From Capable to Trustworthy: Systematic Safety Validation for AI-Native Communication Networks

Zeinab Nezami

AI Data Engineering Lead
Open Telco AI
GSMA

RESEARCH PAPER & FRAMEWORK

https://arxiv.org/abs/2604.09682

Framework: https://main.dfe72y7c9xxrm.amplifyapp.com/

Context & Vision

The Promise of AI-Native Networks

O-RAN programmable architectures meet intelligent automation

WHERE WE ARE HEADING

O-RAN architecture: Disaggregates monolithic nodes into open, programmable software components.

xApps on Near-RT RIC: Optimizes resources in real time at millisecond scales.

rApps on Non-RT RIC: Performs long-term policy learning and strategic guidance.

Collaborating Agents: Multi-agent systems balance competing constraints in real time.

THE TECHNOLOGICAL PROMISE

Zero-touch management: End-to-end network operation with zero manual tuning.

Self-optimizing & self-healing: Infrastructure that responds and dynamically scales.

Multi-objective optimization: Balances energy, QoS, and load simultaneously.

Operational advantage: Reduced costs, faster cycles, and higher uptime.

THE CRITICAL CONTEXT

Life-safety backbones: Powers emergency services and medical systems.

Economic engines: Runs financial backbones and transaction routing.

Billions dependent: Critical utility supporting daily worldwide connectivity.

THE FUNDAMENTAL TENSION

The very autonomy and complexity that make AI-native networks powerful also make them hard to validate, predict, and trust.

"Greater capability demands greater accountability. That accountability starts with safety validation."

The Problem

The Safety Gap We Cannot Ignore

Unvalidated automation in critical infrastructure constitutes an active gamble

REAL ECONOMIC COST

£3.7B

Lost by UK businesses to internet failures in 2023 alone

Unreliable network automation is not a theoretical hazard. As documented by **ENISA in 2024**, telecom reliability incidents continue to scale globally.

And this is before introducing autonomous, real-time multi-agent AI systems into production environments.

The Status Quo: "Deploy and Hope"

01 Unvalidated Deployment

AI models are increasingly integrated into critical telecom environments without systematic pre-deployment validation standards.

02 Static Benchmarks Fail

Standard single-agent benchmarks completely miss system-level, emergent failure modes and multi-agent coordination risks.

03 Passive Monitoring!

Testing in labs, deploying, and tracking issues after they strike is unacceptable for networks supporting lives and financial transactions.

"How do we know these AI systems are safe before they manage networks that serve millions?"

The Challenge

The Research Question & Our Approach

Quantifying safety in multi-agent autonomous critical infrastructure

Three Postulates We Set Out to Prove:

01

Safety is Measurable

Not a subjective feeling or a marketing promise. It must be a concrete, quantifiable numerical representation derived from mathematical foundations.

02

Pre-Deployment Validation

Validation must occur thoroughly prior to deployment in live production, never as a reactive strategy after network failures have happened.

03

System-Level Evaluation

Multi-agent communication topologies require holistic coordination evaluation. Single-component benchmarks are fundamentally insufficient.

The Research Consortium

SAFAR Framework supported by the **UK AI Safety Institute** (inaugural research cohort), in partnership with the University of Leeds, GSMA, and AWS.

