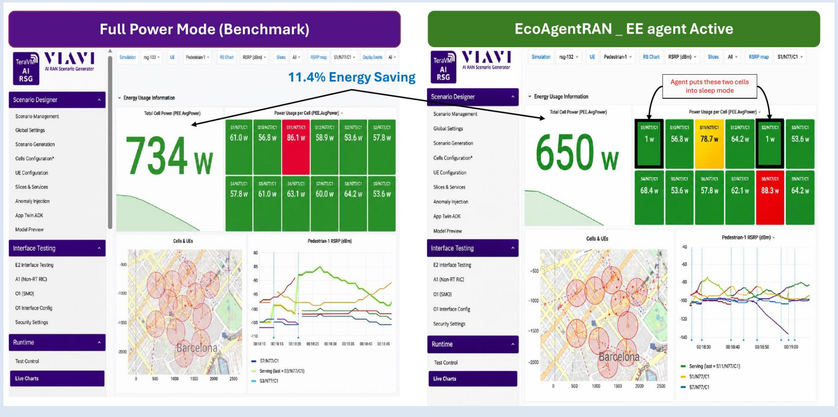
% Energy Saving
11.4%

0% Outage
100% SLA compliance

Testbed A



Testbed B



VIAVI AI-RSG validation — EE Agent sleeps cells S1 & S2: total power **734 W → 650 W (11.4% saving)** with seamless UE re-association and no RSRP degradation

Impact & Significance

Real validation, not just simulation — an agentic AI framework proven on physical O-RAN hardware and an independent commercial emulator
Context beats size — a small 38K-parameter forecasting model with live context awareness outperforms a 126K baseline model without it during real-world events
Generalises without retraining — the same agents ran unchanged and delivered consistent results on both testbeds; code will be released on GitHub upon publication