The Empirical Scope

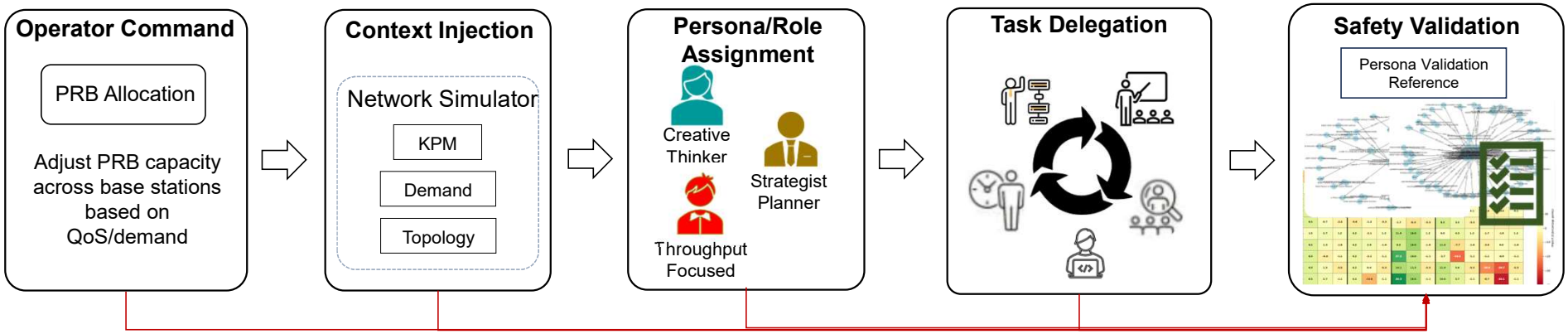
Evaluating **486 persona configurations** systematically across **5 collaborating agents**, running on **2 realistic O-RAN optimization scenarios**.

The Solution

Introducing SAFAR

Safety Assessment Framework for AI in O-RAN

A persona-driven multi-agent framework for O-RAN network automation and a systematic methodology for validating that framework using three-dimensional safety evaluation mechanism.



Transparency by Design

- Every agent decision, refinement cycle, and validation check is logged and traceable
- Enables independent auditing and regulatory compliance verification

Proactive Safety Validation

- Problems detected BEFORE production deployment
- Multi-run evaluation captures stability and convergence dynamics
- Incompatible configurations identified systematically

System Architecture: Agents & Workflow

Modular orchestration, execution, and validation subsystems in SAFAR

ORCHESTRATION

Planner & Coordinator

PlannerAgent: Tree-of-Thoughts reasoning, explores multiple paths, uses GraphRAG over peer-reviewed telecom papers.

CoordinatorAgent: Evaluates and selects optimal solution path, decomposes strategy into execution steps.

EXECUTION

Allocation & Generation

ResourceAllocationAgent: Generates structured pseudocode algorithms, ensures GraphRAG-powered O-RAN compliance.

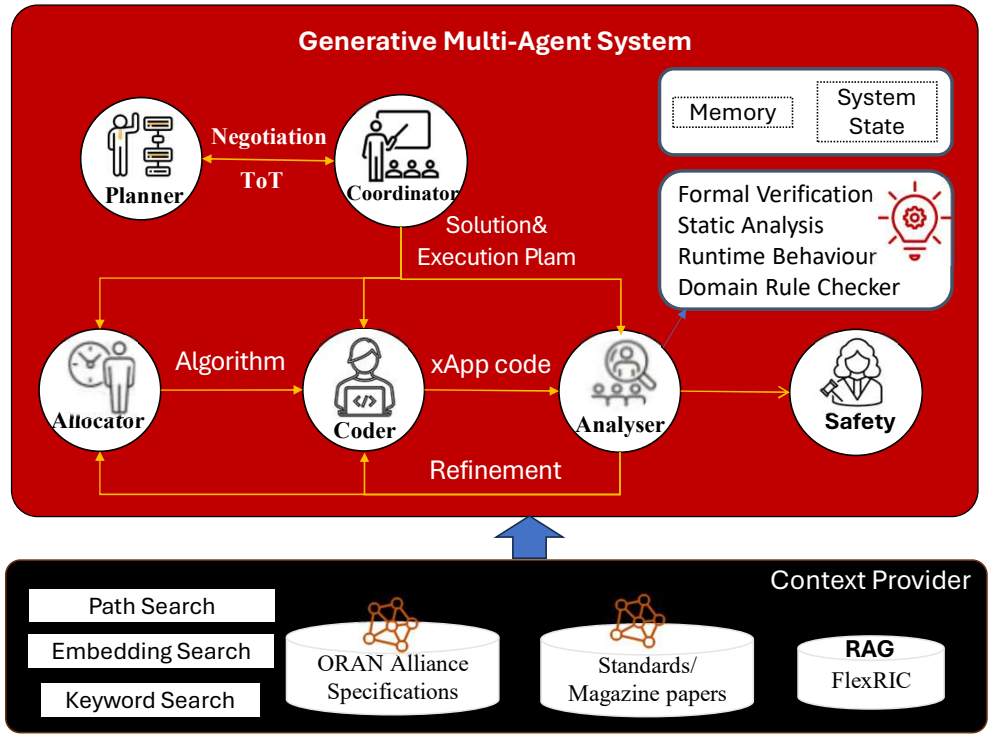
CodeAgent: Generates executable xApp code, utilizes RAG indexing over FlexRIC codebase.

VALIDATION

Analyser & Evaluator

AnalyserAgent: Central validation authority. Evaluates ALL upstream agent outputs simultaneously.

Consistency Checking: Validates code quality, runtime bounds, structural format, and policy guidelines.



Scientific Methodology

Three-Dimensional Evaluation Framework

A comprehensive strategy derived from decision theory models

01 Normative Evaluation

Does it achieve optimal outcomes?

Grounded in **normative decision theory**.

Continuous 0-1 evaluation of outputs using automated structured reasoning judges.

Each persona defines explicit optimality criteria.

02 Prescriptive Evaluation

Does it follow guidelines?

Persona Consistency: stable behavior cross-scenarios.

Linguistic Habits: formality, tone, O-RAN jargon.

Ethical Compliance: boundaries, transparency, fairness.

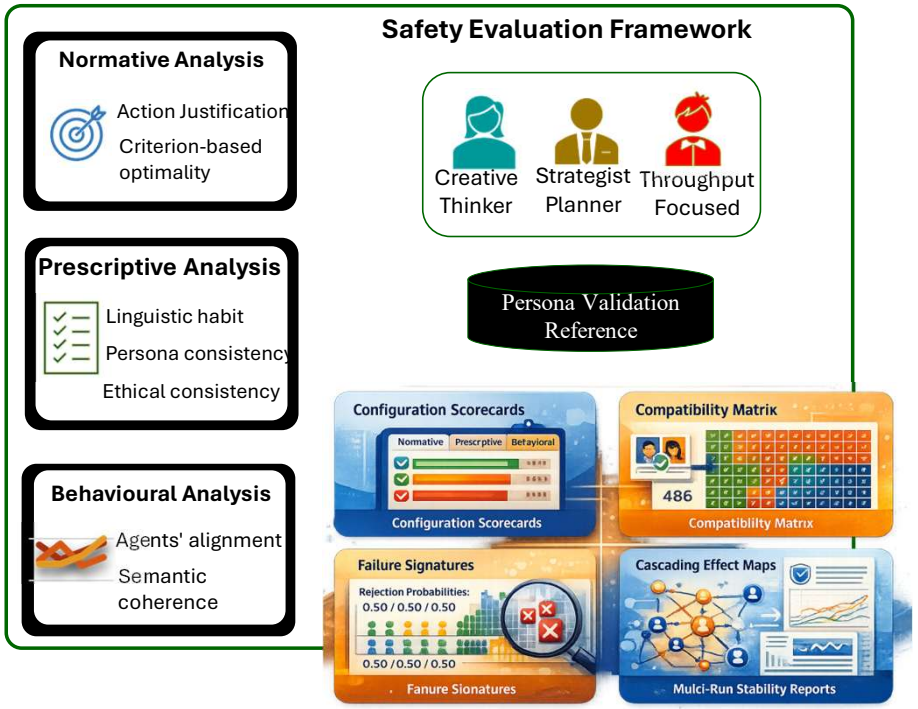
03 Behavioral Evaluation

How does the system evolve?

Semantic Convergence: stability of solutions across runs.

Knowledge Expansion: concept enrichment vs. hallucinated drift.

Code Dynamics: structure, policy, and formal safety guarantees over cycles.



Key Research Findings

Finding 1: Alignment Matters

Task-dependent agent hierarchies, persona variations, and structural retrieval constraints

NORMATIVE ALIGNMENT SWEEP

Task-Dependent Performance

Q1 — SINGLE-OBJECTIVE

Planner & CodeAgent

μ = 0.81

Q2 — MULTI-OBJECTIVE

Allocator Degradation

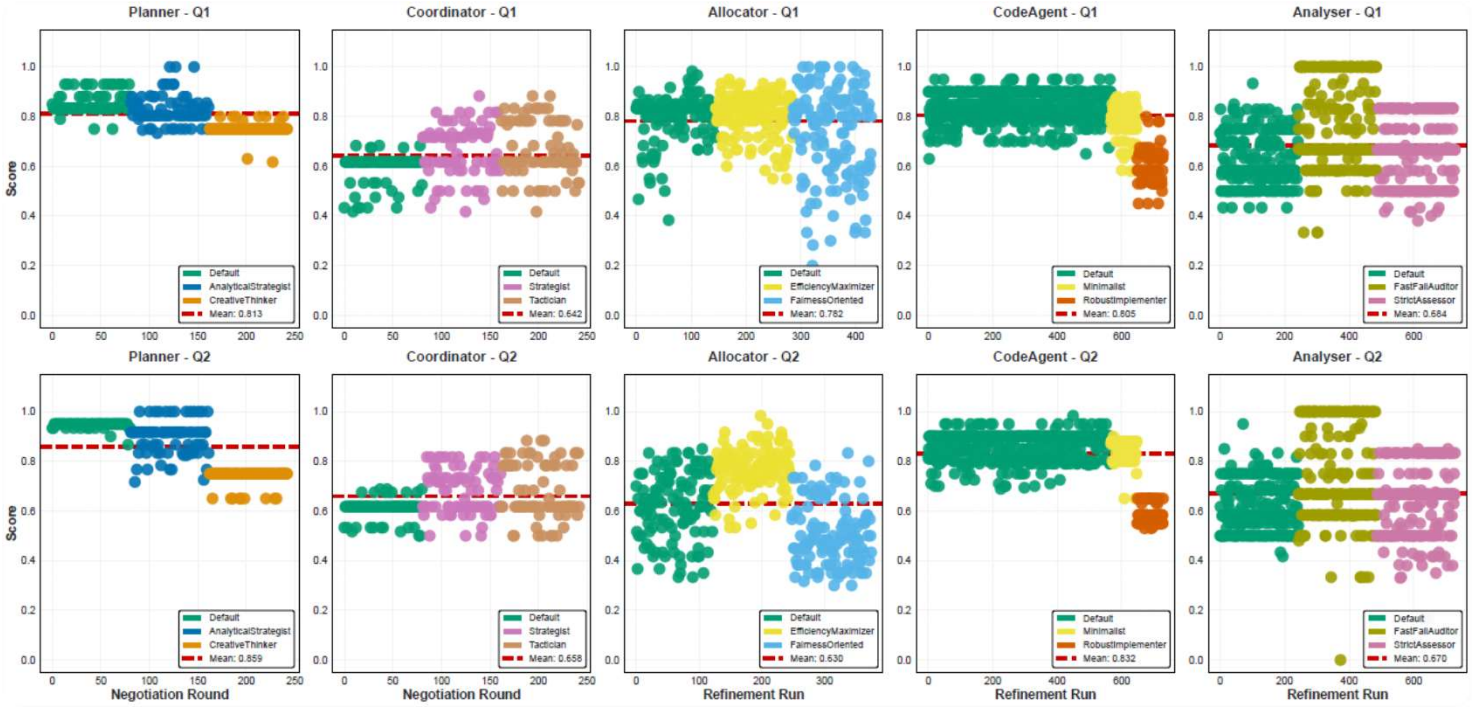
-19% to 0.63

PERSONA & RETRIEVAL LIMITS

Persona Variations & Retrieval Trade-offs

Within-Agent Persona Variance (Up to 53%)

- **Allocator**: EfficiencyMaximizer (μ = 0.77) outscores FairnessOriented (μ = 0.50) by 53% on Q2.
- **Coder**: RobustImplementer scores 28–32% lower than Default (defensive coding limits optimality).
- **Planner**: Default (μ_Q2 = 0.92–0.95) outscores specialized personas by 19–22%.



Frontier LLMs have sufficient O-RAN knowledge for single tasks, but fail multi-objective reasoning.

Key Research Findings

Finding 2: Pre-Deployment Validation Works

Multi-run trajectory and early warning failure signatures detected in testing phase

CODE QUALITY IMPROVEMENT

Evolution Across Iterative Runs

RUN 1 (SNAPSHOT)

72.5%

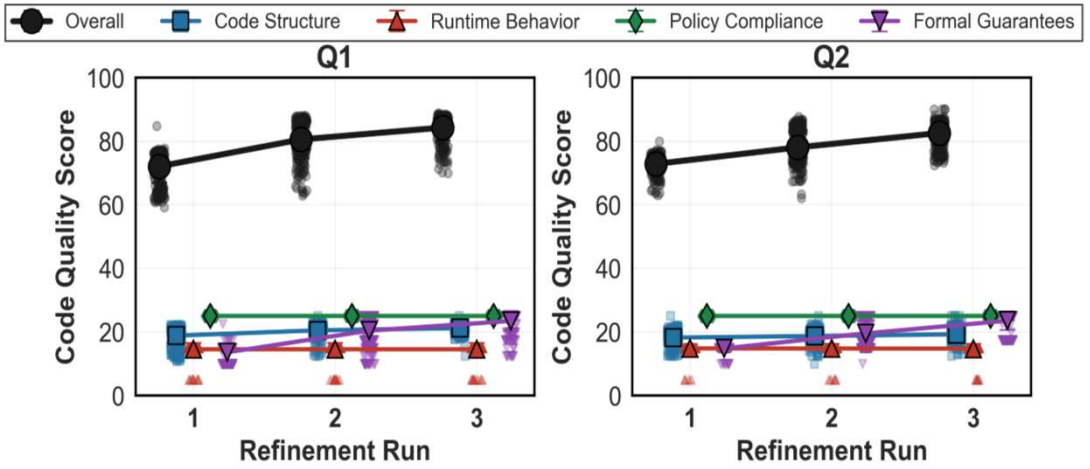
RUN 3 (REFINED)

83.4%

Early Warning Safety Signatures

Formal Guarantees (Type Safety, Errors)	**56.9% → 94.0% (+65%)**
Code Structure Quality	**74.1% → 81.0% (+9%)**
Policy Compliance (O-RAN Specs)	**100% (From Run 1)**
Runtime Behavior Bottleneck	**58.9% (Requires optimization)**

Generated Code Quality



Ecosystem Recommendations

Ecosystem Alignment: Proposed Validation Guidelines

How research-driven frameworks can support open telco-grade AI initiatives

01. Multi-Scenario Validation

We recommend demonstrating AI model performance across a broad operational envelope, rather than relying solely on benign, static lab conditions. (SAFAR: 28.2% performance spread).

02. System-Level Evaluation

Validation frameworks should assess multi-agent coordination dynamics. Standalone benchmarks can be highly misleading for interconnected setups. (SAFAR: +22.4% local vs -1.25% system impact).

03. Multi-Run Stability Testing

Ecosystem participants should track model convergence and behavioral stability across multiple cycles rather than relying on single-run snapshot benchmarks. (SAFAR: 15% quality delta).

04. Early Failure Detection

Establishing guidelines to detect systematic behavioral anomalies such as persona abandonment (0.50 prescriptive score) enables early intervention.

GSMA Open Telco AI focuses on establishing open, telco-grade foundational models and capability benchmarks. Our SAFAR research proposes safety validation methodologies to support a reliable, trusted, and robust telecom ecosystem.

Research Proposal

Conclusion

The Future of AI-Native Networks We Can Trust

"AI safety for critical infrastructure is not a philosophical question, it is quantifiable, measurable, and actionable."

The Target Future

Pre-Validated Deployments: Every system evaluated across realistic network situations before going live.

Open Benchmarks: frameworks accessible to all operators globally, including emerging markets.

Measurable Coordination: Dynamic interaction made as certifiable as latency, throughput, or "five nines" reliability.

Auditability: Safety acts as a proven contract metric, not an unverified vendor promise.

The Open Problems — Join Us

Predictive Models: Can we predict system-wide impact before executing all 486 configurations?

Dynamic Adaptation: Can agents dynamically adjust configurations in response to active load spikes?

Topological Scalability: Scaling evaluations to ultra-dense real-world networks.

Cross-Domain Generalization: Exporting the SAFAR model to the Core and Transport.

Building the Foundations of Telco Grade AI

https://www.open-telco.ai/

GSMA